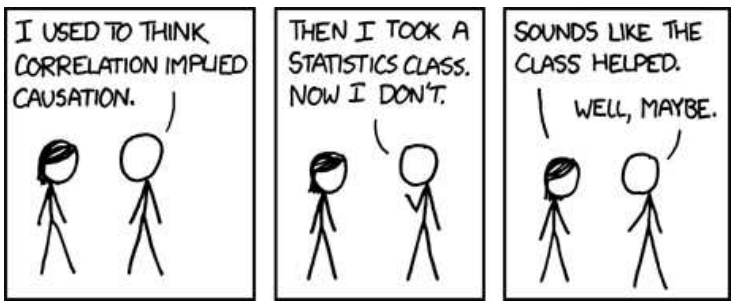
ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

INSTITUT POLYTECHNIQUE DE PARIS

Statistiques et Analyse de Données

An Introduction to Mathematical Statistics and Data Analysis



Julien Reygner

Année universitaire 2025–2026

Avant-propos

Le cours se déroule intégralement en petite classe, selon la répartition suivante.

- Groupe 1 : Loucas Pillaud-Vivien, École des Ponts.
- Groupe 2 : Julien Reygner, École des Ponts.
- Groupe 3 : Geneviève Robin, Université Paris Cité.
- Groupe 4 : Antonin Della Noce, École des Ponts.
- Groupe 5 : Guillaume Perrin, Université Gustave Eiffel.

Toutes les informations relatives au cours sont disponibles sur la page Educnet consacrée :
https://educnet.enpc.fr/course/view.php?id=595

Le programme détaillé des séances est donné p. v.

Exercices obligatoires

Pour chaque séance de cours, il vous est demandé de préparer un ou plusieurs exercices **obligatoires** à l'avance (signalés par le pictogramme dans le programme des séances). Ces exercices portent sur le cours de la séance précédente et sont corrigés en classe. Il est tout-à-fait permis, et même conseillé, d'y travailler à plusieurs, car échanger avec vos camarades sur les difficultés et éventuels problèmes de compréhension de chacun est le meilleur moyen d'assimiler en profondeur le contenu du cours. Il est cependant demandé que chacun rédige sa propre copie, qui peut être **ramassée et notée**, et dont l'évaluation compte dans la note de module finale. Certains exercices peuvent comporter une ou plusieurs questions à résoudre numériquement, soit directement, soit en utilisant un Notebook mis en ligne sur Educnet. Sauf mention explicite du contraire, on ne demande pas d'envoyer le code ou les résultats numériques au chargé de petite classe : le report des valeurs numériques obtenues, une rapide recopie de l'allure des graphes tracés ou une description sommaire du code implémenté suffisent.

Classes inversées

Une spécificité de ce cours est l'organisation de leçons sous la forme de **classes inversées**. Le principe est le suivant : **vous** ferez le cours à vos camarades. Six sujets, répartis dans les notes et résumés dans l'Appendice B, devront être présentés par des groupes de 3 ou 4 élèves, au cours d'un exposé avec slides d'une vingtaine de minutes, suivi d'une discussion d'une dizaine de minutes avec le reste de la classe.

Chaque exposé présente une méthode statistique particulière, et devra comporter deux parties :

- une présentation théorique rapide de la méthode, guidée par les indications données dans les sections concernées — n'hésitez pas à faire vos propres recherches en dehors du poly ;
- la démonstration de l'application de la méthode à un exemple **différent de celui ou ceux traités dans le poly**, issu d'une base de données réelle ou simulée.

Illustration de couverture par Randall Munroe, http://xkcd.com/552.

La seconde partie ne doit pas se limiter à un simple énoncé d'un résultat expérimental, mais doit s'accompagner d'une discussion. Afin de préparer sereinement ces exposés, le calendrier suivant est mis en place :

- pour la Séance #3, constitution des groupes et choix des sujets ;
- On demande à chaque groupe d'envoyer un e-mail à son chargé de petite classe **au moins trois semaines avant la date de l'exposé**, avec le programme envisagé pour l'exposé, ainsi que la ou les bases de données qui seront utilisées pour illustrer la méthode. Le chargé de petite classe répondra par e-mail en validant le programme ou en proposant des modifications ;

L'envoi de l'e-mail est le point de départ d'un échange régulier avec votre chargé de petite classe permettant de valider le contenu de votre présentation. Cette phase de préparation est spécifiquement évaluée dans la note de l'exposé. Quelques conseils supplémentaires pour la préparation des exposés sont donnés au début de l'Appendice B. Les méthodes présentées au cours de ces exposés **seront au programme de l'examen final** : gardez donc à l'esprit que vos camarades attendront de vous le même effort de pédagogie que celui que vous exigez de vos professeurs !

Modalités d'évaluation

La note de module se décompose de la manière suivante :

- 2 points de contrôle continu pour les exercices obligatoires ;
- 5 points pour l'exposé de classe inversée, dont 1 point pour les échanges par e-mail pour préparer l'exposé ;
- 13 points pour l'examen final.

Cours en anglais

Les notes de cours et l'intégralité des documents pédagogiques liés au cours sont rédigés en anglais, afin de sensibiliser les étudiants à la pratique de l'anglais scientifique, ce qui sera très utile pour ceux qui souhaitent partir un semestre ou une année à l'étranger. Nous conseillons de préparer les slides — pardon, les transparents — des exposés de classe inversée en anglais, mais de conserver le français pour l'exposé oral (dans les petites classes francophones). Enfin, les étudiants sont libres de choisir le français ou l'anglais pour rédiger leur copie à l'examen final (dont le sujet est distribué dans les deux langues).

Certains termes spécifiques n'ont pas forcément de correspondance français/anglais : les enseignants de petite classe sont là pour éclairer les possibles ambiguïtés que cela pourrait induire. À titre d'exemple, soulignons dès maintenant une subtilité classique : les termes *positive* et *negative* en anglais désignent respectivement des nombres strictement positifs et strictement négatifs ; pour parler d'un nombre positif ou nul, on emploie le terme *nonnegative* — et évidemment *nonpositive* pour les nombres négatifs ou nuls. La même règle s'applique à la monotonie des fonctions : *increasing* et *decreasing* désignent respectivement des fonctions strictement croissantes et strictement décroissantes, pour les fonctions croissantes et décroissantes au sens large, on utilise *nondecreasing* et *nonincreasing*. Une autre différence notable entre les conventions anglophone et francophone concerne l'emploi d'une parenthèse pour représenter la borne ouverte d'un intervalle : ce que l'on noterait $[0, 1[$ en français est noté $[0, 1)$ en anglais. Enfin, précisons que nous prenons la convention de noter $\log x$, et non $\ln x$, le logarithme naturel de $x > 0$.

Polycopié

Ce polycopié est destiné aux élèves de l'École des Ponts. Une version électronique est disponible sur Educnet, et nous sommes généralement très heureux de la transmettre à tout étudiant à qui elle pourrait être utile. Nous vous remercions cependant de ne pas diffuser la version électronique sur des pages web publiques.

Foreword

The course is taught in small classes, according to the following distribution.

- Group 1: Loucas Pillaud-Vivien, École des Ponts.
- Group 2: Julien Reygner, École des Ponts.
- Group 3: Geneviève Robin, Université Paris Cité.
- Group 4: Antonin Della Noce, École des Ponts.
- Group 5: Guillaume Perrin, Université Gustave Eiffel.

All the important information related to the course is available on the Educnet page:
https://educnet.enpc.fr/course/view.php?id=595

The detailed programme of the sessions is given p. v.

Mandatory exercises

For each class session, you are required to prepare one or more **mandatory** exercises in advance (indicated by the 🏠 symbol in the session programme). These exercises focus on the material covered in the previous Lecture and will be corrected in class. It is entirely allowed, and even encouraged, to work on them together, as discussing difficulties and potential comprehension issues with your classmates is the best way to deeply assimilate the course content. However, it is requested that each student writes their own copy, which may be **collected and graded**, and the evaluation of which is part of the final module grade. Some exercises may contain questions which must be solved numerically, either directly or using a Notebook available on Educnet. Unless it is explicitly stated otherwise, you are not expected to send your codes or numerical results to your lecturer: it is enough to report on your exercise sheet your numerical results, shape of graphs or a short description of your code.

💬 Flipped classrooms

A distinctive feature of this course is the use of **flipped classrooms**. The idea is the following: **you** will teach the course to your classmates. Six topics, distributed throughout the notes and summarized in Appendix B, must be presented by groups of 3 or 4 students, in a 20-minute slide presentation followed by about 10 minutes of discussion with the rest of the class.

Each presentation introduces a statistical method, and will have to contain two parts:

- a theoretical presentation of the method, guided by indications given in the notes — do not hesitate to do your own research as well;
- the illustration of the method on an example **which must not be the same as in the lecture notes**, coming from real or simulated data.

The illustrative part must not be limited to a simple statement of the result of the test, but it must be accompanied by a discussion. To ensure smooth preparation of these presentations, the following schedule has been established:

- By Session #3, formation of groups and selection of topics;
- Each group must send an email to their teaching assistant **at least three weeks before the presentation date**, with the planned outline of the talk and the datasets that will be used to illustrate the method. The teaching assistant will reply by email to validate the plan or suggest changes;

Sending this email marks the start of a regular exchange with your teaching assistant, allowing you to refine and validate the content of your presentation. This preparation phase is explicitly assessed as part of the presentation grade. Additional advice for preparing the talks can be found at the beginning of Appendix B. The methods presented in these talks **will be included in the final exam syllabus**: keep in mind that your classmates will expect from you the same level of pedagogy that you expect from your instructors!

Final mark

The final mark is out of 20 points, and decomposed as follows:

- 2 points for mandatory exercises;
- 5 points for the flipped classroom presentation, including 1 point for the email exchange in preparation;
- 13 points for the final exam.

About these notes

These lecture notes are intended for students of École des Ponts. A PDF version is available on Educnet, and we are generally happy to send it to any student to whom it may be useful. However, we kindly thank you for not distributing it on public webpages.

Program of the course

These notes are divided into 11 Lectures, which roughly correspond to the contents of the 11 first sessions of the course. Each Lecture is organised as follows.

- The body of the Lecture corresponds to the content of the course which must be studied in class. It contains short exercises, which generally consist in a straightforward application of concepts which have just been introduced, and may be solved during the class.
- At the end of the Lecture, different exercises are proposed: training exercises can be made in class at the end of the session; homework is the **mandatory** work to do for the next session; supplementary exercises allow you to go into certain aspects of the course in more depth or provide an introduction to more advanced topics in statistics. The correction of all exercises (except for mandatory homework) is available in Appendix D.
- Parts of some Lectures are signaled by an asterisk: they are not examinable, but you may find them helpful in a future course or job.
- Likewise, the sections marked with the symbol 💬 will be treated during the last two sessions of the course, in flipped classroom format. They are examinable.

There is a specific calendar to respect for the preparation of the flipped classroom sessions. The important steps are indicated below with the icon 💬. Part of the last two Lectures will also be dedicated to revisions for the preparation of the final exam.

In the following detailed program, the main notions to be addressed during each Lecture are listed. All along the semester, do not hesitate to refer to this summary and check whether everything is clear for you. The homework program is also indicated with the icon 🏠.

Session #1. Sept. 26th

🎓 Lecture 1: What you must know in Probability Theory.

• Reminder on abstract random variables: law, expectation and covariance, independence. • Discrete variables. Bernoulli, Binomial, Geometric and Poisson distributions • Random variables with density. Uniform, Exponential and Gaussian variables. CDF and quantiles. Characteristic function. Gaussian vectors. • Different notions of convergences. Law of Large Numbers and Central Limit Theorem. • Handling random variables with SciPy.

Session #2. Oct. 3th

🏠 Homework: Exercise 1.A.7. **There is chocolate to win!**

💬 **Read Appendix B** and choose your group and of subject for flipped classrooms presentations.

🎓 Lecture 2: Parametric Model and Estimators.

• Parametric model, statistic, estimator. Bias, MSE, bias-variance decomposition. • Consistency, asymptotic normality, Delta method. The method of moments.

Session #3. Oct. 10th

Homework: Exercise 2.A.3.

Lecture 3: Statistics in Gaussian Models.

- Gaussian vectors and the Cochran Theorem.
- Chi-square and Student distributions. Joint law of empirical mean and variance.
- Linear regression with Gaussian errors.

Session #4. Oct. 17th

Homework: Exercises 3.A.2 and 3.A.3.

Lecture 4: Confidence Intervals.

- General definitions.
- Construction of exact confidence intervals by the pivotal function method, examples in the Gaussian and Exponential models. Confidence intervals and linear regression in the Gaussian model.
- Construction of asymptotic confidence intervals when an asymptotically normal estimator is available.
- Construction of approximate confidence intervals with concentration inequalities. The Bienaymé–Chebychev and Hoeffding inequalities.

Session #5. Oct. 24th

Homework: Exercises 4.A.3 and 4.A.4.

Lecture 5: Maximum Likelihood Estimation.

- Likelihood, log-likelihood, MLE. Contrast and M-estimators. Examples of the Bernoulli, Exponential, Gaussian and Uniform models.
- Regular model, score, Fisher information. Cramér–Rao bound, efficiency and asymptotic efficiency.
- Formal justification of the asymptotic efficiency of the MLE in regular models.

Session #6. Nov. 14th

Homework: Exercise 5.A.3 and questions 1 to 4 in Exercise 1 of the Revision Sheet 6.2.3.

Lecture 6: Introduction to Bayesian Estimation.

- Prior and posterior distributions. The Beta-Bernoulli example. Conjugate priors.
- PM and MAP. Consistency of Bayesian estimation, sufficient condition.
- Credibility regions.

Flipped classrooms #1: The EM Algorithm for Mixture of Gaussians

Session #7. Nov. 28th

Homework: Exercises 6.A.3 and 6.A.4.

Lecture 7: The Formalism of Statistical Hypothesis Testing.

- Null and alternative hypotheses, rejection region. Type I and type II risks. Level and power. Consistency and asymptotic power. Test statistic and p-value.
- Duality with confidence intervals. Wald's test.
- The look-elsewhere effect, FWER and the Bonferroni correction for multiple testing.

Session #8. Dec. 12th

Homework: Exercise 7.A.2.

Please take your laptop.

Lecture 8: Tests in the Gaussian Model.

- One-sample tests for μ and σ^2.
- Two-sample Student and Fisher tests.
- Tests in the linear regression setting.

Flipped classrooms #2: The Likelihood Ratio Test

Session #9. Dec. 19th

Homework: Exercise 8.A.2.

Lecture 10: χ_2 Tests for Finite State Spaces.

- χ_2 distance on a finite state space. Asymptotic behaviour of the empirical measure.
- χ_2 test of goodness-of-fit to a single distribution and to a parametric family.
- χ_2 tests of independence and homogeneity.

Flipped classrooms #3: The Analysis of Variance in the Gaussian Model

Session #10. Jan. 09th

Homework: Exercise 10.A.3

Lecture 11: Kolmogorov Tests for Nonparametric Models

- Empirical CDF. The Glivenko–Cantelli and Donsker Theorems.
- Asymptotic and nonasymptotic Kolmogorov tests for goodness-of-fit. The Lilliefors correction for goodness-of-fit to a parametric family. Application to the Exponential model.
- Kolmogorov–Smirnov test of homogeneity.

Flipped classrooms #4: Walsd's Test for the Identity of Means

Session #11. Jan. 16th

Homework: Exercise 11.A.4

Flipped classrooms
#5: The Shapiro–Wilk Test
#6: The Kolmogorov–Smirnov Test of Homogeneity

Start Final Revision Sheet p. 153.

Session #11. Jan. 23th: Final exam

- Handwritten notes and printed lecture notes are allowed.
- Other sources (textbooks, lecture notes from other courses) are forbidden.

Contents

Lecture 1

What you must know in Probability Theory

Contents

This Lecture serves as a summary of the notions in measure and probability theory which will be useful during the course. We refer to [2] for an exhaustive introduction. It also contains an introduction to the main commands to deal with random variables in SciPy.

1.1 Abstract random variables

1.1.1 Random variables

A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$ such that Ω is a set, $\mathcal{A}$ is a σ-field[1] and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{A})$.

Exercise 1.1.1. Take a few minutes to make sure that you remember the definitions of a σ-field and of a probability measure.

Given such a space, and a measurable space $(E, \mathcal{E})$, a *random variable* with values in E is a measurable function $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$. The *law* (or equivalently the *distribution*) of X is the probability measure P on $(E, \mathcal{E})$ defined by

$$\forall C \in \mathcal{E}, \qquad P(C) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in C\}).$$

The event $\{\omega \in \Omega : X(\omega) \in C\}$ is often simply denoted by $\{X \in C\}$. Given a probability measure P on E, the notation $X \sim P$ means that P is the law of X.

For any $a \in E$, the *Dirac measure* in a is the probability measure δ_a defined by $\delta_a(C) = \mathbb{1}_{\{a \in C\}}$. A random variable X is called *deterministic* if there exists $a \in E$ such that the law of X is δ_a, that is to say if $X = a$, $\mathbb{P}$-almost surely.

We shall mostly work with two kinds of spaces $(E, \mathcal{E})$.

[1] *Tribu* en français.

- If E is a finite or countably infinite set, then we shall always consider that $\mathcal{E}$ is the set of all subsets of E and call the set $(E, \mathcal{E})$ *discrete*. In this case, all random variables X and probability measures P on E will also be called *discrete*.
- If $E = \mathbb{R}^d$ with $d \geq 1$, then we shall always consider that $\mathcal{E}$ is the Borel σ-field on $\mathbb{R}^d$, which we denote by $\mathcal{B}(\mathbb{R}^d)$.

1.1.2 Expectation and (co)variance

When $E = \mathbb{R}$, the *expectation* of the random variable $X \sim P$ is the Lebesgue integral

$$\mathbb{E}[X] := \int_{\omega \in \Omega} X(\omega) \mathrm{d}\mathbb{P}(\omega) = \int_{x \in \mathbb{R}} x \mathrm{d}P(x),$$

which is well-defined as soon as X is integrable, which we denote by the condition $\mathbb{E}[|X|] < +\infty$. If $\mathbb{E}[X^2] < +\infty$, the *variance* of X is defined by

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

The square root of variance is called *standard deviation*[2].

Lemma 1.1.2 (Jensen inequality)**.** *Let X be a random variable with values in $\mathbb{R}$, such that $\mathbb{E}[|X|] < +\infty$, and let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then*[3]

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Applying this identity with $\phi(x) = x^2$ shows in particular that the condition $\mathbb{E}[X^2] < +\infty$ ensures that $\mathbb{E}[X]$ is well-defined in the definition of $\mathrm{Var}(X)$. More generally, notice that for any $X \in \mathbb{R}$ and $p \geq 1$,

$$\mathbb{E}[|X|] \leq \mathbb{E}[|X|^p]^{1/p}.$$

Exercise 1.1.3. What is the variance of a deterministic random variable?

The *covariance* between two random variables X and Y such that $\mathbb{E}[X^2] < +\infty$ and $\mathbb{E}[Y^2] < +\infty$ is defined by

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

so that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. Notice also the identity

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).$$

When $E = \mathbb{R}^d$ and $X = (X^1, \ldots, X^d)$, we denote by $\mathbb{E}[X]$ the vector $(\mathbb{E}[X^1], \ldots, \mathbb{E}[X^d])$ if these numbers are well-defined. If $\mathbb{E}[\|X\|^2] < +\infty$, the covariance matrix of this random vector is the symmetric $d \times d$ matrix $\mathrm{Cov}[X]$ with coefficients

$$\forall i, j \in \{1, \ldots, d\}, \qquad \mathrm{Cov}[X]_{i,j} = \mathrm{Cov}(X^i, X^j),$$

and its variance is

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

Notice that the covariance matrix of X directly writes

$$\mathrm{Cov}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^\top\right] = \mathbb{E}\left[XX^\top\right] - \mathbb{E}[X]\mathbb{E}[X]^\top.$$

It follows from the next exercise that covariance matrices are *nonnegative*, in the sense that they satisfy $\langle u, \mathrm{Cov}[X]u \rangle \geq 0$ for any $u \in \mathbb{R}^d$.

[2] *Écart-type* en français.

[3] In this statement, we do not assume that $\phi(X)$ is integrable, but this inequality is obviously true if $\mathbb{E}[\phi(X)] = +\infty$.

Exercise 1.1.4 (Properties of covariance matrices)**.** Let $X \in \mathbb{R}^d$ be such that $\mathbb{E}[\|X\|^2] < +\infty$ and set $m = \mathbb{E}[X]$, $K = \mathrm{Cov}[X]$.

1. For any $u \in \mathbb{R}^d$, show that $\mathrm{Var}(\langle u, X\rangle) = \langle u, Ku\rangle$.

2. For any $A \in \mathbb{R}^{k\times d}$ and $b \in \mathbb{R}^k$, compute $\mathbb{E}[AX + b]$ and $\mathrm{Cov}[AX + b]$.

1.1.3 Independence

Definition 1.1.5 (Independence of random variables)**.** *A finite family of random variables $(X_1, \dots, X_n)$, which need not take their values in the same measurable space, is* independent *if for any measurable subsets $(C_1, \dots, C_n)$,*

$$\mathbb{P}(X_1 \in C_1, \dots, X_n \in C_n) = \mathbb{P}(X_1 \in C_1) \cdots \mathbb{P}(X_n \in C_n),$$

or equivalently if the joint law P of $(X_1, \dots, X_n)$ is the product measure $P_1 \otimes \cdots \otimes P_n$ of the respective marginal distributions $P_1, \dots, P_n$ of $X_1, \dots, X_n$.

An infinite family of random variables is independent if any finite subfamily is.

If $X_1, \dots, X_n$ are independent, for any measurable functions $f_1, \dots, f_n$ such that $f_1(X_1), \dots, f_n(X_n)$ are integrable, we have $\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)]$.

Exercise 1.1.6. Show that if $X, Y \in \mathbb{R}$ are independent then $\mathrm{Cov}(X, Y) = 0$, which then implies $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. What about the converse statement?

1.1.4 Correlation

Let X, Y be real-valued random variables, with finite variance. By the Cauchy–Schwarz inequality,

$$\begin{aligned} |\mathrm{Cov}(X, Y)| &= |\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]| \\ &\le \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} = \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}, \end{aligned}$$

which leads to the following definition.

Definition 1.1.7 (Correlation coefficien)**.** *The correlation coefficient between X and Y is*

$$\rho_{X,Y} := \begin{cases} \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1, 1] & \text{if } \mathrm{Var}(X)\,\mathrm{Var}(Y) > 0, \\ 0 & \text{otherwise}. \end{cases}$$

If X and Y are independent then $\rho_{X,Y} = 0$, but the converse is not true, see Exercise 1.1.6. A positive correlation between two variables implies that they typically take simultaneously large or small values; a negative correlation implies that they typically vary in opposite directions; and a correlation close to 0 implies that there is no linear relation between the variations of these variables. This does not mean that if two variables are positively correlated, then there must be a causal relation between them: for example, in seaside cities, the number of ice creams sold per day and the number of sunburns are positively correlated, but icecreams do not cause sunburns (nor are sunburns treated with icecreams). The reason for positive correlation here is the existence of a latent variable (the weather) which is causal to both observed variables.

1.2 Discrete random variables

If E is discrete, any probability measure P on E is characterised by the family of numbers $(p(x))_{x\in E}$ defined by $p(x) = P(\{x\}) = \mathbb{P}(X = x)$ for $X \sim P$. This family is called the *probability mass function* of P.

We recall the definition of elementary discrete distributions.

Definition 1.2.1 (Bernoulli distribution)**.** *A random variable X which takes its values in $\{0, 1\}$ is called a* Bernoulli variable*. Its* parameter *is $p = \mathbb{P}(X = 1)$, and it characterises its law since then $\mathbb{P}(X = 0) = 1 - p$. We write $X \sim \mathcal{B}(p)$.*

Bernoulli variables allow to model the outcome of binary experiments, such as coin tossing. In particular, for any event $A \in \mathcal{A}$, the random variable $\mathbb{1}_A$ defined by

$$\forall \omega \in \Omega, \qquad \mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases}$$

is a Bernoulli random variable with parameter $p = \mathbb{P}(A)$.

Exercise 1.2.2. If $X \sim \mathcal{B}(p)$, compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

Let $(X_i)_{i\geq 1}$ be a sequence of independent Bernoulli variables with parameter p.

Definition 1.2.3 (Binomial distribution)**.** *For any $n \geq 1$, the law of the random variable $S_n = X_1 + \cdots + X_n$ is called the* binomial distribution *with parameters (n, p) and denoted by $\mathcal{B}(n, p)$. It takes its values in $\{0, \ldots, n\}$ and satisfies*

$$\forall k \in \{0, \ldots, n\}, \qquad \mathbb{P}(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Exercise 1.2.4. If $S_n \sim \mathcal{B}(n, p)$, compute $\mathbb{E}[S_n]$ and $\mathrm{Var}(S_n)$.

Definition 1.2.5 (Geometric distribution)**.** *If $p > 0$, the random variable $T = \inf\{i \geq 1 : X_i = 1\}$ is almost surely finite. Its law is called the* geometric distribution *with parameter p, and is denoted by $\mathcal{G}(p)$. It takes its values in $\mathbb{N}^* = \{1, 2, \ldots\}$ and satisfies*

$$\forall i \geq 1, \qquad \mathbb{P}(T = i) = p(1 - p)^{i-1}.$$

The equivalent identity $\mathbb{P}(T \geq i) = (1 - p)^{i-1}$ is also often useful.

Exercise 1.2.6. If $T \sim \mathcal{G}(p)$, compute $\mathbb{E}[T]$ and $\mathrm{Var}(T)$.

As is clear from their definitions, binomial and geometric variables respectively allow to model the number of successes in a set of n independent experiments (how many Tails you get when you toss n times a coin) and the number of experiments after which you observe the first success (how many times you have to toss the coin to get a Tail).

Definition 1.2.7 (Poisson distribution)**.** *For $\lambda > 0$, the Poisson distribution with parameter λ, denoted by $\mathcal{P}(\lambda)$, is the law of a random variable $N \in \mathbb{N} = \{0, 1, \ldots\}$ such that*

$$\forall k \geq 0, \qquad \mathbb{P}(N = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}.$$

This distribution must be understood as an approximation of the binomial distribution when p is small and n is large, as it is a stated in the next exercise.

Exercise 1.2.8. Show that if $p \to 0$ and $n \to +\infty$ with $np \to \lambda$, then for any $k \geq 0$, $\mathbb{P}(S_n = k) \to \mathbb{P}(N = k)$.

To illustrate this approximation, consider a fisherman fishing fish[4] in a lake. The lake contains $n \gg 1$ fish, each of which has a small probability $p \ll 1$ to take the bait, independently from each other. Then, by definition of the model, the number of fish caught by the fisherman follows the binomial distribution with parameters (n, p), but in practice this number can be modelled by a Poisson variable with parameter np.

Exercise 1.2.9. If $N \sim \mathcal{P}(\lambda)$, compute $\mathbb{E}[N]$ and $\mathrm{Var}(N)$.

1.3 Random variables with density

1.3.1 Definition

Definition 1.3.1. *A random variable $X \in \mathbb{R}^d$ is said to have a density p with respect to the Lebesgue measure if*

$$\forall C \in \mathcal{B}(\mathbb{R}^d), \qquad \mathbb{P}(X \in C) = \int_{x \in C} p(x)\mathrm{d}x.$$

A measurable and nonnegative function p on $\mathbb{R}^d$ is a probability density if and only if

$$\int_{x \in \mathbb{R}^d} p(x)\mathrm{d}x = 1.$$

Exercise 1.3.2. Do deterministic random variables in $\mathbb{R}^d$ have a density with respect to the Lebesgue measure?

We shall often omit to refer to the Lebesgue measure and simply refer to p as 'the density of X'.

Remark 1.3.3. *Random variables $X_1, \dots, X_n$ with respective densities $p_1, \dots, p_n$ are independent if and only if the concatenated vector $(X_1, \dots, X_n)$ has density $p_1(x_1) \cdots p_n(x_n)$.*

The *Transfer Theorem*[5] asserts that if X has density p, then for any measurable function f such that $\mathbb{E}[|f(X)|] < +\infty$,

$$\mathbb{E}[f(X)] = \int_{x \in \mathbb{R}^d} f(x)p(x)\mathrm{d}x.$$

1.3.2 One-dimensional examples

Definition 1.3.4 (Uniform distribution)**.** *For $a < b$, the* uniform distribution $\mathcal{U}[a, b]$ *is the probability measure on $\mathbb{R}$ with density*

$$\mathbb{1}_{\{a<x<b\}} \frac{1}{b-a}.$$

Exercise 1.3.5.

1. If $X \sim \mathcal{U}[a, b]$, compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.
2. If $X \sim \mathcal{U}[a, b]$ and $c, d \in \mathbb{R}$, what is the law of $cX + d$?

Definition 1.3.6 (Exponential distribution)**.** *For $\lambda > 0$, the* Exponential distribution $\mathcal{E}(\lambda)$ *is the probability measure on $\mathbb{R}$ with density*

$$\mathbb{1}_{\{x>0\}} \lambda \mathrm{e}^{-\lambda x}.$$

Exercise 1.3.7.

1. If $X \sim \mathcal{E}(\lambda)$, compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.
2. If $X \sim \mathcal{E}(\lambda)$ and $c > 0$, what is the law of cX?

[4] whence the name 'Poisson distribution'?

[5] Also sometimes called *Law Of The Unconscious Statistician*.

Definition 1.3.8 (Gaussian distribution)**.** *For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the* Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ *is the probability measure on $\mathbb{R}$ with density*

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

We also denote by $\mathcal{N}(\mu, 0)$ the Dirac measure in μ.

Exercise 1.3.9. 1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

2. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$, what is the law of $aX + b$?

The Gaussian distribution $\mathcal{N}(0, 1)$ is called the *standard* Gaussian distribution. It follows from Exercise 1.3.9 that if $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$, then

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

We will use this remark repeatedly in the course.

1.3.3 Sum of independent random variables

Proposition 1.3.10 (Density of the sum of independent random variables)**.** *Let X and Y be independent random variables in $\mathbb{R}^d$, with respective densities p and q. The random variable $Z := X + Y$ has density*

$$r(z) := \int_{x \in \mathbb{R}^d} p(x) q(z - x) \mathrm{d}x,$$

which is called the convolution *of p and q.*

Proof. For any $C \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathbb{P}(Z \in C) = \mathbb{E}\left[\mathbb{1}_{\{X+Y \in C\}}\right].$$

Since the pair (X, Y) has density $p(x)q(y)$, the Transfer Theorem yields

$$\mathbb{E}\left[\mathbb{1}_{\{X+Y \in C\}}\right] = \int_{x,y \in \mathbb{R}^d} \mathbb{1}_{\{x+y \in C\}} p(x) q(y) \mathrm{d}x \mathrm{d}y,$$

and by the Fubini Theorem, the change of variable $z = x + y$ yields

$$\int_{x,y \in \mathbb{R}^d} \mathbb{1}_{\{x+y \in C\}} p(x) q(y) \mathrm{d}x \mathrm{d}y = \int_{z \in \mathbb{R}^d} \mathbb{1}_{\{z \in C\}} \left(\int_{x \in \mathbb{R}^d} p(x) q(z - x) \mathrm{d}x \right) \mathrm{d}z$$

so that

$$\mathbb{P}(Z \in C) = \int_{z \in \mathbb{R}^d} \mathbb{1}_{\{z \in C\}} r(z) \mathrm{d}z$$

with $r(z)$ defined by the statement of the proposition. □

An example of application of this proposition is provided by the link between Exponential and Gamma distributions.

Definition 1.3.11 (Gamma distribution)**.** *For $a > 0$ and $\lambda > 0$, the* Gamma distribution *with parameters (a, λ), denoted by $\Gamma(a, \lambda)$, has the density*

$$\mathbb{1}_{\{x>0\}} \frac{\lambda^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-\lambda x},$$

where Γ is Euler's function defined by

$$\Gamma(a) = \int_{t=0}^{+\infty} t^{a-1} \mathrm{e}^{-t} \mathrm{d}t.$$

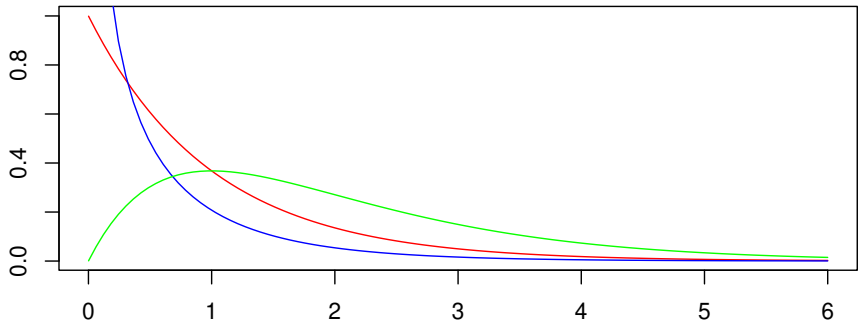


Figure 1.1: The density of $\Gamma(a, 1)$, for $a = 0.5$ (blue), $a = 1$ (red), $a = 2$ (green).

We have $\mathcal{E}(\lambda) = \Gamma(1, \lambda)$, hence the parameter λ remains called the *rate*[6]. The parameter a is called the *shape* parameter, the density of $\Gamma(a, 1)$ is plotted on Figure 1.1 for various values of a.

Exercise 1.3.12. 1. Show that if $X \sim \Gamma(a, \lambda)$ and $c > 0$ then $cX \sim \Gamma(a, \lambda/c)$.

2. Let $X \sim \Gamma(a, \lambda)$ and $Y \sim \Gamma(b, \lambda)$ be independent. Show that $X + Y \sim \Gamma(a + b, \lambda)$.

3. Let $X_1, \dots, X_n$ be independent variables with law $\mathcal{E}(\lambda)$. Show that $\frac{1}{n}(X_1 + \cdots + X_n) \sim \Gamma(n, n\lambda)$.

1.3.4 CDF and quantiles

In the case $E = \mathbb{R}$, we shall also often work with the *Cumulative Distribution Function*[7] (CDF) F of X, defined by

$$\forall x \in \mathbb{R}, \qquad F(x) = \mathbb{P}(X \leq x).$$

Exercise 1.3.13. If X has density p and CDF F, what is the link between F and p?

The related notion of *quantile* shall also play an important role throughout the course.

Definition 1.3.14 (Quantile)**.** *Let X be a real-valued random variable. For any $r \in (0, 1)$, a quantile of order r for X is any number q_r such that*

$$\mathbb{P}(X \leq q_r) = r.$$

In general a quantile need not exist, and it need not be unique either. However in most cases of interest, the variable X possesses a density which is positive on $\mathbb{R}$ (or on an interval), in which case q_r exists and is unique.

Since q_r only depends on the law of X, we shall often directly speak of the quantile of a distribution. For instance, the quantile of order r of the standard Gaussian distribution $\mathcal{N}(0, 1)$ will be denoted by ϕ_r.

Exercise 1.3.15. What is the link between ϕ_r and ϕ_{1-r}?

Exercise 1.3.16 (Quantile of the Exponential distribution)**.** For $\lambda > 0$ and $r \in (0, 1)$, compute the quantile of order r of the Exponential distribution $\mathcal{E}(\lambda)$.

[6] Gamma distributions are sometimes parametrised by the parameter $1/\lambda$, which is called the *scale* parameter.

[7] *Fonction de répartition* en français.

1.3.5 Characteristic function

Definition 1.3.17 (Characteristic function)**.** *Let X be a random vector in $\mathbb{R}^d$. The* characteristic function *of X is the function $\Psi_X : \mathbb{R}^d \to \mathbb{C}$ defined by*

$$\forall u \in \mathbb{R}^d, \qquad \Psi_X(u) := \mathbb{E}\left[e^{i\langle u, X\rangle}\right] = \mathbb{E}[\cos(\langle u, X\rangle)] + i\mathbb{E}[\sin(\langle u, X\rangle)].$$

Since $|e^{i\langle u,X\rangle}| = 1$, the function Ψ_X is always well-defined. Its most important property is that, for any pair of real-valued random vectors X, Y, if $\Psi_X(u) = \Psi_Y(u)$ for any $u \in \mathbb{R}^d$, then X and Y have the same law.

The notion of characteristic function is useful to prove the following statement, which could also be deduced from Proposition 1.3.10 but at the expense of more involved computation.

Proposition 1.3.18 (Additivity of independent Gaussian variables)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$.*

Exercise 1.3.19 (Proof of Proposition 1.3.18)**.** Let $G \sim \mathcal{N}(0, 1)$.

1. Show that Ψ_G is C^1 on $\mathbb{R}$ and satisfies the differential equation
$$\begin{cases} \Psi_G'(u) + u\Psi_G(u) = 0, \\ \Psi_G(0) = 1. \end{cases}$$
2. Deduce that $\Psi_G(u) = \exp(-u^2/2)$.
3. For arbitrary $\mu, \sigma \in \mathbb{R}$, compute the characteristic function of $X \sim \mathcal{N}(\mu, \sigma^2)$.
4. Complete the proof of Proposition 1.3.18.

1.3.6 Gaussian vectors

Definition 1.3.20 (Gaussian vector)**.** *A random vector $X \in \mathbb{R}^d$ is called* Gaussian *if, for any $u \in \mathbb{R}^d$, the random variable $\langle u, X\rangle$ is Gaussian in the sense of Definition 1.3.8.*

Let X be a Gaussian vector, and set $\mathbb{E}[X] = m$ and $\mathrm{Cov}[X] = K$. For any $u \in \mathbb{R}^d$, we deduce from Definition 1.3.20 that $\langle u, X\rangle \sim \mathcal{N}(\langle u, m\rangle, \langle u, Ku\rangle)$ and therefore Exercise 1.3.19 shows that the characteristic function of X is given by

$$\Phi_X(u) = \Phi_{\langle u,X\rangle}(1) = \exp\left(i\langle u, m\rangle - \frac{1}{2}\langle u, Ku\rangle\right).$$

This identity shows that the law of X only depends on m and K, therefore we denote it by $\mathcal{N}_d(m, K)$.

Exercise 1.3.21. If $X \sim \mathcal{N}_d(m, K)$ and $A \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$, what is the law of $AX + b$?

As a consequence of Exercise 1.3.21, we deduce that for any $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d\times d}$ symmetric and nonnegative, one can construct a vector X with law $\mathcal{N}_d(m, K)$ by letting $X = m + \Sigma G$, where G is a vector whose components $G_1, \dots, G_d$ are independent $\mathcal{N}(0, 1)$ variables, and $\Sigma \in \mathbb{R}^{d\times d}$ is such that $\Sigma\Sigma^\top = K$.

Remark 1.3.22. *When K is invertible, the probability measure $\mathcal{N}_d(m, K)$ has density*

$$\frac{1}{\sqrt{(2\pi)^d \det K}} \exp\left(-\frac{1}{2}\langle x - m, K^{-1}(x - m)\rangle\right),$$

but otherwise it does not have a density.

1.4 Convergence and limit theorems

Throughout this section we let $(X_n)_{n\geq 1}$ and X be random variables in $\mathbb{R}^d$.

1.4.1 Convergence almost sure and in probability

Definition 1.4.1 (Convergence almost sure and in probability)**.** *The sequence $(X_n)_{n\geq 1}$ is said to converge to X:*

- almost surely *if* $\mathbb{P}(\lim_{n\to+\infty} X_n = X) = 1$*;*
- in probability *if, for any* $\epsilon > 0$, $\lim_{n\to+\infty} \mathbb{P}(\|X_n - X\| \geq \epsilon) = 0$.

The following statement is the probabilistic formulation of the Dominated Convergence Theorem.

Theorem 1.4.2 (Dominated Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence which converges to X, almost surely. Assume that there exists a random variable $Y \geq 0$ such that $\mathbb{E}[Y] < +\infty$ and, for any $n \geq 1$, $\|X_n\| \leq Y$, almost surely. Then $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$.*

Exercise 1.4.3. Using Theorem 1.4.2, show that if $X_n \to X$ almost surely then $X_n \to X$ in probability. What do you think or know about the converse statement?

The following statement is obvious for almost sure convergence but its proof requires a bit more work for convergence in probability.

Lemma 1.4.4 (Continuous functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be continuous. If $X_n \to X$ almost surely (resp. in probability), then $f(X_n) \to f(X)$ almost surely (resp. in probability).*

1.4.2 Convergence in distribution

Definition 1.4.5 (Convergence in distribution)**.** *The sequence $(X_n)_{n\geq 1}$ is said to converge to X in distribution*[8] *if, for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $\mathbb{E}[f(X_n)]$ converges to $\mathbb{E}[f(X)]$.*

Since the law of X is characterised by the set of values of $\mathbb{E}[f(X)]$ for all continuous and bounded functions $f : \mathbb{R}^d \to \mathbb{R}$, it is clear that $X_n \to X$ in distribution if and only if $X_n \to X'$ in distribution for any random variable X' which has the same law as X. Therefore we shall often directly write $X_n \to P$, with P the law of X.

Proposition 1.4.6 (Properties of convergence in distribution)**.** *(i) If $X_n \to X$ in distribution, then for any continuous function $f : \mathbb{R}^d \to \mathbb{R}^k$, $f(X_n) \to f(X)$ in distribution.*

(ii) $X_n \to X$ in distribution if and only if the associated characteristic functions satisfy $\Psi_{X_n}(u) \to \Psi_X(u)$ for any $u \in \mathbb{R}^d$.

(iii) If $X_n \to X$ in probability then $X_n \to X$ in distribution.

(iv) Conversely, if $X_n \to X$ in distribution and $X = a$ is deterministic, then $X_n \to X$ in probability.

(v) If $E = \mathbb{R}$, the following statements are equivalent:

(a) $X_n \to X$ in distribution,

(b) $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ for any $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$,

(c) $\mathbb{P}(X_n < x) \to \mathbb{P}(X < x)$ for any $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$.

Exercise 1.4.7. Let $(U_n)_{n\geq 1}$ be a sequence of independent random variables with uniform distribution over $[0, 1]$. Let $M_n = \max_{1\leq i\leq n} U_i$.

[8] *En loi* en français.

1. Show that $M_n \to 1$, in probability.
2. Show that, for any $\omega \in \Omega$, the sequence $(M_n(\omega))_{n\geq 1}$ is nondecreasing. Deduce that $M_n \to 1$, almost surely.
3. For any $x \geq 0$, compute $\lim_{n\to+\infty} \mathbb{P}(1 - M_n > x/n)$. Deduce that $n(1 - M_n)$ converges in distribution toward some limit X and describe the law of X.

In the sequel of this course we will also need the following result.

Lemma 1.4.8 (Convergence of quantiles)**.** *Let ζ_n be a sequence of real-valued random variables, which converges in distribution to some limit ζ. Fix $r \in (0, 1)$ and let $q_{n,r}$ be a quantile of order r of ζ_n.*

(i) Assume that there exists an interval I of $\mathbb{R}$ such that $\zeta \in I$, almost surely, and on I, ζ has a positive density. Then ζ has a unique quantile of order r, denoted by q_r, and $q_{n,r}$ converges to q_r.

(ii) Assume that ζ is deterministic. Then $q_{n,r}$ converges to ζ.

The second statement of Lemma 1.4.8 is illustrated on Figure 1.4.8.

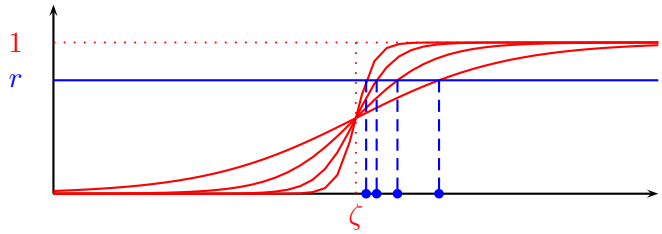


Figure 1.2: Illustration of the convergence of $q_{n,r}$ to ζ when ζ_n converges to the deterministic quantity ζ. Red curves represent the CDF of ζ_n, which converges to a step function in ζ, and blue points the associated values of $q_{n,r}$.

1.4.3 Strong Law of Large Numbers and Central Limit Theorem

We say that a sequence $(X_n)_{n\geq 1}$ is *independent and identically distributed* (iid) if the family $(X_n)_{n\geq 1}$ is independent and all variables X_n have the same law. For any $n \geq 1$, we denote by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

the *empirical mean* of $X_1, \ldots, X_n$. The asymptotic behaviour of $\overline{X}_n$ is described by the strong Law of Large Numbers.

Theorem 1.4.9 (Strong Law of Large Numbers)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of iid random variables in $\mathbb{R}^d$, such that $\mathbb{E}[\|X_1\|] < +\infty$. Then*

$$\lim_{n\to+\infty} \overline{X}_n = \mathbb{E}[X_1], \qquad \textit{almost surely.}$$

The speed of convergence of $\overline{X}_n$ to $\mathbb{E}[X_1]$ is made precise by the Central Limit Theorem.

Theorem 1.4.10 (Multidimensional Central Limit Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of iid random variables in $\mathbb{R}^k$, such that $\mathbb{E}[\|X_1\|^2] < +\infty$. Then*

$$\lim_{n\to+\infty} \sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_1]\right) = \mathcal{N}_k(0, \mathrm{Cov}[X_1]), \qquad \textit{in distribution.}$$

1.5 Random variables with SciPy

For a large class of probability distributions, the module `scipy.stats`[9] makes it very easy to generate realisations of random variables, and to compute densities/mass functions, CDFs and quantiles. A few probability distributions are listed in Table 1.1.

Distribution	Object
Bernoulli	`bernoulli`
Binomial	`binom`
Geometric	`geom`
Poisson	`poisson`
Uniform	`uniform`
Exponential	`expon`
Gaussian	`norm`
Gamma	`gamma`
Gaussian vector	`multivariate_normal`

Table 1.1: Names of a few standard discrete and continuous distributions available in `scipy.stats`.

Given a distribution `distrib`, random variables are generated thanks to the function `distrib.rvs()`. For discrete distributions, the probability mass function is given by `distrib.pmf()`, while for continuous distributions, the density is given by `distrib.pdf()`. For real valued distributions, the CDF is given by `distrib.cdf()` and the quantile is given by `distrib.ppf()`.

As a first example of application, the following code generates $n = 100$ realisations of a geometric random variable with parameter $p = 0.1$, and superposes the plot of the histogram of this sample together with the theoretical probability mass function.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import geom

n = 100 # Number of realisations
p = 0.1 # Parameter of the geometric distribution
x = np.arange(1,50) # Range of x coordinate

# Generate n realisations of a geometric random variable
sample = geom.rvs(p, size=n)

# Plot histogram of the sample
plt.hist(sample, density=True, bins=x, color='g', label='Sample Histogram')

# Calculate and plot the theoretical PMF
plt.plot(x, geom.pmf(x, p), color='b', label='Theoretical PMF')

# Add labels and title
plt.xlabel('Value')
plt.ylabel('Probability')
plt.title('Histogram of Geometric Distribution and Theoretical PMF')
plt.legend()

# Show plot
plt.show()
```

The output plot is represented on Figure 1.3.

As a second example, the following code generates $n = 1000$ realisations $X_1, \ldots, X_n$ of a Gamma random variable X with parameters $a = 3$, $\lambda = 1$ and shows the convergence of the sequence $(\overline{X}_k)_{1 \leq k \leq n}$ toward $\mathbb{E}[X] = a/\lambda$.

[9] See the documentation here: https://docs.scipy.org/doc/scipy/reference/stats.html.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gamma

# Parameters
n = 1000
a = 3
lambda_ = 1
expected_value = a / lambda_

# Generate n realisations of a Gamma random variable
sample = gamma.rvs(a, scale=1/lambda_, size=n)

# Compute the empirical mean for each k = 1, 2, ..., n
empirical_means = np.cumsum(sample) / np.arange(1, n + 1)

# Plot the empirical mean
plt.plot(empirical_means, label='Empirical Mean')

# Plot the theoretical expected value
plt.axhline(y=expected_value, color='r', linestyle='--', label=f'Theoretical
    Mean = {expected_value}')

# Add labels and title
plt.xlabel('Sample Size k')
plt.ylabel('Empirical Mean of $X$')
plt.title('Convergence of Empirical Mean to Theoretical Mean')

# Add legend
plt.legend()

# Show plot
plt.show()
```

The output plot is represented on Figure 1.3.

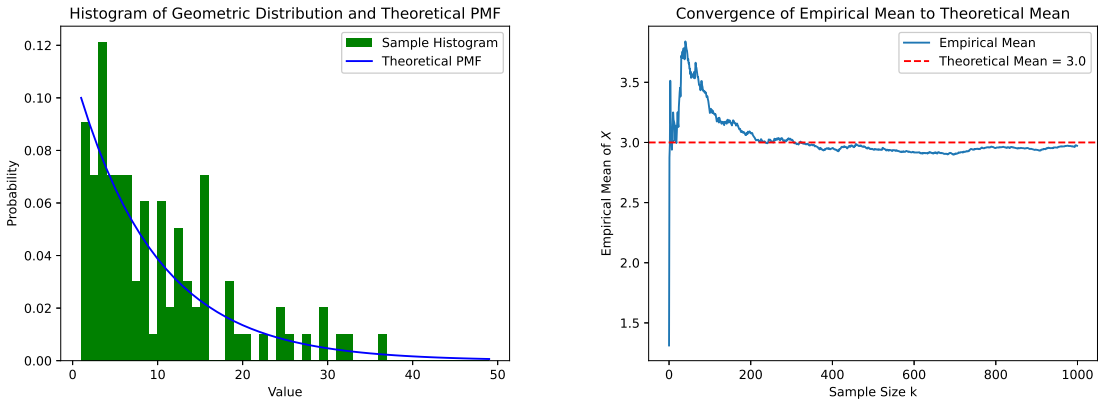


Figure 1.3: Output plots for the two examples.

1.A Exercises

Training exercises

Exercise 1.A.1 (Simulation of basic discrete random variables)**.** Assume that your scientific computing language can only generate a sequence $(U_i)_{i\geq 1}$ of independent random variables which are uniformly

distributed on $[0,1]$. Given $p \in (0,1]$, how to use this sequence to generate:

1. a $\mathcal{B}(p)$ variable?
2. a $\mathcal{B}(n,p)$ variable?
3. a $\mathcal{G}(p)$ variable?

Exercise 1.A.2 (What you read in the news)**.** In a famous newspaper article from 2011[10], two engineers claim that if p is the probability that one nuclear reactor has a serious accidents during one year, then the probability that at least one serious accident occurs among N nuclear reactors, during M years, is $p \times N \times M$.

1. Applying this result with an estimated value $p = 4/14000$, the authors deduce that the probability to have at least one serious accident among the $N = 143$ currently working nuclear reactors in Europe, during the next $M = 30$ years, is equal to 1.23. What do you think of this statement?
2. With the same values for p, N and M, how would you correct this computation?

Exercise 1.A.3 (Unbiasing a coin toss, an exercise attributed to Von Neumann)**.** Assume that you have a random number generator which returns independent Bernoulli variables with an *unknown* parameter $p \in (0,1)$. How to use it to draw a Bernoulli random variable with parameter $1/2$? And can you do the same with an unbalanced dice?

Exercise 1.A.4 (Expectation and variance of Gamma distribution)**.** Let $X \sim \Gamma(a,\lambda)$. Show that $\mathbb{E}[X] = a/\lambda$ and $\mathrm{Var}(X) = a/\lambda^2$.

Exercise 1.A.5 (Higher order moments of Gaussian distributions)**.** Let $G \sim \mathcal{N}(0,1)$. Show that, for any $k \geq 1$,

$$\mathbb{E}[G^{2k-1}] = 0, \qquad \mathbb{E}[G^{2k}] = 1 \times 3 \times \cdots \times (2k-3) \times (2k-1).$$

Exercise 1.A.6 (Beta distribution)**.** For $a, b > 0$, the *Beta distribution* with parameters a and b, denoted by $\beta(a,b)$, is the probability distribution on $[0,1]$ with density

$$u \mapsto \frac{1}{\mathrm{B}(a,b)} u^{a-1}(1-u)^{b-1}, \qquad \mathrm{B}(a,b) := \int_{u=0}^{1} u^{a-1}(1-u)^{b-1}\mathrm{d}u.$$

1. Let $X \sim \Gamma(a,1)$ and $Y \sim \Gamma(b,1)$ be independent. Compute the joint density of the pair (Z,U), where
$$Z := X + Y, \qquad U = \frac{X}{X+Y}.$$
2. Deduce the identity
$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$
3. Show that the variables Z and U are independent and give their distributions.
4. Let $U \sim \beta(a,b)$ and $V \sim \beta(a+b,c)$ be independent. What is the law of UV?

Homework

Exercise 1.A.7 (Mortality rate)**.** Let T be a positive random variable with a continuous and positive density p on $(0,+\infty)$. We interpret the variable T as the lifetime of an individual (in actuarial science) or of the component of a system (in reliability analysis). We let F be the CDF of T.

1. For $t > 0$, the event $\{T > t\}$ means that at age t, the individual is still alive. Conditionally on this event, express the probability $\mathbb{P}(T \in (t, t+h] | T > t)$ that the individual dies in the next h time units, in terms of F.
2. The limit of $\mathbb{P}(T \in (t, t+h] | T > t)/h$, when $h \to 0$, is called the mortality rate[11] and denoted

[10] See the blog post https://www.afis.org/Nouveau-record-du-monde-de-probabilites, in French, for details.

[11] In the context of reliability analysis, it is called the *hazard rate*.

by $\mu(t)$. Express this quantity in terms of F and p.

3. Figure 1.4 shows the mortality rate for the US population with data from 2003[12]. What is your interpretation of this graph?

4. If $T \sim \mathcal{E}(\lambda)$, what is μ?

5. Mortality rates are sometimes modeled using the three-parameter *Gompertz–Makeham law* given by

$$\forall t \geq 0, \qquad \mathbb{P}(T > t) = \exp\left(-\left(\lambda t + \frac{\alpha}{\beta}(\mathrm{e}^{\beta t} - 1)\right)\right),$$

with $\beta, \lambda > 0$ and $\alpha \geq 0$.

   (a) Compute the associated mortality rate $\mu(t)$.

   (b) Which phenomenon, observable on Figure 1.4, is not captured by this model?

   (c) For large values of t, which is assumed to be given in years, express in terms of β by how much mortality is multiplied every year.

   (d) Based on Figure 1.4, to which value would you estimate β?

6. In the notebook `Gompertz.ipynb` and the file `gompertz-dataset.csv`, available on Educnet, a sample of $n = 10^5$ realisations, drawn from the Gompertz–Makeham law with unknown parameters, is given.

   (a) Complete the notebook to plot the associated empirical hazard rate.

   (b) Be creative to find a way to estimate the parameters λ, α and β from these data. In your written sheet, describe your method and give the value of the parameters that you obtained.

Among all students, the 3 closest to the true values will be offered a chocolate box!

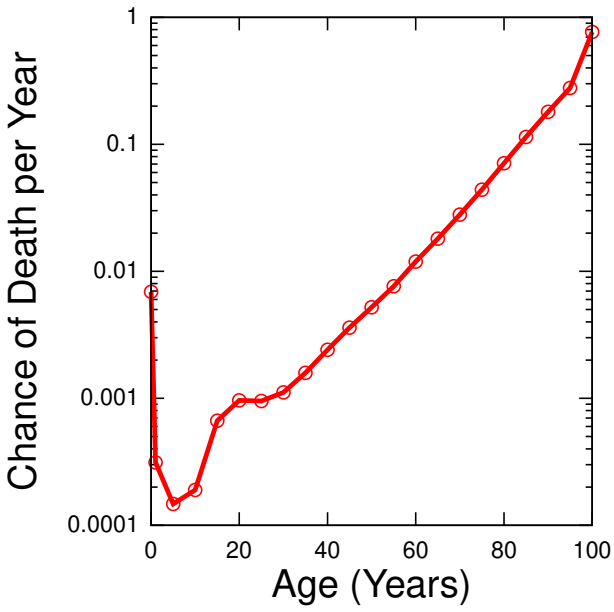


Figure 1.4: The mortality rate for the US population with data from 2003.

[12] Source: https://en.wikipedia.org/wiki/File:USGompertzCurve.svg.

Supplementary exercises

Exercise 1.A.8 (Spearman's coefficient)**.** Let $(x_i^1, x_i^2)_{1\le i\le n}$ be a family of elements of $\mathbb{R}^2$. The *Bravais–Pearson* correlation coefficient between $(x_i^1)_{1\le i\le n}$ and $(x_i^2)_{1\le i\le n}$ is the empirical correlation

$$\mathrm{Corr}(x^1, x^2) := \frac{\mathrm{Cov}\left(x^1, x^2\right)}{\sqrt{\mathrm{Var}\left(x^1\right)}\sqrt{\mathrm{Var}\left(x^2\right)}},$$

where we have denoted

$$\mathrm{Cov}\left(x^1, x^2\right) := \frac{1}{n}\sum_{i=1}^{n}(x_i^1 - \overline{x}_n^1)(x_i^2 - \overline{x}_n^2), \qquad \mathrm{Var}\left(x^k\right) := \frac{1}{n}\sum_{i=1}^{n}(x_i^k - \overline{x}_n^k)^2.$$

For data sets which are related by a monotonic but nonlinear function, the Bravais–Pearson coefficient may be inadequate to represent the dependency between the features.

1. Plot the Bravais–Pearson coefficient between the three points $(0,0)$, $(1-\epsilon, \epsilon)$, $(1,1)$ as a function of $\epsilon \in (0, 1/2]$.

To better describe such nonlinear correlations, Spearman's coefficient is constructed as follows. Take a set of pairs x_i^1, x_i^2, $i \in \{1, \dots, n\}$, such that both vectors of features have pairwise distinct coordinates. For any index $i \in \{1, \dots, n\}$, we let r_i^1 and r_i^2 be the respective rank of x_i^1 and x_i^2 among the series $x_1^1, \dots, x_n^1$ and $x_1^2, \dots, x_n^2$ sorted in increasing order; in other words, if σ^1 and σ^2 are the permutations of $\{1, \dots, n\}$ such that

$$x^1_{\sigma^1(1)} < \cdots < x^1_{\sigma^1(n)}, \qquad x^2_{\sigma^2(1)} < \cdots < x^2_{\sigma^2(n)},$$

then for all $i \in \{1, \dots, n\}$,

$$\sigma^1(r_i^1) = \sigma^2(r_i^2) = i.$$

We now define the Spearman coefficient by

$$r_{\mathrm{s}} = \mathrm{Corr}(r^1, r^2).$$

2. Compute r^1, r^2 and the Spearman coefficient for the set $(0,0)$, $(1-\epsilon, \epsilon)$, $(1,1)$ with $\epsilon \in (0, 1/2]$.
3. More generally, if $x_i^2 = f(x_i^1)$ with f increasing, what is the value of r_{s}? And is f is decreasing?
4. Show that
$$\frac{1}{n}\sum_{i=1}^{n}(r_i^1 - \overline{r}^1)^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i^2 - \overline{r}^2)^2 = \frac{n^2-1}{12},$$
and deduce that
$$r_{\mathrm{s}} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_i^1 - r_i^2)^2.$$

Exercise 1.A.9 (Cauchy distribution)**.** The Cauchy distribution with parameter $a > 0$, denoted by $\mathcal{C}(a)$, is the probability measure on $\mathbb{R}$ with density

$$\frac{a}{\pi}\frac{1}{x^2 + a^2}.$$

1. If $X \sim \mathcal{C}(a)$, what can you say about $\mathbb{E}[X]$?
2. Let $X \sim \mathcal{C}(a)$ and $c \in \mathbb{R}$. What is the law of cX?
3. If $U \sim \mathcal{N}(0, \sigma^2)$ and $V \sim \mathcal{N}(0, \tau^2)$ are independent, with $\sigma^2 > 0$ and $\tau^2 > 0$, show that U/V has a Cauchy distribution and express its parameter in terms of σ and τ.

4. The aim of this question is to show that if $X \sim \mathcal{C}(a)$ and $Y \sim \mathcal{C}(b)$ are independent, then $X + Y \sim \mathcal{C}(a + b)$. The direct computation of the convolution of the associated densities is difficult so we need to use a trick.

   (a) For $a > 0$, the *Laplace* distribution with parameter a, denoted by $\mathcal{L}(a)$, is the probability measure with density $\frac{a}{2}\mathrm{e}^{-a|u|}$ on $\mathbb{R}$. If $U \sim \mathcal{L}(a)$, compute the characteristic function $\Psi_U(x)$ of U, for any $x \in \mathbb{R}$.

   (b) We recall that the Fourier inverse transform states that if Ψ_U is integrable on $\mathbb{R}$, then its density can be recovered by the formula

   $$p(u) = \frac{1}{2\pi}\int_{x\in\mathbb{R}} \mathrm{e}^{-\mathrm{i}ux}\Psi_U(x)\mathrm{d}x.$$

   Deduce from this identity the value of the characteristic transform $\Psi_X(u)$ of $X \sim \mathcal{C}(a)$.

   (c) Conclude that if $X \sim \mathcal{C}(a)$ and $Y \sim \mathcal{C}(b)$ are independent, then $X + Y \sim \mathcal{C}(a + b)$.

5. Let $(X_n)_{n\geq 1}$ be a sequence of iid $\mathcal{C}(1)$ variables. The aim of this question is to describe the law of $\overline{X}_n$ and to show that is does not converge in probability.

   (a) Compute the law of $\overline{X}_n$.

   (b) Compute the law of $\overline{X}_{2n} - \overline{X}_n$.

   (c) Conclude that $\overline{X}_n$ does not converge in probability.

Exercise 1.A.10 (Stronger convergence in the Central Limit Theorem)**.** Under the assumptions of the Central Limit Theorem, say in dimension $d = 1$ to make things simpler, it is a natural question to wonder whether there exists a random variable Z such that $Z_n := \sqrt{n}(\overline{X}_n - \mathbb{E}[X_1])$ converges to Z in probability. Notice that if such a variable exists, then necessarily $Z \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \mathrm{Var}(X_1)$.

1. Set $Y_i := X_i - \mathbb{E}[X_1]$ and let $Z'_n := \frac{1}{\sqrt{n}}\sum_{i=n+1}^{2n} Y_i$. Show that Z'_n converges in distribution to some random variable Z' and explicit the law of Z'.

2. If Z_n converges in probability to some random variable Z, show that Z'_n converges in probability and express its limit in terms of Z.

3. What do you conclude?

Part I

Parameter Inference

Lecture 2

Pointwise Estimation in Parametric Models

Contents

The basis of a *statistical inference* experiment is the observation of a *sample of data* $X_1, \dots, X_n$ in some space E, which are seen as *independent and identically distributed* (iid) realisations of random variables whose law P is unknown. The statistician's goal is then to infer, or estimate, the law P of these variables, or at least some feature of this law, such as the mean of X_1, or the probability that this variable reaches a certain threshold. As the simplest example, you may think of consecutive flips of a coin: you *observe* a sequence of Heads and Tails, and want to *estimate* the probability to get Tail. We first introduce a general formalism in Section 2.1.

We next address the case where, before observing $X_1, \dots, X_n$, we make the assumption that P has a certain predetermined *shape*, such as Exponential or Gaussian. This assumption allows us to reduce the estimation of P, which is a probability measure, to the estimation of a few *parameters*: the number $\lambda > 0$ in the case of the Exponential model, the pair (μ, σ^2) for the Gaussian model, and so on. We provide some more formalism in Section 2.2, and next discuss asymptotic properties in the parametric setting in Section 2.3.

2.1 Estimators

In this Section, we let $X_1, \dots, X_n$ be iid random variables in some measurable space $(E, \mathcal{E})$, with common distribution P. We denote by $\mathbf{X}_n$ the *sample* $(X_1, \dots, X_n) \in E^n$.

2.1.1 General definitions and first examples

We are interested in the estimation, using only the observation of the sample, of some Quantity of Interest (QoI) which depends on P. For instance, QoI can be:

- if $X_1 \in \mathbb{R}$, the mean or the variance of X_1;
- the probability p_A that X_1 takes its values in a measurable subset $A \subset E$;
- the probability distribution P of X_1 itself; or assuming that P has a density p, the pointwise evaluation of the density $p(x_0)$ for some given x_0.

The fact that we only want to use the observation of the sample to estimate QoI is expressed by the following notion.

Definition 2.1.1 (Statistic)**.** *A* statistic T_n *is a random variable which writes*

$$T_n = t_n(\mathbf{X}_n),$$

for some deterministic and measurable function t_n *on* E^n *which does not depend on* P*.*

An *estimator* of QoI is then a statistic which aims at giving an approximate value of QoI. This is a rather loose definition: for example, assume that $E = \mathbb{R}$, and that we are interested in the estimation of $\mathrm{QoI} = \mathbb{E}[X_1]$. Then, the following statistics are estimators of $\mathbb{E}[X_1]$:

- the constant random variable equal to 0;
- the single realisation X_1;
- the *empirical mean* of the sample, defined by $\overline{X}_n := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

The last one has the property that, *whatever the choice of* P, as long as $\mathbb{E}[X_1]$ is well-defined, the strong Law of Large Numbers ensures that $\overline{X}_n$ converges to $\mathbb{E}[X_1]$, almost surely: the larger the sample, the better the estimation. This property is called the *consistency* of the estimator.

Definition 2.1.2 (Consistency of an estimator)**.** *An estimator* Z_n *of* QoI *is called* consistent *if* Z_n *converges to* QoI *in probability when* $n \to +\infty$*. It is* strongly consistent *if this convergence holds almost surely.*

Exercise 2.1.3 (Empirical frequency)**.** Assume that the QoI is the probability $p_A = \mathbb{P}(X_1 \in A)$ for some set $A \in \mathcal{E}$. Show that the *empirical frequency*

$$\widehat{p}_{n,A} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}$$

is strongly consistent.

Exercise 2.1.4 (Consistency of the empirical variance)**.** Assume that $E = \mathbb{R}$ and that $\mathbb{E}[X_1^2] < +\infty$. The *empirical variance* of the sample is the estimator

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

of $\mathrm{Var}(X_1)$.

1. Show that $V_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2$.
2. Deduce that V_n is strongly consistent.

2.1.2 Bias and MSE of an estimator

Consistency is a natural requirement for assessing the quality of an estimator. However, in practice, you will never observe a sample with infinite size. Therefore, nonasymptotic criteria have to be introduced. From now on we assume that $\mathrm{QoI} \in \mathbb{R}^d$, and we denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$.

Definition 2.1.5 (Bias of an estimator)**.** *Let Z_n be an estimator of* QoI *such that $\mathbb{E}[\|Z_n\|] < +\infty$. The* bias *of Z_n is the quantity*

$$b(Z_n) := \mathbb{E}[Z_n] - \mathrm{QoI} \in \mathbb{R}^d.$$

If $b(Z_n) = 0$, then Z_n is called unbiased*.*

The bias determines how far Z_n is, in average, from the quantity QoI that it is supposed to estimate.

Exercise 2.1.6. Check that the empirical mean $\overline{X}_n$ and the empirical frequecy $\widehat{p}_{n,A}$ introduced in the previous subsection are unbiased.

Interestingly, the empirical variance V_n *turns out to be biased*, as is shown in the next exercise.

Exercise 2.1.7 (Bias of the empirical variance)**.** Show that

$$\mathbb{E}[V_n] = \left(1 - \frac{1}{n}\right) \mathrm{Var}(X_1).$$

This motivates the introduction of the following quantity.

Definition 2.1.8 (Unbiased estimator of the variance)**.** *If $E = \mathbb{R}$, the* unbiased estimator of the variance *is the statistic*

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

Exercise 2.1.9 (Properties of S_n^2)**.**

1. Show that, if $\mathbb{E}[X_1^2] < +\infty$, then S_n^2 is strongly consistent.
2. If $E = \mathbb{R}^p$, so $X_1 = (X_1^1, \ldots, X_1^p)$, and $\mathbb{E}[\|X_1\|^2] < +\infty$, construct an unbiased and strongly consistent estimator of the covariance matrix $\mathrm{Cov}[X_1]$.

The bias only measures the gap between Z_n and QoI *in average*. A stronger measure of this gap is provided by the *Mean Squared Error*.

Definition 2.1.10 (Mean Squared Error)**.** *Let Z_n be an estimator of* QoI *such that $\mathbb{E}[\|Z_n\|^2] < +\infty$. The* Mean Squared Error[1] *(MSE) of Z_n is defined by*

$$\mathrm{MSE}(Z_n) = \mathbb{E}[\|Z_n - \mathrm{QoI}\|^2].$$

The MSE has the general shape

$$\mathrm{MSE}(Z_n) = \mathbb{E}[\ell(Z_n; \mathrm{QoI})],$$

with $\ell : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$ given by $\ell(z, z') = \|z - z'\|^2$. In this formulation, ℓ is called a *loss function*. By taking loss functions that are not quadratic, one can construct different *risk functions* $\mathrm{R}(Z_n)$, which provide other measures of the accuracy of the estimator Z_n.

The choice of a quadratic loss function has the advantage to entail a decomposition of the MSE into a term of bias and a term of variance. In the next statement, we define the variance of a random *vector* $Z \in \mathbb{R}^d$ by $\mathrm{Var}(Z) = \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]$; and recall that it is the trace of the covariance matrix of Z.

Proposition 2.1.11 (Bias and variance decomposition of the MSE)**.** *Let Z_n be an estimator of* QoI *such that $\mathbb{E}[\|Z_n\|^2] < +\infty$. We have*

$$\mathrm{MSE}(Z_n) = \|b(Z_n)\|^2 + \mathrm{Var}(Z_n).$$

[1] *Risque quadratique* en français.

Proof. Starting from the definition of the MSE, we write

$$\begin{aligned}\mathrm{MSE}(Z_n) &= \mathbb{E}\left[\|(Z_n - \mathbb{E}[Z_n]) - (\mathrm{QoI} - \mathbb{E}[Z_n])\|^2\right] \\ &= \mathbb{E}\left[\|Z_n - \mathbb{E}[Z_n]\|^2 - 2\langle Z_n - \mathbb{E}[Z_n], \mathrm{QoI} - \mathbb{E}[Z_n]\rangle + \|\mathrm{QoI} - \mathbb{E}[Z_n]\|^2\right] \\ &= \mathbb{E}\left[\|Z_n - \mathbb{E}[Z_n]\|^2\right] + \|\mathrm{QoI} - \mathbb{E}[Z_n]\|^2,\end{aligned}$$

which leads to the claimed decomposition. □

Remark 2.1.12 (Unbiased estimators and covariance matrix comparison)**.** *For any symmetric matrices $K^1, K^2 \in \mathbb{R}^{d\times d}$, let us write $K^1 \preceq K^2$ if $\langle u, K^1 u\rangle \leq \langle u, K^2 u\rangle$ for any $u \in \mathbb{R}^d$. If two estimators Z^1 and Z^2 are unbiased and such that*

$$\mathrm{Cov}[Z_n^1] \preceq \mathrm{Cov}[Z_n^2],$$

then we have[2]

$$\mathrm{MSE}(Z_n^1) = \mathrm{Var}(Z_n^1) = \mathrm{tr}\,\mathrm{Cov}[Z_n^1] \leq \mathrm{tr}\,\mathrm{Cov}[Z_n^2] = \mathrm{Var}(Z_n^2) = \mathrm{MSE}(Z_n^2),$$

so Z_n^1 has a lower MSE than Z_n^2.

Exercise 2.1.13 (Comparison of estimators)**.** Assume that $E = \mathbb{R}$ and $\mathbb{E}[X_1^2] < +\infty$. Compute the MSE of the estimators 0, X_1 and $\overline{X}_n$ of $\mathbb{E}[X_1]$.

In the previous exercise, $\overline{X}_n$ has a lower MSE than X_1, whatever the underlying distribution P. We now present an example of a family of estimators for which it is not possible to minimise the MSE uniformly over P.

Exercise 2.1.14 (On the bias-variance tradeoff in the Bernoulli model)**.** We assume that $E = \{0, 1\}$, so that $X_1, \ldots, X_n$ are iid Bernoulli random variables with a certain parameter $p \in [0, 1]$. We are interested in the estimation of this parameter p. We consider the estimator

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

of p, which is strongly consistent, but also introduce the family of estimators

$$\widehat{p}_n^h = (1-h)\widehat{p}_n + \frac{h}{2}, \qquad h \in [0, 1],$$

which are not consistent as soon as $h > 0$ (unless $p = 1/2$).

1. Compute $b(\widehat{p}_n^h)$, $\mathrm{Var}(\widehat{p}_n^h)$ and $\mathrm{MSE}(\widehat{p}_n^h)$.

2. For several values of p, plot $\mathrm{MSE}(\widehat{p}_n^h)$ as a function of h.

Exercise 2.1.14 shows that in general, the bias and the variance cannot be simultaneously minimised, and a tradeoff between both must be considered. The example depicted here may seem artificial, but from the course *Introduction to Data Science* (see Appendix A) you already know that: (i) there are situations in which you have no choice but introduce bias controlled by a parameter $h > 0$ (think of histograms or Kernel Density Estimation); (ii) introducing bias may in addition help reduce variance and hence overfitting (think of regularisation in linear regression). Another justification of the introduction of such biased estimators will be provided in Lecture 6 in the setting of Bayesian estimation.

[2] You may deduce that $K^1 \preceq K^2$ implies that $\mathrm{tr}\,K^1 \leq \mathrm{tr}\,K^2$ from the inequality $\langle e_j, K^1 e_j\rangle \leq \langle e_j, K^2 e_j\rangle$ for each vector e_j of the canonical basis of $\mathbb{R}^d$.

2.2 Parametric models and moment estimators

In this Section, we focus on the estimation of the whole distribution P of X_1. There are mostly two possible approaches to this purpose: *nonparametric methods*, such as histograms or Kernel Density Estimation, which directly aim at reconstructing P (see Appendix A); and *parametric methods*, which rely on the assumption, *a priori*, that P has a certain shape, such as Exponential or Gaussian. In this course, we mostly focus on the parametric approach.

2.2.1 Parametric model

Definition 2.2.1 (Parametric model)**.** *Given a measurable space $(E, \mathcal{E})$, a* parametric model *on E is a set of probability measures*

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$

on the space E, indexed by a set of parameters $\Theta \subset \mathbb{R}^q$.

The choice of a model depends on the physical features of the phenomenon under study. For example, the choice of the Exponential model to describe the lifetime of a lightbulb prevents one from taking into account *ageing*, namely the fact that the older the lightbulb is, the more likely it is to die soon. In order to accommodate this phenomenon, it is necessary to consider a larger model, as has been seen in Exercise 1.A.7.

Notice that if $P_\theta = P_{\theta'}$ for two different values of θ, θ', then there is no hope to be able to distinguish between the values of θ and θ' from the mere observation of $X_1, \dots, X_n$. Therefore, we shall systematically work under the assumption the map $\theta \mapsto P_\theta$ is injective, in which case the model $\mathcal{P}$ is called *identifiable*.

Exercise 2.2.2 (Identifiable parametrisation of the Gaussian model)**.** We consider the set of all Gaussian measures $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R}$. Which of there parametrisations make the model identifiable?

1. $\theta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}$,
2. $\theta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times [0, +\infty)$,
3. $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times [0, +\infty)$.

2.2.2 Estimator

From now on, a parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^q$, is fixed. For all $\theta \in \Theta$, it is convenient to denote by $\mathbb{P}_\theta$ the probability measure under which, for all $n \geq 1$, the random variables $X_1, \dots, X_n \in E$ are iid according to P_θ. The notation $\mathbb{E}_\theta$, Var_θ, etc. is defined accordingly.

In this context, the Quantities of Interest that we aim at estimating are in fact functions of θ. Therefore, we fix a function $g : \Theta \to \mathbb{R}^d$ and turn our interest to the estimation of $g(\theta)$. In this setting, following Section 2.1, an estimator Z_n of $g(\theta)$ is a statistic which takes its values in $\mathbb{R}^d$.

Remark 2.2.3. *If an estimator $\widehat{\theta}_n$ of θ is available and you want to estimate $g(\theta)$, a natural estimator is $Z_n = g(\widehat{\theta}_n)$. It is called the* plug-in *estimator of $g(\theta)$.*

The definitions of consistency, as well as of the bias and the MSE of an estimator introduced in Section 2.1 remain unchanged. However, to emphasise their dependency upon θ, we now denote them by $b(Z_n; \theta)$ and $\mathrm{MSE}(Z_n; \theta)$, respectively. The bias-variance decomposition of the MSE given by Proposition 2.1.11 thus writes

$$\mathrm{MSE}(Z_n; \theta) = \|b(Z_n; \theta)\|^2 + \mathrm{Var}_\theta(Z_n).$$

To show that an estimator *is* biased, one may often resort to the *strict* Jensen inequality. We recall that a function ϕ on some interval I is *strictly convex* if, for any $x, y \in I$ with $x \neq y$, for any $t \in (0, 1)$,

$$\phi(tx + (1-t)y) < t\phi(x) + (1-t)\phi(y).$$

Proposition 2.2.4 (Strict Jensen inequality)**.** *Let X be a non deterministic random variable in $\mathbb{R}$, such that $\mathbb{E}[|X|] < +\infty$, and which takes its values in some interval I. Let $\phi : I \to \mathbb{R}$ be strictly convex. Then*

$$\phi(\mathbb{E}[X]) < \mathbb{E}[\phi(X)].$$

Exercise 2.2.5 (The Exponential model)**.** We consider the Exponential model $\{\mathcal{E}(\lambda), \lambda > 0\}$.

1. Show that $\overline{X}_n$ is an unbiased estimator of $1/\lambda$.
2. Show that $1/\overline{X}_n$ is a biased estimator of λ.
3. Compute its bias. *You will first need to compute the law of $\overline{X}_n$: have a look back at Exercise 1.3.12!*

As should be clear from this exercise, showing that an estimator is biased can often be done rather shortly, using (strict) inequalities. However, computing the actual bias requires much more effort, and in particular one generally needs to know the exact distribution of the estimator.

2.2.3 The method of moments

The method of moments is a natural procedure to construct estimators. We start by detailing the example of the Exponential model $\{\mathcal{E}(\lambda), \lambda > 0\}$, in which we look for an estimator of λ. By the Law of Large Numbers, for all $\lambda > 0$,

$$\lim_{n\to+\infty} \overline{X}_n = \mathbb{E}_\lambda[X_1] = \frac{1}{\lambda}, \qquad \mathbb{P}_\lambda\text{-almost surely,}$$

so that

$$\lim_{n\to+\infty} \frac{1}{\overline{X}_n} = \lambda, \qquad \mathbb{P}_\lambda\text{-almost surely,}$$

by continuity of the function $x \mapsto 1/x$. As a consequence, $\widehat{\lambda}_n = 1/\overline{X}_n$ is a strongly consistent estimator of the parameter λ.

The abstract generalisation of this procedure is called the *method of moments*. For the estimation of $g(\theta) \in \mathbb{R}^d$ in the model $\{P_\theta, \theta \in \Theta\}$, it consists in finding functions φ and m such that, for all $\theta \in \Theta$,

$$\mathbb{E}_\theta[\varphi(X_1)] = m(g(\theta)).$$

For the Exponential model, we took $\varphi(x) = x$, $g(\lambda) = \lambda$ and $m(\lambda) = 1/\lambda$.

Then the Law of Large Numbers allows us to approximate $m(g(\theta))$ with $\frac{1}{n}\sum_{i=1}^n \varphi(X_i)$, so that as soon as m has a continuous inverse function m^{-1},

$$Z_n = m^{-1}\left(\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\right)$$

is a strongly consistent estimator of $g(\theta)$.

Exercise 2.2.6 (The Pareto distribution)**.** For $\theta > 1$, we denote by P_θ the probability distribution with density

$$p_\theta(x) := \frac{\theta}{x^{\theta+1}} \mathbb{1}_{\{x>1\}}$$

on $\mathbb{R}$.

1. Compute $\mathbb{E}_\theta[X_1]$.
2. Deduce a strongly consistent estimator of θ.

When constructing an estimator with the methods of moments, one usually tries the 'simplest' functions φ, for which the computation of $\mathbb{E}_\theta[\varphi(X_1)]$ is possible or easy: typically, $\varphi(x) = x$ or $\varphi(x) = |x|^2, |x|^3 \ldots$ are natural candidates.

Exercise 2.2.7 (The Pareto distribution, continued)**.** We consider the same example as in Exercise 2.2.6, but now we only assume that $\theta > 0$. Notice that p_θ remains a probability density.

1. What is wrong with the estimator obtained in Exercise 2.2.6?
2. Compute $\mathbb{E}_\theta[1/X_1]$.
3. Deduce a strongly consistent estimator of θ.

The method can also be employed with functions of the form $\varphi(x) = \mathbb{1}_{\{x \leq x_0\}}$ for given x_0, see Exercise 2.A.5 for instance.

2.3 Asymptotic normality

2.3.1 Definition

The construction of estimators by the method of moments depends on the arbitrary choice of the function φ, so in principle, different choices of φ may lead to different estimators Z_n of $g(\theta)$ (see Exercise 2.A.3 for an illustration). By construction, all these estimators are strongly consistent. To discriminate between them, it is then natural to look for the one which converges 'the fastest' toward $g(\theta)$. This rate of convergence can be quantified by the notion of *asymptotic variance*.

Definition 2.3.1 (Asymptotic normality)**.** *A consistent estimator Z_n of $g(\theta)$ is* asymptotically normal *if, for all $\theta \in \Theta$, there exists a symmetric and nonnegative matrix $K(\theta) \in \mathbb{R}^{d\times d}$ such that $\sqrt{n}(Z_n - g(\theta))$ converges in distribution, under $\mathbb{P}_\theta$, to the d-dimensional Gaussian measure $\mathcal{N}_d(0, K(\theta))$. The matrix-valued function $\theta \mapsto K(\theta)$ is called the* asymptotic covariance *of Z_n.*

When $d = 1$, $K(\theta)$ is a nonnegative scalar, called the *asymptotic variance* of Z_n. If several consistent and asymptotically normal estimators of $g(\theta)$ are available, the best one is the one with the smallest asymptotic variance. In the sequel of the course, the notion of asymptotic normality shall play a central role in the construction of asymptotic confidence intervals and tests.

Asymptotic normality results are almost always obtained by applying the Central Limit Theorem, recalled in Theorem 1.4.10.

Exercise 2.3.2. Show that in the Bernoulli model, $\overline{X}_n$ is a strongly consistent and asymptotically normal estimator of p, and compute its asymptotic variance.

Exercise 2.3.3 (The Uniform model)**.** We consider the estimation of θ in the Uniform model $\{\mathcal{U}[0, \theta], \theta > 0\}$.

1. Compute $\mathbb{E}_\theta[X_1]$ and deduce a moment estimator $\widetilde{\theta}_n$. Show that it is strongly consistent, asymptotically normal, and compute its asymptotic variance.
2. We now consider the estimator $\widehat{\theta}_n = \max_{1\leq i\leq n} X_i$. Using Exercise 1.4.7, show that $\widehat{\theta}_n$ is strongly consistent, and describe the limit in distribution of $n(\theta - \widehat{\theta}_n)$. Is $\widehat{\theta}_n$ asymptotically normal?
3. Which estimator converges faster to θ?

2.3.2 Delta method

The Central Limit Theorem provides the asymptotic normality of estimators which write as empirical means of iid variables. In the Exponential model, the estimator $\widehat{\lambda}_n = 1/\overline{X}_n$ of λ derived in Subsection 2.2.3 is a nonlinear function of the empirical mean. The asymptotic normality of such quantities is obtained by the *Delta method*.

Theorem 2.3.4 (Delta Method). *Let $(\zeta_n)_{n\geq 1}$ be a sequence of $\mathbb{R}^k$-valued random variables and $a \in \mathbb{R}^k$ such that $\zeta_n \to a$ in probability and $\sqrt{n}(\zeta_n - a)$ converges in distribution to some random vector $Y \in \mathbb{R}^k$. Let $\phi : \mathcal{U} \to \mathbb{R}^d$ be a C^1 function, defined on an open subset $\mathcal{U}$ of $\mathbb{R}^k$ which contains a. Then*

$$\lim_{n\to+\infty} \sqrt{n}\left(\phi(\zeta_n) - \phi(a)\right) = \nabla\phi(a)Y, \qquad \text{in distribution,}$$

where $\nabla\phi(a)$ is the $d \times k$ matrix with coefficients

$$\left(\nabla\phi(a)\right)_{ij} = \frac{\partial\phi_i}{\partial x_j}(a), \qquad i = 1, \dots, d, \quad j = 1, \dots, k.$$

Remark 2.3.5. *The vector $\nabla\phi(a)Y \in \mathbb{R}^d$ rewrites explicitly*

$$\nabla\phi(a)Y = \begin{pmatrix} \langle \nabla\phi_1(a), Y\rangle \\ \vdots \\ \langle \nabla\phi_d(a), Y\rangle \end{pmatrix}.$$

The proof of Theorem 2.3.4 relies on the use of Slustky's Theorem, which we first recall.

Theorem 2.3.6 (Slutsky's Theorem). *Let (X_n, Y_n) be a sequence of pairs of random variables such that X_n converges in probability to some deterministic variable a, while Y_n converges in distribution to some random variable Y. Then the pair (X_n, Y_n) converges in distribution to (a, Y), and as a consequence, for any continuous function ψ, $\psi(X_n, Y_n)$ converges in distribution to $\psi(a, Y)$.*

Proof of Theorem 2.3.4. In order to avoid technical arguments, we assume that ϕ is C^2 on $\mathcal{U} = \mathbb{R}^k$, with globally bounded second derivatives[3].

Let $i \in \{1, \dots, d\}$. For all $n \geq 1$,

$$\begin{aligned}\sqrt{n}(\phi_i(\zeta_n) - \phi_i(a)) &= \sqrt{n}\int_{t=0}^1 \frac{\mathrm{d}}{\mathrm{d}t}\phi_i((1-t)a + t\zeta_n)\mathrm{d}t \\ &= \left\langle \sqrt{n}(\zeta_n - a), \int_{t=0}^1 \nabla\phi_i((1-t)a + t\zeta_n)\mathrm{d}t \right\rangle.\end{aligned}$$

Owing to our regularity assumption, there exists a constant $C \geq 0$ such that

$$\left\| \int_{t=0}^1 \nabla\phi_i((1-t)a + t\zeta_n)\mathrm{d}t - \nabla\phi_i(a) \right\| \leq C\int_{t=0}^1 \|(1-t)a + t\zeta_n - a\|\mathrm{d}t = \frac{C}{2}\|\zeta_n - a\|.$$

As a consequence

$$\lim_{n\to+\infty} \int_{t=0}^1 \nabla\phi_i((1-t)a + t\zeta_n)\mathrm{d}t = \nabla\phi_i(a), \qquad \text{in probability,}$$

and the conclusion follows from Slutsky's Theorem. □

We apply the combination of the Central Limit Theorem and the Delta Method to the estimator $\widehat{\lambda}_n = 1/\overline{X}_n$ of λ in the Exponential model. Since $\mathrm{Var}_\lambda(X_1) = 1/\lambda^2$, the Central Limit Theorem first asserts that

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \to Y \sim \mathcal{N}(0, 1/\lambda^2) \qquad \text{in distribution.}$$

Second, applying the Delta Method with $\zeta_n = \overline{X}_n$, $\phi(x) = 1/x$ and $a = 1/\lambda$ yields

$$\sqrt{n}\left(\widehat{\lambda}_n - \lambda\right) = \sqrt{n}\left(\phi(\overline{X}_n) - \phi(1/\lambda)\right) \to \phi'(1/\lambda)Y = \lambda^2 Y \qquad \text{in distribution,}$$

from which we conclude that the estimator $\widehat{\lambda}_n$ is asymptotically normal, with asymptotic variance $\phi'(1/\lambda)^2/\lambda^2 = \lambda^2$.

The argument can in fact be used for the abstract formulation of the method of moments introduced in Subsection 2.2.3.

[3] For a proof in a general framework, we refer to [7, Theorem 3.1, p. 26].

Exercise 2.3.7. With the notation introduced in Subsection 2.2.3, write the general expression of the asymptotic variance of Z_n in terms of $\mathbb{E}_\theta[\varphi(X_1)]$ and $\mathrm{Var}_\theta[\varphi(X_1)]$.

Remark 2.3.8. *Assume that Z_n is an estimator of $g(\theta)$ which takes the form*

$$Z_n = \phi\left(\overline{X}_n\right),$$

for some smooth function $\phi : \mathbb{R}^k \to \mathbb{R}^d$. The Central Limit Theorem writes informally

$$\overline{X}_n \simeq \mathbb{E}_\theta[X_1] + \frac{1}{\sqrt{n}} Y,$$

with $Y \sim \mathcal{N}_k(0, \mathrm{Cov}_\theta[X_1])$. The Delta Method then corresponds to the first-order expansion

$$Z_n \simeq \phi\left(\mathbb{E}_\theta[X_1] + \frac{1}{\sqrt{n}} Y\right) \simeq \phi(\mathbb{E}_\theta[X_1]) + \frac{1}{\sqrt{n}} \nabla\phi(\mathbb{E}_\theta[X_1]) Y.$$

Performing the expansion up to second order, we get (taking $d = k = 1$ for the sake of simplicity, but the argument carries over to any dimension)

$$Z_n \simeq \phi(\mathbb{E}_\theta[X_1]) + \frac{1}{\sqrt{n}} \phi'(\mathbb{E}_\theta[X_1]) Y + \frac{1}{2n} \phi''(\mathbb{E}_\theta[X_1]) Y^2,$$

so that the bias of Z_n is approximately

$$b(Z_n; \theta) = \mathbb{E}_\theta[Z_n] - \phi(\mathbb{E}_\theta[X_1]) \simeq \frac{1}{2n} \phi''(\mathbb{E}_\theta[X_1]) \,\mathrm{Var}_\theta(X_1).$$

This shows that the bias is generally of order of magnitude $1/n$, and therefore has a contribution $1/n^2$ in the MSE, while the variance is of order of magnitude $1/n$. Therefore, for large values of n, the main contribution to the MSE is generally due to the variance term.

2.A Exercises

Training exercises

Exercise 2.A.1 (Nonexistence of an unbiased estimator)**.** Let $p \in (0, 1)$ and $X_1, \ldots, X_n$ independent Bernoulli variables with parameter p. The purpose of this exercise is to show that there does not exist an unbiased estimator Z_n of $g(p) = 1/p$.

1. For all $\mathbf{x}_n \in \{0, 1\}^n$, express $\mathbb{P}_p(\mathbf{X}_n = \mathbf{x}_n)$ as a function of $k = x_1 + \cdots + x_n$.
2. We assume that there exists an unbiased estimator Z_n of $g(p)$. Show that there exist real numbers $a_0, \ldots, a_n$ such that
$$\sum_{k=0}^{n} a_k p^k (1-p)^{n-k} = \frac{1}{p},$$
for all $p \in (0, 1)$.
3. Conclude that there cannot exist such an estimator.

Exercise 2.A.2 (Asymptotic variance of the empirical variance)**.** Let $X_1, \ldots, X_n$ be iid random variables, such that $\mathbb{E}[X_1^4] < +\infty$ and $\mathbb{E}[X_1] = 0$. We write

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2.$$

The purpose of the exercise is to show that $\sqrt{n}(V_n - \mathrm{Var}(X_1))$ converges to a Gaussian distribution, and to compute the associated variance. For $k = 2, 3, 4$, we write $\rho_k = \mathbb{E}[X_1^k]$.

1. Define the vectors $Y_i = (X_i, X_i^2)$, $1 \leq i \leq n$, and $y = (0, \rho_2)$ in $\mathbb{R}^2$. Compute the covariance matrix of Y_1.
2. Let $\varphi(x_1, x_2) = x_2 - x_1^2$. Compute $\nabla\varphi(y)$.
3. Express V_n as a function of $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Deduce that $\sqrt{n}(V_n - \rho_2)$ converges to $\mathcal{N}(0, \rho_4 - \rho_2^2)$.

Homework

Exercise 2.A.3 (Poisson model)**.** The Poisson model is the set $\{\mathcal{P}(\lambda), \lambda > 0\}$, under which we recall that, for all $\lambda > 0$,

$$\forall k \in \mathbb{N}, \qquad \mathbb{P}_\lambda(X_1 = k) = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

1. A first moment estimator.
   (a) Compute $\mathbb{E}_\lambda[X_1]$ and deduce a strongly consistent moment estimator $\widetilde{\lambda}_n^{(1)}$.
   (b) What is the bias of this estimator?
   (c) Compute $\mathbb{E}_\lambda[X_1^2]$.
   (d) Show that $\widetilde{\lambda}_n^{(1)}$ is asymptotically normal and compute its asymptotic variance.
2. A second moment estimator.
   (a) For any $\lambda > 0$, let $m(\lambda) = \mathbb{E}_\lambda[X_1^2]$. Show that m is a bijection from $(0, +\infty)$ to itself and compute m^{-1}.
   (b) Deduce another strongly consistent moment estimator $\widetilde{\lambda}_n^{(2)}$ based on m^{-1}.
   (c) Using Jensen's strict inequality, show that this estimator is biased.
   (d) Compute $\mathbb{E}_\lambda[X_1^3]$ and $\mathbb{E}_\lambda[X_1^4]$.
   (e) Show that $\widetilde{\lambda}_n^{(2)}$ is asymptotically normal and compute the asymptotic variance of $\widetilde{\lambda}_n^{(2)}$.
3. Which of the estimators $\widetilde{\lambda}_n^{(1)}$ or $\widetilde{\lambda}_n^{(2)}$ has the smaller asymptotic variance?

Supplementary exercises

Exercise 2.A.4 (Three estimators of μ^2)**.** Let $X \in \mathbb{R}$ be a random variable with finite second order moment and let $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$. We assume that we have two samples $(X_1, \ldots, X_n)$ and $(X_1', \ldots, X_n')$ of independent random variables with the same law as X, and that these two samples are independent from each other. We denote by $\overline{X}_n$ and $\overline{X}_n'$ the associated empirical means, and consider the following three estimators of μ^2:

$$A_n = \left(\frac{\overline{X}_n + \overline{X}_n'}{2}\right)^2, \qquad B_n = \overline{X}_n \times \overline{X}_n', \qquad C_n = \frac{1}{n}\sum_{i=1}^n X_i X_i' = \overline{(XX')}_n.$$

1. Show that A_n, B_n and C_n are strongly consistent.
2. Compute the bias of A_n, B_n and C_n.
3. Show that A_n, B_n and C_n are asymptotically normal and compute their respective asymptotic variances.
4. Using the approximation

$$\mathrm{MSE} \simeq \mathrm{bias}^2 + \frac{1}{n} \times \text{asymptotic variance},$$

which estimator do you prefer?

Exercise 2.A.5 (Translation of Cauchy distributions)**.** For all $\theta \in \mathbb{R}$, we denote by P_θ the probability measure with density

$$p(x;\theta) = \frac{1}{\pi((x-\theta)^2+1)}.$$

We furthermore recall that

$$\forall x \in \mathbb{R}, \qquad \int_{y=-\infty}^{x} \frac{\mathrm{d}y}{y^2+1} = \arctan(x) + \frac{\pi}{2}.$$

1. What is the name of P_θ when $\theta = 0$?
2. Let $X_1, \dots, X_n$ iid random variables with law P_θ. For all $i \in \{1, \dots, n\}$, we define $U_i = \mathbb{1}_{\{X_i \leq 0\}}$. Compute $\mathbb{E}_\theta[U_1]$.
3. Deduce a moment estimator $\widetilde{\theta}_n$ of θ. Show that this estimator is strongly consistent.
4. Show that $\widetilde{\theta}_n$ is asymptotically normal, and compute its asymptotic variance. *Hint: use the relation* $\tan' = 1 + \tan^2$.

Lecture 3

Statistics in Gaussian Models

Contents

In the next Lectures, we will use estimators Z_n of quantities of interest $g(\theta)$ to construct confidence intervals or hypothesis tests. To do so, we will need some information on the law of Z_n. It is however generally difficult to know the exact distribution of Z_n, and one often has to resort to asymptotic properties, such as asymptotic normality, which may be derived from the CLT and the Delta method for most choices of model. Still, there is a model where exact, nonasymptotic distributions may be obtained: this is the Gaussian model[1].

We start by some preliminaries on Gaussian vectors in Section 3.1. We next describe the law of some estimators related to iid Gaussian samples in Section 3.2, and finally extend the study to the linear regression model, with Gaussian errors, in Section 3.3.

3.1 Preliminaries

3.1.1 Standard Gaussian vectors

Definition 3.1.1 (Standard Gaussian vector)**.** *A random vector $G \in \mathbb{R}^n$ is called an n-dimensional* standard Gaussian vector *if its coordinates $G_1, \dots, G_n$ are independent $\mathcal{N}(0,1)$ random variables.*

Lemma 3.1.2. *The characteristic function of an n-dimensional standard Gaussian vector G writes*

$$\forall u \in \mathbb{R}^n, \qquad \Psi_G(u) = \exp\left(-\frac{\|u\|^2}{2}\right).$$

As a consequence, G is a standard Gaussian vector if and only if $G \sim \mathcal{N}_n(0, I_n)$.

Proof. Let $u \in \mathbb{R}^n$. Using the independence of $G_1, \dots, G_n$ and Exercise 1.3.19, we get

$$\Psi_G(u) = \mathbb{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^n u_j G_j\right)\right] = \prod_{j=1}^n \mathbb{E}\left[\exp\left(\mathrm{i}u_j G_j\right)\right] = \prod_{j=1}^n \exp\left(-\frac{u_j^2}{2}\right) = \exp\left(-\frac{\|u\|^2}{2}\right).$$

Since the right-hand side is the characteristic function of $\mathcal{N}_n(0, I_n)$, the conclusion follows. □

[1] This fact was discovered in the early 20th century by William Sealy Gosset, who was working as a statistician at the Guinness brewery in Dublin, and was interested in making reliable inferences from small samples. Writing under the pseudonym 'Student', because the Guinness Board of Directors allowed its scientists to publish research on condition that they do not mention beer, Guinness, or their own surname, he derived what is now known as the Student, or t-distribution, which we shall introduce in this Lecture.

3.1.2 The Cochran Theorem

For any linear subspace[2] E of $\mathbb{R}^n$, we denote by x^E the orthogonal projection of a vector $x \in \mathbb{R}^n$ onto E. We recall that, if $(e_1, \dots, e_k)$ is any orthonormal basis of E, then

$$x^E = \sum_{j=1}^{k} \langle e_j, x \rangle e_j.$$

The main theoretical result of this subsection is the following statement.

Theorem 3.1.3 (Cochran Theorem)**.** *Let $G \sim \mathcal{N}_n(0, I_n)$.*

(i) For any linear subspace E of $\mathbb{R}^n$, with dimension k, the coordinates of G^E in any orthonormal basis of E form a k-dimensional standard Gaussian vector.

(ii) If E_1, E_2 are orthogonal linear subspaces of $\mathbb{R}^n$, then the random vectors G^{E_1} and G^{E_2} are independent.

Proof of (i). Let $(e_1, \dots, e_k)$ be an orthonormal basis of E. The vector of coordinates of G^E in the basis $(e_1, \dots, e_k)$ writes $(\langle e_1, G \rangle, \dots, \langle e_k, G \rangle)$, so that its characteristic function is given, for any $u = (u_1, \dots, u_k) \in \mathbb{R}^k$, by

$$\mathbb{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k} u_j \langle e_j, G \rangle\right)\right] = \Psi_G\left(\sum_{j=1}^{k} u_j e_j\right) = \exp\left(-\frac{1}{2}\left\|\sum_{j=1}^{k} u_j e_j\right\|^2\right),$$

by Lemma 3.1.2. Since the family $(e_1, \dots, e_k)$ is orthonormal,

$$\left\|\sum_{j=1}^{k} u_j e_j\right\|^2 = \sum_{j=1}^{k} |u_j|^2 = \|u\|^2,$$

which by Lemma 3.1.2 again, shows that $(\langle e_1, G \rangle, \dots, \langle e_k, G \rangle) \sim \mathcal{N}_k(0, I_k)$.

Proof of (ii). For $p = 1, 2$, we let k_p be the dimension of E_p, and $(e_{p,1}, \dots, e_{p,k_p})$ be an orthonormal basis of E_p. Since E_1 and E_2 are orthogonal, the concatenated family $(e_{1,1}, \dots, e_{1,k_1}, e_{2,1}, \dots, e_{2,k_2})$ is orthonormal in $\mathbb{R}^n$. Therefore, by (i),

$$(\langle e_{1,1}, G \rangle, \dots, \langle e_{1,k_1}, G \rangle, \langle e_{2,1}, G \rangle, \dots, \langle e_{2,k_2}, G \rangle) \sim \mathcal{N}_{k_1+k_2}(0, I_{k_1+k_2}),$$

which in particular implies that the vectors $(\langle e_{1,1}, G \rangle, \dots, \langle e_{1,k_1}, G \rangle)$ and $(\langle e_{2,1}, G \rangle, \dots, \langle e_{2,k_2}, G \rangle)$ are independent by Lemma 3.1.2 again. Since, for $p = 1, 2$,

$$G^{E_p} = \sum_{j=1}^{k_p} \langle e_{p,j}, G \rangle e_{p,j},$$

we conclude that G^{E_1} and G^{E_2} are independent. □

3.1.3 Chi-square distribution

Definition 3.1.4 (Chi-square distribution)**.** *For $n \geq 1$, the* chi-square distribution *with n degrees of freedom, denoted by $\chi_2(n)$, is the law of the random variable*

$$Z_n := \sum_{i=1}^{n} G_i^2 = \|G\|^2,$$

where $G = (G_1, \dots, G_n) \sim \mathcal{N}_n(0, I_n)$.

[2] *Sous-espace vectoriel* en français.

Remark 3.1.5. *As a consequence of Theorem 3.1.3 (i), we observe that if $G \sim \mathcal{N}_n(0, I_n)$ and E is a linear subspace E of $\mathbb{R}^n$ with dimension k, then*

$$\|G^E\|^2 \sim \chi_2(k).$$

Exercise 3.1.6. If $Z_n \sim \chi_2(n)$, compute $\mathbb{E}[Z_n]$.

Exercise 3.1.7 (A consequence of Cochran's Theorem)**.** Let E be a linear subspace of $\mathbb{R}^n$ with dimension k, and Π be the matrix of the orthogonal projection onto E in the canonical basis of $\mathbb{R}^n$.

1. Show that Π is symmetric and nonnegative.
2. Show that if $X \sim \mathcal{N}_n(0, \Pi)$, then $\|X\|^2 \sim \chi_2(k)$.

3.2 Statistics of Gaussian samples

The Gaussian model $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ is often employed to study real-valued samples. Since the parameters μ and σ^2 respectively coincide with $\mathbb{E}_{\mu,\sigma^2}[X_1]$ and $\mathrm{Var}_{\mu,\sigma^2}(X_1)$, it is natural to estimate them with the empirical mean $\overline{X}_n$ and either the empirical variance V_n or its unbiased version S_n^2, respectively. Clearly, these estimators are strongly consistent.

Exercise 3.2.1 (Asymptotic properties)**.** Show that the estimator $(\overline{X}_n, V_n)$ of (μ, σ^2) is asymptotically normal and compute its asymptotic covariance matrix.

A pleasant property of the Gaussian model is that, in fact, the joint law of the pair $(\overline{X}_n, V_n)$ can be described explicitly for fixed n. This is the aim of this section. It shall actually be more convenient to work with the unbiased estimator of the variance S_n^2.

3.2.1 Joint law of $(\overline{X}_n, S_n^2)$

Proposition 3.2.2 (Joint law of $(\overline{X}_n, S_n^2)$)**.** *Under $\mathbb{P}_{\mu,\sigma^2}$, the estimators $\overline{X}_n$ and S_n^2 are independent, and*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad (n-1)\frac{S_n^2}{\sigma^2} \sim \chi_2(n-1).$$

Proof. The statement on the marginal distribution of $\overline{X}_n$ is an immediate consequence of Proposition 1.3.18.

As a preliminary step for the sequel, we introduce the reduced variables

$$X_i' = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1),$$

and define $\overline{X}_n'$, $(S_n')^2$ as the empirical mean and unbiased estimator of the variance associated with the standard Gaussian vector $G = (X_1', \ldots, X_n')$. We thus have

$$\overline{X}_n = \mu + \sigma \overline{X}_n', \qquad S_n^2 = \sigma^2 (S_n')^2. \tag{3.1}$$

We now denote by E_1 be the linear subspace of $\mathbb{R}^n$ spanned[3] by the vector $\mathbf{1} = (1, \ldots, 1)$, and set $E_2 = E_1^\perp$. On the one hand, $\|\mathbf{1}\|^2 = n$, so that introducing the unit vector $e = \mathbf{1}/\sqrt{n}$, one gets that

$$G^{E_1} = \langle G, e\rangle e = \frac{1}{n}\langle G, \mathbf{1}\rangle \mathbf{1} = \overline{X}_n' \mathbf{1}. \tag{3.2}$$

On the other hand,

$$\|G^{E_2}\|^2 = \|G - G^{E_1}\|^2 = \|G - \overline{X}_n' \mathbf{1}\|^2 = \sum_{i=1}^n (X_i' - \overline{X}_n')^2 = (n-1)(S_n')^2. \tag{3.3}$$

[3] *Engendré* en français.

By Theorem 3.1.3 (ii), G^{E_1} and G^{E_2} are independent. Since (3.2) and (3.3) show that $\overline{X}'_n$ and $(S'_n)^2$ are deterministic functions of G^{E_1} and G^{E_2}, respectively, we deduce that $\overline{X}'_n$ and $(S'_n)^2$ are independent, and finally by (3.1), that $\overline{X}_n$ and S_n^2 are independent.

By Remark 3.1.5, and since it is clear that E_2 has dimension $n-1$, $\|G^{E_2}\|^2 \sim \chi_2(n-1)$, which by (3.1) and (3.3) yields $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi_2(n-1)$. □

3.2.2 The Student distribution

Proposition 3.2.2 yields in particular

$$\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1).$$

For reasons that will become apparent already in the next Lecture, it is of interest to determine what happens to the law of this random variable if one replaces σ^2 by it unbiased estimator S_n^2. The answer is based on the introduction of the *Student distribution*.

Definition 3.2.3 (Student distribution)**.** *For $n \geq 1$, the* Student distribution *with n degrees of freedom, denoted by $\mathrm{t}(n)$, is the law of the random variable*

$$T_n = \frac{Y}{\sqrt{Z_n/n}},$$

where $Y \sim \mathcal{N}(0,1)$ and $Z_n \sim \chi_2(n)$ are independent.

Proposition 3.2.4 (Student variable)**.** *Under $\mathbb{P}_{\mu,\sigma^2}$,*

$$\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \sim \mathrm{t}(n-1).$$

Proof. Let us introduce

$$Y = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1), \qquad Z_{n-1} = \frac{n-1}{\sigma^2} S_n^2 \sim \chi_2(n-1).$$

By Proposition 3.2.2, Y and Z_{n-1} are independent, so we deduce from Definition 3.2.3 that

$$\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{Y}{\sqrt{Z_{n-1}/(n-1)}} \sim \mathrm{t}(n-1),$$

which completes the proof. □

3.3 Linear regression with Gaussian errors

In this Section, we consider pairs $(x_1, y_1), \ldots, (x_n, y_n)$ in $\mathbb{R}^p \times \mathbb{R}$ and we assume that there exist $\beta \in \mathbb{R}^{p+1}$ and random variables $\epsilon_1, \ldots, \epsilon_n$ such that

$$\forall i \in \{1, \ldots, n\}, \qquad y_i = \beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p + \epsilon_i.$$

With the notation

$$\mathbf{x}_n = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix}, \qquad \mathbf{y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\epsilon}_n = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

the model rewrites

$$\mathbf{y}_n = \mathbf{x}_n \beta + \boldsymbol{\epsilon}_n. \tag{3.4}$$

3.3.1 Reminder on Least Square estimation

The *Ordinary Least Square* estimator (OLS) of β is obtained by solving the minimisation problem

$$\min_{\beta\in\mathbb{R}^{p+1}} \|\mathbf{y}_n - \mathbf{x}_n\beta\|^2.$$

We recall (see Appendix A) that, as soon as $n \geq p+1$ and the design matrix $\mathbf{x}_n$ has full rank, the matrix $\mathbf{x}_n^\top \mathbf{x}_n$ is invertible and the OLS is uniquely defined by

$$\widehat{\beta} := (\mathbf{x}_n^\top \mathbf{x}_n)^{-1}\mathbf{x}_n^\top \mathbf{y}_n.$$

As a consequence, if the vector $\boldsymbol{\epsilon}_n$ is assumed to be such that $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$ and $\mathrm{Cov}[\boldsymbol{\epsilon}_n] = \sigma^2 I_n$ for some $\sigma^2 > 0$, then the OLS is unbiased and has covariance $\mathrm{Cov}[\beta] = \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}$. Moreover, if in addition $n > p+1$, then the estimator of σ^2 defined by

$$\widehat{\sigma}^2 := \frac{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2}{n-p-1}, \qquad \widehat{\mathbf{y}}_n := \mathbf{x}_n\widehat{\beta},$$

is unbiased.

3.3.2 The case of Gaussian errors

From now on, we assume that $\epsilon_1, \dots, \epsilon_n$ are independent Gaussian variables, which are centered and have variance σ^2: in other words,

$$\boldsymbol{\epsilon}_n \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

The model is thus parametrised by the pair (β, σ^2) and we shall use the notation $\mathbb{P}_{\beta,\sigma^2}$ accordingly. This assumption strengthens the setting of Subsection 3.3.1 and the results stated there remain in force.

Exercise 3.3.1. Show that the condition that $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$ is necessary for the model to be identifiable.

The case $p = 0$ (so without explanatory variable x) corresponds to the Gaussian model of Section 3.2, with $\mu = \beta_0$.

Exercise 3.3.2. Check that if $p = 0$, $\widehat{\beta}$ is the empirical mean of the sample $(y_1, \dots, y_n)$ and $\widehat{\sigma}^2$ is the associated unbiased estimator of the variance.

In the perspective of the previous exercise, the next statement can be seen as a generalisation of Proposition 3.2.2 to the case $p \geq 1$.

Proposition 3.3.3 (Joint law of $(\widehat{\beta}, \widehat{\sigma}^2)$)**.** *Under $\mathbb{P}_{\beta,\sigma^2}$, the estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent, and*

$$\widehat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}), \qquad (n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1).$$

Proof of Proposition 3.3.3. We already know from Subsection 3.3.1 that $\mathbb{E}_{\beta,\sigma^2}[\widehat{\beta}] = \beta$ and $\mathrm{Cov}_{\beta,\sigma^2}[\widehat{\beta}] = \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}$, so to describe the marginal distribution of $\widehat{\beta}$ we only have to check that this vector is Gaussian. But this fact is immediate since $\widehat{\beta}$ is an affine transform of the Gaussian vector $\boldsymbol{\epsilon}_n$.

We now denote by $\Pi = \mathbf{x}_n \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top$ the orthogonal projection of $\mathbb{R}^n$ onto the range of $\mathbf{x}_n$, and set $\Pi^\perp = I_n - \Pi$. By definition,

$$\mathbf{y}_n - \widehat{\mathbf{y}}_n = \Pi^\perp \mathbf{y}_n = \Pi^\perp \left(\mathbf{x}_n\beta + \boldsymbol{\epsilon}_n\right).$$

Since $\mathbf{x}_n\beta$ is in the range of $\mathbf{x}_n$, $\Pi^\perp \mathbf{x}_n\beta = 0$, so that $\mathbf{y}_n - \widehat{\mathbf{y}}_n = \Pi^\perp \boldsymbol{\epsilon}_n$. We deduce from the definition of $\boldsymbol{\epsilon}_n$ again that $(\mathbf{y}_n - \widehat{\mathbf{y}}_n)/\sigma \sim \mathcal{N}_n(0, \Pi^\perp)$, which by Exercise 3.1.7 shows that $\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2/\sigma^2 \sim \chi_2(n-p-1)$ and gives the marginal distribution of $\widehat{\sigma}^2$.

Finally, the independence statement follows from the following observation. On the one hand, $\mathbf{x}_n(\widehat{\beta} - \beta) = \Pi\boldsymbol{\epsilon}_n$, therefore, since $\mathbf{x}_n^\top \mathbf{x}_n$ is invertible, we have $\widehat{\beta} = \beta + (\mathbf{x}_n^\top \mathbf{x}_n)^{-1}\mathbf{x}_n^\top \Pi\boldsymbol{\epsilon}_n$. On the other hand, $\widehat{\sigma}^2$ is a deterministic function of $\Pi^\perp\boldsymbol{\epsilon}_n$. By Cochran's Theorem 3.1.3, the random vectors $\Pi\boldsymbol{\epsilon}_n$ and $\Pi^\perp\boldsymbol{\epsilon}_n$ are independent, thus the estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent. □

3.A Exercises

Training exercises

Exercise 3.A.1 ($\chi_2(n)$ is a Gamma distribution)**.**

1. Let $G \sim \mathcal{N}(0,1)$. Compute the density of G^2.
2. Deduce that $\chi_2(n) = \Gamma(n/2, 1/2)$.

Homework

Exercise 3.A.2. Let $G \sim \mathcal{N}(0, I_d)$ be a standard Gaussian vector in $\mathbb{R}^d$. Let $u, v \in \mathbb{S}^{d-1}$ be two unit vectors and define $X = G^\top u,\ Y = G^\top v$. Suppose that $\rho := \langle u, v\rangle$ satisfies $|\rho| < 1$.

1. Show that (X, Y) is a centered Gaussian vector in $\mathbb{R}^2$ and compute its covariance matrix.
2. Show that there exist $a, b \in \mathbb{R}$ so that $Z = aX + bY$ and $(X, Z) \sim \mathcal{N}(0, I_2)$.
3. What happens if $|\rho| = 1$?

Exercise 3.A.3. Let $Z_n \sim \chi_2(n)$ and $T_n \sim \mathrm{t}(n)$.

1. Show that $Z_n/n \to 1$ in probability.
2. Show that $T_n \to \mathcal{N}(0,1)$ in distribution.
3. Using SciPy as is presented in Lecture 1, plot on the same graph the densities of the $\mathcal{N}(0,1)$ and $\mathrm{t}(1)$, $\mathrm{t}(5)$ and $\mathrm{t}(10)$ distributions.

Supplementary exercises

Exercise 3.A.4 (A converse to Proposition 3.2.2)**.** Let $n \geq 2$ and $X_1, \dots, X_n$ be iid random variables with finite second-order moment. Set $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$. We assume that the random variables $\overline{X}_n$ and S_n^2 are independent. Our goal is to show that, necessarily, $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. For notational simplicity, we denote by $\psi : \mathbb{R} \to \mathbb{C}$ the characteristic function of X_1. To avoid technicality, it is also simpler to assume that for all $u \in \mathbb{R}$, $\psi(u) \neq 0$, although careful readers will realise that this property will actually be *proved*.

1. For any $u \in \mathbb{R}$, express $\mathbb{E}[S_n^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}]$ in terms of σ^2 and $\psi(u)$.
2. On the other hand, show that
$$\mathbb{E}\left[S_n^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] = \mathbb{E}\left[X_1^2 \mathrm{e}^{\mathrm{i}uX_1}\right] \psi(u)^{n-1} - \left(\mathbb{E}\left[X_1 \mathrm{e}^{\mathrm{i}uX_1}\right]\right)^2 \psi(u)^{n-2}.$$
Hint: you may first check that $S_n^2 = \frac{1}{n}\sum_{j=1}^n X_j^2 - \frac{1}{n(n-1)}\sum_{j\neq k} X_j X_k$.
3. Express the quantities $\mathbb{E}[X_1 \mathrm{e}^{\mathrm{i}uX_1}]$ and $\mathbb{E}[X_1^2 \mathrm{e}^{\mathrm{i}uX_1}]$ in terms of $\psi'(u)$ and $\psi''(u)$.
4. Compute the derivative of the function $f : \mathbb{R} \to \mathbb{C}$ defined by $f(u) = \psi'(u)/\psi(u)$ and express $f(u)$ in terms of μ and σ^2.
5. Conclude.

Lecture 4

Confidence Intervals

Contents

In Lectures 2 and 3, we constructed estimators Z_n of some parameter $g(\theta)$, which to a given realisation $\mathbf{x}_n$ of the sample associate a value $z_n(\mathbf{x}_n)$. However, the mere knowledge of this value is not very useful, as it does not provide any indication on the accuracy of the estimation. In this Lecture, we introduce the notion of *confidence interval* which provides such an indication.

Throughout this Lecture, a parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is given. Our purpose is to estimate the quantity $g(\theta)$, which is assumed to take scalar values.

4.1 General definitions

Let $\alpha \in (0, 1/2)$ be the desired level of precision for our confidence intervals. The classical values for α are 10%, 5% and 1%.

Definition 4.1.1 (Confidence interval)**.** *A* confidence interval *with level* $1-\alpha$ *for* $g(\theta)$ *is an interval* $I_n = [I_n^-, I_n^+]$ *such that:*

- *the boundaries* I_n^- *and* I_n^+ *are statistics,*
- *for all* $\theta \in \Theta$, $\mathbb{P}_\theta(g(\theta) \in I_n) = 1 - \alpha$.

It is to be emphasised that in the event $\{g(\theta) \in I_n\}$, it is the interval I_n which is random, and not the quantity $g(\theta)$.

Sometimes it is difficult, tedious or not possible to construct confidence intervals in the sense of Definition 4.1.1, which are called *exact*, and one has to resort to weaker notions. Hence, an interval I_n whose boundaries I_n^-, I_n^+ are statistics is said to be:

- an *asymptotic* confidence interval if for all $\theta \in \Theta$, $\lim_{n\to+\infty} \mathbb{P}_\theta(g(\theta) \in I_n) = 1 - \alpha$,
- an *approximate* confidence interval[1] if for all $\theta \in \Theta$, $\mathbb{P}_\theta(g(\theta) \in I_n) \geq 1 - \alpha$.

Of course, both definitions can be combined to lead to the notion of *asymptotic approximate confidence interval*, such that $\lim_{n\to+\infty} \mathbb{P}_\theta(g(\theta) \in I_n) \geq 1 - \alpha$.

[1] *Intervalle de confiance par excès* en français.

Remark 4.1.2 (Confidence region)**.** *When $g(\theta) \in \mathbb{R}^d$ with arbitrary d, the multidimensional generalisation of a confidence interval is called a* confidence region*. Formally, a* confidence region *of level $1-\alpha$ is a function C_n from E^n to the set of (measurable) subsets of $\mathbb{R}^d$, such that:*

- *for any $z \in \mathbb{R}^d$, the random variable $\mathbb{1}_{\{z \in C_n(\mathbf{X}_n)\}}$ is a statistic,*
- *for any $\theta \in \Theta$, $\mathbb{P}_\theta(g(\theta) \in C_n(\mathbf{X}_n)) = 1-\alpha$.*

To alleviate notation, it is usual to write $\{g(\theta) \in C_n\}$ rather than $\{g(\theta) \in C_n(\mathbf{X}_n)\}$. Corresponding notions of asymptotic and approximate confidence regions are defined accordingly.

In the sequel of the Lecture we only focus on confidence intervals, where $d = 1$, but most of the ideas and techniques can be adapted to the multidimensional setting without difficulty. An example of a multidimensional confidence region is studied in Exercise 4.2.9.

4.2 Construction of exact confidence intervals

When one possesses an estimator Z_n of $g(\theta)$ whose law under $\mathbb{P}_\theta$ is explicit, it is generally possible to construct exact confidence intervals for $g(\theta)$. We first introduce a few notions, then detail the method on the example of the Gaussian model.

4.2.1 Pivotal function

Definition 4.2.1 (Free random variable)**.** *A random variable Q is* free *if its law under $\mathbb{P}_\theta$ does not depend on θ.*

Definition 4.2.1 does not require the random variable Q to be a statistic, and in general we use free random variables Q *which depend* on the parameter θ. For example, in the Exponential model, the random variables

$$Y_i = \lambda X_i, \qquad i = 1, \dots, n,$$

are iid according to the $\mathcal{E}(1)$ law, and therefore they are free. Likewise, the random variable

$$\overline{Y}_n = \lambda \overline{X}_n$$

has law $\Gamma(n, n)$ and therefore it is free (see Exercise 1.3.12).

Definition 4.2.2 (Pivotal function)**.** *A* pivotal function *for $g(\theta)$ is a function $\pi_n : E^n \times g(\Theta) \to \mathbb{R}$ such that $\pi_n(\mathbf{X}_n; g(\theta))$ is free.*

In the Exponential model, the function π_n defined by $\pi_n(\mathbf{x}_n; \lambda) = \lambda \overline{x}_n$ is pivotal. More examples of pivotal (or not) functions are given in the next subsection.

4.2.2 Example in the Gaussian model

We now detail the construction of a confidence interval for the mean μ in the Gaussian model. We use $\overline{X}_n$ as an estimator of μ. The law of $\overline{X}_n$ under $\mathbb{P}_{\mu,\sigma^2}$ is $\mathcal{N}(\mu, \sigma^2/n)$. As a consequence, the random variable

$$Y_n = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

is free, with law $\mathcal{N}(0,1)$. Still, this does not mean that the function

$$\pi_n(\mathbf{x}_n; \mu) = \frac{\overline{x}_n - \mu}{\sqrt{\sigma^2/n}}$$

is pivotal, because following Definition 4.2.2 it should only depend on the parameter (μ, σ^2) through μ.

Let us momentarily assume that σ^2 is known. Then π_n becomes a pivotal function. As a consequence, for all $a, b \in \mathbb{R}$ such that $a < b$, we have

$$\mathbb{P}_{\mu,\sigma^2}\left(Y_n \in [a,b]\right) = \frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x,$$

which rewrites

$$\mathbb{P}_{\mu,\sigma^2}\left(\overline{X}_n - b\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X}_n - a\sqrt{\frac{\sigma^2}{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

For any choice of a and b such that

$$\frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = 1 - \alpha, \tag{4.1}$$

the interval $[\overline{X}_n - b\sqrt{\sigma^2/n}, \overline{X}_n - a\sqrt{\sigma^2/n}]$ is a confidence interval with level $1-\alpha$ for μ. Thus, we have constructed infinitely many confidence intervals, parametrised by a, b satisfying the constraint (4.1).

Recall that we denote by ϕ_r the quantile of order r of the standard Gaussian distribution $\mathcal{N}(0,1)$. Clearly, a and b satisfy (4.1) if and only if there exists $r \in [0, \alpha]$ such that

$$a = \phi_r, \qquad b = \phi_{r+1-\alpha}.$$

For any such pair (a,b), with the confidence interval $[\overline{X}_n - b\sqrt{\sigma^2/n}, \overline{X}_n - a\sqrt{\sigma^2/n}]$, the probability to *underestimate* μ is

$$\mathbb{P}_{\mu,\sigma^2}\left(\mu > \overline{X}_n - a\sqrt{\sigma^2/n}\right) = r,$$

and the probability to *overestimate* μ is

$$\mathbb{P}_{\mu,\sigma^2}\left(\mu < \overline{X}_n - b\sqrt{\sigma^2/n}\right) = \alpha - r.$$

By construction, the sum of these two probabilities is α. If there is no reason to favour the risk of under- or overestimation, then it is customary to choose a, b so that both probabilities are equal to $\alpha/2$, which amounts to taking $r = \alpha/2$ and thus

$$a = \phi_{\alpha/2} = -\phi_{1-\alpha/2}, \qquad b = \phi_{1-\alpha/2},$$

so the confidence interval finally writes

$$\left[\overline{X}_n - \phi_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{X}_n + \phi_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right].$$

Some commonly employed values of $\phi_{1-\alpha/2}$ are gathered in Figure 4.1.

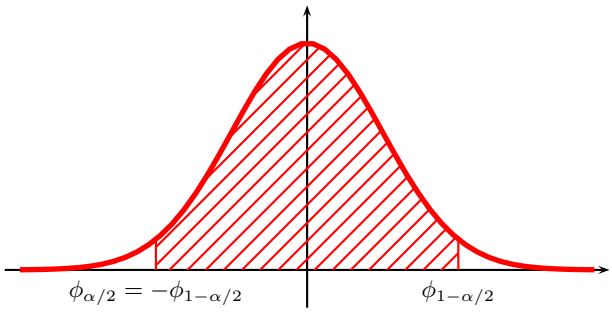
In addition, this choice of (a,b) is the one which provides the *smallest* confidence interval, as is shown in the next exercise.

Exercise 4.2.3. Let $\alpha \in (0, 1/2)$. Let $p : \mathbb{R} \to \mathbb{R}$ be an *even*[2] probability density, which is nonincreasing on $[0, +\infty)$. We consider the minimisation problem

$$\min b - a \qquad \text{such that} \qquad \int_{x=a}^{b} p(x)\mathrm{d}x = 1 - \alpha.$$

1. Write the optimality condition of this problem.

[2] *paire* en français.



α	$\phi_{1-\alpha/2}$
10%	1.65
5%	1.96
1%	2.58

Figure 4.1: Quantiles of the standard Gaussian distribution. The hatched area on the figure is equal to $1-\alpha$.

2. Deduce that the unique solution is given by $-a = b = q_{1-\alpha/2}$, where for any $r \geq 1/2$, q_r is the (lowest) quantile of order r of p.

However, there are situations in which one may prefer to avoid the risk of under- or overestimation in priority. Assume for instance that $X_1, \ldots, X_n$ represent noisy observations of the maximum weight μ that an aircraft can carry safely, obtained as the output of a numerical simulation with uncertain inputs. As an engineer for the airline, if you underestimate μ, then this means that your plane will take fewer passengers than it could, so you lose a bit of profit. But if you overestimate μ, then you will take too many passengers and your plane will crash. So it seems that you should try to avoid overestimation in priority.

Exercise 4.2.4. In the Gaussian model with known variance σ^2, what is the confidence interval for μ which minimises the probability of overestimation of μ?

We now come back to the construction of the confidence interval in the general case where we assume that σ^2 is not known. Then we have to find another pivotal function, which will no longer depend on σ^2. A simple idea consists in replacing σ^2 by an estimator of the variance: taking the unbiased estimator of the variance S_n^2, we are led to considering the random variable

$$Y_n' = \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}},$$

which by Proposition 3.2.4 is free and has distribution $\mathrm{t}(n-1)$. Thus, the function

$$\pi_n'(\mathbf{x}_n; \mu) = \frac{\overline{x}_n - \mu}{\sqrt{s_n^2/n}}, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2,$$

which no longer depends on σ^2, is pivotal. As a consequence, for all $a, b \in \mathbb{R}$ such that $a < b$,

$$\mathbb{P}_{\mu,\sigma^2}\left(Y_n' \in [a,b]\right) = \mathbb{P}_{\mu,\sigma^2}\left(\overline{X}_n - b\sqrt{\frac{S_n^2}{n}} \leq \mu \leq \overline{X}_n - a\sqrt{\frac{S_n^2}{n}}\right) = \int_{x=a}^{b} p_{n-1}(x)\mathrm{d}x,$$

where p_{n-1} is the density of the law $\mathrm{t}(n-1)$. Once again, as soon as a and b are chosen so that

$$\int_{x=a}^{b} p_{n-1}(x)\mathrm{d}x = 1-\alpha,$$

we get a confidence interval with level $1-\alpha$ for μ. The choice which equilibrates the risks of under- and overestimation is

$$a = t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}, \qquad b = t_{n-1,1-\alpha/2},$$

where $t_{n-1,r}$ denotes the quantile of order r of the Student distribution $\mathrm{t}(n-1)$.

As a conclusion, the confidence interval with level $1-\alpha$ for μ, and with equal probability of under- and overestimation, is

$$I_n = \begin{cases} \left[\overline{X}_n - \phi_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{X}_n + \phi_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] & \text{if } \sigma^2 \text{ is known,} \\ \left[\overline{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}, \overline{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{S_n^2}{n}}\right] & \text{if } \sigma^2 \text{ is not known.} \end{cases}$$

Notice that $t_{n-1,1-\alpha/2}$ is larger than $\phi_{1-\alpha/2}$ (this was observed numerically in Exercise 3.A.3 and it is proved in Exercise 4.A.6), so that, up to the fluctuations in the estimation of σ^2 by S_n^2, the latter confidence interval is larger than the former. This is natural: less information is available in the second case, so that there is more uncertainty on the parameter μ.

4.2.3 Summary of the method

We now summarise the method to construct an exact confidence interval for $g(\theta)$.

(i) Find a pivotal function $\pi_n(\mathbf{x}_n; g(\theta))$; denote by Q_n the free random variable $\pi_n(\mathbf{X}_n; g(\theta))$.

(ii) Rewrite the condition $Q_n \in [a,b]$ as $g(\theta) \in I_n$, where the boundaries of I_n are statistics.

(iii) Take $a, b \in \mathbb{R}$ which satisfy the constraint that $\mathbb{P}(Q_n \in [a,b]) = 1-\alpha$, which amounts to taking $a = q_{n,r}$ and $b = q_{n,r+1-\alpha}$, where $r \in [0,\alpha]$ and $q_{n,r}$ is the quantile of order r of Q_n. Then, fix r depending on whether you want your confidence interval to have equal probability of under- and overestimation, or to minimise the probability of under- or overestimation in priority.

In the sequel of this Lecture, for the sake of simplicity, we focus on confidence intervals with equal probability of under- and overestimation, but our computation can always be adapted to nonsymmetric cases.

Exercise 4.2.5. In the Exponential model, find a confidence interval with level $1-\alpha$ for λ, with equal risks of under- and overestimation.

4.2.4 Application: the Gaussian linear model

We see how to apply this method in the case of the Gaussian linear model of Section 3.3. Recall that we observe pairs $(x_i, y_i)_{1\leq i\leq n}$ in $\mathbb{R}^p \times \mathbb{R}$ which are assumed to satisfy the relation

$$\forall i \in \{1,\dots,n\}, \qquad y_i = \beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p + \epsilon_i,$$

where $\beta \in \mathbb{R}^{p+1}$ is unknown and $\epsilon_1,\dots,\epsilon_n$ are independent $\mathcal{N}(0,\sigma^2)$ variables. With the notation of Section 3.3, under the assumption that $n > p+1$, the OLS of β is

$$\widehat{\beta} := (\mathbf{x}_n^\top \mathbf{x}_n)^{-1} \mathbf{x}_n^\top \mathbf{y}_n,$$

and the unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 := \frac{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2}{n-p-1}.$$

We recall from Proposition 3.3.3 that under $\mathbb{P}_{\beta,\sigma^2}$, the estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent, and

$$\widehat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}), \qquad (n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1).$$

Prediction intervals. We want to predict the outcome y corresponding to a value of x which has not been observed yet. Indeed, for a new vector $x_{n+1} \in \mathbb{R}^p$, the LSE provides a natural predictor

$$\widehat{y}_{n+1} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{n+1}^j$$

of the corresponding value

$$y_{n+1} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{n+1}^j + \epsilon_{n+1}.$$

In the Gaussian model, the precision of this predictor can be measured by a *prediction interval*, namely an interval I_n whose boundaries are statistics and such that

$$\mathbb{P}(y_{n+1} \in I_n) = 1 - \alpha.$$

Notice that, in contrast with the notion of confidence interval introduced in this lecture, here the quantity y_{n+1} depends on the unknown parameters β and σ^2, but also on the realisation of the random variable ϵ_{n+1}, therefore we use a different terminology.

Proposition 4.2.6 (Prediction with linear regression)**.** *Given $x_{n+1} = (x_{n+1}^1, \dots, x_{n+1}^p) \in \mathbb{R}^{1\times p}$, let us define the row vector $x'_{n+1} \in \mathbb{R}^{1\times(p+1)}$ by*

$$x'_{n+1} = (1, x_{n+1}^1, \dots, x_{n+1}^p),$$

and write

$$\kappa = 1 + x'_{n+1}\Big(((x'_n)^\top x'_n)^{-1}(x'_{n+1})^\top\Big) \geq 1.$$

A prediction interval of level $1-\alpha$ for y_{n+1} is given by

$$\Big[\widehat{y}_{n+1} - t_{n-p-1,\,1-\alpha/2}\sqrt{\widehat{\sigma}^2\kappa},\ \widehat{y}_{n+1} + t_{n-p-1,\,1-\alpha/2}\sqrt{\widehat{\sigma}^2\kappa}\Big],$$

where $t_{m,r}$ denotes the quantile of order r of Student's distribution with m degrees of freedom.

Proof. Writing $\widehat{y}_{n+1} = x'_{n+1}\widehat{\beta}$ and using Proposition 3.3.3, we get

$$\widehat{y}_{n+1} \sim \mathcal{N}\Big(x'_{n+1}\beta,\ \sigma^2 x'_{n+1}((x'_n)^\top x'_n)^{-1}(x'_{n+1})^\top\Big).$$

Since $\epsilon_{n+1} \sim \mathcal{N}(0,\sigma^2)$ and is independent of $\epsilon_1, \dots, \epsilon_n$, we deduce that

$$y_{n+1} - \widehat{y}_{n+1} = x'_{n+1}\beta + \epsilon_{n+1} - \widehat{y}_{n+1} \sim \mathcal{N}(0, \sigma^2\kappa).$$

By Proposition 3.3.3 again, we deduce that

$$\frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{\widehat{\sigma}^2\kappa}} \sim t_{n-p-1},$$

which completes the proof. □

Remark 4.2.7. *One may also be interested in the prediction of $x'_{n+1}\beta$, without taking into account the noise ϵ_{n+1} associated with the $(n+1)$-th observation. This is the case in the example of polynomial regression where for a given x, one may want to recover the underlying signal $f(x)$ rather than deriving a prediction interval of the noisy observation $f(x) + \epsilon$.*

Exercise 4.2.8 (Prediction interval of the signal)**.** Show that a confidence interval for $x'_{n+1}\beta$ with level $1-\alpha$ is given by

$$\Big[\widehat{y}_{n+1} - t_{n-p-1,\,1-\alpha/2}\sqrt{\widehat{\sigma}^2\lambda},\ \widehat{y}_{n+1} + t_{n-p-1,\,1-\alpha/2}\sqrt{\widehat{\sigma}^2\lambda}\Big],$$

with

$$\lambda = x'_{n+1}((x'_n)^\top x'_n)^{-1}(x'_{n+1})^\top = \kappa - 1.$$

Confidence ellipsoids for β in the linear regression. We now want to construct a confidence region for the unknown parameter $\beta \in \mathbb{R}^{p+1}$.

Exercise 4.2.9 (Confidence ellipsoids for β in the linear regression)**.** Let $\lambda_1 \geq \cdots \geq \lambda_{p+1} > 0$ be the eigenvalues of the matrix $\mathbf{x}_n^\top \mathbf{x}_n$, and let $(e_1, \ldots, e_{p+1})$ be an associated orthonormal basis.

1. Show that the random variables $\sqrt{\lambda_j}\langle \widehat{\beta} - \beta, e_j \rangle$ are iid according to $\mathcal{N}(0, \sigma^2)$.
2. If σ^2 is assumed to be known, find $a > 0$ so that the *ellipsoid*

$$\mathcal{C}(a) = \left\{ \beta \in \mathbb{R}^{p+1} : \sum_{j=1}^{p+1} \frac{\lambda_j}{\sigma^2} \langle \widehat{\beta} - \beta, e_j \rangle^2 \leq a \right\}$$

   is such that $\mathbb{P}(\beta \in \mathcal{C}(a)) = 1 - \alpha$ for a given level $\alpha > 0$.
3. If σ^2 is no longer assumed to be known and has to be estimated by $\widehat{\sigma}^2$, how to modify the definition of $\mathcal{C}(a)$ for the identity $\mathbb{P}(\beta \in \mathcal{C}(a)) = 1 - \alpha$ to remain valid?

4.3 Construction of asymptotic confidence intervals

The construction of exact confidence intervals presented in Section 4.2 requires to find a pivotal function whose law is exactly known. In this Section, we present an asymptotic version of this approach, suited to the case where asymptotic properties of an estimator Z_n of $g(\theta)$ are available, in particular consistency and an expression for the asymptotic variance.

4.3.1 Asymptotically normal estimator

We first state a general result for asymptotically normal estimators. We recall that ϕ_r denotes the quantile of order r of the standard Gaussian distribution $\mathcal{N}(0, 1)$, see Figure 4.1.

Proposition 4.3.1 (Asymptotic confidence intervals)**.** *Let Z_n be a consistent and asymptotically normal estimator of $g(\theta)$, with asymptotic variance $V(\theta)$. Assume that a consistent estimator $\widehat{V}_n$ of $V(\theta)$ is available. Then, for all $\alpha \in (0, 1/2)$,*

$$I_n = \left[Z_n - \phi_{1-\alpha/2} \sqrt{\frac{\widehat{V}_n}{n}}, Z_n + \phi_{1-\alpha/2} \sqrt{\frac{\widehat{V}_n}{n}} \right]$$

is an asymptotic confidence interval with level $1 - \alpha$ for $g(\theta)$.

In general, it is not difficult to find a consistent estimator for $V(\theta)$: as soon as V is continuous and $\widehat{\theta}_n$ is a consistent estimator of θ, one can take the plug-in estimator $\widehat{V}_n = V(\widehat{\theta}_n)$. If there is no such estimator available, the procedure of *variance stabilisation* described in Exercise 4.A.5 can be applied.

Before detailing the proof of Proposition 4.3.1, we first recall a result on the convergence in distribution, which follows from Proposition 1.4.6 (v) and will often be used in the sequel of the course.

Lemma 4.3.2 (Convergence in distribution and convergence of probabilities)**.** *Let Y_n be a sequence of random variables which converges in distribution to a random variable Y which possesses a density. For any interval $[a, b]$, $\mathbb{P}(Y_n \in [a, b])$ converges to $\mathbb{P}(Y \in [a, b])$.*

Proof of Proposition 4.3.1. We start from the asymptotical normality of Z_n, which writes

$$\sqrt{\frac{n}{V(\theta)}} (Z_n - g(\theta)) \to \mathcal{N}(0, 1), \qquad \text{in distribution.}$$

Since $\widehat{V}_n$ converges to $V(\theta)$ in probability, Slutsky's Theorem implies

$$\sqrt{\frac{n}{\widehat{V}_n}}\,(Z_n - g(\theta)) = \sqrt{\frac{V(\theta)}{\widehat{V}_n}} \times \sqrt{\frac{n}{V(\theta)}}\,(Z_n - g(\theta)) \to \mathcal{N}(0,1), \qquad \text{in distribution,}$$

as well. As a consequence, by Lemma 4.3.2 we get, for all $a, b \in \mathbb{R}$ such that $a \leq b$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{\frac{n}{\widehat{V}_n}}\,(Z_n - g(\theta)) \in [a,b]\right) = \frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

Following the same arguments as in the construction of exact confidence intervals for the Gaussian model, we take $b = -a = \phi_{1-\alpha/2}$, which results in the claimed confidence interval. □

Example 4.3.3 (The Bernoulli model)**.** *We want to employ the estimator $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$ to construct an asymptotic confidence interval for p. This estimator is (strongly) consistent, and asymptotically normal with asymptotic variance $V(p) = p(1-p)$. As a consequence, a consistent estimator of the asymptotic variance is given by $\widehat{p}_n\,(1-\widehat{p}_n)$, and we get that*

$$I_n = \left[\widehat{p}_n - \phi_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}, \widehat{p}_n + \phi_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n\,(1-\widehat{p}_n)}{n}}\right]$$

is an asymptotic confidence interval with level $1-\alpha$ for p.

Remark 4.3.4. *For the Bernoulli model, when the size n of the sample is too small for the asymptotic approach to be used, approximate confidence intervals can be constructed based on Chebychev's or Hoeffding's inequality, see Section 4.4. However, there is no pivotal function to obtain exact confidence intervals.*

4.3.2 Asymptotic confidence interval without asymptotic normality

The sketch of the proof of Proposition 4.3.1 allows to construct asymptotic confidence intervals even when the estimator Z_n is not asymptotically normal. As an example, recall that for the Uniform model of Exercise 2.3.3, the estimator $\widehat{\theta}_n = \max_{1\leq i\leq n} X_i$ is strongly consistent, and satisfies

$$n(\theta - \widehat{\theta}_n) \to \mathcal{E}(1/\theta), \qquad \text{in distribution.}$$

As a consequence,

$$n\left(1 - \frac{\widehat{\theta}_n}{\theta}\right) \to \mathcal{E}(1), \qquad \text{in distribution,}$$

which by Lemma 4.3.2 implies, for all $a, b \geq 0$ such that $a \leq b$,

$$\lim_{n\to+\infty} \mathbb{P}\left(n\left(1 - \frac{\widehat{\theta}_n}{\theta}\right) \in [a,b]\right) = \int_{x=a}^{b} \exp(-x)\mathrm{d}x = \exp(-a) - \exp(-b).$$

On the other hand,

$$n\left(1 - \frac{\widehat{\theta}_n}{\theta}\right) \in [a,b] \quad \text{if and only if} \quad \theta \in \left[\frac{\widehat{\theta}_n}{1-a/n}, \frac{\widehat{\theta}_n}{1-b/n}\right].$$

Let us fix a and b such that $\exp(-a) - \exp(-b) = 1-\alpha$, for instance $a = 0$, $b = -\log\alpha$. As a conclusion, an asymptotic confidence interval for θ is

$$I_n = \left[\widehat{\theta}_n, \frac{\widehat{\theta}_n}{1 + (\log\alpha)/n}\right].$$

Remark 4.3.5. *The choice $a = 0$ allows to include $\widehat{\theta}_n$ in the confidence interval. Furthermore, since, by construction, $\widehat{\theta}_n \leq \theta$, it ensures that if θ is not in the interval then it is necessarily larger than the upper-bound of the interval.*

4.4 Construction of approximate confidence intervals

In this Section, we return to nonasymptotic approaches and detail the construction of approximate confidence intervals when pivotal functions are not available, as is the case for the Bernoulli model. In such a case, one can construct approximate confidence intervals, thanks to *concentration inequalities*.

Definition 4.4.1 (Concentration inequality)**.** *A* concentration inequality *for a random variable Y is an inequality of the form*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq r) \leq c_Y(r),$$

for some concentration *function c_Y converging to 0 when $r \to +\infty$.*

If Z_n is an unbiased estimator of $g(\theta)$ and satisfies a concentration inequality with concentration function c_{Z_n}, uniformly in θ — that is to say

$$\forall r > 0, \qquad \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|Z_n - g(\theta)| \geq r \right) \leq c_{Z_n}(r),$$

then any $r_{n,\alpha} > 0$ such that $c_{Z_n}(r_{n,\alpha}) \leq \alpha$ yields the approximate confidence interval

$$[Z_n - r_{n,\alpha}, Z_n + r_{n,\alpha}]$$

for $g(\theta)$.

In the sequel of this Section, we present two concentration inequalities, the Bienaymé–Chebychev inequality and the Hoeffding inequality, and we show how to use them to derive approximate confidence intervals in the Bernoulli model.

4.4.1 The Bienaymé–Chebychev inequality

A famous concentration inequality for random variables Y such that $\mathbb{E}[Y^2] < +\infty$ is the *Bienaymé–Chebychev inequality*

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

which follows from Markov's inequality. We first explain how to obtain an approximate confidence interval from such an inequality, for the Bernoulli model in which we use $\overline{X}_n$ as an estimator of p. For all $r > 0$, the Bienaymé–Chebychev inequality yields

$$\mathbb{P}_p \left(|\overline{X}_n - p| \geq \frac{r}{\sqrt{n}} \right) \leq \frac{\mathrm{Var}_p[\overline{X}_n]}{r^2/n} = \frac{p(1-p)}{r^2}.$$

As a consequence, taking r such that

$$\frac{p(1-p)}{r^2} \leq \alpha \tag{4.2}$$

ensures that

$$\mathbb{P}_p \left(p \in \left[\overline{X}_n - \frac{r}{\sqrt{n}}, \overline{X}_n + \frac{r}{\sqrt{n}} \right] \right) \geq 1 - \alpha.$$

We now have to find such a value of r which does not depend on p. For bounded random variables, the next lemma provides a universal bound on the variance.

Lemma 4.4.2 (Universal bound on the variance)**.** *Let Y be a random variable taking its values in $[0, 1]$. Then*

$$\mathrm{Var}(Y) \leq \frac{1}{4}.$$

Proof. Since $Y \in [0, 1]$, we have $Y^2 \leq Y$, almost surely, and therefore

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \leq \mathbb{E}[Y] - \mathbb{E}[Y]^2 = \mathbb{E}[Y](1 - \mathbb{E}[Y]) \leq \max_{y \in [0,1]} y(1-y) = \frac{1}{4}. \qquad \square$$

Exercise 4.4.3. Find a variable $Y \in [0,1]$ such that $\mathrm{Var}(Y) = 1/4$.

As a consequence of Lemma 4.4.2, it suffices to take

$$r = \frac{1}{2\sqrt{\alpha}} = r^{\mathrm{BC}}(\alpha)$$

to ensure that (4.2) holds whatever the value of p. Thus, we get the approximate confidence interval

$$I_n^{\mathrm{BC}} = \left[\overline{X}_n - \frac{r^{\mathrm{BC}}(\alpha)}{\sqrt{n}}, \overline{X}_n + \frac{r^{\mathrm{BC}}(\alpha)}{\sqrt{n}}\right].$$

As soon as $\mathbb{E}[|Y|^q] < +\infty$ for some $q > 2$, the Bienaymé–Chebychev's inequality can be immediately improved to

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq r) \leq \frac{\mathbb{E}[|Y - \mathbb{E}[Y]|^q]}{r^q},$$

and therefore if a bound on the quantity $\mathbb{E}[|Y - \mathbb{E}[Y]|^q]$, which does not depend on the parameter, is available, then an approximate confidence interval with length of order $\alpha^{1/q}$ can be obtained, and the larger q, the smaller $\alpha^{1/q}$. More generally, the faster c_Y converges to 0, the more useful the inequality is, in the sense that it provides sharper confidence intervals. In this respect, the *Hoeffding inequality* is often employed for bounded random variables.

4.4.2 The Hoeffding inequality

The Bienaymé–Chebychev inequality yields *polynomial concentration*, in $1/r^2$. The more sophisticated Hoeffding inequality yields *Gaussian concentration*, in $\exp(-2r^2)$.

Lemma 4.4.4 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be iid random variables taking their values in $[0,1]$. For all $n \geq 1$, for all $r \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq r\sqrt{n}\right) \leq \exp(-2r^2).$$

Proof. We first define $Y_1 = X_1 - \mathbb{E}[X_1]$ and show that, for all $\lambda \geq 0$,

$$\mathbb{E}[\exp(\lambda Y_1)] \leq \exp(\lambda^2/8).$$

To this aim, let $F(\lambda) = \log \mathbb{E}[\exp(\lambda Y_1)]$. Then

$$F'(\lambda) = \frac{\mathbb{E}[Y_1 \exp(\lambda Y_1)]}{\mathbb{E}[\exp(\lambda Y_1)]},$$

and

$$\begin{aligned} F''(\lambda) &= \frac{\mathbb{E}[Y_1^2 \exp(\lambda Y_1)]\mathbb{E}[\exp(\lambda Y_1)] - \mathbb{E}[Y_1 \exp(\lambda Y_1)]^2}{\mathbb{E}[\exp(\lambda Y_1)]^2} \\ &= \frac{1}{\mathbb{E}[\exp(\lambda Y_1)]}\mathbb{E}\left[\left(Y_1 - \frac{\mathbb{E}[Y_1 \exp(\lambda Y_1)]}{\mathbb{E}[\exp(\lambda Y_1)]}\right)^2 \exp(\lambda Y_1)\right] \\ &= \frac{1}{\mathbb{E}[\exp(\lambda Y_1)]}\mathbb{E}\left[\left(X_1 - \frac{\mathbb{E}[X_1 \exp(\lambda Y_1)]}{\mathbb{E}[\exp(\lambda Y_1)]}\right)^2 \exp(\lambda Y_1)\right]. \end{aligned}$$

Using the same arguments as in the proof of Lemma 4.4.2, we get $F''(\lambda) \leq 1/4$. Noting that $F'(0) = \mathbb{E}[Y_1] = 0$ and $F(0) = \log 1 = 0$, and integrating twice yields $F(\lambda) \leq \lambda^2/8$, whence the claimed result.

To complete the proof, we now write, for all $\lambda > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq r\sqrt{n}\right) &= \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right) \geq \exp(\lambda r\sqrt{n})\right)\\ &\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right]\exp(-\lambda r\sqrt{n})\\ &= \mathbb{E}[\exp(\lambda Y_1)]^n \exp(-\lambda r\sqrt{n})\\ &\leq \exp\left(\frac{\lambda^2 n}{8} - \lambda r\sqrt{n}\right),\end{aligned}$$

where we have used the fact that $y \mapsto \exp(\lambda y)$ is increasing at the first line and the Markov inequality at the second line. The minimum of the quantity $\lambda^2 n/8 - \lambda r\sqrt{n}$ for $\lambda > 0$ is reached at the value $\lambda = 4r/\sqrt{n}$, at which point it equals $-2r^2$, which completes the proof. □

Remark 4.4.5. *The proof of Lemma 4.4.4 conveys two remarks.*

- *It is not surprising that the bound on $F''(\lambda)$ follows from the same arguments as in the proof of Lemma 4.4.2. Indeed, on can define a probability measure $\mathbb{P}^\lambda$ by the identity*

$$\mathbb{P}^\lambda(A) = \frac{\mathbb{E}[\mathbb{1}_A \exp(\lambda Y_1)]}{\mathbb{E}[\exp(\lambda Y_1)]},$$

  for all events A, and then check that $F'(\lambda) = \mathbb{E}^\lambda[Y_1]$, $F''(\lambda) = \mathrm{Var}^\lambda[Y_1] = \mathrm{Var}^\lambda[X_1]$. Since X_1 takes its values in $[0,1]$, the bound on $F''(\lambda)$ is in fact a direct consequence of Lemma 4.4.2.

- *The combination of: the application of the increasing function $y \mapsto \exp(\lambda y)$; the use of Markov's inequality; and the optimisation of the result over the values of $\lambda > 0$; is a classical method, called* Chernoff's method.

Lemma 4.4.4 is often applied under the form of the following corollary.

Corollary 4.4.6 (Application of Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be iid random variables taking their values in $[0,1]$. For all $n \geq 1$, for all $r \geq 0$,*

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geq r/\sqrt{n}\right) \leq 2\exp(-2r^2).$$

Proof. The proof follows from the union bound

$$\begin{aligned}\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geq r/\sqrt{n}\right) &= \mathbb{P}\left(\left\{\overline{X}_n - \mathbb{E}[X_1] \geq r/\sqrt{n}\right\} \cup \left\{\overline{X}_n - \mathbb{E}[X_1] \leq -r/\sqrt{n}\right\}\right)\\ &\leq \mathbb{P}\left(\overline{X}_n - \mathbb{E}[X_1] \geq r/\sqrt{n}\right) + \mathbb{P}\left((1-\overline{X}_n) - \mathbb{E}[1-X_1] \geq r/\sqrt{n}\right)\\ &\leq 2\exp(-2r^2),\end{aligned}$$

where we have applied Hoeffding's inequality to the random variables $X_1, \ldots, X_n$ as well as to the random variables $1 - X_1, \ldots, 1 - X_n$. □

Remark 4.4.7. *When the variables $X_1, \ldots, X_n$ take their values in some interval $[a,b]$, then the conclusion of Corollary 4.4.6 writes*

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geq r/\sqrt{n}\right) \leq 2\exp(-2r^2/(b-a)^2).$$

Applying Corollary 4.4.6 to the Bernoulli model yields, for all $r \geq 0$,

$$\mathbb{P}_p\left(\left|\overline{X}_n - p\right| \geq \frac{r}{\sqrt{n}}\right) \leq 2\exp(-2r^2).$$

As a consequence, taking r such that $2\exp(-2r^2) = \alpha$, that is

$$r = \sqrt{-\frac{1}{2}\log\frac{\alpha}{2}} = r^{\mathrm{H}}(\alpha),$$

we get

$$\mathbb{P}_p\left(|\overline{X}_n - p| \leq \frac{r}{\sqrt{n}}\right) \geq 1 - \alpha.$$

In conclusion,

$$I_n^{\mathrm{H}} = \left[\overline{X}_n - \frac{r^{\mathrm{H}}(\alpha)}{\sqrt{n}}, \overline{X}_n + \frac{r^{\mathrm{H}}(\alpha)}{\sqrt{n}}\right]$$

is an approximate confidence interval for p.

The lengths of the approximate confidence intervals obtained by the Bienaymé–Chebychev and the Hoeffding inequalities exhibit the same dependency upon n, namely they are proportional to $1/\sqrt{n}$. The multiplicative coefficients $r^{\mathrm{BC}}(\alpha)$ and $r^{\mathrm{H}}(\alpha)$ are plotted on Figure 4.2 as functions of α. As is expected, Hoeffding's inequality provides narrower, and thus more precise, confidence intervals.

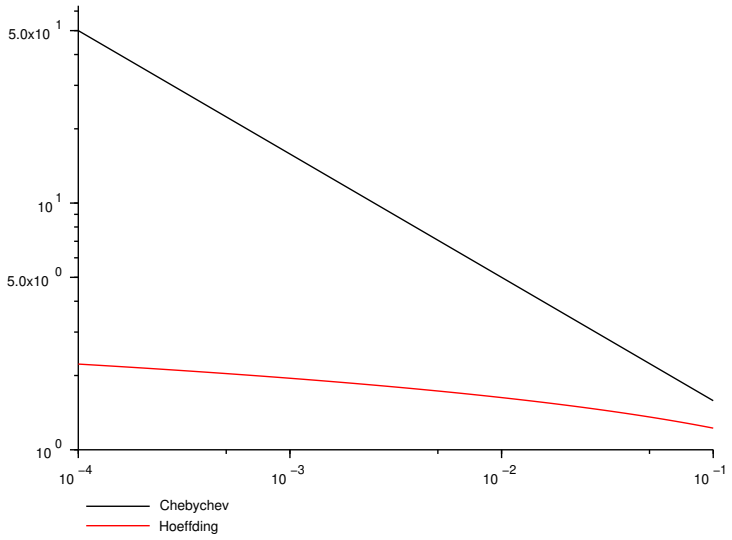


Figure 4.2: Log-log plot of the functions $r^{\mathrm{BC}}(\alpha)$ and $r^{\mathrm{H}}(\alpha)$, for α ranging from 10^{-4} to 10^{-1}.

4.A Exercises

Training exercises

Exercise 4.A.1 (Confidence interval for β_j in linear regression)**.** In the Gaussian linear model introduced in Section 3.3, and studied in Exercise 4.2.9, and assuming that σ^2 is unknown, construct a confidence interval for each β_j, $j \in \{0, \ldots, p\}$, with equal risks of under- and overestimation.

Exercise 4.A.2 (Bonferroni correction for multiple estimation)**.** Assume that, for any α, you are provided with confidence intervals $I_n^1(\alpha) = [I_n^{-,1}(\alpha), I_n^{+,1}(\alpha)]$ and $I_n^2(\alpha) = [I_n^{-,2}(\alpha), I_n^{+,2}(\alpha)]$, with level $1 - \alpha$, for two real-valued quantities $g_1(\theta)$ and $g_2(\theta)$. For a given value of α, how to construct an approximate confidence interval with level $1 - \alpha$ for $g_1(\theta) + g_2(\theta)$?

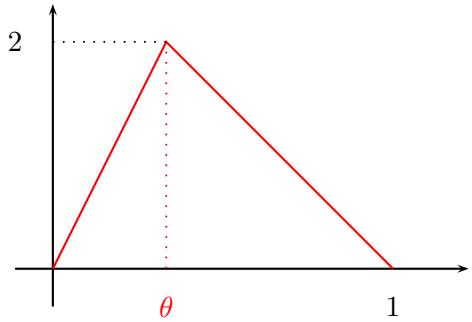
Homework

Exercise 4.A.3. In the Gaussian model, find a confidence interval with level $1-\alpha$ for σ^2, with equal risk of under- and overestimation.

Exercise 4.A.4 (Triangle law on $[0,1]$)**.** The triangle law on $[0,1]$ with parameter $\theta \in (0,1)$ is the probability measure with density

$$p(x;\theta) = 2\left(\mathbb{1}_{\{0<x\leq\theta\}}\frac{x}{\theta} + \mathbb{1}_{\{\theta<x<1\}}\frac{1-x}{1-\theta}\right),$$

which is represented below.



1. Compute $\mathbb{E}_\theta[X_1]$ and deduce a strongly consistent moment estimator $\widetilde{\theta}_n$ for θ.
2. Compute $\mathrm{Var}_\theta(X_1)$, show that $\widetilde{\theta}_n$ is asymptotically normal, and express its asymptotic variance.
3. Construct an asymptotic confidence interval for θ with level 95%.
4. Using Hoeffding's inequality, construct an approximate confidence interval with level $1-\alpha$ for θ.
5. Show that $\sup_{\theta\in(0,1)} \mathrm{Var}_\theta(X_1)$ is smaller than the bound $1/4$ provided by Lemma 4.4.2. Deduce an approximate confidence interval with level $1-\alpha$ for θ, based on the Bienaymé–Chebichev inequality, which is sharper than the interval I_n^{BC} obtained in Section 4.4.
6. Using Python, determine which of the approximate confidence intervals respectively obtained in Questions 4. and 5. is smaller, depending on the value of α.

Supplementary exercises

Exercise 4.A.5 (Variance stabilisation)**.** Let Z_n be a consistent estimator of $g(\theta)$, asymptotically normal with asymptotic variance $V(\theta)$. We assume that a consistent estimator $\widehat{V}_n$ of $V(\theta)$ is available. The level α is fixed throughout the exercise.

1. Recall the width of the asymptotic confidence interval I_n for $g(\theta)$ given by Proposition 4.3.1.
2. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a strictly monotonic and C^1 function. Show that $\Phi(Z_n)$ is a consistent and asymptotically normal estimator of $\Phi(g(\theta))$, and compute its asymptotic variance.
3. Using Proposition 4.3.1, construct an asymptotic confidence interval for $\Phi(g(\theta))$. Using the fact that Φ is strictly monotonic, deduce an asymptotic confidence interval I_n^Φ for $g(\theta)$.
4. Express the width of I_n^Φ in terms of the function φ_n defined by

$$\varphi_n(s) = \Phi^{-1}\left(\Phi(Z_n) + s\Phi'(Z_n)\right) - \Phi^{-1}\left(\Phi(Z_n) - s\Phi'(Z_n)\right),$$

and compute $\varphi_n(0)$ and $\varphi_n'(0)$. If the function φ_n is either convex or concave, which of the intervals I_n and I_n^Φ is the smallest?

5. Assume that Φ satisfies the relation $\Phi'(g(\theta)) = 1/\sqrt{V(\theta)}$. What is the expression for the corresponding confidence interval I_n^Φ? Such a choice of Φ is called *variance stabilisation*.

6. Consider the estimator $\widehat{\lambda}_n = 1/\overline{X}_n$ of λ in the Exponential model. We recall from Section 2.2.3 that this estimator is asymptotically normal, with asymptotic variance $V(\lambda) = \lambda^2$. Find a function Φ such that $\Phi'(\lambda) = 1/\sqrt{V(\lambda)}$, and compute the associated asymptotic confidence interval for λ. Does variance stabilisation reduce the width of the confidence interval?

Exercise 4.A.6. The purpose of this exercise is to prove that for all $n \geq 1$, for all $r \in (1/2, 1)$, $t_{n,r} \geq \phi_r$.

1. Show that this is equivalent to an inequality on the cumulative distribution functions of $\mathrm{t}(n)$ and $\mathcal{N}(0,1)$ on $[0, +\infty)$.

2. Show that if $x > 0$ and $T_n \sim \mathrm{t}(n)$, $\mathbb{P}(T_n \leq x) = \mathbb{E}[\Phi(x\sqrt{Y_n/n})]$, where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$ and $Y_n \sim \chi_2(n)$.

3. Conclude using Jensen's inequality.

Lecture 5

Maximum Likelihood Estimation

Contents

The method of moments introduced in Lecture 2 provides consistent estimators of $g(\theta)$, but has an arbitrary aspect in the choice of the function φ on which it is based, and if one is given two estimators computed from two different functions φ, then it is not clear how to decide *a priori* which one is better — but, *a posteriori*, you should choose the one with the smallest asymptotic variance, since it will give you sharper asymptotic confidence intervals for example. In the present Lecture, we explain how to construct an estimator based on the maximisation of a certain criterion, and study its optimality properties.

5.1 The Maximum Likelihood Estimator

5.1.1 Likelihood of a realisation

Throughout the Lecture, we assume that:

- (continuous case) either $E \subset \mathbb{R}^l$, $l \geq 1$, and for all $\theta \in \Theta$, the probability measure P_θ possesses a density $p(x;\theta)$ with respect to the Lebesgue measure on $\mathbb{R}^l$;
- (discrete case) or E is a countable space, and we define $p(x;\theta) = P_\theta(\{x\}) = \mathbb{P}_\theta(X_1 = x)$ for all $x \in E$.

We recall that, in the second case, the function $x \mapsto p(x;\theta)$ is called the *probability mass function* of P_θ.

Remark 5.1.1. *A more general and abstract setting, which includes these two cases, is to assume that there exists a σ-finite measure μ on the space E such that for any $\theta \in \Theta$, P_θ has a density $p(x;\theta)$ with respect to μ. In the first case, μ is the Lebesgue measure on $\mathbb{R}^l$; in the second case, it is the* counting measure $\sum_{x\in E} \delta_x$. *All results in this section may be formulated in this abstract setting.*

Definition 5.1.2 (Likelihood of an observation)**.** *Let $\mathbf{x}_n = (x_1, \ldots, x_n) \in E^n$ be a possible value of the sample $\mathbf{X}_n = (X_1, \ldots, X_n)$. The* likelihood[1] *of this realisation is the function $L_n(\mathbf{x}_n; \cdot) : \Theta \to [0, +\infty)$ defined by*

$$L_n(\mathbf{x}_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta).$$

[1] *Vraisemblance* en français.

In the continuous case, $L_n(\cdot;\theta)$ is the density with respect to the Lebesgue measure of the vector $(X_1,\dots,X_n)$ under $\mathbb{P}_\theta$; while in the discrete case, it is its probability mass function, that is to say $\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n) = L_n(\mathbf{x}_n;\theta)$.

5.1.2 Maximum Likelihood Estimator

The Maximum Likelihood Estimation relies on the principle of selecting the parameter which makes the observed realisation of the sample the most likely.

Definition 5.1.3 (Maximum Likelihood Estimator)**.** *Assume that, for all $\mathbf{x}_n = (x_1,\dots,x_n) \in E^n$, the function $\theta \mapsto L_n(\mathbf{x}_n;\theta)$ reaches a global maximum at $\theta = \theta_n(\mathbf{x}_n)$. The* Maximum Likelihood Estimator *(MLE) of θ is the statistic defined by*

$$\widehat{\theta}_n = \theta_n(\mathbf{X}_n).$$

Remark 5.1.4 (On the notation)**.** *In this definition, the notation θ_n refers to the* deterministic function $E^n \to \Theta$*, while $\widehat{\theta}_n$ denotes the* random variable *obtained by applying the function θ_n to the random vector $\mathbf{X}_n$.*

When the function $\theta \mapsto L_n(\mathbf{x}_n;\theta)$ is differentiable, $\theta_n(\mathbf{x}_n)$ can be computed by looking for the points at which the derivative of $L_n(\mathbf{x}_n;\theta)$ vanishes — without forgetting to check that these points actually correspond to a maximum! In this perspective, it may be more convenient to take the derivative of the *log-likelihood*

$$\ell_n(\mathbf{x}_n;\theta) = \log L_n(\mathbf{x}_n;\theta) = \sum_{i=1}^{n} \ell_1(x_i;\theta), \tag{5.1}$$

rather that $L_n(\mathbf{x}_n;\theta)$, because the product over i in the definition of the likelihood is turned into a sum. Since the logarithm is increasing, both approaches are equivalent.

Exercise 5.1.5 (The Bernoulli model)**.** Compute the MLE of p in the Bernoulli model $\{\mathcal{B}(p), p \in [0,1]\}$.

5.1.3 More examples

Example 5.1.6 (The Exponential model)**.** *The likelihood of a realisation $\mathbf{x}_n = (x_1,\dots,x_n) \in (0,+\infty)^n$ in the Exponential model $\{\mathcal{E}(\lambda), \lambda > 0\}$ writes*

$$L_n(\mathbf{x}_n;\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

The log-likelihood is

$$\ell_n(\mathbf{x}_n;\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i,$$

it is C^1 on $(0,+\infty)$ and

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\, \ell_n(\mathbf{x}_n;\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

vanishes if and only if λ is equal to

$$\lambda_n(\mathbf{x}_n) = \frac{n}{\sum_{i=1}^{n} x_i}.$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\, \ell_n(\mathbf{x}_n;\lambda) > 0 \quad \text{if } \lambda < \lambda_n(\mathbf{x}_n), \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda}\, \ell_n(\mathbf{x}_n;\lambda) < 0 \quad \text{if } \lambda > \lambda_n(\mathbf{x}_n),$$

the log-likelihood — and therefore the likelihood — actually attains its maximum at $\lambda_n(\mathbf{x}_n)$. As a consequence, the MLE of λ is

$$\widehat{\lambda}_n = \lambda_n(\mathbf{X}_n) = \frac{1}{\overline{X}_n}.$$

The MLE coincides with moment estimator obtained in Section 2.2.3. We shall see in Example 5.1.9 that it is not always the case.

Example 5.1.7 (The Gaussian model)**.** *The likelihood of a realisation* $\mathbf{x}_n = (x_1, \dots, x_n) \in \mathbb{R}^n$ *in the Gaussian model* $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$ *writes*

$$L_n(\mathbf{x}_n; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right).$$

The log-likelihood is

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

it is C^1 *on* $\mathbb{R} \times (0, +\infty)$ *and satisfies*

$$\frac{\partial}{\partial\mu}\,\ell_n(\mathbf{x}_n; \mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu), \qquad \frac{\partial}{\partial\sigma^2}\,\ell_n(\mathbf{x}_n; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i-\mu)^2.$$

Both derivatives vanish if and only

$$\mu = \mu_n(\mathbf{x}_n) = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \sigma^2 = \sigma_n^2(\mathbf{x}_n) = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_n(\mathbf{x}_n))^2,$$

and the fact that the log-likelihood actually attains its maximum at the point $(\mu_n(\mathbf{x}_n), \sigma_n^2(\mathbf{x}_n))$ *can be checked by studying the sign of the Hessian matrix of* $\ell_n(\mathbf{x}_n; \mu, \sigma^2)$ *at this point, which is left as an exercise to the reader. Alternatively, one may observe that* $\ell_n(\mathbf{x}_n; \mu, \sigma^2) \to -\infty$ *when* $|\mu| \to +\infty$ *and* $\sigma^2 \to +\infty$ *or* $\sigma^2 \to 0$*, so that the only critical point* $(\mu_n(\mathbf{x}_n), \sigma_n^2(\mathbf{x}_n))$ *is necessarily a global maximum.*

As a conclusion, the MLE of (μ, σ^2) *is given by*

$$\widehat{\mu}_n = \mu_n(\mathbf{X}_n) = \overline{X}_n, \qquad \widehat{\sigma}_n^2 = \sigma_n^2(\mathbf{X}_n) = V_n,$$

with the notation of p. 20.

Remark 5.1.8. *By Proposition 3.2.2, the estimators* $\widehat{\mu}_n$ *and* $\widehat{\sigma}_n^2$ *are independent.*

Example 5.1.9 (The Uniform model)**.** *We consider the set of uniform distributions* $\{\mathcal{U}([0, \theta]), \theta > 0\}$*. For a given* $\theta > 0$*, the random variables* $X_1, \dots, X_n$ *take their values in* $[0, \theta]$*,* $\mathbb{P}_\theta$*-almost surely; but since* θ *can a priori take any positive value, we have to work with the state space* $E = [0, +\infty)$*. For any* $\mathbf{x}_n = (x_1, \dots, x_n) \in [0, +\infty)^n$*, the likelihood of the realisation* $\mathbf{x}_n$ *writes*

$$L_n(\mathbf{x}_n; \theta) = \prod_{i=1}^n \frac{1}{\theta}\mathbb{1}_{\{x_i \le \theta\}} = \theta^{-n}\mathbb{1}_{\{\max_{1\le i\le n} x_i \le \theta\}}.$$

As a function of $\theta > 0$*, the likelihood is not differentiable at the point* $\max_{1\le i\le n} x_i$*, therefore the method of the previous examples cannot be applied. On the other hand, the study of the variations of the function* $\theta \mapsto L_n(\mathbf{x}_n; \theta)$ *is straightforward (see Figure 5.1) and shows that the latter attains its maximum for* $\theta = \theta_n(\mathbf{x}_n) = \max_{1\le i\le n} x_i$*. As a consequence, the MLE is given by*

$$\widehat{\theta}_n = \max_{1\le i\le n} X_i.$$

The asymptotic properties of $\widehat{\theta}_n$ *were studied in Exercise 2.3.3, where we observed in particular that it is strongly consistent but not asymptotically normal.*

In the Uniform model of Example 5.1.9, the support of the law of X_1 depends on the parameter θ, which makes the likelihood not differentiable with respect to θ. Trying to differentiate the likelihood, or the log-likelihood, with respect to θ, then leads to wrong results. Models for which the support depends on the parameter generally belong to the class of *nonregular models*, as introduced in Section 5.2.

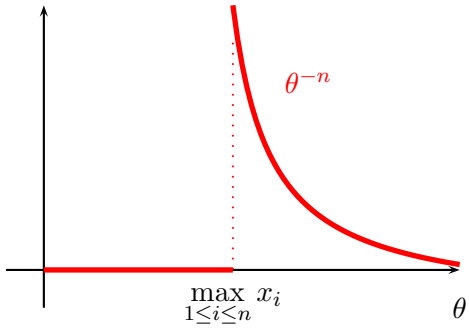


Figure 5.1: The likelihood of the Uniform model.

5.1.4 Contrast and M-estimators

In this Subsection, we provide a theoretical justification to the use of the MLE.

Definition 5.1.10 (Contrast function)**.** *A* contrast *is a function $\psi : E \times \Theta \to \mathbb{R}$ such that, for any $\theta_* \in \Theta$, the function $\theta \mapsto \mathbb{E}_{\theta_*}[\psi(X_1;\theta)]$ is maximal at $\theta = \theta_*$.*

Given a contrast function, to estimate the parameter θ_* under which the sample is distributed, it suffices to maximise $\theta \mapsto \mathbb{E}_{\theta_*}[\psi(X_1;\theta)]$. However, in general this quantity is unknown, but by the Law of Large Numbers, the approximation

$$\mathbb{E}_{\theta_*}[\psi(X_1;\theta)] \simeq \frac{1}{n}\sum_{i=1}^{n}\psi(X_i;\theta)$$

is unbiased and strongly consistent under $\mathbb{P}_{\theta_*}$. Therefore, it is natural to estimate θ_* by maximising the right-hand side in θ.

Definition 5.1.11 (M-estimator)**.** *Assume that, for any $\mathbf{x}_n \in E^n$, the function $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}\psi(x_i;\theta)$ reaches a global maximum at $\theta = \theta_n^{\psi}(\mathbf{x}_n)$. The* M-estimator *of θ_* associated with the contrast function ψ is the statistic $\widehat{\theta}_n^{\psi} = \theta_n^{\psi}(\mathbf{X}_n)$.*

Since the MLE maximises the right-hand side of (5.1), to show that it is an M-estimator it suffices to check that the function ℓ_1 is a contrast. This is the purpose of the next statement.

Proposition 5.1.12 (MLE as an M-estimator)**.** *The function ℓ_1 is a contrast.*

Proof. We assume for simplicity that the space E is discrete but the proof is the same in the continuous case. Then, for any θ, θ_*, the convexity inequality $\log(u) \leq u - 1$ yields

$$\begin{aligned}\mathbb{E}_{\theta_*}\left[\ell_1(X_1;\theta)\right] - \mathbb{E}_{\theta_*}\left[\ell_1(X_1;\theta_*)\right] &= \sum_{x\in E}\log\left(\frac{p(x;\theta)}{p(x;\theta_*)}\right)p(x;\theta_*)\\ &\leq \sum_{x\in E}\left(\frac{p(x;\theta)}{p(x;\theta_*)} - 1\right)p(x;\theta_*) = 0,\end{aligned}$$

which shows that $\mathbb{E}_{\theta_*}[\ell_1(X_1;\theta)] \leq \mathbb{E}_{\theta_*}[\ell_1(X_1;\theta_*)]$ and completes the proof. □

Remark 5.1.13. *The quantity $-\mathbb{E}_{\theta_*}[\ell_1(X_1;\theta)]$ is called the* cross-entropy *between the distributions P_{θ_*} and P_θ. In statistical learning, it is often taken as a loss function to be minimised, which is therefore just equivalent to likelihood maximisation.*

Exercise 5.1.14 (Moment estimators as M-estimators)**.** Let Z_n be an estimator of $g(\theta_*)$ obtained by the methods of moments with $m(z) = z$; namely,

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i), \qquad \mathbb{E}_{\theta_*}\left[\varphi(X_1)\right] = g(\theta_*).$$

Check that Z_n is an M-estimator for the contrast function $\psi(x;\theta) = -(\varphi(x) - m(g(\theta)))^2$.

Expressing both MLE and moment estimators as M-estimators allows to develop an abstract asymptotic theory for them, which provides in particular consistency, asymptotic normality and a general formula for the asymptotic covariance, under regularity assumptions on the underlying model. We shall present a simplified version of this approach in Section 5.2 for the MLE, and refer to [7] for an extensive treatment.

5.2 Optimality of the MLE

This Section aims at justifying that, under some conditions, the MLE is an *optimal* estimator in a certain sense. To proceed, we introduce in Subsection 5.2.1 the notion of *Fisher information*, which is used in Subsection 5.2.2 to define the notion of *(asymptotic) efficient* estimator. We finally argue in Subsection 5.2.3 that in *regular* models, the MLE is indeed asymptotically efficient.

5.2.1 Regular models and Fisher information

We recall that $L_1(x_1;\theta) = p(x_1;\theta)$ (respectively, $\ell_1(x_1;\theta) = \log p(x_1;\theta)$) denotes the likelihood (respectively, the log-likelihood) of a single observation x_1. We denote by $(\theta_1, \dots, \theta_q)$ the coordinates of the parameter $\theta \in \Theta \subset \mathbb{R}^q$, and for any smooth function $\varphi : \Theta \to \mathbb{R}$, we shall write

$$\nabla_\theta \varphi(\theta) = \left(\frac{\partial \varphi}{\partial \theta_1}(\theta), \dots, \frac{\partial \varphi}{\partial \theta_q}(\theta)\right) \in \mathbb{R}^q.$$

Definition 5.2.1 (Regular model)**.** *A parametric model is* regular *if:*

(i) Θ *is an open subset of* $\mathbb{R}^q$*,*

(ii) for all $x_1 \in E$*, for all* $\theta \in \Theta$*,* $L_1(x_1;\theta) > 0$*,*

(iii) for all $x_1 \in E$*, the function* $\theta \mapsto \ell_1(x_1;\theta)$ *is* C^1 *on* Θ*,*

(iv) for all $\theta \in \Theta$*,* $\mathbb{E}_\theta[\|\nabla_\theta \ell_1(X_1;\theta)\|^2] < +\infty$*.*

Example 5.2.2 (Regular models)**.** *The Bernoulli model* $\{\mathcal{B}(p), p \in (0,1)\}$*, the Exponential model* $\{\mathcal{E}(\lambda), \lambda > 0\}$ *and the Gaussian model* $\{\mathcal{N}(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ *are regular, but the Uniform model* $\{\mathcal{U}[0,\theta], \theta > 0\}$ *is not (why?).*

Definition 5.2.3 (Score and Fisher information)**.** *The* score *of a regular model is the random vector*

$$\nabla_\theta \ell_1(X_1;\theta) \in \mathbb{R}^q,$$

with coordinates

$$\frac{\partial}{\partial \theta_i}\ell_1(X_1;\theta) = \frac{1}{p(X_1;\theta)}\frac{\partial}{\partial \theta_i}p(X_1;\theta), \qquad i \in \{1,\dots,q\}.$$

The Fisher information $I(\theta)$ *of a regular model is the covariance of the score, namely*

$$I(\theta) := \mathrm{Cov}_\theta\left[\nabla_\theta \ell_1(X_1;\theta)\right] \in \mathbb{R}^{q\times q}.$$

Remark 5.2.4. *In the sequel, we shall repeatedly compute derivatives, with respect to θ, of quantities which write as an expectation or a sum with respect to x. Justifying that the derivative of the integral is the integral of the derivative would in principle require to make a priori assumptions on the (local) domination of the derivative of the integrand with respect to θ in order to apply Lebesgue's Theorem. For the sake of clarity, we systematically omit to formulate such assumptions. The reader interested in the rigorous validity of our statements for a given model $\mathcal{P}$ should have no difficulty to reproduce our computation and check whether they are legitimate.*

Proposition 5.2.5 (Expression of $I(\theta)$)**.** *We have, for any $\theta \in \Theta$,*

$$\mathbb{E}_\theta \left[\nabla_\theta \ell_1(X_1;\theta) \right] = 0,$$

and the coefficients of $I(\theta)$ admit the following two expressions:

$$\begin{aligned} I_{ij}(\theta) &= \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \ell_1(X_1;\theta) \frac{\partial}{\partial \theta_j} \ell_1(X_1;\theta) \right] \\ &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell_1(X_1;\theta) \right]. \end{aligned}$$

Proof. Assume for example that P_θ possesses a density $p(x;\theta)$ on $\mathbb{R}^l$. The computation

$$0 = \frac{\partial}{\partial \theta_i} \underbrace{\int_{x \in \mathbb{R}^l} p(x;\theta) \mathrm{d}x}_{=1} = \int_{x \in \mathbb{R}^l} \frac{\partial}{\partial \theta_i} p(x;\theta) \mathrm{d}x = \int_{x \in \mathbb{R}^l} \left(\frac{\partial}{\partial \theta_i} \log p(x;\theta) \right) p(x;\theta) \mathrm{d}x \tag{5.2}$$

yields the first identity. As a consequence,

$$\mathrm{Cov}_\theta \left[\nabla_\theta \ell_1(X_1;\theta) \right] = \mathbb{E}_\theta \left[\nabla_\theta \ell_1(X_1;\theta)^\top \nabla_\theta \ell_1(X_1;\theta) \right],$$

which yields the first expression for $I(\theta)$. To get the second expression, we take the derivative with respect to θ_j of (5.2) to get

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_j} \int_{x \in \mathbb{R}^l} \left(\frac{\partial}{\partial \theta_i} \log p(x;\theta) \right) p(x;\theta) \mathrm{d}x \\ &= \int_{x \in \mathbb{R}^l} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x;\theta) \right) p(x;\theta) \mathrm{d}x + \int_{x \in \mathbb{R}^l} \left(\frac{\partial}{\partial \theta_i} \log p(x;\theta) \right) \frac{\partial}{\partial \theta_j} p(x;\theta) \mathrm{d}x, \end{aligned}$$

and complete the proof by observing that the second term in the right-hand side is

$$\int_{x \in \mathbb{R}^l} \left(\frac{\partial}{\partial \theta_i} \log p(x;\theta) \right) \frac{\partial}{\partial \theta_j} p(x;\theta) \mathrm{d}x = \int_{x \in \mathbb{R}^l} \left(\frac{\partial}{\partial \theta_i} \ell_1(x;\theta) \right) \left(\frac{\partial}{\partial \theta_j} \ell_1(x;\theta) \right) p(x;\theta) \mathrm{d}x. \qquad \square$$

Example 5.2.6. *In the Exponential model, we have*

$$\forall x_1 > 0, \qquad \ell_1(x_1;\lambda) = \log \lambda - \lambda x_1,$$

so that the score writes $\lambda^{-1} - X_1$. As a consequence, $I(\lambda) = 1/\lambda^2$.

Exercise 5.2.7. Compute the Fisher information $I(\mu, \sigma^2)$ of the Gaussian model.

5.2.2 Efficient estimator

In this paragraph, we use the Fisher information to define a notion of *efficiency* of unbiased estimators. We recall from Remark 2.1.12 that for symmetric matrices $A, B \in \mathbb{R}^{q \times q}$, we write $A \succeq B$ is $A - B$ is nonnegative, that is to say if $\langle u, Au \rangle \geq \langle u, Bu \rangle$ for any $u \in \mathbb{R}^q$.

Theorem 5.2.8 (Cramér–Rao[2] bound)**.** *For a regular model such that $I(\theta)$ is positive for all $\theta \in \Theta$, let $\widetilde{\theta}_n = t_n(\mathbf{X}_n)$ be an unbiased estimator of θ, with covariance matrix $K_n(\theta) = \mathrm{Cov}_\theta[\widetilde{\theta}_n]$. For all $\theta \in \Theta$,*

$$K_n(\theta) \succeq \frac{I^{-1}(\theta)}{n}.$$

Proof. We assume that we are in the case where E is a subset of $\mathbb{R}^l$ and for all $\theta \in \Theta$, the probability measure P_θ possesses a density with respect to the Lebesgue measure on $\mathbb{R}^l$. The adaptation of the proof to the case of discrete random variables is straightforward.

For any $u \in \mathbb{R}^q$, the unbiasedness of $\widetilde{\theta}_n$ implies that

$$\langle u, \theta\rangle = \langle u, \mathbb{E}_\theta[\widetilde{\theta}_n]\rangle = \mathbb{E}_\theta[\langle u, \widetilde{\theta}_n\rangle] = \int_{\mathbf{x}_n \in E^n} \langle u, t_n(\mathbf{x}_n)\rangle \prod_{i=1}^n p(x_i;\theta)\mathrm{d}\mathbf{x}_n.$$

Taking the gradient, with respect to θ, of this identity yields

$$u = \int_{\mathbf{x}_n \in E^n} \langle u, t_n(\mathbf{x}_n)\rangle \sum_{i=1}^n \frac{\nabla_\theta p(x_i;\theta)}{p(x_i;\theta)} \prod_{i=1}^n p(x_i;\theta)\mathrm{d}\mathbf{x}_n = \mathbb{E}_\theta\left[\langle u, \widetilde{\theta}_n\rangle \sum_{i=1}^n \nabla_\theta \ell_1(X_i;\theta)\right],$$

and therefore, for any $v \in \mathbb{R}^d$, we have

$$\langle u, v\rangle = \mathbb{E}_\theta\left[\langle u, \widetilde{\theta}_n\rangle \sum_{i=1}^n \langle \nabla_\theta \ell_1(X_i;\theta), v\rangle\right] = \mathrm{Cov}_\theta\left(\langle u, \widetilde{\theta}_n\rangle, \sum_{i=1}^n \langle \nabla_\theta \ell_1(X_i;\theta), v\rangle\right),$$

since, by Proposition 5.2.5, the second term in the covariance has expectation 0. Therefore, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\langle u, v\rangle^2 &\le \mathrm{Var}_\theta(\langle u, \widetilde{\theta}_n\rangle)\,\mathrm{Var}_\theta\left(\sum_{i=1}^n \langle \nabla_\theta \ell_1(X_i;\theta), v\rangle\right) \\ &= \langle u, K_n u\rangle \cdot n\,\mathrm{Var}_\theta\left(\langle \nabla_\theta \ell_1(X_1;\theta), v\rangle\right) = n\langle u, K_n u\rangle\langle v, I(\theta) v\rangle,\end{aligned}$$

using the definition of $I(\theta)$. We conclude by taking $v = I^{-1}(\theta)u$. □

Definition 5.2.9 (Efficient estimator)**.** *An estimator $\widetilde{\theta}_n$ of θ is called* efficient[3] *if it is unbiased and its covariance $K_n(\theta)$ satisfies*

$$K_n(\theta) = \frac{I^{-1}(\theta)}{n}.$$

Recall that if $\widetilde{\theta}_n$ is an unbiased estimator of θ, then its MSE writes

$$\mathrm{MSE}(\widetilde{\theta}_n;\theta) = \mathrm{Var}_\theta(\widetilde{\theta}_n) = \mathrm{tr}\,\mathrm{Cov}_\theta[\widetilde{\theta}_n],$$

and therefore Theorem 5.2.8 provides the lower bound

$$\mathrm{MSE}(\widetilde{\theta}_n;\theta) \ge \frac{1}{n}\,\mathrm{tr}\,I^{-1}(\theta).$$

As a consequence,

(i) in regular models, the MLE of unbiased estimators is always of order at least $1/n$;

(ii) efficient estimators have the smallest possible MSE among the class of unbiased estimators.

Exercise 5.2.10. Check that in the Bernoulli model, the estimator $\overline{X}_n$ of p is efficient.

[2] *Fréchet–Darmois* en français.

[3] *Efficace* en français.

Although it is necessary for the Cramér–Rao bound to hold, the condition that $\widetilde{\theta}_n$ be unbiased makes the definition of efficiency a bit restrictive. We may relax this assumption by introducing the notion of *asymptotic* efficiency.

Definition 5.2.11 (Asymptotically efficient estimator)**.** *A consistent estimator $\widetilde{\theta}_n$ of θ which is asymptotically normal, with asymptotic covariance matrix $K(\theta)$, is called* asymptotically efficient *if $K(\theta) = I^{-1}(\theta)$.*

Remark 5.2.12 (Cramér–Rao bound for the estimation of $g(\theta)$)**.** *Let $g : \Theta \to \mathbb{R}^d$ be a C^1 function. If a consistent estimator $\widetilde{\theta}_n$ of θ is asymptotically efficient, then by the Delta method, the plug-in estimator $g(\widetilde{\theta}_n)$ of $g(\theta)$ is asymptotically normal, with asymptotic covariance matrix $\nabla g(\theta)^\top I^{-1}(\theta)\nabla g(\theta)$, where $\nabla g(\theta)$ is the $q \times d$ matrix with columns $\nabla g_1(\theta), \dots, \nabla g_d(\theta)$. It turns out that this bound remains optimal, in the sense that the proof of Theorem 5.2.8 can be extended to show that any unbiased estimator Z_n of $g(\theta)$ satisfies the inequality*

$$\mathrm{Cov}_\theta[Z_n] \succeq \frac{1}{n}\nabla g(\theta)^\top I^{-1}(\theta)\nabla g(\theta).$$

5.2.3 Asymptotic efficiency of the MLE

In general, the MLE is biased (think of the Exponential model, or the estimator of σ^2 in the Gaussian model), and therefore it cannot be expected to be efficient. However, in regular models, the following statement is generally true:

The MLE $\widehat{\theta}_n$ is asymptotically efficient.

This fact justifies the interest of Maximum Likelihood Estimation in regular models. It is most easily observed case-by-case by:

(i) showing that $\widehat{\theta}_n$ is asymptotically normal and computing its asymptotic covariance $K(\theta)$ (often using the Central Limit Theorem and the Delta Method) on the one hand;

(ii) computing the Fisher information $I(\theta)$ of the model on the other hand;

(iii) finally checking that $K(\theta) = I^{-1}(\theta)$.

Exercise 5.2.13. Apply this procedure to the Gaussian model.

Assuming that $\widehat{\theta}_n$ is consistent, the fact that, in general, it is asymptotically normal with asymptotic covariance matrix $I^{-1}(\theta)$, can be derived thanks to the following heuristic computation[4]. For the sake of simplicity, we assume that $q = 1$ so that $\Theta \subset \mathbb{R}$.

By the definition of $\widehat{\theta}_n$ and the regularity assumption on $\ell_1(x_1; \cdot)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n; \widehat{\theta}_n) = 0.$$

Assuming that $\widehat{\theta}_n$ is close to θ, we perform the Taylor approximation

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n; \widehat{\theta}_n) \simeq \frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n; \theta) + \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\mathbf{X}_n; \theta)(\widehat{\theta}_n - \theta),$$

and thus deduce that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \simeq -\sqrt{n}\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n; \theta)}{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\mathbf{X}_n; \theta)} = -\frac{\frac{1}{\sqrt{n}}\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n; \theta)}{\frac{1}{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\mathbf{X}_n; \theta)}.$$

[4] We refer to [7, Theorem 5.39, p. 65] for a complete proof.

On the one hand, Proposition 5.2.5 shows that

$$\frac{1}{\sqrt{n}}\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{X}_n;\theta) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_1(X_i;\theta) - \mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_1(X_1;\theta)\right]\right),$$

which by the Central Limit Theorem converges in distribution, under $\mathbb{P}_\theta$, to $\mathcal{N}(0, I(\theta))$. On the other hand, by the Law of Large Numbers,

$$\frac{1}{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\mathbf{X}_n;\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_1(X_i;\theta)$$

converges $\mathbb{P}_\theta$-almost surely to

$$\mathbb{E}_\theta\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_1(X_1;\theta)\right] = -I(\theta).$$

By Slutsky's Theorem, we conclude that $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in distribution to $\mathcal{N}(0, 1/I(\theta))$.

Remark 5.2.14 (Misspecified case)**.** *So far, we have always worked under the assumption that the observed sample is actually drawn under some measure P_{θ_*} which belongs to our parametric model $\mathcal{P}$, in which case we have shown that the MLE generally converges to the true parameter θ_*. One may now wonder what happens if one observes $X_1, \dots, X_n$ iid according to some probability measure P which does* not *belong to $\mathcal{P}$. Such a case is called* misspecified*. In this setting, one may still define the MLE $\widehat{\theta}_n := \theta_n(\mathbf{X}_n)$ and then ask whether it converges, and if so, toward what. Coming back to the interpretation in Subsection 5.1.4 of $\widehat{\theta}_n$ as an empirical approximation of the maximiser of the function $\theta \mapsto \mathbb{E}[\ell_1(X_1;\theta)]$, one may show that under similar regularity assumptions as above, the MLE converges to*

$$\theta_* := \operatorname*{arg\,max}_{\theta\in\Theta} \mathbb{E}[\ell_1(X_1;\theta)],$$

where in the right-hand side, $X_1 \sim P$. The probability measure P_{θ_*} can therefore be understood as the projection, for the cross-entropy distance, of P onto the model $\mathcal{P}$, see [7, Remark 5.25] for details.

5.3 Advanced examples

5.3.1 Logistic and linear regression

The definition of the likelihood of a realisation is given in Definition 5.1.2 in the case of an iid sample, but its interpretation as the density, or the probability mass function, of the sample $(X_1, \dots, X_n)$ under $\mathbb{P}_\theta$ allows to generalise this definition to samples which are neither identically distributed, nor independent.

Consider for example the setting of logistic regression, whose principle is recalled in Section A.6, where you observe pairs $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \{0, 1\}$: these pairs are assumed to be such that, for all $i \in \{1, \dots, n\}$, y_i is the realisation of a Bernoulli random variable Y_i with parameter $p_\beta(x_i) = \Psi(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p)$, with $\Psi(u) = 1/(1+\mathrm{e}^{-u})$, and these realisations are independent from each other. The joint probability mass function of $(Y_1, \dots, Y_n)$ therefore writes

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^{n} p_\beta(x_i)^{y_i}(1 - p_\beta(x_i))^{1-y_i},$$

which thus defines the likelihood $L_n(\mathbf{x}_n, \mathbf{y}_n; \beta)$ of the realisation $(\mathbf{x}_n, \mathbf{y}_n)$ as a function of β. The MLE of β is then obtained by maximising this function. In general, this can only be done with numerical solvers.

Consider now the linear regression model

$$\mathbf{y}_n = \mathbf{x}_n\beta + \boldsymbol{\epsilon}_n,$$

where $\mathbf{x}_n \in \mathbb{R}^{n\times(p+1)}$, $\mathbf{y}_n \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$ and

$$\boldsymbol{\epsilon}_n = \mathcal{N}_n(0, \sigma^2 I_n)$$

as in Section 3.3. Then $\mathbf{y}_n$ can be seen as a realisation of a random vector $\mathbf{Y}_n \sim \mathcal{N}_n(\mathbf{x}_n\beta, \sigma^2 I_n)$, which has density

$$L_n(\mathbf{x}_n, \mathbf{y}_n; \beta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2\right).$$

The MLE of β is therefore obtained by minimising the quantity $\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2$, which leads to the conclusion that

in the linear Gaussian model with $\epsilon_1, \dots, \epsilon_n$ iid, the OLS coincides with the MLE.

The study of the MLE of β can in fact be carried out in a more general setting, which allows for the variables $\epsilon_1, \dots, \epsilon_n$ to be neither independent nor identically distributed. Assume indeed that the vector $\boldsymbol{\epsilon}_n$ has a Gaussian distribution

$$\boldsymbol{\epsilon}_n = \mathcal{N}_n(0, V_n)$$

for some covariance matrix $V_n \in \mathbb{R}^{n\times n}$ which we assume to be positive. Then $\mathbf{y}_n$ is a now realisation of a random vector $\mathbf{Y}_n \sim \mathcal{N}_n(\mathbf{x}_n\beta, V_n)$, whose density writes

$$L_n(\mathbf{x}_n, \mathbf{y}_n; \beta, V_n) = \frac{1}{\sqrt{(2\pi)^n \det V_n}} \exp\left(-\frac{1}{2}\langle \mathbf{y}_n - \mathbf{x}_n\beta, V_n^{-1}(\mathbf{y}_n - \mathbf{x}_n\beta)\rangle\right).$$

The MLE of β is therefore obtained by minimising the quantity $\langle \mathbf{y}_n - \mathbf{x}_n\beta, V_n^{-1}(\mathbf{y}_n - \mathbf{x}_n\beta)\rangle$, which is called Generalised Least Square estimation[5].

Proposition 5.3.1 (Generalized Least Square)**.** *Under the assumptions of Proposition A.4.1, the unique minimiser of the Generalised Least Square problem*

$$\min_{\beta\in\mathbb{R}^{p+1}} \left\langle \mathbf{y}_n - \mathbf{x}_n\beta, V_n^{-1}\left(\mathbf{y}_n - \mathbf{x}_n\beta\right)\right\rangle$$

is given by

$$\widehat{\beta}_{V_n} := \left(\mathbf{x}_n^\top V_n^{-1}\mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top V_n^{-1}\mathbf{y}_n,$$

which is called the GLS.

Proof. Since V_n is symmetric and positive, by the Spectral Theorem one may find an invertible matrix $A_n \in \mathbb{R}^{n\times n}$ such that $A_n^\top A_n = V_n^{-1}$. Introducing $\mathbf{x}'_n := A_n\mathbf{x}_n$ and $\mathbf{y}'_n := A_n\mathbf{y}_n$ allows us to turn the Generalised Least Square problem with data $\mathbf{x}_n, \mathbf{y}_n$ into a standard Least Square problem with data $\mathbf{x}'_n, \mathbf{y}'_n$. Since A_n is invertible, the matrices $\mathbf{x}_n$ and $\mathbf{x}'_n$ have the same rank, therefore Proposition A.4.1 applies and yields

$$\widehat{\beta}_{V_n} = \left(\mathbf{x}_n'^\top\mathbf{x}'_n\right)^{-1}\mathbf{x}_n'^\top\mathbf{y}'_n = \left(\mathbf{x}_n^\top V_n^{-1}\mathbf{x}_n\right)^{-1}\mathbf{x}_n^\top V_n^{-1}\mathbf{y}_n. \qquad \square$$

Exercise 5.3.2 (Generalisation of the Gauss–Markov Theorem)**.** Assuming only that $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$, $\mathrm{Cov}[\boldsymbol{\epsilon}_n] = V_n$ but not necessarily that $\boldsymbol{\epsilon}_n$ is Gaussian, show that $\widehat{\beta}_{V_n}$ identified in Proposition 5.3.1 as the solution to the Generalised Least Square problem is the BLUE of β (recall Theorem A.5.4).

[5] A. C. Aitken. On Least Squares and Linear Combinations of Observations. *Proceedings of the Royal Society of Edinburgh*, 1935.

5.3.2 Mixture models and the EM algorithm

Weldon's crabs

In 1893, English evolutionary biologist Raphael Weldon measured the ratio between the forehead width and the body length of $n = 1000$ crabs[6]. The empirical frequencies of these ratios as well as smoothed densities are displayed on Figure 5.2, for the whole sample as well as for the separate subsamples of male and female crabs. These distributions look Gaussian for males and females, but with different parameters; in contrast, the overall distribution is apparently skewed[7] so it is harder to recognise a Gaussian distribution.

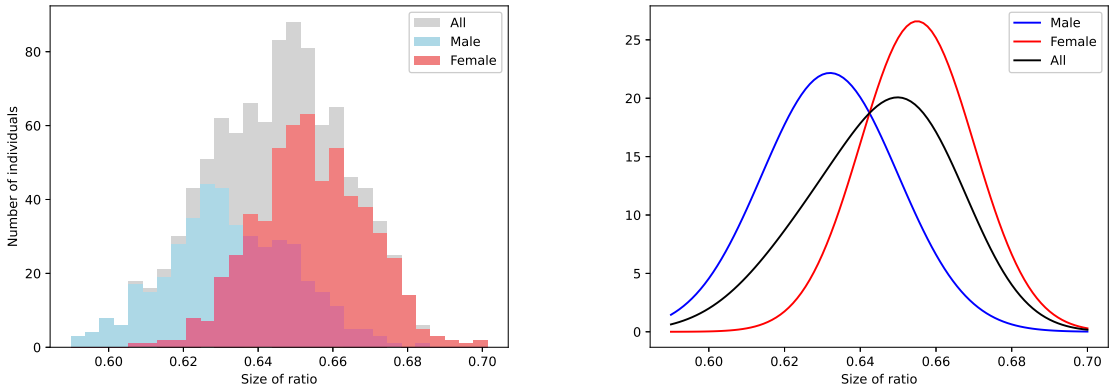


Figure 5.2: Histogram and smoothed density estimation of the distribution of ratios in Weldon's data.

A simple statistical model to describe this sample is therefore the following:

- the ratios for male crabs are Gaussian with parameters $(\mu_{\mathrm{m}}, \sigma^2_{\mathrm{m}})$;
- the ratios for female crabs are Gaussian with parameters $(\mu_{\mathrm{f}}, \sigma^2_{\mathrm{f}})$;
- there is a proportion $\pi(\mathrm{m})$ of male crabs, and $\pi(\mathrm{f}) = 1 - \pi(\mathrm{m})$ of female crabs.

Denoting by $p_{\mu,\sigma^2}(x)$ the density of the Gaussian distribution with parameters (μ, σ^2), it is then an easy computation to show that if one catches a crab at random, the density of its ratio writes

$$p(x) = \pi(\mathrm{m}) p_{\mu_{\mathrm{m}},\sigma^2_{\mathrm{m}}}(x) + \pi(\mathrm{f}) p_{\mu_{\mathrm{f}},\sigma^2_{\mathrm{f}}}(x).$$

We now consider the estimation of the parameters $(\mu_{\mathrm{m}}, \sigma^2_{\mathrm{m}})$, $(\mu_{\mathrm{f}}, \sigma^2_{\mathrm{f}})$, and $\pi(\mathrm{m}), \pi(\mathrm{f})$. Let $X_i \in \mathbb{R}$ denote the ratio for the i-th crab, and $Y_i \in \{\mathrm{m}, \mathrm{f}\}$ denote its sex. If both variables X_i and Y_i are observed, then, for any $y \in \{\mathrm{m}, \mathrm{f}\}$, setting

$$n_y := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}},$$

it is clear that the natural estimators

$$\widehat{\pi}_n(y) := \frac{n_y}{n}, \qquad \widehat{\mu}_{y,n} := \frac{1}{n_y} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}} X_i, \qquad \widehat{\sigma}^2_{y,n} := \frac{1}{n_y} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}} (X_i - \widehat{\mu}_{y,n})^2$$

are strongly consistent. But there are situations where the variable Y_i remains *unobserved* – for example, Weldon might have released the crabs before realising that the ratios could differ between males and

[6] The data are reported in Section 5.6 of Delmas, Jourdain, *Modèles aléatoires*, 2006.
[7] *asymétrique* en français.

females. In this situation, only the sample $(X_1, \dots, X_n)$ is available, so that one can only plot, on Figure 5.2, the histogram and the smoothed density for the whole population. However, this should not scare you: the density $p(x)$ of X_i is given above, it is parametrised by $\{(\mu_{\mathrm{m}}, \sigma^2_{\mathrm{m}}), (\mu_{\mathrm{f}}, \sigma^2_{\mathrm{f}}), (\pi(\mathrm{m}), \pi(\mathrm{f}))\}$. So you may write the likelihood of the realisation $\mathbf{x}_n$ and look for maximisers. Let us directly write the log-likelihood

$$\ell_n(\mathbf{x}_n; \{(\mu_{\mathrm{m}}, \sigma^2_{\mathrm{m}}), (\mu_{\mathrm{f}}, \sigma^2_{\mathrm{f}}), (\pi(\mathrm{m}), \pi(\mathrm{f}))\}) = \sum_{i=1}^{n} \log\left(\pi(\mathrm{m}) p_{\mu_{\mathrm{m}}, \sigma^2_{\mathrm{m}}}(x_i) + \pi(\mathrm{f}) p_{\mu_{\mathrm{f}}, \sigma^2_{\mathrm{f}}}(x_i)\right).$$

This quantity may be differentiated with respect to all parameters, but the analytical resolution of the resulting system is intractable[8]. Therefore, this is a typical situation where a numerical algorithm is required to compute the MLE. The *Expectation-Maximisation algorithm* does this. We present this algorithm in a general setting, therefore we first properly introduce the notion of *mixture model*.

Mixture models

Definition 5.3.3 (Mixture model)**.** *Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be a parametric model on some space E. Let F be a finite set*[9]*, let $\pi = (\pi(y))_{y \in F}$ be a probability measure on F, and let $\boldsymbol{\theta} = (\theta_y)_{y \in F} \in \Theta^F$ be a collection of parameters in Θ. The probability measure*

$$P_{\boldsymbol{\theta}, \pi} = \sum_{y \in F} \pi(y) P_{\theta_y}$$

is called a mixture *of elements of $\mathcal{P}$, with* weights *π. This defines a new parametric family on E, namely*

$$\mathcal{P}_F = \{P_{\boldsymbol{\theta}, \pi}, \boldsymbol{\theta} \in \Theta^F, \pi \text{ probability on } F\}.$$

The probability measure $P_{\boldsymbol{\theta}, \pi}$ has a very simple interpretation. First, draw a random variable $Y \in F$, with distribution π. Next, given the value of Y, draw a random variable $X \in E$ with distribution P_{θ_Y}. Then X has distribution $P_{\boldsymbol{\theta}, \pi}$.

Remark 5.3.4 (Classification)**.** *Assuming that the parameters $\boldsymbol{\theta}$ and π are known, and that only the variable X is observed, it is a natural question to try to* predict *the value of the variable Y: for example, in Weldon's case, can the sex of the crab be determined from the mere measure of the ratio? From the graphs of Figure 5.2, it is clear that if a crab has a ratio X equal to 0.60, it is more likely to be a male, while a ratio of 0.70 rather indicates that the crab is a female. These statements can be made quantitative thanks to Bayes' formula.*

Let us first assume that the space E, in which X takes its values, is discrete. Then recall that the pmf of P_θ is denoted by $(p(x;\theta))_{x \in E}$. Bayes' formula states that, for any $x \in E$ and $y \in F$,

$$\mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y | X = x) = \frac{\mathbb{P}_{\boldsymbol{\theta}, \pi}(X = x | Y = y) \mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y)}{\sum_{y' \in F} \mathbb{P}_{\boldsymbol{\theta}, \pi}(X = x | Y = y') \mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y')} = \frac{\pi(y) p(x; \theta_y)}{\sum_{y' \in F} \pi(y') p(x; \theta_{y'})}.$$

Now, if E is continuous and P_θ has density $p(x;\theta)$, the conditional probability $\mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y | X = x)$ does not a priori make sense, since the event $X = x$ has probability 0. It is however still possible to give a rigorous meaning to the expression above[10]*, and we therefore take the identity*

$$\mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y | X = x) = \frac{\pi(y) p(x; \theta_y)}{\sum_{y' \in F} \pi(y') p(x; \theta_{y'})}$$

[8] Alternatively, Pearson constructed explicit moment estimators by computing the first five moments of X_1, which required to compute the roots of a polynomial of degree 9.

[9] We take F to be finite for simplicity but the definition could be extended to arbitrary (measurable) sets.

[10] You can for instance check that if $p(\cdot;\theta)$ is continuous and positive at x, then the right-hand side is the limit, when $\epsilon \to 0$, of the well-defined conditional probability $\mathbb{P}_{\boldsymbol{\theta}, \pi}(Y = y | |X - x| \leq \epsilon)$.

as the definition of the conditional probability of the event $Y = y$ given that $X = x$. The application of this formula to Weldon's example is show in Figure 5.3.

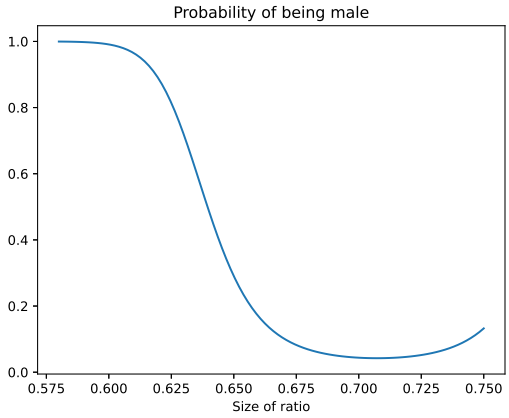


Figure 5.3: Conditional probability of being male as a function of the observed ratio in Weldon's data.

Remark 5.3.5 (Nonidentifiability)**.** *In general, the model $\mathcal{P}_F$ is not identifiable, because for any permutation σ of F, the parameters $((\theta_y)_{y\in F}, (\pi(y))_{y\in F})$ and $((\theta_{\sigma(y)})_{y\in F}, (\pi(\sigma(y)))_{y\in F})$ define the same probability distribution $P_{\boldsymbol{\theta},\pi}$. Said otherwise, in Weldon's data, it may be possible to identify parameters $\{(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), (\pi(1), \pi(2))\}$ which best fit the data, but then $\{(\mu_2, \sigma_2^2), (\mu_1, \sigma_1^2), (\pi(2), \pi(1))\}$ will be equally good, and there is no way to tell which parameters correspond to male and female populations.*

The EM algorithm

Given a realisation $\mathbf{x}_n = (x_1, \ldots, x_n)$, the goal is now to compute an estimator of the parameter $(\boldsymbol{\theta}, \pi)$. The log-likelihood of this realisation is

$$\ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi) = \sum_{i=1}^{n} \log \left(\sum_{y\in F} \pi(y) p(x_i; \theta_y) \right),$$

and as was already observed for the case of a mixture of two Gaussian distributions, in general its maximiser cannot be computed analytically. In contrast, if one assumes that the variables Y_i are also observed, the log-likelihood of the realisation $(\mathbf{x}_n, \mathbf{y}_n)$ becomes

$$\ell_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) = \sum_{i=1}^{n} \log \left(\pi(y_i) p(x_i; \theta_{y_i}) \right).$$

This quantity rewrites under the form

$$\ell_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) = \sum_{y\in F} \sum_{i=1}^{n} \mathbb{1}_{\{y=y_i\}} \{\log \pi(y) + \log p(x_i; \theta_y)\}, \tag{5.3}$$

from which one immediately deduces that the MLE of π is given by

$$\forall y \in F, \qquad \widehat{\pi}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i=y\}}, \tag{5.4}$$

while the MLE of each θ_y can be computed by solving the equation

$$0 = \nabla_{\theta_y} \ell_n(\mathbf{x}_n, \mathbf{y}_n; \widehat{\boldsymbol{\theta}}_n, \pi) = \sum_{i=1}^{n} \mathbb{1}_{\{y_i=y\}} \nabla_\theta \log \left(p(x_i; \widehat{\theta}_{n,y}) \right). \tag{5.5}$$

Exercise 5.3.6. Check that the MLE of π is indeed given by $\widehat{\pi}_n$ as defined above, and, for a mixture of Gaussian distributions, compute the MLE of the parameters (μ_y, σ_y^2).

If, instead of knowing exactly the *value* y_i of Y_i, you had a guess of its *distribution*, you could still define meaningful estimators of π and $\boldsymbol{\theta}$ by replacing, in (5.4) and (5.5), the indicator $\mathbb{1}_{\{y=y_i\}}$ by an estimate of the probability that $Y_i = y$. The Expectation-Maximisation (EM) algorithm is exactly based on this idea. It starts from an arbitrary choice of estimators $\boldsymbol{\theta}^{(0)}$ and $\pi^{(0)}$. Assuming that, at the beginning of the $(t+1)$-th iteration, the current estimators are $\boldsymbol{\theta}^{(t)}$ and $\pi^{(t)}$, it then proceeds in two steps.

1. For each $i \in \{1, \dots, n\}$, set

$$\forall y \in F, \qquad \rho^{(t+1)}(y|x_i) = \frac{\pi^{(t)}(y) p(x_i; \theta_y^{(t)})}{\sum_{y' \in F} \pi^{(t)}(y') p(x_i; \theta_{y'}^{(t)})},$$

   where we recall that $p(x;\theta)$ is either the pmf or the density of the probability distribution P_θ. By Remark 5.3.4, this quantity represents the conditional probability that Y_i takes the value y given the event $X_i = x_i$, under $\mathbb{P}_{\boldsymbol{\theta}^{(t)}, \pi^{(t)}}$.

2. For each $y \in F$, update the estimators by setting

$$\pi^{(t+1)}(y) = \frac{1}{n} \sum_{i=1}^{n} \rho^{(t+1)}(y|x_i),$$

   and letting $\theta_y^{(t+1)}$ solve

$$\sum_{i=1}^{n} \rho^{(t+1)}(y|x_i) \nabla_\theta \log \left(p(x_i; \theta_y^{(t+1)}) \right) = 0.$$

   These are the same expressions as (5.4) and (5.5), but the indicators $\mathbb{1}_{\{y=y_i\}}$ are replaced by the probabilities $\rho^{(t+1)}(y|x_i)$.

The first step is called *Expectation* because it allows to compute the *expected log-likelihood*

$$Q^{(t+1)}(\boldsymbol{\theta}, \pi) := \sum_{y \in F} \sum_{i=1}^{n} \rho^{(t+1)}(y|x_i) \{\log \pi(y) + \log p(x_i; \theta_y)\},$$

which is obtained by replacing $\mathbb{1}_{\{y=y_i\}}$ by $\rho^{(t+1)}(y|x_i)$ in (5.3). The second step is called *Maximisation* because $\boldsymbol{\theta}^{(t+1)}$ and $\pi^{(t+1)}$ are obtained by maximising $Q^{(t+1)}(\boldsymbol{\theta}, \pi)$.

The structure of the EM algorithm is very close to the one of the k-means algorithm for clustering, seen in the course *Introduction to Data Science*. Indeed, at each iteration:

- in the first step, instead of assigning each point of the dataset to a single cluster, the EM algorithm rather *estimates* the probability for this point to belong to each cluster (from a classification point of view, you should indeed think of the variable Y_i as indicating the class to which X_i belongs);
- in the second step, instead of computing the barycenter of each cluster, it estimates the associated parameters $\pi(y)$ and θ_y.

The miraculous thing about the k-means algorithm is that it decreases the WCSS at each step. The EM algorithm has a similar property: the log-likelihood increases at each step.

Proposition 5.3.7 (Increase in the log-likelihood)**.** *For any* $t \geq 0$*,*

$$\ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \leq \ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}).$$

Proof. We first rewrite the likelihood of a realisation $(\mathbf{x}_n, \mathbf{y}_n)$ under the form

$$\begin{aligned} L_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) &= \prod_{i=1}^{n} \pi(y_i) p(x_i; \theta_{y_i}) \\ &= \prod_{i=1}^{n} \frac{\pi(y_i) p(x_i; \theta_{y_i})}{\sum_{y \in F} \pi(y) p(x_i; \theta_y)} \times \prod_{i=1}^{n} \sum_{y \in F} \pi(y) p(x_i; \theta_y) \\ &= \mathbb{P}_{\boldsymbol{\theta}, \pi}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) \times L_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi), \end{aligned}$$

where the identification of the first term in the last line follows from Remark 5.3.4. This yields the identity

$$\ell_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) = \log \mathbb{P}_{\boldsymbol{\theta}, \pi}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) + \ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi).$$

Let us multiply both sides of this identity by

$$\mathbb{P}_{\boldsymbol{\theta}^{(t)}, \pi^{(t)}}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) = \prod_{i=1}^{n} \rho^{(t+1)}(y_i | x_i)$$

and sum over $\mathbf{y}_n$. Since the quantity $\ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi)$ does not depend on $\mathbf{y}_n$, the right-hand side writes

$$\begin{aligned} &\sum_{\mathbf{y}_n \in F^n} \left(\log \mathbb{P}_{\boldsymbol{\theta}, \pi}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) + \ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi) \right) \mathbb{P}_{\boldsymbol{\theta}^{(t)}, \pi^{(t)}}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) \\ &= H^{(t+1)}(\boldsymbol{\theta}, \pi) + \ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi), \end{aligned}$$

with

$$H^{(t+1)}(\boldsymbol{\theta}, \pi) := \sum_{\mathbf{y}_n \in F^n} \mathbb{P}_{\boldsymbol{\theta}^{(t)}, \pi^{(t)}}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n) \log \mathbb{P}_{\boldsymbol{\theta}, \pi}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n).$$

The left-hand side writes

$$\begin{aligned} &\sum_{\mathbf{y}_n \in F^n} \ell_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) \prod_{i=1}^{n} \rho^{(t+1)}(y_i | x_i) \\ &= \sum_{\mathbf{y}_n \in F^n} \left(\sum_{i=1}^{n} \log(\pi(y_i) p(x_i; \theta_{y_i})) \right) \left(\prod_{i=1}^{n} \rho^{(t+1)}(y_i | x_i) \right) \\ &= \sum_{\mathbf{y}_n \in F^n} \sum_{i=1}^{n} \left(\log(\pi(y_i) p(x_i; \theta_{y_i})) \rho^{(t+1)}(y_i | x_i) \prod_{j \neq i} \rho^{(t+1)}(y_j | x_j) \right) \\ &= \sum_{i=1}^{n} \left(\sum_{\mathbf{y}_n \in F^n} \log(\pi(y_i) p(x_i; \theta_{y_i})) \rho^{(t+1)}(y_i | x_i) \prod_{j \neq i} \rho^{(t+1)}(y_j | x_j) \right). \end{aligned}$$

For $i \in \{1, \dots, n\}$, we write each $\mathbf{y}_n \in F^n$ under the form $(y_i, \mathbf{y}_{-i})$ where $\mathbf{y}_{-i} = (y_j)_{j \neq i} \in F^{n-1}$, so that

$$\begin{aligned} &\sum_{\mathbf{y}_n \in F^n} \log(\pi(y_i) p(x_i; \theta_{y_i})) \rho^{(t+1)}(y_i | x_i) \prod_{j \neq i} \rho^{(t+1)}(y_j | x_j) \\ &= \sum_{y_i \in F} \log(\pi(y_i) p(x_i; \theta_{y_i})) \rho^{(t+1)}(y_i | x_i) \sum_{\mathbf{y}_{-i} \in F^{n-1}} \prod_{j \neq i} \rho^{(t+1)}(y_j | x_j). \end{aligned}$$

Remark that

$$\sum_{\mathbf{y}_{-i} \in F^{n-1}} \prod_{j \neq i} \rho^{(t+1)}(y_j | x_j) = \sum_{\mathbf{y}_{-i} \in F^{n-1}} \mathbb{P}_{\boldsymbol{\theta}^{(t)}, \pi^{(t)}}(\mathbf{Y}_{-i} = \mathbf{y}_{-i} | \mathbf{X}_n = \mathbf{x}_n) = 1,$$

so the expressions above reduce to the identity

$$\begin{aligned}\sum_{\mathbf{y}_n \in F^n} \ell_n(\mathbf{x}_n, \mathbf{y}_n; \boldsymbol{\theta}, \pi) \prod_{i=1}^n \rho^{(t+1)}(y_i|x_i) &= \sum_{i=1}^n \sum_{y_i \in F} \log(\pi(y_i) p(x_i; \theta_{y_i})) \rho^{(t+1)}(y_i|x_i) \\ &= Q^{(t+1)}(\boldsymbol{\theta}, \pi).\end{aligned}$$

As an intermediate conclusion, we therefore get

$$Q^{(t+1)}(\boldsymbol{\theta}, \pi) = H^{(t+1)}(\boldsymbol{\theta}, \pi) + \ell_n(\mathbf{x}_n; \boldsymbol{\theta}, \pi).$$

In particular,

$$\begin{aligned}\ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) &= Q^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)}) - H^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)}), \\ \ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) &= Q^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - H^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}),\end{aligned}$$

so

$$\begin{aligned}\ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - \ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) &= \left(Q^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - Q^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)})\right) \\ &\quad - \left(H^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - H^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)})\right).\end{aligned}$$

On the one hand, since $(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)})$ is defined as a maximiser of $Q^{(t+1)}(\boldsymbol{\theta}, \pi)$, we have

$$Q^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - Q^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)}) \geq 0.$$

On the other hand, introducing the notation $q(\mathbf{y}_n; \boldsymbol{\theta}, \pi) := \mathbb{P}_{\boldsymbol{\theta},\pi}(\mathbf{Y}_n = \mathbf{y}_n | \mathbf{X}_n = \mathbf{x}_n)$, which is such that for any $(\boldsymbol{\theta}, \pi)$, $\sum_{\mathbf{y}_n \in F^n} q(\mathbf{y}_n; \boldsymbol{\theta}, \pi) = 1$, we have

$$\begin{aligned}&H^{(t+1)}(\boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - H^{(t+1)}(\boldsymbol{\theta}^{(t)}, \pi^{(t)}) \\ &= \sum_{\mathbf{y}_n \in F^n} q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \log q(\mathbf{y}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - \sum_{\mathbf{y}_n \in F^n} q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \log q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \\ &= \sum_{\mathbf{y}_n \in F^n} q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \log \frac{q(\mathbf{y}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)})}{q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)})} \\ &\leq \sum_{\mathbf{y}_n \in F^n} q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)}) \left(\frac{q(\mathbf{y}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)})}{q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)})} - 1\right) \\ &= \sum_{\mathbf{y}_n \in F^n} \left(q(\mathbf{y}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) - q(\mathbf{y}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)})\right) \\ &= 0\end{aligned}$$

where we have used the convexity inequality $\log u \leq u - 1$, similarly to the proof of Proposition 5.1.12. This finally shows that $\ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t+1)}, \pi^{(t+1)}) \geq \ell_n(\mathbf{x}_n; \boldsymbol{\theta}^{(t)}, \pi^{(t)})$ and thus completes the proof. □

Just like the k-means algorithm, the fact that the log-likelihood increases along the EM algorithm does not ensure that it will converge to the global maximiser of the log-likelihood (which is anyway not unique by Remark 5.3.5). But in practice it is observed to converge quickly to some good enough estimator of $(\boldsymbol{\theta}, \pi)$.

5.A Exercises

Training exercises

Exercise 5.A.1 (Poisson model, continued)**.** We continue the study of the Poisson model initiated in Exercise 2.A.3.

1. Write the likelihood of the model.
2. Compute the MLE of λ.
3. Compute the Fisher information of the model.
4. Show that the MLE is efficient.
5. Using the MLE of λ, construct an asymptotic confidence interval for λ.

Exercise 5.A.2 (Geometric model)**.** We consider the geometric model $\{\mathcal{G}(p), p \in (0,1)\}$.

1. Write the likelihood of the model.
2. Compute the MLE $\widehat{p}_n$ of p, and show that it is strongly consistent.
3. Show that $\widehat{p}_n$ is asymptotically normal and compute its asymptotic variance.
4. Is $\widehat{p}_n$ asymptotically efficient?

Homework

Exercise 5.A.3 (MLE in the Weibull model)**.** A random variable $X > 0$ is said to follow the *Weibull* distribution with shape parameter $m > 0$ if

$$\forall x > 0, \qquad \mathbb{P}(X > x) = \exp\left(-x^m\right).$$

1. Warm-up.
   (a) Do you know a particular case of Weibull distribution?
   (b) If X follows the Weibull distribution with parameter m, what is the law of X^k for $k > 0$?
2. Maximum Likelihood Estimation.
   (a) Write the likelihood and the log-likelihood of a realisation $\mathbf{x}_n = (x_1, \dots, x_n)$.
   (b) Let us define
   $$\varphi(\mathbf{x}_n; m) = \frac{1}{m} + \frac{1}{n}\sum_{i=1}^{n}(1 - x_i^m)\log x_i.$$
   Show that $m \mapsto \varphi(\mathbf{x}_n; m)$ is decreasing and that, if $\mathbf{x}_n \neq (1, \dots, 1)$, there exists a unique $m_n(\mathbf{x}_n) > 0$ such that $\varphi(\mathbf{x}_n; m_n(\mathbf{x}_n)) = 0$. What about the case $\mathbf{x}_n = (1, \dots, 1)$?
   (c) Conclude that $\widehat{m}_n = m_n(\mathbf{X}_n)$ is the MLE of m.
3. To check consistency empirically, complete the notebook `Weibull.ipynb` available on Educnet. You will have to implement the numerical computation of $\widehat{m}_n$.
4. *Optional.* Proof of consistency.
   (a) For all $m > 0$, compute the limit, when $n \to +\infty$, of $\varphi(\mathbf{X}_n; m)$.
   You may admit, or prove, the identity
   $$\int_{y=0}^{+\infty} ((1-y)\log y)\exp(-y)\mathrm{d}y = -1.$$

For $a \in \mathbb{R}$, we define $[a]_+ = \max(a, 0)$, $[a]_- = \max(-a, 0)$, and recall that $|a| = [a]_+ + [a]_-$. To prove that $\widehat{m}_n$ is a consistent estimator m, we study $[\widehat{m}_n - m]_-$ and $[\widehat{m}_n - m]_+$ separately.

(b) *Study of* $[\widehat{m}_n - m]_-$.

i. Check that $\varphi'(\mathbf{x}_n; m) \leq -1/m^2$ for all $m > 0$.

ii. If $m_n(\mathbf{x}_n) \leq m$, show that $m - m_n(\mathbf{x}_n) \leq m^2 |\varphi(\mathbf{x}_n; m)|$.

iii. Deduce that $[\widehat{m}_n - m]_-$ converges almost surely to 0.

(c) *Study of* $[\widehat{m}_n - m]_+$.

i. Let $\epsilon > 0$. Show that if $m_n(\mathbf{x}_n) - m \geq \epsilon$, then

$$0 \leq \frac{1}{m+\epsilon} + \frac{1}{n}\sum_{i=1}^{n}(1 - x_i^m)\log x_i.$$

ii. Deduce that $[\widehat{m}_n - m]_+$ converges in probability to 0.

(d) What do you conclude?

Supplementary exercises

Exercise 5.A.4 (Parameter estimation in the Black–Scholes model)**.** In financial mathematics, the price of an asset at time $t \geq 0$ is modelled by a random variable $S_t > 0$. The Black–Scholes theory predicts that there exist $\lambda \in \mathbb{R}$ and $\sigma^2 > 0$ such that, for any $0 \leq t_0 < t_1 < \cdots < t_n$, the increments $(S_{t_i}/S_{t_{i-1}})_{1 \leq i \leq n}$ of this price are independent, with law

$$\forall i \in \{1, \ldots, n\}, \qquad \log \frac{S_{t_i}}{S_{t_{i-1}}} \sim \mathcal{N}\left(\lambda(t_i - t_{i-1}), \sigma^2(t_i - t_{i-1})\right).$$

In this exercise, a finite final time $T > 0$ is fixed and we assume that we observe the price of the asset at times $t_i = iT/n$, $i \in \{0, \ldots, n\}$, for a given observation frequency n. We then set

$$\forall i \in \{1, \ldots, n\}, \qquad X_i = \log \frac{S_{t_i}}{S_{t_{i-1}}} \sim \mathcal{N}\left(\lambda\frac{T}{n}, \sigma^2\frac{T}{n}\right).$$

1. Compute the MLE $(\widehat{\lambda}_n, \widehat{\sigma}_n^2)$ of (λ, σ^2).
2. Compute $\mathbb{E}_{\lambda,\sigma^2}[\widehat{\lambda}_n]$ and $\mathrm{Var}_{\lambda,\sigma^2}(\widehat{\lambda}_n)$. Is the estimator $\widehat{\lambda}_n$ biased?
3. When the observation frequency n goes to $+\infty$, is the estimator $\widehat{\lambda}_n$ consistent?
4. Is the estimator $\widehat{\sigma}_n^2$ biased?
5. Express $\widehat{\sigma}_n^2$ as a function of the random variables G_i defined by

$$\forall i \in \{1, \ldots, n\}, \qquad G_i = \frac{X_i - \lambda T/n}{\sigma\sqrt{T/n}}.$$

Is this estimator consistent?

Exercise 5.A.5 (Nonuniqueness of MLE)**.** For $\theta \in \mathbb{R}$, we denote by P_θ the uniform distribution on the interval $[\theta, \theta + 1]$.

1. Construct a moment estimator for θ.
2. Write the likelihood $L_n(\mathbf{x}_n; \theta)$ of the model. What do you observe?

3. For all $t \in [0,1]$, we introduce the estimator

$$\widehat{\theta}_n^t = (1-t)\left(\max_{1\leq i\leq n} X_i - 1\right) + t \min_{1\leq i\leq n} X_i.$$

   Show that $\widehat{\theta}_n^t$ is strongly consistent.

4. Compute the marginal distributions of $U = \min_{1\leq i\leq n} X_i - \theta$ and $V = \max_{1\leq i\leq n} X_i - \theta$, and deduce the bias of $\widehat{\theta}_n^t$.

5. Write the MSE $R(\widehat{\theta}_n^t;\theta)$ as a function of t, $\alpha = \mathbb{E}[U^2]$ and $\gamma = \mathbb{E}[(1-V)U]$. Which value of t minimises this quantity?

6. Compute the joint density of (U,V) and deduce an expression for the MSE of the optimal choice of $\widehat{\theta}_n^t$.

Lecture 6

Introduction to Bayesian Estimation

Contents

Introduction: frequentist and Bayesian approaches

The techniques of statistical inference which have been studied so far essentially find their justification in the fact that estimators are *consistent*, and therefore if enough data is collected, then the *true* value of θ is recovered. This approach is called *frequentist*. In certain situations, especially if one cannot collect so much data, it may be desirable to include a prior knowledge on the value of θ, independent from the observed data, in the estimation of θ. This is the foundational principle of *Bayesian* inference.

To illustrate this approach, assume that you do not feel very well and go to the doctor[1]. There are finitely many diseases from which you may suffer, denote by Θ the set thereof. To keep things simple, assume that the doctor's examination of your symptoms returns a random vector $\mathbf{X}_n = (X_1, \dots, X_n)$ of Bernoulli variables, where X_1 describes whether you have a high fever, X_2 describes whether your head aches, X_3 describes whether you sneeze, etc. Given your disease θ, this vector is random, because two people with the same disease may not exactly have the same symptoms. With the notation of Lecture 5, set $L_n(\mathbf{x}_n; \theta) = \mathbb{P}_\theta(\mathbf{X}_n = \mathbf{x}_n)$ the probability, for the disease θ, to observe the vector of symptoms $\mathbf{x}_n \in \{0,1\}^n$. This quantity depends on θ, because symptoms are generally not equally likely for all diseases.

Given the observation $\mathbf{x}_n$ of your symptoms, a frequentist doctor will diagnose you with the MLE, that is to say the disease which makes the realisation $\mathbf{x}_n$ the most likely. With the data of Table 6.1, if you have both fever and headache, you will thus come back home with medication for the bubonic plague[2]. And probably a bit of anxiety.

In contrast, the Bayesian doctor will include the knowledge that all diseases are not equally spread among the population. Before examining you, she already has the information that there is a probability $Q(\theta)$ that you suffer from the disease θ, as in Table 6.2. This probability measure Q on Θ is called the *prior* distribution.

To take the realisation $\mathbf{x}_n$ of your symptoms into account, the Bayesian doctor will then compute, for each disease $\theta \in \Theta$, what is the *conditional* probability that you suffer from θ, *given the fact that* $\mathbf{X}_n = \mathbf{x}_n$. Seeing your disease as a random variable in Θ, which we denote by D and has distribution

[1] This example was suggested by A. Parmentier.

[2] This example is obviously, and voluntarily, caricatural; in particular, it does not take into account the fact that the frequentist doctor is also able to compute confidence intervals to quantify how imprecise his diagnosis is.

Symptoms	FH	F	H	$\varnothing$
Flu	.6	.2	.15	.05
Bubonic plague	.7	.2	.08	.02
...				
No disease	.01	.005	.005	.98

Table 6.1: An example of (fictitious) law of symptoms for various diseases. The signification of the columns is the following: FH=Fever and Headache, F=Fever and no Headache, H=Headache and no Fever, $\varnothing$=neither Fever nor Headache. Therefore, among the population of people suffering from the flu, 60% have both Fever and Headache, 20% have only Fever, 15% have only Headache, 5% have neither Fever nor Headache.

Disease	Repartition in the population
Flu	.03
Bubonic plague	.00001
...	
No disease	.92

Table 6.2: The (fictitious) repartition of diseases in the population, which is taken as the prior distribution Q on Θ by the Bayesian doctor. Here, 3% of the population has the flu.

Q, she uses the Bayes formula to write, for any $\theta \in \Theta$,

$$\begin{aligned} Q(\theta|\mathbf{x}_n) := \mathbb{P}(D=\theta|\mathbf{X}_n=\mathbf{x}_n) &= \frac{\mathbb{P}(D=\theta, \mathbf{X}_n=\mathbf{x}_n)}{\mathbb{P}(\mathbf{X}_n=\mathbf{x}_n)} \\ &= \frac{\mathbb{P}(\mathbf{X}_n=\mathbf{x}_n|D=\theta)\mathbb{P}(D=\theta)}{\sum_{\vartheta\in\Theta}\mathbb{P}(\mathbf{X}_n=\mathbf{x}_n|D=\vartheta)\mathbb{P}(D=\vartheta)} \\ &= \frac{L(\mathbf{x}_n;\theta)Q(\theta)}{\sum_{\vartheta\in\Theta}L(\mathbf{x}_n;\vartheta)Q(\vartheta)}. \end{aligned}$$

She will then diagnose you with the disease θ which maximises the *posterior* distribution $Q(\theta|\mathbf{x}_n)$, thereby achieving a tradeoff between the prior knowledge on θ provided by Q and the information brought by the observation of the data $\mathbf{x}_n$.

Exercise 6.0.1. With the data of Tables 6.1 and 6.2, what is the diagnosis of the Bayesian doctor if you have both fever and headache? *Hint: you may remark that the denominator of the right-hand side in the Bayes formula does not depend on θ.*

6.1 The formalism of Bayesian estimation

We still work in the context of parametric estimation and therefore assume that we have fixed a parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, with Θ being either a finite set or a subset of $\mathbb{R}^q$.

6.1.1 Prior and posterior distribution

The *prior distribution* is a probability measure Q on Θ, which measures the credibility given to each value θ of the parameter before observing the data.

Definition 6.1.1 (Posterior distribution)**.** *For a prior distribution $Q(\cdot)$ on Θ and a realisation $\mathbf{x}_n$ of the sample, the* posterior distribution *is the probability measure $Q(\cdot|\mathbf{x}_n)$ defined on Θ by the formula*

$$Q(\mathrm{d}\theta|\mathbf{x}_n) := \frac{L_n(\mathbf{x}_n;\theta)Q(\mathrm{d}\theta)}{\displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q(\mathrm{d}\vartheta)},$$

where $L_n(\mathbf{x}_n;\theta)$ is the likelihood of the observation $\mathbf{x}_n$ as defined in Lecture 5.

The symbolic notation above means that, for any measurable subset $B \subset \Theta$,

$$Q(B|\mathbf{x}_n) = \frac{\displaystyle\int_{\theta\in B} L_n(\mathbf{x}_n;\theta)Q(\mathrm{d}\theta)}{\displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q(\mathrm{d}\vartheta)}.$$

In other words, the posterior $Q(\cdot|\mathbf{x}_n)$ has a density proportional to the likelihood $L_n(\mathbf{x}_n;\theta)$ with respect to the prior $Q(\cdot)$. The formula is rather abstract, but in practice:

- if the set Θ is countable, then for any $\theta \in \Theta$,

$$Q(\theta|\mathbf{x}_n) = \frac{L(\mathbf{x}_n;\theta)Q(\theta)}{\displaystyle\sum_{\vartheta\in\Theta} L(\mathbf{x}_n;\vartheta)Q(\vartheta)};$$

- if $Q(\cdot)$ has a density $q(\theta)$ with respect to the Lebesgue measure on $\mathbb{R}^q$, then $Q(\cdot|\mathbf{x}_n)$ has density

$$q(\theta|\mathbf{x}_n) := \frac{L_n(\mathbf{x}_n;\theta)q(\theta)}{\displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)q(\vartheta)\mathrm{d}\vartheta}.$$

The posterior distribution has to be interpreted as the credibility of θ given the observation $\mathbf{x}_n$.

6.1.2 The Beta-Bernoulli example

Assume that you toss a coin in order to know if it is biased, and you want to put a prior on the probability p to hit Head which favours values of p which are closer to $1/2$. A suitable family of such priors is the *Beta distribution*, which has already been studied in Exercise 1.A.6.

Definition 6.1.2 (Beta distribution)**.** *For any $a, b > 0$, the* Beta distribution *with parameters a and b, denoted by $\beta(a,b)$, is the probability measure on $[0,1]$ with density*

$$q_{a,b}(u) = \frac{1}{\mathrm{B}(a,b)} u^{a-1}(1-u)^{b-1},$$

where $\mathrm{B}(a,b)$ is the function

$$\mathrm{B}(a,b) = \int_{u=0}^{1} u^{a-1}(1-u)^{b-1}\mathrm{d}u = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

with Γ being Euler's function[3].

Using the fundamental property that, for any $a > 0$, $\Gamma(a+1) = a\Gamma(a)$, one may check that if $U \sim \beta(a,b)$, then

$$\mathbb{E}[U] = \frac{a}{a+b}, \qquad \mathrm{Var}(U) = \frac{ab}{(a+b)^2(a+b+1)}.$$

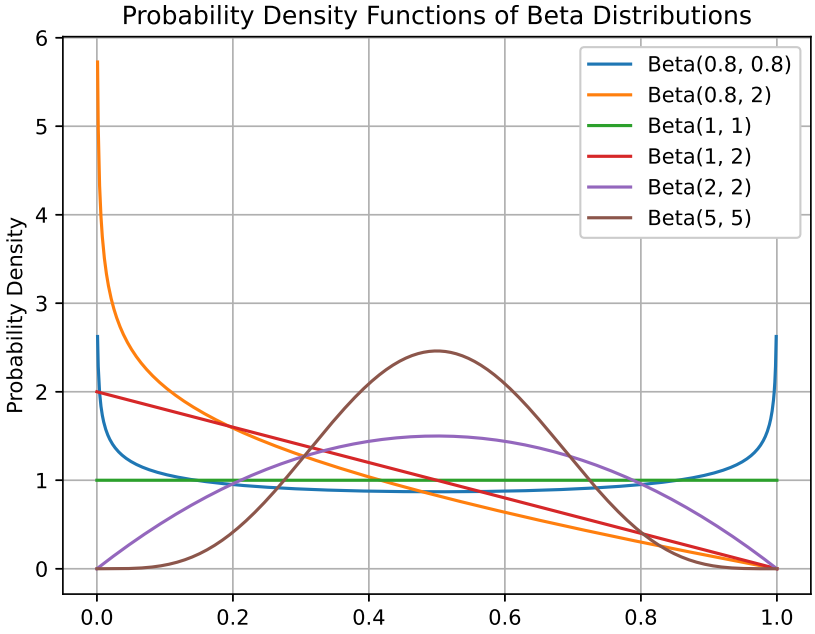


Figure 6.1: Density of the β distribution for various values of (a, b). If $a < 1$ (resp. $b < 1$) then the density has a mode in 0 (resp. 1); if $a + b > 2$ the mode is located at $(a - 1)/(a + b - 2)$.

The density $q_{a,b}$ is plotted for several values of (a, b) on Figure 6.1.

Using $\beta(a, b)$ as a prior for p in the Bernoulli model, following Definition 6.1.1 the posterior has density

$$\begin{aligned} q_{a,b}(p|\mathbf{x}_n) &= \frac{L_n(\mathbf{x}_n; p) q_{a,b}(p)}{\displaystyle\int_{\rho\in[0,1]} L_n(\mathbf{x}_n; \rho) q_{a,b}(\rho) \mathrm{d}\rho} \\ &= \frac{\left(\displaystyle\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\right) p^{a-1}(1-p)^{b-1}}{\displaystyle\int_{\rho\in[0,1]} \left(\prod_{i=1}^{n} \rho^{x_i}(1-\rho)^{x_i}\right) \rho^{a-1}(1-\rho)^{b-1} \mathrm{d}\rho} \\ &= \frac{p^{s+a-1}(1-p)^{n-s+b-1}}{\displaystyle\int_{\rho\in[0,1]} \rho^{s+a-1}(1-\rho)^{n-s+b-1} \mathrm{d}\rho}, \end{aligned}$$

with $s = x_1 + \cdots + x_n$. By definition of the Beta distribution,

$$\int_{\rho\in[0,1]} \rho^{s+a-1}(1-\rho)^{n-s+b-1} \mathrm{d}\rho = \mathrm{B}(s + a, n - s + b),$$

and therefore the posterior distribution satisfies

$$Q_{a,b}(\cdot|\mathbf{x}_n) = \beta(s + a, n - s + b).$$

On Figure 6.2, we plot the density of the posterior distribution $Q_{a,b}(\cdot|\mathbf{X}_n)$ for two different choices of Beta priors, where the data $\mathbf{X}_n$ are drawn according to the Bernoulli law with a certain parameter

[3] This identity is proved in Exercise 1.A.6.

p. In both cases, when n increases, the posterior becomes more and more peaked around the value of p. In fact, we shall prove in Exercise 6.2.10 below that, for this example, the posterior converges to the Dirac measure in p, *whatever the choice of the paramaters* a, b. In other words, when the size of the sample becomes large enough, the Bayesian estimation recovers the true value of the parameter, and the (arguably arbitrary) choice of the prior no longer influences the estimation of θ. This phenomenon is called the *consistency* of Bayesian estimation and studied in more detail in Subsection 6.2.2.

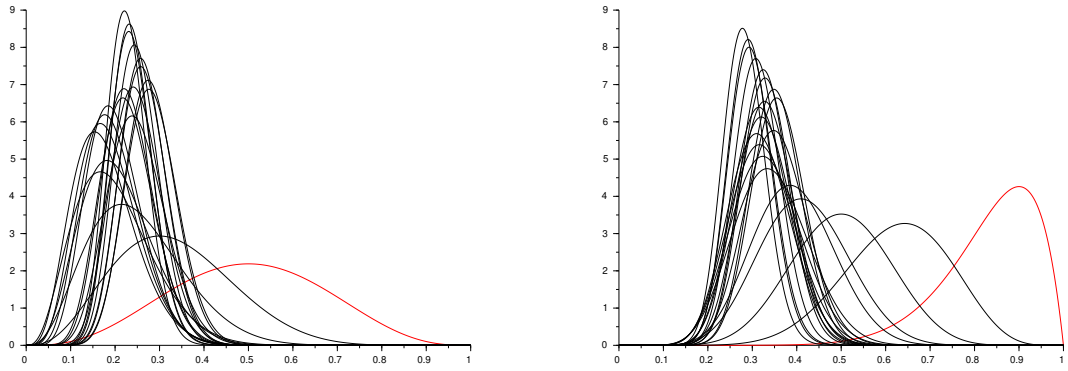


Figure 6.2: Plot of the density of the prior (in red) $\beta(a, b)$, then of the posterior (in black) for increasing values of n, with two different priors ($a = b = 4$ on the left, $a = 10$, $b = 2$ on the right) but the same realisation of $\mathbf{X}_n$, drawn with parameter $p = 0.3$. In both cases, the posterior concentrates around the value 0.3 as n increases.

6.1.3 Conjugate priors

In the Beta-Bernoulli example, the prior is chosen in a parametric family. In such a case, the parameters of the prior (a, b in the Beta-Bernoulli example) are called *hyperparameters*.

It is also remarkable that, in this example, the posterior distribution remains in the same parametric family as the prior. It is an instance of a *conjugate prior*.

Definition 6.1.3 (Conjugate prior)**.** *Consider the Bayesian estimation of a parameter $\theta \in \Theta$ for a parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ on a state space E. A parametric family $\mathcal{Q} = \{Q_h, h \in H\}$ of probability measures on Θ is called a* conjugate prior *for the model $\mathcal{P}$ if, for any realisation $\mathbf{x}_n \in E^n$ and for any hyperparameter $h \in H$, the posterior distribution $Q_h(\cdot|\mathbf{x}_n)$ on Θ still belongs to the family $\mathcal{Q}$.*

Exercise 6.1.4. Show that the Gamma family is a conjugate prior for the Poisson model.

6.2 Bayesian estimators

The posterior distribution encodes all relevant information on the Bayesian estimation of θ. However, since it is a probability distribution, it may not be easy to manipulate, in particular if the prior is not conjugate and thus the posterior cannot be characterised by a low-dimensional hyperparameter. Therefore, it is also useful to introduce 'frequentist-like' notions such as (Bayesian) *estimators* and *credible intervals*.

6.2.1 PM and MAP

Let $Q(\cdot|\mathbf{x}_n)$ denote the posterior distribution for the Bayesian estimation of a parameter $\theta \in \Theta$. For the next definition to make sense, we assume that Θ is a convex subset of $\mathbb{R}^q$.

Definition 6.2.1 (Posterior mean)**.** *The* Posterior Mean *(PM) is the mean of the posterior distribution, namely*

$$\widehat{\theta}_n^{\mathrm{PM}} := \theta_n^{\mathrm{PM}}(\mathbf{X}_n), \qquad \textit{with} \quad \theta_n^{\mathrm{PM}}(\mathbf{x}_n) := \int_{\theta\in\Theta} \theta Q(\mathrm{d}\theta|\mathbf{x}_n).$$

The random variable $\widehat{\theta}_n^{\mathrm{PM}}$ is a statistic which takes its values in Θ, therefore it is an estimator, in the frequentist sense, of θ.

In the next definition, we assume that either the set Θ is discrete and we set $q(\theta) := Q(\{\theta\})$ (resp. $q(\theta|\mathbf{x}_n) := Q(\{\theta\}|\mathbf{x}_n)$), or $Q(\mathrm{d}\theta)$ (resp. $Q(\mathrm{d}\theta|\mathbf{x}_n)$) has a density $q(\theta)$ (resp. $q(\theta|\mathbf{x}_n)$) with respect to the Lebesgue measure on $\mathbb{R}^q$.

Definition 6.2.2 (Maximum A Posteriori)**.** *Assume that for any $\mathbf{x}_n \in E^n$, the function $\theta \mapsto q(\theta|\mathbf{x}_n)$ has a unique maximum reached at $\theta = \theta_n^{\mathrm{MAP}}(\mathbf{x}_n)$. Then the* Maximum A Posteriori *(MAP) is defined by*

$$\widehat{\theta}_n^{\mathrm{MAP}} := \theta_n^{\mathrm{MAP}}(\mathbf{X}_n).$$

Similarly to the PM, the MAP is an estimator of Θ in the frequentist sense.

Remark 6.2.3. *To compute the MAP, it suffices to maximise the numerator $L_n(\mathbf{x}_n;\theta)q(\theta)$ of the posterior density, and therefore there is no need to evaluate the normalisation constant $\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)q(\vartheta)\mathrm{d}\vartheta$ in the denominator. In contrast, to compute the PM, this evaluation is necessary. This is often a nontrivial computational issue, which is discussed in more detail in Subsection 6.2.3.*

Remark 6.2.4. *If the prior is uniform (either on a finite set Θ or on a compact subset $\Theta \subset \mathbb{R}^q$) then it is easily checked that the MAP simply coincides with the MLE. More generally, taking the logarithm of the posterior, the MAP is observed to satisfy*

$$\theta_n^{\mathrm{MAP}}(\mathbf{x}_n) = \arg\max_{\theta\in\Theta} \left\{\ell_n(\mathbf{x}_n;\theta) + \log q(\theta)\right\}.$$

Therefore, the contribution of the prior can be seen as a penalisation *term in the optimisation leading to MLE, so that the MAP has to achieve simultaneously a high likelihood and a high credibility under the prior (the analogy with the notion of penalisation/regularisation in linear regression, discussed in Section A.5, should ring a bell!).*

Exercise 6.2.5. In the Beta-Bernoulli model from Subsection 6.1.2, compute the PM and the MAP and show that they are strongly consistent.

Remark 6.2.6. *When $a = b > 1$, the prior $\beta(a,b)$ has a unique mode (that is to say, its density has a unique maximum) at $p = 1/2$. Therefore, in the coin tossing experiment, if your prior belief is that the coin is not too much biased, and you have no reason to think that the bias is specifically oriented toward Head or Tail, the choice $a = b$ is natural. Then both the PM and the MAP rewrite under the form*

$$\widehat{p}_n^{\mathrm{PM/MAP}} = (1 - h_n^{\mathrm{PM/MAP}})\overline{X}_n + \frac{h_n^{\mathrm{PM/MAP}}}{2},$$

for some $h_n^{\mathrm{PM/MAP}} \in [0,1]$ which vanishes when $n \to +\infty$. You already encountered this biased estimator in Exercise 2.1.14. Therefore, the use of a prior in Bayesian estimation can be seen as a formalised way to introduce bias in frequentist estimators.

6.2.2 Consistency of Bayesian estimation

As is illustrated on Figure 6.2, it is generally expected that the posterior distribution $Q(\cdot|\mathbf{X}_n)$ concentrates around θ when $n \to +\infty$. When Θ is discrete, the definition of this phenomenon is straightforward.

Definition 6.2.7 (Consistency of the posterior for discrete parameters)**.** *If Θ is discrete, the Bayesian estimation of θ with the prior distribution Q is* consistent *if, for any $\theta \in \Theta$,*

$$\lim_{n\to+\infty} Q(\theta|\mathbf{X}_n) = 1, \qquad \mathbb{P}_\theta\textit{-almost surely.}$$

The condition in Definition 6.2.7 also rewrites

$$\forall \vartheta \neq \theta, \qquad \lim_{n\to+\infty} Q(\vartheta|\mathbf{X}_n) = 0, \qquad \mathbb{P}_\theta\text{-almost surely.}$$

In general, consistency can be checked directly from the computation of the posterior, see Exercise 6.A.4 for an application.

In the case where $\Theta \subset \mathbb{R}^q$, the definition is more technical but does not really differ.

Definition 6.2.8 (Consistency of the posterior)**.** *If $\Theta \subset \mathbb{R}^q$, the Bayesian estimation of θ with the prior distribution Q is* consistent *if, for any $\theta \in \Theta$, for any $\epsilon > 0$,*

$$\lim_{n\to+\infty} Q(\{\vartheta \in \Theta : \|\vartheta - \theta\| \geq \epsilon\}|\mathbf{X}_n) = 0, \qquad \mathbb{P}_\theta\textit{-almost surely.}$$

In other words, $\mathbb{P}_\theta$-almost surely, a sequence of random variables distributed under $Q(\cdot|\mathbf{X}_n)$ converges in probability to θ. A sufficient condition for consistency is given by the following proposition, which is particularly useful if the prior is conjugate and analytical formulas for the variance of the posterior are available.

Proposition 6.2.9 (Sufficient condition for consistency)**.** *For any $\mathbf{x}_n \in E^n$, let $V_n(\mathbf{x}_n)$ be the variance of the posterior distribution, defined by*

$$V_n(\mathbf{x}_n) := \int_{\theta\in\Theta} \|\theta - \theta_n^{\text{PM}}(\mathbf{x}_n)\|^2 \, Q(\mathrm{d}\theta|\mathbf{x}_n).$$

If the following two conditions hold:

(i) the PM $\widehat{\theta}_n^{\text{PM}}$ is strongly consistent,

(ii) for any $\theta \in \Theta$, $V_n(\mathbf{X}_n) \to 0$, $\mathbb{P}_\theta$-almost surely,

then the Bayesian estimation of θ with the prior distribution Q is consistent.

Proof. Let $\epsilon > 0$. By the Bienaymé–Chebychev inequality,

$$Q(\{\vartheta \in \Theta : \|\vartheta - \theta\| \geq \epsilon\}|\mathbf{X}_n) \leq \frac{1}{\epsilon^2} \int_{\vartheta\in\Theta} \|\vartheta - \theta\|^2 Q(\mathrm{d}\vartheta|\mathbf{X}_n).$$

The integral in the right-hand side is a Mean-Squared Error (see Lecture 2), whose bias-variance decomposition reads

$$\int_{\vartheta\in\Theta} \|\vartheta - \theta\|^2 Q(\mathrm{d}\vartheta|\mathbf{X}_n) = V_n(\mathbf{X}_n) + \|\widehat{\theta}_n^{\text{PM}} - \theta\|^2.$$

By the assumptions of the proposition, both terms in the right-hand converge to 0, almost surely. □

The empirical observation of Figure 6.2 may now be theoretically justified.

Exercise 6.2.10. Show that, in the Beta-Bernoulli model of Subsection 6.1.2, the Bayesian estimation of p with any Beta prior $Q_{a,b}$, $a, b > 0$, is consistent.

6.2.3 Credible regions

Assume that Θ is a subset of $\mathbb{R}$. Then a *credible interval* with level $1 - \alpha$ for θ is an interval I, whose bounds are statistics, such that

$$Q(I|\mathbf{X}_n) = 1 - \alpha.$$

Credible intervals are the Bayesian equivalent to confidence intervals. When $\Theta \subset \mathbb{R}^q$ with $q \geq 2$, the corresponding notion of *credible region* is defined similarly.

It is clear that a credible interval necessarily takes the form $[a, b]$, where a and b are the quantiles of respective order r and s of the posterior distribution, with $s - r = 1 - \alpha$. Just like for confidence intervals, the smaller $b - a$, the more informative the credible interval. When the prior is conjugate then computing credible intervals often reduces to analytical quantile calculations, otherwise the task may be dramatically more complicated.

In this case, which includes in particular situations where the parameter θ may be high-dimensional, numerical methods have to be employed, such as the Monte Carlo method which requires one to be able to *sample* iid realisations of θ from the posterior distribution. The latter task may however not be trivial, and dedicated methods, such as the Markov Chain Monte Carlo method, have been profusely developed in this purpose since the 1950's[4].

6.A Exercises

Training exercises

Exercise 6.A.1 (Gaussian model)**.** We consider the problem of estimating the mean μ in the Gaussian model $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}\}$, where $\sigma^2 > 0$ is assumed to be known. We consider for μ the Gaussian prior $\mathcal{Q} = \{\mathcal{N}(m, s^2), m \in \mathbb{R}, s^2 > 0\}$.

1. Compute the posterior distribution $Q(\cdot|\mathbf{x}_n)$ for $\mathbf{x}_n \in \mathbb{R}^n$.
2. Is the Gaussian prior conjugate?
3. Compute the PM and the MAP.
4. Show that the Bayesian estimation is consistent.
5. Construct a credible interval for μ with level $1 - \alpha$.

Exercise 6.A.2 (Conjugate hyperprior)**.** Let $\mathcal{Q} = \{Q_h, h \in H\}$ be a parametric family of probability measures on Θ. The set of (finite) mixtures of $\mathcal{Q}$ is the set $\mathcal{Q}^*$ of probability measures on Θ which write under the form

$$Q^* = t_1 Q_{h_1} + \cdots + t_m Q_{h_m},$$

with $m \geq 1$, $h_1, \ldots, h_m \in H$, and $t_1, \ldots, t_n \geq 0$, $t_1 + \cdots + t_n = 1$. Show that if $\mathcal{Q}$ is a conjugate prior for the estimation of θ, then so is $\mathcal{Q}^*$.

Homework

Exercise 6.A.3. We work with the Gamma-Poisson model of Exercise 6.1.4, which means that we observe integer random variables $X_1, \ldots, X_n$ which are assumed to have a Poisson distribution with parameter $\theta > 0$, and we put a $\Gamma(a, \lambda)$ prior on θ, which we denote by $Q_{a,\lambda}(\mathrm{d}\theta)$.

Recall that the expressions for the expectation and variance of Gamma distributions are given in Exercise 1.A.4.

1. Compute the MAP and the PM of θ.
2. Show that they are strongly consistent.
3. Show that the Bayesian estimation is consistent.

[4] The Markov Chain Monte Carlo (MCMC) method is presented in the course *Stochastic Processes and their Applications* by J.-F. Delmas. To learn more about the use of MCMC in Bayesian statistics, the reader is referred to the book [5].

4. With Python, choose two arbitrary pairs of hyperparameters (a_1, λ_1) and (a_2, λ_2), and plot the densities of the associated prior $Q_{a_1,\lambda_1}(\mathrm{d}\theta) = \Gamma(a_1, \lambda_1)$ and $Q_{a_2,\lambda_2}(\mathrm{d}\theta) = \Gamma(a_2, \lambda_2)$. Then fix a value of θ and draw a sample $(X_1, \dots, X_N)$ of independent $\mathcal{P}(\theta)$ variables. Now, for $n = 1, \dots, N$, plot the densities of the two posterior distributions $Q_{a_1,\lambda_1}(\mathrm{d}\theta|\mathbf{X}_n)$ and $Q_{a_2,\lambda_2}(\mathrm{d}\theta|\mathbf{X}_n)$. Comment on the result.

Exercise 6.A.4. You are in charge of detecting gravitational waves emitted by accelerated masses (say, black holes) in the universe. There are $K \geq 1$ black holes likely to emit waves, and when the k-th one emits a wave, each one of the n sensors of your experiment returns a one-dimensional signal of the form $X_i = \mu_k + \epsilon_i$, where $\mu_k \in \mathbb{R}$ is the signature of the k-th black hole, and ϵ_i is a random variable which represents the measure noise associated with the i-th sensor. All noise variables $\epsilon_1, \dots, \epsilon_n$ are assumed to be independent and identically distributed according to the law $\mathcal{N}(0, \sigma^2)$ for some known variance parameter σ^2.

Thus, given the observation of the vector $\mathbf{X}_n = (X_1, \dots, X_n)$, your purpose is to determine which of the K black holes has emitted the signal. Furthermore, you know from previous observations that the prior probability for the k-th black hole to emit a wave is $Q(\mu_k)$, where Q is a probability measure on the finite set $\Theta = \{\mu_k, 1 \leq k \leq K\}$. We assume that $Q(\mu_k) > 0$ for any $k \in \{1, \dots, K\}$. The signatures μ_k are assumed to be pairwise distinct, so that you can identify k from the estimation of μ_k.

1. Compute the posterior distribution $Q(\cdot|\mathbf{x}_n)$ associated with a possible realisation $\mathbf{x}_n \in \mathbb{R}^n$ of the sample $\mathbf{X}_n$.

2. *Optional.* For any $k, k' \in \{1, \dots, K\}$, we define

$$U^n_{k,k'} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_{k'})^2 - (X_i - \mu_k)^2.$$

   (a) Express $Q(\mu_k|\mathbf{X}_n)$ in terms of the quantities $U^n_{k,k'}$.

   (b) Compute the almost sure limit of $U^n_{k,k'}$ when $n \to +\infty$ under $\mathbb{P}_{\mu_k}$.

   (c) Deduce that the Bayesian estimation of μ is consistent.

3. In the notebook `BlackHoles.ipynb` available on Educnet, you can visualise the consistency of the estimation of μ. Depending on K, what seems to be the order of magnitude for n at which the true value of μ is recovered?

Supplementary exercises

Exercise 6.A.5 (Bayesian regression)**.** We consider the Gaussian linear model from Section 3.3, where $\epsilon_1, \dots, \epsilon_n$ are independent $\mathcal{N}(0, \sigma^2)$ variables with variance σ^2 which is assumed to be known. We consider the Bayesian estimation of β. We take as a prior the Gaussian measure $\mathcal{N}_{p+1}(0, kI_{p+1})$ for some parameter $k > 0$, which amounts to assuming that $\beta_0, \dots, \beta_p$ are iid centered Gaussian variables with variance k.

1. Show that the MAP $\widehat{\beta}^{\mathrm{MAP}}$ is the minimiser of $\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2 + h\|\beta\|^2$ for some $h > 0$ to be expressed in terms of k and σ^2. The Bayesian estimation of β is therefore equivalent to the ridge regularisation introduced in Section A.5.

2. Which prior on β should one take in order to recover the LASSO penalisation, also described in Section A.5?

☑ Intermediate Revision Sheet

Exercise 1 (Parametric estimation)**.** For any $\theta > 0$, we consider the probability density

$$p(x;\theta) = \mathbb{1}_{\{x>0\}} \sqrt{\frac{2}{\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right).$$

1. Show that, under $\mathbb{P}_\theta$, X_1 has the same law as $\sqrt{\theta}|G|$, where $G \sim \mathcal{N}(0,1)$.
2. (a) Compute $\mathbb{E}[|G|]$ and deduce a strongly consistent moment estimator $\widetilde{\theta}_n$ of θ.
   (b) Show that $\widetilde{\theta}_n$ is biased.
   (c) Show that $\widetilde{\theta}_n$ is asymptotically normal and compute its asymptotic variance.
3. (a) Compute the MLE $\widehat{\theta}_n$ of θ.
   (b) Show that $\widehat{\theta}_n$ is unbiased and strongly consistent.
   (c) Show that $\widehat{\theta}_n$ is asymptotically normal and compute its asymptotic variance.
   (d) Which of the estimators $\widetilde{\theta}_n$ and $\widehat{\theta}_n$ has the smaller asymptotic variance?
   (e) Show that the MLE $\widehat{\theta}_n$ is efficient.
4. (a) Construct an asymptotic confidence interval with level $1-\alpha$ for θ.
   (b) Show that under $\mathbb{P}_\theta$, the random variable $\widehat{\theta}_n/\theta$ is free and deduce an exact confidence interval with level $1-\alpha$ for θ, first with equal risk of under- and overestimation, and then with zero risk of overestimation.
5. The *inverse Gamma distribution* with parameters $a, b > 0$ is the probability distribution on $(0, +\infty)$ with density

$$q_{a,b}(\theta) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\theta}\right)^{a+1} \exp\left(-\frac{b}{\theta}\right).$$

   Its mean is $b/(a-1)$ (for $a > 1$) and its variance is $b^2/(a-1)^2(a-2)$ (for $a > 2$).

   (a) If $q_{a,b}$ is a prior for θ, what is the posterior?
   (b) What can you deduce on the family of inverse Gamma distributions?
   (c) Compute the posterior mean $\widehat{\theta}_n^{\mathrm{PM}}$ and show that it is strongly consistent.
   (d) Show that the Bayesian estimation is consistent.

Exercise 2 (Prediction interval in the linear Gaussian model)**.** We consider the linear Gaussian model as in Section 3.3. Recall the notations of subsection 4.2.4 and that $\widehat{\beta}$ denotes the LSE of the model.

1. Given a new feature $x'_{n+1} = (1, x^1_{n+1}, \dots, x^p_{n+1}) \in \mathbb{R}^{1\times(p+1)}$, derive a confidence interval with level $1-\alpha$ for $x'_{n+1}\beta$, which corresponds to the underlying signal of this new prediction.
2. In the next table, we report the selling price of $n = 10$ cars of the same model as a function of their age and number of kilometres travelled. Compute a 95% prediction interval for the selling price of a 4 year old car with 80000 km. Compute similarly a 95% confidence interval for the signal of the selling price of this car.

Selling price	Kilometres	Age (in months)
18500	32000	24
14500	50000	36
22000	18000	12
9500	87000	60
7800	120000	84
16500	40000	30
11200	70000	48
19900	25000	18
8800	95000	72
13200	60000	42

Part II

Hypothesis Testing

Lecture 7

The Formalism of Statistical Hypothesis Testing

Contents

Introduction: the Lady Tasting Tea experiment[1]

In 1919, Ronald Fisher worked as a statistician at Rothamsted Research, an agricultural research institute in the UK. He once proposed a cup of tea to his colleague Muriel Bristol, who was a phycologist. She declined and argued that she preferred having the milk poured into the cup before the tea. Fisher did not believe that whether the milk was poured before or after the tea could affect the flavour, and thus designed the following experiment: he prepared eight cups, four with the milk poured first and four with the tea poured first, and had Bristol blindly taste the eight cups. She was able to decide correctly in which order the milk and the tea had been poured for all cups.

Fisher then developed the following argument. Assume that the lady is not actually able to tell whether the milk or the tea have been poured first, and thus answers at random. Fisher called this assumption the *null hypothesis*, because it assumes the absence of the effect that the experiment is trying to evidence — namely, Bristol's ability to distinguish between the two preparations. Since Bristol was aware that four cups of each type had been prepared, the probability under the null hypothesis that she gave all correct answers is

$$p = \frac{1}{\binom{8}{4}} = \frac{1}{70} \simeq 1,4\%,$$

which is called the *p-value*. Since this probability is small, Fisher declared that the result of the experiment was *too unlikely* under the null hypothesis, and thus rejected the latter, hence concluding that Bristol was actually able to tell whether the milk or the tea was poured first.

Fisher's argument is considered as one of the first attempts to formalise the design and statistical analysis of scientific experiments. It is the basis of the theory of *hypothesis testing*, which was then developed in particular by Neyman and Pearson in the 1930's. In this Lecture, we present this theory in the framework of parametric estimation.

[1] This experiment is reported in Fisher's book *The design of experiments* (1935).

7.1 General formalism

In this Section, a parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is fixed, with $\Theta \subset \mathbb{R}^q$. The state space for the sample $\mathbf{X}_n = (X_1, \dots, X_n)$ is denoted by E^n.

7.1.1 Null and alternative hypotheses, test

Let H_0, H_1 be a partition of Θ into two subsets:

- H_0 is the *null hypothesis*,
- H_1 is the *alternative hypothesis*.

An hypothesis is called *simple* if it contains a single element, otherwise it is called *composite*.

Definition 7.1.1 (Test)**.** *A* test *of H_0 against H_1 is a decision rule determining, given an observation $\mathbf{x}_n \in E^n$, whether $\theta \in H_0$ or $\theta \in H_1$.*

In other words, a test is a deterministic function $E^n \to \{H_0, H_1\}$. It is characterised by its *region of rejection*, or *critical region*.

Definition 7.1.2 (Region of rejection)**.** *The* region of rejection *of a test is the set W_n of realisations $\mathbf{x}_n \in E^n$ for which H_0 is rejected.*

Example 7.1.3 (Bernoulli model)**.** *A sequence of coin tosses is modelled by Bernoulli random variables with parameter $p \in [0, 1]$. The experimenter wants to know whether the coin is biased or not. She sets:*

- *$H_0 = \{p = 1/2\}$: the coin is not biased,*
- *$H_1 = \{p \neq 1/2\}$: the coin is biased.*

The null hypothesis is simple, while the alternative hypothesis is composite.

For a sufficiently large sample size n, the Law of Large Numbers asserts that $\overline{X}_n$ is close to p. As a consequence, an intuitive test consists in rejecting H_0 as soon as $\overline{X}_n$ is 'far enough' from $1/2$. Formally, this amounts to taking

$$W_n = \{\mathbf{x}_n \in \{0, 1\}^n : |\overline{x}_n - 1/2| \geq a\},$$

for some $a \in (0, 1/2)$, which has to be determined in order to control the risk of taking a wrong decision — this notion will be made precise below.

Remark 7.1.4. *In Example 7.1.3, the* null *hypothesis is that under which there is* no *bias. This is a general fact: when one wants to show that a certain effect is present in the data, one defines the null hypothesis as the* absence *of this effect.*

7.1.2 Type I and type II errors, level and statistical power

Since the sample $\mathbf{x}_n$ is the realisation of a random variable $\mathbf{X}_n$, it may happen that the test returns an incorrect result. Two types of errors are distinguished. We recall that W_n is the region of rejection defined in Definition 7.1.2.

Definition 7.1.5 (Type I and type II errors)**.** *A* type I error[2] *is the incorrect rejection of H_0. It is measured by the* type I risk $\theta \in H_0 \mapsto \mathbb{P}_\theta(\mathbf{X}_n \in W_n)$.

A type II error[3] *is the incorrect acceptance of H_0. It is measured by the* type II risk $\theta \in H_1 \mapsto \mathbb{P}_\theta(\mathbf{X}_n \notin W_n)$.

[2] *Erreur de première espèce* en français.
[3] *Erreur de seconde espèce* en français.

Remark 7.1.6. *With the interpretation that the null hypothesis is the absence of the effect that one wants to identify, the type I error corresponds to a* false positive*, as the test incorrectly concludes to the presence of the effect. On the contrary, the type II error corresponds to a* false negative*.*

In Example 7.1.3, taking a very small leads the test to reject H_0 as soon as $\overline{X}_n$ is not very close to $1/2$, and therefore increases the type I risk. On the contrary, taking a close to $1/2$ makes the test reject H_0 for few values of the sample set, and thus increases the type II risk. This example shows that one cannot minimise both risks simultaneously. In order to select a, the *Neymann–Pearson approach* consists in:

(i) fixing a *level* $\alpha \in (0,1)$, usually 1%, 5% or 10%;

(ii) defining the rejection region which minimises the type II risk under the constraint that, for any $\theta \in H_0$, the type I risk is lower than α.

In general, the minimisation of the type II risk makes the supremum over H_0 of the type I risk become equal to α. The *statistical power* of the test is the function

$$\theta \in H_1 \mapsto \mathbb{P}_\theta(\mathbf{X}_n \in W_n) = 1 - \text{type II error}.$$

Remark 7.1.7. *This procedure induces a dissymmetry between type I and type II errors: fixing a level first shows that, in the experimenter's eyes, it is more important to control the type I error than the type II error. This dissymmetry must be taken into account when the null and alternative hypotheses are defined. For instance, if one looks for the presence of Higgs' boson at CERN, defining H_0 as the absence of the particle allows to control the risk of being wrong when claiming that the particle has been discovered, at the price of accepting a possibly large type II error, that is to say not being able to claim that the particle has been detected with sufficient significance.*

To compute the type I and type II risks in the Bernoulli model of Example 7.1.3, we use the approximation

$$\overline{X}_n \simeq p + \sqrt{\frac{p(1-p)}{n}} G, \qquad G \sim \mathcal{N}(0,1),$$

based on the Central Limit Theorem. Then the type I risk is

$$\mathbb{P}_{1/2}(\mathbf{X}_n \in W_n) \simeq \mathbb{P}(|G| \geq 2a\sqrt{n}),$$

so that the level of the test is α if and only if a is chosen so that $2\sqrt{n}a = \phi_{1-\alpha/2}$, with ϕ_r the quantile of order r of the standard Gaussian distribution. With this choice, the statistical power of the test is plotted for several values of n on Figure 7.1. Obviously, its minimum value is α because the function $p \mapsto \mathbb{P}_p(\mathbf{X}_n \in W_n)$ is continuous on $[0,1]$ and the test is designed for this function to take the value α at $p = 1/2$. However, for any $p \neq 1/2$, that is to say for any $p \in H_1$, the statistical power increases with n.

That the test behaves better when the size of the sample increases is a natural requirement, and motivates the following definition.

Definition 7.1.8 (Asymptotic properties)**.** *A test with rejection region W_n is* consistent *if*

$$\forall \theta \in H_1, \qquad \lim_{n \to +\infty} \mathbb{P}_\theta(\mathbf{X}_n \in W_n) = 1.$$

The test is of asymptotic level α *if*[4]

$$\alpha = \sup_{\theta \in H_0} \limsup_{n \to +\infty} \mathbb{P}_\theta(\mathbf{X}_n \in W_n).$$

[4]We recall that the limsup and liminf of a sequence $(a_n)_{n \geq 1}$ are defined by $\limsup_{n\to+\infty} a_n = \lim_{n\to+\infty} \sup_{k \geq n} a_k$ and $\liminf_{n\to+\infty} a_n = \lim_{n\to+\infty} \inf_{k \geq n} a_k$.

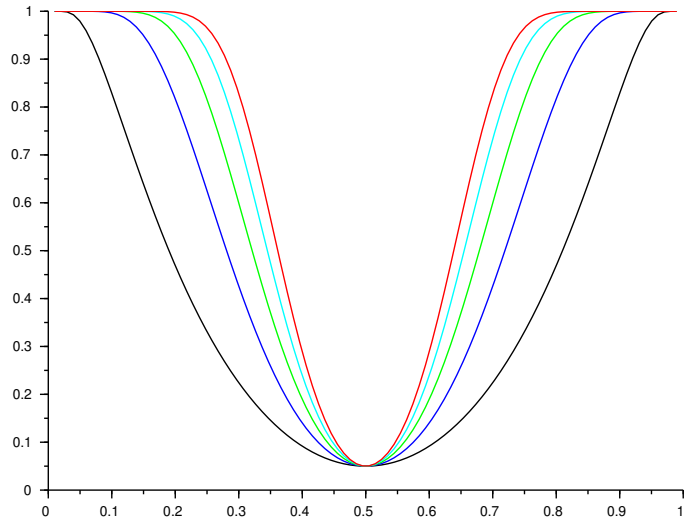


Figure 7.1: Statistical power of the test of level $\alpha = 5\%$ in the Bernoulli model, for $n = 10, 20, 30, 40, 50$. The larger n, the more peaked the curve.

Exercise 7.1.9. Using the Dominated Convergence Theorem 1.4.2, show that, in the Bernoulli model of Example 7.1.3, the test with rejection region $W_n = \{\mathbf{x}_n \in \{0,1\}^n : |\overline{x}_n - 1/2| \geq \phi_{1-\alpha/2}/(2\sqrt{n})\}$ is consistent.

Remark 7.1.10. *Tests with a poor statistical power face the risk of returning a type II error with a large probability. As Figure 7.1 shows, increasing the size of the sample allows to reduce this risk. A standard value for the acceptable power of a test is* 80%. *Yet, Figure 7.1 also shows that this value can generally not be reached uniformly over* H_1. *For hypotheses of the form* $H_0 = \{\theta = \theta_0\}$, $H_1 = \{\theta \neq \theta_0\}$, *a possible approach consists in fixing a power level* ρ *(say* $\rho = 0.8$*) and a threshold* $\delta > 0$, *and looking for* n *such that*

$$\forall \theta \in H_1^\delta = \{|\theta - \theta_0| \geq \delta\}, \qquad \mathbb{P}_\theta(\mathbf{X}_n \in W_n) \geq \rho.$$

Remark 7.1.11. *To alleviate notation, we shall often write rejection regions as* events, *depending on* $\mathbf{X}_n$, *rather than subsets of* E^n*: for instance, in the Bernoulli model of Example 7.1.3, we may write*

$$W_n = \{|\overline{X}_n - 1/2| \geq \phi_{1-\alpha/2}/(2\sqrt{n})\},$$

rather than the longer expression

$$W_n = \{\mathbf{x}_n \in \{0,1\}^n : |\overline{x}_n - 1/2| \geq \phi_{1-\alpha/2}/(2\sqrt{n})\}.$$

Accordingly, we shall also sometimes write such expressions as $\mathbb{P}_\theta(W_n)$ *instead of* $\mathbb{P}_\theta(\mathbf{X}_n \in W_n)$.

A natural question then arises when you have several tests available for the same pair of null and alternative hypotheses, and you want to know which one is the best. According to the approach described above,

(i) only tests with the same level α can be compared;

(ii) a test is then preferable to another one if it is more powerful, that is to say that it has lower type II risk.

7.1.3 General procedure to construct a test

First steps

The first steps to construct a test consist in specifying the model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ in which you work, and then defining the null hypothesis H_0 and the alternative hypothesis H_1. Do not forget that H_0 and H_1 do not play symmetric roles: they must be chosen so that type I errors be avoided as a priority. In particular, you must fix H_1 first: it must contain the effect that the experience aims at evidencing. Then define H_0 as the absence of this effect.

You next have to define the rejection region W_n of your test. To proceed, you need to look at H_1. We focus on the case where there exist some function $g : \Theta \to \mathbb{R}$ and some given value $g_0 \in \mathbb{R}$ such that H_1 can be written under one of the following two forms:

- $H_1 = \{g(\theta) > g_0\}$ or $H_1 = \{g(\theta) < g_0\}$, in which case the test is called *one-sided*[5];
- $H_1 = \{g(\theta) \neq g_0\}$, in which case the test is called *two-sided*[6].

Assume in addition that an estimator Z_n of $g(\theta)$ is available: then you should reject H_0 if Z_n takes large values in the one-sided case with $H_1 = \{g(\theta) > g_0\}$, small values in the one-sided case with with $H_1 = \{g(\theta) < g_0\}$, and values which are far from g_0 in the two-sided case. This naturally leads to rejection regions W_n with respective forms:

- $\{Z_n \geq g_0 + a_n\}$ in the one-sided case with $H_1 = \{g(\theta) > g_0\}$;
- $\{Z_n \leq g_0 - a_n\}$ in the one-sided case with $H_1 = \{g(\theta) < g_0\}$;
- $\{Z_n \notin (g_0 - a_n, g_0 + b_n)\}$, in the two-sided case;

for some *thresholds* $a_n, b_n \geq 0$.

Fixing a level $\alpha \in (0, 1/2)$, the whole game then lies in the determination of these thresholds to ensure that the test has smallest possible type II risk under the constraint that the type I risk remains below α.

One-sided tests

We first address one-sided tests, with $H_1 = \{g(\theta) > g_0\}$. For any $a_n \geq 0$, the type I risk is

$$\mathbb{P}_\theta(W_n) = \mathbb{P}_\theta(Z_n \geq g_0 + a_n), \qquad \theta \in H_0.$$

Let us make the assumption that there exists $\theta_0 \in H_0$ such that, whatever the choice of $a_n \geq 0$,

$$\sup_{\theta \in H_0} \mathbb{P}_\theta(Z_n \geq g_0 + a_n) = \mathbb{P}_{\theta_0}(Z_n \geq g_0 + a_n).$$

Then, denoting by $z_{\theta_0,n,r}$ the quantile of order r of the law of Z_n under $\mathbb{P}_{\theta_0}$, we deduce that

$$\sup_{\theta \in H_0} \mathbb{P}_\theta(W_n) \leq \alpha \qquad \text{if and only if} \qquad g_0 + a_n \geq z_{\theta_0,n,1-\alpha}.$$

Moreover, for any $\theta \in H_1$, the type II risk $\mathbb{P}_\theta(Z_n < g_0 + a_n)$ is a nondecreasing function of $g_0 + a_n$, so it is minimised when $g_0 + a_n$ takes the minimal admissible value $z_{\theta_0,n,1-\alpha}$. We conclude that the final rejection region, with maximal power under the constraint that the level be lower than α, is

$$W_n = \{Z_n \geq z_{\theta_0,n,1-\alpha}\}.$$

If $H_1 = \{g(\theta) > g_0\}$, then the final rejection region writes symmetrically $W_n = \{Z_n \leq z_{\theta_0,n,\alpha}\}$.

This construction is illustrated on the following example.

[5] *Unilatéral* en français.
[6] *Bilatéral* en français.

Example 7.1.12 (The Exponential model)**.** *A smartphone manufacturer claims that the average lifespan of its products is at least of* 3 *years. A consumer organisation carries out a study over* 1000 *devices and observes an average lifespan of* 2.8 *years. At which level can it be concluded that the manufacturer is lying?*

We answer this question by following the guideline described above.

(i) In order to keep the computation simple, we assume that the lifespan of a device is exponentially distributed, and take as a model

$$\mathcal{P} = \{\mathcal{E}(\lambda), \lambda > 0\}.$$

(ii) Since the goal is to determine whether the lifespan of the products is significantly below the claimed value of 3*, we set* $H_1 = \{\mathbb{E}_\lambda[X_1] < 3\}$ *and* $H_0 = \{\mathbb{E}_\lambda[X_1] \geq 3\}$*. This rewrites*

$$H_0 = \{g(\lambda) \geq g_0\}, \qquad H_1 = \{g(\lambda) < g_0\},$$

with $g(\lambda) = \mathbb{E}_\lambda[X_1] = 1/\lambda$ *and* $g_0 = 3$*.*

(iii) We take $\overline{X}_n$ *as an estimator of* $g(\lambda)$*, and thus take a rejection region of the form*

$$W_n = \{\overline{X}_n \leq g_0 - a_n\}.$$

Notice that since $\overline{X}_n > 0$*, we must necessarily take* $a_n < g_0$*.*

(iv) Let us compute the type I risk. For all $\lambda \in H_0 = (0, \lambda_0]$ *with* $\lambda_0 = 1/g_0$*, for all* $a_n \in [0, g_0)$*,*

$$\mathbb{P}_\lambda(W_n) = \mathbb{P}_\lambda(\overline{X}_n \leq g_0 - a_n) = \mathbb{P}(S_n/\lambda \leq g_0 - a_n) = \mathbb{P}(S_n \leq \lambda(g_0 - a_n)),$$

where the random variable $S_n = \lambda\overline{X}_n$ *is* free *in the sense of Definition 4.2.1, with law* $\Gamma(n, n)$*. As a consequence, the type I risk* $\lambda \mapsto \mathbb{P}_\lambda(W_n)$ *is nondecreasing (because* $g_0 - a_n > 0$*), so that*

$$\sup_{\lambda \in H_0} \mathbb{P}_\lambda(W_n) = \mathbb{P}_{\lambda_0}(W_n).$$

We thus conclude that the rejection region of the test writes

$$W_n = \{\overline{X}_n \leq x_{\lambda_0,n,\alpha}\},$$

where $x_{\lambda_0,n,\alpha}$ *is the quantile of order* α *of* $\overline{X}_n$ *under* $\mathbb{P}_{\lambda_0}$*.*

(v) To compute the quantile, we still use the fact that, under $\mathbb{P}_{\lambda_0}$*,* $S_n = \lambda_0\overline{X}_n \sim \Gamma(n, n)$*, so that for any* $x > 0$*,*

$$\mathbb{P}_{\lambda_0}(\overline{X}_n \leq x) = \mathbb{P}(S_n \leq \lambda_0 x) \leq \alpha \qquad \textit{if and only if} \qquad \lambda_0 x \leq \gamma_{n,\alpha},$$

where $\gamma_{n,r}$ *the quantile of order* r *of the law* $\Gamma(n, n)$*. We deduce that* $x_{\lambda_0,n,\alpha} = \gamma_{n,\alpha}/\lambda_0 = g_0\gamma_{n,\alpha}$*, and therefore conclude that the rejection region of the test is*

$$W_n = \{\overline{X}_n \leq g_0\gamma_{n,\alpha}\}.$$

With the values $n = 1000$*,* $\alpha = 0.05$ *and* $g_0 = 3$*, we get* $g_0\gamma_{n,\alpha} = 2.85$*.*

(vi) Since the observed value $\overline{x}_n = 2.8$ *is lower than* 2.85*, we are in the rejection region and* H_0 *is rejected at the level* 5%*.*

As should be clear from this example, the main two technical points of the construction of the test are:

- the verification of the existence of $\theta_0 \in H_0$ such that $\sup_{\theta \in H_0} \mathbb{P}_\theta(W_n) = \mathbb{P}_{\theta_0}(W_n)$;

- the computation of the quantile of order $1-\alpha$ (or α, depending on the form of the rejection region) of the estimator Z_n of $g(\theta)$.

Both steps rely on the use of a free random variable which depends on both Z_n and $g(\theta)$, that is to say on the use of a *pivotal function* in the very same sense as in the construction of confidence intervals in Lecture 4.

There are however situations where there is no such pivotal function, so it may be impossible to compute the quantiles of Z_n under $\mathbb{P}_{\theta_0}$. In this case, one may resort to the construction of tests with only *asymptotic* level α, by leveraging asymptotic properties of Z_n, such as consistency and asymptotic normality.

Exercise 7.1.13. Using the Central Limit Theorem for $\overline{X}_n$ in place of the free random variable S_n in Example 7.1.12, construct a test with asymptotic level α.

Two-sided tests

We now address the determination of the thresholds a_n and b_n for two-sided tests. As for one-sided tests, we assume that there exists a $\theta_0 \in H_0$ such that, whatever the values of a_n and b_n,

$$\sup_{\theta\in H_0} \mathbb{P}_\theta(Z_n \notin (g_0 - a_n, g_0 + b_n)) = \mathbb{P}_{\theta_0}(Z_n \notin (g_0 - a_n, g_0 + b_n)).$$

Then it is clear that the type I risk is bounded from above by α as soon as a_n and b_n are chosen such that there exists $r \in [0, \alpha]$ such that

$$g_0 - a_n \leq z_{\theta_0,n,r} \qquad \text{and} \qquad z_{\theta_0,n,1-\alpha+r} \leq g_0 + b_n.$$

Moreover, for any $\theta \in H_1$, the type II risk is necessarily reduced if the inequalities above are turned into equalities, but then the value of r which minimises the type II risk may, in general, depend on θ (which was not the case in the construction of one-sided tests). It is then customary to choose $r = \alpha/2$, so that the final rejection region writes

$$W_n = \{Z_n \leq z_{\theta_0,n,\alpha/2} \quad \text{or} \quad Z_n \geq z_{\theta_0,n,1-\alpha/2}\}.$$

Remark 7.1.14 (Symmetric distributions)**.** *It is sometimes the case that the law of $Z_n - g_0$ is symmetric under $\mathbb{P}_{\theta_0}$, so*

$$z_{\theta_0,n,\alpha/2} = g_0 - z_{n,1-\alpha/2}, \qquad z_{\theta_0,n,1-\alpha/2} = g_0 + z_{n,1-\alpha/2},$$

with $z_{n,1-\alpha/2}$ the quantile of order $1-\alpha/2$ of the law of $Z_n - g_0$ under $\mathbb{P}_{\theta_0}$, and we may rewrite

$$W_n = \{|Z_n - g_0| \geq z_{n,1-\alpha/2}\}.$$

The next example provides an illustration of such a symmetric situation.

Example 7.1.15 (The Gaussian model)**.** *We assume that we observe $X_1, \dots, X_n$ iid random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, with σ^2 known. For a given value μ_0, we want to construct a test for the hypotheses*

$$H_0 = \{\mu = \mu_0\}, \qquad H_1 = \{\mu \neq \mu_0\}.$$

We take $\overline{X}_n$ as an estimator of μ, and therefore our rejection region writes

$$W_n = \{\overline{X}_n \notin (\mu_0 - a_n, \mu_0 + b_n)\}.$$

Since the null hypothesis is simple, it is a trivial statement that

$$\sup_{\mu\in H_0} \mathbb{P}_\mu(W_n) = \mathbb{P}_{\mu_0}(W_n).$$

Moreover, under $\mathbb{P}_{\mu_0}$, $\overline{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n)$, so its quantile of order r is $\mu_0 + \frac{\sigma}{\sqrt{n}}\phi_r$. Thus, the rejection region writes

$$W_n = \left\{\overline{X}_n \notin \left(\mu_0 + \frac{\sigma}{\sqrt{n}}\phi_{\alpha/2}, \mu_0 + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}\right)\right\} = \left\{\left|\overline{X}_n - \mu_0\right| \geq \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}\right\}.$$

Consistency

For both one- and two-sided tests constructed following the procedure above, consistency follows from the consistency of the estimator Z_n.

Proposition 7.1.16 (Consistency)**.** *Let W_n be the rejection region of either a one-sided test*

$$W_n = \{Z_n \geq z_{\theta_0,n,1-\alpha}\} \qquad \text{or} \qquad W_n = \{Z_n \leq z_{\theta_0,n,\alpha}\},$$

or a two-sided test

$$W_n = \{Z_n \leq z_{\theta_0,n,\alpha/2} \quad \text{or} \quad Z_n \geq z_{\theta_0,n,1-\alpha/2}\}.$$

If Z_n is a consistent estimator of $g(\theta)$, then the test is consistent.

Proof. We address the case of a one-sided test with $H_1 = \{g(\theta) > g_0\}$ and rejection region $W_n = \{Z_n \geq z_{\theta_0,n,1-\alpha}\}$, but the arguments are the same for the other two cases.

We first describe the limit of the quantile $z_{\theta_0,n,1-\alpha}$: since Z_n is consistent, we know that under $\mathbb{P}_{\theta_0}$, Z_n converges in probability to $g(\theta_0)$, thus by Lemma 1.4.8, $z_{\theta_0,n,1-\alpha}$ converges to $g(\theta_0)$.

We now fix $\theta \in H_1$. Since Z_n is consistent, it converges in probability to $g(\theta)$ under $\mathbb{P}_\theta$, which implies that $Z_n - z_{\theta_0,n,1-\alpha}$ converges in probability to

$$g(\theta) - g(\theta_0) = \underbrace{(g(\theta) - g_0)}_{> 0 \text{ since } \theta \in H_1} + \underbrace{(g_0 - g(\theta_0)}_{\geq 0 \text{ since } \theta_0 \in H_0} > 0.$$

Defining $\epsilon = (g(\theta) - g(\theta_0))/2$ and noting that

$$\begin{aligned}\mathbb{P}_\theta(W_n) &= \mathbb{P}_\theta(Z_n \geq z_{\theta_0,n,1-\alpha}) \\ &\geq \mathbb{P}_\theta(|(Z_n - z_{\theta_0,n,1-\alpha}) - (g(\theta) - g(\theta_0))| < \epsilon) \\ &= 1 - \mathbb{P}_\theta(|(Z_n - z_{\theta_0,n,1-\alpha}) - (g(\theta) - g(\theta_0))| \geq \epsilon),\end{aligned}$$

we conclude that

$$\lim_{n\to+\infty} \mathbb{P}_\theta(W_n) = 1,$$

which shows that the test is consistent. □

7.1.4 p-value

In the Lady Tasting Tea experiment described in the introduction of this Lecture, the probability that under the null hypothesis, Bristol got all answers correct was called the p-value. We first give a precise definition of this quantity in the particular case where there exist a statistic $\zeta_n(\mathbf{X}_n)$ and $\theta_0 \in H_0$ such that $W_n = \{\zeta_n(\mathbf{X}_n) \geq \zeta_{\theta_0,n,1-\alpha}\}$, where $\zeta_{\theta_0,n,1-\alpha}$ is the quantile of order $1-\alpha$ of $\zeta_n(\mathbf{X}_n)$ under $\mathbb{P}_{\theta_0}$. This situation covers the following cases:

- one-sided tests constructed in Subsection 7.1.3, with $\zeta_n(\mathbf{X}_n) = Z_n$ if $H_1 = \{g(\theta) > g_0\}$ and $\zeta_n(\mathbf{X}_n) = -Z_n$ if $H_1 = \{g(\theta) < g_0\}$;
- two-sided tests constructed in Subsection 7.1.3 in the specific symmetric case described in Remark 7.1.14, with $\zeta_n(\mathbf{X}_n) = |Z_n - g_0|$.

Definition 7.1.17 (p-value)**.** *In the setting described above, for all $\mathbf{x}_n \in E^n$, the p-value of an observation $\mathbf{x}_n$ is*

$$p\text{-value} = \mathbb{P}_{\theta_0}(\zeta_n(\mathbf{X}_n) \geq \zeta_n(\mathbf{x}_n)).$$

The p-value must be understood as the probability, under H_0, that the test statistic takes values more unfavourable for the acceptance of H_0 than the observed value in the data. Consider for instance the (two-sided) test for the Bernoulli model of Example 7.1.3: under H_0, the empirical mean $\overline{X}_n$ should take values concentrated around $1/2$. However, due to the randomness of the sample, it may happen that $\overline{X}_n$ takes values which are far from $1/2$. For a given value $\overline{x}_n$ of the test statistic, the p-value assesses how likely it is that this value is due to random fluctuations of the sample: the smaller the p-value, the more unlikely the realisation $\overline{x}_n$ under H_0, and the more the experimenter is encouraged to reject H_0. This remark is clarified by the following fundamental property of the p-value.

Proposition 7.1.18 (p-value and level)**.** *In the setting of Definition 7.1.17, we have*

$$H_0 \text{ is rejected} \qquad \text{if and only if} \qquad p\text{-value} \leq \alpha.$$

Proof. For a given observation $\mathbf{x}_n$, H_0 is rejected if and only if $\zeta_n(\mathbf{x}_n) \geq \zeta_{\theta_0,n,1-\alpha}$, which by the definition of $\zeta_{\theta_0,n,1-\alpha}$ is equivalent to $\mathbb{P}_{\theta_0}(\zeta_n(\mathbf{X}_n) \geq \zeta_n(\mathbf{x}_n)) \leq \alpha$, and the left-hand side of the latter inequality is precisely the p-value. □

As a consequence, the p-value indicates *all levels at which H_0 will be rejected*. It is therefore a highly informative and readily interpretable quantity: **the p-value should always be presented with the result of a test**.

For instance, in the Exponential model studied in Example 7.1.12, the p-value associated with the observation $\overline{x}_n = 2.8$ is

$$\mathbb{P}_{\lambda_0}(\overline{X}_n \leq 2.8) = 0.016,$$

which confirms the rejection of H_0 at the level 5% but also shows that at the level 1%, H_0 would not have been rejected.

Example 7.1.19. *In the Bernoulli model considered in Example 7.1.3, we give the p-values of the observation $\overline{x}_n = 0.6$ for different values of n in Table 7.1. For small values of n, the approximation $\overline{X}_n \simeq 1/2 + G/(2\sqrt{n})$ under H_0 is not valid and we rather used exact computations with the Binomial distribution.*

n	1	10	100	1000
p-value	1	0.75	0.046	$2.5\ 10^{-10}$

Table 7.1: p-values of the observation $\overline{x}_n = 0.6$ in the Bernoulli model, for various values of n. At the level $\alpha = 5\%$, H_0 is rejected for $n = 100$, $n = 1000$ but not for $n = 1$, $n = 10$.

Exercise 7.1.20. In the Lady Tasting Tea experiment, what would have been the p-value of the experiment outcome if Bristol had made one single mistake? At the level 5%, what would then have been Fisher's conclusion?

In the more general setting where you have two hypotheses H_0, H_1 and a family of rejection regions $(W_n^\alpha)_\alpha$ such that each W_n^α has level α, Proposition 7.1.18 allows to extend the definition of the p-value of an observation as the minimal level at which H_0 is rejected:

$$p\text{-value} = \inf\{\alpha : H_0 \text{ is rejected at level } \alpha\}.$$

For example, in the case of general two-sided tests constructed in Subsection 7.1.3, with rejection region

$$W_n^\alpha = \{Z_n \notin (z_{\theta_0,n,\alpha/2}, z_{\theta_0,n,1-\alpha/2})\},$$

the p-value of an observed value z_n of the estimator Z_n is $\inf\{\alpha : z_n \notin (z_{\theta_0,n,\alpha/2}, z_{\theta_0,n,1-\alpha/2})\}$.

Remark 7.1.21 (Monte Carlo computation of a p-value)**.** *Consider a test with rejection region of the form* $\{\zeta_n(\mathbf{X}_n) \geq \zeta_{\theta_0,n,1-\alpha}\}$*, so by Definition 7.1.17, the p-value of an observation* $\mathbf{x}_n$ *is* $\mathbb{P}_{\theta_0}(\zeta_n(\mathbf{X}_n) \geq \zeta_n(\mathbf{x}_n))$*. To compute this p-value, it is necessary to compute the value of the cumulative distribution function of* $\zeta_n(\mathbf{X}_n)$ *under* $\mathbb{P}_{\theta_0}$ *at the point* $\zeta_n(\mathbf{x}_n)$ *– as we did above for the Example 7.1.12 of the Exponential model. This is not always possible analytically. If the law of* $\zeta_n(\mathbf{X}_n)$ *under* $\mathbb{P}_{\theta_0}$ *has a classical distribution, then statistical tables or the function* `distrib.cdf()` *in* `scipy.stats` *(see Lecture 1) can be employed. Otherwise, the* Monte Carlo method *can be employed, assuming that one is able to sample independent realisations of* $\mathbf{X}_n$ *under* $\mathbb{P}_{\theta_0}$*. It works as follows:*

- *fix a large number* M *of copies;*
- *draw* M *independent realisations* $\mathbf{X}_n^{(1)}, \ldots, \mathbf{X}_n^{(M)}$ *of* $\mathbf{X}_n$ *under* $\mathbb{P}_{\theta_0}$ *(which therefore requires to simulate* M *times a sample of size* n*);*
- *estimate the p-value by the empirical frequency*

$$\widehat{p}_M = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{\zeta_n(\mathbf{X}_n^{(m)}) \geq \zeta_n(\mathbf{x}_n)\}}.$$

An example of such an approach will be seen in Lecture 11.

7.1.5 Duality between tests and confidence regions

Consider null and alternative hypotheses of the form

$$H_0 = \{g(\theta) = g_0\}, \qquad H_1 = \{g(\theta) \neq g_0\},$$

with $g : \Theta \to \mathbb{R}^d$, $g_0 \in g(\Theta)$. Assume in addition that a confidence region C_n (not necessarily an interval) with level $1 - \alpha$ is available for $g(\theta)$. Then, by definition, for any $\theta \in H_0$,

$$\mathbb{P}_\theta(g_0 \notin C_n) = \mathbb{P}_\theta(g(\theta) \notin C_n) = \alpha,$$

so that the test with rejection region $W_n = \{g_0 \notin C_n\}$ has level α. This computation readily extends to the case of asymptotic confidence regions and yields a test with asymptotic level α, which is then consistent if C_n has the property that for any $\theta \in \Theta$, for any $z \neq g(\theta)$, $\mathbb{P}_\theta(z \notin C_n)$ converges to 1.

Exercise 7.1.22 (Test from a confidence interval for the Bernoulli model)**.** We consider the Bernoulli model $\{\mathcal{B}(p), p \in [0,1]\}$ with null and alternative hypotheses

$$H_0 = \{p = p_0\}, \qquad H_1 = \{p \neq p_0\},$$

for some $p_0 \in [0,1]$.

1. Recall the rejection region W_n of a consistent test with asymptotic level α following the steps described in Subsection 7.1.3.
2. Recall the expression of an asymptotic confidence interval I_n with level $1 - \alpha$ for p, and deduce the rejection region W_n' of a second consistent test with asymptotic level α.
3. If $p_0 = 1/2$, show that the test with rejection region W_n' *is always*[7] *more powerful* than the test with rejection region W_n.
4. What do you think of this assertion when $p_0 \neq 1/2$?

[7] that is to say, for any value of $p \in H_1$

Remark 7.1.23. *The test constructed in Exercise 7.1.22 is an example of a* Wald test*, which is to hypothesis tests what Proposition 4.3.1 is to confidence intervals. The general principle of this test is presented in Lecture 9.*

Conversely, there are situations where one is naturally provided with a rejection region $W_n(g_0)$ for any $g_0 \in g(\Theta)$. Then, a confidence region for $g(\theta)$ may be defined by letting

$$C_n := \{g_0 \in g(\Theta) : \mathbf{X}_n \notin W_n(g_0)\}.$$

This approach allows to consider cases in which $g(\theta)$ is no longer real-valued, but vector-valued or even infinite-dimensional. For example, it is typically useful in the nonparametric setting which will be studied in Lecture 11, where $X_1, \dots, X_n$ are iid under some probability measure P on which no parametric assumption is made. Then, given a probability measure P_0, *goodness-of-fit tests* are available for the hypotheses

$$H_0 = \{P = P_0\}, \qquad H_1 = \{P \neq P_0\},$$

and can therefore be employed to construct a confidence region for P in the set of probability measures on E.

7.2 Multiple comparisons

7.2.1 The look-elsewhere effect

Assume that you build a test with rejection region W_n and level α for some hypotheses H_0 and H_1, and for simplicity assume that $H_0 = \{\theta = \theta_0\}$ is simple. If you observe data $\mathbf{X}_n$ which are distributed under H_0, then by construction, the probability of a false positive (that is to say, to reject H_0) is α. Now if you have m independent samples $\mathbf{X}_{1,n}, \dots, \mathbf{X}_{m,n}$, to each of which you apply your test, under H_0 the probability to return *at least one* false positive becomes

$$\mathbb{P}_{\theta_0}\left(\cup_{k=1}^m \{\mathbf{X}_{k,n} \in W_n\}\right) = 1 - \prod_{k=1}^m \mathbb{P}_{\theta_0}(\mathbf{X}_{k,n} \notin W_n) = 1 - (1-\alpha)^m.$$

When m grows, this probability goes to 1. So even if you data are distributed under H_0, which means that the effect that your experiment is trying to show is not present, if you repeat your experiment sufficiently often, you will finally conclude (wrongly!) to the *presence* of this effect – because, at each experiment, you have a small probability α to get a false positive. This is the *look-elsewhere effect*, which is also illustrated on Figure 7.2.

7.2.2 Family-wise error rate and the Bonferroni method

The framework of multiple comparisons can be formalised by considering, for a given choice of null and alternative hypotheses H_0 and H_1, a family of m rejection regions $W_n^1, \dots, W_n^m$, such that the test with rejection region W_n^k has level α_k. The sample set is $\mathbf{X}_n = (X_1, \dots, X_n)$, but each realisation takes its values in a possibly high-dimensional state space E. In the example of the previous section, the events $\{\mathbf{X}_n \in W_n^k\}$, $k \in \{1, \dots, m\}$ were assumed to be independent, in which case we shall call the tests *independent*, but this assumption may be relaxed.

For multiple comparisons, the type I risk can be measured by the *Family-Wise Error Rate*.

Definition 7.2.1 (Family-Wise Error Rate)**.** *The* Family-Wise Error Rate *(FWER) of the family of tests $W_n^1, \dots, W_n^m$ is defined by*

$$\mathrm{FWER} = \sup_{\theta \in H_0} \mathbb{P}_\theta\left(\exists k \in \{1, \dots, m\} : \mathbf{X}_n \in W_n^k\right).$$

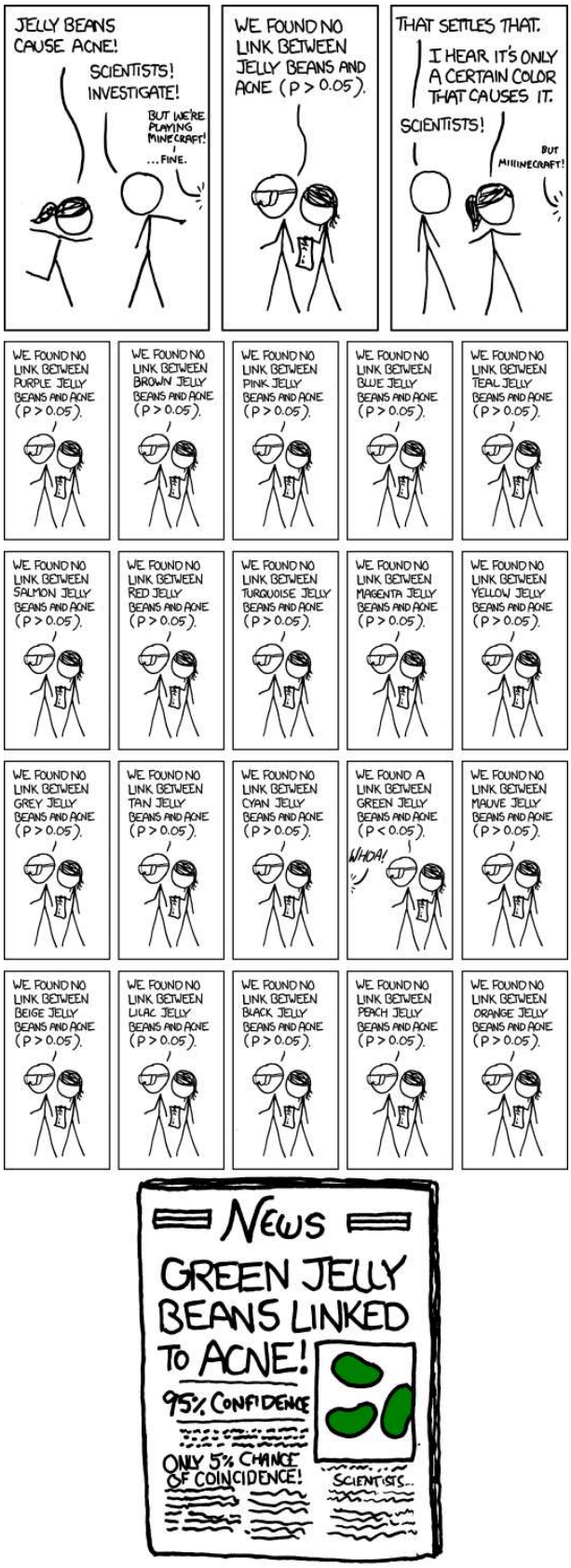


Figure 7.2: An illustration of the look-elsewhere effect. Taken from XKCD by Randall Munroe: http://www.xkcd.com/882.

Exercise 7.2.2. If the tests are independent and H_0 is simple, compute FWER in terms of the individual levels $\alpha_1, \dots, \alpha_m$.

The *Bonferroni method* consists in rejecting H_0 at the level $\alpha \in (0,1)$ as soon as at least one of the p-values $p_1, \dots, p_m$ is lower than α/m. It is based on the next result, which does not require the tests to be independent.

Lemma 7.2.3 (Bonferroni correction)**.** *We have*

$$\mathrm{FWER} \leq \sum_{k=1}^{m} \alpha_k.$$

As a consequence, if one takes $\alpha_k = \alpha/m$ for all $k \in \{1, \dots, m\}$, then FWER *is lower than α.*

Proof. Using the *union bound* $\mathbb{P}(\cup_{k=1}^n A_k) \leq \sum_{k=1}^n \mathbb{P}(A_k)$, we have, for all $\theta \in H_0$,

$$\mathbb{P}_\theta\left(\exists k \in \{1, \dots, m\} : \mathbf{X}_n \in W_n^k\right) \leq \sum_{k=1}^{m} \mathbb{P}_\theta\left(\mathbf{X}_n \in W_n^k\right) \leq \sum_{k=1}^{m} \alpha_k,$$

which leads to the expected bound by taking the supremum of the left-hand side over $\theta \in H_0$. □

The union bound employed in the proof of Lemma 7.2.3 can be very rough, and as a consequence the Bonferroni method has the counterpart of generally increasing the type II risk severely. A more modern approach, which enables to control the *False Discovery Rate*, is presented in Exercise 7.A.3.

7.A Exercises

Training exercises

Exercise 7.A.1 (Mortality rate for Covid-19)**.** In an article available on the Internet (published in 2020), the following sentence can be found:

The mortality rate of Covid-19 is far below 0.5%. There were 1046 infected sailors on the aircraft carrier Charles de Gaulle, and none of them died.

In order to quantify the statistical foundation of this statement, we set $n = 1046$ and, for $i \in \{1, \dots, n\}$, we denote by X_i the random variable equal to 1 if the i-th infected sailor died and 0 otherwise. We assume that $X_1, \dots, X_n$ are independent Bernoulli variables with parameter $p \in [0,1]$ and we let $p_0 = 0.005$. We finally introduce the hypotheses

$$H_0 = \{p \geq p_0\}, \qquad H_1 = \{p < p_0\}.$$

1. Recall (without proof) the expression of the maximum likelihood estimator $\widehat{p}_n$ of p in this model.
2. Let $\alpha \in (0,1)$. Construct a consistent test with asymptotic level α for the hypotheses H_0 and H_1.
3. (a) Give an explicit expression of the p-value associated with the observation 'none of the n infected sailors died' as a function of p_0 and n.
   (b) Compute this p-value. What do you conclude?
4. Which critical comment can you make on the choice of the sample with respect to the hypotheses to be tested?

Homework

Exercise 7.A.2 (Power and sample size)**.** The probability for an individual to be infected by a virus is denoted by p_0, and assumed to be known. A new vaccine is tested on a sample of n individuals. We denote by $p \in [0, p_0]$ the probability to be infected after the vaccine (the probability to be infected cannot be increased by the vaccine). For all $i \in \{1, \dots, n\}$, we define $X_i = 1$ if the i-th individual is infected by the virus after the vaccine and $X_i = 0$ otherwise, so that $X_i \sim \mathcal{B}(p)$. We introduce the hypotheses

$$H_0 = \{p = p_0\}, \qquad H_1 = \{p < p_0\}.$$

It is recommended to use Python to perform the numerical computations. Moreover, throughout the exercise, you may use the Gaussian approximation of $\overline{X}_n$ given by the Central Limit Theorem.

1. Following the steps described in Subsection 7.1.3, construct a consistent test with asymptotic level α for this model.

2. For $p_0 = 18\%$, an experiment carried over $n = 100$ individuals yields $\overline{x}_n = 16\%$. At the level $\alpha = 5\%$, what is the conclusion of the experiment regarding the efficiency of the vaccine?

3. It turns out that $p = 15\%$, so that the vaccine actually reduces the number of infected individuals by $1/6$. What is the probability that an experiment carried over $n = 100$ individuals succeed in detecting the efficiency of the vaccine at the level $\alpha = 5\%$? What do you think of this result?

4. What should be the minimum size of the sample in order to detect, with probability at least 80%, a diminution of the number of infected people by $1/6$? By $1/3$?

Supplementary exercises

Exercise 7.A.3 (False Discovery Rate and the Benjamini–Hochberg procedure)**.** The use of the *False Discovery Rate*, defined below, is an alternative approach to the control of the FWER for multiple comparisons. It is concerned with the expected number of false positives rather than the probability of returning at least one false positive, and therefore provides methods which have a larger type I risk than FWER-based procedures, but in turn have a better statistical power.

We extend the framework of Section 7.2 by assuming that the m tests may have different null and alternative hypotheses, respectively denoted by H_0^k, H_1^k, $k \in \{1, \dots, m\}$. We introduce the notation of Table 7.2, where for instance V is the number of tests for which the null hypothesis is rejected while true, that is to say V is the number of false positives.

	H_0 is true	H_0 is false	Total
H_0 is rejected	V	S	R
H_0 is not rejected	U	T	$m - R$
Total	m_0	$m - m_0$	m

Table 7.2: Notations for multiple comparisons. V, S, R, U, T are random variables, among which only R is a statistic. m_0 is not random but depends on θ.

Notice that, in Table 7.2, R is a statistic because it is directly observed from the data. On the contrary, in order to know V, S, U and T, one has to know which of the hypotheses are actually true, and this information depends on the value of θ.

With this notation, the *False Discovery Rate* (FDR) is

$$\mathrm{FDR}(\theta) = \mathbb{E}_\theta[V/R],$$

with the convention that $V/R = 0$ when $V = R = 0$. The *Benjamini–Hochberg procedure*[8] allows to control the FDR. It was published in 1995 and has become a standard in the analysis of large data sets. It works as follows:

(i) Perform the m tests and let $p_1, \dots, p_m$ be the associated p-values.

(ii) Denote by $p_{(1)} \leq \cdots \leq p_{(m)}$ the increasing reordering of the p-values.

(iii) For $\alpha \in (0,1)$, define
$$J = \max\left\{ j \in \{1, \dots, m\} : p_{(j)} \leq \frac{j}{m}\alpha \right\}.$$

(iv) Reject H_0^k for all k such that $p_k \leq p_{(J)}$.

1. Assuming that the p-values $p_1, \dots, p_m$ are pairwise distinct, show that with this procedure, the number of discoveries is $R = J$.

The purpose of the sequel of the exercise is to show that, if the tests are independent, then with the Benjamini–Hochberg procedure,
$$\forall \theta \in \Theta, \qquad \mathrm{FDR}(\theta) = \frac{m_0}{m}\alpha \leq \alpha.$$
To proceed, we let $\theta \in \Theta$, $\mathcal{I}$ be the set of indices k for which $\theta \in H_0^k$, and for all $k \in \{1, \dots, m\}$,
$$V_k = \mathbb{1}_{\{H_0^k \text{ is rejected}\}} = \mathbb{1}_{\{p_k \leq p_{(J)}\}}.$$

2. Show that $\mathrm{FDR}(\theta) = \sum_{k \in \mathcal{I}} \sum_{j=1}^{m} \frac{1}{j} \mathbb{E}_\theta[V_k \mathbb{1}_{\{J=j\}}]$.

3. For all $k \in \mathcal{I}$, let J_k denote the number of discoveries of the procedure if the p-value of the k-th test is replaced with the value 0. Show that $\mathrm{FDR}(\theta) = \sum_{k \in \mathcal{I}} \sum_{j=1}^{m} \frac{1}{j} \mathbb{E}_\theta[\mathbb{1}_{\{p_k \leq \alpha j/m\}} \mathbb{1}_{\{J_k = j\}}]$.

4. Show that $\mathbb{E}_\theta[\mathbb{1}_{\{p_k \leq \alpha j/m\}} \mathbb{1}_{\{J_k=j\}}] = \frac{\alpha j}{m} \mathbb{E}_\theta[\mathbb{1}_{\{J_k=j\}}]$, and conclude.

[8] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995.

Lecture 8

Tests in the Gaussian Model

Contents

In this Lecture, we focus on samples which are assumed to be Gaussian and construct (nonasymptotic) tests in various contexts: one-sample tests on μ and σ^2, but also tests in the case where several samples are given, or tests in the context of linear regression.

Before entering into the details of the constructions of these tests, let us mention that of course, since these tests assume a Gaussian distribution for samples, one should check this assumption before applying these tests. Deciding whether a sample is distributed according to a Gaussian distribution is the object of nonparametric *goodness-of-fit* tests, which will be seen in Lecture 11.

8.1 One-sample tests

In this Section, we assume that we observe a sample $\mathbf{X}_n = (X_1, \dots, X_n)$ which is iid according to $\mathcal{N}(\mu, \sigma^2)$.

8.1.1 Two-sided tests for the mean

We first fix $\mu_0 \in \mathbb{R}$ and consider the hypotheses

$$H_0 = \{\mu = \mu_0\}, \qquad H_1 = \{\mu \neq \mu_0\}.$$

Just like for the construction of confidence intervals in Lecture 4, we first assume that σ^2 **is known**. Then, following the steps described in Lecture 7 (see Example 7.1.15 in particular), we obtain that the test with rejection region

$$W_n = \left\{ |\overline{X}_n - \mu_0| \geq \sqrt{\frac{\sigma^2}{n}} \phi_{1-\alpha/2} \right\}$$

is consistent and has level α. Let us slightly modify the expression of W_n and rewrite

$$W_n = \left\{ |Z_n| \geq \phi_{1-\alpha/2} \right\},$$

where the statistic

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}}$$

is called the *Z-statistic*, or the *Z-score*. It has the property that, under H_0, it has distribution $\mathcal{N}(0,1)$ – in particular, it is *free*.

This test is called a *two-sided Z-test*. The p-value of an observation whose Z-score is z_n writes

$$\mathbb{P}(|Z_n| \geq |z_n|) = 2\left(1 - \Phi(|z_n|)\right),$$

where $Z_n \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z_n.

Remark 8.1.1 (The '5σ-rule'). *Levels of* 1%, 5% *or* 10% *are considered standard in many fields of applications, such as biology, medicine, and social sciences. In particle physics, the standard rule, called the* 5σ-rule*, is much more conservative: the null hypothesis, namely the nonexistence of the sought particle, is usually rejected if the observed value of the test statistic is larger than* 5 *times its standard deviation. For a two-sided Z-test, this rule leads to reject* H_0 *if the p-value is lower than* $2(1-\Phi(5)) \simeq 6\ 10^{-7}$.

If we now assume that σ^2 **is unknown**, we know from Proposition 3.2.4 that, under H_0, the statistic

$$T_n = \frac{\overline{X}_n - \mu_0}{\sqrt{S_n^2/n}}$$

has distribution $\mathrm{t}(n-1)$. This quantity is called the t-*statistic*, or the t-*score*. We then deduce that the test with rejection region

$$W_n = \left\{|T_n| \geq t_{n-1,1-\alpha/2}\right\},$$

where $t_{n,r}$ is the quantile of order r of the $\mathrm{t}(n)$ distribution, has level α. This test is called a *two-sided* t-*test*.

Exercise 8.1.2 (Consistency).
1. Show that, for any $r \in (0,1)$, $\lim_{n\to+\infty} t_{n,r} = \phi_r$. *Hint: recall Lemma 1.4.8 from Lecture 1, and Exercise 3.A.3 from Lecture 3.*
2. Deduce that the t-test above is consistent.

The p-value of an observation whose t-score is t_n writes

$$\mathbb{P}(|T_n| \geq |t_n|) = 2\left(1 - F_{n-1}(|t_n|)\right),$$

where $T_n \sim \mathrm{t}(n-1)$ and F_{n-1} is the CDF of T_n.

8.1.2 One-sided tests for the mean

We now consider the hypotheses

$$H_0 = \{\mu \leq \mu_0\}, \qquad H_1 = \{\mu > \mu_0\},$$

where μ_0 is still fixed in $\mathbb{R}$. If σ^2 **is known**, the test with rejection region

$$W_n = \{Z_n \geq \phi_{1-\alpha}\}, \qquad Z_n = \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}},$$

is consistent and has level α. It is called a *one-sided Z-test*, and the p-value of an observation whose Z-score is z_n writes

$$\mathbb{P}(Z_n \geq z_n) = 1 - \Phi(z_n),$$

where $Z_n \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z_n. If σ^2 **is unknown**, the test with rejection region

$$W_n = \{T_n \geq t_{n-1,1-\alpha}\}, \qquad T_n = \frac{\overline{X}_n - \mu_0}{\sqrt{S_n^2/n}},$$

is consistent and has level α. It is called a *one-sided* t-*test*, and the p-value of an observation whose t-score is t_n writes

$$\mathbb{P}(T_n \geq t_n) = 1 - F_{n-1}(t_n),$$

where $T_n \sim \mathrm{t}(n-1)$ and F_{n-1} is the CDF of T_n.

Exercise 8.1.3. Write the rejection region and the p-value of one-sided Z- and t-tests for the hypotheses $H_0 = \{\mu \geq \mu_0\}$, $H_1 = \{\mu < \mu_0\}$.

Exercise 8.1.4 (Ozone concentration)**.** A network of n sensors measures the concentration of ozone in various places of Paris. Due to local variability and measure noise, on a given day, the ozone concentration recorded by the i-th sensor is representation by a random variable $X_i = \mu + \varepsilon_i$, where μ is the mean ozone concentration on this day and the variables $(\varepsilon_i)_{1 \leq i \leq n}$ are independent $\mathcal{N}(0, \sigma^2)$ variables. In the block below, we give the measurements (in $\mu\text{g}/\text{m}^3$) corresponding to 3 different days.

Day 1	24.5, 48.0, 32.0, 45.8, 32.1, 35.7, 41.1, 30.4, 30.8, 27.5
Day 2	48.0, 57.3, 56.1, 43.2, 37.9, 39.9, 40.3, 28.7, 45.7, 57.2
Day 3	56.2, 54.1, 53.4, 46.2, 53.5, 48.0, 57.5, 65.2, 49.3, 58.5

The maximal value of mean ozone concentration fixed by WHO is $\mu_0 = 40\ \mu\text{g}/\text{m}^3$. Our goal is to use the sample $(X_1, \ldots, X_n)$ to determine whether μ is above μ_0.

1. Assuming that σ is known and equal to $10\ \mu\text{g}/\text{m}^3$, apply a Z-test to each of the three samples.
2. Without assuming that σ is known, apply a t-test to each of the three samples. You may implement it directly or use the `scipy.stats` function `ttest_1samp()`.

8.1.3 Tests for the variance

Exercise 8.1.5 (Test for the variance)**.** For $\sigma_0^2 > 0$, we consider the hypotheses

$$H_0 = \{\sigma^2 \leq \sigma_0^2\}, \qquad H_1 = \{\sigma^2 > \sigma_0^2\},$$

without assuming that the mean is known.

1. Construct a test with level α.
2. Show that this test is consistent. *Hint: use again Lemma 1.4.8 from Lecture 1, and Exercise 3.A.3 from Lecture 3.*

8.2 Two-sample tests

In this Section, we are interested in the problem of testing the *homogeneity* of two populations: one observes independent samples $\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1})$ and $\mathbf{X}_{2,n_2} = (X_{2,1}, \ldots, X_{2,n_2})$, with respective laws P_1 and P_2, and wants to known whether $P_1 = P_2$ or not. Applications of such homogeneity tests are ubiquitous and we shall encounter several of them in the sequel of the course.

Here, we assume that the two samples $\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1})$ and $\mathbf{X}_{2,n_2} = (X_{2,1}, \ldots, X_{2,n_2})$ are Gaussian:

$$\forall i_1 \in \{1, \ldots, n_1\}, \quad X_{1,i_1} \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad \forall i_2 \in \{1, \ldots, n_2\}, \quad X_{2,i_2} \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

We shall study two tests:

- Fisher's test of *homoscedasticity* with null and alternative hypotheses

$$H_0 = \{\sigma_1^2 = \sigma_2^2\}, \qquad H_1 = \{\sigma_1^2 \neq \sigma_2^2\};$$

- Student's test of homogeneity, in which *it is assumed that* $\sigma_1^2 = \sigma_2^2$ (but the value of the variance remains unknown) and either

$$H_0 = \{\mu_1 = \mu_2\}, \qquad H_1 = \{\mu_1 \neq \mu_2\},$$

or

$$H_0 = \{\mu_1 \leq \mu_2\}, \qquad H_1 = \{\mu_1 > \mu_2\}.$$

Example 8.2.1 (Grades of IMI and SEGF students in 2023)**.** *The statistics associated with the grades at the final exam of the course* Statistics and Data Analysis *in 2023 are reported in Table 8.1. Assuming that these samples are Gaussian, we want to known whether there is a statistically significant difference between IMI and SEGF students.*

	IMI	SEGF
Number of students	$n_1 = 54$	$n_2 = 26$
Average	14.074	13.076
Standard deviation ($\sqrt{s_n^2}$)	2.323	3.084

Table 8.1: Statistics of the grades at the final exam of the course *Statistics and Data Analysis* in 2023.

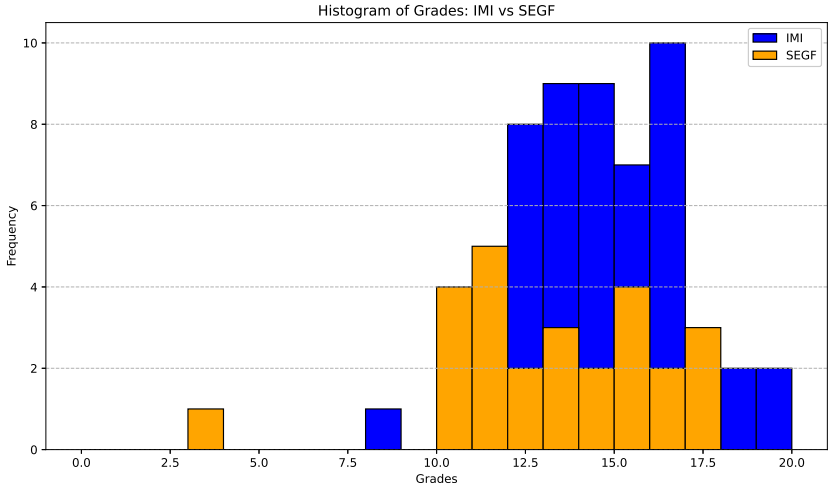


Figure 8.1: Histograms of the grades of IMI and SEGF students.

8.2.1 Student's test of homogeneity

Here, we suppose that it is already known that the two samples have the same variance. The latter is denoted by σ^2, however its value is not assumed to be known. Then we have

$$\forall i_1 \in \{1, \dots, n_1\}, \quad X_{1,i_1} \sim \mathcal{N}(\mu_1, \sigma^2), \qquad \forall i_2 \in \{1, \dots, n_2\}, \quad X_{2,i_2} \sim \mathcal{N}(\mu_2, \sigma^2).$$

We first consider the case where the null and alternative hypotheses write

$$H_0 = \{\mu_1 = \mu_2\}, \qquad H_1 = \{\mu_1 \neq \mu_2\}.$$

Then, by Proposition 3.2.2, under H_0,

$$\overline{X}_{1,n_1} - \overline{X}_{2,n_2} \sim \mathcal{N}(0, \sigma^2/n_1 + \sigma^2/n_2),$$

while

$$\frac{(n_1 - 1)S_{1,n_1}^2 + (n_2 - 1)S_{2,n_2}^2}{\sigma^2} \sim \chi_2(n_1 + n_2 - 2),$$

and these two variables are independent. As a consequence,

$$T_{n_1,n_2} = \frac{\frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}}{\sqrt{\frac{(n_1-1)S_{1,n_1}^2 + (n_2-1)S_{2,n_2}^2}{(n_1+n_2-2)\sigma^2}}} \sim \mathrm{t}(n_1 + n_2 - 2),$$

which shows that the test with rejection region

$$W_{n_1,n_2} = \{|T_{n_1,n_2}| \geq t_{n_1+n_2-2,1-\alpha/2}\}$$

has level α. It is moreover easily seen to be consistent when both n_1 and n_2 go to infinity. This test is called a *two-sample two-sided* t*-test*. The p-value of an observation whose t-score is t_{n_1,n_2} writes

$$\mathbb{P}(|T_{n_1,n_2}| \geq |t_{n_1,n_2}|) = 2\left(1 - F_{n_1+n_2-2}(|t_{n_1,n_2}|)\right),$$

where $T_{n_1,n_2} \sim \mathrm{t}(n_1+n_2-2)$ and $F_{n_1+n_2-2}$ is the CDF of T_{n_1,n_2}. This test can be applied using the `scipy.stats` function `ttest_ind()`, with the parameter `alternative='two-sided'`.

In the case where the null and alternative hypotheses write

$$H_0 = \{\mu_1 \leq \mu_2\}, \qquad H_1 = \{\mu_1 > \mu_2\},$$

similar arguments show that the test with rejection region

$$W_{n_1,n_2} = \{T_{n_1,n_2} \geq t_{n_1+n_2-2,1-\alpha}\}$$

has level α and is consistent when both n_1 and n_2 go to infinity. This test is called a *two-sample one-sided* t*-test*. The p-value of an observation whose t-score is t_{n_1,n_2} writes

$$\mathbb{P}(T_{n_1,n_2} \geq t_{n_1,n_2}) = \left(1 - F_{n_1+n_2-2}(t_{n_1,n_2})\right),$$

where $T_{n_1,n_2} \sim \mathrm{t}(n_1+n_2-2)$ and $F_{n_1+n_2-2}$ is the CDF of T_{n_1,n_2}. This test can be applied using the `scipy.stats` function `ttest_ind()`, with the parameter `alternative='greater'`.

Exercise 8.2.2. With the data of Example 8.2.1, can you affirm that IMI students are better (at doing statistics) than SEGF students?

8.2.2 Fisher's test of homoscedasticity

We no longer assume that the two samples have the same variance, and work with the hypotheses

$$H_0 = \{\sigma_1^2 = \sigma_2^2\}, \qquad H_1 = \{\sigma_1^2 \neq \sigma_2^2\}.$$

Since the variances σ_1^2 and σ_2^2 are estimated by $S_{n_1}^2$ and $S_{n_2}^2$, we should reject H_0 when these two quantities are far from each other. In order to construct a statistic which measures how far they are from each other, and which is free under H_0, we introduce the following distribution.

Definition 8.2.3 (Fisher distribution)**.** *Let $Y_1 \sim \chi_2(n_1)$ and $Y_2 \sim \chi_2(n_2)$ be independent random variables. The law of the random variable*

$$Z = \frac{Y_1/n_1}{Y_2/n_2}$$

is called the Fisher distribution *with parameters n_1 and n_2, and denoted by $\mathrm{F}(n_1,n_2)$.*

By Proposition 3.2.2, under H_0 we have

$$F_{n_1,n_2} = \frac{S_{1,n_1}^2}{S_{2,n_2}^2} \sim \mathrm{F}(n_1-1, n_2-1).$$

Therefore the test with rejection region

$$W_{n_1,n_2} = \left\{F_{n_1,n_2} \notin [f_{n_1-1,n_2-1,\alpha/2}, f_{n_1-1,n_2-1,1-\alpha/2}]\right\},$$

where $f_{n_1-1,n_2-1,r}$ is the quantile of order r of $\mathrm{F}(n_1-1, n_2-2)$, has level α. This test is called a *two-sided* F*-test*.

Exercise 8.2.4. 1. Show that for any $r \in (0,1)$, $\lim_{n_1,n_2\to+\infty} f_{n_1-1,n_2-1,r} = 1$. *Hint: use again Lemma 1.4.8 from Lecture 1, and Exercise 3.A.3 from Lecture 3.*

2. Deduce that the F test above is consistent when both n_1 and n_2 go to infinity.

With the data of Example 8.2.1, the F-statistic takes the value 0.567, while for $\alpha = 0.05$, the interval $[f_{n_1-1,n_2-1,\alpha/2}, f_{n_1-1,n_2-1,1-\alpha/2}]$ is $[0.525, 2.069]$. Therefore the homoscedasticity assumption is not rejected.

8.3 Tests in linear regression

In this Section, we consider the Linear Regression model with Gaussian errors $\boldsymbol{\epsilon}_n \sim \mathcal{N}_n(0, \sigma^2 I_n)$ for some unknown $\sigma^2 > 0$, as in Section 3.3. In this setting, we recall that the OLS $\widehat{\beta}$ has law $\mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})$, and it is independent from the estimator $\widehat{\sigma}^2 = \frac{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2}{n-p-1}$ of σ^2, which is such that $(n-p-1)\widehat{\sigma}^2/\sigma^2 \sim \chi_2(n-p-1)$. We are interested in determining whether a feature x^j has an actual influence over y, which amounts to deciding where $\beta^j = 0$ or not.

8.3.1 Student's test

If a feature x^j has no influence on y, then the corresponding coefficient β_j must be 0. The corresponding test therefore has the null and alternative hypotheses

$$H_0 = \{\beta_j = 0\}, \qquad H_1 = \{\beta_j \neq 0\}.$$

We denote by ρ_j the j-th diagonal coefficient of the matrix $(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}$ (the coefficients being indexed from 0 to p).

Lemma 8.3.1 (Student's test)**.** *The test rejecting H_0 as soon as*

$$\left| \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 \rho_j}} \right| \geq t_{n-p-1,1-\alpha/2},$$

where we write $\widehat{\beta} = (\widehat{\beta}_0, \dots, \widehat{\beta}_p)$, has level α.

Proof. Let $e^j = (0, 0, \dots, 1, \dots 0)$ seen as a row vector of $\mathbb{R}^{p+1}$, where the j-th coefficient is equal to 1 and the coefficients are indexed from 0 to p. By Proposition 3.3.3,

$$\widehat{\beta}_j = e^j \widehat{\beta} \sim \mathcal{N}(\beta_j, \sigma^2 \rho_j),$$

and this variable is independent from $\widehat{\sigma}^2$. As a consequence, under H_0 we have

$$Y = \frac{e^j \widehat{\beta}}{\sqrt{\sigma^2 \rho_j}} \sim \mathcal{N}(0,1), \qquad \text{independent from} \quad Z = \frac{\widehat{\sigma}^2}{\sigma^2}(n-p-1) \sim \chi_2(n-p-1),$$

so that

$$\frac{e^j \widehat{\beta}}{\sqrt{\widehat{\sigma}^2 \rho_j}} = \frac{Y}{\sqrt{Z/(n-p-1)}} \sim \mathrm{t}(n-p-1),$$

from which the construction of the Student test follows. □

8.3.2 Fisher's test

Fisher's test allows to test the joint influence of several features. Up to applying a permutation to the indices of the features, the corresponding null and alternative hypotheses may be written

$$H_0 = \{\beta_{q+1} = \cdots = \beta_p = 0\}, \qquad H_1 = \{\text{there exists } j \geq q+1 \text{ such that } \beta_j \neq 0\},$$

where $q \in \{0, \ldots, p-1\}$. Denoting by $\widehat{\mathbf{y}}_n^0$ the orthogonal projection of $\widehat{\mathbf{y}}_n$ onto the range of the matrix

$$(\mathbf{x}_n)^0 = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^q \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^q \end{pmatrix} \in \mathbb{R}^{n \times (q+1)},$$

we introduce the statistic

$$F = \frac{\|\widehat{\mathbf{y}}_n - \widehat{\mathbf{y}}_n^0\|^2/(p-q)}{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2/(n-p-1)}.$$

Lemma 8.3.2 (Fisher's test)**.** *The test rejecting H_0 as soon as*

$$F \geq f_{p-q,n-p-1,1-\alpha},$$

where $f_{p-q,n-p-1,1-\alpha}$ is the quantile of order $1-\alpha$ of the Fisher distribution $\mathrm{F}(p-q, n-p-1)$, has level α.

As for Student's test above, the proof of Lemma 8.3.2 reduces to showing that under H_0, $F \sim \mathrm{F}(p-q, n-p-1)$. It is left as an exercise.

Remark 8.3.3. *Fisher's test plays the same role for Linear Regression as the Likelihood Ratio Test for Logistic Regression introduced in Subsection 9.2.3.*

8.4 Analysis of variance

The technique of *Analysis of Variance* (ANOVA) allows to test the homogeneity of $k \geq 3$ samples, and thereby generalises Student's test of Section 8.2. Let us be given k samples

$$\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1}), \ldots, \mathbf{X}_{k,n_k} = (X_{k,1}, \ldots, X_{k,n_k}),$$

which are assumed to be Gaussian and have the same (unknown) variance:

$$\forall \ell \in \{1, \ldots, k\}, \quad \forall i \in \{1, \ldots, n_\ell\}, \qquad X_{\ell,i} \sim \mathcal{N}(\mu_\ell, \sigma^2).$$

The purpose of ANOVA is to construct a test for the null and alternative hypotheses

$$H_0 = \{\mu_1 = \cdots = \mu_k\}, \qquad H_1 = \{\exists \ell, m : \mu_\ell \neq \mu_m\}.$$

We will also need the notation

$$n = \sum_{\ell=1}^{k} n_\ell, \qquad \mathbf{X}_n = (\mathbf{X}_{1,n_1}, \ldots, \mathbf{X}_{k,n_k}) \in \mathbb{R}^n.$$

Example 8.4.1 (Fertiliser for lettuce plants)**.** *Consider an experiment aiming at determining if the use of fertiliser has an influence on the number of leaves of lettuce plants: the higher this number, the better it is for consumers*[1]*. Lettuces are grown in $k = 4$ groups, in the same greenhouse, with different soil types:*

[1] This example is taken from L. Clement and J. Gilis' tutorial for the course Practical Statistics for the Life Sciences from the Instituto Gulbenkian de Ciência (Portugal) in 2020: https://gtpb.github.io/PSLS20/. It was adapted by Q. Duchemin.

- *no fertiliser;*
- *biochar;*
- *compost;*
- *cobc (a combination of biochar and compost).*

Each group contains 7 *lettuce plants. The mean and standard deviation of the freshweight for each group are reported in Table 8.2. We want to know whether there is a statistically significant difference between these distributions, assuming that they are Gaussian.*

Group	No fertiliser	Biochar	Compost	Cobc
Mean	38	36.14286	58.85714	54.14286
Standard deviation	5.09902	3.760699	5.047394	6.440201

Table 8.2: Statistic of freshweight for Example 8.4.1.

The intuitive idea to construct a test for the hypotheses H_0 and H_1 first consists in estimating $\mu_1, \dots, \mu_k$ by the respective empirical means

$$\overline{X}_{1,\cdot} := \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \quad \dots, \quad \overline{X}_{k,\cdot} := \frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,i},$$

and then to reject H_0 if these estimators are 'far from each other'. A good way to measure this is for instance to introduce the global empirical mean

$$\overline{X}_{\cdot,\cdot} = \frac{1}{n}\sum_{\ell=1}^{k}\sum_{i=1}^{n_\ell} X_{\ell,i} = \frac{1}{n}\sum_{\ell=1}^{k} n_\ell \overline{X}_{\ell,\cdot},$$

and then to consider the statistic

$$\mathrm{SSM} = \sum_{\ell=1}^{k} n_\ell \left(\overline{X}_{\ell,\cdot} - \overline{X}_{\cdot,\cdot}\right)^2,$$

which is called the *Sum of Squares of the Model*. Indeed, the closer $\overline{X}_{1,\cdot}, \dots, \overline{X}_{k,\cdot}$ are from each other, the smaller SSM is. Thus, the test that we aim to construct has a rejection region of the form

$$W_n = \{\mathrm{SSM} \geq a_n\},$$

and it remains to determine the threshold a_n. To proceed, we compute the law of SSM under H_0.

Lemma 8.4.2 (Law of SSM)**.** *Under* H_0, $\dfrac{\mathrm{SSM}}{\sigma^2} \sim \chi_2(k-1)$.

Lemma 8.4.2 confirms that, to construct our test, we need to estimate σ^2. Within each sample, the empirical variance writes

$$\widehat{\sigma}^2_{\ell,n_\ell} = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(X_{\ell,i} - \overline{X}_{\ell,\cdot}\right)^2,$$

so one may aggregate all these estimators to define the *Sum of Squares for the Error*

$$\mathrm{SSE} = \sum_{\ell=1}^{k} n_\ell \widehat{\sigma}^2_{\ell,n_\ell}.$$

Lemma 8.4.3 (Law of SSE). *Under H_0, $\dfrac{\mathrm{SSE}}{\sigma^2} \sim \chi_2(n-k)$ and this variable is independent from* SSM.

As a consequence of Lemmas 8.4.2 and 8.4.3, we deduce that under H_0, the statistic

$$F_n = \frac{\mathrm{SSM}/(k-1)}{\mathrm{SSE}/(n-k)}$$

is distributed according to the Fisher distribution with $k-1$ and $n-k$ degrees of freedom. We conclude that the test rejecting H_0 as soon as $F_n \geq f_{k-1,n-k,1-\alpha}$ has level α: this is the F-test for ANOVA.

We now detail the proofs of Lemmas 8.4.2 and 8.4.3, which rely on a geometric interpretation and the use of Cochran's Theorem.

Proof of Lemmas 8.4.2 and 8.4.3. Let us denote by E the k-dimensional linear subspace of $\mathbb{R}^n$ spanned by the vectors

$$(\mathbf{1}_{n_1}, \mathbf{0}_{n_2}, \ldots, \mathbf{0}_{n_k}), (\mathbf{0}_{n_1}, \mathbf{1}_{n_2}, \ldots, \mathbf{0}_{n_k}), \ldots, (\mathbf{0}_{n_1}, \mathbf{0}_{n_2}, \ldots, \mathbf{1}_{n_k}),$$

where for all $\ell \in \{1, \ldots, k\}$,

$$\mathbf{1}_{n_\ell} = (1, \ldots, 1) \in \mathbb{R}^{n_\ell}, \qquad \mathbf{0}_{n_\ell} = (0, \ldots, 0) \in \mathbb{R}^{n_\ell}.$$

We also denote by H the one-dimensional linear subspace of $\mathbb{R}^n$ spanned by the vector

$$\mathbf{1}_n = (\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_k}) \in \mathbb{R}^n,$$

and notice that $H \subset E$. The orthogonal projections of $\mathbf{X}_n$ onto E and H are respectively denoted by $\mathbf{X}_n^E$ and $\mathbf{X}_n^H$. Then it is easily checked that

$$\mathrm{SSM} = \|\mathbf{X}_n^E - \mathbf{X}_n^H\|^2, \qquad \mathrm{SSE} = \|\mathbf{X}_n - \mathbf{X}_n^E\|^2.$$

Under H_0, let $\boldsymbol{\mu}_n \in \mathbb{R}^n$ be the vector $\mu \mathbf{1}_n$, where μ is the common value of $\mu_1, \ldots, \mu_k$. Then

$$\mathbf{X}_n = \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n, \qquad \boldsymbol{\epsilon}_n \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

Let us denote by $\boldsymbol{\epsilon}_n^E$ and $\boldsymbol{\epsilon}_n^H$ the respective orthogonal projections of $\boldsymbol{\epsilon}_n$ onto E and H, so that

$$\mathbf{X}_n^E = \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n^E, \qquad \mathbf{X}_n^H = \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n^H.$$

As a consequence,

$$\mathrm{SSM} = \|\mathbf{X}_n^E - \mathbf{X}_n^H\|^2 = \|\boldsymbol{\epsilon}_n^E - \boldsymbol{\epsilon}_n^H\|^2, \qquad \mathrm{SSE} = \|\mathbf{X}_n - \mathbf{X}_n^E\|^2 = \|\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}_n^E\|^2.$$

Combining Cochran's Theorem 3.1.3 with the orthogonal decomposition

$$\boldsymbol{\epsilon}_n = \underbrace{\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}_n^E}_{\in E^\perp} + \underbrace{\boldsymbol{\epsilon}_n^E - \boldsymbol{\epsilon}_n^H}_{\in H'} + \underbrace{\boldsymbol{\epsilon}_n^H}_{\in H},$$

where H' denotes the orthogonal of H in E, we deduce that SSM and SSE are independent, with

$$\frac{1}{\sigma^2}\mathrm{SSM} \sim \chi_2(k-1), \qquad \frac{1}{\sigma^2}\mathrm{SSE} \sim \chi_2(n-k),$$

which completes the proof. □

Remark 8.4.4 (SST). *Under H_0, the whole sample $\mathbf{X}_n$ is a standard Gaussian sample. Its empirical variance (multiplied by n) is called the* Sum of Squares for the Total

$$\mathrm{SST} = \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} \left(X_{\ell,i} - \overline{X}_{\cdot,\cdot}\right)^2,$$

and it has law $\frac{\mathrm{SST}}{\sigma^2} \sim \chi_2(n-1)$. In the geometric interpretation of the proof of Lemmas 8.4.2 and 8.4.3, it rewrites

$$\mathrm{SST} = \|\mathbf{X}_n - \mathbf{X}_n^H\|^2.$$

As a consequence, since $\mathbf{X}_n - \mathbf{X}_n^E \in E^\perp$ while $\mathbf{X}_n^E - \mathbf{X}_n^H \in E$, Pythagoras' Theorem yields

$$\|\mathbf{X}_n - \mathbf{X}_n^H\|^2 = \|\mathbf{X}_n - \mathbf{X}_n^E\|^2 + \|\mathbf{X}_n^E - \mathbf{X}_n^H\|^2,$$

which rewrites

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSM}.$$

In the module `scipy.stats`, ANOVA is performed using the function `f_oneway()`. Its application on the data of Example 8.4.1 writes as follows.

```
from scipy.stats import f_oneway

control = [38,34,41,43,43,29,38]
biochar = [38,34,30,42,35,36,38]
compost = [59,64,57,56,50,64,62]
cobc = [57,49,52,43,59,61,58]

f_oneway(control, biochar compost, cobc)
```

We obtain a p-value about $8.308\ 10^{-9}$, so H_0 is rejected at all usual levels.

8.A Exercises

Training exercises

Exercise 8.A.1 (Tests for simple regression)**.** In the simple regression case $p = 1$, you want to test whether y actually depends on x or not. You may use the two tests of Section 8.3: Student's test with $j = 1$, or Fisher's test with $q = 0$. But do they really differ?

Homework

Exercise 8.A.2. In order to be labelled as 'organic food', a food producer has to ensure that each of his products contains less than 1% of GMO[2]. He takes a sample of $n = 25$ products and computes the percentage of GMO in each of these products. We denote by X_i the logarithm of this percentage for the i-th product and assume that $X_1, \dots, X_n$ are iid under the law $\mathcal{N}(\mu, 1)$.

1. For the producer, the products do not contain GMO unless the contrary is proved. He therefore sets
$$H_0 = \{\mu \leq 0\}, \qquad H_1 = \{\mu > 0\}.$$
Write the rejection region of the test to use with level $\alpha = 0.05$.

2. An environmental organisation wants to make sure that the products do not contain GMO. In particular, they worry about the ability of the test to detect products with a quantity of GMO which is 50% larger than what is allowed. Compute the probability that the test does not detect the presence of GMO when the actual percentage of GMO is 1.5% (that is to say with $\mu = \log 1.5$).

[2] *OGM* en français.

3. Outraged by this result, the organisation wants to modify the producer's test. To them, the products do contain GMO unless the contrary is proved. Which test should be constructed? With $n = 25$ and $\alpha = 0.05$, what is now the probability to conclude to the absence of GMO when $\mu = \log 1.5$?

Supplementary exercises

Exercise 8.A.3 (The Cobb–Douglas model[3])**.** The *Cobb–Douglas model* in economics relates the total production Y of a company (the real value of all goods produced) in a year with the labour input L (the total number of person-hours worked) and the capital input K (a measure of all machinery, equipment, and buildings) through the formula

$$Y = AL^{\beta_1}K^{\beta_2}.$$

The variables Y, L and K are expressed in M\$.

The constant A is called the *total factor productivity*. We let $y = \log Y$, $\beta_0 = \log A$, $x^1 = \log L$, $x^2 = \log K$ and assume the linear model

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Applying linear regression on a database of $n = 1658$ companies for which Y, L and K are given, we obtain the following results:

$\widehat{\beta}_0$	3.136
$\widehat{\beta}_1$	0.738
$\widehat{\beta}_2$	0.282
R^2	0.945
$\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2$	148.27

$$(\mathbf{x}_n^\top \mathbf{x}_n)^{-1} = \begin{pmatrix} 0.0288 & 0.0012 & -0.0034 \\ 0.0012 & 0.0016 & 0.0010 \\ -0.0034 & 0.0010 & 0.0009 \end{pmatrix}$$

1. Compute the value of the unbiased estimator $\widehat{\sigma}^2$ of σ^2.
2. Give a confidence interval with level 95% for the total factor productivity.
3. Give a prediction interval with level 95% for the total production of a company with labour input $L = 100$ M\$ and capital input $K = 50$ M\$.
4. Determine, at the level 5%, if the capital input is a significative factor.
5. The model is said to display *constant returns to scale*[4] if multiplying L and K by $\lambda > 0$ results in multiplying Y by λ.
   (a) Express this condition in terms of β_1 and β_2.
   (b) Construct and apply a t-test deciding if the model displays constant returns to scale.

[3] Taken from the lecture notes *Régression linéaire* by A. Guyader.

[4] *Rendements d'échelle constants* en français.

Lecture 9

The Wald and Likelihood Ratio Tests

Contents

In this Lecture, we present two families of tests which may be applied for large classes of parametric models: Wald's asymptotic tests in Section 9.1, and the Likelihood Ratio test in Section 9.2. Both tests exhibit the interesting feature to be reminiscent of notions seen in the previous lectures: Wald's tests are a general class of tests which may be applied as soon as one has a consistent and asymptotically normal estimator, in a very similar way to the construction of asymptotic confidence intervals detailed in Lecture 4; while the Likelihood Ratio test is based on the notion of likelihood introduced in Lecture 5, and just like the Maximum Likelihood Estimator, it has a certain optimality property – namely, it is *Uniformy More Powerful* than any other test with the same level.

9.1 Wald's asymptotic tests

9.1.1 The one-sample Wald test

We consider null and alternative hypotheses which write under one of the following forms:

- one-sided: $H_0 = \{g(\theta) \leq g_0\}$, $H_1 = \{g(\theta) > g_0\}$, or $H_0 = \{g(\theta) \geq g_0\}$, $H_1 = \{g(\theta) < g_0\}$;
- two-sided: $H_0 = \{g(\theta) = g_0\}$, $H_1 = \{g(\theta) \neq g_0\}$;

for some function $g : \Theta \to \mathbb{R}$ and some fixed value $g_0 \in \mathbb{R}$.

Assume that a consistent estimator Z_n of $g(\theta) \in \mathbb{R}$ is available, that this estimator is asymptotically normal, and that a consistent estimator $\widehat{V}_n$ of its asymptotic variance $V(\theta)$ is also available. In this situation, Proposition 4.3.1 provides a systematic construction of asymptotic confidence intervals. The Wald test plays a similar role for hypothesis testing.

Proposition 9.1.1 (Two-sided Wald test)**.** *Consider the hypotheses $H_0 = \{g(\theta) = g_0\}$ and $H_1 = \{g(\theta) \neq g_0\}$ for some $g_0 \in g(\Theta)$. Under the assumptions made above, the test with rejection region*

$$W_n = \left\{ \sqrt{\frac{n}{\widehat{V}_n}} \, |Z_n - g_0| \geq \phi_{1-\alpha/2} \right\}$$

is consistent and has asymptotic level α.

Proof. By Slutsky's Lemma, we know that under H_0,

$$\lim_{n \to +\infty} \sqrt{\frac{n}{\widehat{V}_n}} \, (Z_n - g_0) = \mathcal{N}(0, 1), \qquad \text{in distribution,}$$

which by Lemma 4.3.2 shows that

$$\lim_{n\to+\infty} \mathbb{P}_\theta(W_n) = \alpha,$$

for any $\theta \in H_0$, and therefore that the test has asymptotic level α. Now, for any $\theta \in H_1$,

$$\lim_{n\to+\infty} \sqrt{\frac{n}{\widehat{V}_n}}\,|Z_n - g_0| = +\infty, \qquad \mathbb{P}_\theta\text{-almost surely,}$$

and therefore by Theorem 1.4.2, the test is consistent. □

Since, under H_0, the random variable $\sqrt{\frac{n}{\widehat{V}_n}}(Z_n - g_0)$ has asymptotic distribution $\mathcal{N}(0,1)$, the p-value associated with an observation $(z_n, \widehat{v}_n)$ is

$$\mathbb{P}\left(|G| \geq \sqrt{\frac{n}{\widehat{v}_n}}\,|z_n - g_0|\right) = 2\left(1 - \Phi\left(\sqrt{\frac{n}{\widehat{v}_n}}\,|z_n - g_0|\right)\right),$$

with $G \sim \mathcal{N}(0,1)$ and Φ the CDF of G.

Exercise 9.1.2. Under the same assumptions as in Proposition 9.1.1, construct a consistent test with asymptotic level α for the hypotheses $H_0 = \{g(\theta) \leq g_0\}$ and $H_1 = \{g(\theta) > g_0\}$.

A major interest of the Wald test is that it may be applied in a nonparametric setting, in the sense that it does not necessarily require one to specify a parametric model for the law of the sample. Consider for instance the setting of Example 7.1.12: writing the hypotheses under the form

$$H_0 = \{\mathbb{E}[X_1] \geq g_0\}, \quad H_1 = \{\mathbb{E}[X_1] < g_0\}, \qquad g_0 = 3,$$

one may observe that the test with rejection region

$$W_n = \left\{\sqrt{\frac{n}{V_n}}\,(\overline{X}_n - g_0) \leq -\phi_{1-\alpha}\right\},$$

with $V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ the empirical variance, is consistent and has asymptotic level α, *whatever the law* of the sample. Therefore, is contrast with the approach of Example 7.1.12, *there is no need to assume that the $X_1, \ldots, X_n$ are Exponential*. The price to pay is then that, in addition to the empirical mean, one must also compute the empirical variance of the sample.

9.1.2 Wald's test for the identity of means

The goal of this Section is to present the adaptation of Wald's test for the situation in which two samples are available. The framework is as follows: we observe two independent real-valued samples $\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1})$ and $\mathbf{X}_{2,n_2} = (X_{2,1}, \ldots, X_{2,n_2})$, and we denote

$$\mu_1 = \mathbb{E}[X_{1,i}], \qquad \mu_2 = \mathbb{E}[X_{2,i}].$$

Our aim is to construct a test for hypotheses made on μ_1 and μ_2. Depending on the considered application, we may consider either a two-sided test with hypotheses

$$H_0 = \{\mu_1 = \mu_2\}, \qquad H_1 = \{\mu_1 \neq \mu_2\},$$

or a one-sided test with hypotheses

$$H_0 = \{\mu_1 \leq \mu_2\}, \qquad H_1 = \{\mu_1 > \mu_2\}.$$

This framework includes the following particular cases.

Example 9.1.3 (Comparison of proportions)**.** *If both samples are composed of Bernoulli variables, then the test reduces to whether success has the same probability in both samples. This situation is ubiquitous in basic statistical studies, and called the problem of* comparison of proportions.

For instance, after the penaltygate *during the football game PSG/Montpellier in August 2022, the French newspaper* L'Équipe *released on social media the following statistics: in official games, Neymar scored 70 penalty kicks out of 85 tries (so the frequency of success is* 82%*), while Mbappé's rate is 20 out 25 (*80%*). The implicit claim was that Neymar has better success than Mbappé. Denoting by* p_1 *(resp.* p_2*) the probability for Neymar (resp. Mbappé) to score a penalty kick, is the estimated difference between* p_1 *and* p_2 *statistically significant?*

In this example, since the hypothesis to be tested is that Neymar is better than Mbappé (and not just that they have different efficiencies), it is appropriate to work with the hypotheses

$$H_0 = \{p_1 \leq p_2\}, \qquad H_1 = \{p_1 > p_2\},$$

which results in a one-sided test.

Example 9.1.4 (Gaussian model)**.** *If both samples are Gaussian, then the hypotheses*

$$H_0 = \{\mu_1 = \mu_2\}, \qquad H_1 = \{\mu_1 \neq \mu_2\},$$

are the same as for the Student test from Lecture 7. However, here, it is not required that both samples have the same variance σ^2*. The counterpart is that the test becomes asymptotic, which is not the case for Student's test.*

Construction of the test

The basic assumption is that within each sample, the variables are iid, with laws respectively denoted by P_1 and P_2, and that the quantities

$$\mu_j = \mathbb{E}[X_{j,1}], \qquad \sigma_j^2 = \mathrm{Var}(X_{j,1}), \qquad j \in \{1,2\},$$

are well-defined, with $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$. We denote by $\overline{X}_{1,n_1}$ and $\overline{X}_{2,n_2}$ the empirical means of the samples. Clearly,

$$Z_{n_1,n_2} = \overline{X}_{1,n_1} - \overline{X}_{2,n_2}$$

is a strongly consistent estimator of $\mu_1 - \mu_2$ when $n_1, n_2 \to +\infty$. Besides, it has variance

$$\mathrm{Var}(Z_{n_1,n_2}) = \mathrm{Var}\left(\overline{X}_{1,n_1}\right) + \mathrm{Var}\left(\overline{X}_{2,n_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Proposition 9.1.5 (Asymptotic normality)**.** *The estimator* Z_{n_1,n_2} *is asymptotically normal, that is to say,*

$$\lim_{n_1,n_2 \to +\infty} \frac{Z_{n_1,n_2} - (\mu_1 - \mu_2)}{\sqrt{\mathrm{Var}(Z_{n_1,n_2})}} = \mathcal{N}(0,1), \qquad \textit{in distribution.}$$

The proof of Proposition 9.1.5 is postponed to Exercise 9.1.8. A bit of technicality is required to handle the double index (n_1, n_2), but you should not be surprised by the overall statement of the proposition.

To construct the test, which definitely starts to look like a two-sample version of Wald's test, it remains to estimate $\mathrm{Var}(Z_{n_1,n_2})$. Assuming that two strongly consistent estimators $\widehat{\sigma}_{1,n_1}^2$ and $\widehat{\sigma}_{2,n_1}^2$ of σ_1^2 and σ_2^2 are available, we get the following intermediary result.

Lemma 9.1.6 (Strong consistency of variance estimation)**.** *We have*

$$\lim_{n_1,n_2 \to +\infty} \frac{\sqrt{\dfrac{\widehat{\sigma}_{1,n_1}^2}{n_1} + \dfrac{\widehat{\sigma}_{2,n_2}^2}{n_2}}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = 1, \qquad \textit{almost surely.}$$

Proof. Let $\epsilon \in (0,1)$. Almost surely, there exists $N \geq 1$ such that, for any $n_1, n_2 \geq N$,

$$1-\epsilon \leq \frac{\widehat{\sigma}^2_{1,n_1}}{\sigma_1^2} \leq 1+\epsilon, \qquad 1-\epsilon \leq \frac{\widehat{\sigma}^2_{2,n_1}}{\sigma_2^2} \leq 1+\epsilon.$$

Thus, for $n_1, n_2 \geq N$,

$$(1-\epsilon)\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right) \leq \frac{\widehat{\sigma}^2_{1,n_1}}{n_1}+\frac{\widehat{\sigma}^2_{2,n_2}}{n_2} \leq (1+\epsilon)\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right),$$

whence

$$\sqrt{1-\epsilon} \leq \frac{\sqrt{\dfrac{\widehat{\sigma}^2_{1,n_1}}{n_1}+\dfrac{\widehat{\sigma}^2_{2,n_2}}{n_2}}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \leq \sqrt{1+\epsilon},$$

which proves the claim. □

By Slutsky's Lemma, we deduce that

$$\lim_{n_1,n_2\to+\infty} \frac{Z_{n_1,n_2}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\widehat{\sigma}^2_{1,n_1}}{n_1}+\dfrac{\widehat{\sigma}^2_{2,n_2}}{n_2}}} = \mathcal{N}(0,1), \qquad \text{in distribution.}$$

This yields the following final result.

Proposition 9.1.7 (Asymptotic test of mean homogeneity)**.** *The two-sided test with rejection region*

$$W_{n_1,n_2} = \left\{ \frac{|Z_{n_1,n_2}|}{\sqrt{\dfrac{\widehat{\sigma}^2_{1,n_1}}{n_1}+\dfrac{\widehat{\sigma}^2_{2,n_2}}{n_2}}} \geq \phi_{1-\alpha/2} \right\}$$

is consistent and has asymptotic level α for the hypotheses

$$H_0 = \{\mu_1 = \mu_2\}, \qquad H_1 = \{\mu_1 \neq \mu_2\}.$$

The one-sided test with rejection region

$$W_{n_1,n_2} = \left\{ \frac{Z_{n_1,n_2}}{\sqrt{\dfrac{\widehat{\sigma}^2_{1,n_1}}{n_1}+\dfrac{\widehat{\sigma}^2_{2,n_2}}{n_2}}} \geq \phi_{1-\alpha} \right\}$$

is consistent and has asymptotic level α for the hypotheses

$$H_0 = \{\mu_1 \leq \mu_2\}, \qquad H_1 = \{\mu_1 > \mu_2\}.$$

Exercise 9.1.8 (Proof of Proposition 9.1.5)**.** In the setting of Proposition 9.1.5, we set

$$G_{1,n_1} = \frac{\overline{X}_{1,n_1}-\mu_1}{\sqrt{\sigma_1^2/n_1}}, \qquad G_{2,n_2} = \frac{\overline{X}_{2,n_2}-\mu_2}{\sqrt{\sigma_2^2/n_2}}.$$

1. Describe the limit in distribution of the pair (G_{1,n_1}, G_{2,n_2}) when $n_1, n_2 \to +\infty$.

2. Construct two deterministic sequences ρ_{n_1,n_2} and θ_{n_1,n_2} such that for any n_1, n_2,

$$\frac{Z_{n_1,n_2} - (\mu_1 - \mu_2)}{\sqrt{\mathrm{Var}(Z_{n_1,n_2})}} = \rho_{n_1,n_2} G_{1,n_1} + \theta_{n_1,n_2} G_{2,n_2}.$$

3. Prove Proposition 9.1.5 in the case where n_1, n_2 go to $+\infty$ with the ratio n_1/n_2 converging to some limit $\ell \in (0, +\infty)$.

4. Prove Proposition 9.1.5 in the case where n_1 and n_2 go to $+\infty$ without any particular condition. *You may admit (or prove!) that when a real-valued random variable converges in distribution, then its characteristic function converges uniformly on any bounded interval of* $\mathbb{R}$.

To conclude on Example 9.1.3 in the Bernoulli model, we have

$$\overline{X}_{1,n_1} = \frac{70}{85}, \qquad \overline{X}_{2,n_2} = \frac{20}{25}, \qquad \widehat{\sigma}^2_{1,n_1} = \overline{X}_{1,n_1}\left(1 - \overline{X}_{1,n_1}\right), \qquad \widehat{\sigma}^2_{2,n_2} = \overline{X}_{2,n_2}\left(1 - \overline{X}_{2,n_2}\right),$$

and $n_1 = 85$, $n_2 = 25$, so that the test statistics has value

$$\frac{Z_{n_1,n_2}}{\sqrt{\dfrac{\widehat{\sigma}^2_{1,n_1}}{n_1} + \dfrac{\widehat{\sigma}^2_{2,n_2}}{n_2}}} \simeq 0.26.$$

Since $\phi_{1-\alpha} = 1.65$ for $\alpha = 0.05$, the null hypothesis is not rejected at the level 5% and therefore the observed higher efficiency of Neymar is not statistically significant. In fact, the p-value of the observation is $\mathbb{P}(G \geq 0.26) \simeq 0.40$ for $G \sim \mathcal{N}(0,1)$, which shows that the observation is not significant at any usual confidence level.

Application to the geometric model

A study[1] published on December, 29th 2021, compares the clinical severity of waves due to the Delta variant (May 2021) and Omicron variant (November 2021) of Covid-19 in South Africa. Among severity indicators, the average length of stay in hospital[2] is reported in the table below.

Variant	Number of hospital admissions	Average length of stay in hospital
Delta	$n_1 = 4574$	$\overline{X}_{1,n_1} = 8.5$ days
Omicron	$n_2 = 4438$	$\overline{X}_{2,n_2} = 4$ days

From these data, we want to test whether the mean length of stay in hospital is significantly shorter for the Omicron wave than for the Delta wave. Denoting by μ_1 and μ_2 these mean lengths, the null and alternative hypotheses thus write

$$H_0 = \{\mu_1 \leq \mu_2\}, \qquad H_1 = \{\mu_1 > \mu_2\}.$$

We assume that the lengths of stay $\mathbf{X}_{1,n_1} = (X_{1,i})_{1 \leq i \leq n_1}$ and $\mathbf{X}_{2,n_2} = (X_{2,i})_{1 \leq i \leq n_2}$ for these two waves are independent samples, and that each sample contains independent geometric variables with respective parameters p_1 and p_2.

Exercise 9.1.9 (Application of the test).

1. Express the variances σ_1^2 and σ_2^2 for each sample as a function of μ_1 and μ_2.

2. Deduce strongly consistent estimators $\widehat{\sigma}^2_{1,n_1}$ and $\widehat{\sigma}^2_{2,n_2}$ of σ_1^2 and σ_2^2 which write as functions of $\overline{X}_{1,n_1}$ and $\overline{X}_{2,n_2}$.

[1] http://dx.doi.org/10.2139/ssrn.3996320

[2] Actually, only quartiles and the median are given in the study. We shall admit that the numbers in the present table are a realistic extrapolation of these data.

3. We recall that the quantile of order 0.95 of the law $\mathcal{N}(0,1)$ is 1.65. At the level $\alpha = 0.05$, is the decrease in the length of stay in hospital significant between the two waves?

4. The data from the same study, but restricted to patients with age less than 20 years, are reported below. At the same level, is the decrease in the length of stay in hospital significant?

Variant	Number of hospital admissions	Average length of stay in hospital
Delta	$n_1 = 161$	$\overline{X}_{1,n_1} = 3.5$ days
Omicron	$n_2 = 844$	$\overline{X}_{2,n_2} = 3.1$ days

9.2 The likelihood ratio test

The *likelihood ratio test* is useful to derive a test statistic, with a certain optimality property, when there is no obvious estimator of θ available. We present this test in the case where both the null and the alternative hypotheses are simple: $\Theta = \{\theta_0, \theta_1\}$ and $H_0 = \{\theta = \theta_0\}$, $H_1 = \{\theta = \theta_1\}$.

Example 9.2.1 (Unimodal versus bimodal model)**.** *Consider a population of micro-organisms which may be either monomorphic or dimorphic, and for which you observe a certain trait $X \in \mathbb{R}$. If the population is monomorphic then $X \sim P_0$, with $P_0 = \mathcal{N}(0,1)$ and if the population is dimorphic then $X \sim P_1$, with $P_1 = \frac{1}{2}\mathcal{N}(-b,1) + \frac{1}{2}\mathcal{N}(b,1)$, for some known parameter $b > 0$; see Figure 9.1. You observe the traits $X_1, \dots, X_n$ of a population and want to test whether the population is dimorphic. Therefore you set $H_0 = \{P = P_0\}$ and $H_1 = \{P = P_1\}$, where P is the law of your sample.*

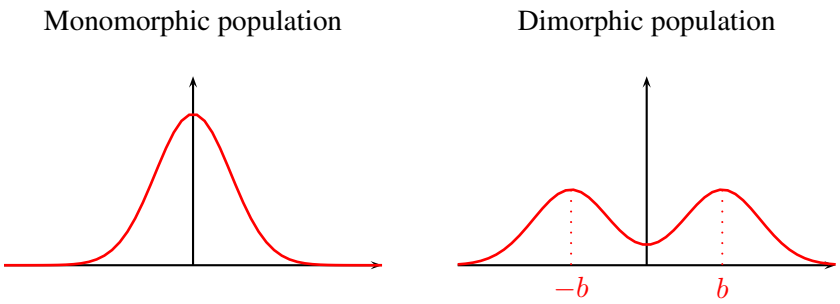


Figure 9.1: Distribution P of the trait X for monomorphic and dimorphic populations in Example 9.2.1.

Exercise 9.2.2. In the example above, compute the variance of X_1 under H_0 and under H_1. Deduce a consistent test, with level α, for these hypotheses.

9.2.1 Presentation of the test

The likelihood ratio test is based on the following statistic.

Definition 9.2.3 (Likelihood ratio)**.** *With simple hypotheses $H_0 = \{\theta = \theta_0\}$ and $H_1 = \{\theta = \theta_1\}$, the* likelihood ratio *is the statistic $\Lambda_n(\mathbf{X}_n)$, where*

$$\Lambda_n(\mathbf{x}_n) = \frac{L_n(\mathbf{x}_n; \theta_1)}{L_n(\mathbf{x}_n; \theta_0)}.$$

By the definition of the likelihood of a realisation, the likelihood ratio typically takes larger values under H_1 than under H_0, which leads to defining a test with rejection region

$$W_n^{\mathrm{LR}} = \{\Lambda_n(\mathbf{X}_n) \geq a_n\},$$

with a threshold $a_n \geq 0$ selected for the test to have level α. This test is called the *likelihood ratio test*.

Example 9.2.4 (Continuation of Example 9.2.1)**.** *In the problem of Example 9.2.1, the likelihood ratio writes*

$$\Lambda_n(\mathbf{x}_n) = \frac{\displaystyle\prod_{i=1}^n \left(\frac{1}{2} \frac{\mathrm{e}^{-(x_i+b)^2/2}}{\sqrt{2\pi}} + \frac{1}{2} \frac{\mathrm{e}^{-(x_i-b)^2/2}}{\sqrt{2\pi}} \right)}{\displaystyle\prod_{i=1}^n \frac{\mathrm{e}^{-x_i^2/2}}{\sqrt{2\pi}}} = \mathrm{e}^{-nb^2/2} \prod_{i=1}^n \cosh(bx_i).$$

Therefore, the likelihood ratio test W_n^{LR} has level α if the threshold a_n writes

$$a_n = \mathrm{e}^{-nb^2/2} q_{n,1-\alpha},$$

where $q_{n,r}$ is the quantile of order r of the law of the random variable

$$\prod_{i=1}^n \cosh(bG_i),$$

with iid standard Gaussian variables $G_1, \dots, G_n$.

In the example, to complete the practical construction of the test, one has to compute the quantile $q_{n,1-\alpha}$. When it is not possible to do so analytically, the standard approach consists in using the Monte Carlo method, see Remark 7.1.21. Alternatively, one may use the following asymptotic result.

Proposition 9.2.5 (Asymptotics of the likelihood ratio test)**.** *Assume that the quantities*

$$h := \mathbb{E}_{\theta_0} \left[\ell_1(X_1;\theta_1) - \ell_1(X_1;\theta_0) \right], \qquad s^2 := \mathrm{Var}_{\theta_0} \left(\ell_1(X_1;\theta_1) - \ell_1(X_1;\theta_0) \right),$$

are well-defined. Then, the test with rejection region

$$W_n^{\mathrm{LR,asymp}} = \left\{ \frac{1}{n} \log(\Lambda_n(\mathbf{X}_n)) \geq h + \sqrt{\frac{s^2}{n}} \phi_{1-\alpha} \right\}$$

is has asymptotic level α. Besides, if the model is identifiable[3]*, then this test is consistent.*

Proof. By definition of the likelihood ratio,

$$\frac{1}{n} \log(\Lambda_n(\mathbf{X}_n)) = \frac{1}{n} \sum_{i=1}^n \left(\ell_1(X_i;\theta_1) - \ell_1(X_i;\theta_0) \right).$$

Therefore, under H_0, the Central Limit Theorem yields

$$\sqrt{\frac{n}{s^2}} \left(\frac{1}{n} \log(\Lambda_n(\mathbf{X}_n)) - h \right) \to \mathcal{N}(0,1), \qquad \text{in distribution,}$$

which shows that the test has asymptotic level α.

To show consistency, we observe that under H_1,

$$\lim_{n\to+\infty} \frac{1}{n} \log(\Lambda_n(\mathbf{X}_n)) = \mathbb{E}_{\theta_1} \left[\ell_1(X_1;\theta_1) - \ell_1(X_1;\theta_0) \right], \qquad \mathbb{P}_{\theta_1}\text{-almost surely.}$$

To complete the proof, it thus suffices to show that the right-hand side is strictly larger than h. By Proposition 5.1.12, we have

$$\mathbb{E}_{\theta_1} \left[\ell_1(X_1;\theta_1) - \ell_1(X_1;\theta_0) \right] \geq 0 \geq \mathbb{E}_{\theta_0} \left[\ell_1(X_1;\theta_1) - \ell_1(X_1;\theta_0) \right] = h,$$

and it is not difficult to show that if the model is identifiable then both inequalities are strict. □

[3] We recall that it means that the probability measures P_{θ_0} and P_{θ_1} do not coincide.

9.2.2 The Neyman–Pearson Lemma

So far, we have constructed two tests for Example 9.2.1: a somehow intuitive test based on the variance in Exercise 9.2.2, and the likelihood ratio test which uses as a test statistic the product $\prod_{i=1}^n \cosh(bX_i)$ and may arguably be considered as less intuitive. There is a parallel to be drawn here with the construction of estimators. Indeed, the variance-based test shares common features with the method of moments: it relies on the arbitrary but natural choice of a moment (here, the variance) to be compared between H_0 and H_1, but there is no reason why this test should be better than another one. On the other hand, the likelihood ratio test is similar to the MLE principle as its construction obeys a systematic rule, based on the notion of likelihood. Given the efficiency results for the MLE exposed in Lecture 5, one might therefore expect the likelihood ratio test to satisfy a certain *optimality* property. This is indeed the case.

Proposition 9.2.6 (Neyman–Pearson Lemma)**.** *Among all tests of level α for the simple hypotheses $H_0 = \{\theta = \theta_0\}$ and $H_1 = \{\theta = \theta_1\}$, the likelihood ratio test is the most powerful.*

Proof. Let W_n be a subset of E^n such that $\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W_n) \leq \alpha$. We want to show that

$$\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n) \leq \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n^{\mathrm{LR}}).$$

To this aim, we first write

$$\begin{aligned}
&\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n^{\mathrm{LR}}) - \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n) \\
&\quad = \int_{\mathbf{x}_n \in W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n - \int_{\mathbf{x}_n \in W_n} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n \\
&\quad = \int_{\mathbf{x}_n \in W_n^{\mathrm{LR}} \setminus W_n} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n - \int_{\mathbf{x}_n \in W_n \setminus W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n.
\end{aligned}$$

Since

$$\mathbf{x}_n \in W_n^{\mathrm{LR}} \qquad \text{if and only if} \qquad \Lambda_n(\mathbf{x}_n) = \frac{L_n(\mathbf{x}_n;\theta_1)}{L_n(\mathbf{x}_n;\theta_0)} \geq a_n,$$

we get

$$\begin{aligned}
\int_{\mathbf{x}_n \in W_n^{\mathrm{LR}} \setminus W_n} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n &\geq a_n \int_{\mathbf{x}_n \in W_n^{\mathrm{LR}} \setminus W_n} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n, \\
\int_{\mathbf{x}_n \in W_n \setminus W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_1)\mathrm{d}\mathbf{x}_n &\leq a_n \int_{\mathbf{x}_n \in W_n \setminus W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n,
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n^{\mathrm{LR}}) - \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W_n) \\
&\quad \geq a_n \left(\int_{\mathbf{x}_n \in W_n^{\mathrm{LR}} \setminus W_n} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n - \int_{\mathbf{x}_n \in W_n \setminus W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n \right) \\
&\quad = a_n \left(\int_{\mathbf{x}_n \in W_n^{\mathrm{LR}}} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n - \int_{\mathbf{x}_n \in W_n} L_n(\mathbf{x}_n;\theta_0)\mathrm{d}\mathbf{x}_n \right) \\
&\quad = a_n \left(\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W_n^{\mathrm{LR}}) - \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W_n) \right).
\end{aligned}$$

Since the level of the likelihood ratio test is α while $\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W_n) \leq \alpha$, we deduce that the right-hand side above is nonnegative, which proves the claimed inequality. □

To conclude the study of Example 9.2.1, it may be an interesting exercise to illustrate Proposition 9.2.6 by computing (by numerical simulation) the power of the variance-based test and of the likelihood ratio test, and possibly plotting both results as a function of the parameter b.

9.2.3 The case of composite hypotheses

In the case of composite hypotheses, a variant of the likelihood ratio test is based on the statistic $\widetilde{\Lambda}_n(\mathbf{X}_n)$, where

$$\widetilde{\Lambda}_n(\mathbf{x}_n) = \frac{\sup_{\theta\in\Theta} L_n(\mathbf{x}_n;\theta)}{\sup_{\theta\in H_0} L_n(\mathbf{x}_n;\theta)}.$$

Notice that in the case of simple hypotheses, $\widetilde{\Lambda}_n(\mathbf{x}_n)$ does *not* reduce to $\Lambda_n(\mathbf{x}_n)$, because in the numerator, there remains a maximum of $L_n(\mathbf{x}_n;\theta)$ over $\{\theta_0,\theta_1\}$. In particular, since $H_0 \subset \Theta$, we have $\widetilde{\Lambda}_n(\mathbf{X}_n) \geq 1$.

To construct a test with this statistic, we need to describe its (asymptotic) behaviour under H_0 and H_1. This is the content of Wilks' Theorem[4].

Theorem 9.2.7 (Wilks' Theorem)**.** *Assume that Θ has nonempty interior in $\mathbb{R}^p$ and H_0 has nonempty interior in $\mathbb{R}^q$ for some $q < p$.*

- *Under H_1, $\widetilde{\Lambda}_n(\mathbf{X}_n) \to +\infty$, almost surely.*
- *Under H_0, $2\log\widetilde{\Lambda}_n(\mathbf{X}_n) \to \chi_2(p-q)$, in distribution.*

As a consequence of Wilks' Theorem, we deduce that the test with rejection region

$$W_n = \{2\log\widetilde{\Lambda}_n(\mathbf{X}_n) \geq \chi^2_{p-q,1-\alpha}\}$$

is consistent and has asymptotic level α.

This test is particularly useful in the context of logistic regression. Assume that you want to test if y actually depends on the features $x^{q+1},\dots,x^p$ for some $q \in \{0,\dots,p-1\}$. Thus, we introduce the null and alternative hypotheses

$$H_0 = \{\beta_{q+1} = \cdots = \beta_p = 0\}, \qquad H_1 = \mathbb{R}^{p+1} \setminus H_0.$$

To apply the Likelihood Ratio Test, one now needs to compute the MLE $\widehat{\beta}$ in the whole model (that is to say with all features $x^1,\dots,x^p$), and the MLE $\widehat{\beta}_0$ in the restricted model (only with features $x^1,\dots,x^q$). The likelihood ratio statistic is then

$$\widetilde{\Lambda}_n(\mathbf{x}_n,\mathbf{y}_n) = \frac{L_n(\mathbf{x}_n,\mathbf{y}_n;\widehat{\beta})}{L_n(\mathbf{x}_n,\mathbf{y}_n;\widehat{\beta}_0)},$$

and the test rejecting H_0 if $2\log(\widetilde{\Lambda}_n(\mathbf{x}_n,\mathbf{y}_n)) \geq \chi^2_{p-q,1-\alpha}$ is consistent and has asymptotic level α.

9.A Exercises

Training exercises

Exercise 9.A.1 (Bias in coin toss)**.** In a 2007 scientific article[5], a precession phenomenon is evidenced in the coin toss experiment, and the authors conclude that it is more likely that the coin lands the same way up as it started. In order to experimentally verify this claim, we design an experiment in which a coin is tossed n times, alternatively starting from Heads and Tails, and we denote by p the probability to get the same position as initially. We set

$$H_0 = \left\{p = \frac{1}{2}\right\}, \qquad H_1 = \left\{p > \frac{1}{2}\right\},$$

therefore discarding the possibility of a bias toward the opposite side to the initial one.

[4] S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of Mathematical Statistics*, 1938. See also [7, Chapter 16].

[5] Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical Bias in the Coin Toss. *SIAM Review*.

1. Justify the choice of these hypotheses.
2. Describe the rejection region of a consistent test with asymptotic level α for these hypotheses.
3. This experiment was carried out in 2009 by two students of the University of Berkeley[6], who made $40\,000$ tosses and got $20\,245$ times the same side as the initial one. Compute the p-value of this result and give the conclusion of the test.
4. The 2007 article predicts a probability $p = 50.8\%$ to land on the same side as the initial one. With this value, what is the power of the test in the 2009 experiment?
5. We now wonder whether the probability to get the same side as the initial one depends on this side. We set p_{H} the probability to get Heads starting from Heads, p_{T} the probability to get Tails starting from Tails, and
$$H_0 = \{p_{\mathrm{H}} = p_{\mathrm{T}}\}, \qquad H_1 = \{p_{\mathrm{H}} \neq p_{\mathrm{T}}\}.$$
The data of the 2009 experiment are reported below:

	Heads obtained	Tails obtained
Heads initial	10 231	9 769
Tails initial	9 986	10 014

Choose a suitable test and compute the associated p-value.

Remark. In spite of its result, the 2009 experiment suffers a few weaknesses in its protocol. It has been replicated in 2023 by a team of 50 researchers who made $350\,757$ throws and confirmed the value 50.8% by getting as a confidence interval $[0.506, 0.509]$[7].

Exercise 9.A.2 (Correlation test)**.** Let $(X_i, Y_i)_{1\leq i\leq n}$ be a sample of iid pairs of random variables, such that $\mathbb{E}[X_1^2 + Y_1^2] < +\infty$. We set
$$C = \mathrm{Cov}(X_1, Y_1) = \mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1],$$
and consider the hypotheses
$$H_0 = \{C = 0\}, \qquad H_1 = \{C \neq 0\}.$$

1. Find a strongly consistent estimator $\widehat{C}_n$ of C.
2. Show that $\widehat{C}_n$ is asymptotically normal and compute its asymptotic variance V. *Hint: you may start by computing the covariance matrix of the vector* (X_1, Y_1, X_1Y_1).
3. Find a strongly consistent estimator $\widehat{V}_n$ of V and deduce a consistent test with asymptotic level α for H_0 and H_1.
4. You observe an empirical correlation $\widehat{\rho}_n \neq 0$ in your data. What is the value of n (depending on $\widehat{\rho}_n$ and α) starting from which you may conclude that your data are indeed correlated?

[6] See https://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html.
[7] See https://www.pourlascience.fr/sd/mathematiques/un-biais-dans-le-pile-ou-face-25906.php.

Lecture 10

χ_2 Tests for Finite State Spaces

Contents

In this Lecture, we address models where the observed data $X_1, \dots, X_n$ take their values in a *finite* state space E, with cardinality m. In this situation, a probability measure P is characterised by its probability mass function $(p_x)_{x \in E}$, which can be interpreted as a vector of $\mathbb{R}^m$. Therefore, it is technically possible to use the formalism of parametric estimation with $\theta = (p_x)_{x \in E}$ directly.

In this Lecture, we adopt this point of view and introduce several χ_2 tests with various purposes.

- *Goodness-of-fit* tests[1] allow to test whether a sample is distributed according to some given probability measure P_0 on E. For example, they allow to test whether a random digit generator returns integers which are uniformly distributed in $\{0, \dots, 9\}$.

- *Goodness-of-fit* tests to a family of distributions[2] allow to test whether the law of a sample belongs to a given, low-dimensional, parametric family. For example, they allow to test whether the numbers of students skipping a lecture follows a Binomial distribution or not.

- *Independence* tests concern samples which take the form $(X_i, Y_i)_{1 \leq i \leq n}$ in some finite product space $E \times F$ and allow to test whether X and Y are independent. For example, they allow to determine whether the gender of your second child depends on the gender of your first child.

We shall also see that *homogeneity tests* between $k \geq 2$ samples can be recast as independence tests and are therefore covered by χ_2 tests. In other words, whatever you want to test in a model with finite state space, there is always a χ_2 test for you!

While we use the technical framework of parametric estimation, with Θ being the set of all probability measures on E (or on $E \times F$ for independence tests), we do not make any specific assumption on the law of the data. From this point of view, the tests which we shall construct in this Lecture are therefore of a *nonparametric* nature, and their study will be continued with the introduction of nonparametric tests on the real line in Lecture 11.

[1] *Tests d'adéquation* en français.

[2] *Tests d'adéquation à une famille de lois* en français.

10.1 Empirical distribution in the finite setting

In this Section, we consider iid random variables $X_1, \dots, X_n$ which take their values in a finite state space E with cardinality m. The basic idea for χ_2 tests consists in approximating P with the *empirical measure* of the sample, which is the probability measure $\widehat{P}_n$ on E with probability mass function $(\widehat{p}_{n,x})_{x\in E}$ defined by

$$\forall x \in E, \qquad \widehat{p}_{n,x} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}}.$$

It is clear that $\widehat{P}_n$ is an unbiased and strongly consistent estimator of P. Therefore, if one fixes a probability measure P_0 on E and wants to construct a goodness-of-fit test for the hypotheses

$$H_0 = \{P = P_0\}, \qquad H_1 = \{P \neq P_0\},$$

a natural idea consists in choosing a distance d on the space of probability measures on E (seen as a subset of $\mathbb{R}^m$) and taking a rejection region of the form

$$W_n = \{d(\widehat{P}_n, P_0) \geq a_n\}$$

for some threshold a_n to be chosen appropriately. In this Section, we introduce and study a specific distance-like function d, the χ_2 *distance*, which has the advantage to make the computation of the threshold a_n *independent from* P_0, a property which is recurrent in the construction of nonparametric tests.

10.1.1 The χ_2 distance

Let P and Q be two probability measures on E, with respective probability mass functions $(p_x)_{x\in E}$ and $(q_x)_{x\in E}$.

Definition 10.1.1 (Absolute continuity and χ_2 distance)**.** *The probability measure P is said to be absolutely continuous with respect to Q, which we denote by $P \ll Q$, if $q_x = 0$ implies $p_x = 0$. The χ_2 distance between P and Q, denoted by $\chi_2(P|Q)$, is defined by*

$$\chi_2(P|Q) := \begin{cases} \displaystyle\sum_{x\in E} \left(\frac{p_x}{q_x} - 1\right)^2 q_x & \text{if } P \ll Q, \\ +\infty & \text{otherwise}, \end{cases}$$

with the convention that $p_x/q_x = 0$ when $p_x = q_x = 0$.

When $P \ll Q$, the quantity $\chi_2(P|Q)$ rewrites more explicitly as

$$\chi_2(P|Q) = \sum_{x\in E} \frac{(p_x - q_x)^2}{q_x},$$

with the convention that $(p_x - q_x)^2/q_x = 0$ when $p_x = q_x = 0$.

Remark 10.1.2. *The χ_2 distance is not a distance, because it is not symmetric. However, it can be checked that $\chi_2(P|Q) = 0$ if and only if $P = Q$, so that $\chi_2(P|Q)$ can still be understood as a measure of how close P and Q are.*

10.1.2 Limit theorems for the empirical distribution

In this Subsection, we consider a sample $X_1, \dots, X_n$ iid according to some probability measure P on E, and we denote by $\widehat{P}_n$ the associated empirical measure.

Remark 10.1.3. *If there exists $x \in E$ such that $p_x = 0$, then almost surely, $\widehat{p}_{n,x} = 0$ for all $n \geq 1$. Therefore up to removing x from E, we shall always assume that $p_x > 0$ for all $x \in E$ in the sequel.*

Proposition 10.1.4 (Asymptotic behaviour of $\widehat{P}_n$)**.** *Let $X_1, \dots, X_n \in E$ be independent random variables, identically distributed according to P. When $n \to +\infty$,*

(i) the empirical distribution $\widehat{P}_n$ converges almost surely to P,

(ii) $\sqrt{n}(\widehat{P}_n - P)$ (seen as a random vector of $\mathbb{R}^m$) converges in distribution to a Gaussian vector $\mathcal{N}_m(0, K)$ with covariance matrix $K = (K_{x,y})_{x,y\in E}$ given by

$$K_{x,y} = \begin{cases} p_x(1-p_x) & \text{if } x = y, \\ -p_x p_y & \text{if } x \neq y. \end{cases}$$

Proof. We first notice that

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} R_i, \qquad R_i = (r_{i,x})_{x\in E},$$

where $R_1, \dots, R_n$ are the iid vectors of $\mathbb{R}^m$ such that $r_{i,x} = \mathbb{1}_{\{X_i = x\}}$. By the strong Law of Large Numbers, we deduce that

$$\lim_{n\to+\infty} \widehat{P}_n = \mathbb{E}[R_1] = P, \qquad \text{almost surely.}$$

Now by the Central Limit Theorem,

$$\lim_{n\to+\infty} \sqrt{n}(\widehat{P}_n - P) = \mathcal{N}_m(0, K), \qquad \text{in distribution,}$$

where K is the covariance matrix of R_1. Its coefficients $(K_{x,y})_{x,y\in E}$ write

$$K_{x,y} = \mathrm{Cov}(r_{1,x}, r_{1,y}) = \mathbb{E}\left[\mathbb{1}_{\{X_1=x\}}\mathbb{1}_{\{X_1=y\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{X_1=x\}}\right]\mathbb{E}\left[\mathbb{1}_{\{X_1=y\}}\right],$$

and we have $\mathbb{E}[\mathbb{1}_{\{X_1=x\}}]\mathbb{E}[\mathbb{1}_{\{X_1=y\}}] = p_x p_y$ while

$$\mathbb{E}\left[\mathbb{1}_{\{X_1=x\}}\mathbb{1}_{\{X_1=y\}}\right] = \begin{cases} \mathbb{E}\left[\mathbb{1}_{\{X_1=x\}}\right] = p_x & \text{if } x = y, \\ 0 & \text{if } x \neq y, \end{cases}$$

which yields the claimed expression for $K_{x,y}$. □

Corollary 10.1.5 (Asymptotic behaviour of $\chi_2(\widehat{P}_n|Q)$)**.** *Under the assumptions of Proposition 10.1.4,*

(i) for any probability measure Q on E, with a probability mass function $(q_x)_{x\in E}$ such that $q_x > 0$ for all $x \in E$, $\chi_2(\widehat{P}_n|Q)$ converges to $\chi_2(P|Q)$,

(ii) $n\chi_2(\widehat{P}_n|P)$ converges in distribution to a random variable with distribution $\chi_2(m-1)$.

Proof. The first point follows from the continuity of $P \mapsto \chi_2(P|Q)$ and Proposition 10.1.4 (i). In order to prove the second point, let us introduce the diagonal matrix $M \in \mathbb{R}^{m\times m}$ with diagonal coefficients $(1/\sqrt{p_x})_{x\in E}$, so that

$$n\chi_2(\widehat{P}_n|P) = n\sum_{x\in E}\left(\frac{\widehat{p}_{n,x} - p_x}{\sqrt{p_x}}\right)^2 = \left\|M\sqrt{n}(\widehat{P}_n - P)\right\|^2.$$

By Proposition 10.1.4 (ii), $n\chi_2(\widehat{P}_n|P)$ converges in distribution to $\|MU\|^2$, where $U \sim \mathcal{N}_m(0, K)$. As a consequence, $MU \sim \mathcal{N}_m(0, \Pi)$ with $\Pi = MKM^\top$. The coefficients $(\Pi_{x,y})_{x,y\in E}$ are easy to compute and write

$$\Pi_{x,y} = \frac{1}{\sqrt{p_x}\sqrt{p_y}}K_{x,y} = \begin{cases} 1 - p_x & \text{if } x = y, \\ -\sqrt{p_x p_y} & \text{if } x \neq y. \end{cases}$$

It is now straightforward to check that

$$\Pi = I_m - ee^\top, \qquad e = (\sqrt{p_x})_{x\in E}, \qquad \|e\| = 1,$$

so that Π is the orthogonal projection of $\mathbb{R}^m$ onto the $m-1$ dimensional space $e^\perp$. But now, if G is a standard m-dimensional Gaussian vector as in Definition 3.1.1, then $\Pi G \sim \mathcal{N}(0, \Pi\Pi^\top) = \mathcal{N}(0,\Pi)$ so MU has the same law as ΠG. Finally, the Cochran Theorem 3.1.3 and Remark 3.1.5 yield that $\Pi G \sim \chi_2(m-1)$. □

As already mentioned in the introduction of this Section, it is remarkable that the limit law of $n\chi_2(\widehat{P}_n|P)$ does not depend on P. This would not necessarily be the case with other choices of distances, for instance it is clear from Proposition 10.1.4 (ii) that $\sqrt{n}\|\widehat{P}_n - P\|$ converges in distribution to the norm of a Gaussian vector whose covariance depends on P.

10.2 Goodness-of-fit χ_2 test

In this Section, we fix a probability measure $P_0 = (p_{0,x})_{x\in E}$ and address the goodness-of-fit test for the hypotheses

$$H_0 = \{P = P_0\}, \qquad H_1 = \{P \neq P_0\}.$$

Following Remark 10.1.3, we assume, up to removing some elements from E, that $p_{0,x} > 0$ for all $x \in E$.

Example 10.2.1 (Uniform number generator)**.** *In order to assess the quality of a random number generator, we draw $n = 1000$ digits at random between 0 and 9 (for instance, letting $X_i = \lfloor 10U_i \rfloor$ where $(U_i)_{1\le i\le n}$ are independent uniform variables on $[0,1]$) and obtain the following results.*

Digit	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Number of occurrences	*90*	*90*	*100*	*98*	*104*	*107*	*111*	*84*	*102*	*114*

We would like to know whether the distribution is uniform in $\{0,\dots,9\}$ or not.

Definition 10.2.2 (Pearson's statistic)**.** *The statistic*

$$d_n = n\chi_2(\widehat{P}_n|P_0) = n\sum_{x\in E}\frac{(\widehat{p}_{n,x} - p_{0,x})^2}{p_{0,x}}$$

is called Pearson's statistic.

For all $\ell \ge 1$ and $r \in (0,1)$, we denote by $\chi^2_{\ell,r}$ the quantile of order r of the $\chi_2(\ell)$ distribution. The χ_2 test is presented in the next proposition.

Proposition 10.2.3 (χ_2 goodness-of-fit test)**.** *For all $\alpha \in (0,1)$, the test with rejection region*

$$W_n = \{d_n \ge \chi^2_{m-1,1-\alpha}\}$$

is consistent and has asymptotic level α.

Proof. We first check that this test is consistent, and therefore assume that $P \neq P_0$. Then by Corollary 10.1.5 (i), $\chi_2(\widehat{P}_n|P_0)$ converges almost surely to $\chi_2(P|P_0) > 0$, therefore $d_n \to +\infty$. As consequence, $\mathbb{P}(W_n)$ converges to 1 and the test is consistent.

We now compute the asymptotic level, and therefore take $P = P_0$. By Corollary 10.1.5 (ii), d_n converges in distribution to a random variable $\xi \sim \chi_2(m-1)$, so that

$$\lim_{n\to+\infty}\mathbb{P}(W_n) = \mathbb{P}(\xi \ge \chi^2_{m-1,1-\alpha}) = \alpha,$$

which completes the proof. □

Remark 10.2.4 (Validity of the asymptotic approximation)**.** *The χ_2 test is based on the approximation of the law of Pearson's statistic under H_0 by the χ_2 distribution, which theoretically only holds when n goes to $+\infty$. In practice, this approximation is considered to be legitimate when the property*

$$\forall x \in E, \qquad np_{0,x}(1-p_{0,x}) \geq 5$$

holds. Notice that this property holds if and only if

$$n\left(\min_{x\in E} p_{0,x}\right)\left(1-\left(\min_{x\in E} p_{0,x}\right)\right) \geq 5 \qquad \text{and} \qquad n\left(\max_{x\in E} p_{0,x}\right)\left(1-\left(\max_{x\in E} p_{0,x}\right)\right) \geq 5,$$

so that only 2 computations are necessary to check this property.

Exercise 10.2.5. Apply the χ_2 test to answer the question of Example 10.2.1 (with the command `chi2.ppf(0.95, df=9)` in `scipy.stats`, one obtains $\chi^2_{9,.95} = 16.9$).

There is in fact no need to compute Pearson's statistic by hand. Let us indeed define the data from Example 10.2.1, the distribution P_0 and perform the χ_2 test as follows:

```
import numpy as np
from scipy.stats import chisquare

# Empirical occurrences
empirical_occurrences = np.array([90, 90, 100, 98, 104, 107, 111, 84, 102, 114])

# Theoretical frequencies (assuming equal probabilities)
theoretical_freq = np.array([0.1] * 10)

# Multiply the theoretical frequencies by the sum of empirical occurrences to
    get expected frequencies
expected_freq = theoretical_freq * empirical_occurrences.sum()

# Perform the chi-squared goodness-of-fit test
gof_test_statistic, p_value = chisquare(f_obs=empirical_occurrences, f_exp=
    expected_freq)
```

Then the value of d_n and the associated p-value are respectively given by:

```
# Output the test statistic and p-value
print("Test Statistic (d_n):", gof_test_statistic)
print("P-Value:", p_value)
```

Remark 10.2.6 (Application to general models)**.** *When the state space E is infinite, the χ_2 test cannot be applied as the number of degrees of freedom $m-1$ of the χ_2 statistic should be infinite. However, it can be adapted by partitioning the state space E into a finite number of classes $\widetilde{E} = \{A_1, \dots, A_m\}$ and testing whether the random variables $\widetilde{X}_1, \dots, \widetilde{X}_n$, defined from the sample $X_1, \dots, X_n$ by*

$$\widetilde{X}_i = A_j \qquad \text{if } X_i \in A_j,$$

are distributed according to the probability measure $\widetilde{P}_0$ on $\widetilde{E}$ defined by

$$\widetilde{P}_0(A_j) = P_0(\{x \in E : x \in A_j\}).$$

The same preliminary procedure can be applied with goodness-of-fit tests to a family of distributions, independence and homogeneity tests which will be seen in the sequel.

In any case, the result of the test will of course depend on the choice of the partition. In practice, the latter should be chosen so that under P_0, all classes $A_1, \dots, A_m$ have approximately the same probability $1/m$; in this case, following Remark 10.2.4, the number m of classes must be chosen such that

$$n\frac{1}{m}\left(1-\frac{1}{m}\right) \geq 5.$$

This procedure is particularly adapted to countably infinite state spaces, in which case P_0 remains represented by its probability mass function $(p_{0,x})_{x\in E}$ and

$$\widetilde{P}_0(A_j) = \sum_{x\in A_j} p_{0,x}.$$

In fact, it is also adapted to the case where the condition of Remark 10.2.4 ensuring the validity of the χ_2 approximation in finite state spaces is not satisfied because p_x takes too small values. Then, it is a common practice to aggregate the classes for which p_x is small into larger classes.

However, for continuous probability measures on $\mathbb{R}$, nonparametric tests such as those introduced in Lecture 11 should be preferred.

10.3 χ_2 test of goodness-of-fit to a family of distributions

Let $\mathcal{P}_0$ be a subset of the set of probability measures on E. We are now interested in the null and alternative hypotheses

$$H_0 = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}.$$

Example 10.3.1 (Binomial model)**.** *In a group of N students, the number of students missing the i-th session of the course is denoted by X_i, $i = 1, \dots, n$, where n is the total number of sessions. If each student chooses to skip a class independently of each other, the variables X_i should follow a binomial distribution $\mathcal{B}(N, p)$ where p is unknown. To test whether this independence property holds (null hypothesis), or whether there is a contagion effect in absenteeism (alternative hypothesis), one may set $\mathcal{P}_0 = \{\mathcal{B}(N, p), p \in [0, 1]\}$.*

Notice that if $\mathcal{P}_0 = \{P_0\}$ is a singleton, then we are in the framework of the previous Section. If this is not the case, then Pearson's statistic is ill-defined as one does not know a priori which of the measures $P_0 \in \mathcal{P}_0$ should be compared with $\widehat{P}_n$. In order to circumvent this issue, we shall assume that $\mathcal{P}_0$ is a parametric family with low dimension, which writes

$$\mathcal{P}_0 = \{P_{0,\theta}, \theta \in \Theta\},$$

where $\Theta \subset \mathbb{R}^q$ with $q < m$. Example 10.3.1 fulfills this assumption, with $q = 1$ (the unknown parameter p has dimension 1) and $m = N + 1$ (a $\mathcal{B}(N, p)$ variable can take the $N + 1$ values $0, \dots, N$).

Proposition 10.3.2 (Pearson's statistic for the goodness-of-fit to a family of distributions)**.** *Assume that Θ has a nonempty interior in $\mathbb{R}^q$, and that the model $\{P_{0,\theta}, \theta \in \Theta\}$ is identifiable[3]. Let $\widehat{\theta}_n$ be a consistent estimator of θ under H_0. Consider the statistic*

$$d'_n = n\chi_2(\widehat{P}_n | P_{0,\widehat{\theta}_n}) = n \sum_{x\in E} \frac{(\widehat{p}_{n,x} - p_{0,\widehat{\theta}_n,x})^2}{p_{0,\widehat{\theta}_n,x}}.$$

Under H_0, d'_n converges in distribution to $\chi_2(m - q - 1)$, while under H_1, $d'_n \to +\infty$ almost surely.

We omit the proof of Proposition 10.3.2 (and refer for example to [7, Section 17.5]) but insist on the fact that the number of degrees of freedom of the limiting χ_2 distribution under H_0 **is not the same** as in the case of a simple null hypothesis: the larger the dimension q of the parameter set Θ, the lower the number of degrees of freedom!

Corollary 10.3.3 (χ_2 goodness-of-fit test for a family of distributions)**.** *The test with rejection region*

$$W_n = \{d'_n \geq \chi^2_{m-q-1,1-\alpha}\}$$

is consistent and has asymptotic level α.

[3] See Subsection 2.2.1.

To know whether n is large enough for the χ_2 approximation to be valid, the adaptation of the criterion of Remark 10.2.4 writes

$$\forall x \in E, \qquad np_{0,\widehat{\theta}_n,x}\left(1 - p_{0,\widehat{\theta}_n,x}\right) \geq 5.$$

Exercise 10.3.4 (Continuation of Example 10.3.1)**.** The number of students missing each session of the course during the year 2017/2018 is reported below, for a group of $N = 20$ students.

Number of absent students for each of the 12 sessions[4]: $3, 3, 0, 4, 3, 0, 1, 1, 0, 1, 0, 2$.

Was there a contagion effect in absenteeism?

The `scipy.stats` function `chisquare()` introduced in the previous Section has an argument `ddof`, for *delta degrees of freedom*, which allows you to apply the χ_2 test by modifying the number of degrees of freedom: if `f_obs` and `f_exp` are two arrays of size m respectively containing the quantities $(n\widehat{p}_{n,x})_{x\in E}$ and $(np_{0,\widehat{\theta}_n,x})_{x\in E}$, then the test with $m - q - 1$ degrees of freedom is performed using the command `chisquare(f_obs, f_exp, ddof=q)`.

10.4 The χ_2 test of independence

In this Section, we assume that we observe independent realisations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of pairs of random variables taking their values in the product space $E \times F$, where E and F are finite spaces, with respective cardinality m and l. In other words, for each experiment $i \in \{1, \ldots, n\}$, two features $X_i \in E$ and $Y_i \in F$ are collected. A natural question is whether these features are independent or not.

Example 10.4.1 (Is your second child more likely to be a boy when the first one is a boy?)**.** *Over a population of $n = 5983$ families with two children, the gender of both children is recorded*[5]*. The results are represented in the* contingency table *of Table 10.1. We want to check whether the gender of the first child has an influence on the gender of the second child.*

	Second child: male	Second child: female	Total
First child: male	1686	1444	3130
First child: female	1428	1435	2853
Total	3114	2869	5983

Table 10.1: Contingency table for the study of Example 10.4.1. The cells contain the numbers of individuals with the corresponding features.

The distribution of the pair (X_1, Y_1) is denoted by P. It is a probability measure on the finite space $E \times F$, with probability mass function $(p_{x,y})_{(x,y)\in E\times F}$ defined by

$$\forall (x, y) \in E \times F, \qquad p_{x,y} = \mathbb{P}(X_1 = x, Y_1 = y).$$

The marginal distribution of X_1 and Y_1 are respectively denoted by P^X and P^Y, and their probability mass functions $(p^X_x)_{x\in E}$ and $(p^Y_y)_{y\in F}$ satisfy

$$p^X_x = \mathbb{P}(X_1 = x) = \sum_{y\in F} \mathbb{P}(X_1 = x, Y_1 = y) = \sum_{y\in F} p_{x,y},$$
$$p^Y_y = \mathbb{P}(Y_1 = y) = \sum_{x\in E} \mathbb{P}(X_1 = x, Y_1 = y) = \sum_{x\in E} p_{x,y}.$$

[4] Do not take this as a challenge to beat...

[5] Source: Table 2 in M. E. Bernstein. Studies in the human sex ratio: 2. The proportion of unisexual sibships. *Human Biology*, 1952. To know more, see the discussion in https://www.biorxiv.org/content/10.1101/031344v3.

The variables X_1 and Y_1 are independent if and only if

$$\forall (x,y) \in E \times F, \qquad p_{x,y} = \mathbb{P}(X_1 = x, Y_1 = y) = \mathbb{P}(X_1 = x)\mathbb{P}(Y_1 = y) = p_x^X p_y^Y. \tag{10.1}$$

We denote by $\mathcal{P}_0$ the set of all probability measures on $E \times F$ satisfying the condition (10.1).

Lemma 10.4.2 (Parametrisation of $\mathcal{P}_0$)**.** *The set $\mathcal{P}_0$ is parametrised by a subset Θ of $\mathbb{R}^{(m-1)+(l-1)}$ with nonempty interior, and this model is identifiable.*

Proof. Assume that the elements of E and F are labeled: $E = \{x_1, \dots, x_m\}$ and $F = \{y_1, \dots, y_l\}$. Set

$$\Theta := \Theta^X \times \Theta^Y,$$

with

$$\Theta^X := \{(p_i)_{1\le i\le m-1} : p_i \ge 0, p_1 + \ldots + p_{m-1} \le 1\},$$
$$\Theta^Y := \{(q_j)_{1\le j\le l-1} : q_j \ge 0, q_1 + \ldots + q_{l-1} \le 1\}.$$

The set Θ has non-empty interior in $\mathbb{R}^{(m-1)+(l-1)}$. Moreover, an element $(p_i)_{1\le i\le m-1} \in \Theta^X$ can be completed to define a probability mass function $(p_x)_{x\in E}$ on E by setting

$$p_x = \begin{cases} p_i & \text{if } x = x_i \text{ with } i \le m-1, \\ 1 - (p_1 + \ldots + p_{m-1}) & \text{if } x = x_m, \end{cases}$$

and similarly for $(q_j)_{1\le j\le l-1} \in \Theta^Y$. Now for any $\theta = ((p_i)_{1\le i\le m-1}, (q_j)_{1\le j\le l-1}) \in \Theta$, define the probability measure $P_{0,\theta} \in \mathcal{P}_0$ by its probability mass function

$$p_{0,\theta,x,y} = p_x q_y.$$

It is easily seen that the mapping $\theta \in \Theta \mapsto P_{0,\theta} \in \mathcal{P}_0$ is a bijection, which completes the proof. □

Let us now introduce the statistics

$$\widehat{p}_{n,x,y} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i=x, Y_i=y\}}, \qquad \widehat{p}^X_{n,x} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i=x\}}, \qquad \widehat{p}^Y_{n,y} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i=y\}},$$

which are respective estimators of $p_{x,y}$, p_x^X and p_y^Y. Let us denote by $\widehat{P}_n$ and $\widehat{P}_n^X \otimes \widehat{P}_n^Y$ the random probability measures on $E \times F$ with respective probability mass functions $(\widehat{p}_{n,x,y})_{(x,y)\in E\times F}$ and $(\widehat{p}^X_{n,x}\widehat{p}^Y_{n,y})_{(x,y)\in E\times F}$. With the notation of the proof of Lemma 10.4.2,

$$\widehat{P}_n^X \otimes \widehat{P}_n^Y = P_{0,\widehat{\theta}_n},$$

where

$$\widehat{\theta}_n = \left((\widehat{p}^X_{n,x_i})_{1\le i\le m-1}, (\widehat{p}^Y_{n,y_j})_{1\le j\le l-1}\right).$$

Under H_0, $\widehat{\theta}^n$ is a consistent estimator of θ. Therefore, Proposition 10.3.2 allows to construct a test of independence, for the hypotheses

$$H_0 = \{X \text{ and } Y \text{ are independent}\} = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}.$$

The test statistic writes

$$d'_n = n\chi_2(\widehat{P}_n | \widehat{P}_n^X \otimes \widehat{P}_n^Y) = n \sum_{x\in E, y\in F} \frac{(\widehat{p}_{n,x,y} - \widehat{p}^X_{n,x}\widehat{p}^Y_{n,y})^2}{\widehat{p}^X_{n,x}\widehat{p}^Y_{n,y}},$$

and under H_0, it converges in distribution to the χ_2 distribution with

$$\underbrace{(ml-1)}_{\text{cardinality of } E\times F - 1} - \underbrace{((m-1)+(l-1))}_{\text{dimension of } \mathcal{P}_0} = (m-1)(l-1),$$

which yields the following result.

Proposition 10.4.3 (χ_2 test of independence)**.** *The test rejecting H_0 as soon as*

$$d'_n \geq \chi^2_{(m-1)(l-1),1-\alpha}$$

is consistent and has asymptotic level α.

For this independence test, the adaptation of the criterion of Remark 10.2.4 writes

$$\forall (x,y) \in E \times F, \qquad n\widehat{p}^X_{n,x}\widehat{p}^Y_{n,y}\left(1-\widehat{p}^X_{n,x}\widehat{p}^Y_{n,y}\right) \geq 5.$$

In order to apply this test to the case of Example 10.4.1, we take $E = F = \{\mathrm{m}, \mathrm{f}\}$ to encode the gender of the first and second child, respectively, and compute the empirical frequencies in Table 10.2.

	Second child: male	Second child: female	Total
First child: male	$\widehat{p}_{(\mathrm{m},\mathrm{m})} = 0.282$	$\widehat{p}_{(\mathrm{m},\mathrm{f})} = 0.241$	$\widehat{p}^X_{\mathrm{m}} = 0.523$
First child: female	$\widehat{p}_{(\mathrm{f},\mathrm{m})} = 0.239$	$\widehat{p}_{(\mathrm{f},\mathrm{f})} = 0.238$	$\widehat{p}^X_{\mathrm{f}} = 0.477$
Total	$\widehat{p}^Y_{\mathrm{m}} = 0.520$	$\widehat{p}^Y_{\mathrm{f}} = 0.480$	1

Table 10.2: Empirical frequencies for the study of Example 10.4.1. For the sake of legibility, we omit the dependence upon $n = 5983$ in the notation of the empirical frequencies.

The value of the test statistic in this example is then

$$d'_{5983} = 5983 \times \left(\frac{(0.282 - 0.52 \times 0.523)^2}{0.52 \times 0.523} + \frac{(0.241 - 0.48 \times 0.523)^2}{0.48 \times 0.523} + \frac{(0.239 - 0.52 \times 0.477)^2}{0.52 \times 0.477} + \frac{(0.238 - 0.48 \times 0.477)^2}{0.48 \times 0.477} \right) \simeq 8.70.$$

The p-value of this observation is obtained by the command `1-chi2.cdf(8.7, df=1)` and approximately equal to $3.2\ 10^{-3}$. Therefore, at all usual levels, the independence assumption is *rejected*: the gender of the second child is not independent from the gender of the first child.

Exercise 10.4.4. Using the data from Example 10.4.1, test whether it is equally likely to have a boy or a girl.

To complete this Section, we point out the fact that this independence test can be used to test the *homogeneity* of $k \geq 2$ samples $\mathbf{X}_{j,n_j} = (X_{j,1}, \ldots, X_{j,n_j})$, $1 \leq j \leq k$, which take their values in E. We assume that in the j-th sample, the variables are iid according to some probability measure P_j on E, and we want to construct a test for the hypotheses

$$H_0 = \{P_1 = \cdots = P_k\}, \qquad H_1 = \{\text{all samples do not have the same law}\}.$$

Setting $n = n_1 + \cdots + n_k$ and defining the bivariate sample $(\mathbf{X}_n, \mathbf{Y}_n)$ in $E \times \{1, \ldots, k\}$ by

$$\begin{array}{lllll} \mathbf{X}_n & = (X_{1,1}, \ldots, X_{1,n_1}, & X_{2,1}, \ldots, X_{2,n_2}, & \ldots, & X_{k,1}, \ldots, X_{k,n_k}), \\ \mathbf{Y}_n & = (1, \ldots, 1, & 2, \ldots, 2, & \ldots, & k, \ldots, k), \end{array}$$

the null hypothesis is equivalent to stating that the variables X_i and Y_i are independent. Therefore, homogeneity can be tested using the χ_2 test of independence with the $E \times \{1, \ldots, k\}$ contingency table whose (x, j) cell is $\sum_{i=1}^{n_j} \mathbb{1}_{\{X_{j,i}=x\}}$.

Remark 10.4.5 (Link with Wald's test of identity of means)**.** *In the case where $E = \{0, 1\}$ and $k = 2$, Wald's test of identity of means, described in Subsection 9.1.2, also allows to test the homogeneity of the two samples. Is this test equivalent to the χ_2 test?*

Let us represent the data by the following contingency table, with $s_j = \sum_{i=1}^{n_j} X_{j,i}$, $j = 1, 2$.

	Sample 1	*Sample 2*
$x=0$	$n_1 - s_1$	$n_2 - s_2$
$x=1$	s_1	s_2

We set $\widehat{p}_{1,n_1} = s_1/n_1$ *and* $\widehat{p}_{2,n_2} = s_2/n_2$*, so that with the notation of Subsection 9.1.2,* $Z_{n_1,n_2} = \widehat{p}_{1,n_1} - \widehat{p}_{2,n_2}$*. A (rather tedious, but elementary) computation shows that the Pearson statistic of this test writes*

$$d'_n = \frac{(\widehat{p}_{1,n_1} - \widehat{p}_{2,n_2})^2}{\widehat{p}_n(1-\widehat{p}_n)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \qquad \widehat{p}_n := \frac{s_1+s_2}{n_1+n_2},$$

and the independence hypothesis is rejected as soon as $d'_n \geq \chi^2_{1,1-\alpha}$*. It then turns out that this is not exactly the same rejection region as Wald's test, but as a slight modification thereof.*

Indeed, remark that under H_0*, the concatenation of both samples* $\mathbf{X}_{1,n_1}$ *and* $\mathbf{X}_{2,n_2}$ *yields a sample of* $n = n_1 + n_2$ *independent Bernoulli variables with common parameter* p*, which can be estimated by* $\widehat{p}_n$ *introduced above, so that*

$$\mathrm{Var}(Z_{n_1,n_2}) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$$

can be estimated by $\widehat{p}_n(1-\widehat{p}_n)(1/n_1 + 1/n_2)$*. Therefore, another*[6] *consistent test with asymptotic level* α *is obtained by rejecting* H_0 *as soon as*

$$\frac{|\widehat{p}_{1,n_1} - \widehat{p}_{2,n_2}|}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \geq \phi_{1-\alpha/2}.$$

Since it is easily checked that $\chi^2_{1,1-\alpha} = \phi^2_{1-\alpha/2}$*, we conclude that the latter test actually coincides with the* χ_2 *test of independence.*

10.A Exercises

Training exercises

Exercise 10.A.1. In the setting of Section 10.1, show that $\widehat{P}_n$ is the MLE of P.

Exercise 10.A.2. During 300 minutes, the number of clients entering a shop per minute is recorded. The results are reported below:

Number of clients	0	1	2	3	4	5
Number of minutes	23	75	68	51	53	30

Does the number of clients entering the shop per minute follow a Poisson distribution? Following Remark 10.2.6, you may start by partitioning the state space $\mathbb{N}$ for the Poisson distribution into seven classes $\{0\}, \{1\}, \ldots, \{5\}, \{6, 7, \ldots\}$. Then you may use the command `1-chi2.cdf(x, df=n)` to compute the probability that a $\chi_2(n)$-distributed variable takes values larger than x, and thus return a p-value.

Homework

Exercise 10.A.3 (Weldon's dice)**.** This is a true story: in 1894, English evolutionary biologist Walter Weldon rolled a set of 12 dice $n = 26306$ times, and for each roll, recorded the number X of dice showing 5 or 6 points. The data[7] are contained in the notebook `Weldon.ipynb` available on Educnet.

[6] The difference between this test and Wald's test from Proposition 9.1.7 is similar to the difference between the tests with rejection regions W_n and W'_n in Exercise 7.1.22 for the one-sample Bernoulli model.

[7] A. W. Kemp and C.D. Kemp. Weldon's dice data revisited, *The American Statistician*, 1991.

The null hypothesis to be tested here is that the dice are independent, identical and perfectly equilibrated, so that you should have no difficulty to figure out that

$$H_0 = \{P = \mathcal{B}(12, 1/3)\}, \qquad H_1 = \{P \neq \mathcal{B}(12, 1/3)\},$$

with P the law of X. On Figure 10.1, we superpose the histogram of the data with the probability mass function of X under H_0.

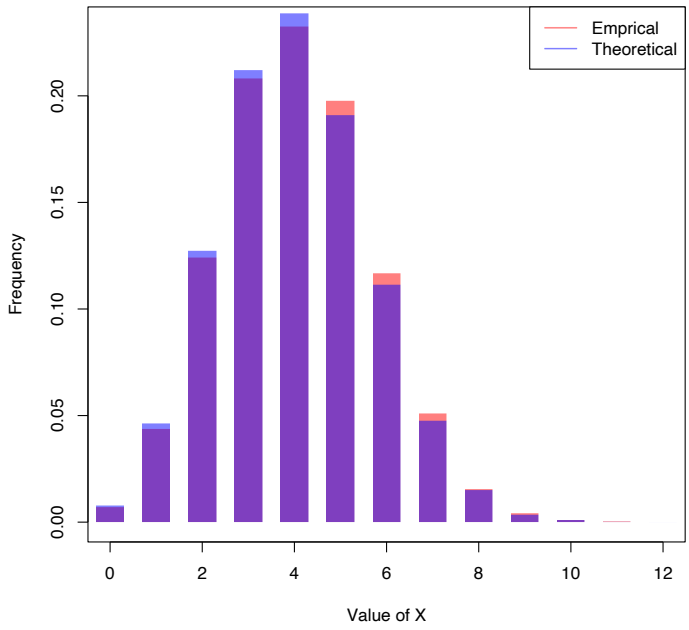


Figure 10.1: Comparison of empirical and theoretical frequencies for Weldon's experiment.

1. Compute the p-value of Weldon's observation. You may refer to Section 10.2 for the use of the function `chisquare()`.
2. Compute a 95% asymptotic confidence interval for $\mathbb{E}[X]/12$, and compare the bounds of the interval with what you expect $\mathbb{E}[X]/12$ to be under H_0.
3. How do you explain these results?
4. Even if the dice are biased, one may still want to test whether they are independent and identically distributed (but not necessarily equilibrated).
   (a) Write the corresponding hypotheses under the form of a goodness-of-fit test to a family of distributions.
   (b) Compute the associated p-value. You may use the parameter `ddof` in the function `chisquare()` to adjust the number of degrees of freedom of the test.

Weldon's experiment was used in 1900 by Pearson to construct of the χ_2 goodness-of-fit test.

Lecture 11

Nonparametric Tests for Continuous Data

Contents

This Lecture addresses goodness-of-fit and two-sample homogeneity tests in the case where the space E in which the data $X_1, \dots, X_n$ take their values is $\mathbb{R}$. As in Lecture 10, we do not assume that the law P of the data belongs to a predetermined parametric family $\mathcal{P}$, and the overall idea consists in approximating the law P by the empirical measure $\widehat{P}_n$ of the sample. For regularity reasons, it is more convenient to work with Cumulative Distribution Functions (CDFs) than with probability measures, therefore we denote by F and $\widehat{F}_n$ the respective CDFs of P and $\widehat{P}_n$. In particular, it is easily checked that

$$\forall x \in \mathbb{R}, \qquad \widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

The empirical CDF of a sample $\mathbf{X}_n$ is plotted on Figure 11.1, together with the actual CDF F of the sample.

11.1 Asymptotic Kolmogorov test

In this section we fix a law P_0 on $\mathbb{R}$ with CDF F_0 and aim at constructing an asymptotic test for the hypotheses

$$H_0 = \{P = P_0\} = \{F = F_0\}, \qquad H_1 = \{P \neq P_0\} = \{F \neq F_0\}.$$

Since one may naturally expect F to be approximated by $\widehat{F}_n$ when $n \to +\infty$, the construction of the test requires to describe the asymptotic distribution of the distance between $\widehat{F}_n$ and F. This is the object of Subsection 11.1.1. The construction of the test is next described in Subsection 11.1.2.

11.1.1 Limit theorems for the empirical CDF

As is intuitively expected, if $X_1, \dots, X_n$ are iid random variables in $\mathbb{R}$ with common CDF F, then $\widehat{F}_n$ approximates F when the size of the sample increases. A first justification of this approximation is that, for a fixed $x \in \mathbb{R}$, the strong Law of Large Numbers yields

$$\lim_{n \to +\infty} \widehat{F}_n(x) = \mathbb{E}[\mathbb{1}_{\{X_1 \leq x\}}] = F(x), \qquad \text{almost surely.} \tag{11.1}$$

This result is strengthened in the next theorem.

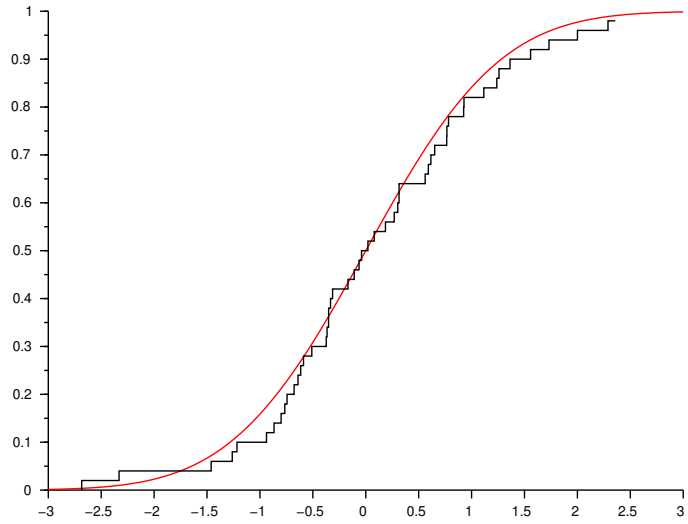


Figure 11.1: Plot of the empirical CDF $\widehat{F}_n$ of a sample, superposed with its actual CDF F.

Theorem 11.1.1 (Glivenko–Cantelli Theorem)**.** *Let F be a CDF on $\mathbb{R}$ and $(X_i)_{i\geq 1}$ be a family of independent random variables with CDF F. We have*

$$\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| = 0, \qquad \textit{almost surely.}$$

Before starting the proof, the reader should wonder in which sense the statement of the Glivenko–Cantelli is stronger than the convergence result (11.1) asserted above.

Proof. For any $x \in \mathbb{R}$, we denote by $\widehat{F}_n(x^-)$ and $F(x^-)$ the respective left limits of $\widehat{F}_n$ and F at x; since these functions are right continuous, there is no need to introduce a notation for the right limits. By the strong Law of Large Numbers, for all $x \in \mathbb{R}$, in addition to (11.1) we also get

$$\lim_{n\to+\infty} \widehat{F}_n(x^-) = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i<x\}} = \mathbb{E}[\mathbb{1}_{\{X_1<x\}}] = F(x^-), \qquad \text{almost surely.} \tag{11.2}$$

Let $\epsilon > 0$. Since F is nondecreasing and bounded, there is only a finite number of points $x \in \mathbb{R}$ such that $F(x) - F(x^-) > \epsilon$. Thus, there exist $k \geq 1$ and $-\infty = x_0 < x_1 < \cdots < x_k = +\infty$ such that $F(x_\ell^-) - F(x_{\ell-1}) \leq \epsilon$ for all $\ell \in \{1, \ldots, k\}$. Therefore, using the fact that $\widehat{F}_n$ and F are nondecreasing, for all $x \in [x_{\ell-1}, x_\ell)$, it holds

$$\widehat{F}_n(x) \leq \widehat{F}_n(x_\ell^-) \qquad \text{and} \qquad F(x) \geq F(x_{\ell-1}) \geq F(x_\ell^-) - \epsilon,$$

so that

$$\widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_\ell^-) - F(x_\ell^-) + \epsilon.$$

With similar arguments, we get

$$\widehat{F}_n(x_{\ell-1}) - F(x_{\ell-1}) - \epsilon \leq \widehat{F}_n(x) - F(x).$$

By (11.1) and (11.2) for $x = x_0, \ldots, x_k$, we deduce that almost surely,

$$\limsup_{n\to+\infty} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \leq \epsilon,$$

which completes the proof since the left-hand side does not depend on ϵ. □

It should be clear that the Glivenko–Cantelli Theorem is a result of the same nature as the strong Law of Large Numbers, stated in a functional space. It is accompanied by a Central Limit Theorem, which relies on a random process[1] $(\beta(t))_{t\in[0,1]}$ called the *Brownian bridge*.

Definition 11.1.2 (Brownian bridge)**.** *A* Brownian bridge *is a random process* $(\beta(t))_{t\in[0,1]}$ *such that:*

(i) almost surely, the mapping $t \mapsto \beta(t)$ *is continuous on* $[0,1]$*, and* $\beta(0) = \beta(1) = 0$*;*

(ii) for any $d \geq 1$ *and* $t_1, \ldots, t_d \in [0,1]$*, the random vector* $(\beta(t_1), \ldots, \beta(t_d))$ *is Gaussian, centered, with covariance matrix given by* $\mathrm{Cov}(\beta(t_i), \beta(t_j)) = \min\{t_i, t_j\} - t_i t_j$*.*

At a heuristic level, a Brownian bridge must be understood as 'a Brownian motion on $[0,1]$ conditioned to take the value 0 at the points 0 and 1', see Figure 11.2. Since β is almost surely continuous on $[0,1]$, the random variable

$$\beta_\infty := \sup_{t\in[0,1]} |\beta(t)|$$

is well-defined, and its CDF is known to be given by the formula[2]

$$\forall y > 0, \qquad \mathbb{P}(\beta_\infty \leq y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp\left(-2k^2y^2\right).$$

This identity allows in particular to compute quantiles of β_∞, which shall be denoted by $b_{\infty,r}$.

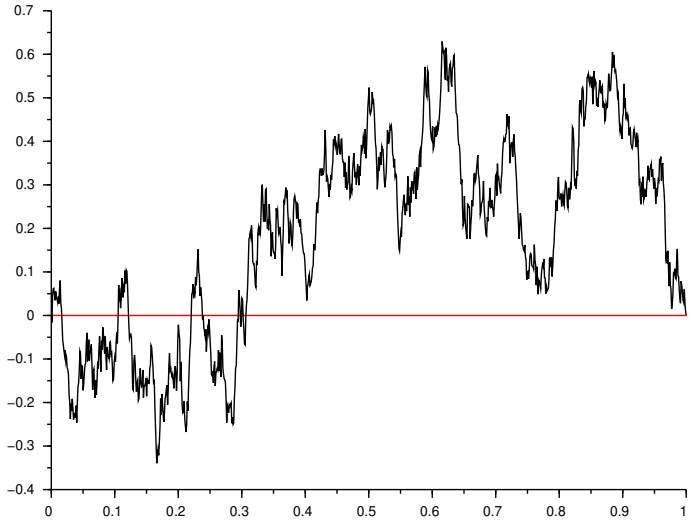


Figure 11.2: A realisation of a Brownian bridge.

Theorem 11.1.3 (Donsker Theorem)**.** *Let* F *be a CDF on* $\mathbb{R}$ *and* $(X_i)_{i\geq 1}$ *be a family of independent random variables with CDF* F*. The random function*

$$G_n : x \in \mathbb{R} \mapsto \sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)$$

[1] That is to say, a random variable in the space of functions.

[2] A. N. Kolmogorov, Sulla Determinazione Empirica di una Legge di Distribuzione, *Giornale dell'Istituto Italiano degli Attuari*, 1933.

converges in distribution (in an appropriate functional space[3]*) to the random function*

$$G : x \in \mathbb{R} \mapsto \beta(F(x)),$$

where $\beta : [0,1] \to \mathbb{R}$ *is a Brownian bridge.*

Donsker's Theorem is presented because of its importance in nonparametric statistics, but we shall not elaborate on its contents (the interested reader may have a look at Exercise 11.A.1). Still, an important consequence of Donsker's Theorem is the following asymptotic behaviour of the supremum of G_n.

Corollary 11.1.4 (Asymptotic of the supremum)**.** *Under the assumptions of Theorem 11.1.3, the random variable*

$$g_n = \sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right|$$

converges in distribution to the random variable

$$g_\infty = \sup_{x \in \mathbb{R}} |\beta(F(x))|.$$

When F *is* continuous*, one then has*

$$g_\infty = \sup_{t \in [0,1]} |\beta(t)| = \beta_\infty.$$

11.1.2 The asymptotic Kolmogorov test

We now fix a probability measure P_0 on $\mathbb{R}$, with CDF F_0, and address the test of the hypotheses

$$H_0 = \{P = P_0\} = \{F = F_0\}, \qquad H_1 = \{P \neq P_0\} = \{F \neq F_0\}.$$

Definition 11.1.5 (Kolmogorov's statistic)**.** *The Kolmogorov statistic is the statistic*

$$\zeta_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|.$$

The Glivenko–Cantelli and Donsker Theorems then imply that:

- under H_0, $\zeta_n \to 0$ almost surely and $\sqrt{n}\zeta_n \to \sup_{x \in \mathbb{R}} |\beta(F_0(x))|$ in distribution;
- under H_1, $\zeta_n \to \sup_{x \in \mathbb{R}} |F(x) - F_0(x)| > 0$, almost surely and therefore $\sqrt{n}\zeta_n \to +\infty$.

We deduce the following statement.

Corollary 11.1.6 (Asymptotic Kolmogorov test)**.** *If* P_0 *has a continuous CDF* F_0 *on* $\mathbb{R}$*, the test with rejection region*

$$W_n = \{\sqrt{n}\zeta_n \geq b_{\infty,1-\alpha}\},$$

with $b_{\infty,r}$ *the quantile of order* r *of the random variable* β_∞ *defined in Subsection 11.1.1, is consistent and has asymptotic level* α*.*

Remark 11.1.7 (Computation of the Kolmogorov statistic)**.** *To implement Kolmogorov's test, one needs to compute the value of the statistic* ζ_n*, which a priori requires to evaluate* $F_0(x)$ *at all points* $x \in \mathbb{R}$*. However, the monotonicity of* F_0 *together with the fact that* $\widehat{F}_n$ *is piecewise constant show that the supremum is necessarily reached at one of the* n *points* $X_1, \ldots, X_n$ *(see Figure 11.3), so that if* F_0 *is continuous,*

$$\zeta_n = \max_{1 \leq k \leq n} \max \left\{ \left| \frac{k-1}{n} - F_0(X_{(k)}) \right|, \left| \frac{k}{n} - F_0(X_{(k)}) \right| \right\},$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ *denotes the nondecreasing reordering of* $X_1, \ldots, X_n$*. Therefore, only* n *evaluations of* F_0 *are necessary.*

[3] For example, in the linear space of bounded right-continuous and left-limited functions, endowed with the norm $\sup_{x \in \mathbb{R}} |g(x)|$.

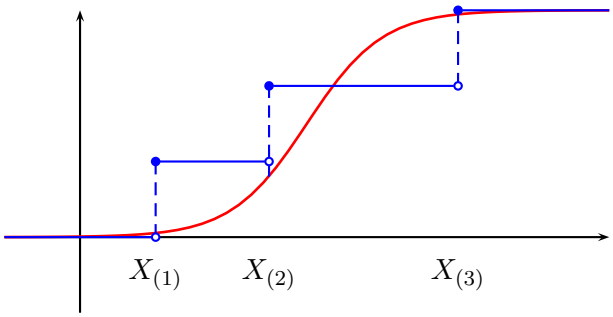


Figure 11.3: The maximum of $|\widehat{F}_n - F|$ is reached at the point $X_{(2)}$.

Exercise 11.1.8 (Cramér–von Mises test)**.** As an alternative to the Kolmogorov test, the Cramér–von Mises test is based on the statistic

$$\xi_n = \int_{x\in\mathbb{R}} \left(\widehat{F}_n(x) - F_0(x)\right)^2 F_0'(x)\mathrm{d}x,$$

where F_0 is assumed to be C^1 and such that $F_0' > 0$. This statistic is another measure of the distance between $\widehat{F}_n$ and F_0. Thanks to heuristic computations based on the Donsker Theorem[4], describe the region of rejection of this test.

11.2 The nonasymptotic Kolmogorov test

Unlike the case of Pearson's statistic for finite state space models, the Kolmogorov statistic enjoys a peculiar property allowing to derive nonasymptotic tests: it is *free*[5] under H_0. This makes it possible to construct nonasymptotic tests.

11.2.1 Freeness of Kolmogorov's statistic

The purpose of this subsection is to prove the following statement.

Proposition 11.2.1 (Freeness of Kolmogorov's statistic)**.** *In the setting of Subsection 11.1.2, if P_0 has a continuous CDF F_0 on $\mathbb{R}$, the law of Kolmogorov's statistic ζ_n under H_0 only depends on n and not on F_0.*

The proof of Proposition 11.2.1 first relies on the notion of *pseudo-inverse* of a CDF.

Definition 11.2.2 (Pseudo-inverse)**.** *Let $F : \mathbb{R} \to [0,1]$ be a CDF. The* pseudo-inverse $F^{-1} : [0,1] \to [-\infty, +\infty]$ *is defined by*

$$\forall u \in [0,1], \qquad F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\},$$

where we take the conventions that $\inf \mathbb{R} = -\infty$ *and* $\inf \varnothing = +\infty$.

Exercise 11.2.3 (Some precautions to be taken)**.** Construct:

1. a CDF F such that there exists $u \in (0,1)$ for which $F(F^{-1}(u)) \neq u$,
2. a CDF F such that there exists $x \in \mathbb{R}$ for which $F^{-1}(F(x)) \neq x$.

Lemma 11.2.4 (Properties of the pseudo-inverse)**.** *Let $F : \mathbb{R} \to [0,1]$ be a CDF.*

[4] You may admit that the mapping $g \mapsto \int g(x)^2 F_0'(x)\mathrm{d}x$ is continuous for the topology on which the convergence in distribution of Donsker's Theorem is stated.

[5] We recall the Definition 4.2.1 in Lecture 4 of the freeness of a statistic in the parametric context.

(i) For all $u \in (0,1)$, $x \in \mathbb{R}$, $F^{-1}(u) \leq x$ if and only if $u \leq F(x)$.

(ii) Let U be a uniform random variable on $[0,1]$. Then F is the CDF of the random variable $F^{-1}(U)$.

Proof. Since F is right-continuous, for any $u \in]0,1[$, the set $\{x \in \mathbb{R} : F(x) \geq u\}$ is closed, so that $F(F^{-1}(u)) \geq u$. Since F is nondecreasing, we deduce that if $F^{-1}(u) \leq x$, then $u \leq F(F^{-1}(u)) \leq F(x)$. Reciprocally, if $u \leq F(x)$, then by the definition of F^{-1}, $F^{-1}(u) \leq x$: this proves the first point.

We check the second point by writing

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = \int_{u=0}^{F(x)} \mathrm{d}u = F(x),$$

where the first identity follows from the first part of the lemma. □

We may now detail the proof of Proposition 11.2.1.

Proof of Proposition 11.2.1. Let $U_1, \dots, U_n$ be independent uniform random variables on $[0,1]$. By Lemma 11.2.4, under H_0, the vectors $(X_1, \dots, X_n)$ and $(F_0^{-1}(U_1), \dots, F_0^{-1}(U_n))$ have the same law. Therefore, under H_0, ζ_n has the same law as

$$\begin{aligned} Z_n := \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{F_0^{-1}(U_i) \leq x\}} - F_0(x) \right| &= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq F_0(x)\}} - F_0(x) \right| \\ &= \sup_{u \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq u\}} - u \right|, \end{aligned}$$

where we have set $u = F_0(x)$ and used the continuity of F_0 to ensure that $F_0(x)$ takes all values $u \in (0,1)$ in the second inequality. The law of the right-hand side does not depend on F_0, which is the announced statement. □

Proposition 11.2.1 motivates the following definition.

Definition 11.2.5 (Kolmogorov's law)**.** *Let $(U_i)_{i \geq 1}$ be a sequence of independent random variables uniformly distributed on $[0,1]$. For all $n \geq 1$, the law of the random variable*

$$Z_n = \sup_{u \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq u\}} - u \right|$$

is called the Kolmogorov law *with parameter n.*

11.2.2 Goodness-of-fit test

An immediate consequence of Proposition 11.2.1 is the construction of the following nonasymptotic test for the hypotheses

$$H_0 = \{P = P_0\}, \qquad H_1 = \{P \neq P_0\}.$$

Corollary 11.2.6 (Nonasymptotic Kolmogorov test)**.** *Under the assumptions of Proposition 11.2.1, the test with rejection region*

$$W_n = \{\zeta_n \geq z_{n,1-\alpha}\}$$

where $z_{n,r}$ is the quantile of order r of Kolmogorov's law with parameter n, has level α.

With `scipy.stats`, this test is performed with the command `ks_1samp()`.

Exercise 11.2.7. Using the Glivenko–Cantelli Theorem and Lemma 1.4.8, show that the test is consistent.

Exercise 11.2.8. Construct a nonasymptotic version of the Cramér–von Mises test from Exercise 11.1.8.

11.2.3 Goodness-of-fit to a family of distributions: the Lilliefors correction

Let $\mathcal{P}_0$ be a subset of the space of probability measures (with a continuous CDF) on $\mathbb{R}$. Similarly to the χ_2 test in Section 10.3, the Kolmogorov test may generally be adapted to the set of hypotheses

$$H_0 = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}.$$

We shall study the specific case (once again, similar to that of Section 10.3) where $\mathcal{P}_0$ is a *parametric family*, which thus writes

$$\mathcal{P}_0 = \{P_{0,\theta}, \theta \in \Theta\}, \qquad \text{with } \Theta \subset \mathbb{R}^q.$$

Following the lines of Section 10.3, a natural approach consists in:

(i) finding a consistent estimator $\widehat{\theta}_n$ of θ under H_0;

(ii) comparing the distance between the empirical CDF $\widehat{F}_n$ of the sample, and the CDF $F_{0,\widehat{\theta}_n}$ of the probability measure in $\mathcal{P}_0$ corresponding to the estimated value $\widehat{\theta}_n$ of θ.

In some cases, it may then be proved that the law of the statistic

$$\zeta'_n = \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F_{0,\widehat{\theta}_n}(x)\right|$$

is *free* under H_0, that is to say that it depends on n (and on the model $\mathcal{P}_0$), but non on the underlying value θ of the parameter. In order to check this property, it is useful to mimic the proof of Proposition 11.2.1 and to express both $\widehat{F}_n(x)$ and $\widehat{\theta}_n$ in terms of independent uniform random variables $U_1, \dots, U_n$, see Example 11.2.9 below.

Just like the estimation of θ by $\widehat{\theta}_n$ changes the number of degrees of freedom of the limiting χ_2 distribution in the context of Section 10.3 (see Proposition 10.3.2), **the law of ζ'_n under H_0 will generally not be the Kolmogorov law from Definition 11.2.5**, but a certain probability measure, depending on the model $\mathcal{P}_0$, whose quantiles $z'_{n,r}$ need to be computed, often by numerical simulation.

This procedure is called the *Lilliefors correction*. It was initially designed to test whether a sample is Gaussian or not[6]. Here, we detail the example of the Exponential model, and refer to Exercise 11.A.4 for the Gaussian model. Another popular Gaussian fit test, the Shapiro–Wilk test, is presented in Section 11.4.

Example 11.2.9 (The Lilliefors correction in the Exponential model)**.** *We observe positive random variables $X_1, \dots, X_n$ which are assumed to be iid under some probability measure P, and want to test whether these random variables are exponentially distributed. We therefore set*

$$\mathcal{P}_0 = \{\mathcal{E}(\lambda), \lambda > 0\}, \qquad H_0 = \{P \in \mathcal{P}_0\}, \quad H_1 = \{P \notin \mathcal{P}_0\}.$$

In other words, under H_0, there exists $\lambda > 0$ such that $P = \mathcal{E}(\lambda)$. Then a consistent estimator of λ is given by the MLE $\widehat{\lambda}_n = 1/\overline{X}_n$, so that, for any $x \geq 0$,

$$F_{0,\widehat{\lambda}_n}(x) = 1 - \exp\left(-\widehat{\lambda}_n x\right) = 1 - \exp\left(-\frac{x}{\overline{X}_n}\right),$$

and therefore

$$\zeta'_n := \sup_{x\geq 0} \left|\widehat{F}_n(x) - F_{0,\widehat{\lambda}_n}(x)\right| = \sup_{x\geq 0} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} - \left(1 - \exp\left(-\frac{x}{\overline{X}_n}\right)\right)\right|.$$

[6] H. W. Lilliefors. On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 1967.

Under H_0, the vector $(X_1, \ldots, X_n)$ has the same law as $(-\frac{1}{\lambda}\log(1-U_1), \ldots, -\frac{1}{\lambda}\log(1-U_n))$, with $U_1, \ldots, U_n$ independent uniform variables on $[0,1]$. Therefore, the law of ζ_n' under H_0 is the law of the random variable

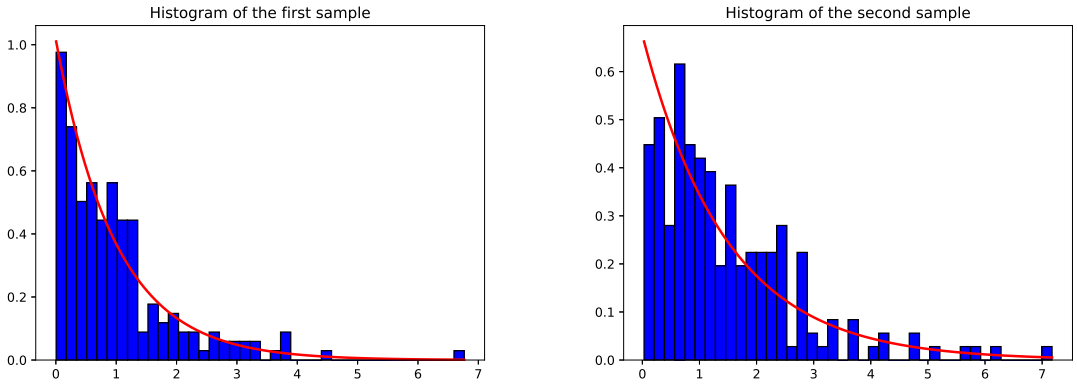
$$Z_n' = \sup_{x \geq 0} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{-\frac{1}{\lambda}\log(1-U_i) \leq x\}} - \left(1 - \exp\left(-\frac{x}{-\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda}\log(1-U_i)}\right)\right)\right|,$$

which rewrites, after the change of variable $x = -\frac{1}{\lambda}\log(1-u)$,

$$\begin{aligned} Z_n' &= \sup_{u \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{-\frac{1}{\lambda}\log(1-U_i) \leq -\frac{1}{\lambda}\log(1-u)\}} - \left(1 - \exp\left(-\frac{-\frac{1}{\lambda}\log(1-u)}{-\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda}\log(1-U_i)}\right)\right)\right| \\ &= \sup_{u \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq u\}} - \left(1 - \exp\left(-\frac{\log(1-u)}{\frac{1}{n}\sum_{i=1}^{n} \log(1-U_i)}\right)\right)\right|. \end{aligned}$$

The second expression no longer involves λ and only depends on the vector $(U_1, \ldots, U_n)$. Therefore, under H_0, ζ_n' is free and a test with level α is obtained by rejecting H_0 whenever ζ_n' is larger than the quantile or order $1-\alpha$ of Z_n'.

To apply this example, we consider two samples of $n = 200$ values each, whose histograms superposed with the density of the associated $\mathcal{E}(\widehat{\lambda}_n)$ distribution are reported below.



For the realisation pictured in the graphs, the respective values of the statistic ζ_n' are 0.06207 and 0.10320. To conclude whether H_0 can be rejected or not for these values, we compute the associated p-values $\mathbb{P}(Z_n' \geq 0.06207)$ and $\mathbb{P}(Z_n' \geq 0.10320)$ through Monte Carlo approximation.

```
import numpy as np

# Function to sample from the law of Z'
def sample_from_z(n):
    u = np.random.uniform(0, 1, size=n)
    u_sort = np.sort(u)
    r = 1 / np.mean(-np.log(1 - u))
    m = 0
    for k in range(1, n + 1):
        f = 1 - np.exp(r * np.log(1 - u_sort[k - 1]))
        m = max(m, abs((k - 1) / n - f), abs(k / n - f))
    return m

# Monte Carlo simulations to compute p-values
Nsim = 10000  # Number of Monte-Carlo simulations

z1 = 0.06207  # Value of the statistic zeta' on the first sample
```

```
z2 = 0.10320  # Value of the statistic zeta' on the second sample

ns = 200  # Size of the samples

# Generate samples and compute the statistics
mysample = np.array([sample_from_z(ns) for _ in range(Nsim)])

# Calculate p-values
pval1 = np.mean(mysample >= z1)
pval2 = np.mean(mysample >= z2)

print("P-value for z1:", pval1)
print("P-value for z2:", pval2)
```

With this code, we obtain that the p-value associated with the first sample is 0.19 while the p-value associated with the second sample is 0.0025. Therefore, at the level 5%, H_0 is rejected for the second sample but not for the first sample.

Remark 11.2.10 (Consistency)**.** *To check that the test which rejects H_0 whenever ζ_n' is larger than the quantile or order $1-\alpha$ of Z_n', let us prove the following two statements:*

1. *under H_1, $\liminf_{n\to+\infty} \zeta_n' > 0$, almost surely;*
2. *the quantile of order $1-\alpha$ of Z_n' converges to 0.*

For the first point, one may write, under H_1,

$$\zeta_n' = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_{0,\widehat{\lambda}_n}(x)| \geq \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_{0,1/m}(x)| - \sup_{x\in\mathbb{R}} |F_{0,\widehat{\lambda}_n}(x) - F_{0,1/m}(x)|,$$

by the triangle inequality, where $m = \mathbb{E}[X]$. By the Glivenko–Cantelli Theorem,

$$\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_{0,1/m}(x)| = \sup_{x\in\mathbb{R}} |F(x) - F_{0,1/m}(x)| > 0,$$

where F is the CDF of X. On the other hand, by the strong Law of Large Numbers, $\widehat{\lambda}_n \to 1/m$ and therefore by Dini's Theorem[7]*,*

$$\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} |F_{0,\widehat{\lambda}_n}(x) - F_{0,1/m}(x)| = 0.$$

The same arguments allow to show the second point: one starts by writing

$$\begin{aligned} Z_n' &= \sup_{u\in(0,1)} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{U_i\leq u\}} - \left(1 - \exp\left(-\frac{\log(1-u)}{\frac{1}{n}\sum_{i=1}^n \log(1-U_i)}\right)\right)\right| \\ &\leq \sup_{u\in(0,1)} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{U_i\leq u\}} - u\right| + \sup_{u\in(0,1)} \left| u - \left(1 - \exp\left(-\frac{\log(1-u)}{\frac{1}{n}\sum_{i=1}^n \log(1-U_i)}\right)\right)\right|. \end{aligned}$$

By the Glivenko–Cantelli Theorem, the first term in the right-hand side converges to 0, while the convergence to 0 of the second term follows from the fact that, by the strong Law of Large Numbers,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \log(1-U_i) = 1,$$

combined with Dini's Theorem.

[7] A sequence of CDFs which converges pointwise to a continuous limit converges uniformly on $\mathbb{R}$.

11.3 The Kolmogorov–Smirnov test of homogeneity

In this Section, we consider nonparametric tests of homogeneity, aiming at checking whether two independent samples $\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1})$ and $\mathbf{X}_{2,n_2} = (X_{2,1}, \ldots, X_{2,n_2})$ have the same distribution, without making any parametric assumption on this distribution.

Example 11.3.1 (Efficiency of a vaccine)**.** *In order to study the efficiency of a vaccine, a group of* 200 *people is split into two groups of* $n_1 = n_2 = 100$ *people. The first group is treated with the vaccine while the other group receives a placebo. One week after, the concentration of antibodies in the patients' blood is measured for both groups. The corresponding histograms are plotted on Figure 11.4. Clearly,*

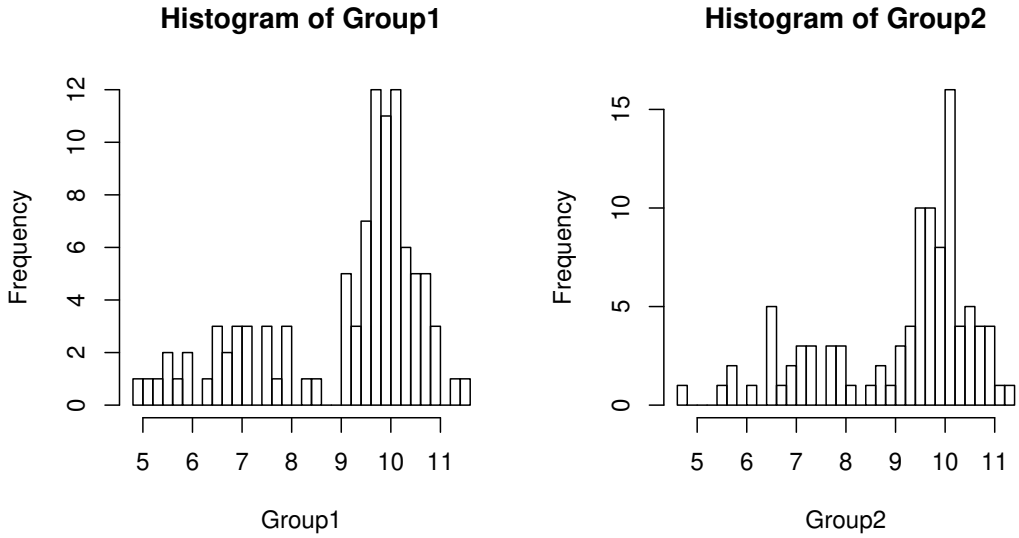


Figure 11.4: Concentration in antibodies for both groups.

the distribution of the concentration in antibodies is not Gaussian. Therefore Student's and Fisher's tests from Lecture 8 are not appropriate, and a nonparametric homogeneity test must be employed.

We first present the QQ-plot, which is a heuristic tool allowing to assess the closeness of two empirical distributions. We then describe the principle of the Kolmogorov–Smirnov test, which relies on ideas which are similar to Kolmogorov one-sample tests.

11.3.1 QQ-plot

Consider two real-valued random variables X and Y, with respective CDFs F and G. The QQ-plot (for *Quantile-Quantile*) of F and G is the parametric curve $u \in (0,1) \mapsto (F^{-1}(u), G^{-1}(u))$, where F^{-1} and G^{-1} are the respective pseudo-inverses of F and G. It provides a visual representation of 'how different' F and G are. In particular, it is supported by the diagonal $\{x = y\}$ if and only if $F = G$, and more generally, the QQ-plot of F and G has equation $y = ax + b$ (with $a > 0$) if and only if Y has the same law as $aX + b$.

QQ-plots are also useful to study empirical samples:

- for the goodness-of-fit test for a sample $X_1, \ldots, X_n$ and null hypothesis $\{F = F_0\}$, the QQ-plot of F_0 and $\widehat{F}_n$ allows to visually determine whether both distributions are close or not (see Remark 11.3.2 below);
- for two-sample tests of homogeneity, the QQ-plot of the corresponding empirical CDFs is also informative.

Examples of applications are given on Figure 11.5.

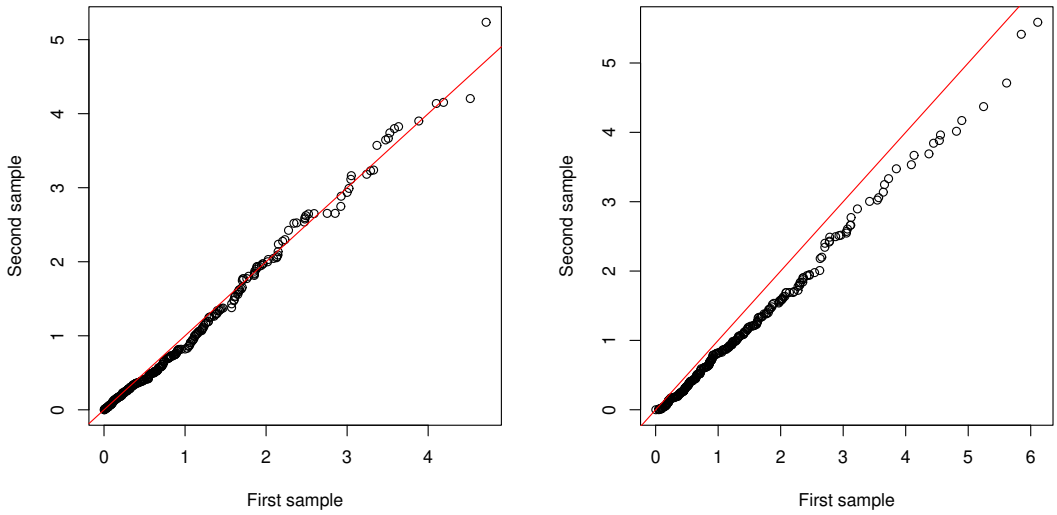


Figure 11.5: Two-sample QQ-plots. On the left-hand figure, the two samples are $\mathcal{E}(1)$-distributed; on the right-hand figure, the samples have respective distribution $\mathcal{E}(1)$ and $\Gamma(1.3, 1)$. On both figures, the diagonal $x = y$ is added in red.

Remark 11.3.2 (QQ-plot of a sample and a continuous distribution)**.** *In practice, when one wants to draw the QQ-plot of a sample of size n and of a CDF F_0, one draws the points with coordinates $(X_{(i)}, F_0^{-1}(i/(n+1)))$ for $i \in \{1, \dots, n\}$, where $X_{(1)} \leq \cdots \leq X_{(n)}$ denotes the order statistic of $X_1, \dots, X_n$.*

11.3.2 Kolmogorov–Smirnov test

Let us denote by F_1 and F_2 the respective CDFs of the samples $\mathbf{X}_{1,n_1}$ and $\mathbf{X}_{2,n_2}$. The null and alternative hypotheses for homogeneity tests are

$$H_0 = \{F_1 = F_2\}, \qquad H_1 = \{F_1 \neq F_2\}.$$

The Kolmogorov–Smirnov test for these hypotheses is a variation of the Kolmogorov test studied in the previous section. It is based on the Kolmogorov–Smirnov statistic

$$\xi_{n_1,n_2} = \sup_{x \in \mathbb{R}} \left| \widehat{F}_{1,n_1}(x) - \widehat{F}_{2,n_2}(x) \right|,$$

which can be computed with similar arguments as those detailed in Remark 11.1.7. The test is based on the following result.

Lemma 11.3.3 (Freeness of the Kolmogorov–Smirnov statistic)**.** *Assume that F_1 and F_2 are continuous. Under H_0, the statistic ξ_{n_1,n_2} is free: its law only depends on n_1 and n_2.*

Exercise 11.3.4. Prove Lemma 11.3.3.

The law of ξ_{n_1,n_2} under H_0 is called the *Kolmogorov–Smirnov distribution with parameters n_1 and n_2*, its quantile of order r is denoted by $x_{n_1,n_2,r}$.

Corollary 11.3.5 (Kolmogorov–Smirnov test)**.** *The test rejecting H_0 when $\xi_{n_1,n_2} \geq x_{n_1,n_2,1-\alpha}$ is consistent and has level α.*

The application of this test to the data of Example 11.3.1, thanks to the command `ks_2samp()`, yields a p-value of 0.7, which allows not to reject H_0. Based on Donsker's Theorem, it is also possible to design an asymptotic version of the Kolmogorov–Smirnov test. Another popular nonparametric test of homogeneity is the *Mann–Whitney U-test* [7, Example 12.7, p. 66]. Wilcoxon's test, which addresses *matched* samples, is studied in Exercise 11.A.5.

11.4 The Shapiro–Wilk test

The Shapiro–Wilk test[8] is a popular alternative to Lilliefors' test to test whether a sample $\mathbf{X}_n = (X_1, \dots, X_n)$ is Gaussian. It is experimentally observed to have better power than other tests, such as Lilliefors or Anderson–Darling[9].

Let P denote the law of $X_1, \dots, X_n$. Our aim is to test the hypotheses

$$H_0 = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}, \qquad \mathcal{P}_0 = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

The intuition behind the test is based on the QQ-plot of the standard Gaussian distribution, with CDF Φ, and of the sample $\mathbf{X}_n$. For large samples, two instances of this plot are represented on Figure 11.6. Following the introduction of the QQ-plot made in the previous Section, under H_0 it should be a straight line, with intercept μ and slope σ (because then $X = \mu + \sigma G$ where G has CDF Φ), while under H_1 it should be *anything but* a straight line. So a natural idea to test whether the sample is Gaussian is to apply linear regression to the QQ-plot and to reject H_0 if the coefficient of determination R^2 (see Appendix A) is far enough from 1: indeed, it may be proved that

$$\lim_{n \to +\infty} R^2 \begin{cases} = 1 & \text{under } H_0, \\ < 1 & \text{under } H_1, \end{cases} \tag{11.3}$$

in probability. This idea is exactly that of the Shapiro–Francia test[10], which was published in 1972.

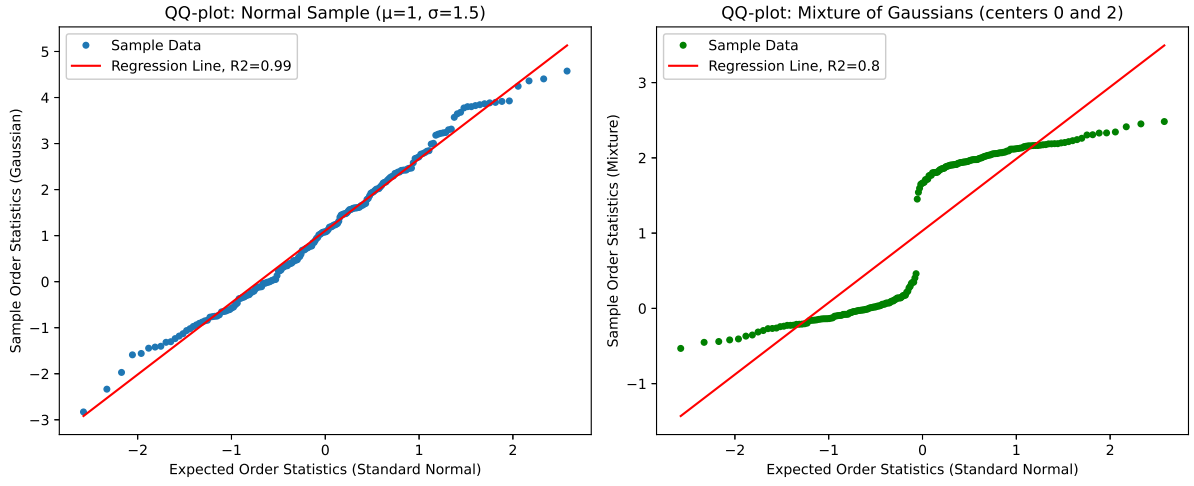


Figure 11.6: The QQ-plot of the standard Gaussian distribution and of the sample $\mathbf{X}_n$, together with the linear regression line, respectively under H_0 and H_1. Under H_0, $R^2 \simeq 1$ while under H_1, $R^2 < 1$.

[8] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 1965. The author M. B. Wilk must not be confused with S. S. Wilks, to whom we owe the Wilks' Theorem for the likelihood ratio test.

[9] N. Razali, Y. B. Wah. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, 2011.

[10] S. S. Shapiro and R. S. Francia. An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, 1972.

Remark 11.4.1 (A formal derivation of (11.3))**.** *By Remark 11.3.2, the QQ-plot of the standard Gaussian distribution and of the sample $\mathbf{X}_n$ represents the set of points with coordinates*

$$\left(\Phi^{-1}(\tfrac{i}{n+1}), X_{(i)}\right), \qquad i \in \{1, \dots, n\}.$$

Thus, noting that $X_{(i)} = \widehat{F}_n^{-1}(i/n)$ with $\widehat{F}_n$ the empirical CDF of the sample $\mathbf{X}_n$, we get

$$\begin{aligned} R^2 &= \mathrm{Corr}\left((\Phi^{-1}(\tfrac{i}{n+1}))_{1\le i\le n}, (\widehat{F}_n^{-1}(\tfrac{i}{n}))_{1\le i\le n}\right)^2 \\ &= \frac{\mathrm{Cov}\left((\Phi^{-1}(\tfrac{i}{n+1}))_{1\le i\le n}, (\widehat{F}_n^{-1}(\tfrac{i}{n}))_{1\le i\le n}\right)^2}{\mathrm{Var}\left((\Phi^{-1}(\tfrac{i}{n+1}))_{1\le i\le n}\right)\mathrm{Var}\left((\widehat{F}_n^{-1}(\tfrac{i}{n}))_{1\le i\le n}\right)}, \end{aligned}$$

where the notation Cov *and* Var *refers to the empirical covariance and variance. Assuming that we may replace $\widehat{F}_n^{-1}$ with the CDF F^{-1} of X_1 when n is large, and seeing the empirical variance and covariance as Riemann sums, we deduce that R^2 should converge to*

$$\frac{\left(\displaystyle\int_{u=0}^1 \Phi^{-1}(u)F^{-1}(u)\mathrm{d}u - \int_{u=0}^1 \Phi^{-1}(u)\mathrm{d}u \int_{u=0}^1 F^{-1}(u)\mathrm{d}u\right)^2}{\left(\displaystyle\int_{u=0}^1 \Phi^{-1}(u)^2\mathrm{d}u - \left(\int_{u=0}^1 \Phi^{-1}(u)\mathrm{d}u\right)^2\right)\left(\displaystyle\int_{u=0}^1 F^{-1}(u)^2\mathrm{d}u - \left(\int_{u=0}^1 F^{-1}(u)\mathrm{d}u\right)^2\right)}$$

and it is easy to see that this quantity rewrites as

$$\rho^2 := \mathrm{Corr}(\Phi^{-1}(U), F^{-1}(U))^2,$$

where $U \sim \mathcal{U}[0,1]$. Notice that $F^{-1}(U)$ has CDF F and $\Phi^{-1}(U)$ is a standard Gaussian variable. The identity (11.3), namely the fact that $\rho^2 = 1$ or $\rho^2 < 1$ depending on whether the sample is Gaussian or not, next follows from the Cauchy–Schwarz inequality.

The Shapiro–Wilk test relies on a variant of this approach, which takes into account the covariance structure of the noise in the linear regression. It is based on a slightly different QQ-plot, namely the set of points with coordinates

$$\left(m_i, X_{(i)}\right), \qquad i \in \{1, \dots, n\},$$

where $m_i = \mathbb{E}[G_{(i)}]$ for $G_{(1)} \le \cdots \le G_{(n)}$ the order statistics of iid standard Gaussian variables $G_1, \dots, G_n$. In fact, m_i and $\Phi(\frac{i}{n+1})$ are very close, so we are essentially working with the same QQ-plot. But the point is now that, under H_0, we may write

$$\forall i \in \{1, \dots, n\}, \qquad X_{(i)} = \mu + \sigma G_{(i)} = \mu + \sigma m_i + \epsilon_i,$$

with $\epsilon_i := \sigma(G_{(i)} - m_i)$ such that the vector $\boldsymbol{\epsilon}_n = (\epsilon_i)_{1\le i\le n}$ satisfies

$$\mathbb{E}[\boldsymbol{\epsilon}_n] = 0, \qquad \mathrm{Cov}[\boldsymbol{\epsilon}_n] = \sigma^2 V, \qquad V := \mathrm{Cov}\left[G_{(1)}, \dots, G_{(n)}\right].$$

We are therefore in the situation of the Generalised Least Square estimation described in Proposition 5.3.1, with $\beta_0 = \mu$ and $\beta_1 = \sigma$. In particular, if one denotes by $\widehat{\beta}_1$ the GLS of β_1, one gets that $\widehat{\sigma}^2$ is an estimator of σ^2. It turns out that, under H_0, this estimator is consistent, while under H_1, it is not. Therefore, the Shapiro–Wilk test uses the ratio $\widehat{\sigma}^2/S_n^2$, with a suitable normalisation constant, in the same way as the Shapiro–Francia test uses the coefficient R^2 as a test statistic.

To present the Shapiro–Wilk test with more detail, we first state the following Lemma regarding the covariance matrix V.

Lemma 11.4.2 (Relation between m and V)**.** *Let m be the vector with coefficients $(m_i)_{1\le i\le n}$. We have*

$$\mathbf{1}^\top V^{-1} m = 0.$$

Proof. Since the vectors $(G_1, \dots, G_n)$ and $(-G_1, \dots, -G_n)$ have the same law, their respective order statistics $(G_{(1)}, \dots, G_{(n)})$ and $(-G_{(n)}, \dots, -G_{(1)})$ have the same law. We deduce that, for any i, j,

$$m_i = -m_{n-i+1}, \qquad V_{i,j} = V_{n-i+1,n-j+1},$$

which rewrites in matrix form

$$Jm = -m, \qquad JVJ = V,$$

where

$$J = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ 0 & \vdots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix}$$

is such that $J^{-1} = J$. As a consequence, J and V^{-1} commute and

$$JV^{-1}m = V^{-1}Jm = -V^{-1}m.$$

Therefore the coefficients of $V^{-1}m$ satisfy $(V^{-1}m)_i = -(V^{-1}m)_{n-i+1}$, so they sum to 0. □

As a consequence of Lemma 11.4.2, we first deduce explicit expressions for the GLS of μ and σ.

Corollary 11.4.3 (GLS of μ and σ)**.** *In the linear model*

$$\forall i \in \{1, \dots, n\}, \qquad X_{(i)} = \mu + \sigma m_i + \epsilon_i,$$

with $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$ *and* $\mathrm{Cov}[\boldsymbol{\epsilon}_n] = \sigma^2 V$*, the GLS of* μ *and* σ *is given by*

$$\widehat{\mu} = \sum_{i=1}^{n} u_i X_{(i)}, \qquad \widehat{\sigma} = \sum_{i=1}^{n} v_i X_{(i)},$$

where $u, v \in \mathbb{R}^n$ *are defined by*

$$u := \frac{V^{-1}\mathbf{1}}{\mathbf{1}^\top V^{-1}\mathbf{1}}, \qquad v := \frac{V^{-1}m}{m^\top V^{-1}m}.$$

Proof. This is an application of Proposition 5.3.1, using Lemma 11.4.2 to simplify the computation. □

We may now study the ratio between the estimators $\widehat{\sigma}^2$ and S_n^2 of σ^2, with an appropriate normalisation.

Proposition 11.4.4 (Shapiro–Wilk statistic)**.** *Let us define the* Shapiro–Wilk *statistic by*

$$W_n = \frac{1}{(n-1)\|v\|^2}\frac{\widehat{\sigma}^2}{S_n^2} = \mathrm{Corr}\left(V^{-1}m, (X_{(i)})_{1\le i\le n}\right)^2 \in [0,1].$$

Under H_0*,* W_n *is free.*

Proof. The identity between the two expressions given for W_n is straightforward to check and explains in particular the reason for the normalisation of the ratio $\widehat{\sigma}^2/S_n^2$ in the first expression. Now, since under H_0, $X_{(i)} = \mu + \sigma G_{(i)}$, we get using the invariance of empirical correlation under affine transforms that

$$W_n = \mathrm{Corr}\left(V^{-1}m, (G_{(i)})_{1\le i\le n}\right)^2,$$

which only depends on the standard Gaussian vector $(G_1, \dots, G_n)$ and no longer on μ and σ, so it is free. □

To conclude the construction of the Shapiro–Wilk test, we now study its asymptotic behaviour under H_0 and H_1.

Lemma 11.4.5 (Asymptotic behaviour of W_n)**.** *Let*

$$\rho^2 \begin{cases} = 1, & \text{under } H_0, \\ < 1, & \text{under } H_1, \end{cases}$$

be defined as in Remark 11.4.1. We have

$$\lim_{n \to +\infty} W_n = \rho^2, \qquad \text{in probability.}$$

Proof. The key argument of the proof is the fact that

$$\lim_{n \to +\infty} \|V^{-1} m - 2m\| = 0,$$

which is admitted[11]. It implies that asymptotically, the Shapiro–Wilk statistic behaves like the Shapiro–Francia statistic, because $m_i \simeq \Phi(\frac{i+1}{n})$, which gives the result. □

We are now in position to fully describe the Shapiro–Wilk test. To proceed, we call the law of W_n under H_0 the *Shapiro–Wilk distribution* with parameter n, and denote by $w_{n,r}$ its quantile of order r.

Proposition 11.4.6 (Shapiro–Wilk test)**.** *The test with rejection region $\{W_n \leq w_{n,\alpha}\}$ is consistent and has level α.*

Proof. That this test has level α follows from Proposition 11.4.4. For consistency, we may first notice that Lemma 11.4.5 applied under H_0 implies that $w_{n,\alpha} \to 1$ for any $\alpha > 0$. Therefore, under H_1, since $W_n \to \rho^2 < 1$, we get that $\mathbb{P}(W_n \leq w_{n,\alpha}) \to 1$. □

11.A Exercises

Training exercises

Exercise 11.A.1 (On Donsker's Theorem)**.** We recall the Definition 11.1.2 of the Brownian bridge. The purpose of this exercise is to give a proof of a weak form of the convergence stated in Theorem 11.1.3.

1. For all $t \in [0, 1]$, compute the variance of $\beta(t)$.
2. Let F be a CDF on $\mathbb{R}$, and let $x_1 \leq \cdots \leq x_d$ in $\mathbb{R}$. What is the law of the random vector $\mathbf{G} = (\beta(F(x_1)), \ldots, \beta(F(x_d)))$?
3. With the notation of Theorem 11.1.3, show that when $n \to +\infty$, the random vector $\mathbf{G}_n = (G_n(x_1), \ldots, G_n(x_d))$ converges in distribution to $\mathbf{G}$.

The property which we established is called the *convergence in finite-dimensional distribution*.

Exercise 11.A.2 (Nonasymptotic Kolmogorov test)**.** Let $(X_i)_{i \geq 1}$ be a sequence of iid random variables with CDF F on $\mathbb{R}$.

1. Using Hoeffding's inequality, show that for all $a > 0$,

$$\sup_{x \in \mathbb{R}} \mathbb{P}\left(\sqrt{n} |\widehat{F}_n(x) - F(x)| \geq a\right) \leq 2 \exp(-2a^2).$$

[11] It is proved in J. R. Leslie, M. A. Stephens, S. Fotopoulos. Asymptotic distribution of the Shapiro–Wilk W for testing for normality. *Annals of Statistics*, 1986.

In 1956, Dvoretzky, Kiefer and Wolfowitz[12] proved that there exists a constant C such that

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\sqrt{n}|\widehat{F}_n(x)-F(x)|\geq a\right)\leq C\exp(-2a^2).$$

In 1958, Birnbaum and McCarty[13] conjectured that the best constant was $C=2$, and this conjecture was proved by Massart[14] in 1990.

2. Deduce a nonasymptotic version of the Kolmogorov test, with level at most α, which does not require to compute the quantiles of the Kolmogorov statistic. What is another benefit of this test?

Exercise 11.A.3 (Sum of twelve uniform variables)**.** A student is asked to write a random number generator returning iid realisations of the standard normal distribution using only uniform random variables. He is clever, but a bit lazy, and he therefore decides to write a program which returns $X=\sum_{i=1}^{12}U_i-6$, where $U_1,\dots,U_{12}$ are independent uniform variables on $[0,1]$.

1. Compute the mean and variance of X.
2. Plot the density of X together with the density of the standard normal distribution.
3. The student's teacher applies the Kolmogorov test, with level $\alpha=5\%$, to determine whether the student's code actually returns standard Gaussian variables. Using numerical simulations, estimate the probability that she detects the student's cheat, for $n=10^4$, $n=10^5$, $n=10^6$.

Homework

Exercise 11.A.4 (Lilliefors test for the Gaussian model)**.** A series of 1000 values $X_1,\dots,X_n\in\mathbb{R}$ is measured. The empirical mean and variance are

$$\overline{X}_n=1.05,\qquad V_n=1.42^2.$$

The corresponding histogram is plotted on Figure 11.7, together with the density of the $\mathcal{N}(1.05,1.42^2)$ distribution, and we want to know whether the law P under which $X_1,\dots,X_n$ have been drawn is a Gaussian.

1. We set $\mathcal{P}_0=\{\mathcal{N}(\mu,\sigma^2),\mu\in\mathbb{R},\sigma^2>0\}$ and consider the hypotheses $H_0=\{P\in\mathcal{P}_0\}$, $H_1=\{P\notin\mathcal{P}_0\}$. We recall that we denote by Φ the CDF of the standard Gaussian distribution, and for any $\mu\in\mathbb{R}$ and $\sigma^2>0$, we denote by F_{0,μ,σ^2} the CDF of the law $\mathcal{N}(\mu,\sigma^2)$.
   (a) For any $x\in\mathbb{R}$, write $F_{0,\overline{X}_n,V_n}(x)$ in terms of Φ and $(X_1,\dots,X_n)$.
   (b) Deduce that, under H_0, the statistic

$$\zeta_n'=\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x)-F_{0,\overline{X}_n,V_n}(x)\right|$$

   has the same law as the random variable

$$Z_n'=\sup_{u\in(0,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{U_i\leq u\}}-\Psi_n(u;U_1,\dots,U_n)\right|,$$

   for some function $\Psi_n(u;U_1,\dots,U_n)$ to explicit in terms of Φ, where $(U_1,\dots,U_n)$ are independent uniform random variables on $[0,1]$.

[12] A. Dvoretzky, J. Kiefer and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 1956.

[13] Z. W. Binbaum and R. McCarty. A distribution-free upper confidence bound for $\mathbb{P}(Y<X)$, based on independent samples of X and Y. *Annals of Mathematical Statistics*, 1958.

[14] P. Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 1990.

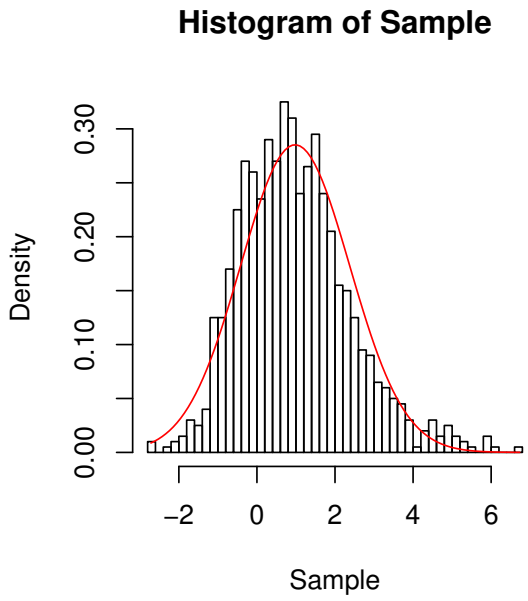


Figure 11.7: Histogram of the sample, and density with estimated parameters, for Exercise 11.A.4.

(c) Conclude by describing the rejection region of a test with level α for the hypotheses H_0 and H_1.

2. With the data of Figure 11.7, ζ_n' takes the value 0.05. Using Monte Carlo simulation, compute the p-value associated with this observation. What do you conclude? *You may refer to the Python code included in Subsection 11.2.3 for the general Monte Carlo procedure.*

Supplementary exercises

Exercise 11.A.5 (Wilcoxon's test)**.** The *Wilcoxon signed-rank test* is adapted to a slightly different framework than homogeneity tests described so far, as it addresses *matched* samples[15]. In this context, it is assumed that the sample is a series of independent pairs $(X_{1,i}, X_{2,i})_{1\leq i\leq n}$, such that the differences $Z_i = X_{1,i} - X_{2,i}$ are identically distributed. We recall that the law of a random variable ζ is said to be *symmetric* if ζ and $-\zeta$ have the same law, and define the following null and alternative hypotheses:

$$H_0 = \{\text{the law of } Z_1 \text{ is symmetric}\}, \qquad H_1 = \{\text{the law of } Z_1 \text{ is not symmetric}\}.$$

An example. The papers of n students at an exam are graded successively by two professors, yielding the grades $X_{1,i}$ and $X_{2,i}$ for the i-th student. We want to know whether there is a 'professor effect' resulting from different grading methods. We assume that each paper has an intrinsic value x_i, and that the grade given by each professor is a random fluctuation around this intrinsic value, so that

$$X_{1,i} = x_i + \epsilon_{1,i}, \qquad X_{2,i} = x_i + \epsilon_{2,i},$$

where the sequences $(\epsilon_{1,i})_{1\leq i\leq n}$ and $(\epsilon_{2,i})_{1\leq i\leq n}$ are independent, and within each of these sequences, the variables are iid with respective distributions P_1 and P_2. In this context, what is the relation between the null hypothesis H_0 introduced above and the 'natural' homogeneity hypothesis $H_0' = \{P_1 = P_2\}$?

We shall make the technical assumption:

$$\forall z \in \mathbb{R}, \qquad \mathbb{P}(Z_1 = z) = 0. \tag{11.4}$$

[15] *Échantillons appariés* en français.

This ensures that, almost surely, the values of $Z_1, \dots, Z_n$ are pairwise distinct, and therefore there is a unique permutation π of $\{1, \dots, n\}$, the set of which is denoted by $\mathfrak{S}_n$, such that $|Z_{\pi(1)}| < \cdots < |Z_{\pi(n)}|$, and thus allows us to introduce the statistic

$$T^+ = \sum_{k=1}^{n} k \mathbb{1}_{\{Z_{\pi(k)} > 0\}},$$

on which Wilcoxon's test for H_0 and H_1 is based. We insist on the fact that the permutation π depends on the realisation of the sample $(Z_1, \dots, Z_n)$, and therefore is *random*.

1. Let ζ be a random variable with symmetric law, satisfying (11.4). Show that the random variables[16] $\operatorname{sign}(\zeta) \in \{-1, 1\}$ and $|\zeta| > 0$ are independent, and that $\operatorname{sign}(\zeta)$ is a *Rademacher* variable (that is to say that $\mathbb{P}(\operatorname{sign}(\zeta) = -1) = \mathbb{P}(\operatorname{sign}(\zeta) = 1) = 1/2$).

2. Let $\zeta_1, \dots, \zeta_n$ be iid random variables, with symmetric law and satisfying (11.4). We define the random permutation $\pi \in \mathfrak{S}_n$ to be such that $|\zeta_{\pi(1)}| < \cdots < |\zeta_{\pi(n)}|$. Show that the random variables $\operatorname{sign}(\zeta_{\pi(1)}), \dots, \operatorname{sign}(\zeta_{\pi(n)})$ are independent Rademacher variables.

3. Deduce that under H_0, the statistic T^+ is free, and describe a nonasymptotic two-sided test for H_0.

4. Compute the expectation t_n and the variance σ_n^2 of T^+ under H_0.

5. Show that, under H_0, $(T^+ - t_n)/\sigma_n$ converges in distribution to $\mathcal{N}(0, 1)$. Deduce an asymptotic test for H_0.

[16] The condition (11.4) ensures that $\zeta \neq 0$, almost surely, so that there is no need to take a convention to define the value of $\operatorname{sign}(0)$.

☑ Final Revision Sheet

Exercise 1 (Detection of the acceleration of glacier melting)**.** Over the last decades, glaciers throughout the world have been estimated to lose an average $1\ \%$ of their mass per year. We want to know if this loss has accelerated in 2022. To this aim, we have measured the masses $M_1^{2021}, \dots, M_n^{2021}$ and $M_1^{2022}, \dots, M_n^{2022}$ of n glaciers in 2021 and 2022, and we define, for every $i \in \{1, \dots, n\}$,

$$X_i = -\log\left(\frac{M_i^{2022}}{M_i^{2021}}\right) \in \mathbb{R}.$$

We denote by $\mathbb{P}_\mu$ the probability measure under which $X_1, \dots, X_n$ are iid according to the $\mathcal{N}(\mu, \sigma_0^2)$ distribution, for some known variance σ_0^2. Neglecting the Gaussian fluctuations around μ, an average mass loss by $1\ \%$ corresponds to μ taking the value

$$\mu_0 = -\log(0.99) \simeq 0.01.$$

Thus, to detect the acceleration of melting, we introduce the hypotheses

$$H_0 = \{\mu \leq \mu_0\}, \qquad H_1 = \{\mu > \mu_0\}.$$

1. Under $\mathbb{P}_\mu$, what is the law of the variable G defined by the identity $\overline{X}_n = \mu + \dfrac{\sigma_0}{\sqrt{n}} G$?
2. Is G a statistic?
3. Construct a consistent test with level α for the hypotheses H_0 and H_1. *Keep in mind that we assume that the variance σ_0^2 of $X_1, \dots, X_n$ is known.*
4. We recall that for $\alpha = 0.05$, $\phi_{1-\alpha} = 1.65$. At this confidence level, and taking $\sigma_0 = \mu_0$, is an observed value of $\overline{X}_n = 1.1\ \%$ in 2022 a significant increase of the melting rate if the sample has size $n = 100$? And if $n = 400$?

Exercise 2 (Experimental verification of Mendel's inheritance law)**.** A type of plant has two genetical characters, each of which can take two forms:

- the first character can take the forms A or a;
- the second character can take the forms B or b.

We take a population of first generation plants in which the phenotypes AB, aB, Ab, ab are equally distributed. In order to test the hypothesis H_0 that A and B are dominant and a and b are recessive, we perform random interbreedings in the population and obtain a second generation. Under H_0, the theoretical probability of each phenotype in the second generation, predicted by Mendel's theory, is given below. For a population of $n = 160$ plants in the second population, the actual number of occurrences of each phenotype are also given. What do you conclude?

Phenotype	AB	aB	Ab	ab
Theoretical probability	$9/16$	$3/16$	$3/16$	$1/16$
Experimental observation	100	18	24	18

Exercise 3 (Log-normal distribution)**.** For any $\theta > 0$, we denote by P_θ the law of the random variable

$$X = \mathrm{e}^{\theta G}, \qquad \text{for } G \sim \mathcal{N}(0,1).$$

We denote by $\mathbb{P}_\theta$ the probability measure under which $X_1, \dots, X_n$ are iid according to P_θ.

1. **Frequentist estimation.**

   (a) For any integer $k \geq 1$, compute $\mathbb{E}_\theta[X_1^k]$. *Hint: think of the identity*

   $$\forall z \in \mathbb{R}, \qquad k\theta z - \frac{z^2}{2} = -\frac{1}{2}(z - k\theta)^2 + \frac{k^2\theta^2}{2}.$$

   (b) Deduce from the value of $\mathbb{E}_\theta[X_1]$ a moment estimator $\widetilde{\theta}_n$ of θ.

   (c) Show that this estimator is strongly consistent and asymptotically normal, and compute its asymptotic variance.

   (d) Compute the density of X_1 under $\mathbb{P}_\theta$, and then deduce the MLE $\widehat{\theta}_n$ of θ.

   (e) Show that the random variable $\widehat{\theta}_n^2/\theta^2$ is free.

   (f) Let $\theta_0 > 0$. Construct a consistent test with level 5% for the hypotheses $H_0 = \{\theta \leq \theta_0\}$, $H_1 = \{\theta > \theta_0\}$, and then for the hypotheses $H_0 = \{\theta = \theta_0\}$, $H_1 = \{\theta \neq \theta_0\}$.

2. **Nonparametric test.** We now assume that we observe a sample $X_1, \dots, X_n$ of positive variables which are iid under some probability measure P, and want to test if P belongs to the parametric family $\mathcal{P}_0 = \{P_\theta : \theta > 0\}$. We thus consider the nonparametric hypotheses

   $$H_0 = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}.$$

   To make the computation easier we shall work with the sample $(Y_1, \dots, Y_n)$ defined by $Y_i = \log(X_i)$. We denote by Q the common law of $Y_1, \dots, Y_n$, and by Q_θ the law of $\log(X)$ when $X \sim P_\theta$.

   (a) Rewrite H_0 and H_1 in terms of Q.

   (b) A natural idea is to apply a normality test, such as the Lilliefors test or the Shapiro–Wilk test, to the sample $(Y_1, \dots, Y_n)$. We denote by W_n' the rejection region, associated with the level α, of either of these tests. Recall the null hypothesis H_0' of these tests, compare it with our null hypothesis H_0, and deduce that this procedure is not consistent, in the sense that one can find a probability measure $Q \notin H_0$ such that $\mathbb{P}_Q(W_n')$ does not converge to 1 when $n \to +\infty$.

   (c) For any $\theta > 0$, express the CDF F_θ of Q_θ as a function of θ and the CDF Φ of the standard Gaussian distribution $\mathcal{N}(0,1)$.

   (d) We denote by $\widehat{F}_n$ the empirical CDF of the sample $(Y_1, \dots, Y_n)$. Show that, under H_0, the random variable

   $$\zeta_n' = \sup_{y \in \mathbb{R}} \left| \widehat{F}_n(y) - F_{\widehat{\theta}_n}(y) \right|$$

   is free. *You may start by writing $\widehat{\theta}_n$ in terms of $Y_1, \dots, Y_n$.*

   (e) Conclude by describing the rejection region of a test with level α for the hypotheses H_0 and H_1.

Training for the Exam

Exercise 1 (Quota method for surveys)**.** We consider a binary survey in which people have to answer either 0 or 1. We denote by p the probability that a randomly chosen individual answers 1.

We first interview n persons chosen uniformly and independently in the population. We denote by $X_1, \dots, X_n$ the answers, which are iid realisations with law $\mathcal{B}(p)$, and we set $\widehat{p}_n = \overline{X}_n$.

1. Recall the value of $\mathrm{Var}(\widehat{p}_n)$ and the expression of an asymptotic confidence interval with level $1-\alpha$ for p.

The *quota method*, which is used in practice by survey companies, consists in dividing the population into $K \geq 2$ classes, depending on several features (sex, age, socioeconomic class). For each class $k \in \{1, \dots, K\}$, we denote by p_k the probability that a member of the class k answers 1, and $u_k \in (0,1)$ the proportion of the population which belongs to this class. We therefore have

$$\sum_{k=1}^{K} u_k = 1, \qquad \sum_{k=1}^{K} u_k p_k = p.$$

Moreover, we assume that the numbers p_k are unknown, while the numbers u_k are known.

For a given total sample size n, we set[17] $n_k = u_k n$. For each class k, we interview n_k members of the class. We denote by $X_{k,1}, \dots, X_{k,n_k}$ the obtained variables, which are therefore iid with law $\mathcal{B}(p_k)$, and we set $\widehat{p}_{k,n_k}$ the associated empirical mean. The samples $(X_{1,1}, \dots, X_{1,n_1}), \dots, (X_{K,1}, \dots, X_{K,n_K})$ are independent. We finally define

$$\widehat{p}_n^K = \sum_{k=1}^{K} u_k \widehat{p}_{k,n_k} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_{k,i}.$$

2. Show that $\widehat{p}_n^K$ is strongly consistant and asymptotically normal, and express the asymptotic variance in terms of $(u_k)_{1\leq k\leq K}$ and $(p_k)_{1\leq k\leq K}$.

3. Deduce an asymptotic confidence interval with level $1-\alpha$ for p using the estimators $\widehat{p}_{1,n_1}, \dots, \widehat{p}_{K,n_K}$.

4. Show that
$$\sum_{k=1}^{K} u_k \widehat{p}_{k,n_k}(1-\widehat{p}_{k,n_k}) \leq \widehat{p}_n^K(1-\widehat{p}_n^K)$$
and comment on the interest of the quota method.

5. We have asked the question: *should one say « pain au chocolat » or « chocolatine »?* to a sample of $n = 1000$ French people, divided into 3 groups: Aquitains, Occitans, and the remainder of France. The results are reported below.

Area	Size of the sample	Proportion of « pain au chocolat »
Aquitaine	88	37%
Occitanie	85	53%
Remainder	827	90%

[17] We shall always assume that $u_k n$ is an integer.

Compute the confidence intervals (with level 95%) on the proportion of people who say « pain au chocolat » obtained without the quota method, and with it.

6. Justify that the method is all the more efficient that the classes are homogeneous, that is to say that the members of each class give similar answers.

Exercise 2 (Sex of first- and second-born children)**.** This exercise is dedicated to a further analysis of the data from Tables 10.1 and 10.2, on the sex of first- and second-born children. We consider a sample of n families with two children, and in the i-th family, we denote by $X_i \in \{\mathrm{m}, \mathrm{f}\}$ the sex of the first child, and by $Y_i \in \{\mathrm{m}, \mathrm{f}\}$ the sex of the second child. The pairs $(X_i, Y_i)_{1 \leq i \leq n}$ are assumed to be iid, and the notation for the probability mass function of their distribution is summarised in Table 11.1.

	$Y_1 = \mathrm{m}$	$Y_1 = \mathrm{f}$	Total
$X_1 = \mathrm{m}$	$p_{\mathrm{m},\mathrm{m}}$	$p_{\mathrm{m},\mathrm{f}}$	$p^X_{\mathrm{m}} = p_{\mathrm{m},\mathrm{m}} + p_{\mathrm{m},\mathrm{f}}$
$X_1 = \mathrm{f}$	$p_{\mathrm{f},\mathrm{m}}$	$p_{\mathrm{f},\mathrm{f}}$	$p^X_{\mathrm{f}} = p_{\mathrm{f},\mathrm{m}} + p_{\mathrm{f},\mathrm{f}}$
Total	$p^Y_{\mathrm{m}} = p_{\mathrm{m},\mathrm{m}} + p_{\mathrm{f},\mathrm{m}}$	$p^Y_{\mathrm{f}} = p_{\mathrm{m},\mathrm{f}} + p_{\mathrm{f},\mathrm{f}}$	

Table 11.1: Notation for the probability mass functions of (X_1, Y_1) and of X_1 and Y_1.

For each $(x, y) \in \{\mathrm{m}, \mathrm{f}\} \times \{\mathrm{m}, \mathrm{f}\}$, the estimator $\widehat{p}_{n,x,y}$ of $p_{x,y}$ is defined by

$$\widehat{p}_{n,x,y} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x, Y_i = y\}},$$

and the respective estimators $\widehat{p}^X_{n,x}$ and $\widehat{p}^Y_{n,y}$ of p^X_x and p^Y_y are defined by

$$\widehat{p}^X_{n,x} := \widehat{p}_{n,x,\mathrm{m}} + \widehat{p}_{n,x,\mathrm{f}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}}, \qquad \widehat{p}^Y_{n,y} := \widehat{p}_{n,\mathrm{m},y} + \widehat{p}_{n,\mathrm{f},y} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = y\}}.$$

The values of these estimators for the dataset, with $n = 5983$, are reported on Table 10.2.

1. We recall that the p-value of the χ_2 test of independence between the variables X_1 and Y_1 is approximately $3.2\ 10^{-3}$. So what is the conclusion of the test?

2. (a) What is the law of the random variable $\mathbb{1}_{\{X_1 = \mathrm{m}\}}$?
   (b) Describe the rejection region of a consistent test, with asymptotic level α, for the hypotheses $H_0 = \{p^X_{\mathrm{m}} = 1/2\}$, $H_1 = \{p^X_{\mathrm{m}} \neq 1/2\}$.
   (c) For $\alpha = 0.05$ and the data from Table 10.2, what is the conclusion of the test?

3. We consider the estimation of the conditional probabilities

$$q_{\mathrm{m}} = \mathbb{P}(Y_1 = \mathrm{m} | X_1 = \mathrm{m}), \qquad q_{\mathrm{f}} = \mathbb{P}(Y_1 = \mathrm{f} | X_1 = \mathrm{f}).$$

   (a) Express q_{m} and q_{f} in terms of the parameters in Table 11.1, and deduce strongly consistent estimators $\widehat{q}_{n,\mathrm{m}}$ and $\widehat{q}_{n,\mathrm{f}}$ of q_{m} and q_{f}.
   (b) What is their value for the data of Table 10.2? Comment on these results.
   (c) For $x \in \{\mathrm{m}, \mathrm{f}\}$ and $i \in \{1, \dots, n\}$, we define

$$R_{x,i} = \begin{pmatrix} \mathbb{1}_{\{X_i = x, Y_i = x\}} \\ \mathbb{1}_{\{X_i = x\}} \end{pmatrix} \in \mathbb{R}^2.$$

   Express the vector $\mathbb{E}[R_{x,1}] \in \mathbb{R}^2$ and the matrix $\mathrm{Cov}[R_{x,1}] \in \mathbb{R}^{2 \times 2}$ in terms of the parameters from Table 11.1.

(d) Let $\phi : (0, +\infty)^2 \to (0, +\infty)$ be defined by

$$\forall r_1, r_2 > 0, \qquad \phi\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \frac{r_1}{r_2}.$$

Compute the coefficients of the 1×2 matrix

$$\nabla\phi(\mathbb{E}[R_{x,1}]) := \left(\frac{\partial\phi}{\partial r_1}(\mathbb{E}[R_{x,1}]) \quad \frac{\partial\phi}{\partial r_2}(\mathbb{E}[R_{x,1}])\right).$$

(e) Deduce that $\widehat{q}_{n,x}$ is asymptotically normal and express its asymptotic variance v_x in terms of the parameters from Table 11.1.

(f) Give an asymptotic confidence interval with level $1 - \alpha$ for q_x, and compute the bounds of this interval for q_{m} and q_{f} with the data from Table 10.2, for $\alpha = 0.05$.

Exercise 3 (Quantile estimation)**.** Let X be a random variable in $\mathbb{R}$, with Cumulative Distribution Function (CDF) $F : \mathbb{R} \to [0, 1]$. For $r \in (0, 1)$, we set

$$q_r := F^{-1}(r) = \inf\{x \in \mathbb{R} : F(x) \geq r\}.$$

In general, we have $F(q_r) = \mathbb{P}(X \leq q_r) \geq r$, and throughout this problem we shall actually assume that $F(q_r) = r$. The goal of the problem is to study the estimation of q_r.

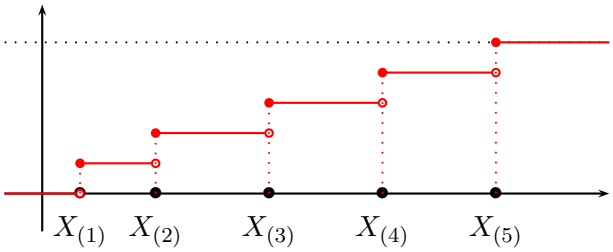
Given a sample $(X_1, \ldots, X_n)$ of independent and identically distributed random variables with CDF F, we define $X_{(1)} \leq \cdots \leq X_{(n)}$ by

$$\forall k \in \{1, \ldots, n\}, \qquad X_{(k)} := \widehat{F}_n^{-1}\left(\frac{k}{n}\right),$$

where $\widehat{F}_n$ is the empirical CDF of the sample, defined by

$$\forall x \in \mathbb{R}, \qquad \widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

In other words, $(X_{(1)}, \ldots, X_{(n)})$ is the nondecreasing reordering of $(X_1, \ldots, X_n)$, see the picture below.



CDF of $X_{(k)}$ and Wilks' estimator. Let $k \in \{1, \ldots, n\}$ and $x \in \mathbb{R}$.

1. What is the law of the random variable $n\widehat{F}_n(x)$?
2. Deduce that $\mathbb{P}\left(X_{(k)} \leq x\right) = \displaystyle\sum_{\ell=k}^{n} \binom{n}{\ell} F(x)^{\ell}(1 - F(x))^{n-\ell}$.
3. In this question, we aim at constructing an approximate confidence interval for q_r of the form $(-\infty, X_{(n)})$; in other words, we fix some $\alpha \in (0, 1)$ and look for the smallest sample size n such that

$$\mathbb{P}(X_{(n)} > q_r) \geq 1 - \alpha.$$

Finding confidence intervals of this form is crucial in many applications where the risk of underestimating q_r (that is to say, of having $X_{(n)} \leq q_r$) must be controlled.

(a) For any $n \geq 1$, express $\mathbb{P}(X_{(n)} > q_r)$ as a function of r.

(b) Express $n_{r,\alpha} := \min\{n \geq 1 : \mathbb{P}(X_{(n)} > q_r) \geq 1 - \alpha\}$ as a function of r and α.

(c) Compute $n_{r,\alpha}$ for $r = 0.95$ and $\alpha = 0.05$.

The estimator $X_{(n_{r,\alpha})}$ of q_r constructed in the last question is called *Wilks' estimator*.

Consistency and asymptotic normality of the empirical quantile. We now study the asymptotic behaviour of $\widehat{F}_n^{-1}(r) = X_{(\lceil nr \rceil)}$, where we recall that for any $z \in \mathbb{R}$, $\lceil z \rceil$ denotes the unique integer such that $\lceil z \rceil - 1 < z \leq \lceil z \rceil$. We temporarily admit the following statement.

Proposition 11.1.6 (Asymptotic behaviour of $U_{(\lceil nr \rceil)}$)**.** *Let $U_1, \dots, U_n$ be independent random variables with uniform distribution on $[0,1]$, and denote by $U_{(1)} \leq \cdots \leq U_{(n)}$ their nondecreasing reordering. For any $r \in (0,1)$, when $n \to +\infty$,*

(i) $U_{(\lceil nr \rceil)}$ converges in probability to r,

(ii) $\sqrt{n}(U_{(\lceil nr \rceil)} - r)$ converges in distribution to $\mathcal{N}(0, r(1-r))$.

We first use Proposition 11.1.6 to describe the asymptotic behaviour of $X_{(\lceil nr \rceil)}$ in terms of the quantile q_r and the CDF F.

1. Show that, for any $k \in \{1, \dots, n\}$, $X_{(k)}$ has the same law as $F^{-1}(U_{(k)})$.

2. Deduce that, if F^{-1} is continuous, then $X_{(\lceil nr \rceil)}$ is a consistent estimator of q_r.

3. We admit that, if X has a positive and continuous density p on some interval $I \subset \mathbb{R}$, then F^{-1} is C^1 on $(0,1)$ and its derivative writes $(F^{-1})'(u) = 1/p(F^{-1}(u))$ for any $u \in (0,1)$. Under this assumption, show that $X_{(\lceil nr \rceil)}$ is asymptotically normal and compute its asymptotic variance.

The sequel of this problem is dedicated to the proof of Proposition 11.1.6. So from now on we only work with the sample $(U_1, \dots, U_n)$.

4. Show that, for any $k \in \{1, \dots, n\}$, the random variable $U_{(k)}$ has density

$$p_{k,n}(u) = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k}, \qquad u \in [0,1].$$

5. Let $(Y_n)_{n \geq 1}$ and $(Z_n)_{n \geq 1}$ be two independent sequences of independent random variables with exponential distribution of parameter 1. For $k \in \{1, \dots, n\}$, write the density of the pair $(k\overline{Y}_k, (n-k+1)\overline{Z}_{n-k+1})$, where we recall that

$$\overline{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i, \qquad \overline{Z}_{n-k+1} = \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} Z_i.$$

6. Deduce that $U_{(k)}$ has the same law as $V_{k,n} := \dfrac{k\overline{Y}_k}{k\overline{Y}_k + (n-k+1)\overline{Z}_{n-k+1}}$.
You may use the fact that the function $(y,z) \mapsto (\frac{y}{y+z}, y+z)$ is a C^1-diffeomorphism from $(0,+\infty)^2$ to $(0,1) \times (0,+\infty)$.

7. From now on, we fix $r \in (0,1)$ and define

$$\xi_n := \frac{\lceil nr \rceil}{n} \overline{Y}_{\lceil nr \rceil}, \qquad \zeta_n := \frac{n - \lceil nr \rceil + 1}{n} \overline{Z}_{n - \lceil nr \rceil + 1}.$$

Show that

$$\lim_{n\to+\infty}\begin{pmatrix}\xi_n\\ \zeta_n\end{pmatrix}=\begin{pmatrix}r\\ 1-r\end{pmatrix},\qquad\text{almost surely,}$$

and that

$$\lim_{n\to+\infty}\sqrt{n}\left(\begin{pmatrix}\xi_n\\ \zeta_n\end{pmatrix}-\begin{pmatrix}r\\ 1-r\end{pmatrix}\right)=\mathcal{N}_2\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}r & 0\\ 0 & 1-r\end{pmatrix}\right),\qquad\text{in distribution.}$$

8. Complete the proof of Proposition 11.1.6.

Appendix A

Reminder on Nonparametric Estimation and Regression Models

Contents

This Appendix is extracted from the first-year course *Introduction to Data Science* and contains complements on nonparametric estimation, as well as on the linear and logistic regression models, to which some parts of the course refer.

A.1 Nonparametric Estimation

In this Section, we assume that we observe iid random variables $X_1, \dots, X_n$, and focus on the estimation of their common distribution P. Unlike the approach studied in Lecture 2, we do not want here to do *a priori* assumptions on the shape of P, and we therefore focus on the *nonparametric* approach.

If E is a finite, or countably infinite space, estimating P is, in principle, elementary. Indeed, denoting by $(p_x)_{x\in E}$ the Probability Mass Function (PMF) of P, which is defined by

$$\forall x \in E, \qquad p_x = P(\{x\}) = \mathbb{P}(X_1 = x),$$

then it is immediate that the *empirical PMF* of the sample $(\widehat{p}_{n,x})_{x\in E}$, defined by

$$\forall x \in E, \qquad \widehat{p}_{n,x} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}},$$

is an unbiased and strongly consistent estimator of $(p_x)_{x\in E}$.

However, if we now assume that $E = \mathbb{R}^d$ and that P has a density p, estimating this density is a delicate issue: the natural equivalent of the empirical PMF is the *empirical distribution* of the sample $X_1, \dots, X_n$, which is the probability measure on $\mathbb{R}^d$ defined by

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where, for any $x \in \mathbb{R}^d$, we denote by δ_x is the *Dirac measure* at x. While the theoretical properties of $\widehat{P}_n$ are very interesting, from the practical point of view it has the disadvantage of *not* being a probability density, since it is a measure but not a function.

In order to construct estimators of p which remain probability densities, there exist several approaches which all rely on the introduction of a *smoothing* parameter $h > 0$, which plays a very similar role, in terms of bias and variance, to the parameter h from Exercise 2.1.14.

A.1.1 Histogram

For the sake of simplicity we assume here that $X \in \mathbb{R}$, and let $h > 0$. Let us divide $\mathbb{R}$ into a family of intervals $(I_k)_{k\in\mathbb{Z}}$ with width h, which we call *bins*. The associated histogram of the sample $X_1, \dots, X_n$ is the function $\widehat{p}_n^h : \mathbb{R} \to [0, +\infty)$, which is constant on each interval I_k and satisfies

$$\forall x \in I_k, \qquad \widehat{p}_n^h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in I_k\}}.$$

It has the advantage of being a probability density. However, it is (in general) biased and not consistent, in the sense that, for $k \in \mathbb{Z}$ and $x \in I_k$, one has

$$\mathbb{E}\left[\widehat{p}_n^h(x)\right] = p^h(x), \qquad \lim_{n\to+\infty} \widehat{p}_n^h(x) = p^h(x), \quad \text{almost surely,}$$

where $p^h : \mathbb{R} \to [0, +\infty)$ is the piecewise constant function defined by

$$\forall x \in I_k, \qquad p^h(x) = \frac{1}{h} \int_{y\in I_k} p(y)\mathrm{d}y.$$

One may then show that, when $h \to 0$, p^h converges to p (in various senses, and depending on the regularity of p). The quantity h, while necessary for the histogram to be defined, therefore necessarily introduces a bias in the estimation of p. To remove this bias, and recover consistency, it is then possible to let h depend on n and tend to 0 when $n \to 0$. This convergence is illustrated on Figure A.1.

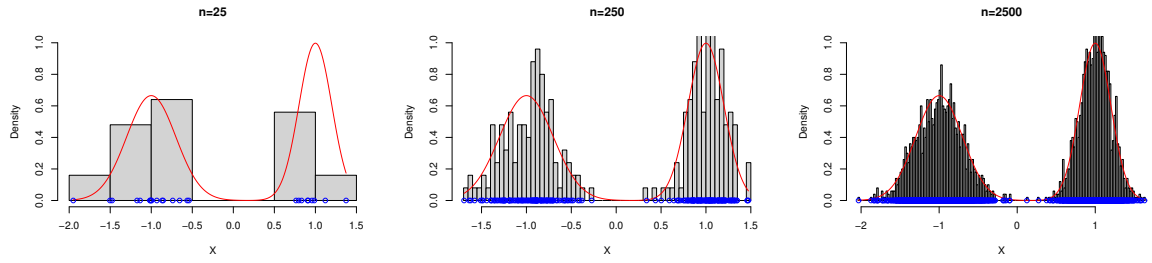


Figure A.1: Histograms of a sample with increasing size n and decreasing bin width h. The blue points are the values of $X_1, \dots, X_n$ and the red line is the density p under which these variables have been drawn.

A.1.2 Moving histogram and kernel smoothing

Histograms are functions, and even probability densities, but they are not very smooth since, by construction, they are piecewise constant on a set of predetermined bins. In order to construct estimators of the density p which are more regular, *moving histograms* are based on the following idea: instead of fixing the bin I_k a priori, and approximating $p(x)$, for all $x \in I_k$, by the number of elements of the sample which fall into this bin (appropriately rescaled by nh to ensure that $\widehat{p}_n^h$ remains a probability density), one may, for each $x \in \mathbb{R}$, approximate $p(x)$ by the number of elements of the sample which fall into an

interval of width h, centered at x (and still appropriately rescaled by nh). In other words, the *moving histogram* with bin width h is the probability density $\widehat{p}_n^h$ defined by

$$\forall x \in \mathbb{R}, \qquad \widehat{p}_n^h(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in (x-h/2, x+h/2]\}}.$$

An example of a moving histogram is given on Figure A.2.

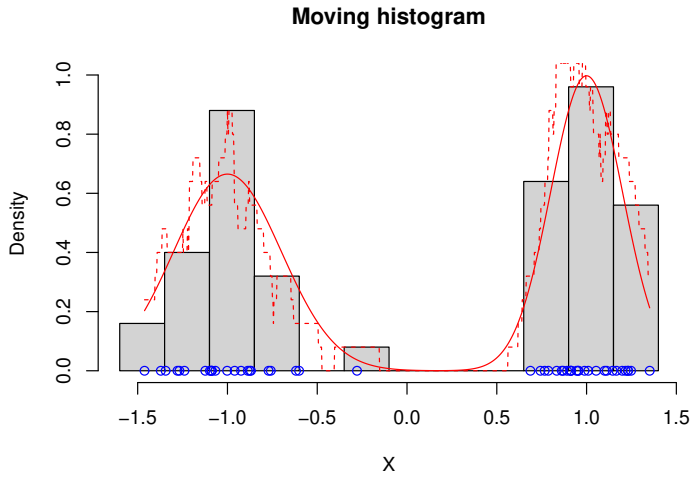


Figure A.2: Histogram and moving histogram (in dotted red) estimation of a probability density.

The idea of moving histogram estimation can be pushed further by noting that $\widehat{p}_n^h(x)$ rewrites under the abstract form

$$\forall x \in \mathbb{R}, \qquad \widehat{p}_n^h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \tag{A.1}$$

where

$$K(u) = \mathbb{1}_{\{u \in (-1/2, 1/2]\}}$$

is the density of the uniform distribution on $(-1/2, 1/2]$. In fact, for *any* probability density K on $\mathbb{R}$, which in this context is called a *kernel*, one may define the function $\widehat{p}_n^h$ by the formula above, and observe that it is a probability density. This density, as a function of x, is now as smooth as the kernel K, therefore its is called the *kernel smoothing* of p (also *Kernel Density Estimator*), see Figure A.3.

As for histograms, it is easy to show that, for all $x \in \mathbb{R}$,

$$\mathbb{E}\left[\widehat{p}_n^h(x)\right] = \lim_{n \to +\infty} \widehat{p}_n^h(x) = p^h(x) := \int_{y \in \mathbb{R}} \frac{1}{h} K\left(\frac{x - y}{h}\right) p(y) \mathrm{d}y,$$

and that p^h converges to p when $h \to 0$. It is also possible to remove the bias induced by h, and recover consistency, by letting h tend to 0 when $n \to +\infty$.

Exercise A.1.1. Show that

$$p^h(x) = \int_{u \in \mathbb{R}} p(x - hu) K(u) \mathrm{d}u,$$

and deduce that if p is bounded on $\mathbb{R}$ and continuous at x, then $p^h(x)$ converges to $p(x)$ when $h \to 0$.

The kernel smoothing estimator can be readily extended to the case where $X \in \mathbb{R}^d$ with $d \geq 2$; in this case, in order to be a probability density on $\mathbb{R}^d$, it must be defined by

$$\forall x \in \mathbb{R}^d, \qquad \widehat{p}_n^h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K_d\left(\frac{x - X_i}{h}\right),$$

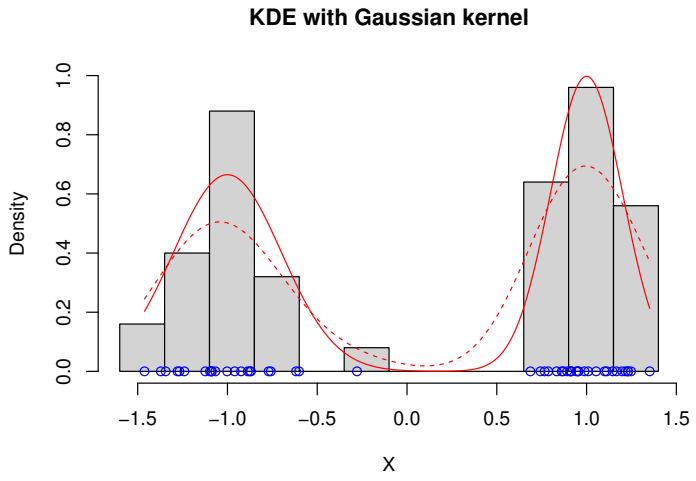


Figure A.3: Kernel density estimator of p (in dotted red), with Gaussian kernel, for the same data as on Figure A.2.

where K_d is a probability density on $\mathbb{R}^d$. The latter can for example be constructed by *tensorisation*, namely by setting

$$K_d(x_1, \dots, x_d) = K(x_1) \cdots K(x_d)$$

for some given kernel K on $\mathbb{R}$.

A.1.3 Bias, variance, over- and underfitting

Following Definition 2.1.10 and Proposition 2.1.11, the MSE of the estimator $\widehat{p}_n^h(x)$ of $p(x)$, defined by (A.1), satisfies

$$\mathrm{MSE}(\widehat{p}_n^h(x)) = \mathbb{E}\left[\left|\widehat{p}_n^h(x) - p(x)\right|^2\right] = \left(p(x) - p^h(x)\right)^2 + \mathrm{Var}\left(\widehat{p}_n^h(x)\right),$$

and one may then check that

$$\begin{aligned}\mathrm{Var}\left(\widehat{p}_n^h(x)\right) &= \frac{1}{n}\left(\frac{1}{h}\int_{u\in\mathbb{R}} p(x-hu)K(u)^2\mathrm{d}u - (p^h(x))^2\right)\\ &\sim \frac{1}{nh}\left(p(x)\int_{u\in\mathbb{R}} K(u)^2\mathrm{d}u\right), \qquad \text{when } h \to 0 \text{ at fixed } n.\end{aligned}$$

As in Exercise 2.1.14, we therefore see that the bias and variance terms vary in opposite directions with h: when h decreases, the bias decreases but the variance increases. This may in fact be observed empirically on Figure A.4: when h is too small, the KDE of p is very peaked around the points $X_1, \dots, X_n$ of the sample. If one draws a new sample, then the KDE will be different, therefore the large variance. This is the *overfitting* phenomenon: the estimation of p is too dependent on the observation of the sample to be informative. On the other hand, if h is too large, then we actually essentially observe the graph of the kernel K, with a large bandwidth. This estimator has a low variance, in the sense that if we draw a new realisation of the sample, it will almost remain unchanged. But it is not informative either: this is the *underfitting* phenomenon. As a conclusion, you may remember that balancing bias and variance allows your estimator to be the most informative, and that the two extreme cases to avoid can be summarised as follows:

low bias, high variance $\Leftrightarrow$ **overfitting**
high bias, low variance $\Leftrightarrow$ **underfitting**

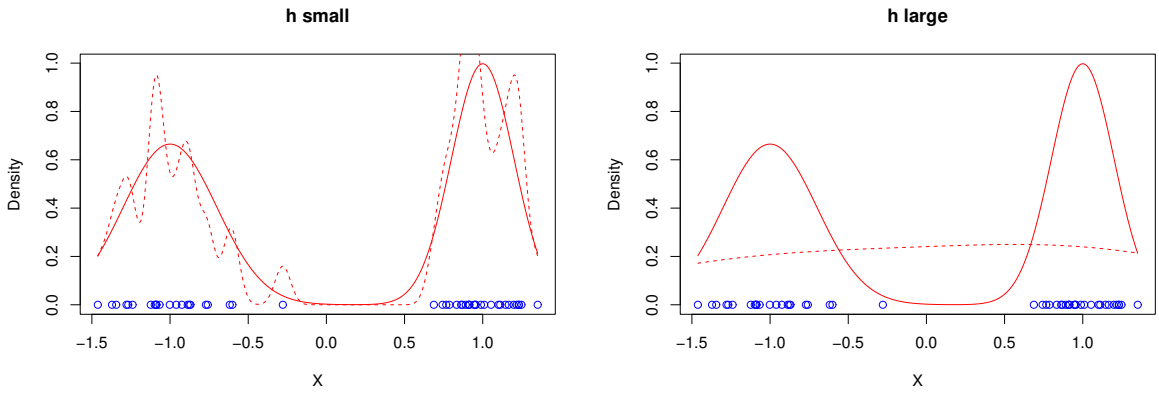


Figure A.4: Over- and underfitting for KDE with h too small, or h too large: the solid red line is the true density, and the dotted red line its KDE.

Let us finally note that computing the MSE of $\widehat{p}_n^h(x)$ for a fixed value of x provides a pointwise information which is rather poor. In general, it is preferred to measure the quality of the estimator $\widehat{p}_n^h$ of p globally. To do so, one may for instance study the *Mean Integrated Squared Error*, defined by

$$\mathrm{MISE}(\widehat{p}_n^h) := \mathbb{E}\left[\int_{x\in\mathbb{R}} |\widehat{p}_n^h(x) - p(x)|^2 \mathrm{d}x\right] = \int_{x\in\mathbb{R}} \mathrm{MSE}(\widehat{p}_n^h(x))\mathrm{d}x.$$

Finding the value of h, depending on n, which minimises this quantity is the starting point of quantitative results on KDE. We refer to [6] for more information.

A.1.4 Nadaraya–Watson nonparametric regression

Anticipating on the sequel of this Appendix, we present here an application of nonparametric estimation to the regression problem.

Exercise A.1.2 (Nadaraya–Watson estimator for nonparametric regression)**.** Let (X, Y) be a pair of random variables in $\mathbb{R}\times\mathbb{R}$ which has a density $p_{(X,Y)}(x, y)$ with respect to the Lebesgue measure.

1. Express the marginal density $p_X(x)$ of the random variable X in terms of the function $p_{(X,Y)}$.

From now on, we assume that:

- $\mathbb{E}[|Y|] < +\infty$;
- the functions $p_{(X,Y)}$ and p_X are continuous;
- for any $x \in \mathbb{R}$, $p_X(x) > 0$.

2. For any measurable subset $B \subset \mathbb{R}$ such that $\mathbb{P}(X \in B) > 0$, we define the *conditional expectation* of Y given the event $\{X \in B\}$ by

$$\mathbb{E}[Y|X \in B] = \frac{\mathbb{E}[Y\,\mathbb{1}_{\{X\in B\}}]}{\mathbb{P}(X \in B)}.$$

Write this quantity as the ratio of two integrals, involving the functions $p_{(X,Y)}$ and p_X, respectively.

We admit that under the assumptions made above, for any $\delta > 0$ and $x \in \mathbb{R}$, we have

$$\lim_{\delta\to 0} \mathbb{E}[Y|X \in [x-\delta, x+\delta]] = m(x) = \int_{y\in\mathbb{R}} y\, p_{Y|X}(y|x)\,\mathrm{d}y,$$

where the function

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

is called the *conditional density* of Y given $\{X = x\}$.

3. Compute $\mathbb{E}[m(X)]$.

4. We define the random variable ε by the identity

$$Y = m(X) + \varepsilon.$$

Show that $\mathbb{E}[\varepsilon] = 0$, and that if $\mathbb{E}[|Y|^2] < +\infty$, then $\mathrm{Cov}(m(X), \varepsilon) = 0$.

5. The aim of *nonparametric regression* is to estimate the function $m(x)$, given a sample $(x_i, y_i)_{1\leq i\leq n}$ of iid realisations of the pair (X, Y). To proceed, we fix two positive kernels K^X and K^Y on $\mathbb{R}$, which satisfy

$$\int_{u\in\mathbb{R}} K^X(u)\,\mathrm{d}u = \int_{v\in\mathbb{R}} K^Y(v)\,\mathrm{d}v = 1, \qquad \int_{v\in\mathbb{R}} v\,K^Y(v)\,\mathrm{d}v = 0,$$

and two positive numbers h^X and h^Y. We then consider the kernel estimator for $p_{(X,Y)}(x,y)$ defined by

$$\widehat{p}_{(X,Y)}(x,y) = \frac{1}{nh^Xh^Y}\sum_{i=1}^{n} K^X\left(\frac{x-x_i}{h^X}\right) K^Y\left(\frac{y-y_i}{h^Y}\right).$$

(a) Compute the marginal density $\widehat{p}_X(x)$ of $\widehat{p}_{(X,Y)}(x,y)$.

(b) For $x \in \mathbb{R}$, compute $\displaystyle\int_{y\in\mathbb{R}} y\,\widehat{p}_{(X,Y)}(x,y)\,\mathrm{d}y$.

(c) Conclude by giving the expression of the function $\widehat{m}(x)$ defined by

$$\widehat{m}(x) = \int_{y\in\mathbb{R}} y\,\widehat{p}_{Y|X}(y|x)\,\mathrm{d}y,$$

where $\widehat{p}_{Y|X}(y|x) = \dfrac{\widehat{p}_{(X,Y)}(x,y)}{\widehat{p}_X(x)}$.

The function $\widehat{m}(x)$ is called the *Nadaraya–Watson* estimator of $m(x)$.

A.2 Introduction to regression

A.2.1 The regression problem

In the framework of *regression*, you observe data which are pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, and you want to find a function $f : \mathbb{R}^p \to \mathbb{R}$ which best represents the relation between x and y, in the sense that you would like $f(x_i)$ to be close to y_i for all $i \in \{1, \ldots, n\}$. A *nonparametric* approach to this problem is studied in Exercise A.1.2. We now focus on *parametric* approaches; namely, we fix a family of functions $\{f_\beta, \beta \in \mathbb{R}^d\}$ such that for each $\beta \in \mathbb{R}^d$,

$$f_\beta : \left\{ \begin{array}{ccc} \mathbb{R}^p & \to & \mathbb{R}, \\ x & \mapsto & y = f_\beta(x) \end{array} \right.$$

and our goal is then to find the parameter β which makes $f_\beta(x_i)$ the closest to y_i for all $i \in \{1, \ldots, n\}$, with the objective to predict the value of y given new values of x.

In this setting, the coordinates $x^1, \ldots, x^p$ of the variable x are called predictors, features, or explanatory variables. The variable y is called the response variable or the explained variable. In econometry, x and y are respectively called independent and dependent variables, but this terminology may be misleading since, if you see $x^1, \ldots, x^p$ as random variables, they need not be statistically independent.

A.2.2 Example: polynomial regression

Assume that $p = 1$, so x and y are real numbers. A simple parametric model is to try to fit the data $(x_i, y_i)_{1 \leq i \leq n}$ with a polynomial curve, which amounts to taking f_β of the form

$$f_\beta(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p,$$

for a vector of parameters $\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$. This looks like a good idea: the Stone–Weierstrass Theorem asserts that on bounded intervals, every continuous function can be uniformly approximated by polynomial functions; and the Lagrange polynomials provide polynomial functions which exactly interpolate data. Polynomial regression is illustrated on Figure A.5. We observe that the order p of the polynomial function plays an important role: if p is too low, the polynomial function fits the data poorly, this is an underfitting situation. On the contrary, if p is too large, the polynomial function almost exaclty interpolates the data, but exhibits a highly fluctuating behaviour outside the data and is likely to predict very poorly the value of y for new values of x. This is an example of overfitting. Therefore, the number of parameters p plays a role similar to the bandwidth h in density estimation as it allows to look for a trade-off between bias (small p) and variance (large p).

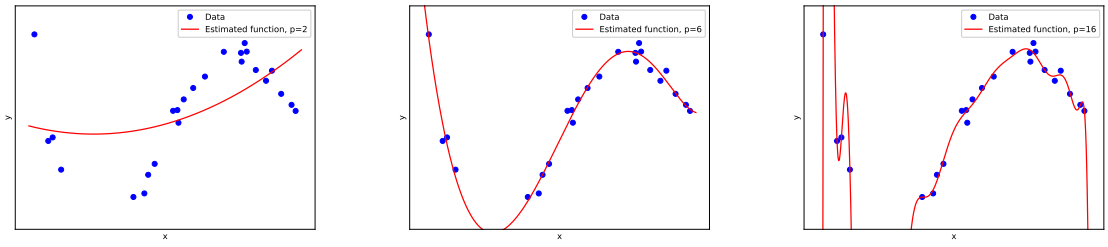


Figure A.5: Polynomial regression with increasing order p. A small order yields underfitting: the curve does not fit the data; a large order yields overfitting: the curve has no predictive capability.

A.2.3 The linear regression model

The sequel of this Appendix is largely dedicated to the case where $x \in \mathbb{R}^p$ with $p \geq 1$ and f_β is an *affine* function of x:

$$f_\beta(x) = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p,$$

with $\beta = (\beta_0, \ldots, \beta_p)$. This is the *linear regression model*, in which the coefficient β_0 is called the *intercept* or *offset*.

This model may seem restrictive: if you observe the free fall of a body and want to predict the position as a function of time, an affine relation will poorly fit your data. But you can always add the square of time as a new feature in your model, which should drastically improve the goodness of fit. Selecting relevant features to incorporate in a model is called *feature engineering*. In particular, it is important to observe that polynomial regression as is presented in the previous subsection is a specific case of linear model, with features $x, x^2, \ldots, x^p$ given by the increasing powers of $x \in \mathbb{R}$.

A.3 Simple linear regression

In this Section, we focus on the case $p = 1$, so x and y are real numbers, and our goal is to find the affine function which best approximates the scatter plot[1] $(x_i, y_i)_{1 \leq i \leq n}$, see Figure A.6.

[1] *Nuage de points* en français.

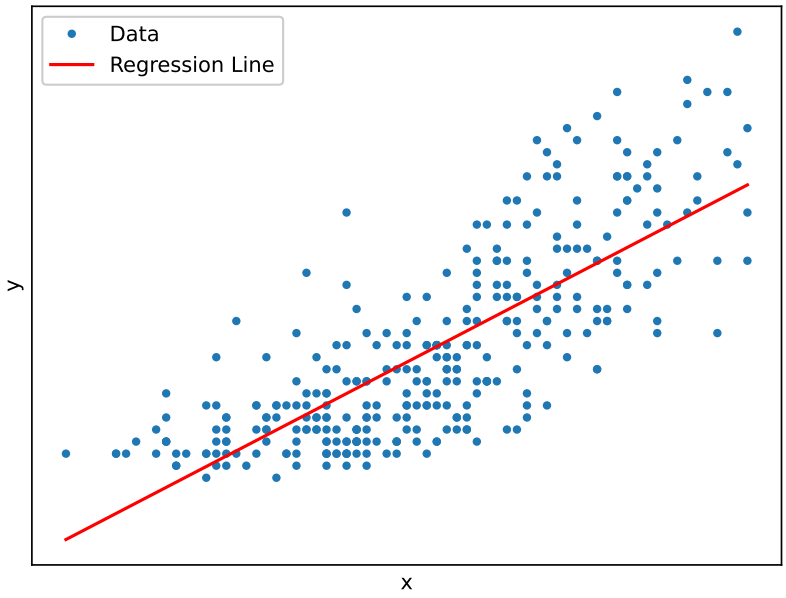


Figure A.6: Simple linear regression.

In the sequel of this Section, we will use the vector notation

$$\mathbf{x}_n = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \qquad \mathbf{y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

and given two vectors $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^n$,

$$\overline{a}_n = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad \mathrm{Var}(\mathbf{a}_n) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a}_n)^2,$$

$$\mathrm{Cov}(\mathbf{a}_n, \mathbf{b}_n) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a}_n)(b_i - \overline{b}_n), \quad \mathrm{Corr}(\mathbf{a}_n, \mathbf{b}_n) = \frac{\mathrm{Cov}(\mathbf{a}_n, \mathbf{b}_n)}{\sqrt{\mathrm{Var}(\mathbf{a}_n)}\sqrt{\mathrm{Var}(\mathbf{b}_n)}}.$$

A.3.1 Ordinary Least Square estimator

The *Least Square Problem*[2] is the minimisation problem

$$\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

Its solution is called the *Ordinary Least Square* estimator[3] (OLS) of (β_0, β_1). In this model, the coefficient β_0 is called the *intercept*.

Proposition A.3.1 (OLS for the simple linear regression)**.** *Assume that $n \geq 2$ and that there exist i, j such that $x_i \neq x_j$. The OLS of (β_0, β_1) is given by*

$$\widehat{\beta}_0 = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n, \qquad \widehat{\beta}_1 = \frac{\mathrm{Cov}(\mathbf{x}_n, \mathbf{y}_n)}{\mathrm{Var}(\mathbf{x}_n)}.$$

[2] *Problème des moindres carrés* en français.
[3] *Estimateur des Moindres Carrés Ordinaire (MCO)* en français.

Exercise A.3.2. Prove Proposition A.3.1.

Remark A.3.3. *The condition that there exist i, j such that $x_i \neq x_j$ simply means that all points (x_i, y_i) are not aligned vertically.*

A.3.2 Coefficient of determination

Once the OLS $(\widehat{\beta}_0, \widehat{\beta}_1)$ is estimated, it may be interesting to quantify the quality of the linear model by measuring whether the points (x_i, y_i) are far from the line with equation $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ or not. To do so, let us define

$$\widehat{\mathbf{y}}_n = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} := \widehat{\beta}_0 \mathbf{1}_n + \widehat{\beta}_1 \mathbf{x}_n \in \mathbb{R}^n, \qquad \mathbf{1}_n := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n,$$

and define the *residual error*[4] between $\mathbf{y}_n$ (the data) and $\widehat{\mathbf{y}}_n$ (the model) by

$$\begin{aligned}
\frac{1}{n}\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2 &= \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_n + \widehat{\beta}_1 \overline{x}_n - \widehat{\beta}_1 x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_n)^2 - 2\widehat{\beta}_1 (y_i - \overline{y}_n)(x_i - \overline{x}_n) + \widehat{\beta}_1^2 (x_i - \overline{x}_n)^2 \\
&= \mathrm{Var}(\mathbf{y}_n) - 2\widehat{\beta}_1 \,\mathrm{Cov}(\mathbf{x}_n, \mathbf{y}_n) + \widehat{\beta}_1^2 \,\mathrm{Var}(\mathbf{x}_n) \\
&= \mathrm{Var}(\mathbf{y}_n) - \frac{\mathrm{Cov}(\mathbf{x}_n, \mathbf{y}_n)^2}{\mathrm{Var}(\mathbf{x}_n)} \\
&= \mathrm{Var}(\mathbf{y}_n)\left(1 - \mathrm{Corr}(\mathbf{x}_n, \mathbf{y}_n)^2\right).
\end{aligned}$$

Since $\widehat{\mathbf{y}}_n = \widehat{\beta}_0 \mathbf{1}_n + \widehat{\beta}_1 \mathbf{x}_n$, one may next rewrite

$$\mathrm{Corr}(\mathbf{x}_n, \mathbf{y}_n)^2 = \mathrm{Corr}(\widehat{\mathbf{y}}_n, \mathbf{y}_n)^2,$$

because up to a sign change, the correlation between two samples is invariant under affine transforms of each sample. This yields the following definition.

Definition A.3.4 (Coefficient of determination)**.** *The* coefficient of determination *is defined by*

$$R^2 = \mathrm{Corr}(\widehat{\mathbf{y}}_n, \mathbf{y}_n)^2 \in [0, 1].$$

We deduce from the identity

$$\frac{1}{n}\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2 = \mathrm{Var}(\mathbf{y}_n)\left(1 - R^2\right) \tag{A.2}$$

derived above that the closer R^2 is to 1, the smaller the residual error and therefore the better the linear model fits the data, see Figure A.7.

[4] In the context of regression, the square root of $\frac{1}{n}\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2$ is often referred to as Root Mean Squared Error (RMSE); it may indeed be understood as an empirical version of the MSE defined in Lecture 2.

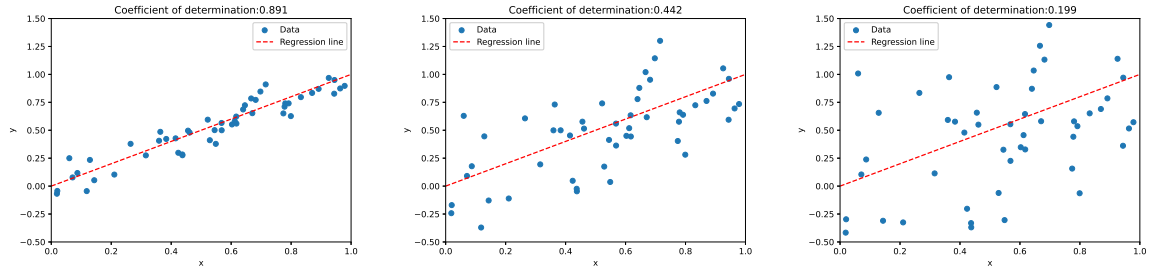


Figure A.7: Values of R^2 for various datasets with the same regression line.

A.4 Multiple linear regression

In this Section, we consider the linear model in the case $p \geq 1$, so features are denoted by $x = (x^1, \dots, x^p) \in \mathbb{R}^p$. We slightly change the notation with respect to Section A.3, and now write

$$\mathbf{x}_n = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^p \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}, \qquad \mathbf{y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}.$$

The matrix $\mathbf{x}_n$ is called the *design matrix*.

A.4.1 Ordinary Least Square estimator

The Least Square Problem

$$\min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i^1 + \cdots + \beta_p x_i^p)\right)^2$$

rewrites

$$\min_{\beta\in\mathbb{R}^{p+1}} \|\mathbf{y}_n - \mathbf{x}_n\beta\|^2.$$

It is clear that if we let $\widehat{\mathbf{y}}_n$ be the orthogonal projection of $\mathbf{y}_n$ onto the range[5] of $\mathbf{x}_n$, any vector $\widehat{\beta} \in \mathbb{R}^{p+1}$ such that $\widehat{\mathbf{y}}_n = \mathbf{x}_n\widehat{\beta}$ is an OLS for β. We give a more precise statement in the next proposition.

Proposition A.4.1 (OLS for multiple linear regression)**.** *Assume that $p+1 \leq n$, and that $\mathbf{x}_n$ has full rank*[6] *$p+1$.*

(i) The matrix $\mathbf{x}_n^\top\mathbf{x}_n \in \mathbb{R}^{(p+1)\times(p+1)}$ is invertible.

(ii) The Least Square Problem has a unique minimiser, given by

$$\widehat{\beta} = \left(\mathbf{x}_n^\top\mathbf{x}_n\right)^{-1}\mathbf{x}_n^\top\mathbf{y}_n. \tag{A.3}$$

The vector $\widehat{\beta}$ is called the OLS of β.

Proof. By the Rank-nullity Theorem[7], since the matrix $\mathbf{x}_n \in \mathbb{R}^{n\times(p+1)}$ has rank $p+1$, it is injective. Therefore, there is a unique $\widehat{\beta} \in \mathbb{R}^{p+1}$ such that $\widehat{\mathbf{y}}_n = \mathbf{x}_n\widehat{\beta}$. Besides, Lemma A.4.2 below shows that $\mathbf{x}_n^\top\mathbf{x}_n$ is invertible and that

$$\widehat{\beta} = \left(\mathbf{x}_n^\top\mathbf{x}_n\right)^{-1}\mathbf{x}_n^\top\widehat{\mathbf{y}}_n.$$

[5] *Image* en français.
[6] *Est de rang complet* en français.
[7] *Théorème du rang* en français.

It now remains to check that one may replace $\widehat{\mathbf{y}}_n$ with $\mathbf{y}_n$ in the right-hand side. To proceed, we remark that by the definition of $\widehat{\mathbf{y}}_n$, for any $u \in \mathbb{R}^{p+1}$,

$$0 = \langle \mathbf{x}_n u, \mathbf{y}_n - \widehat{\mathbf{y}}_n \rangle = \langle u, \mathbf{x}_n^\top (\mathbf{y}_n - \widehat{\mathbf{y}}_n) \rangle,$$

which shows that $\mathbf{x}_n^\top \mathbf{y}_n = \mathbf{x}_n^\top \widehat{\mathbf{y}}_n$ and completes the proof. □

Lemma A.4.2 (Inversion of injective matrices on their range)**.** *Let $\mathbf{x}_n \in \mathbb{R}^{n \times (p+1)}$ be an injective matrix. The matrix $\mathbf{x}_n^\top \mathbf{x}_n \in \mathbb{R}^{(p+1)\times(p+1)}$ is invertible, and for any vector v in the range of $\mathbf{x}_n$, there is a unique vector $u \in \mathbb{R}^{p+1}$ such that $v = \mathbf{x}_n u$. This vector is given by the formula*

$$u = \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top v.$$

Proof. We first prove that the matrix $\mathbf{x}_n^\top \mathbf{x}_n$ is invertible. To this aim, we let $u \in \mathbb{R}^{p+1}$ be such that $\mathbf{x}_n^\top \mathbf{x}_n u = 0$. Then $\|\mathbf{x}_n u\|^2 = u^\top \mathbf{x}_n^\top \mathbf{x}_n u = 0$, so that $u = 0$ since $\mathbf{x}_n$ is injective. Therefore $\mathbf{x}_n^\top \mathbf{x}_n$ is injective and thus invertible. We now take v in the range of $\mathbf{x}_n$, so there exists $u \in \mathbb{R}^{p+1}$ such that $v = \mathbf{x}_n u$. As a consequence, $\mathbf{x}_n^\top v = \mathbf{x}_n^\top \mathbf{x}_n u$ and since $\mathbf{x}_n^\top \mathbf{x}_n$ is invertible, we conclude that $u = (\mathbf{x}_n^\top \mathbf{x}_n)^{-1} \mathbf{x}_n^\top v$. □

Remark A.4.3. *The formula for $\widehat{\beta}$ can also be obtained directly from the first-order optimality condition for the Least Square Problem: we have*

$$\nabla_\beta \|\mathbf{y}_n - \mathbf{x}_n \beta\|^2 = -2\mathbf{x}_n^\top \left(\mathbf{y}_n - \mathbf{x}_n \beta\right),$$

and the right-hand side vanishes if and only if $\mathbf{x}_n^\top \mathbf{y}_n = \mathbf{x}_n^\top \mathbf{x}_n \beta$, which yields the claimed expression.

Remark A.4.4. *In the case of simple linear regression, the design matrix is*

$$\mathbf{x}_n = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}.$$

The condition in Proposition A.3.1 that there exist $i \neq j$ such that $x_i \neq x_j$ is therefore exactly the condition that the two columns of the design matrix be linearly independent, that is to say that $\mathbf{x}_n$ has full rank.

A.4.2 Coefficient of determination

The coefficient of determination introduced in Definition A.3.4 for simple linear regression can be extended to the case of multiple linear regression.

Proposition A.4.5 (Coefficient of determination)**.** *Under the assumptions of Proposition A.4.1, the coefficient of determination defined by*

$$R^2 = \mathrm{Corr}(\widehat{\mathbf{y}}_n, \mathbf{y}_n)^2 \in [0, 1]$$

satisfies

$$\frac{1}{n}\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2 = \mathrm{Var}(\mathbf{y}_n)\left(1 - R^2\right).$$

Proof. We recall that we denote by $\mathbf{1}_n \in \mathbb{R}^n$ the first column of $\mathbf{x}_n$. Since this vector belongs to the range of $\mathbf{x}_n$, we have $\langle \mathbf{y}_n - \widehat{\mathbf{y}}_n, \widehat{\mathbf{y}}_n - \overline{y}_n \mathbf{1}_n \rangle = 0$. In other words, the triangle between $\mathbf{y}_n$, $\widehat{\mathbf{y}}_n$ and $\overline{y}_n \mathbf{1}_n$ is rectangle in $\widehat{\mathbf{y}}_n$, see Figure A.8. Let α denote the nonoriented value of the angle at $\overline{y}_n \mathbf{1}_n$.

On the one hand,

$$\sin^2 \alpha = \left(\frac{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|}{\|\mathbf{y}_n - \overline{y}_n \mathbf{1}_n\|}\right)^2 = \frac{\frac{1}{n}\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2}{\mathrm{Var}(\mathbf{y}_n)}.$$

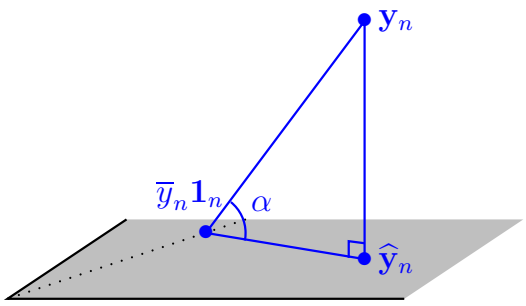


Figure A.8: The gray plane is the range of $\mathbf{x}_n$. The dashed line is the span of $\mathbf{1}_n$. The span of $\mathbf{1}_n$ is a subspace of the range of $\mathbf{x}_n$. Therefore, $\widehat{\mathbf{y}}_n$ and $\mathbf{y}_n$ have the same orthogonal projection $\overline{y}_n\mathbf{1}_n$ onto the span of $\mathbf{1}_n$, and the triangle between $\mathbf{y}_n$, $\widehat{\mathbf{y}}_n$ and $\overline{y}_n\mathbf{1}_n$ is rectangle in $\widehat{\mathbf{y}}_n$.

On the other hand,

$$\cos\alpha = \frac{\langle \mathbf{y}_n - \overline{y}_n\mathbf{1}_n, \widehat{\mathbf{y}}_n - \overline{y}_n\mathbf{1}_n\rangle}{\|\mathbf{y}_n - \overline{y}_n\mathbf{1}_n\|\,\|\widehat{\mathbf{y}}_n - \overline{y}_n\mathbf{1}_n\|},$$

and since, using the fact that $\langle \widehat{\mathbf{y}}_n - \mathbf{y}_n, \mathbf{1}_n\rangle = 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \frac{1}{n}\langle \widehat{\mathbf{y}}_n, \mathbf{1}_n\rangle = \frac{1}{n}\langle \mathbf{y}_n, \mathbf{1}_n\rangle = \overline{y}_n,$$

we deduce that we may rewrite $\cos\alpha = \mathrm{Corr}(\mathbf{y}_n, \widehat{\mathbf{y}}_n)$. The expected identity now follows from the fact that $\sin^2\alpha = 1 - \cos^2\alpha = 1 - R^2$. □

The coefficient R^2 still measures the fit between the model and the data, and therefore a low value is a good indicator of underfitting. However, it does not prevent overfitting. Indeed, assume that you add a feature to your model, that is to say that you add a column to the design matrix $\mathbf{x}_n$, and denote by $\mathbf{x}'_n$ the resulting design matrix. By construction, the range of $\mathbf{x}_n$ is included in the range of $\mathbf{x}'_n$, and therefore the orthogonal projection $\widehat{\mathbf{y}}'_n$ of $\mathbf{y}_n$ onto the range of $\mathbf{x}'_n$ is necessarily closer to $\mathbf{y}_n$ than $\widehat{\mathbf{y}}_n$, namely

$$\|\mathbf{y}_n - \widehat{\mathbf{y}}'_n\| \leq \|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|.$$

By Proposition A.4.5, this implies that the coefficient of determination R'^2 of the linear regression on $\mathbf{x}'_n$ satisfies $R'^2 \geq R^2$. In summary, *adding features to the model will always increase the coefficient* R^2. The values of R^2 as a function of the order of the polynomial for the example of polynomial regression of Section A.2 are reported in Table A.1: clearly, models which achieve the highest R^2 are overfitting.

Order p	2	4	6	8	10
R^2	0.28	0.971	0.974	0.979	0.980

Table A.1: Values of R^2 for polynomial regression on the example of Figure A.5.

A.5 Variance and regularisation

A.5.1 Statistical properties of the OLS

So far, we have only considered regression as a model-fitting problem: the data points $(x_i, y_i)_{1\leq i\leq n}$ are given and we look for a parameter β which makes the linear model closest to these points. In the statistical approach introduced in Lecture 2, one assumes that the data points $(x_i, y_i)_{1\leq i\leq n}$ are generated

by a random procedure, namely that there exists a parameter $\beta \in \mathbb{R}^{p+1}$ and random variables $\epsilon_1, \dots, \epsilon_n$ such that

$$\forall i \in \{1, \dots, n\}, \qquad y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i,$$

which rewrites in vector notation

$$\mathbf{y}_n = \mathbf{x}_n \beta + \boldsymbol{\epsilon}_n, \qquad \boldsymbol{\epsilon}_n := \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{A.4}$$

The variables ϵ_i are not observed, but the observation of the pairs (x_i, y_i) allows to infer the value of β on the one hand, and quantities of interest related to the distribution of $\boldsymbol{\epsilon}_n$ on the other hand.

Let us make the following assumptions on the variables $\epsilon_1, \dots, \epsilon_n$:

- for all $i \in \{1, \dots, n\}$, $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$;
- if $i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$.

In other words, $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$ and $\mathrm{Cov}[\boldsymbol{\epsilon}_n] = \sigma^2 I_n$.

Proposition A.5.1 (Bias and covariance of the OLS)**.** *In the setting above, and under the assumptions of Proposition A.4.1, the OLS is unbiased, in the sense that for all $\beta \in \mathbb{R}^{p+1}$, if $(x_i, y_i)_{1 \leq i \leq n}$ satisfy (A.4), then $\widehat{\beta}$ defined by (A.3) satisfies*

$$\mathbb{E}[\widehat{\beta}] = \beta.$$

Moreover, its covariance matrix is given by

$$\mathrm{Cov}[\widehat{\beta}] = \sigma^2 \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1}.$$

Proof. Combining (A.4) with Proposition A.4.1 we get

$$\widehat{\beta} = \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top \left(\mathbf{x}_n \beta + \boldsymbol{\epsilon}_n\right) = \beta + \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top \boldsymbol{\epsilon}_n.$$

By the linearity of the expectation, we deduce

$$\mathbb{E}[\widehat{\beta}] = \beta + \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top \mathbb{E}[\boldsymbol{\epsilon}_n] = \beta,$$

since, by assumption, $\mathbb{E}[\boldsymbol{\epsilon}_n] = 0$. Moreover, by standard properties of covariance matrices[8],

$$\mathrm{Cov}[\widehat{\beta}] = \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top \mathrm{Cov}[\boldsymbol{\epsilon}_n] \left(\left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1} \mathbf{x}_n^\top\right)^\top = \sigma^2 \left(\mathbf{x}_n^\top \mathbf{x}_n\right)^{-1},$$

since, by assumption, $\mathrm{Cov}[\boldsymbol{\epsilon}_n] = \sigma^2 I_n$. □

Exercise A.5.2. Assume that $p + 1 < n$. Show that

$$\widehat{\sigma}^2 := \frac{1}{n - (p+1)} \|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2$$

is an unbiased estimator of σ^2.

Our purpose is now to show an optimality property of the OLS. We recall from Remark 2.1.12 that if two unbiased estimators $\widetilde{\beta}^1$ and $\widetilde{\beta}^2$ of β satisfy $\mathrm{Cov}[\widetilde{\beta}^1] \preceq \mathrm{Cov}[\widetilde{\beta}^2]$, then it follows that $\mathrm{MSE}(\widetilde{\beta}^1) \leq \mathrm{MSE}(\widetilde{\beta}^2)$. We first define the class of estimators in which we will show the optimality of the OLS.

[8] Recall that if X is a d-dimensional random vector with $\mathrm{Cov}[X] = K$ and $A \in \mathbb{R}^{p \times d}$, then $\mathrm{Cov}[AX] = AKA^\top$.

Definition A.5.3 (Linear estimators)**.** *An estimator $\widetilde{\beta}$ of β is called* linear *if there exists a matrix $\mathbf{A}_n \in \mathbb{R}^{(p+1)\times n}$, which may depend on $\mathbf{x}_n$ but not on β, such that $\widetilde{\beta} = \mathbf{A}_n \mathbf{y}_n$.*

Theorem A.5.4 (Gauss–Markov Theorem)**.** *For any an unbiased linear estimator $\widetilde{\beta}$ of β, we have*

$$\mathrm{Cov}[\widehat{\beta}] \preceq \mathrm{Cov}[\widetilde{\beta}].$$

Therefore, the OLS is called the Best Linear Unbiased Estimator *of β in this model.*

Proof. Let $\widetilde{\beta} = \mathbf{A}_n \mathbf{y}_n = \mathbf{A}_n(\mathbf{x}_n\beta + \boldsymbol{\epsilon}_n)$ be a linear unbiased estimator of β. Since $\widetilde{\beta}$ is unbiased, we first get $\mathbb{E}[\widetilde{\beta}] = \mathbf{A}_n\mathbf{x}_n\beta = \beta$. This identity holds for any $\beta \in \mathbb{R}^{p+1}$ (see the statement of Proposition A.5.1), therefore $\mathbf{A}_n\mathbf{x}_n = I_{p+1}$. On the other hand, we have $\mathrm{Cov}[\widetilde{\beta}] = \sigma^2 \mathbf{A}_n\mathbf{A}_n^\top$. Introducing the matrix $\mathbf{D}_n = \mathbf{A}_n - (\mathbf{x}_n^\top\mathbf{x}_n)^{-1}\mathbf{x}_n^\top$, a straightforward computation yields

$$\begin{aligned}
\mathbf{D}_n\mathbf{D}_n^\top &= \left(\mathbf{A}_n - (\mathbf{x}_n^\top\mathbf{x}_n)^{-1}\mathbf{x}_n^\top\right)\left(\mathbf{A}_n^\top - \mathbf{x}_n(\mathbf{x}_n^\top\mathbf{x}_n)^{-1}\right) \\
&= \mathbf{A}_n\mathbf{A}_n^\top - \mathbf{A}_n\mathbf{x}_n(\mathbf{x}_n^\top\mathbf{x}_n)^{-1} - (\mathbf{x}_n^\top\mathbf{x}_n)^{-1}\mathbf{x}_n^\top\mathbf{A}_n^\top + (\mathbf{x}_n^\top\mathbf{x}_n)^{-1} \\
&= \mathbf{A}_n\mathbf{A}_n^\top - (\mathbf{x}_n^\top\mathbf{x}_n)^{-1},
\end{aligned}$$

where we have used the fact that $\mathbf{A}_n\mathbf{x}_n = I_{p+1}$ at the last line. Since, by construction, $\mathbf{D}_n\mathbf{D}_n^\top \succeq 0$, we conclude that $\mathbf{A}_n\mathbf{A}_n^\top \succeq (\mathbf{x}_n^\top\mathbf{x}_n)^{-1}$, which by Proposition A.5.1 completes the proof. □

A.5.2 Regularisation

In Proposition A.4.1, if $p+1 \leq n$ but the rank of $\mathbf{x}_n$ is lower than $p+1$, then at least one of the columns of $\mathbf{x}_n$ is a linear combination of the other ones, so it may be removed from $\mathbf{x}_n$ without loss of information, and the process can be iterated until the matrix $\mathbf{x}_n$ recovers full rank. Still, this method requires the number of features p to be lower than the number of observations n.

However, in high-dimensional statistics, it is often the case that $p > n$, so that the matrix $\mathbf{x}_n^\top\mathbf{x}_n$ is necessarily degenerate. In such a situation, the OLS is no longer well-defined as there is a (possibly high-dimensional) space of vectors $\widehat{\beta}$ such that $\mathbf{x}_n\widehat{\beta} = \widehat{\mathbf{y}}_n$. To recover uniqueness, a usual approach consists in adding a *penalisation* in the OLS problem, so as to bias the selection of β toward certain directions. For example, adding the squared Euclidean norm of β to the sum of squares leads to the minimisation problem

$$\min_{\beta\in\mathbb{R}^{p+1}} \left\{\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2 + h\|\beta\|^2\right\}, \tag{A.5}$$

which is known as *ridge regression*[9], or *Tikhonov regularisation*. Here, the parameter $h > 0$ determines the intensity of this bias: the larger it is, the more the optimisation will tend to select an estimator with a small Euclidean norm. It therefore plays a similar role to bandwidth in kernel smoothing.

Exercise A.5.5.

1. Show that, even if the matrix $\mathbf{x}_n^\top\mathbf{x}_n \in \mathbb{R}^{(p+1)\times(p+1)}$ is not invertible (which is in particular the case if $n < p+1$), then the minimiser of (A.5) exists and is unique, and give its expression. It is denoted by $\widehat{\beta}_h$.

2. In the setting of Subsection A.5.1, show that $\widehat{\beta}_h$ is biased, and that if the OLS $\widehat{\beta}$ is well-defined, $\mathrm{Cov}_\beta[\widehat{\beta}_h] \preceq \mathrm{Cov}_\beta[\widehat{\beta}]$. *Hint: you may remark that $\mathrm{Cov}_\beta[\widehat{\beta}_h]$ and $\mathrm{Cov}_\beta[\widehat{\beta}]$ have the same eigenvectors and compare the associated eigenvalues.*

Exercise A.5.5 shows that the introduction of penalisation has two main consequences:

- it makes the minimisation problem well-posed even if $p > n$ (therefore, the addition of the penalisation is also called *regularisation*);

[9] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1958.

- even in cases where the OLS is well-defined, it yields an estimator which is biased, but has a smaller covariance matrix than the OLS[10]. As a consequence, regularisation is helpful to reduce overfitting.

The second point is definitely reminiscent of the situation of Exercise 2.1.14, where the introduction of bias allows to reduce the variance and thus to globally decrease the MSE. Similarly, it finds a natural interpretation in the context of Bayesian inference seen in Lecture 6.

There are of course other possible penalisations. For instance, it is generally assumed that the vector β is *sparse*[11], that is to say that many of its coordinates are 0. The underlying paradigmatic idea is that at each experiment or observation, many data are collected, but only a few of these data are actually relevant. To compute an estimator of β in such problems, it is natural to add a term taking the sparsity constraint into account, and thus one may try to solve the minimisation problem

$$\min_{\beta\in\mathbb{R}^{p+1}} \left\{\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2 + h\|\beta\|_0\right\},$$

where $h > 0$ still determines the intensity of the penalisation, and $\|\beta\|_0$ denotes the number of nonzero coordinates of β: an optimal solution should then have a low $\|\cdot\|_0$-norm. This problem is actually computationally difficult, and it may be fruitfully relaxed to the minimisation problem

$$\min_{\beta\in\mathbb{R}^{p+1}} \left\{\|\mathbf{y}_n - \mathbf{x}_n\beta\|^2 + h\|\beta\|_1\right\},$$

where $\|\beta\|_1 = \sum_{j=0}^{p} |\beta_j|$. The obtained estimator is called *LASSO* (for *Least Absolute Shrinkage and Selection Operator*), and was introduced by Tibshirani[12] in 1996. It is nowadays considered as a standard variable selection tool.

A.6 Logistic regression

Unlike what you could expect from its name, logistic regression is not a method of regression as introduced in Section A.2, but a method of *binary classification*. Indeed, in this Section, we assume that we observe data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \{0, 1\}$, that is to say that the explained variable y is not continuous but binary.

In this setting, it is not very appropriate to look for a function f such that $f(x_i) \simeq y_i$, because in general y_i is not expected to be a deterministic function of x_i, but rather a random variable whose law depends on x_i. More concretely, logistic regression (and, more generally, probabilistic binary classification methods) typically aims at answering questions such as: how does the probability that a pedestrian be killed in a collision with a car depend on the weight of the car?[13] In this example, the death of the pedestrian certainly depends on the weight of the car, but also on many other factors, so it makes sense to see it as a Bernoulli variable, with a parameter $p(x)$ which depends on the weight x of the car.

Just like linear regression is a particular choice of a parametric family of functions $\{f_\beta, \beta \in \mathbb{R}^{p+1}\}$ to approximate the relation between explanatory variables $x^1, \dots, x^p$ and a continuous variable y, logistic regression is a particular choice of functions $\{p_\beta, \beta \in \mathbb{R}^{p+1}\}$ to approximate the probability $p(x)$ that y takes the value 1. This model writes

$$\forall \beta \in \mathbb{R}^{p+1}, \quad \forall x \in \mathbb{R}^p, \qquad p_\beta(x) = \Psi\left(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p\right),$$

where $\Psi : \mathbb{R} \to (0, 1)$ is the *logistic* (or *logit*) function

$$\forall u \in \mathbb{R}, \qquad \Psi(u) = \frac{\exp(u)}{1 + \exp(u)} = \frac{1}{1 + \exp(-u)},$$

whose graph is plotted on Figure A.9.

[10] Since $\widehat{\beta}$ is biased, there is no contradiction with the Gauss–Markov Theorem.

[11] *Parcimonieux* en français.

[12] R. Tibshirani. Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[13] The answer is here: https://doi.org/10.1016/j.ecotra.2024.100342.

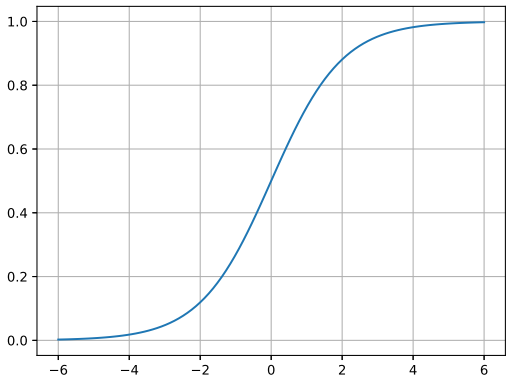


Figure A.9: The logistic function Ψ, which takes its values between 0 and 1.

Now that we have specified our parametric model $p_\beta(x)$ for $p(x)$, the question is how to estimate β from the data $\mathbf{x}_n$, $\mathbf{y}_n$? A possible approach is based on the *maximum likelihood principle*. If we let $Y_1, \dots, Y_n$ be independent Bernoulli variables with respective parameters $p_\beta(x_1), \dots, p_\beta(x_n)$, then the probability to observe the realisation $y_1, \dots, y_n$ is

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} p_\beta(x_i)^{y_1}(1 - p_\beta(x_i))^{1-y_i}.$$

This quantity is denoted by $L_n(\mathbf{x}_n, \mathbf{y}_n; \beta)$ and called the *likelihood* of the realisation $(\mathbf{x}_n, \mathbf{y}_n)$. The maximum likelihood principle consists in estimating β by

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\max}\, L_n(\mathbf{x}_n, \mathbf{y}_n; \beta),$$

which amounts to taking the value of the parameter which makes the observed realisation the most probable. A much more general study of this principle will be carried out in Lecture 5. The estimator $\widehat{\beta}$ is called the *Maximum Likelihood Estimator* (MLE). Unlike for the OLS in linear regression, there is no closed-form expression for the MLE, but it can be computed by numerical algorithm.

Appendix B

Flipped Classrooms

General recommandations

Your presentation must be about twenty-minute long, and contain a theoretical part: you will teach your classmates the contents of the lecture notes just as you would expect your lecturer to do; and an applicative part: you will report on the experimental application of your method that you made by yourself. The presentation has to be accompanied by slides, both for the theoretical and the applicative parts.

In the theoretical part, provide a global picture of the method (what is the overall purpose, in which framework is it employed, how is it applied) rather than insisting on mathematical details: they are overwhelmingly long to be presented, so it is better to refer your audience to the lecture notes.

There are several possible starting points to determine which illustrative application you will present.

- Based on the descriptions below, you may look for a real-world dataset on the Internet which you think might fit your method well. A list of dataset repositories is available on Educnet.

- You may also directly browse specialised websites, such as Kaggle or Towards Data Science, to find suitable datasets.

- If you ask ChatGPT to provide you with some examples of application of your method, it will likely return made-up but in principle good ideas, unfortunately without reliable sources. You may try Perplexity.ai to get sources to blogs or tutorials, which may be an interesting starting point.

- You may look for scientific papers which apply your method and present, discuss or even reproduce their results. Good sources to do so are Wikipedia and the website Papers With Code.

- It is also possible to make a more theoretical discussion with synthetic data, that you simulated by yourself, so as to investigate the limits of your method. For example, if your method relies on the assumption that samples are Gaussian, it may be interesting to show that it 'works' on data that you have simulated with a Gaussian distribution, then to check whether it continues to work if the data are 'less and less' Gaussian.

During the preparation of your presentation, it is also useful to check the tips and common errors to avoid described in Appendix C of these notes. Last, keep in mind that when you present the result of a hypothesis test, it is much more indicative to give the p-value, which is immediately interpretable, than the value of a test statistic (which depends on the test). When it is possible, do not hesitate to discuss the power of your test as well.

B.1 The EM Algorithm for Mixture of Gaussians

B.1.1 Introduction

The Expectation-Maximization (EM) algorithm is an iterative procedure for maximum likelihood estimation when the data involve hidden or unobserved variables. A central application is in mixture models, where each observation is assumed to come from one of several latent components, but the component labels are not observed.In this setting, EM alternates between an *expectation step* and a *maximization step*. Applied to mixture models, EM provides a natural way to "softly" assign data points to components while refining the parameters of each distribution. It is particularly attractive because it turns a seemingly intractable estimation problem into a sequence of simpler, interpretable steps.

B.1.2 Instructions

Your presentation will be based on Section 5.3.2 in Lecture 5. You will present the framework in which the Expectation-Maximization test is typically applied. To illustrate the algorithm, you may begin to work with the historical example of Weldon's crab classification, but are also strongly encouraged to find your own examples. Generally speaking, you are expected to present the motivation for EM as a way to perform maximum likelihood estimation in latent-variable models, emphasizing that a direct maximization of the likelihood is often intractable due to the coupling between hidden labels and parameters. In your example of application, you will try to compare the EM algorithm with other unsupervised clustering methods seen in the data analysis course such as k-means.

B.2 The Likelihood Ratio Test

B.2.1 Introduction

The Likelihood Ratio Test (LRT) is a statistical method used for comparing two statistical models to determine which model provides a better fit to the data. It is commonly employed in hypothesis testing when you want to assess whether a more complex model significantly improves the fit compared to a simpler model. It is based on comparing the likelihoods of two models: a null model (simpler) and an alternative model (more complex). The likelihood represents how well each model explains the observed data. The LRT calculates a test statistic, often denoted as the likelihood ratio statistic, which quantifies the improvement in fit between the two models.

B.2.2 Instructions

Your presentation will be based on Section 9.2 in Lecture 7. You will present the framework of the test, when the two hypotheses are simple but no obvious test statistic is available. You may use Example 9.2.1 or find another one. You will present the likelihood ratio and its asymptotic behaviour, and deduce the construction of the likelihood ratio test. You will then state the Neyman–Pearson optimality result. In your example of application, you will try to compare the power of the likelihood ratio test with other possible tests (similarly to the idea of the test constructed in Exercise 9.2.2).

B.3 Analysis of Variance in the Gaussian Model

B.3.1 Introduction

ANOVA is a statistical technique used to analyse the variation between two or more groups or treatments to determine if there are statistically significant differences among them. It is particularly useful when you want to compare means from multiple groups simultaneously. ANOVA calculates two types of variation: variation between groups and variation within groups. It then compares these variations to determine if there is evidence of significant differences among the groups.

B.3.2 Instructions

Your presentation will be based on Section 8.4 in Lecture 8. You will present the context and the hypotheses of the test. You will then explain the interpretation of the quantities SSM and SSE, and describe the law of these quantities under H_0. You will then present the geometric interpretation of SSM and SSE, and derive the construction of the Fisher test. In your example of application, you will be careful to the fact that your samples need to be Gaussian and have the same variance.

B.4 Wald's Test for the Identity of Means

B.4.1 Introduction

Wald's test is a statistical method used to compare the means of two independent samples to determine if they are significantly different from each other. This test is used when you want to assess whether there is evidence to suggest that the means of two groups are equal or not.

B.4.2 Instructions

Your presentation will be based on Subsection 9.1.2 in Lecture 7. You will present the framework of the test and the hypotheses to be tested. You will detail the construction of the test and emphasise the similarity with Wald's one sample test from Subsection 9.1.1. You will discuss the specific application of the test to the problem of comparison of proportions in the Bernoulli model. Your example of application may cover the Bernoulli model or another one.

B.5 The Shapiro–Wilk Test

B.5.1 Introduction

The Shapiro-Wilk test is a statistical test used to assess whether a given dataset follows a Gaussian distribution. It is a common test for checking the normality assumption, which is important in many statistical analyses. The test provides a p-value that indicates whether there is enough evidence to conclude that the data significantly deviate from a normal distribution.

B.5.2 Instructions

Your presentation will be based on Section 11.4 in Lecture 11. You will introduce in detail the construction of the estimator $\widehat{\sigma}^2$ based on the Generalised Least Square formulation. You will then describe the properties of the Shapiro–Wilk statistic on which the test is based. Your example of application may include a comparison of this test with the Lilliefors correction for the Gaussian goodness-of-fit.

B.6 Kolmogorov–Smirnov Test of Homogeneity

B.6.1 Introduction

The two-sample Kolmogorov-Smirnov test, often referred to as the KS test, is a non-parametric statistical test used to compare the distributions of two independent samples. The KS test assesses the similarity between two datasets by comparing their empirical cumulative distribution functions (CDFs). It calculates a test statistic that quantifies the maximum vertical distance between the two empirical CDFs. The larger the test statistic, the more dissimilar the distributions are.

B.6.2 Instructions

Your presentation will be based on Section 11.3 in Lecture 11. You will first present the global issue of nonparametric homogeneity tests. You will then introduce the notion of QQ-plot and explain how it helps to visually assess whether two samples are drawn from the same distribution or not. You will finally introduce the Kolmogorov–Smirnov statistic, and explain how the Kolmogorov–Smirnov test works. In your example of application it will be important both to show QQ-plots and compute p-values.

Appendix C

Tips for your Statistical Studies

Respect the chronological path of decisions. Make sure that you have well understood the structure of the data and that you have well established the question that you want to investigate (i.e. clearly write the null hypothesis and the alternative). Otherwise, it is likely that you will choose the wrong test for your particular data and/or for the proposed hypothesis.

Do not forget that H_0 and H_1 do not play symmetric roles (and understand why). Testing procedures are designed to provide a theoretical control of the type I error. Then, deriving theoretical guarantees regarding the type II error is not impossible but more involved. In most cases, the theoretical guarantees that can be provided regarding the control type II error are asymptotic (and typically correspond to a power of the test tending to 1 as n goes to $+\infty$). Non asymptotic theoretical control on the power is in general difficult to obtain and often requires to work on a specific class of alternatives.

Missing the concrete difference between homogeneity tests and independence tests. Keep in mind that homogeneity tests aim at understanding if two (or even more) samples have been drawn from the same population, whereas independent tests are concerned with whether one attribute is independent of the other and involve a single sample from the population.

Double-dipping in the data. Make sure that you are not evaluating the prediction performance of a model with instances used to train it.

Correlation is not causation. Just because frogs appear after the rain does not mean that frogs have rained!

Do not mix up the notions of bias and consistency. Let us consider random variables $(X_i)_{1\leq i\leq n}$ iid with $X_1 \sim \mathcal{B}(p)$ (for some $p \in (0,1)$). Then,

- X_1 is an unbiased but not consistent estimator of p;
- $\frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased and consistent estimator of p;
- $\frac{1}{n+1}\sum_{i=1}^n X_i$ is a biased and consistent estimator of p;
- $X_1/2$ is a biased and not consistent estimator of p.

Appendix written by Q. Duchemin.

Make sure that your dataset fits with the assumptions of the statistical test used. The typical assumption that you may check is the normality assumption. But do not forget that for ANOVAs for example, a continuous dependent variable is an important assumption. If the dependent variable only takes a few different values, it may not be legitimate to analyse it as a continuous variable. Another possible error is the inappropriate use of χ_2 test when numerical value in a cell is less than 5.

Make sure you did not apply unpaired t-test for paired data. A common error is made during the computation of paired and unpaired data. It is necessary for the measurements of two different groups that unpaired observations should be distinguished — for example, patients receiving alternative therapeutic regimens — from that of paired observations, when the comparison is done between two measurements made on the same individuals at different time intervals. For unpaired data, two sample t-test, Mann–Whitney U-test and χ_2 test are useful whereas for paired data the common paired t-test, Wilcoxon test and McNemar test are used.

Be curious about how to put theory into practice. Some possible questions that you should think about are the following.

- How does a random number generator work?
- Which optimisation algorithm is used to estimate the coefficients in a logistic regression?
- Is this test robust to model misspecification?

How to deal with a qualitative variable in a regression model? The standard technique is the so-called *One-Hot Encoding* method. If some feature X_j takes values in a finite space E of small cardinality d, one can create d dummy variables taking values in $\{0, 1\}$ so that every unique value in E will be added as a feature.

How to deal with missing values? There are several usual ways to handle missing data:

- delete rows with missing data;
- Mean/Median/Mode imputation;
- assigning a unique value;
- predicting the missing values;
- using an algorithm which supports missing values, like random forests.

The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. Indeed, let us describe the drawbacks of the mean imputation.

Mean imputation is the practice of replacing null values in a data set with the mean of the data. Mean imputation is generally bad practice because it does not take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should. Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

How to deal with outliers? Keep in mind that theoretical results provided in the course are more or less robust to model misspecification. Let us consider the example of outliers. Suppose that we have a sample of size n where each variable is drawn from a standard normal distribution except that 1/20 of the dataset is replaced by samples drawn from a normal distribution with unit variance but a mean of 100. If we want to estimate the mean value 0 of the original samples, using the empirical mean is not the best choice possible since it is highly sensitive to outliers. One possible solution consists in using the so-called **median of means technique** which works as follows.

1. Shuffle the data points and then splits them into k groups of $\lfloor n/k \rfloor$ data points.
2. Compute the Arithmetic Mean of each group.
3. Finally, calculate the median over the resulting k Arithmetic Means.

How to select the testing procedure? Let us give a motivation with a concrete example. The t-test and the Wilcoxon Rank Sum test can be both used to compare the mean of two samples. The difference is that the t-test assumes that the samples being tested are drawn from a normal distribution, while Wilcoxon Rank Sum test does not. Hence, a natural question is why would we use the t-test if the Wilcoxon Rank Sum test addresses the same problem with fewer model constraints? One short answer to this question is the usual *No free lunch* principle. If the Wilcoxon Rank Sum test can be used beyond the Gaussian model, this should lead to a less powerful test in the Gaussian model where both tests can be used.

A (non exhaustive) list of common errors in statistics and related fields

Sampling errors. A sample is a subset of the targeted population. This implies that every dataset may contain groups that are over/under represented. Such unbalance will influence (sometimes dramatically) the results of an analysis. As a consequence, this is a powerful tool to manipulate opinions (in medias or in politics...). Note that unbalanced dataset is not always due to malicious people who want to use the data for their benefit. Dealing with unbalanced datasets is often intrinsic to the problem tackled. This is typically the case when you try to identify rare cases in rather big datasets (e.g. in fraud detection). Confidence intervals or p-values are tools that allow to relieve the uncertainty due to such unbalanced structure. However, there are some cases where finding confidence intervals can be tricky (for example for regression coefficients). Methods have been proposed to cope with such situations. One can mention the example of the bootstrapping method.

Using methods beyond their theoretical framework of application. For example in multiple regression, distributions of variables need to be normal and relationships need to be linear. Another problem comes from the potential mulcolinearity of variables. These errors are commom in the scientific community[1].

Errors due to bias.

1. *Cherry picking* consists in selecting people, facts or situations that correspond to what we want to show and then to generalise the results.
2. Error in measurements.
3. Error in observations: in a scientific experiment, it can be tempting to see what we want to see. To avoid this kind of issue, one can use the double blind method.

[1] P. Schrodt, Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 2014.

Explicative errors.

1. *Approximation error*. A model is always a simplified representation of a complex real phenomenon. There is always an error lying in the part non-explained by the model. Pearson and Kendall's R, Cohen's D or Cramer's V are tools allowing to measure this error.

2. *Specification error*. It is likely that at least one of the key features of the problem is not available. In consequence, estimation of the model may yield results that are incorrect or misleading.

Appendix D

Correction of Exercises

D.1 What you must know in Probability Theory

Correction of Exercise 1.A.1 The following functions respectively return a realisation of $\mathcal{B}(p)$, $\mathcal{B}(n,p)$ and $\mathcal{G}(p)$ variables, using `if`, `for` and `while` statements.

```
from scipy.stats import uniform

def bernoulli(p):
    # p must be between 0 and 1
    U = uniform.rvs()  # Generate a random variable from the uniform
        distribution on [0, 1)
    if U <= p:
        return 1
    else:
        return 0

def binomial(n, p):
    # n must be an integer, p must be between 0 and 1
    s = 0
    for _ in range(n):
        s += bernoulli(p)
    return s

def geometric(p):
    # p must be between 0 (strictly) and 1
    i = 1
    while bernoulli(p) == 0:
        i += 1
    return i
```

Correction of Exercise 1.A.2

1. A probability larger than 1 is not possible.

2. The underlying assumption is that the $N \times M$ events 'one nuclear reactor has a serious accidents during one year' are independent, with identical probability p. So the probability that no incident occurs in $(1-p)^{NM}$, and thus the probability that at least one incident occurs is $1-(1-p)^{NM}$. Remark that in the $p \to 0$ limit, the Taylor expansion yields the formula $p \times N \times M$ used by the authors in the article — but this is only an approximation. With this corrected formula and for the values $p = 4/14000$, $N = 143$ and $M = 30$ one gets a probability of 0.7. This is still very large! However, both the estimation of p and the assumptions on the events are questionnable.

Correction of Exercise 1.A.3 Von Neumann's solution consists in throwing the coin twice at each toss: you thus obtain iid pairs $(X_i, Y_i)_{i\geq 1}$, which have law $\mathcal{B}(p)\otimes\mathcal{B}(p)$. You then keep the first toss for which $X_i \neq Y_i$, that is to say that you set $N = \inf\{i \geq 1 : X_i \neq Y_i\}$. Notice that N is a geometric random variable, with parameter $\mathbb{P}(X \neq Y) = \mathbb{P}(X = 0, Y = 1) + \mathbb{P}(X = 1, Y = 0) = 2p(1-p)$. Then it turns out that $X_N \sim \mathcal{B}(1/2)$. Indeed,

$$\begin{aligned}\mathbb{P}(X_N = 1) &= \sum_{n=1}^{+\infty} \mathbb{P}(X_n = 1, n = N)\\ &= \sum_{n=1}^{+\infty} \mathbb{P}(X_1 = Y_1, \dots, X_{n-1} = Y_{n-1}, X_n = 1, Y_n = 0)\\ &= \sum_{n=1}^{+\infty} \mathbb{P}(X_1 = Y_1)^{n-1}\mathbb{P}(X_1 = 1, Y_1 = 0),\end{aligned}$$

where we have used the fact that the pairs (X_i, Y_i) are iid. Now, on the one hand,

$$\mathbb{P}(X_1 = Y_1) = \mathbb{P}(X_1 = 0, Y_1 = 0) + \mathbb{P}(X_1 = 1, Y_1 = 1) = (1-p)^2 + p^2,$$

while on the other hand, $\mathbb{P}(X_1 = 1, Y_1 = 0) = p(1-p)$. We deduce that

$$\mathbb{P}(X_N = 1) = p(1-p)\sum_{n=1}^{+\infty}\left((1-p)^2 + p^2\right)^{n-1} = \frac{p(1-p)}{1-((1-p)^2+p^2)} = \frac{1}{2}.$$

With an unbalanced dice, one may do the same. More generally, assume that you have a device which returns iid random variables X_i in a finite set E with cardinality m, whose law P is unknown. The goal is to use this device to draw a random variable Z in E with uniform distribution. To proceed, sample independent batches $(X_i^{(1)}, \dots, X_i^{(m)})_{i\geq 1}$ and let N be the first index i for which all values of $\{X_i^{(1)}, \dots, X_i^{(m)}\}$ are different from each other. Then $(X_N^{(1)}, \dots, X_N^{(m)})$ equals all $m!$ permutations of E with the same probability, and in particular, $Z = X_N^{(1)}$ is uniformly distributed in E.

Correction of Exercise 1.A.4 We first fix $Y \sim \Gamma(a, 1)$ and compute $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$. First,

$$\mathbb{E}[Y] = \int_{y=0}^{+\infty} y\frac{1}{\Gamma(a)}y^{a-1}\mathrm{e}^{-y}\mathrm{d}y = \frac{\Gamma(a+1)}{\Gamma(a)}.$$

Since $\Gamma(a+1) = a\Gamma(a)$, we deduce that $\mathbb{E}[Y] = a$. With similar arguments, we now have

$$\mathbb{E}[Y^2] = \int_{y=0}^{+\infty} y^2\frac{1}{\Gamma(a)}y^{a-1}\mathrm{e}^{-y}\mathrm{d}y = \frac{\Gamma(a+2)}{\Gamma(a)} = (a+1)a.$$

Therefore, $\mathrm{Var}(Y) = (a+1)a - a^2 = a$. Finally, for any $\lambda > 0$, by Exercise 1.3.12, if $X \sim \Gamma(a, \lambda)$ then $Y := \lambda X \sim \Gamma(a, 1)$, so that we conclude that $\mathbb{E}[X] = \mathbb{E}[Y]/\lambda = a/\lambda$ and $\mathrm{Var}(X) = \mathrm{Var}(Y)/\lambda^2 = a/\lambda^2$.

Correction of Exercise 1.A.5 For any integer $k \geq 1$, we have

$$\mathbb{E}[G^{2k-1}] = \frac{1}{\sqrt{2\pi}}\int_{x\in\mathbb{R}} x^{2k-1}\mathrm{e}^{-x^2/2}\mathrm{d}x.$$

It is clear that the integrand is odd and therefore the integral vanishes. We now write

$$\begin{aligned}\mathbb{E}[G^{2k}] &= \frac{1}{\sqrt{2\pi}}\int_{x\in\mathbb{R}} x^{2k}\mathrm{e}^{-x^2/2}\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{x\in\mathbb{R}} x^{2k-1}\times x\mathrm{e}^{-x^2/2}\mathrm{d}x\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\left\{\left[x^{2k-1}\times(-\mathrm{e}^{-x^2/2})\right]_{-\infty}^{+\infty} - \int_{x\in\mathbb{R}}(2k-1)x^{2k-2}\times(-\mathrm{e}^{-x^2/2})\mathrm{d}x\right\} \\ &= \frac{2k-1}{\sqrt{2\pi}}\int_{x\in\mathbb{R}} x^{2k-2}\mathrm{e}^{-x^2/2}\mathrm{d}x \\ &= (2k-1)\mathbb{E}[G^{2k-2}].\end{aligned}$$

By induction, and since $\mathbb{E}[G^0]=1$, we conclude that $\mathbb{E}[G^{2k}] = (2k-1)\times(2k-3)\times 3\times 1$.

Correction of Exercise 1.A.6

1. For any measurable and bounded function $f:\mathbb{R}^2\to\mathbb{R}$,

$$\mathbb{E}[f(Z,U)] = \int_{x,y>0} f\left(x+y,\frac{x}{x+y}\right)\frac{1}{\Gamma(a)}x^{a-1}\mathrm{e}^{-x}\frac{1}{\Gamma(b)}y^{b-1}\mathrm{e}^{-y}\mathrm{d}x\mathrm{d}y.$$

   Let us set $z=x+y$, $u=x/(x+y)$, so that $x=uz$, $y=(1-u)z$. Then $(x,y)\in(0,+\infty)^2$ if and only if $(z,u)\in(0,+\infty)\times(0,1)$, and $\mathrm{d}x\mathrm{d}y = z\mathrm{d}u\mathrm{d}z$. Therefore,

$$\mathbb{E}[f(Z,U)] = \frac{1}{\Gamma(a)\Gamma(b)}\int_{u=0}^{1}\int_{z=0}^{+\infty} f(z,u)\,u^{a-1}(1-u)^{b-1}z^{a+b-1}\mathrm{e}^{-z}\mathrm{d}z\mathrm{d}u.$$

   This shows that (Z,U) has density

$$\frac{1}{\Gamma(a+b)}z^{a+b-1}\mathrm{e}^{-z}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}u^{a-1}(1-u)^{b-1}$$

   on $(0,+\infty)\times(0,1)$.

2. Taking $f=1$ we get the desired identity.

3. The density of (Z,U) now writes as the product of the density of $\Gamma(a+b,1)$ (in z) and $\beta(a,b)$ (in u). So Z and U are independent, with respective distributions $\Gamma(a+b,1)$ and $\beta(a,b)$.

4. Let $X\sim\Gamma(a,1)$, $Y\sim\Gamma(b,1)$, $Z\sim\Gamma(c,1)$, independent from each other. Let us set

$$U=\frac{X}{X+Y},\qquad V=\frac{X+Y}{X+Y+Z}.$$

   By the previous question, $U\sim\beta(a,b)$, $V\sim\beta(a+b,c)$ and these variables are independent, because V is a function of $(X+Y,Z)$ and U is independent from $X+Y$ as well as from Z. We now remark that

$$UV=\frac{X}{X+Y+Z}$$

   so $UV\sim\beta(a,b+c)$.

Correction of Exercise 1.A.8

1. We have $x_1^1 = 15$, $x_2^1 = 12$ and $x_3^1 = 18$, so that $x_2^1 < x_1^1 < x_3^1$. As a consequence, $r_1^1 = 2$ (because x_1^1 is ranked second), $r_2^1 = 1$ and $r_3^1 = 3$. Likewise, $x_1^2 = 9$, $x_2^2 = 7$ and $x_3^2 = 8$, so that $x_2^2 < x_3^2 < x_1^2$ and $r_1^2 = 3$, $r_2^2 = 1$ and $r_3^2 = 2$. In the definition

$$r_{\mathrm{s}} = \frac{\frac{1}{3}\sum_{i=1}^3 (r_i^1 - \overline{r}^1)(r_i^2 - \overline{r}^2)}{\sqrt{\frac{1}{3}\sum_{i=1}^3 (r_i^1 - \overline{r}^1)^2}\sqrt{\frac{1}{3}\sum_{i=1}^3 (r_i^2 - \overline{r}^2)^2}},$$

we immediately have that

$$\overline{r}^1 = \overline{r}^2 = \frac{1+2+3}{3} = 2,$$

and

$$\frac{1}{3}\sum_{i=1}^3 (r_i^1 - \overline{r}^1)^2 = \frac{1}{3}\sum_{i=1}^3 (r_i^2 - \overline{r}^2)^2 = \frac{1^2+0^2+1^2}{3} = \frac{2}{3},$$

while the numerator writes

$$\frac{1}{3}\sum_{i=1}^3 (r_i^1 - \overline{r}^1)(r_i^2 - \overline{r}^2) = \frac{1}{3}\left[(2-2)(3-2) + (1-2)(1-2) + (3-2)(2-2)\right] = \frac{1}{3}.$$

Therefore $r_{\mathrm{s}} = 1/2$.

We may already notice for further purpose that since, in the general case, r^1 and r^2 are permutations of $\{1, \dots, n\}$, it always holds

$$\overline{r}^1 = \overline{r}^2 = \frac{1 + \cdots + n}{n} = \frac{n+1}{2},$$

and

$$\frac{1}{n}\sum_{i=1}^n (r_i^1 - \overline{r}^1)^2 = \frac{1}{n}\sum_{i=1}^n (r_i^2 - \overline{r}^2)^2 = \frac{1}{n}\sum_{k=1}^n \left(k - \frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.$$

2. Since f is increasing, the series $x_1^1, \dots, x_n^1$ and $x_1^2, \dots, x_n^2$ are ranked in the same order, therefore $r^1 = r^2$ and it is immediate that $r_{\mathrm{s}} = 1$. On the contrary, if f is decreasing, the series $x_1^1, \dots, x_n^1$ and $x_1^2, \dots, x_n^2$ are ranked in the opposite order, therefore $r_i^1 = n + 1 - r_i^2$ for any $i \in \{1, \dots, n\}$. As a consequence, the numerator in the definition of Spearman's coefficient writes

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (r_i^1 - \overline{r}^1)(r_i^2 - \overline{r}^2) &= \frac{1}{n}\sum_{i=1}^n \left(r_i^1 - \frac{n+1}{2}\right)\left(n+1-r_i^1 - \frac{n+1}{2}\right) \\
&= \frac{1}{n}\sum_{k=1}^n \left(k - \frac{n+1}{2}\right)\left(n+1-k - \frac{n+1}{2}\right) \\
&= -\frac{1}{n}\sum_{k=1}^n \left(k - \frac{n+1}{2}\right)^2 \\
&= -\frac{n^2-1}{12},
\end{aligned}$$

where we have used the computation made at the end of the previous question. We conclude that $r_{\mathrm{s}} = -1$.

3. In general, the numerator in the definition of r_{s} writes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(r_i^1-\overline{r}^1)(r_i^2-\overline{r}^2) &= \frac{1}{n}\sum_{i=1}^{n}\left(r_i^1-\frac{n+1}{2}\right)\left(r_i^2-\frac{n+1}{2}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(r_i^1r_i^2-(r_i^1+r_i^2)\frac{n+1}{2}+\frac{(n+1)^2}{4}\right).\end{aligned}$$

Let us compute separately the sums of the three terms appearing in the right-hand side, starting from the last two. We get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(n+1)^2}{4}=\frac{(n+1)^2}{4},$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(r_i^1+r_i^2)\frac{n+1}{2}=\frac{n+1}{2}\frac{2}{n}\sum_{k=1}^{n}k=\frac{(n+1)^2}{2}.$$

Last,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}r_i^1r_i^2 &= \frac{1}{2n}\sum_{i=1}^{n}\left((r_i^1)^2+(r_i^2)^2-(r_i^1-r_i^2)^2\right)\\ &= \frac{1}{2n}\left(2\sum_{k=1}^{n}k^2-\sum_{i=1}^{n}(r_i^1-r_i^2)^2\right)\\ &= \frac{(n+1)(2n+1)}{6}-\frac{1}{2n}\sum_{i=1}^{n}(r_i^1-r_i^2)^2.\end{aligned}$$

Putting these results together, we get

$$r_{\mathrm{s}}=\frac{\frac{(n+1)(2n+1)}{6}-\frac{1}{2n}\sum_{i=1}^{n}(r_i^1-r_i^2)^2-\frac{(n+1)^2}{2}+\frac{(n+1)^2}{4}}{\frac{n^2-1}{12}},$$

which simplifies into

$$r_{\mathrm{s}}=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_i^1-r_i^2)^2.$$

Correction of Exercise 1.A.9

1. We have

$$\mathbb{E}[|X|]=\int_{x\in\mathbb{R}}|x|\frac{a}{\pi}\frac{1}{x^2+a^2}\mathrm{d}x=+\infty,$$

so $\mathbb{E}[X]$ is not defined.

2. If $c=0$ then $X\sim\delta_0$. Otherwise, for any measurable and bounded $f:\mathbb{R}\to\mathbb{R}$, we have

$$\mathbb{E}[f(cX)]=\int_{x\in\mathbb{R}}f(cx)\frac{a}{\pi}\frac{1}{x^2+a^2}\mathrm{d}x=\int_{y\in\mathbb{R}}f(y)\frac{a}{\pi}\frac{1}{(y/c)^2+a^2}\frac{\mathrm{d}y}{|c|}=\int_{y\in\mathbb{R}}f(y)\frac{a|c|}{\pi}\frac{1}{y^2+|c|^2}\mathrm{d}y,$$

so $cX\sim\mathcal{C}(a|c|)$.

3. First, notice that $\mathbb{P}(V=0)=0$ so, almost surely, U/V is well defined. Let $f:\mathbb{R}\to\mathbb{R}$ be measurable and bounded. We have

$$\mathbb{E}\left[f\left(\frac{U}{V}\right)\right]=\int_{u,v\in\mathbb{R}}f\left(\frac{u}{v}\right)\exp\left(-\left(\frac{u^2}{2\sigma^2}+\frac{v^2}{2\tau^2}\right)\right)\frac{\mathrm{d}u\mathrm{d}v}{2\pi}.$$

Let us perform the change of variable $x = u/v$ in the integral term in u. We get

$$\mathbb{E}\left[f\left(\frac{U}{V}\right)\right] = \int_{v\in\mathbb{R}}\int_{x\in\mathbb{R}} f(x)\exp\left(-\left(\frac{(xv)^2}{2\sigma^2}+\frac{v^2}{2\tau^2}\right)\right)\frac{|v|\mathrm{d}x\mathrm{d}v}{2\pi},$$

so that, setting

$$\frac{1}{\rho(x)^2} = \frac{x^2}{\sigma^2}+\frac{1}{\tau^2},$$

we get that the density of U/V is

$$\begin{aligned}\int_{v\in\mathbb{R}}\exp\left(-\frac{v^2}{2\rho(x)^2}\right)\frac{|v|\mathrm{d}v}{2\pi} &= -\int_{v=-\infty}^{0} v\exp\left(-\frac{v^2}{2\rho(x)^2}\right)\frac{\mathrm{d}v}{2\pi}+\int_{v=0}^{+\infty} v\exp\left(-\frac{v^2}{2\rho(x)^2}\right)\frac{\mathrm{d}v}{2\pi}\\ &= \frac{\rho(x)^2}{\pi} = \frac{a}{\pi}\frac{1}{x^2+a^2},\end{aligned}$$

with $a = \sigma/\tau$. We deduce that U/V has a Cauchy distribution.

4. (a) We have, for any $x\in\mathbb{R}$,

$$\begin{aligned}\Psi_U(x) &= \int_{u\in\mathbb{R}} \mathrm{e}^{\mathrm{i}ux}\frac{a}{2}\mathrm{e}^{-a|u|}\mathrm{d}u\\ &= \int_{u=-\infty}^{0}\frac{a}{2}\mathrm{e}^{\mathrm{i}ux+au}\mathrm{d}u+\int_{u=0}^{+\infty}\frac{a}{2}\mathrm{e}^{\mathrm{i}ux-au}\mathrm{d}u\\ &= \frac{a}{2}\left(\frac{1}{\mathrm{i}x+a}-\frac{1}{\mathrm{i}x-a}\right)\\ &= \frac{a^2}{x^2+a^2}.\end{aligned}$$

   (b) On the one hand, the definition of $\Psi_X(u)$ writes

$$\Psi_X(u) = \int_{x\in\mathbb{R}}\mathrm{e}^{\mathrm{i}ux}\frac{a}{\pi}\frac{1}{x^2+a^2}\mathrm{d}x.$$

   On the other hand, the Fourier inverse transform applied to the Laplace distribution yields

$$\frac{a}{2}\mathrm{e}^{-a|u|} = \frac{1}{2\pi}\int_{x\in\mathbb{R}}\mathrm{e}^{-\mathrm{i}ux}\frac{a^2}{x^2+a^2}\mathrm{d}x,$$

   which finally leads to

$$\Psi_X(u) = \mathrm{e}^{-a|u|}.$$

   (c) We conclude that the characteristic function of $X+Y$ writes

$$\Psi_{X+Y}(u) = \Psi_X(u)\Psi_Y(u) = \mathrm{e}^{-(a+b)|u|}$$

   which shows that $X+Y\sim\mathcal{C}(a+b)$.

5. (a) From the previous questions we get that $\overline{X}_n\sim\mathcal{C}(1)$.

   (b) We have

$$\overline{X}_{2n}-\overline{X}_n = \frac{1}{2n}\sum_{i=n+1}^{2n}X_i-\frac{1}{2n}\sum_{i=1}^{n}X_i = \frac{1}{2n}\sum_{i=1}^{2n}X_i',$$

   with

$$X_i' = \begin{cases}X_i & \text{if } i\in\{n+1,\dots,2n\},\\ -X_i & \text{if } i\in\{1,\dots,n\}.\end{cases}$$

   By Question 2, the variables X_i' are iid with law $\mathcal{C}(a)$. Therefore by the previous question, $\overline{X}'_{2n}\sim\mathcal{C}(a)$.

   (c) If the sequence $\overline{X}_n$ converges in distribution to some random variable Z then $\overline{X}_{2n}-\overline{X}_n$ converges in probability, and hence in distribution, to $Z-Z=0$. This is in contradiction with the fact that the law of $\overline{X}_{2n}-\overline{X}_n$ is equal to $\mathcal{C}(a)$.

Correction of Exercise 1.A.10

1. Z'_n is $\sqrt{n}$ times the empirical mean of n iid centered random variables with finite variance σ^2, so by the Central Limit Theorem we have $Z'_n \to \mathcal{N}(0, \sigma^2)$ in distribution.

2. We have $Z_n + Z'_n = \sqrt{2}Z_{2n}$ so if $Z_n \to Z$ in probability we have $Z'_n = \sqrt{2}Z_{2n} - Z_n \to (\sqrt{2}-1)Z$ in probability.

3. Under the assumption made above that $Z_n \to Z$ in probability, we deduce from the previous questions that necessarily, $(\sqrt{2}-1)Z \sim \mathcal{N}(0, \sigma^2)$. But on the other hand, since $Z_n \to Z$ in distribution, we also have $Z \sim \mathcal{N}(0, \sigma^2)$. We deduce that for the assumption to hold true, it is necessary that $\sigma^2 = 0$. And it is straightforward to check that, conversely, if $\sigma^2 = 0$ then indeed $Z_n = 0$ converges in probability.

   As a conclusion, we have established that Z_n converges in probability if and only if the variables X_i are deterministic.

D.2 Pointwise Estimation in Parametric Models

Correction of Exercise 2.A.1

1. Since the random variables $X_1, \dots, X_n$ are iid according to the Bernoulli distribution with parameter p, we get

$$\mathbb{P}_p(\mathbf{X}_n = \mathbf{x}_n) = \prod_{i=1}^{n} \mathbb{P}_p(X_i = x_i) = p^k(1-p)^{n-k},$$

   where $k = x_1 + \cdots + x_n$.

2. The assumption that Z_n is unbiased writes $\mathbb{E}_p[Z_n] = g(p) = 1/p$ for all $p \in (0,1)$. On the other hand, since Z_n is a statistic, there exists a function $z_n : \{0,1\}^n \to (0, +\infty)$ such that $Z_n = z_n(\mathbf{X}_n)$. Thus, the expectation of Z_n writes

$$\mathbb{E}_p[Z_n] = \sum_{\mathbf{x}_n \in \{0,1\}^n} z_n(\mathbf{x}_n) \mathbb{P}_p(\mathbf{X}_n = \mathbf{x}_n).$$

   Regrouping the terms of this sum according to the value of $x_1 + \cdots + x_n$ and using Question 1, we get

$$\mathbb{E}_p[Z_n] = \sum_{k=0}^{n} \underbrace{\left(\sum_{\substack{\mathbf{x}_n \in \{0,1\}^n \\ x_1 + \cdots + x_n = k}} z_n(\mathbf{x}_n) \right)}_{a_k} p^k (1-p)^{n-k},$$

   which leads to the expected result.

3. Multiplying both sides of the result of Question 2 by p yields the identity

$$\forall p \in (0,1), \qquad \sum_{k=0}^{n} a_k p^{k+1} (1-p)^{n-k} - 1 = 0.$$

   The left-hand side is a polynomial function of p with order at most $n+1$, and its zero-th order coefficient is -1. Therefore this identity cannot hold for more than $n+1$ values of p, which provides a contradiction.

Correction of Exercise 2.A.2

1. The covariance matrix of Y_1 writes

$$K = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_1^2) \\ \mathrm{Cov}(X_1, X_1^2) & \mathrm{Var}(X_1^2) \end{pmatrix},$$

and we have

$$\begin{aligned} \mathrm{Var}(X_1) &= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \rho_2, \\ \mathrm{Cov}(X_1, X_1^2) &= \mathbb{E}[X_1^3] - \mathbb{E}[X_1]\mathbb{E}[X_1^2] = \rho_3, \\ \mathrm{Var}(X_1^2) &= \mathbb{E}[X_1^4] - \mathbb{E}[X_1^2]^2 = \rho_4 - \rho_2^2. \end{aligned}$$

2. For all $x_1, x_2 \in \mathbb{R}$, we have

$$\nabla\varphi(x_1, x_2) = \begin{pmatrix} \frac{\partial\varphi}{\partial x_1}(x_1, x_2) \\ \frac{\partial\varphi}{\partial x_2}(x_1, x_2) \end{pmatrix} = \begin{pmatrix} -2x_1 \\ 1 \end{pmatrix}.$$

As a consequence,

$$\nabla\varphi(y) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

3. We have $V_n = \varphi(\overline{Y}_n)$ and $\rho_2 = \varphi(y)$, so that by the Delta Method, $\sqrt{n}(V_n - \rho_2)$ converges in distribution to $\mathcal{N}(0, v)$ with $v = \nabla\varphi(y)^\top K \nabla\varphi(y) = \rho_4 - \rho_2^2$.

Correction of Exercise 2.A.4

1. By the strong law of large numbers, $\overline{X}_n \to \mu$ and $\overline{X}'_n \to \mu$, almost surely. By continuity, this proves that A_n and B_n are strongly consistent estimators of μ^2. Using the strong law of large numbers again, as well as the independence between the two samples, we have

$$\lim_{n\to+\infty} C_n = \mathbb{E}[X_1 X'_1] = \mathbb{E}[X_1]\mathbb{E}[X'_1] = \mu^2.$$

2. We first compute

$$\mathbb{E}\left[A_n\right] = \frac{1}{4}\mathbb{E}\left[\overline{X}_n^2 + 2\overline{X}_n\overline{X}'_n + \overline{X}_n'^2\right] = \frac{1}{2}\left(\mathbb{E}\left[\overline{X}_n^2\right] + \mathbb{E}\left[\overline{X}_n\right]^2\right),$$

where we have used the fact that $\overline{X}_n$ and $\overline{X}'_n$ are independent and have the same law. On the one hand, $\mathbb{E}[\overline{X}_n]^2 = \mu^2$. On the other hand, we recall that the empirical variance $V_n = \overline{X^2}_n - \overline{X}_n^2$ satisfies the identity

$$\mathbb{E}[V_n] = \frac{n-1}{n}\sigma^2,$$

which yields

$$\mathbb{E}\left[\overline{X}_n^2\right] = \mathbb{E}\left[\overline{X^2}_n\right] - \frac{n-1}{n}\sigma^2 = \mathbb{E}[X_1^2] - \frac{n-1}{n}\sigma^2 = \mu^2 + \frac{\sigma^2}{n}.$$

As a conclusion,

$$b(A_n; \mu^2) = \mathbb{E}\left[A_n\right] - \mu^2 = \frac{1}{2}\left(\mu^2 + \frac{\sigma^2}{n} + \mu^2\right) - \mu^2 = \frac{\sigma^2}{2n}.$$

On the other hand, it easily follows from the independence of the two samples that B_n and C_n are unbiased.

3. By the Central Limit Theorem and the independence of the samples (or equivalently the multidimensional Central Limit Theorem),

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n\\ \overline{X}'_n\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\to\begin{pmatrix}Y\\ Y'\end{pmatrix},\qquad\text{in distribution,}$$

where Y and Y' are independent $\mathcal{N}(0,\sigma^2)$ variables (or equivalently $\begin{pmatrix}Y\\ Y'\end{pmatrix}\sim\mathcal{N}_2(0,\sigma^2 I_2)$).

On the one hand, we have

$$A_n=\phi\begin{pmatrix}\overline{X}_n\\ \overline{X}'_n\end{pmatrix},\qquad\phi\begin{pmatrix}x\\ x'\end{pmatrix}=\left(\frac{x+x'}{2}\right)^2.$$

Therefore, by the Delta method,

$$\sqrt{n}\left(A_n-\mu^2\right)=\sqrt{n}\left(\phi\begin{pmatrix}\overline{X}_n\\ \overline{X}'_n\end{pmatrix}-\phi\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\to\phi'\begin{pmatrix}\mu\\ \mu\end{pmatrix}\cdot\begin{pmatrix}Y\\ Y'\end{pmatrix},\qquad\text{in distribution,}$$

with

$$\phi'\begin{pmatrix}x\\ x'\end{pmatrix}=\begin{pmatrix}\frac{\partial\phi}{\partial x}\begin{pmatrix}x\\ x'\end{pmatrix} & \frac{\partial\phi}{\partial x'}\begin{pmatrix}x\\ x'\end{pmatrix}\end{pmatrix}=\begin{pmatrix}\frac{x+x'}{2} & \frac{x+x'}{2}\end{pmatrix},\qquad\phi'\begin{pmatrix}\mu\\ \mu\end{pmatrix}=\begin{pmatrix}\mu & \mu\end{pmatrix}.$$

We deduce that A_n is asymptotically normal, with asymptotic variance

$$\mathrm{Var}\left(\begin{pmatrix}\mu & \mu\end{pmatrix}\cdot\begin{pmatrix}Y\\ Y'\end{pmatrix}\right)=\mathrm{Var}(\mu Y+\mu Y')=2\mu^2\sigma^2.$$

On the other hand, we have

$$B_n=\psi\begin{pmatrix}\overline{X}_n\\ \overline{X}'_n\end{pmatrix},\qquad\psi\begin{pmatrix}x\\ x'\end{pmatrix}=xx'.$$

Therefore, by the Delta method,

$$\sqrt{n}\left(B_n-\mu^2\right)=\sqrt{n}\left(\phi\begin{pmatrix}\overline{X}_n\\ \overline{X}'_n\end{pmatrix}-\psi\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\to\psi'\begin{pmatrix}\mu\\ \mu\end{pmatrix}\cdot\begin{pmatrix}Y\\ Y'\end{pmatrix},\qquad\text{in distribution,}$$

with

$$\psi'\begin{pmatrix}x\\ x'\end{pmatrix}=\begin{pmatrix}\frac{\partial\psi}{\partial x}\begin{pmatrix}x\\ x'\end{pmatrix} & \frac{\partial\psi}{\partial x'}\begin{pmatrix}x\\ x'\end{pmatrix}\end{pmatrix}=\begin{pmatrix}x' & x\end{pmatrix},\qquad\psi'\begin{pmatrix}\mu\\ \mu\end{pmatrix}=\begin{pmatrix}\mu & \mu\end{pmatrix}.$$

We deduce that B_n is asymptotically normal, with asymptotic variance

$$\mathrm{Var}\left(\begin{pmatrix}\mu & \mu\end{pmatrix}\cdot\begin{pmatrix}Y\\ Y'\end{pmatrix}\right)=\mathrm{Var}(\mu Y+\mu Y')=2\mu^2\sigma^2.$$

Last, the Central Limit Theorem shows that

$$\sqrt{n}\left(C_n-\mu^2\right)=\sqrt{n}\left(\overline{(XX')}_n-\mathbb{E}[X_1X'_1]\right)\to\mathcal{N}(0,\mathrm{Var}(X_1X'_1)),\qquad\text{in distribution,}$$

so that C_n is asymptotically normal with asymptotic variance

$$\begin{aligned}\mathrm{Var}(X_1X'_1)&=\mathbb{E}[X_1^2X_1'^2]-\mathbb{E}[X_1X'_1]^2\\&=\mathbb{E}[X_1^2]^2-\mu^4\\&=(\mu^2+\sigma^2)^2-\mu^4\\&=\sigma^2(2\mu^2+\sigma^2).\end{aligned}$$

4. C_n has a larger asymptotic variance than A_n and B_n, and among the two latter estimators, A_n is biased while B_n is not. Therefore B_n has the lowest approximated MSE.

Correction of Exercise 2.A.5

1. P_0 is the Cauchy distribution with parameter 1.

2. By definition,

$$\begin{aligned}\mathbb{E}_\theta[U_1] &= \mathbb{P}_\theta(X_1 \leq 0) \\ &= \int_{x=-\infty}^{0} \frac{\mathrm{d}x}{\pi((x-\theta)^2+1)} \\ &= \int_{y=-\infty}^{-\theta} \frac{\mathrm{d}y}{\pi(y^2+1)} \qquad \text{letting } y = x-\theta \\ &= \frac{1}{\pi}\left(\arctan(-\theta)+\frac{\pi}{2}\right) = \frac{1}{\pi}\left(-\arctan(\theta)+\frac{\pi}{2}\right).\end{aligned}$$

3. The identity obtained in Question 2 rewrites

$$\arctan(\theta) = \pi\left(\frac{1}{2} - \mathbb{E}_\theta[U_1]\right).$$

Since $\mathbb{E}_\theta[U_1] \in (0,1)$, the right-hand side belongs to the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$, therefore

$$\theta = g(\mathbb{E}_\theta[U_1]),$$

where g is defined on $(0,1)$ by

$$g(u) = \tan\left(\pi\left(\frac{1}{2}-u\right)\right).$$

Approximating $\mathbb{E}_\theta[U_1]$ by $\frac{1}{n}\sum_{i=1}^n U_i$, we get the estimator

$$\widetilde{\theta}_n = g\left(\frac{1}{n}\sum_{i=1}^n U_i\right).$$

Since the function g is continuous on $(0,1)$ on the one hand, and by the strong Law of Large Numbers, $\frac{1}{n}\sum_{i=1}^n U_i$ converges almost surely to $\mathbb{E}_\theta[U_1]$ on the other hand, we deduce that this estimator is strongly consistent.

4. By the Central Limit Theorem,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n U_i - \mathbb{E}_\theta[U_1]\right) \to \mathcal{N}(0, \mathrm{Var}_\theta[U_1]), \qquad \text{in distribution,}$$

and since U_1 is a Bernoulli random variable,

$$\begin{aligned}\mathrm{Var}_\theta[U_1] &= \mathbb{E}_\theta[U_1](1-\mathbb{E}_\theta[U_1]) \\ &= \frac{1}{\pi^2}\left(-\arctan(\theta)+\frac{\pi}{2}\right)\left(\arctan(\theta)+\frac{\pi}{2}\right) \\ &= \frac{1}{\pi^2}\left(\frac{\pi^2}{4} - \arctan(\theta)^2\right).\end{aligned}$$

Besides, g is C^1 on $(0,1)$, and for all $u \in (0,1)$,

$$g'(u) = -\pi\tan'\left(\pi\left(\frac{1}{2}-u\right)\right) = -\pi\left[1+\tan\left(\pi\left(\frac{1}{2}-u\right)\right)^2\right],$$

so that

$$\begin{aligned} g'(\mathbb{E}_\theta[U_1]) &= -\pi\left[1+\tan\left(\pi\left(\frac{1}{2}-\frac{1}{\pi}\left(-\arctan(\theta)+\frac{\pi}{2}\right)\right)\right)^2\right] \\ &= -\pi\left[1+\tan\left(\arctan(\theta)\right)^2\right] \\ &= -\pi(1+\theta^2). \end{aligned}$$

Applying the Delta Method, we get that

$$\sqrt{n}\left(\widetilde{\theta}_n-\theta\right)=\sqrt{n}\left(g\left(\frac{1}{n}\sum_{i=1}^{n}U_i\right)-g\left(\mathbb{E}_\theta[U_1]\right)\right)$$

converges in distribution to a centered Gaussian variable, with variance

$$v(\theta)=g'(\mathbb{E}_\theta[U_1])^2\,\mathrm{Var}_\theta[U_1]=(1+\theta^2)^2\left(\frac{\pi^2}{4}-\arctan(\theta)^2\right).$$

5. The function v introduced in Question 4 is continuous, therefore $v(\widetilde{\theta}_n)$ converges to $v(\theta)$, $\mathbb{P}_\theta$-almost surely. By Slutsky's Theorem, we deduce that $\sqrt{n/v(\widetilde{\theta}_n)}(\widetilde{\theta}_n-\theta)$ converges in distribution to $\mathcal{N}(0,1)$. In particular, if $\phi_{1-\alpha/2}$ denotes the quantile of order $1-\alpha/2$ of the standard Gaussian distribution, we have for all $\theta\in\mathbb{R}$,

$$\lim_{n\to+\infty}\mathbb{P}_\theta\left(\sqrt{\frac{n}{v(\widetilde{\theta}_n)}}|\widetilde{\theta}_n-\theta|\geq\phi_{1-\alpha/2}\right)=\alpha,$$

whence

$$\left[\widetilde{\theta}_n-\phi_{1-\alpha/2}\sqrt{\frac{v(\widetilde{\theta}_n)}{n}},\widetilde{\theta}_n+\phi_{1-\alpha/2}\sqrt{\frac{v(\widetilde{\theta}_n)}{n}}\right]$$

is an asymptotic confidence interval with level $1-\alpha$ for θ.

D.3 Statistics in Gaussian Models

Correction of Exercise 3.A.1

1. For any $x>0$, $\mathbb{P}(G^2\leq x)=\mathbb{P}(-\sqrt{x}\leq G\leq\sqrt{x})=\Phi(\sqrt{x})-\Phi(-\sqrt{x})=2\Phi(\sqrt{x})-1$. We deduce that G^2 has density

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(G^2\leq x)=\frac{\mathrm{d}}{\mathrm{d}x}\left(2\Phi(\sqrt{x})-1\right)=\frac{1}{\sqrt{x}}\Phi'(\sqrt{x})=\frac{1}{\sqrt{2\pi}}x^{-1/2}\mathrm{e}^{-x/2}.$$

Since this density is proportional to the $\Gamma(1/2,1/2)$ distribution, we deduce that the normalisation constants are the same (and therefore $\Gamma(1/2)=\sqrt{\pi}$), and $G^2\sim\Gamma(1/2,1/2)$.

2. We conclude from Exercise 1.3.12 and the definition of the $\chi_2(n)$ distribution that $\chi_2(n)=\Gamma(n/2,1/2)$.

Correction of Exercise 3.A.4

1. By independence,
$$\mathbb{E}\left[S_n^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] = \mathbb{E}\left[S_n^2\right]\mathbb{E}\left[\mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] = \sigma^2\psi(u)^n,$$
since S_n^2 is an unbiased estimator of σ^2.

2. Using the formula given for S_n^2, we have
$$\mathbb{E}\left[S_n^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[X_j^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] - \frac{1}{n(n-1)}\sum_{j\neq k}\mathbb{E}\left[X_j X_k \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right].$$
Since $X_1, \dots, X_n$ are iid, we have for each $j \in \{1, \dots, n\}$,
$$\begin{aligned}\mathbb{E}\left[X_j^2 \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] &= \mathbb{E}\left[X_j^2 \mathrm{e}^{\mathrm{i}u(X_1+\cdots+X_j+\cdots+X_n)}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_1}\right]\cdots\mathbb{E}\left[X_j^2\mathrm{e}^{\mathrm{i}uX_j}\right]\cdots\mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_n}\right] \\ &= \mathbb{E}\left[X_1^2\mathrm{e}^{\mathrm{i}uX_1}\right]\psi(u)^{n-1},\end{aligned}$$
and similarly, for $j \neq k$,
$$\begin{aligned}\mathbb{E}\left[X_j X_k \mathrm{e}^{\mathrm{i}un\overline{X}_n}\right] &= \mathbb{E}\left[X_j X_k \mathrm{e}^{\mathrm{i}u(X_1+\cdots+X_j+\cdots+X_k+\cdots+X_n)}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_1}\right]\cdots\mathbb{E}\left[X_j\mathrm{e}^{\mathrm{i}uX_j}\right]\cdots\mathbb{E}\left[X_k\mathrm{e}^{\mathrm{i}uX_k}\right]\cdots\mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_n}\right] \\ &= \left(\mathbb{E}\left[X_1\mathrm{e}^{\mathrm{i}uX_1}\right]\right)^2\psi(u)^{n-2}.\end{aligned}$$

3. Using the Leibniz differentiation theorem and the fact that X_1 has a finite second-order moment, we have that the function ψ is C^2 and, for any $u \in \mathbb{R}$,
$$\psi'(u) = \mathbb{E}\left[\mathrm{i}X_1\mathrm{e}^{\mathrm{i}uX_1}\right], \qquad \psi''(u) = \mathbb{E}\left[(\mathrm{i}X_1)^2\mathrm{e}^{\mathrm{i}uX_1}\right].$$
Therefore, $\mathbb{E}[X_1\mathrm{e}^{\mathrm{i}uX_1}] = -\mathrm{i}\psi'(u)$ and $\mathbb{E}[X_1^2\mathrm{e}^{\mathrm{i}uX_1}] = -\psi''(u)$.

4. Putting together the results of the previous questions, we deduce the identity
$$\sigma^2\psi(u)^n = -\psi''(u)\psi(u)^{n-1} + \psi'(u)^2\psi(u)^{n-2}.$$
Since $\psi(u)$ is assumed to be nonzero, we deduce that for any $u \in \mathbb{R}$,
$$\sigma^2\psi(u)^2 = -\psi''(u)\psi(u) + \psi'(u)^2.$$
As a consequence, the function f defined by $f(u) = \psi'(u)/\psi(u)$ satisfies
$$f'(u) = \frac{\psi''(u)\psi(u) - \psi'(u)^2}{\psi(u)^2} = -\sigma^2.$$
Thus, there exists $C \in \mathbb{C}$ such that, for any $u \in \mathbb{R}$,
$$f(u) = -\sigma^2 u + C.$$
Since $f(0) = \psi'(0)/\psi(0) = \mathrm{i}\mu$, we get $C = \mathrm{i}\mu$.

5. As a consequence of the previous result, ψ satisfies the differential equation
$$\begin{cases}\psi'(u) = (-\sigma^2 u + \mathrm{i}\mu)\psi(u), \\ \psi(0) = 1,\end{cases}$$
whose solution is given by $\psi(u) = \mathrm{e}^{-\sigma^2\frac{u^2}{2} + \mathrm{i}\mu u}$. By Exercise 1.3.19, we conclude that $X_1 \sim \mathcal{N}(\mu, \sigma^2)$.

To complete the exercise without assuming that ψ does not vanish, one may proceed as follows. Since ψ is continuous on $\mathbb{R}$ and $\psi(0) = 1$, one has

$$u_- := \sup\{u \in (-\infty, 0) : \psi(u) = 0\} \in [-\infty, 0), \quad u_+ := \inf\{u \in (0, +\infty) : \psi(u) = 0\} \in (0, +\infty].$$

Our goal here is to show that, necessarily, $u_- = -\infty$ and $u_+ = +\infty$. The arguments used in Question 4 allow to show that f is well-defined and satisfies $f'(u) = -\sigma^2$ on the nonempty interval (u_-, u_+), so that the identities $\psi'(u) = (-\sigma^2 u + \mathrm{i}\mu)\psi(u)$ and then $\psi(u) = \mathrm{e}^{-\sigma^2 \frac{u^2}{2} + \mathrm{i}\mu u}$ hold true on this interval. Assuming that $u_- > -\infty$ and using the continuity of both ψ and $u \mapsto \mathrm{e}^{-\sigma^2 \frac{u^2}{2} + \mathrm{i}\mu u}$ on $\mathbb{R}$ would then lead to the contradictory statement that $0 = \mathrm{e}^{-\sigma^2 \frac{u_-^2}{2} + \mathrm{i}\mu u_-}$, therefore u_- must be $-\infty$ and for the same reason, $u_+ = +\infty$.

D.4 Confidence Intervals

Correction of Exercise 4.A.1 Since $\widehat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})$, we have that

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \rho_j}} \sim \mathcal{N}(0, 1),$$

where $\rho_j > 0$ is the i-th diagonal coefficient of the matrix $(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}$. On the other hand, $\widehat{\sigma}^2$ is independent of $\widehat{\beta}$ and satisfies $(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1)$, which implies that

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 \rho_j}} \sim \mathrm{t}(n-p-1).$$

We deduce that the interval

$$\left[\widehat{\beta}_j - t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2 \rho_j}, \widehat{\beta}_j + t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2 \rho_j}\right]$$

is a confidence interval for β_j with equal risks of under- and overestimation.

Correction of Exercise 4.A.2 Let us fix α' and introduce the following events:

$$\begin{aligned}
A^1 &= \{I_n^{-,1}(\alpha') \le g_1(\theta) \le I_n^{+,1}(\alpha')\}, \\
A^2 &= \{I_n^{-,2}(\alpha') \le g_2(\theta) \le I_n^{+,2}(\alpha')\}, \\
B &= \{I_n^{-,1}(\alpha') + I_n^{-,2}(\alpha') \le g_1(\theta) + g_2(\theta) \le I_n^{+,1}(\alpha') + I_n^{+,2}(\alpha')\},
\end{aligned}$$

and note that

$$A^1 \cap A^2 \subset B,$$

which implies that

$$B^c \subset \left(A^1 \cap A^2\right)^c = \left(A^1\right)^c \cup \left(A^2\right)^c.$$

We deduce that

$$\begin{aligned}
\mathbb{P}_\theta(B) &= 1 - \mathbb{P}_\theta(B^c) \\
&\ge 1 - \mathbb{P}_\theta\left(\left(A^1\right)^c \cup \left(A^2\right)^c\right) \\
&\ge 1 - \left(\mathbb{P}_\theta\left(\left(A^1\right)^c\right) + \mathbb{P}_\theta\left(\left(A^2\right)^c\right)\right) \\
&= 1 - \left(1 - \mathbb{P}_\theta(A^1) + 1 - \mathbb{P}_\theta(A^2)\right) \\
&= \mathbb{P}_\theta(A^1) + \mathbb{P}_\theta(A^2) - 1 \\
&= 1 - 2\alpha'.
\end{aligned}$$

As a consequence, taking $\alpha' = \alpha/2$ yields $\mathbb{P}_\theta(B) \ge 1 - \alpha$. Therefore we deduce that $[I_n^{-,1}(\alpha/2) + I_n^{-,2}(\alpha/2), I_n^{+,1}(\alpha/2) + I_n^{+,2}(\alpha/2)]$ is an approximate confidence interval with level $1 - \alpha$ for $g_1(\theta) + g_2(\theta)$.

Correction of Exercise 4.A.5

1. The confidence interval I_n provided by Proposition 4.3.1 has width $2\phi_{1-\alpha/2}\sqrt{\widehat{V}_n/n}$.

2. By the Delta Method, $\Phi(Z_n)$ is a consistent and asymptotically normal estimator of $\Phi(g(\theta))$, with asymptotic variance $\Phi'(g(\theta))^2 V(\theta)$.

3. Since Φ is C^1, $\Phi'(Z_n)$ converges to $\Phi'(g(\theta))$ in probability, therefore $\Phi'(Z_n)^2\widehat{V}_n$ is a consistent estimator of $\Phi'(g(\theta))^2 V(\theta)$. As a consequence, by Proposition 4.3.1 an asymptotic confidence interval for $\Phi(g(\theta))$ is

$$J_n = \left[\Phi(Z_n) - \phi_{1-\alpha/2}\sqrt{\frac{\Phi'(Z_n)^2\widehat{V}_n}{n}}, \Phi(Z_n) + \phi_{1-\alpha/2}\sqrt{\frac{\Phi'(Z_n)^2\widehat{V}_n}{n}}\right].$$

   Using the monotonicity of Φ, we have

$$\Phi(Z_n) - \phi_{1-\alpha/2}\sqrt{\frac{\Phi'(Z_n)^2\widehat{V}_n}{n}} \leq \Phi(g(\theta)) \leq \Phi(Z_n) + \phi_{1-\alpha/2}\sqrt{\frac{\Phi'(Z_n)^2\widehat{V}_n}{n}}$$

   if and only if

$$\Phi^{-1}\left(\Phi(Z_n) - \Phi'(Z_n)\phi_{1-\alpha/2}\sqrt{\frac{\widehat{V}_n}{n}}\right) \leq g(\theta) \leq \Phi^{-1}\left(\Phi(Z_n) + \Phi'(Z_n)\phi_{1-\alpha/2}\sqrt{\frac{\widehat{V}_n}{n}}\right),$$

   so that

$$I_n^\Phi = \left[\Phi^{-1}\left(\Phi(Z_n) - \Phi'(Z_n)\phi_{1-\alpha/2}\sqrt{\frac{\widehat{V}_n}{n}}\right), \Phi^{-1}\left(\Phi(Z_n) + \Phi'(Z_n)\phi_{1-\alpha/2}\sqrt{\frac{\widehat{V}_n}{n}}\right)\right]$$

   is a second approximate confidence interval for $g(\theta)$.

4. Let us define

$$s_n = \phi_{1-\alpha/2}\sqrt{\frac{\widehat{V}_n}{n}},$$

   so that the width of the interval I_n^Φ writes $\varphi_n(s_n)$, while the width of I_n is $2s_n$.

   It is obvious that $\varphi_n(0) = 0$, and

$$\varphi_n'(s) = \Phi'(Z_n)(\Phi^{-1})'\left(\Phi(Z_n) + \Phi'(Z_n)s\right) + \Phi'(Z_n)(\Phi^{-1})'\left(\Phi(Z_n) - \Phi'(Z_n)s\right).$$

   As a consequence,

$$\varphi_n'(0) = 2\Phi'(Z_n)(\Phi^{-1})'\left(\Phi(Z_n)\right) = 2.$$

   We deduce that the width of I_n rewrites $\varphi_n(0) + \varphi_n'(0)s$, from which we conclude that:

   - if φ_n is convex, then I_n is smaller than I_n^Φ;
   - if φ_n is concave, then I_n^Φ is smaller than I_n.

5. If $\Phi'(g(\theta)) = 1/\sqrt{V(\theta)}$, then the asymptotic variance of $\Phi(Z_n)$ is 1 and no longer depends on the parameter θ. As a consequence, there is no need to estimate the variance and we get the asymptotic confidence interval

$$J_n = \left[\Phi(Z_n) - \frac{\phi_{1-\alpha/2}}{\sqrt{n}}, \Phi(Z_n) + \frac{\phi_{1-\alpha/2}}{\sqrt{n}}\right]$$

   for $\Phi(g(\theta))$, from which we deduce the asymptotic confidence interval

$$I_n^\Phi = \left[\Phi^{-1}\left(\Phi(Z_n) - \frac{\phi_{1-\alpha/2}}{\sqrt{n}}\right), \Phi^{-1}\left(\Phi(Z_n) + \frac{\phi_{1-\alpha/2}}{\sqrt{n}}\right)\right]$$

   for $g(\theta)$.

6. As an example of a function Φ such that

$$\Phi'(\lambda) = \frac{1}{\sqrt{V(\lambda)}} = \frac{1}{\lambda},$$

one can take $\Phi(\lambda) = \log \lambda$. With this choice, the function φ_n introduced in Question 4 writes

$$\varphi_n(s) = \exp\left(\log \widehat{\lambda}_n + \frac{s}{\widehat{\lambda}_n}\right) - \exp\left(\log \widehat{\lambda}_n - \frac{s}{\widehat{\lambda}_n}\right) = 2\widehat{\lambda}_n \sinh\left(\frac{s}{\widehat{\lambda}_n}\right),$$

which is convex on $[0, +\infty)$. As a consequence, the confidence interval obtained with variance stabilisation is larger than the original one.

Correction of Exercise 4.A.6

1. Let us denote by Φ_n the CDF of $\mathrm{t}(n)$, so that $t_{n,r} = \Phi_n^{-1}(r)$ and $\phi_r = \Phi^{-1}(r)$. Since both Φ_n and Φ are increasing, with $\Phi_n(0) = \Phi(0) = 1/2$, it is clear that $t_{n,r} \geq \phi_r$ on $(1/2, 1)$ if and only if $\Phi_n(x) \leq \Phi(x)$ for any $x > 0$.

2. Setting $T_n = G/\sqrt{Y_n/n}$ with $G \sim \mathcal{N}(0,1)$, $Y_n \sim \chi_2(n)$ independent, we get, for any $x > 0$,

$$\begin{aligned}
\Phi_n(x) &= \mathbb{P}\left(\frac{G}{\sqrt{Y_n/n}} \leq x\right) \\
&= \int_{z\in\mathbb{R}} \int_{y>0} \mathbb{1}_{\{z/\sqrt{y/n} \leq x\}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} f_n(y) \mathrm{d}y \mathrm{d}z \\
&= \int_{y>0} \left(\int_{z\in\mathbb{R}} \mathbb{1}_{\{z \leq x\sqrt{y/n}\}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z \right) f_n(y) \mathrm{d}y \\
&= \int_{y>0} \Phi\left(x\sqrt{y/n}\right) f_n(y) \mathrm{d}y \\
&= \mathbb{E}\left[\Phi\left(x\sqrt{\frac{Y_n}{n}}\right)\right],
\end{aligned}$$

where f_n is the density of the $\chi_2(n)$ distribution.

3. Since Φ' is decreasing on $[0, +\infty)$, Φ is concave, and thus for any $x > 0$, the function $y \mapsto \Phi(x\sqrt{y/n})$ is concave on $[0, +\infty)$. By Jensen's inequality, we thus deduce that

$$\Phi_n(x) = \mathbb{E}\left[\Phi\left(x\sqrt{\frac{Y_n}{n}}\right)\right] \leq \Phi\left(x\sqrt{\frac{\mathbb{E}\left[Y_n\right]}{n}}\right) = \Phi(x).$$

D.5 Maximum Likelihood Estimation

Correction of Exercise 5.A.1

1. Let $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathbb{N}^n$. The likelihood of the realisation $\mathbf{x}_n$ writes

$$L_n(\mathbf{x}_n; \lambda) = \prod_{i=1}^{n} \exp(-\lambda) \frac{\lambda^{x_i}}{x_i!}.$$

2. The log-likelihood of the realisation $\mathbf{x}_n$ writes

$$\ell_n(\mathbf{x}_n; \lambda) = -n\lambda + s_n(\mathbf{x}_n) \log \lambda + \log\left(\frac{1}{\prod_{i=1}^{n} x_i!}\right),$$

where $s_n(\mathbf{x}_n) = \sum_{i=1}^n x_i$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell_n(\mathbf{x}_n;\lambda) = -n + \frac{s_n(\mathbf{x}_n)}{\lambda},$$

which vanishes when λ takes the value

$$\lambda_n(\mathbf{x}_n) = \frac{s_n(\mathbf{x}_n)}{n}.$$

The study of the sign of the derivative of $\ell_n(\mathbf{x}_n;\lambda)$ shows that this function reaches its maximum at $\lambda_n(\mathbf{x}_n)$, therefore the MLE of λ is

$$\widehat{\lambda}_n = \lambda_n(\mathbf{X}_n) = \overline{X}_n.$$

Notice that it coincides with the moment estimator $\widetilde{\lambda}_n^{(1)}$ constructed in Exercise 2.A.3. As a consequence, it is unbiased, strongly consistent, and asymptotically normal with asymptotic variance λ.

3. The score of the model writes
$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell_1(X_1;\lambda) = -1 + \frac{X_1}{\lambda},$$
whose variance is
$$I(\lambda) = \frac{1}{\lambda}.$$

4. The estimator $\widehat{\lambda}_n$ is unbiased, and its variance writes
$$\mathrm{Var}_\lambda[\widehat{\lambda}_n] = \frac{1}{n}\mathrm{Var}_\lambda[X_1] = \frac{\lambda}{n} = \frac{1}{nI(\lambda)},$$
which is the Cramér–Rao bound. As a consequence, the MLE is efficient.

5. By Proposition 4.3.1, $[\widehat{\lambda}_n - \phi_{1-\alpha/2}\sqrt{\frac{\widehat{\lambda}_n}{n}}, \widehat{\lambda}_n + \phi_{1-\alpha/2}\sqrt{\frac{\widehat{\lambda}_n}{n}}]$ is an asymptotic confidence interval for λ with level $1-\alpha$.

Correction of Exercise 5.A.2

1. The likelihood of a realisation $\mathbf{x}_n = (x_1,\dots,x_n) \in (\mathbb{N}^*)^n$ writes
$$L_n(\mathbf{x}_n;p) = \prod_{i=1}^n p(1-p)^{x_i-1} = p^n(1-p)^{n(\overline{x}_n-1)}, \qquad \overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i.$$

2. For all $\mathbf{x}_n = (x_1,\dots,x_n) \in (\mathbb{N}^*)^n$, let us denote by
$$\ell_n(\mathbf{x}_n;p) = \log L_n(\mathbf{x}_n;p) = n\log p + n(\overline{x}_n - 1)\log(1-p)$$
the log-likelihood. The function $p \mapsto \ell_n(\mathbf{x}_n;p)$ is differentiable on $(0,1]$ and
$$\frac{\mathrm{d}}{\mathrm{d}p}\ell_n(\mathbf{x}_n;p) = \frac{n}{p} - \frac{n(\overline{x}_n-1)}{1-p},$$
and the right-hand side is null if and only if p takes the value
$$p_n(\mathbf{x}_n) = \frac{1}{\overline{x}_n}.$$

Notice that

$$\frac{\mathrm{d}}{\mathrm{d}p}\ell_n(\mathbf{x}_n;p)\begin{cases}<0 & \text{if } p>p_n(\mathbf{x}_n),\\ >0 & \text{if } p<p_n(\mathbf{x}_n),\end{cases}$$

so that $p\mapsto \ell_n(\mathbf{x}_n;p)$ reaches its global maximum in $p_n(\mathbf{x}_n)$. As a consequence, the MLE of p is

$$\widehat{p}_n = p_n(\mathbf{X}_n) = \frac{1}{\overline{X}_n}.$$

By the strong Law of Large Numbers, for all $p\in(0,1]$,

$$\lim_{n\to+\infty}\overline{X}_n = \mathbb{E}_p[X_1] = \frac{1}{p}, \qquad \mathbb{P}_p\text{-almost surely,}$$

so that the continuity of the mapping $x\mapsto 1/x$ at $1/p$ ensures that

$$\lim_{n\to+\infty}\widehat{p}_n = \frac{1}{1/p} = p, \qquad \mathbb{P}_p\text{-almost surely,}$$

and the estimator $\widehat{p}_n$ is strongly consistent.

3. We shall combine the Central Limit Theorem with the Delta Method. First, we write

$$\sqrt{n}\,(\widehat{p}_n - p) = \sqrt{n}\left(g(\overline{X}_n) - g(\mathbb{E}_p[X_1])\right),$$

with $g(x)=1/x$. On the one hand, since $\mathrm{Var}_p[X_1]=(1-p)/p^2$, the Central Limit Theorem yields

$$\sqrt{n}\left(\overline{X}_n - 1/p\right)\to Z\sim\mathcal{N}(0,(1-p)/p^2), \qquad \text{in distribution.}$$

On the other hand,

$$g'(\mathbb{E}_p[X_1]) = -\frac{1}{(1/p)^2} = -p^2.$$

As a consequence, the Delta Method asserts that

$$\sqrt{n}\left(g(\overline{X}_n) - g(\mathbb{E}_p[X_1])\right)\to -p^2 Z, \qquad \text{in distribution,}$$

which shows that $\widehat{p}_n$ is asymptotically normal, with asymptotic variance

$$V(p) = (-p^2)^2\frac{1-p}{p^2} = p^2(1-p).$$

4. Let us first compute the Fisher information of the model. We have

$$\frac{\partial}{\partial p}\ell_1(x_1;p) = \frac{1}{p} - \frac{X_1-1}{1-p}, \qquad \frac{\partial^2}{\partial p^2}\ell_1(x_1;p) = -\frac{1}{p^2} - \frac{X_1-1}{(1-p)^2}.$$

As a consequence,

$$I(p) = -\mathbb{E}_p\left[\frac{\partial^2}{\partial p^2}\ell_1(X_1;p)\right] = \frac{1}{p^2} + \frac{1/p-1}{(1-p)^2} = \frac{1}{p^2(1-p)}.$$

Since the Fisher information and the asymptotic variance of the MLE satisfy the identity $I(p)=1/V(p)$, we deduce that the MLE is asymptotically efficient.

Correction of Exercise 5.A.4

1. Since the model is Gaussian, $\frac{T}{n}\widehat{\lambda}_n = \overline{X}_n$ and $\frac{T}{n}\widehat{\sigma}_n^2 = V_n$, therefore
$$\widehat{\lambda}_n = \frac{1}{T}\sum_{i=1}^n X_i, \qquad \widehat{\sigma}_n^2 = \frac{1}{T}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2.$$

2. We have
$$\mathbb{E}_{\lambda,\sigma^2}[\widehat{\lambda}_n] = \frac{n}{T}\mathbb{E}_{\lambda,\sigma^2}[X_1] = \lambda,$$
thus $\widehat{\lambda}_n$ is not biased. Moreover, by independence of the X_i,
$$\mathrm{Var}_{\lambda,\sigma^2}(\widehat{\lambda}_n) = \frac{n}{T^2}\,\mathrm{Var}_{\lambda,\sigma^2}(X_1) = \frac{\sigma^2}{T}.$$

3. Since the variables X_i are independent and Gaussian, their sum is Gaussian and
$$\widehat{\lambda}_n \sim \mathcal{N}\left(\lambda, \frac{\sigma^2}{T}\right).$$
If the estimator $\widehat{\lambda}_n$ was consistent, then it would in particular converge in distribution to the deterministic variable λ, which is not the case since its law remains constant (and with positive variance).

4. Since $\mathbb{E}_{\lambda,\sigma^2}[V_n] = (1-\frac{1}{n})\,\mathrm{Var}_{\lambda,\sigma^2}(X_1)$, we deduce that $\mathbb{E}_{\lambda,\sigma^2}[\widehat{\sigma}_n^2] = (1-\frac{1}{n})\sigma^2$, so $\widehat{\sigma}_n^2$ is biased.

5. We have
$$\widehat{\sigma}_n^2 = \frac{\sigma^2}{n}\sum_{i=1}^n \left(G_i - \overline{G}_n\right)^2.$$
Since the variables G_i are independent and with law $\mathcal{N}(0,1)$, we have
$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \left(G_i - \overline{G}_n\right)^2 = \mathrm{Var}(G_1) = 1, \qquad \text{in probability,}$$
therefore the estimator $\widehat{\sigma}_n^2$ is consistent.

Remark. One may show that even if we observe the whole trajectory $(S_t)_{t\in[0,T]}$, in continuous time, there exists no consistent estimator of λ. The Black–Scholes theory however states that to price and hedge an option on the asset, it is not necessary to know the value of this parameter: only the estimation of σ^2, which is called the *volatility*, is required.

Correction of Exercise 5.A.5

1. We have $\mathbb{E}_\theta[X_1] = \theta + 1/2$, so that $\widetilde{\theta}_n = \overline{X}_n - 1/2$ is a moment estimator. It is unbiased, strongly consistent, asymptotically normal with asymptotic variance $1/12$.

2. The likelihood of a realisation $\mathbf{x}_n \in \mathbb{R}^n$ writes
$$L_n(\mathbf{x}_n;\theta) = \prod_{i=1}^n \mathbb{1}_{\{\theta\le x_i\le\theta+1\}} = \mathbb{1}_{\{\max_{1\le i\le n} x_i - 1\le\theta\le\min_{1\le i\le n} x_i\}},$$
which is plotted on Figure D.1. We observe that any choice of θ between $\max_{1\le i\le n} x_i - 1$ and $\min_{1\le i\le n} x_i$ maximises this likelihood: the MLE is not unique.

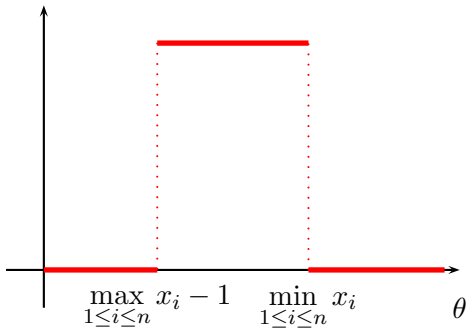


Figure D.1: The likelihood in Exercise 5.A.5.

3. By the same arguments as in Exercise 1.4.7, we have, $\mathbb{P}_\theta$-almost surely, $\min_{1\le i\le n} X_i \to \theta$ and $\max_{1\le i\le n} X_i \to \theta + 1$. As a consequence,

$$\lim_{n\to+\infty} \widehat{\theta}_n^t = (1-t)(\theta + 1 - 1) + t\theta, \qquad \mathbb{P}_\theta\text{-almost surely,}$$

so that $\widehat{\theta}_n^t$ is strongly consistent.

4. Let $x \in [0,1]$. For all $n \ge 1$,

$$\begin{aligned}\mathbb{P}_\theta\left(\min_{1\le i\le n} X_i - \theta \le x\right) &= 1 - \mathbb{P}_\theta\left(\min_{1\le i\le n} X_i - \theta > x\right)\\ &= 1 - \mathbb{P}_\theta(X_1 - \theta > x)^n\\ &= 1 - (1-x)^n,\end{aligned}$$

so that under $\mathbb{P}_\theta$, $U = \min_{1\le i\le n} X_i - \theta \sim \beta(1,n)$. By symmetry, $V = \max_{1\le i\le n} X_i - \theta \sim \beta(n,1)$. Recalling that the expectation of a $\beta(a,b)$ random variable is $a/(a+b)$, we deduce that

$$\mathbb{E}_\theta[\widehat{\theta}_n^t] = (1-t)\left(\theta + \frac{n}{n+1} - 1\right) + t\left(\theta + \frac{1}{n+1}\right) = \theta + \frac{2t-1}{n+1}.$$

As a consequence, the estimator $\widehat{\theta}_n^t$ is unbiased for $t = 1/2$.

5. The MSE writes

$$\begin{aligned}R(\widehat{\theta}_n^t;\theta) &= \mathbb{E}_\theta[(\widehat{\theta}_n^t - \theta)^2]\\ &= \mathbb{E}[((1-t)(V-1) + tU)^2]\\ &= (1-t)^2\mathbb{E}[(V-1)^2] + 2t(1-t)\mathbb{E}[(V-1)U] + t^2\mathbb{E}[U^2].\end{aligned}$$

Since $V \sim \beta(n,1)$, we have $1 - V \sim \beta(1,n)$ and therefore $\mathbb{E}[(1-V)^2] = \mathbb{E}[U^2] = \alpha$, so that

$$R(\widehat{\theta}_n^t;\theta) = \alpha((1-t)^2 + t^2) - \gamma 2t(1-t) =,$$

which is easily seen to reach its minimum over t for $t = 1/2$.

6. For $t = 1/2$, the value of the MSE is

$$R(\widehat{\theta}_n^{1/2};\theta) = \frac{\alpha - \gamma}{2}.$$

On the one hand, a straightforward computation for Beta distributions shows that

$$\alpha = \mathbb{E}[U^2] = \int_{u=0}^{1} u^2 n(1-u)^{n-1}\mathrm{d}u = \frac{2}{(n+1)(n+2)}.$$

On the other hand, to compute β we need to compute the joint distribution of (U,V). To this aim, we write, for any bounded function $f:[0,1]^2\to\mathbb{R}$,

$$\begin{aligned}\mathbb{E}[f(U,V)] &= \int_{u_1,\dots,u_n=0}^{1} f\left(\min_{1\le i\le n} u_i, \max_{1\le i\le n} u_i\right)\mathrm{d}u_1\cdots\mathrm{d}u_n\\ &= n!\int_{0\le u_1\le\cdots\le u_n\le 1} f(u_1,u_n)\mathrm{d}u_1\cdots\mathrm{d}u_n\\ &= n!\int_{u_1=0}^{1}\int_{u_n=0}^{1} f(u_1,u_n)I_n(u_1,u_n)\mathrm{d}u_1\mathrm{d}u_n,\end{aligned}$$

where

$$\begin{aligned}I_n(u_1,u_n) &= \int_{u_2,\dots,u_{n-1}=0}^{1}\mathbb{1}_{\{u_1\le u_2\le\cdots\le u_n-1\le u_n\}}\mathrm{d}u_2\cdots\mathrm{d}u_{n-1}\\ &= \mathbb{1}_{\{u_1\le u_n\}}\int_{u_2=u_1}^{u_n}\int_{u_3=u_2}^{u_n}\cdots\int_{u_{n-1}=u_{n-2}}^{u_n}\mathrm{d}u_{n-1}\cdots\mathrm{d}u_2\\ &= \mathbb{1}_{\{u_1\le u_n\}}\frac{(u_n-u_1)^{n-2}}{(n-2)!}.\end{aligned}$$

We deduce that the density of the pair (U,V) writes

$$q_n(u,v)=\mathbb{1}_{\{u\le v\}}n(n-1)(v-u)^{n-2}$$

on $[0,1]^2$. We may now compute

$$\gamma=\mathbb{E}[(1-V)U]=\int_{u,v=0}^{1}(1-v)uq_n(u,v)\mathrm{d}u\mathrm{d}v=\frac{1}{(n+1)(n+2)},$$

from which we finally deduce that

$$R(\widehat{\theta}_n^{1/2};\theta)=\frac{1}{2(n+1)(n+2)}.$$

We notice that the MSE is of order $1/n^2$, while for regular model, it is generally expected to be of order $1/n$.

D.6 Introduction to Bayesian Estimation

Correction of Exercise 6.A.1

1. By Definition 6.1.1, the posterior $Q(\mathrm{d}\mu|\mathbf{x}_n)$ has density

$$\begin{aligned}q(\mu|\mathbf{x}_n) &\propto L_n(\mathbf{x}_n;\mu)q(\mu)\\ &\propto\prod_{i=1}^{n}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu-m)^2}{2s^2}\right)\\ &\propto\exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}+\frac{(\mu-m)^2}{s^2}\right)\right).\end{aligned}$$

Let us rewrite

$$\begin{aligned}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}+\frac{(\mu-m)^2}{s^2} &= \mu^2\left(\frac{n}{\sigma^2}+\frac{1}{s^2}\right)-2\mu\left(\frac{n\overline{x}}{\sigma^2}+\frac{m}{s^2}\right)+\left(\frac{n\overline{x^2}}{\sigma^2}+\frac{m^2}{s^2}\right)\\ &= \frac{1}{v_n}\left(\mu-\rho(\mathbf{x}_n)\right)^2+\mathrm{Cte},\end{aligned}$$

with

$$\frac{1}{v_n} = \frac{n}{\sigma^2} + \frac{1}{s^2}, \qquad \rho(\mathbf{x}_n) = v_n \left(\frac{n\overline{x}}{\sigma^2} + \frac{m}{s^2} \right).$$

We deduce that

$$q(\mu|\mathbf{x}_n) \propto \exp\left(-\frac{(\mu - \rho(\mathbf{x}_n))^2}{2v_n} \right),$$

which implies that $Q(\cdot|\mathbf{x}_n) = \mathcal{N}(\rho(\mathbf{x}_n), v_n)$.

2. Since the posterior remains in $\mathcal{Q}$, the Gaussian prior is conjugate.

3. Since the Gaussian density reaches its maximum in its expectation, we have

$$\widehat{\mu}_n^{\text{PM}} = \widehat{\mu}_n^{\text{MAP}} = \rho(\mathbf{X}_n) = \frac{\frac{n\overline{X}_n}{\sigma^2} + \frac{m}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}} = (1 - h_n)\overline{X}_n + h_n m, \qquad h_n := \frac{1}{\frac{ns^2}{\sigma^2} + 1}.$$

4. Since $h_n \to 0$ when $n \to +\infty$, the PM is strongly consistent. Besides, $v_n \to 0$. Therefore by Proposition 6.2.9, the Bayesian estimation is consistent.

5. Since $Q(\cdot|\mathbf{x}_n) = \mathcal{N}(\rho(\mathbf{x}_n), v_n)$, a credible region with level $1 - \alpha$ for μ is provided by

$$\left[\rho(\mathbf{X}_n) - \phi_{1-\alpha/2}\sqrt{v_n}, \rho(\mathbf{X}_n) + \phi_{1-\alpha/2}\sqrt{v_n} \right].$$

We notice that, when $n \to +\infty$, the length of this interval is of order $2\phi_{1-\alpha/2}\sigma/\sqrt{n}$, which is the same as the frequentist confidence interval.

Correction of Exercise 6.A.2 Let $Q^* = t_1 Q_{h_1} + \cdots + t_m Q_{h_m} \in \mathcal{Q}^*$. The associated posterior writes

$$Q^*(\mathrm{d}\theta|\mathbf{x}_n) = \frac{L_n(\mathbf{x}_n;\theta)Q^*(\mathrm{d}\theta)}{\displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q^*(\mathrm{d}\vartheta)} = \frac{\displaystyle\sum_{k=1}^m t_k L_n(\mathbf{x}_n;\theta)Q_{h_k}(\mathrm{d}\theta)}{\displaystyle\sum_{l=1}^m t_l \int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q_{h_l}(\mathrm{d}\vartheta)}.$$

Since the family $\mathcal{Q}$ is conjugate, for any $k \in \{1, \ldots, m\}$, the posterior

$$Q_{h_k}(\mathrm{d}\theta|\mathbf{x}_n) = \frac{L_n(\mathbf{x}_n;\theta)Q_{h_k}(\mathrm{d}\theta)}{\displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q_{h_k}(\mathrm{d}\vartheta)}$$

remains in $\mathcal{Q}$. We therefore deduce that

$$Q^*(\mathrm{d}\theta|\mathbf{x}_n) = \sum_{k=1}^m t'_k Q_{h_k}(\mathrm{d}\theta|\mathbf{x}_n) \in \mathcal{Q}^*, \qquad t'_k := \frac{t_k \displaystyle\int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q_{h_k}(\mathrm{d}\vartheta)}{\displaystyle\sum_{l=1}^m t_l \int_{\vartheta\in\Theta} L_n(\mathbf{x}_n;\vartheta)Q_{h_l}(\mathrm{d}\vartheta)},$$

which shows that $\mathcal{Q}^*$ is conjugate.

Correction of Exercise 6.A.5

1. The posterior is proportional to

$$L_n(\mathbf{x}'_n, \mathbf{y}_n; \beta) q_k(\beta) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}_n - \mathbf{x}'_n\beta\|^2\right) \exp\left(-\frac{\|\beta\|^2}{2k}\right),$$

which shows that the MAP minimises

$$\frac{1}{2\sigma^2}\|\mathbf{y}_n - \mathbf{x}'_n\beta\|^2 + \frac{\|\beta\|^2}{2k} = \frac{1}{2\sigma^2}\left(\|\mathbf{y}_n - \mathbf{x}'_n\beta\|^2 + h\|\beta\|^2\right),$$

with $h = \sigma^2/k$.

2. The LASSO penalisation is obtained by taking as a prior for β a product of Laplace distributions

$$q(\beta) = \prod_{j=0}^{p} \frac{h}{2} \exp(-h|\beta_j|).$$

☑ Intermediate Revision Sheet

Correction of Exercise 1

5. (a) The posterior has density, for any $\mathbf{x}_n \in (0, +\infty)^n$ and $\theta > 0$,

$$\begin{aligned} q_{a,b}(\theta|\mathbf{x}_n) &= \frac{L_n(\mathbf{x}_n|\theta) q_{a,b}(\theta)}{\int_{\vartheta=0}^{+\infty} L_n(\mathbf{x}_n|\vartheta) q_{a,b}(\vartheta) \mathrm{d}\vartheta} \\ &= \frac{\left(\prod_{i=1}^{n} \sqrt{\frac{2}{\pi\theta}} \exp\left(-\frac{x_i^2}{2\theta}\right)\right)\left(\frac{b^a}{\Gamma(a)}\left(\frac{1}{\theta}\right)^{a+1} \exp\left(-\frac{b}{\theta}\right)\right)}{\int_{\vartheta=0}^{+\infty}\left(\prod_{i=1}^{n} \sqrt{\frac{2}{\pi\vartheta}} \exp\left(-\frac{x_i^2}{2\vartheta}\right)\right)\left(\frac{b^a}{\Gamma(a)}\left(\frac{1}{\vartheta}\right)^{a+1} \exp\left(-\frac{b}{\vartheta}\right)\right) \mathrm{d}\vartheta} \\ &= \text{Constant} \times \left(\frac{1}{\theta}\right)^{a'+1} \exp\left(-\frac{b'}{\theta}\right) = q_{a',b'}(\theta), \end{aligned}$$

with

$$a' = a + \frac{n}{2}, \qquad b' = b + \frac{1}{2}\sum_{i=1}^{n} x_i^2.$$

(b) In particular, the inverse Gamma distribution is a conjugate prior for the model $\{P_\theta, \theta > 0\}$.

(c) The posterior mean is

$$\widehat{\theta}_n^{\mathrm{PM}} = \frac{b + \frac{1}{2}\sum_{i=1}^{n} X_i^2}{a + \frac{n}{2} - 1},$$

so by the strong Law of Large Numbers, $\widehat{\theta}_n^{\mathrm{PM}}$ converges to θ, $\mathbb{P}_\theta$-almost surely.

(d) To show the consistency of Bayesian estimation, by Proposition 6.2.9 it suffices now to show that

$$V_n(\mathbf{X}_n) = \frac{\left(b + \frac{1}{2}\sum_{i=1}^{n} X_i^2\right)^2}{\left(a + \frac{n}{2} - 1\right)^2\left(a + \frac{n}{2} - 2\right)}$$

converges to 0, $\mathbb{P}_\theta$-almost surely. To this aim we rewrite

$$V_n(\mathbf{X}_n) = \frac{1}{n} \frac{\left(\frac{b}{n} + \frac{1}{2n}\sum_{i=1}^n X_i^2\right)^2}{\left(\frac{a-1}{n} + \frac{1}{2}\right)^2 \left(\frac{a-2}{n} + \frac{1}{2}\right)} \sim \frac{(\theta/2)^2}{n(1/2)^3},$$

which concludes.

Correction of Exercise 2

1. On the one hand, $y_{n+1} = x'_{n+1}\beta + \epsilon_{n+1} \sim \mathcal{N}(x'_{n+1}\beta, \sigma^2)$, and this variable only depends on ϵ_{n+1}. On the other hand, we know from Proposition 3.3.3 that $\widehat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})$ and $(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1)$, and this variables are independent. As a consequence, $\widehat{y}_{n+1} = x'_{n+1}\widehat{\beta} \sim \mathcal{N}(x'_{n+1}\beta, \sigma^2 x'_{n+1}(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})(x'_{n+1})^\top)$, and since $\widehat{\beta}, \widehat{\sigma}^2$ only depend on $\epsilon_1, \ldots, \epsilon_n$, they are independent from y_{n+1}.

2. We first deduce from the previous question that

$$y_{n+1} - \widehat{y}_{n+1} \sim \mathcal{N}(0, \underbrace{\sigma^2 + \sigma^2 x'_{n+1}(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})(x'_{n+1})^\top}_{\sigma^2\kappa}),$$

and that this variable is independent from $\widehat{\sigma}^2$. As a consequence,

$$\frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{\widehat{\sigma}^2\kappa}} \sim \mathrm{t}(n-p-1).$$

3. We deduce from the previous question that

$$\mathbb{P}_{\beta,\sigma^2}\left(\left|\frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{\widehat{\sigma}^2\kappa}}\right| \leq t_{n-p-1,1-\alpha/2}\right) = 1 - \alpha,$$

which implies that

$$\left[\widehat{y}_{n+1} - t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2\kappa}, \widehat{y}_{n+1} + t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2\kappa}\right]$$

is a prediction interval for y_{n+1}.

4. With the code below we find a predicted price of 11779, with interval $[5393, 18164]$.

```
import numpy as np
from sklearn.linear_model import LinearRegression
from scipy import stats

# Original data
data = np.array([
    [18500, 32000, 24],
    [14500, 50000, 36],
    [22000, 18000, 12],
    [9500,  87000, 60],
    [7800,  120000, 84],
    [16500, 40000, 30],
    [11200, 70000, 48],
    [19900, 25000, 18],
    [8800,  95000, 72],
    [13200, 60000, 42]
])
```

```
X = data[:, 1:]  # [Kilometers, Age]
y = data[:, 0]   # Selling Price

# Fit model
model = LinearRegression()
model.fit(X, y)

# Prediction point
x_new = np.array([[80000, 48]])
y_pred = model.predict(x_new)[0]

# Calculate residuals and standard error
y_hat = model.predict(X)
residuals = y - y_hat
n = len(y)
p = X.shape[1]
df = n - p - 1  # degrees of freedom

# Standard error of estimate
s_err = np.sqrt(np.sum(residuals**2) / df)

# Add a column of ones for intercept
X_augmented = np.hstack([np.ones((n, 1)), X])
x_new_augmented = np.array([1, 80000, 48])

# (X^T X)^(-1)
xtx_inv = np.linalg.inv(X_augmented.T @ X_augmented)

# Compute prediction standard error
pred_var = s_err**2 * (1 + x_new_augmented @ xtx_inv @ x_new_augmented.T)
pred_std = np.sqrt(pred_var)

# 95% prediction interval
t_val = stats.t.ppf(0.975, df)
interval = t_val * pred_std

lower = y_pred - interval
upper = y_pred + interval

# Output
print(f"Predicted price: ${y_pred:.2f}")
print(f"95% prediction interval: (${lower:.2f}, ${upper:.2f})")
```

D.7 The Formalism of Statistical Hypothesis Testing

Correction of Exercise 7.A.1

1. In the Bernoulli model, the MLE of p is $\widehat{p}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

2. The estimator $\widehat{p}_n$ takes smaller values under H_1 than under H_0. We therefore choose a rejection region of the form $W_n = \{\widehat{p}_n \leq a\}$. For $p \geq p_0$, the type I risk writes

$$\mathbb{P}_p\left(W_n\right) = \mathbb{P}_p\left(\widehat{p}_n \leq a\right) \leq \mathbb{P}_{p_0}\left(\widehat{p}_n \leq a\right),$$

whence

$$\sup_{p \geq p_0} \mathbb{P}_p\left(W_n\right) = \mathbb{P}_{p_0}\left(\widehat{p}_n \leq a\right) = \mathbb{P}_{p_0}\left(\overline{X}_n \leq a\right).$$

Under $\mathbb{P}_{p_0}$, $\sqrt{n}(\overline{X}_n - p_0)$ converges in law to $\mathcal{N}(0, p_0(1-p_0))$, therefore

$$\lim_{n \to +\infty} \mathbb{P}_{p_0}\left(\sqrt{\frac{n}{p_0(1-p_0)}}(\overline{X}_n - p_0) \leq \phi_\alpha\right) = \alpha.$$

We deduce that taking

$$a = p_0 + \phi_\alpha \sqrt{\frac{p_0(1-p_0)}{n}},$$

we get a test with asymptotic level α. Moreover, if $p < p_0$,

$$\mathbb{P}_p(W_n) = \mathbb{P}_p\left(\overline{X}_n \leq p_0 + \phi_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right) = \mathbb{E}_p\left[\mathbb{1}_{\{\overline{X}_n - p_0 \leq \phi_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\}}\right].$$

By the strong Law of Large Numbers, $\overline{X}_n$ converges $\mathbb{P}_p$-almost surely toward $p < p_0$, therefore the indicator in the right-hand side goes to 1. By the Dominated Convergence Theorem, we conclude that

$$\lim_{n \to +\infty} \mathbb{P}_p(W_n) = 1,$$

that is to say that the test is consistent.

3. (a) Since the rejection region writes $\{\widehat{p}_n \leq a\}$, the p-value associated with an observation $\widehat{p}_n^{\text{obs}}$ writes

$$p\text{-value} = \sup_{p \geq p_0} \mathbb{P}_p\left(\widehat{p}_n \leq \widehat{p}_n^{\text{obs}}\right) = \mathbb{P}_{p_0}\left(\widehat{p}_n \leq \widehat{p}_n^{\text{obs}}\right).$$

Here, $\widehat{p}_n^{\text{obs}} = 0$ so

$$p\text{-value} = \mathbb{P}_{p_0}\left(\widehat{p}_n = 0\right) = \mathbb{P}_{p_0}\left(\forall i \in \{1, \ldots, n\}, X_i = 0\right) = (1-p_0)^n.$$

(b) We get that the p-value is equal to $(1 - 0.005)^{1046} = 5.3\ 10^{-3}$. The claim that 'the mortality rate is larger than 0.5%' is therefore rejected at all usual levels, and the sentence extracted from the article is not statistically incorrect, for the considered sample.

4. It may however be objected that sailors on the aircraft carrier are essentially young and in good shape. The mortality rate for this specific population may therefore be assumed to be much lower than among the overall population.

Correction of Exercise 7.A.3

1. R is the number of experiments for which H_0 is rejected, that is to say

$$R = \sum_{k=1}^{m} \mathbb{1}_{\{p_k \leq p_{(J)}\}},$$

which by definition of $p_{(1)} < \cdots < p_{(m)}$ ensures that $R = J$.

2. We have $V = \sum_{k \in \mathcal{I}} V_k$ and

$$\text{FDR}(\theta) = \mathbb{E}_\theta\left[\frac{V}{R}\right] = \sum_{k \in \mathcal{I}} \mathbb{E}_\theta\left[\frac{V_k}{J}\right] = \sum_{k \in \mathcal{I}} \sum_{j=1}^{m} \frac{1}{j} \mathbb{E}_\theta\left[V_k \mathbb{1}_{\{J=j\}}\right].$$

3. It follows from the definition of the Benjamini–Hochberg procedure that, for all $j \in \{1, \ldots, n\}$,

$$V_k \mathbb{1}_{\{J=j\}} = V_k \mathbb{1}_{\{J_k=j\}} = \mathbb{1}_{\{p_k \leq \alpha j/m\}} \mathbb{1}_{\{J_k=j\}},$$

so that

$$\text{FDR}(\theta) = \sum_{k \in \mathcal{I}} \sum_{j=1}^{m} \frac{1}{j} \mathbb{E}_\theta\left[\mathbb{1}_{\{p_k \leq \alpha j/m\}} \mathbb{1}_{\{J_k=j\}}\right].$$

4. Since J_k only depends on $(p_1, \dots, p_{k-1}, p_{k+1}, \dots, p_m)$, it is independent of p_k so that
$$\mathbb{E}_\theta\left[\mathbb{1}_{\{p_k \le \alpha j/m\}} \mathbb{1}_{\{J_k = j\}}\right] = \mathbb{E}_\theta\left[\mathbb{1}_{\{p_k \le \alpha j/m\}}\right] \mathbb{E}_\theta\left[\mathbb{1}_{\{J_k = j\}}\right],$$
and the fact that $k \in \mathcal{I}$ implies that under $\mathbb{P}_\theta$, the p-value p_k is uniformly distributed on $[0,1]$, therefore
$$\mathbb{E}_\theta\left[\mathbb{1}_{\{p_k \le \alpha j/m\}}\right] = \frac{\alpha j}{m}.$$
We deduce that
$$\mathrm{FDR}(\theta) = \sum_{k \in \mathcal{I}} \sum_{j=1}^{m} \frac{1}{j} \frac{\alpha j}{m} \mathbb{E}_\theta\left[\mathbb{1}_{\{J_k = j\}}\right] = \alpha \frac{m_0}{m},$$
which completes the proof.

D.8 Tests in the Gaussian Model

Correction of Exercise 8.A.1 In the context of simple regression, we recall that
$$|\widehat{\beta}_1| = \frac{|\operatorname{Cov}(\mathbf{x}_n, \mathbf{y}_n)|}{\operatorname{Var}(\mathbf{x}_n)} = \sqrt{\frac{\operatorname{Var}(\mathbf{y}_n)}{\operatorname{Var}(\mathbf{x}_n)}} |R|,$$
and
$$\sigma^2 = \frac{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2}{n-2} = \frac{n \operatorname{Var}(\mathbf{y}_n)(1-R^2)}{n-2}.$$
Besides,
$$\mathbf{x}_n'^{\top} \mathbf{x}'_n = n \begin{pmatrix} 1 & \overline{x}_n \\ \overline{x}_n & \overline{x^2}_n \end{pmatrix},$$
from which we deduce that
$$(\mathbf{x}_n'^{\top} \mathbf{x}'_n)^{-1} = \frac{1}{n \operatorname{Var}(\mathbf{x}_n)} \begin{pmatrix} \overline{x^2}_n & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix}$$
and therefore
$$\rho_1 = \frac{1}{n \operatorname{Var}(\mathbf{x}_n)}.$$
As a consequence, Student's statistic writes
$$T = \left| \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2 \rho_1}} \right| = \sqrt{\frac{(n-2)R^2}{1-R^2}},$$
and H_0 is rejected if this quantity is larger than $t_{n-2,1-\alpha/2}$.

On the other hand, Fisher's statistics writes
$$F = \frac{\|\widehat{\mathbf{y}}_n - \widehat{\mathbf{y}}_n^0\|^2}{\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2/(n-2)},$$
and H_0 is rejected if it is larger than $f_{1,n-2,1-\alpha}$. As has already been noted, in the denominator we have
$$\|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2 = n \operatorname{Var}(\mathbf{y}_n)(1-R^2),$$
while by definition of R^2,
$$\|\widehat{\mathbf{y}}_n - \widehat{\mathbf{y}}_n^0\|^2 = n \operatorname{Var}(\widehat{\mathbf{y}}_n) = nR^2 \operatorname{Var}(\mathbf{y}_n).$$
We deduce that
$$F = \frac{(n-2)R^2}{1-R^2} = T^2,$$
and conclude that since it is easily checked that
$$f_{1,n-2,1-\alpha} = t^2_{n-2,1-\alpha/2},$$
both tests actually coincide.

Correction of Exercise 8.A.3

1. We have $\widehat{\sigma} = \|\mathbf{y}_n - \widehat{\mathbf{y}}_n\|^2/(n-p-1) \simeq 0.0895$.

2. Since $\widehat{\beta} \sim \mathcal{N}_3(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})$, the random variable $\widehat{\beta}_0$ is Gaussian, with mean $\beta_0 = \log A$ and variance $\sigma^2\rho_0$, where ρ_0 is the first diagonal coefficient of the matrix $(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}$. In addition, $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent and $(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1)$, so that

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\widehat{\sigma}^2\rho_0}} \sim \mathrm{t}(n-p-1).$$

   As a consequence, a confidence interval with level α for β_0 is given by

$$\widehat{\beta}_0 \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2\rho_0},$$

   which yields the confidence interval $[20.82950, 25.42238]$ for A.

3. We let $x' = (1, \log 100, \log 50)$ and use the formula

$$\widehat{y} \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\sigma}^2\kappa}, \qquad \widehat{y} = x'\widehat{\beta}, \quad \kappa = 1 + x'(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}x'^\top,$$

   for the prediction interval for the associated value of y. Taking the exponential, we obtain the prediction interval $[1122, 3837]$ for Y.

4. We apply Student's test. We have

$$t_{n-p-1,1-\alpha/2} = 1.96, \qquad \left|\frac{\widehat{\beta}_2}{\sqrt{\widehat{\sigma}^2\rho_2}}\right| = 31.4,$$

   so that the hypothesis $\{\beta_2 = 0\}$ is rejected at the level 5%.

5. (a) With $Y = AL^{\beta_1}K^{\beta_2}$, we have $A(\lambda L)^{\beta_1}(\lambda K)^{\beta_2} = \lambda^{\beta_1+\beta_2}Y$, so that the model displays constant returns to scale if and only if $\beta_1 + \beta_2 = 1$.

   (b) We have $\widehat{\beta} \sim \mathcal{N}_3(\beta, \sigma^2(\mathbf{x}_n^\top \mathbf{x}_n)^{-1})$ and want to test the hypotheses

$$H_0 = \{\beta_1 + \beta_2 = 1\}, \qquad H_1 = \{\beta_1 + \beta_2 \neq 1\}.$$

   Let us define $u = (0, 1, 1)$, so that

$$\widehat{\beta}_1 + \widehat{\beta}_2 = \langle u, \widehat{\beta}\rangle \sim \mathcal{N}\left(\beta_1 + \beta_2, \sigma^2 u(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}u^\top\right).$$

   Under H_0, we therefore deduce that

$$\frac{\widehat{\beta}_1 + \widehat{\beta}_2 - 1}{\sqrt{\widehat{\sigma}^2 u(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}u^\top}} \sim \mathrm{t}(n-p-1).$$

   This test statistic takes the value

$$\frac{|\widehat{\beta}_1 + \widehat{\beta}_2 - 1|}{\sqrt{\widehat{\sigma}^2 u(\mathbf{x}_n^\top \mathbf{x}_n)^{-1}u^\top}} = 0.9960,$$

   so that the p-value of the t-test is 0.319. As a consequence, the hypothesis of constant returns to scale is not rejected.

The whole code for our computation is reported below.

```
import numpy as np
from scipy.stats import t

# Define the coefficients
beta0 = 3.136
beta1 = 0.738
beta2 = 0.282
R2 = 0.945
SCR = 148.27

# Define the matrix XnTXn_inv
XnTXn_inv = np.array([[0.0288, 0.0012, -0.0034],
                      [0.0012, 0.0016,  0.0010],
                      [-0.0034, 0.0010, 0.0009]])

n = 1658
p = 2

# Define the significance level and the quantile for the t-distribution
alpha = 0.05
quantile_for_t = t.ppf(1 - alpha/2, df=n - p - 1)

# Value of sigma^2
sigma2 = SCR / (n - p - 1)
print("sigma2:', sigma2)

# Confidence interval for A = exp(beta0)
inf_for_beta0 = beta0 - quantile_for_t * np.sqrt(sigma2 * XnTXn_inv[0, 0])
sup_for_beta0 = beta0 + quantile_for_t * np.sqrt(sigma2 * XnTXn_inv[0, 0])
CI_for_A = np.exp([inf_for_beta0, sup_for_beta0])
print("Confidence interval for A:", CI_for_A)

# Prediction interval for a new observation
new_x = np.array([1, np.log(100), np.log(50)])
kappa = 1 + new_x @ XnTXn_inv @ new_x
predictor_of_new_y = new_x @ np.array([beta0, beta1, beta2])
inf_for_new_y = predictor_of_new_y - quantile_for_t * np.sqrt(sigma2 * kappa)
sup_for_new_y = predictor_of_new_y + quantile_for_t * np.sqrt(sigma2 * kappa)
CI_for_new_Y = np.exp([inf_for_new_y, sup_for_new_y])
print("Prediction interval for new Y:", CI_for_new_Y)

# Test of significance for beta2
t_stat = abs(beta2 / np.sqrt(sigma2 * XnTXn_inv[2, 2]))
print("t-statistic for beta2:", t_stat)

# Test for constant returns to scale
u = np.array([0, 1, 1])
crs_stat = abs(beta1 + beta2 - 1) / np.sqrt(sigma2 * (u @ XnTXn_inv @ u))
p_val = 2 * (1 - t.cdf(crs_stat, df=n - p - 1))
print("CRS test statistic:", crs_stat)
print("p-value for CRS test:", p_val)
```

D.9 The Wald and Likelihood Ratio Tests

Correction of Exercise 9.A.1

1. Since one wants to evidence the existence of a bias toward the face up at the beginning, one takes as the null hypothesis the absence of this bias.

2. Let us denote by X_i the binary variable equal to 1 if, at the i-th throw, the coin has landed on the same face as the starting position, and 0 else. The empirical mean $\overline{X}_n$ is a strongly consistent

estimator of p. We therefore set

$$W_n = \left\{ \overline{X}_n \geq \frac{1}{2} + a_n \right\},$$

and we look for the threshold $a_n \geq 0$ which minimises the type II risk under the constraint that the type I risk be (asymptotically) smaller than α. Using the Gaussian approximation

$$\overline{X}_n \simeq p + \sqrt{\frac{p(1-p)}{n}} G, \qquad G \sim \mathcal{N}(0,1),$$

the type I risk is

$$\mathbb{P}_{1/2}(W_n) \simeq \mathbb{P}\left(\frac{1}{2} + \sqrt{\frac{1}{4n}} G \geq \frac{1}{2} + a_n \right) = \mathbb{P}(G \geq 2\sqrt{n} a_n),$$

and the latter quantity is smaller than α if and only if $a_n \geq \phi_{1-\alpha}/2\sqrt{n}$. Since the type II risk is an increasing function of a_n, it is minimal when a_n takes the smallest value satisfying the constraint on the type I risk, that is to say $a_n = \phi_{1-\alpha}/2\sqrt{n}$. We deduce that the test with rejection region

$$W_n = \left\{ \overline{X}_n \geq \frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}} \right\}$$

has asymptotic level α. Moreover, if $p > 1/2$,

$$\mathbb{P}_p\left(W_n\right) = \mathbb{P}_p\left(\overline{X}_n \geq \frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}} \right).$$

Since $\overline{X}_n \to p$, $\mathbb{P}_p$-almost surely, and $\frac{\phi_{1-\alpha}}{2\sqrt{n}} \to 0$, we have

$$\lim_{n \to +\infty} \mathbb{1}_{\{\overline{X}_n \geq \frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}}\}} = \mathbb{1}_{\{p \geq \frac{1}{2}\}} = 1, \qquad \mathbb{P}_p\text{-almost surely,}$$

and therefore, by the dominated convergence theorem, the test is consistent.

3. The observed value of $\overline{X}_n$ is $20\,245/40\,000 \simeq 0.506$. Using the Gaussian approximation, the p-value of this result thus writes

$$\begin{aligned}
\mathbb{P}_{1/2}\left(\overline{X}_n \geq 20\,245/40\,000\right) &\simeq \mathbb{P}\left(\frac{1}{2} + \frac{1}{2\sqrt{n}} G \geq 20\,245/40\,000 \right) \\
&= \mathbb{P}(G \geq 2 \times (20\,245/40\,000) \times \sqrt{40\,000}) \\
&= 1 - \Phi(2.45) \simeq 0.007.
\end{aligned}$$

Since the p-value is smaller than the level 5%, one rejects H_0, and therefore the conclusion is the existence of the bias.

4. For a fixed value of $p > 1/2$, the power of the test is

$$\mathbb{P}_p(W_n) \simeq \mathbb{P}\left(p + \sqrt{\frac{p(1-p)}{n}} G \geq \frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}} \right) = 1 - \Phi\left(\sqrt{\frac{n}{p(1-p)}} \left(\frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}} - p \right) \right).$$

With $n = 40\,000$, $p = 0.508$ and $\alpha = 0.05$, one has

$$\sqrt{\frac{n}{p(1-p)}} \left(\frac{1}{2} + \frac{\phi_{1-\alpha}}{2\sqrt{n}} - p \right) \simeq -1.55.$$

The power of the test is therefore about 94%.

5. We are dealing here with a test of comparison of proportions. One may employ Wald's test for the identity of means, or anticipating on Lecture 10, the χ_2 test of independence (see Remark 10.4.5 for a detailed discussion of the link between both tests). In any case, one needs to be cautious in the application of the test.

   Wald's test of identity of parameters in the Bernoulli model. The parameters p_{H} and p_{T} are estimated by
   $$\widehat{p}_{\mathrm{H}} = \frac{10\,231}{20\,000}, \qquad \widehat{p}_{\mathrm{T}} = \frac{10\,014}{20\,000}.$$
   Wald's test statistic is
   $$\frac{\widehat{p}_{\mathrm{H}} - \widehat{p}_{\mathrm{T}}}{\sqrt{\dfrac{\widehat{p}_{\mathrm{H}}(1-\widehat{p}_{\mathrm{H}})}{20\,000} + \dfrac{\widehat{p}_{\mathrm{T}}(1-\widehat{p}_{\mathrm{T}})}{20\,000}}} \simeq 2.17.$$
   The p-value of the test writes
   $$\mathbb{P}(|G| > 2.17) = 2(1 - \Phi(2.17)) \simeq 0.03.$$

   χ_2 test of independence. Here, we want to test the independence between the variable 'starting side' and the event 'the landing side is the same as the starting side'. The contingency table to which the test has to be applied is the following:

	Same landing side	Different landing side	Total
Heads initial	10 231	9 769	20 000
Tails initial	10 014	9 986	20 000
Total	20 245	19 755	40 000

   Pearson's statistic is equal to $d'_n \simeq 4.7$. The p-value of the test is $\mathbb{P}(Z > 4.7)$ with $Z \sim \chi_2(1)$, that is to say
   $$\mathbb{P}(G^2 > 4.7) = \mathbb{P}(|G| > \sqrt{4.7}) = 2(1 - \Phi(\sqrt{4.7})) \simeq 0.03.$$

   For both tests, at the confidence level 5%, the null hypothesis is therefore rejected.

Correction of Exercise 9.A.2

1. We obviously take the empirical correlation
   $$\widehat{C}_n = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n}\sum_{i=1}^n X_i \times \frac{1}{n}\sum_{i=1}^n Y_i,$$
   which is strongly consistent by the strong Law of Large Numbers.

2. Our goal is to use the Delta method, since $\widehat{C}_n$ rewrites under the form
   $$\widehat{C}_n = \phi\left(\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} X_i \\ Y_i \\ X_i Y_i \end{pmatrix}\right), \qquad \phi\begin{pmatrix} x \\ y \\ z \end{pmatrix} = z - xy.$$
   On the one hand, we have (writing X, Y for X_1, Y_1 for simplicity)
   $$\mathrm{Cov}\begin{bmatrix} X \\ Y \\ XY \end{bmatrix} = \begin{pmatrix} \mathrm{Var}(X) & 0 & \mathrm{Var}(X)\mathbb{E}[Y] \\ 0 & \mathrm{Var}(Y) & \mathrm{Var}(Y)\mathbb{E}[X] \\ \mathrm{Var}(X)\mathbb{E}[Y] & \mathrm{Var}(Y)\mathbb{E}[X] & \mathrm{Var}(XY) \end{pmatrix}.$$
   On the other hand,
   $$\nabla\phi\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -y \\ -x \\ 1 \end{pmatrix}.$$

By the Delta method, we know that

$$\sqrt{n}\left(\widehat{C}_n - C\right) = \sqrt{n}\left(\phi\left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} X_i \\ Y_i \\ X_iY_i \end{pmatrix}\right) - \phi\left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \\ \mathbb{E}[XY] \end{pmatrix}\right)\right)$$

converges in distribution toward a centered Gaussian random variable with variance

$$\nabla\phi\begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \\ \mathbb{E}[XY] \end{pmatrix}^{\top} \mathrm{Cov}\begin{bmatrix} X \\ Y \\ XY \end{bmatrix} \nabla\phi\begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \\ \mathbb{E}[XY] \end{pmatrix} = \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

3. A strongly consistent estimator of V is given by the product of the empirical variances

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \times \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2.$$

We may therefore conclude that Wald's two-sided test, with rejection region

$$W_n = \left\{ \frac{|\widehat{C}_n|}{\sqrt{\widehat{V}_n/n}} \geq \phi_{1-\alpha/2} \right\},$$

is consistent and has asymptotic level α.

4. Interestingly, the test statistic rewrites

$$\frac{|\widehat{C}_n|}{\sqrt{\widehat{V}_n/n}} = \sqrt{n}|\widehat{\rho}_n|,$$

where $\widehat{\rho}_n$ is the empirical correlation between the samples $(X_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$. So to conclude to the rejection of H_0, n must be larger than $\phi^2_{1-\alpha/2}/\widehat{\rho}^2_n$.

D.10 χ_2 Tests for Finite State Spaces

Correction of Exercise 10.A.1 With $\theta = (p_x)_{x\in E}$, the likelihood of a realisation $\mathbf{x}_n = (x_1, \ldots, x_n) \in E^n$ writes

$$L_n(\mathbf{x}_n; \theta) = \prod_{i=1}^{n} p_{x_i}.$$

Let $\widehat{P}_n$ be the empirical measure associated with the realisation $\mathbf{x}_n$. In the product in the right-hand above, each $x \in E$ appears $n\widehat{p}_{n,x}$ times, so that the likelihood rewrites

$$L_n(\mathbf{x}_n; \theta) = \prod_{x\in E} p_x^{n\widehat{p}_{n,x}},$$

and the associated log-likelihood is

$$\ell_n(\mathbf{x}_n; \theta) = \sum_{x\in E} n\widehat{p}_{n,x}\log(p_x).$$

We now check that this function reaches its maximum on Θ for $\theta = \widehat{P}_n$, by writing, for any $\theta \in \Theta$,

$$\begin{aligned}\ell_n(\mathbf{x}_n; \widehat{P}_n) - \ell_n(\mathbf{x}_n; \theta) &= \sum_{x \in E} n\widehat{p}_{n,x} \log(\widehat{p}_{n,x}) - \sum_{x \in E} n\widehat{p}_{n,x} \log(p_x) \\ &= n \sum_{x \in E} \widehat{p}_{n,x} \log\left(\frac{\widehat{p}_{n,x}}{p_x}\right) \\ &= n \sum_{x \in E} p_x \phi\left(\frac{\widehat{p}_{n,x}}{p_x}\right),\end{aligned}$$

where $\phi(u) = u \log u$. Since this function is convex, Jensen's inequality yields

$$\sum_{x \in E} p_x \phi\left(\frac{\widehat{p}_{n,x}}{p_x}\right) \geq \phi\left(\sum_{x \in E} \frac{\widehat{p}_{n,x}}{p_x} p_x\right) = \phi(1) = 0,$$

which completes the computation.

Correction of Exercise 10.A.2 We perform a χ_2 goodness-of-fit test for a family of distribution. A consistent estimator of λ in the Poisson model is the empirical mean $\widehat{\lambda}_n$, which with the present data takes the value 2.42.

```
import numpy as np

# Define the values of x and the observed counts
x = np.arange(7)
numbers = np.array([23, 75, 68, 51, 53, 30, 0])

# Calculate the total number of observations
n = np.sum(numbers)

# Calculate the estimated probabilities (hat_p)
hat_p = numbers / n

# Calculate lambda as the weighted sum of x with hat_p
lambda_ = np.sum(hat_p * x)
```

Partitioning the state space $\mathbb{N}$ into the 7 classes $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, and $\{6, \ldots\}$, we first compute the probability mass function of $P_{0,\widehat{\lambda}_n}$:

Class	0	1	2	3	4	5	≥ 6
Value of $p_{0,\widehat{\lambda}_n}$	0.09	0.22	0.26	0.21	0.13	0.06	0.03

```
from scipy.special import factorial

# Calculate the probabilities p_0 for each x
p_0 = np.exp(-lambda_) * lambda_**x / factorial(x)

# Adjust the last element of p_0
p_0[6] = 1 - np.sum(p_0[:6])
```

Pearson's statistic then takes the value 29.9, which has to be compared with the quantiles of the $\chi_2(5)$ distribution, as $m = 7$ and $q = 1$ here. We obtain a p-value of $1.6\ 10^{-5}$, which allows the reject the hypothesis that the data are Poisson-distributed at all usual levels.

```
from scipy.stats import chi2

# Compute the statistic dn
dn = n * np.sum((hat_p - p_0)**2 / p_0)
```

```
# Calculate the p-value using the chi-squared distribution
p_value = chi2.sf(dn, df=5)  # sf is the survival function, equivalent to 1 -
    cdf

print("dn:", dn)
print("p_value:", p_value)
```

Alternatively, one may directly adjust the number of degrees of freedom in the χ_2 test using the parameter `ddof` of the function `chisquare()`.

```
from scipy.stats import chisquare

# Perform the chi-squared test
chi2_stat, p_value = chisquare(f_obs=numbers, f_exp=p_0 * n, ddof=1)

print("Chi-squared statistic:", chi2_stat)
print("P-value:", p_value)
```

D.11 Nonparametric Tests for Continuous Data

Correction of Exercise 11.A.1

1. By Definition 11.1.2, for any $t \in [0,1]$, $\beta(t)$ is a centered Gaussian variable, with variance $t - t^2 = t(1-t)$.

2. By Definition 11.1.2, $\mathbf{G}$ is a centered Gaussian vector with covariance matrix given by

$$\begin{aligned}\mathrm{Cov}(\beta(F(x_i)), \beta(F(x_j))) &= \min\{F(x_i), F(x_j)\} - F(x_i)F(x_j) \\ &= F(\min\{x_i, x_j\}) - F(x_i)F(x_j),\end{aligned}$$

   since F is nondecreasing.

3. The random vector $\mathbf{G}_n$ writes

$$\mathbf{G}_n = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} S_i - \mathbb{E}[S_1]\right),$$

   where $S_1, \dots, S_n$ are iid random vectors in $\mathbb{R}^d$ defined by

$$S_i = \left(\mathbb{1}_{\{X_i \le x_1\}}, \dots, \mathbb{1}_{\{X_i \le x_d\}}\right).$$

   By the multidimensional Central Limit Theorem, $\mathbf{G}_n$ converges in distribution to $\mathcal{N}_d(0, K)$ where K is the covariance matrix of S_1. The coefficients of this matrix are given by

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\{X_1 \le x_i\}}\mathbb{1}_{\{X_1 \le x_j\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{X_1 \le x_i\}}\right]\mathbb{E}\left[\mathbb{1}_{\{X_1 \le x_j\}}\right] \\ &= \mathbb{E}\left[\mathbb{1}_{\{X_1 \le \min\{x_i, x_j\}\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{X_1 \le x_i\}}\right]\mathbb{E}\left[\mathbb{1}_{\{X_1 \le x_j\}}\right] \\ &= F(\min\{x_i, x_j\}) - F(x_i)F(x_j),\end{aligned}$$

   so that $\mathbf{G}_n$ converges in distribution to $\mathbf{G}$.

Correction of Exercise 11.A.2

1. Let $x \in \mathbb{R}$. For all $n \ge 1$, the iid random variables $\mathbb{1}_{\{X_1 \le x\}}, \dots, \mathbb{1}_{\{X_n \le x\}}$ take their values in $[0,1]$ and satisfy $\mathbb{E}[\mathbb{1}_{\{X_1 \le x\}}] = F(x)$, therefore by Corollary 4.4.6,

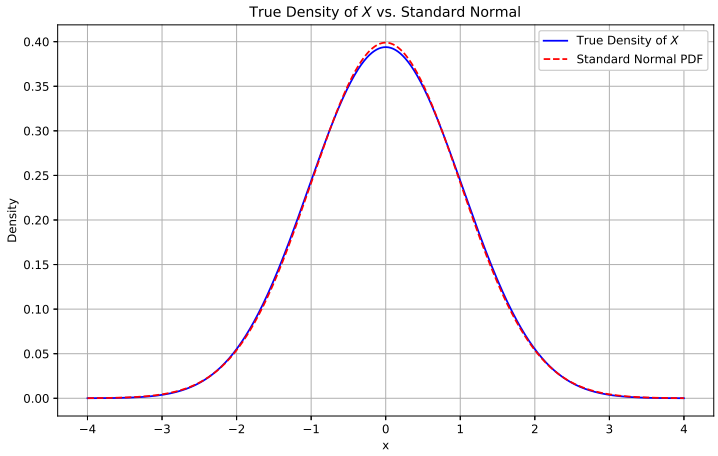
$$\mathbb{P}\left(\sqrt{n}|\widehat{F}_n(x) - F(x)| \ge a\right) \le 2\exp(-2a^2).$$

2. For $\alpha \in (0,1)$, we let
$$a = \sqrt{-\frac{1}{2}\log\frac{\alpha}{2}}.$$
Then the test rejecting H_0 as soon as
$$\sup_{x\in\mathbb{R}} \sqrt{n}|\widehat{F}_n(x) - F_0(x)| \geq a$$
has level
$$\mathbb{P}_{H_0}\left(\sup_{x\in\mathbb{R}} \sqrt{n}|\widehat{F}_n(x) - F_0(x)| \geq a\right) \leq 2\exp(-2a^2) = \alpha.$$
In addition to provide an easily computable threshold a, this test does not require the CDF F_0 to be continuous.

Correction of Exercise 11.A.3

1. $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = 1$: X has the same mean and variance as the standard normal distribution.

2. The law of $\sum_{i=1}^{12} U_i$ is called the Irwin–Hall distribution. We obtain the following plot for its density:



3. With the following code, we get that the power of the test is about 5% if $n = 10^4$, 40% if $n = 10^5$ and 100% if $n = 10^6$.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import kstest

# Parameters
M = 1000  # Number of repetitions per sample size
n = 1000000  # Sample size

# Function to generate samples of X = sum_{i=1}^{12} U_i - 6
def generate_sample(size):
    U = np.random.uniform(0, 1, (size, 12))
    return U.sum(axis=1) - 6

# Perform M Kolmogorov-Smirnov tests for each sample size n
rejections = 0
for _ in range(M):
```

```
    sample = generate_sample(n)
    stat, p_value = kstest(sample, 'norm')  # Null: sample ~ N(0,1)
    if p_value < 0.05:
        rejections += 1
rejection_proba = rejections / M

# Return the results
print("Rejection probability:", rejection_proba)
```

Correction of Exercise 11.A.5 In the example, $Z_i = X_{1,i} - X_{2,i} = \epsilon_{1,i} - \epsilon_{2,i}$. Under H_0', the pairs $(\epsilon_{1,i}, \epsilon_{2,i})$ and $(\epsilon_{2,i}, \epsilon_{1,i})$ have the same distribution, so that $\epsilon_{1,i} - \epsilon_{2,i}$ and $\epsilon_{2,i} - \epsilon_{1,i}$ have the same distribution, which implies that the law of Z_i is symmetric. Therefore, $H_0' \subset H_0$, so that if W_n is the rejection region for a test of level α for the null hypothesis H_0, then

$$\sup_{H_0'} \mathbb{P}(W_n) \leq \sup_{H_0} \mathbb{P}(W_n) = \alpha.$$

In other words, the test rejecting H_0' on the event W_n has level at most α.

1. Let $f : \{-1, 1\} \to \mathbb{R}$ and $g : (0, +\infty) \to \mathbb{R}$ be bounded functions. The symmetry of the law of ζ yields
$$\mathbb{E}[f(\operatorname{sign}(\zeta))g(|\zeta|)] = \mathbb{E}[f(\operatorname{sign}(-\zeta))g(|-\zeta|)] = \mathbb{E}[f(-\operatorname{sign}(\zeta))g(|\zeta|)],$$
so that
$$\begin{aligned}\mathbb{E}[f(\operatorname{sign}(\zeta))g(|\zeta|)] &= \frac{1}{2}\left(\mathbb{E}[f(\operatorname{sign}(\zeta))g(|\zeta|)] + \mathbb{E}[f(-\operatorname{sign}(\zeta))g(|\zeta|)]\right) \\ &= \mathbb{E}\left[\frac{1}{2}\underbrace{\left(f(\operatorname{sign}(\zeta)) + f(-\operatorname{sign}(\zeta))\right)}_{=f(1)+f(-1)} g(|\zeta|)\right] \\ &= \frac{f(1)+f(-1)}{2}\mathbb{E}[g(|\zeta|)],\end{aligned}$$
which shows that $\operatorname{sign}(\zeta)$ and $|\zeta|$ are independent, with $\mathbb{P}(\operatorname{sign}(\zeta) = 1) = \mathbb{P}(\operatorname{sign}(\zeta) = -1) = 1/2$. (Notice that we have used the assumption that $\mathbb{P}(\zeta = 0) = 0$, which follows from (11.4), to ensure that $f(\operatorname{sign}(\zeta)) + f(-\operatorname{sign}(\zeta)) = f(1) + f(-1)$, almost surely.)

2. Let $\epsilon_1, \ldots, \epsilon_n \in \{-1, 1\}$. We write
$$\begin{aligned}&\mathbb{P}\left(\operatorname{sign}(\zeta_{\pi(1)}) = \epsilon_1, \ldots, \operatorname{sign}(\zeta_{\pi(n)}) = \epsilon_n\right) \\ &\quad= \sum_{\sigma\in\mathfrak{S}_n} \mathbb{P}\left(\operatorname{sign}(\zeta_{\pi(1)}) = \epsilon_1, \ldots, \operatorname{sign}(\zeta_{\pi(n)}) = \epsilon_n, \pi = \sigma\right) \\ &\quad= \sum_{\sigma\in\mathfrak{S}_n} \mathbb{P}\left(\operatorname{sign}(\zeta_{\sigma(1)}) = \epsilon_1, \ldots, \operatorname{sign}(\zeta_{\sigma(n)}) = \epsilon_n, |\zeta_{\sigma(1)}| < \cdots < |\zeta_{\sigma(n)}|\right).\end{aligned}$$
By the result of Question 1, the vectors $(\operatorname{sign}(\zeta_1), \ldots, \operatorname{sign}(\zeta_n))$ and $(|\zeta_1|, \ldots, |\zeta_n|)$ are independent, therefore for any $\sigma \in \mathfrak{S}_n$,
$$\begin{aligned}&\mathbb{P}\left(\operatorname{sign}(\zeta_{\sigma(1)}) = \epsilon_1, \ldots, \operatorname{sign}(\zeta_{\sigma(n)}) = \epsilon_n, |\zeta_{\sigma(1)}| < \cdots < |\zeta_{\sigma(n)}|\right) \\ &= \mathbb{P}\left(\operatorname{sign}(\zeta_{\sigma(1)}) = \epsilon_1, \ldots, \operatorname{sign}(\zeta_{\sigma(n)}) = \epsilon_n\right)\mathbb{P}\left(|\zeta_{\sigma(1)}| < \cdots < |\zeta_{\sigma(n)}|\right) \\ &= \frac{1}{2^n}\mathbb{P}(\pi = \sigma),\end{aligned}$$

where at the last line, we have used the fact that the variables $\mathrm{sign}(\zeta_1), \dots, \mathrm{sign}(\zeta_n)$ are independent Rademacher variables. We deduce that

$$\mathbb{P}\left(\mathrm{sign}(\zeta_{\pi(1)}) = \epsilon_1, \dots, \mathrm{sign}(\zeta_{\pi(n)}) = \epsilon_n\right) = \frac{1}{2^n} \sum_{\sigma \in \mathfrak{S}_n} \mathbb{P}(\pi = \sigma) = \frac{1}{2^n},$$

which shows that $\mathrm{sign}(\zeta_{\pi(1)}), \dots, \mathrm{sign}(\zeta_{\pi(n)})$ are independent Rademacher variables.

3. According to the previous question, under H_0, T^+ has the same law as the random variable

$$\tau^+ = \sum_{k=1}^{n} k\ell_k,$$

where $\ell_1, \dots, \ell_n$ are independent $\mathcal{B}(1/2)$ variables. This variable does not depend on the law of Z_1, so that the statistic T^+ is free under H_0. Denoting by $t^+_{n,r}$ the quantile of order r of τ^+, we deduce that the test rejecting H_0 as soon as $T^+ \notin [t^+_{n,\alpha/2}, t^+_{n,1-\alpha/2}]$ has level α. These quantiles may be computed by numerical simulation.

4. Under H_0, the expectation of T^+ is equal to

$$t_n = \mathbb{E}[\tau^+] = \sum_{k=1}^{n} k\mathbb{E}[\ell_k] = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4},$$

and the variance of T^+ is equal to

$$\sigma_n^2 = \mathrm{Var}(\tau^+) = \sum_{k=1}^{n} k^2 \,\mathrm{Var}(\ell_k) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}.$$

5. We proceed as for the proof of the Central Limit Theorem, and compute the characteristic function Φ_{w_n} of the reduced variable

$$w_n = \frac{\tau^+ - t_n}{\sigma_n}.$$

For all $u \in \mathbb{R}$,

$$\begin{aligned}
\Phi_{w_n}(u) &= \mathbb{E}\left[\exp(\mathrm{i}uw_n)\right] \\
&= \mathbb{E}\left[\exp\left(\mathrm{i}u\frac{\tau^+ - t_n}{\sigma_n}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\mathrm{i}u}{\sigma_n}\sum_{k=1}^{n} k(\ell_k - 1/2)\right)\right] \\
&= \prod_{k=1}^{n} \mathbb{E}\left[\exp\left(\frac{\mathrm{i}u}{\sigma_n} k(\ell_k - 1/2)\right)\right] \\
&= \prod_{k=1}^{n} \left(\frac{1}{2}\exp\left(\frac{\mathrm{i}uk}{2\sigma_n}\right) + \frac{1}{2}\exp\left(-\frac{\mathrm{i}uk}{2\sigma_n}\right)\right).
\end{aligned}$$

The expression of σ_n^2 computed above shows that $k/\sigma_n = \mathrm{O}(1/\sqrt{n})$, uniformly over $k \le n$, so that for any $k \in \{1, \dots, n\}$,

$$\begin{aligned}
&\frac{1}{2}\exp\left(\frac{\mathrm{i}uk}{2\sigma_n}\right) + \frac{1}{2}\exp\left(-\frac{\mathrm{i}uk}{2\sigma_n}\right) \\
&= \frac{1}{2}\left(1 + \frac{\mathrm{i}uk}{2\sigma_n} - \frac{u^2k^2}{8\sigma_n^2} + \mathrm{o}\left(\frac{1}{n}\right)\right) + \frac{1}{2}\exp\left(1 - \frac{\mathrm{i}uk}{2\sigma_n} - \frac{u^2k^2}{8\sigma_n^2} + \mathrm{o}\left(\frac{1}{n}\right)\right) \\
&= 1 - \frac{u^2k^2}{8\sigma_n^2} + \mathrm{o}\left(\frac{1}{n}\right).
\end{aligned}$$

Pretending not to see that we are taking the logarithm of a complex number, we then write

$$\begin{aligned}\log \Phi_{w_n}(u) &= \sum_{k=1}^{n} \log\left(1 - \frac{u^2k^2}{8\sigma_n^2} + \mathrm{o}\left(\frac{1}{n}\right)\right)\\ &= \sum_{k=1}^{n} -\frac{u^2k^2}{8\sigma_n^2} + \mathrm{o}(1)\\ &= -\frac{u^2}{8\sigma_n^2}\frac{n(n+1)(2n+1)}{6} + \mathrm{o}(1)\\ &= -\frac{u^2}{2} + \mathrm{o}(1).\end{aligned}$$

We deduce that $\Phi_{w_n}(u)$ converges to the characteristic function $\exp(-u^2/2)$ of the $\mathcal{N}(0,1)$ distribution. This allows us to construct an asymptotic test, which rejects H_0 as soon as $|w_n| \geq \phi_{1-\alpha/2}$.

☑ Final Revision Sheet

Correction of Exercise 1

1. Under $\mathbb{P}_\mu$, $G \sim \mathcal{N}(0,1)$.
2. G depends on μ and thus it is not a statistic.
3. The parameter μ is estimated by $\overline{X}_n$, and since the alternative hypothesis writes $H_1 = \{\mu > \mu_0\}$, we take a rejection region of the form $W_n = \{\overline{X}_n \geq \mu_0 + a_n\}$. To determine the threshold a_n, for any $\mu \leq \mu_0$ we write

$$\mathbb{P}_\mu(W_n) = \mathbb{P}_\mu(\overline{X}_n \geq \mu_0 + a_n) = \mathbb{P}\left(\mu + \frac{\sigma_0}{\sqrt{n}}G \geq \mu_0 + a_n\right) = \mathbb{P}\left(G \geq \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu + a_n)\right),$$

with the notation from Question 1. The right-hand side above is an increasing function of μ, therefore

$$\sup_{\mu \leq \mu_0} \mathbb{P}_\mu(W_n) = \mathbb{P}_{\mu_0}(W_n) = \mathbb{P}\left(G \geq \frac{\sqrt{n}}{\sigma_0}a_n\right),$$

which is equal to α if and only if $\frac{\sqrt{n}}{\sigma_0}a_n = \phi_{1-\alpha}$, that is to say

$$a_n = \frac{\sigma_0}{\sqrt{n}}\phi_{1-\alpha}.$$

With this choice, the test with rejection region W_n has level α. To check that it is consistent, we observe that for any $\mu > \mu_0$,

$$\mathbb{P}_\mu(W_n) = \mathbb{P}\left(G \geq \frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu) + \phi_{1-\alpha}\right) \to 1$$

since $\frac{\sqrt{n}}{\sigma_0}(\mu_0 - \mu) \to -\infty$.

4. For $\alpha = .05$ and $\sigma_0 = \mu_0 = 10^{-2}$, we have

$$\mu_0 + \frac{\sigma_0}{\sqrt{n}}\phi_{1-\alpha} = \begin{cases} 10^{-2}\left(1 + \frac{1.65}{10}\right) = 1.165\ \% & \text{if } n = 100,\\ 10^{-2}\left(1 + \frac{1.65}{20}\right) = 1.0825\ \% & \text{if } n = 400.\end{cases}$$

Therefore, if $\overline{X}_n = 1.1\ \%$, H_0 is not rejected for $n = 100$ and the observed increase in the melting rate is concluded to be not statistically significant, while for $n = 400$, H_0 is rejected and the observed increase in the melting rate is concluded to be statistically significant.

Correction of Exercise 2 We perform a χ_2 test with the following code:

```
import numpy as np
from scipy.stats import chisquare

# Definition of distributions
theoretical_proba = np.array([9/16, 3/16, 3/16, 1/16])
experimental_distrib = np.array([100, 18, 24, 18])

# Chi-square test
chi2_stat, p_value = chisquare(f_obs=experimental_distrib, f_exp=
    theoretical_proba * np.sum(experimental_distrib))

print("Pearson's statistic:", chi2_stat)
print("p-value:", p_value)
```

We obtain $d_n \simeq 13.5$ and the p-value is approximately $3.7\ 10^{-3}$, so that the null hypothesis is rejected at all usual levels and we conclude that it is not true that A and B are dominant and a and b are recessive.

Correction of Exercise 3

1. **Frequentist estimation.**

   (a) For any $k \geq 1$,

$$\begin{aligned}
\mathbb{E}_\theta[X_1^k] &= \mathbb{E}[\mathrm{e}^{k\theta G}] \\
&= \frac{1}{\sqrt{2\pi}} \int_{z\in\mathbb{R}} \exp\left(k\theta z - \frac{z^2}{2}\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{z\in\mathbb{R}} \exp\left(-\frac{k^2\theta^2}{2} + k\theta z - \frac{z^2}{2}\right) \mathrm{d}z \exp\left(\frac{k^2\theta^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_{z\in\mathbb{R}} \exp\left(-\frac{1}{2}(k\theta - z)^2\right) \mathrm{d}z \exp\left(\frac{k^2\theta^2}{2}\right) \\
&= \exp\left(\frac{k^2\theta^2}{2}\right).
\end{aligned}$$

   (b) We deduce that $\mathbb{E}_\theta[X_1] = \mathrm{e}^{\theta^2/2}$, which yields the moment estimator

$$\widetilde{\theta}_n = \sqrt{2\log(\overline{X}_n)}.$$

   Notice that $\widetilde{\theta}_n$ is only defined if $\overline{X}_n > 1$, which by the strong law of large numbers holds true for n large enough, $\mathbb{P}_\theta$-almost surely, for any $\theta > 0$.

   (c) By the strong law of large numbers, $\overline{X}_n \to \mathrm{e}^{\theta^2/2}$, $\mathbb{P}_\theta$-almost surely, which by construction yields the strong consistency of $\widetilde{\theta}_n$. We then have, by the Central Limit Theorem,

$$\lim_{n\to+\infty} \sqrt{n}\left(\overline{X}_n - \mathbb{E}_\theta[X_1]\right) = Y \sim \mathcal{N}(0, \mathrm{Var}_\theta(X_1)), \qquad \text{in distribution,}$$

   with $\mathrm{Var}_\theta(X_1) = \mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \mathrm{e}^{2\theta^2} - \mathrm{e}^{\theta^2}$. We then deduce from the Delta method that

$$\sqrt{n}\left(\widetilde{\theta}_n - \theta\right) = \sqrt{n}\left(\phi\left(\overline{X}_n\right) - \phi\left(\mathbb{E}_\theta[X_1]\right)\right) \to \phi'\left(\mathbb{E}_\theta[X_1]\right) Y, \qquad \text{in distribution,}$$

   with $\phi(x) = \sqrt{2\log(x)}$ and $\phi'(x) = 1/\sqrt{2x^2\log(x)}$, so that

$$\phi'\left(\mathbb{E}_\theta[X_1]\right) = \frac{1}{\mathrm{e}^{\theta^2/2}\theta}.$$

We conclude that $\widetilde{\theta}_n$ is asymptotically normal with asymptotic variance equal to

$$\left[\frac{1}{\mathrm{e}^{\theta^2/2}\theta}\right]^2 \left(\mathrm{e}^{2\theta^2} - \mathrm{e}^{\theta^2}\right) = \frac{\mathrm{e}^{\theta^2}-1}{\theta^2}.$$

(d) For any measurable and bounded function $f : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_\theta[f(X_1)] &= \frac{1}{\sqrt{2\pi}} \int_{z\in\mathbb{R}} f(\mathrm{e}^{\theta z})\mathrm{e}^{-z^2/2}\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int_{x=0}^{+\infty} f(x) \exp\left(-\frac{(\log x)^2}{2\theta^2}\right) \frac{\mathrm{d}x}{\theta x},\end{aligned}$$

so that the density of X_1 under $\mathbb{P}_\theta$ writes

$$p(x;\theta) = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{(\log x)^2}{2\theta^2}\right)}{\theta x}, \qquad x > 0.$$

We deduce that the log-likelihood of a realisation $\mathbf{x}_n = (x_1, \ldots, x_n)$ writes

$$\ell_n(\mathbf{x}_n;\theta) = -\frac{n}{2}\log(2\pi) - n\log(\theta) - \sum_{i=1}^n \frac{(\log x_i)^2}{2\theta^2} - \sum_{i=1}^n \log(x_i),$$

which yields

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{x}_n;\theta) = -\frac{n}{\theta} + \sum_{i=1}^n \frac{(\log x_i)^2}{\theta^3}.$$

The right-hand side vanishes when θ takes the value

$$\theta_n(\mathbf{x}_n) = \sqrt{\frac{1}{n}\sum_{i=1}^n (\log x_i)^2}$$

and the study of the sign of $\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\mathbf{x}_n;\theta)$ shows that the likelihood reaches its maximum there. Therefore the MLE of θ is

$$\widehat{\theta}_n = \theta_n(\mathbf{X}_n) = \sqrt{\frac{1}{n}\sum_{i=1}^n (\log X_i)^2}.$$

We then remark that if one works with the sample $Y_1, \ldots, Y_n$ defined by $Y_i = \log X_i$, then the law of Y_i is $\mathcal{N}(0,\theta^2)$ and $\widehat{\theta}_n^2$ is the usual MLE of the variance in this Gaussian model.

(e) We have

$$\widehat{\theta}_n^2 = \frac{1}{n}\sum_{i=1}^n (\log X_i)^2,$$

and by construction, the variables $\log X_i$ are iid under the $\mathcal{N}(0,\theta^2)$ distribution. Therefore,

$$n\frac{\widehat{\theta}_n^2}{\theta^2} \sim \chi_2(n),$$

so the variable $\widehat{\theta}_n^2/\theta^2$ is free.

(f) We follow the approach from Subsection 7.1.3 in Lecture 7, and first consider the one-sided case $H_0 = \{\theta \leq \theta_0\}$, $H_1 = \{\theta > \theta_0\}$. The rejection region takes the form

$$W_n = \{\widehat{\theta}_n \geq \theta_0 + a_n\},$$

and for any such region, using the fact that $n\frac{\widehat{\theta}_n^2}{\theta^2} \sim \chi_2(n)$, we deduce that for all $\theta \in H_0$,

$$\begin{aligned}\mathbb{P}_\theta(W_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n \geq \theta_0 + a_n\right) \\ &= \mathbb{P}_\theta\left(n\frac{\widehat{\theta}_n^2}{\theta^2} \geq n\frac{(\theta_0+a_n)^2}{\theta^2}\right) \\ &= \mathbb{P}\left(\zeta_n \geq n\frac{(\theta_0+a_n)^2}{\theta^2}\right),\end{aligned}$$

for $\zeta_n \sim \chi_2(n)$. This is a nondecreasing function of θ, and thus we get

$$\sup_{\theta\in H_0} \mathbb{P}_\theta(W_n) = \mathbb{P}_{\theta_0}(W_n).$$

So the rejection region is

$$W_n = \{\widehat{\theta}_n \geq t_{\theta_0,n,1-\alpha}\},$$

where $t_{\theta_0,n,1-\alpha}$ is the quantile of order $1-\alpha$ of $\widehat{\theta}_n$ under $\mathbb{P}_{\theta_0}$. To compute this quantile, we use again the fact that $n\frac{\widehat{\theta}_n^2}{\theta^2} \sim \chi_2(n)$, and write for any $t > 0$,

$$\mathbb{P}_{\theta_0}(\widehat{\theta}_n \leq t) = \mathbb{P}_{\theta_0}\left(n\frac{\widehat{\theta}_n^2}{\theta^2} \leq n\frac{t^2}{\theta^2}\right) = \mathbb{P}\left(\zeta_n \leq n\frac{t^2}{\theta^2}\right),$$

so, denoting by $\chi^2_{n,1-\alpha}$ the quantile of order $1-\alpha$ of $\zeta_n \sim \chi_2(n)$, we finally get

$$n\frac{t^2_{\theta_0,n,1-\alpha}}{\theta_0^2} = \chi^2_{n,1-\alpha}$$

and

$$W_n = \left\{n\frac{\widehat{\theta}_n^2}{\theta_0^2} \geq \chi^2_{n,1-\alpha}\right\}.$$

In the two-sided case $H_0 = \{\theta = \theta_0\}$, $H_1 = \{\theta \neq \theta_0\}$, we deduce from the same arguments that the rejection region writes

$$W_n = \left\{n\frac{\widehat{\theta}_n^2}{\theta_0^2} \leq \chi^2_{n,\alpha/2} \quad \text{or} \quad n\frac{\widehat{\theta}_n^2}{\theta_0^2} \geq \chi^2_{n,1-\alpha/2}\right\}.$$

In both cases, since $\widehat{\theta}_n$ is consistent, we deduce from Proposition 7.1.16 that the tests are consistent.

2. **Nonparametric test**

(a) $H_0 = \{Q \in \{\mathcal{N}(0,\theta^2), \theta > 0\}\}$ and $H_1 = \{Q \notin \{\mathcal{N}(0,\theta^2), \theta > 0\}\}$.

(b) The null hypothesis of Gaussian goodness-of-fit tests is $H_0' = \{Q \in \{\mathcal{N}(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}\}$. Therefore H_0 is a strict subset of H_0'. As a consequence, if $Q = \mathcal{N}(\mu,\sigma^2)$ for some $\mu \neq 0$, then by the definition of W_n', $\mathbb{P}_Q(W_n') \leq \alpha$ so this quantity cannot converge to 1.

(c) For any $y \in \mathbb{R}$, $F_\theta(y) = \mathbb{P}(\theta G \leq y) = \Phi(y/\theta)$.

(d) We first rewrite

$$\zeta_n' = \sup_{y\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i\le y\}} - \Phi\left(\frac{y}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right)\right|.$$

Under H_0, there exist $\theta>0$ such that $Y_1,\dots,Y_n$ are iid under $\mathcal{N}(0,\theta^2)$ and thus the vector $(Y_1,\dots,Y_n)$ has the same law as $(F_\theta^{-1}(U_1),\dots,F_\theta^{-1}(U_n)) = (\theta\Phi^{-1}(U_1),\dots,\theta\Phi^{-1}(U_n))$, where $U_1,\dots,U_n$ are iid under $\mathcal{U}[0,1]$. We deduce that ζ_n' has the same law as the random variable

$$\begin{aligned} Z_n' &= \sup_{y\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\theta\Phi^{-1}(U_i)\le y\}} - \Phi\left(\frac{y}{\sqrt{\frac{1}{n}\sum_{i=1}^n (\theta\Phi^{-1}(U_i))^2}}\right)\right| \\ &= \sup_{y\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\Phi^{-1}(U_i)\le y/\theta\}} - \Phi\left(\frac{y/\theta}{\sqrt{\frac{1}{n}\sum_{i=1}^n \Phi^{-1}(U_i)^2}}\right)\right| \\ &= \sup_{u\in(0,1)} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{U_i\le u\}} - \Phi\left(\frac{\Phi^{-1}(u)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \Phi^{-1}(U_i)^2}}\right)\right|, \end{aligned}$$

where we have set $u=\Phi(y/\theta)$ at the last line. This shows that the law of Z_n' does not depend on θ.

(e) Denoting by $z_{n,r}'$ the quantile of order r of Z_n', the test which rejects H_0 if $\zeta_n' \ge z_{n,1-\alpha}'$ has level α.

Training for the Exam

Correction of Exercise 1

1. We have $\mathrm{Var}(\widehat{p}_n) = \frac{1}{n}p(1-p)$, and the asymptotic confidence interval is $[\widehat{p}_n \pm \phi_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}]$.

2. For each $k\in\{1,\dots,K\}$, the strong LLN ensures that $\widehat{p}_{k,n_k}$ converges almost surely to p_k, so $\widehat{p}_n^K$ is strongly consistent. The CLT then shows that

$$\lim_{n\to+\infty} \sqrt{u_k n}\,(\widehat{p}_{k,n_k} - p_k) = \mathcal{N}(0, p_k(1-p_k)), \qquad \text{in distribution.}$$

Since the variables $\widehat{p}_{1,n_1},\dots,\widehat{p}_{K,n_K}$ are independent, we deduce that the vector

$$(\sqrt{u_1 n}(\widehat{p}_{1,n_1}-p_1),\dots,\sqrt{u_K n}(\widehat{p}_{K,n_K}-p_K))$$

converges in distribution to $(Z_1,\dots,Z_K)$, where the variables Z_k are independent and with respective laws $\mathcal{N}(0,p_k(1-p_k))$. As a consequence,

$$\sqrt{n}\left(\widehat{p}_n^K - p\right) = \sum_{k=1}^K \sqrt{u_k}\sqrt{u_k n}(\widehat{p}_{k,n_k}-p_k) \to \sum_{k=1}^K \sqrt{u_k} Z_k \sim \mathcal{N}\left(0, \sum_{k=1}^K u_k p_k(1-p_k)\right),$$

in distribution. The estimator $\widehat{p}_n^K$ is therefore asymptotically normal, with asymptotic variance $\sum_{k=1}^K u_k p_k(1-p_k)$.

3. A consistent estimator of the asymptotic variance $\widehat{p}_n^K$ is given by $\sum_{k=1}^K u_k \widehat{p}_{k,n_k}(1-\widehat{p}_{k,n_k})$, we thus deduce the asymptotic confidence interval

$$\left[\widehat{p}_n^K \pm \phi_{1-\alpha/2}\sqrt{\frac{1}{n}\sum_{k=1}^K u_k \widehat{p}_{k,n_k}(1-\widehat{p}_{k,n_k})}\right].$$

4. The function $\phi : p \in [0,1] \mapsto p(1-p)$ is concave. Therefore

$$\sum_{k=1}^{K} u_k \widehat{p}_{k,n_k}(1-\widehat{p}_{k,n_k}) = \sum_{k=1}^{K} u_k \phi(\widehat{p}_{k,n_k}) \leq \phi\left(\sum_{k=1}^{K} u_k \widehat{p}_{k,n_k}\right) = \phi(\widehat{p}_n^K).$$

The confidence interval obtained by the method of quotas is therefore always smaller than the one obtained without taking the repartition of the sample into classes.

5. With these data,

$$\widehat{p}_n^K = \frac{1}{1000}\left(88 \times 0.37 + 85 \times 0.53 + 827 \times 0.9\right) \simeq 0.823.$$

The confidence interval obtained without classes writes

$$\left[\widehat{p}_n^K \pm 1.96\sqrt{\frac{\widehat{p}_n^K(1-\widehat{p}_n^K)}{n}}\right] = [0.798, 0.850],$$

and the one with classes is

$$\left[\widehat{p}_n^K \pm 1.96\sqrt{\frac{1}{n}\sum_{k=1}^{K} u_k \widehat{p}_{k,n_k}(1-\widehat{p}_{k,n_k})}\right] = [0.801, 0.843].$$

The second one is therefore more accurate.

6. In the ideal case where the members of each class all give the same answer, then each p_k is either 0 or 1. The squared length of the interval obtained with quotas is then of order of magnitude

$$\frac{1}{n}\sum_{k=1}^{K} u_k p_k(1-p_k) = 0,$$

that is to say that the estimation is exact, without uncertainty. On the contrary, if the classes are as heterogeneous as the overall population, that is to say that each p_k actually equals p, then the squared length of the interval obtained with quotas is of order of magnitude

$$\frac{1}{n}\sum_{k=1}^{K} u_k p_k(1-p_k) = \frac{1}{n}\sum_{k=1}^{K} u_k p(1-p) = \frac{1}{n}p(1-p),$$

that is to say that the estimation of p is not improved with the use of this method.

Correction of Exercise 2

1. The null hypothesis, under which X_1 and Y_1 are independent, is rejected at all usual levels: the sex of the first- and second-born children are not independent.

2. (a) The variable $\mathbb{1}_{\{X_1=\mathrm{m}\}}$ is a Bernoulli variable with parameter p_{m}^X.

   (b) In the Bernoulli model, we have seen that a consistent test, with asymptotic level α, is given by the rejection region

$$W_n = \left\{\left|\widehat{p}_{n,\mathrm{m}}^X - \frac{1}{2}\right| \geq \frac{\phi_{1-\alpha/2}}{2\sqrt{n}}\right\},$$

   where ϕ_r is the quantile of order r of the standard Gaussian distribution.

(c) With the data from Table 10.2 and $\alpha = 0.05$, we have $\phi_{1-\alpha/2} = 1.96$ and then

$$\left|\widehat{p}^X_{n,\mathrm{m}} - \frac{1}{2}\right| = 0.023, \qquad \frac{\phi_{1-\alpha/2}}{2\sqrt{n}} \simeq 0.013.$$

Therefore we reject H_0 and conclude that the probability to have a male or female first-born child is not equilibrated.

3. (a) We have

$$q_{\mathrm{m}} = \frac{\mathbb{P}(X_1 = \mathrm{m}, Y_1 = \mathrm{m})}{\mathbb{P}(X_1 = \mathrm{m})} = \frac{p_{\mathrm{m,m}}}{p^X_{\mathrm{m}}}, \qquad q_{\mathrm{f}} = \frac{\mathbb{P}(X_1 = \mathrm{f}, Y_1 = \mathrm{f})}{\mathbb{P}(X_1 = \mathrm{f})} = \frac{p_{\mathrm{f,f}}}{p^X_{\mathrm{f}}}.$$

As a consequence, by the strong Law of Large Numbers,

$$\widehat{q}_{n,\mathrm{m}} := \frac{\widehat{p}_{n,\mathrm{m,m}}}{\widehat{p}^X_{n,\mathrm{m}}} \qquad \text{and} \qquad \widehat{q}_{n,\mathrm{f}} := \frac{\widehat{p}_{n,\mathrm{f,f}}}{\widehat{p}^X_{n,\mathrm{f}}}$$

are strongly consistent estimators of q_{m} and q_{f}.

(b) With the data of Table 10.2,

$$\widehat{q}_{n,\mathrm{m}} = \frac{0.282}{0.523} \simeq 0.539, \qquad \widehat{q}_{n,\mathrm{f}} = \frac{0.238}{0.477} \simeq 0.499.$$

We observe that $\widehat{q}_{n,\mathrm{m}} > \widehat{p}^Y_{n,\mathrm{m}}$ and $\widehat{q}_{n,\mathrm{f}} > \widehat{p}^Y_{n,\mathrm{f}}$, which seems to indicate that having a first-born child of a given sex increases the probability of having a second-born child of the same sex.

(c) We have $\mathbb{E}[R_{x,1}] = \begin{pmatrix} p_{x,x} \\ p^X_x \end{pmatrix}$ and $\mathrm{Cov}[R_{x,1}] = \begin{pmatrix} p_{x,x}(1-p_{x,x}) & p_{x,x}(1-p^X_x) \\ p_{x,x}(1-p^X_x) & p^X_x(1-p^X_x) \end{pmatrix}$.

(d) We have $\dfrac{\partial\phi}{\partial r_1}\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \dfrac{1}{r_2}$ and $\dfrac{\partial\phi}{\partial r_2}\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = -\dfrac{r_1}{r_2^2}$, thus $\nabla\phi(\mathbb{E}[R_{x,1}]) = \begin{pmatrix} \dfrac{1}{p^X_x} & -\dfrac{p_{x,x}}{(p^X_x)^2} \end{pmatrix}$.

(e) The estimator $\widehat{q}_{n,x}$ writes $\widehat{q}_{n,x} = \phi\left(\frac{1}{n}\sum_{i=1}^n R_{x,i}\right)$, and it converges almost surely to $q_x = \phi(\mathbb{E}[R_{x,1}])$. Therefore, by the Delta method[1],

$$\sqrt{n}\left(\widehat{q}_{n,x} - q_x\right) = \sqrt{n}\left(\phi\left(\frac{1}{n}\sum_{i=1}^n R_{x,i}\right) - \phi(\mathbb{E}[R_{x,1}])\right) \to \nabla\phi(\mathbb{E}[R_{x,1}])Y$$

in distribution, where $Y \sim \mathcal{N}_2(0, \mathrm{Cov}[R_{x,1}])$. The limit is a one-dimensional centered Gaussian distribution, with variance

$$v_x = \nabla\phi(\mathbb{E}[R_{x,1}])\,\mathrm{Cov}[R_{x,1}])\nabla\phi(\mathbb{E}[R_{x,1}])^\top.$$

The computation of this double matrix product yields $v_x = \dfrac{p_{x,x}}{(p^X_x)^2}\left(1 - \dfrac{p_{x,x}}{p^X_x}\right)$.

(f) A consistent estimator of v_x is given by

$$\widehat{v}_{n,x} = \frac{\widehat{p}_{n,x,x}}{(\widehat{p}^X_{n,x})^2}\left(1 - \frac{\widehat{p}_{n,x,x}}{\widehat{p}^X_{n,x}}\right),$$

and thus

$$I_{n,x} = \left[\widehat{q}_{n,x} \pm \sqrt{\frac{\widehat{v}_{n,x}}{n}}\phi_{1-\alpha/2}\right]$$

is an asymptotic confidence interval with level $1-\alpha$ for q_x.

[1] You may safely assume that all parameters from Table 11.1 are positive...

With the data of Table 10.2, we obtain

$$I_{n,\mathrm{m}} \simeq [0.522, 0.557], \qquad I_{n,\mathrm{f}} \simeq [0.481, 0.517].$$

Remark. The interval $I_{n,x}$ rewrites

$$I_{n,x} = \left[\widehat{q}_{n,x} \pm \sqrt{\frac{\widehat{q}_{n,x}(1-\widehat{q}_{n,x})}{n_x^X}} \phi_{1-\alpha/2} \right], \qquad n_x^X := n\widehat{p}_{n,x}^X = \sum_{i=1}^n \mathbb{1}_{\{X_i = x\}},$$

which is what you would have obtained if you had only worked, from the beginning, with the subsample (of size n_x^X) of families with a first-born child of sex x, and had considered the estimation of q_x as the probability that the second-born child has sex x in this sample as a simple parameter estimation problem in the Bernoulli model.

Correction of Exercise 3

CDF of $X_{(k)}$ and Wilks' estimator.

1. Since $X_1, \dots, X_n$ are iid with CDF F, the variables $\mathbb{1}_{\{X_i \le x\}}$ are independent Bernoulli variables with parameter $F(x)$, and therefore $n\widehat{F}_n(x) = \sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}} \sim \mathcal{B}(n, F(x))$.

2. We recall that the pseudo-inverse functions satisfy $\widehat{F}_n^{-1}(u) \le x$ if and only if $u \le \widehat{F}_n(x)$. Therefore,

$$\mathbb{P}\left(X_{(k)} \le x\right) = \mathbb{P}\left(\widehat{F}_n^{-1}\left(\frac{k}{n}\right) \le x\right) = \mathbb{P}\left(\frac{k}{n} \le \widehat{F}_n(x)\right) = \sum_{\ell=k}^n \binom{n}{\ell} F(x)^\ell (1-F(x))^{n-\ell}.$$

3. (a) We have $\mathbb{P}(X_{(n)} > q_r) = 1 - \mathbb{P}(X_{(n)} \le q_r) = 1 - F(q_r)^n = 1 - r^n$.
   (b) From the previous question we get $\mathbb{P}(X_{(n)} > q_r) = 1 - r^n$, so this quantity is larger than $1-\alpha$ if and only if $n \ge \log\alpha / \log r$. Therefore $n_{r,\alpha} = \lceil \log\alpha / \log r \rceil$, with the definition of $\lceil \cdot \rceil$ given below.
   (c) With $r = 0.95$ and $\alpha = 0.05$, we get that $n_{r,\alpha} = 59$.

Consistency and asymptotic normality of the empirical quantile.

1. Since the vectors $(X_1, \dots, X_n)$ and $(F^{-1}(U_1), \dots, F^{-1}(U_n))$ have the same law, in particular $X_{(k)}$ has the same law as the k-th ranked element of the vector $(F^{-1}(U_1), \dots, F^{-1}(U_n))$. And since the function F^{-1} is nondecreasing, this k-th ranked element is $F^{-1}(U_{(k)})$.

2. If F^{-1} is continuous, using the first part of Proposition 11.1.6 we get that $F^{-1}(U_{(\lceil nr \rceil)})$ converges in probability to $F^{-1}(r) = q_r$. Since $F^{-1}(U_{(\lceil nr \rceil)})$ has the same law as $X_{(\lceil nr \rceil)}$, and therefore $\mathbb{P}(|F^{-1}(U_{(\lceil nr \rceil)}) - q_r| \ge \epsilon) = \mathbb{P}(|X_{(\lceil nr \rceil)} - q_r| \ge \epsilon)$ for any $\epsilon > 0$, this completes the argument.

3. By the Delta method, using Proposition 11.1.6,

$$\lim_{n \to +\infty} \sqrt{n}\left(F^{-1}(U_{(\lceil nr \rceil)}) - F^{-1}(r)\right) = (F^{-1})'(r) Y, \qquad \text{in distribution,}$$

with $Y \sim \mathcal{N}(0, r(1-r))$. This shows that

$$\lim_{n \to +\infty} \sqrt{n}\left(F^{-1}(U_{(\lceil nr \rceil)}) - F^{-1}(r)\right) = \mathcal{N}\left(0, \frac{r(1-r)}{p(q_r)^2}\right), \qquad \text{in distribution,}$$

and since $\sqrt{n}(F^{-1}(U_{(\lceil nr \rceil)}) - F^{-1}(r)) = \sqrt{n}(F^{-1}(U_{(\lceil nr \rceil)}) - q_r)$ has the same law as $\sqrt{n}(X_{(\lceil nr \rceil)} - q_r)$, we conclude that $X_{(\lceil nr \rceil)}$ is asymptotically normal with asymptotic variance $r(1-r)/p(q_r)^2$.

4. By construction, the random variable $U_{(k)}$ takes its values in $[0,1]$, and by the first part of the problem, its CDF writes

$$\forall u \in [0,1], \qquad \mathbb{P}(U_{(k)} \leq u) = \sum_{\ell=k}^{n} \binom{n}{\ell} u^\ell (1-u)^{n-\ell}.$$

We deduce that $U_{(k)}$ has density

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}u} \sum_{\ell=k}^{n} \binom{n}{\ell} u^\ell (1-u)^{n-\ell} \\
&= \sum_{\ell=k}^{n} \frac{n!}{\ell!(n-\ell)!} \left\{ \ell u^{\ell-1}(1-u)^{n-\ell} - (n-\ell) u^\ell (1-u)^{n-\ell-1} \right\} \\
&= \sum_{\ell=k}^{n} \frac{n!}{(\ell-1)!(n-\ell)!} u^{\ell-1}(1-u)^{n-\ell} - \sum_{\ell=k}^{n-1} \frac{n!}{\ell!(n-\ell-1)!} u^\ell (1-u)^{n-\ell-1} \\
&= \sum_{\ell=k}^{n} \frac{n!}{(\ell-1)!(n-\ell)!} u^{\ell-1}(1-u)^{n-\ell} - \sum_{\ell=k+1}^{n} \frac{n!}{(\ell-1)!(n-\ell)!} u^{\ell-1} (1-u)^{n-\ell} \\
&= \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k}
\end{aligned}$$

on $[0,1]$.

5. Since the sequences $(Y_n)_{n\geq 1}$ and $(Z_n)_{n\geq 1}$ are independent, the variables $k\overline{Y}_k$ and $(n-k+1)\overline{Z}_{n-k+1}$ are independent. Moreover, they are sums of independent $\mathcal{E}(1)$ variables, so they have Gamma distributions, namely $k\overline{Y}_k \sim \Gamma(k,1)$ and $(n-k+1)\overline{Z}_{n-k+1} \sim \Gamma(n-k+1,1)$. As a consequence, the pair $(k\overline{Y}_k, (n-k+1)\overline{Z}_{n-k+1})$ has density

$$q_{k,n}(y,z) = \frac{1}{\Gamma(k)} y^{k-1} \mathrm{e}^{-y} \frac{1}{\Gamma(n-k+1)} z^{n-k} \mathrm{e}^{-z}, \qquad y,z>0.$$

6. We compute the density of $V_{k,n}$. For any measurable and bounded function $f : [0,1] \to \mathbb{R}$,

$$\mathbb{E}[f(V_{k,n})] = \int_{y,z>0} f\left(\frac{y}{y+z}\right) \frac{1}{\Gamma(k)} y^{k-1} \mathrm{e}^{-y} \frac{1}{\Gamma(n-k+1)} z^{n-k} \mathrm{e}^{-z} \mathrm{d}y \mathrm{d}z.$$

Using the indication, we perform the change of variable $(u,s) = (\frac{y}{y+z}, y+z)$. Using the fact that

$$y = su, \qquad z = s(1-u),$$

we get $\mathrm{d}y\mathrm{d}z = s\mathrm{d}u\mathrm{d}s$ and therefore

$$\begin{aligned}
\mathbb{E}[f(V_{k,n})] &= \frac{1}{\Gamma(k)\Gamma(n-k+1)} \int_{u=0}^{1} \int_{s=0}^{+\infty} f(u)(su)^{k-1} \mathrm{e}^{-su} (s(1-u))^{n-k} \mathrm{e}^{-s(1-u)} s \mathrm{d}u \mathrm{d}s \\
&= \frac{1}{\Gamma(k)\Gamma(n-k+1)} \int_{u=0}^{1} f(u) u^{k-1}(1-u)^{n-k} \left(\int_{s=0}^{+\infty} s^n \mathrm{e}^{-s} \mathrm{d}s \right) \mathrm{d}u \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \int_{u=0}^{1} f(u) u^{k-1}(1-u)^{n-k} \mathrm{d}u,
\end{aligned}$$

where at the last line we have used the fact that $\int_{s=0}^{+\infty} s^n \mathrm{e}^{-s} \mathrm{d}s = \Gamma(n+1)$ which follows from the formula for the density of the $\Gamma(n+1,1)$ distribution. We conclude that $V_{k,n}$ has density

$$\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} u^{k-1}(1-u)^{n-k} = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k} = p_{k,n}(u)$$

on $[0,1]$. Here we have used the fact that for any integer $m \geq 1$, $\Gamma(m) = (m-1)!$, but you can conclude simply using the fact that the density of $V_{k,n}$ is proportional to $p_{k,n}(u)$, and therefore these two densities necessarily coincide.

7. By the strong Law of Large Numbers, $\overline{Y}_{\lceil nr\rceil} \to 1$ and $\overline{Z}_{n-\lceil nr\rceil+1} \to 1$, almost surely. On the other hand, it is straightforward that $\frac{\lceil nr\rceil}{n} \to r$ and $\frac{n-\lceil nr\rceil+1}{n} \to 1-r$. This yields the first statement. Now the Central Limit Theorem yields

$$\lim_{n\to+\infty} \sqrt{\lceil nr\rceil}\left(\overline{Y}_{\lceil nr\rceil} - 1\right) = \mathcal{N}(0,1), \quad \lim_{n\to+\infty} \sqrt{n-\lceil nr\rceil+1}\left(\overline{Z}_{n-\lceil nr\rceil+1} - 1\right) = \mathcal{N}(0,1),$$

in distribution. Rewriting

$$\sqrt{n}(\xi_n - r) = \sqrt{\frac{\lceil nr\rceil}{n}} \cdot \sqrt{\lceil nr\rceil}\left(\overline{Y}_{\lceil nr\rceil} - 1\right) + \frac{\lceil nr\rceil - nr}{\sqrt{n}},$$

and using the fact that $\sqrt{\frac{\lceil nr\rceil}{n}} \to \sqrt{r}$, $\frac{\lceil nr\rceil - nr}{\sqrt{n}} \to 0$ (because $0 \le \lceil nr\rceil - nr \le 1$), we deduce from Slutsky's Lemma that

$$\lim_{n\to+\infty} \sqrt{n}(\xi_n - r) = \mathcal{N}(0,r), \qquad \text{in distribution.}$$

By the same arguments,

$$\lim_{n\to+\infty} \sqrt{n}(\zeta_n - (1-r)) = \mathcal{N}(0,1-r), \qquad \text{in distribution,}$$

and the conclusion follows from the fact that ξ_n and ζ_n are independent.

8. We first note that

$$V_{\lceil nr\rceil,n} = \phi\begin{pmatrix}\xi_n\\ \zeta_n\end{pmatrix}, \qquad \phi\begin{pmatrix}y\\ z\end{pmatrix} := \frac{y}{y+z}.$$

We deduce from the continuity of ϕ that

$$\lim_{n\to+\infty} V_{\lceil nr\rceil,n} = \phi\begin{pmatrix}r\\ 1-r\end{pmatrix} = r, \qquad \text{almost surely.}$$

Since $U_{(\lceil nr\rceil)}$ has the same law as $V_{\lceil nr\rceil,n}$, we deduce that for any $\epsilon > 0$,

$$\mathbb{P}(|U_{(\lceil nr\rceil)} - r| \ge \epsilon) = \mathbb{P}(|V_{\lceil nr\rceil,n} - r| \ge \epsilon) \to 0,$$

by the Dominated Convergence Theorem. This shows the consistency of $U_{(\lceil nr\rceil)}$.

We now use the Delta method to write

$$\lim_{n\to+\infty} \sqrt{n}\left(V_{\lceil nr\rceil,n} - r\right) = \lim_{n\to+\infty} \sqrt{n}\left(\phi\begin{pmatrix}\xi_n\\ \zeta_n\end{pmatrix} - \phi\begin{pmatrix}r\\ 1-r\end{pmatrix}\right) = \nabla\phi\begin{pmatrix}r\\ 1-r\end{pmatrix} Y,$$

in distribution, with $Y \sim \mathcal{N}_2\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}r & 0\\ 0 & 1-r\end{pmatrix}\right)$ and $\nabla\phi\begin{pmatrix}r\\ 1-r\end{pmatrix} = \begin{pmatrix}1-r & -r\end{pmatrix}$. We deduce that $\sqrt{n}\left(V_{\lceil nr\rceil,n} - r\right)$ converges to the centered Gaussian distribution with variance

$$\begin{pmatrix}1-r & -r\end{pmatrix}\begin{pmatrix}r & 0\\ 0 & 1-r\end{pmatrix}\begin{pmatrix}1-r\\ -r\end{pmatrix} = r(1-r).$$

Since $U_{(\lceil nr\rceil)}$ has the same law as $V_{\lceil nr\rceil,n}$, this completes the proof of Proposition 11.1.6.

Bibliography

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009.

[2] B. Jourdain. *Probabilités et statistiques*. Ellipses, 2016.

[3] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[4] A. Reinhart. *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press, 2015.

[5] C. P. Robert et G. Casella. *Monte Carlo Statistical Methods*. Seconde édition. Springer Texts in Stistics, Springer, 2004.

[6] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[7] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.