

# **Statistiques Fondamentales**

Simon Coste

2024-01-22

# Table des matières

<b>Organisation</b>	<b>3</b>
Utiliser ce site . . . . .	3
<b>1 Introduction</b>	<b>4</b>
1.1 Un exemple pour fixer les idées . . . . .	4
1.2 Qu'est-ce qu'un problème statistique ? . . . . .	5
1.3 Qu'est-ce qu'un estimateur ? . . . . .	6
1.4 Points de vue . . . . .	7
<b>2 Estimation de paramètre</b>	<b>8</b>
2.1 Précision d'un estimateur . . . . .	8
2.2 Convergence . . . . .	8
2.3 Normalité asymptotique . . . . .	9
2.4 Trois outils sur la normalité asymptotique . . . . .	9
2.5 Deux estimateurs importants . . . . .	10
<b>3 La méthode des moments</b>	<b>12</b>
3.1 Qu'est-ce qu'un <i>moment</i> ? . . . . .	12
3.2 Estimateur des moments . . . . .	12
<b>4 Exercices</b>	<b>14</b>
4.1 Questions . . . . .	14
4.2 Exercices . . . . .	14
<b>5 Intervalles de confiance</b>	<b>16</b>
5.1 Principe . . . . .	16
5.2 Exemples gaussiens . . . . .	16
5.2.1 Estimation de $\mu$ . . . . .	16
5.2.2 Estimation de $\sigma$ . . . . .	18
5.3 Exemples asymptotiques . . . . .	19
5.3.1 Estimation du paramètre $p$ dans un modèle de Bernoulli. . . . .	19
5.3.2 Estimation de moyenne dans un modèle non-gaussien. . . . .	20
<b>6 Outils pour les IC</b>	<b>21</b>
6.1 Quantiles . . . . .	21
6.2 Calculs de lois . . . . .	22
6.2.1 Lois Gamma . . . . .	22
6.2.2 Loi du chi-deux . . . . .	22
6.2.3 Loi de Student . . . . .	23
6.2.4 Loi de la statistique de Student . . . . .	23
6.3 Inégalités de concentration . . . . .	24

6.4	Inégalité de Bienaymé-Tchebychev . . . . .	24
6.5	Inégalité de Hoeffding . . . . .	25
<b>7</b>	<b>Exercices</b>	<b>27</b>
7.1	Questions . . . . .	27
7.2	Exercices . . . . .	27
<b>8</b>	<b>Test d'hypothèses</b>	<b>29</b>
8.1	Exemples de tests gaussiens . . . . .	30
8.1.1	Construction du test . . . . .	30
8.1.2	Calcul de la puissance et hypothèse alternative . . . . .	31
8.2	La notion de $p$ -valeur . . . . .	31
<b>9</b>	<b>Théorie des tests simples</b>	<b>33</b>
9.1	La distance en variation totale . . . . .	33
9.2	Test optimal au sens de l'affinité . . . . .	34
9.3	Théorème de Neyman-Pearson . . . . .	35
9.4	Bornes sur la variation totale . . . . .	36
<b>10</b>	<b>Tests du <math>\chi_2</math></b>	<b>38</b>
10.1	Loi multinomiale . . . . .	38
10.2	Test d'adéquation . . . . .	39
10.3	Test d'indépendance . . . . .	40
<b>11</b>	<b>Exercices</b>	<b>42</b>
11.1	Questions . . . . .	42
11.2	Tests élémentaires . . . . .	42
11.3	Exercices . . . . .	42
<b>12</b>	<b>Moindres carrés</b>	<b>45</b>
12.1	Ajustement affine en une dimension. . . . .	45
12.2	Moindres carrés généraux . . . . .	46
<b>13</b>	<b>Modèles linéaires</b>	<b>48</b>
13.1	Modèle gaussien . . . . .	48
13.2	Modèle linéaire général . . . . .	49
<b>14</b>	<b>Outils gaussiens</b>	<b>50</b>
14.1	Vecteurs gaussiens . . . . .	50
14.2	Conditionnement gaussien . . . . .	50
14.3	Théorème de Cochran . . . . .	51
14.4	Loi de Fisher . . . . .	52
<b>15</b>	<b>Tests linéaires</b>	<b>53</b>
15.1	Significativité d'un coefficient . . . . .	53
15.2	Significativité de la régression . . . . .	53
15.3	Test d'une contrainte . . . . .	53
<b>16</b>	<b>Exercices</b>	<b>54</b>
16.1	Questions . . . . .	54

<b>17 Modèles exponentiels</b>	<b>55</b>
<b>18 Maximum de vraisemblance</b>	<b>56</b>
<b>19 Information et entropie</b>	<b>57</b>
<b>20 Estimation de densité</b>	<b>58</b>
<b>Et après ?</b>	<b>59</b>
<b>Références</b>	<b>60</b>

# Organisation

Bienvenue sur la page du cours de Statistiques Fondamentales (STF8) du master Mathématiques Fondamentales et Appliquées de l'Université Paris-Cité. Les notes de cours sont accessibles à tous. De nombreux auteurs s'y sont succédés ; je suis le dernier en date, mais les versions précédentes ont été travaillées par Clément Levrard, Stéphane Boucheron, Stéphane Gaïffas, Pierre Youssef.

- Les CM ont lieu les jeudi à (8h30 - 10h30), et les vendredi (10h45 - 12h45) **sauf le premier cours qui a lieu lundi 8 janvier à 10h45-12h45.**
- Les TD ont lieu lundi (13h45 - 16h45) et vendredi (13h30 - 15h30), de lundi 8 janvier à vendredi 16 février.
- Il y aura deux contrôles de 2h, le vendredi 26 janvier et lundi 12 février.
- L'examen a lieu le 1er mars de 13h30 à 16h30.
- Il y aura une interro de 5 minutes chaque semaine le jeudi.

## Utiliser ce site

Chaque chapitre de ce livre contient une page dédiée au cours théorique, et contiendra dans un futur proche une page d'exercices.

La saveur du cours est essentiellement mathématique et nous n'aurons pas de TP d'info ; cependant, je vous recommande vraiment d'essayer d'appliquer tout ça via votre langage de programmation favori, c'est-à-dire ~~Python R SAS C++ Julia~~ Python R SAS C++ Julia. J'essaierai autant que possible de fournir des mini-jeux de données avec des petits challenges pour appliquer ce que vous apprenez en cours.

Ces notes sont mises en lignes et totalement accessibles via [Quarto](#). Si vous savez comment utiliser `git`, n'hésitez pas à corriger toutes les erreurs que vous pourriez voir (et Dieu sait qu'elles seront nombreuses) via des pull requests.

# 1 Introduction

Les outils des statistiques furent créés pour analyser des phénomènes quantitatifs dans lesquels la présence de *bruit* ou de *hasard* rendait l'analyse classique moins opérante. Il peut typiquement s'agir de problèmes dans lesquels de nombreuses données indépendantes ont été générées par le même phénomène. Dans cette section, nous allons développer un exemple pour bien comprendre les questions qui se posent et la façon de les résoudre, puis nous poserons quelques bases qui nous permettront d'utiliser le langage des mathématiques.

## 1.1 Un exemple pour fixer les idées

Une grande enseigne de distribution possède  $n = 100$  magasins identiques, qui génèrent chaque année un chiffre d'affaire annuel (CA, en millions d'euros). Ce chiffre oscille autour d'une valeur de référence  $\mu$ . Cette valeur n'est pas observée ; ce qui est observé, ce sont tous les chiffres d'affaires des  $n$  magasins, qui fluctuent tous autour de la vraie valeur  $\mu$ . Ces fluctuations proviennent de nombreuses sources : erreurs comptables, perturbations des ventes dues aux fournisseurs ou aux prix, etc. Ce qu'on observe, c'est donc des chiffres  $x_1, \dots, x_n$  qui ne sont pas tous égaux ; comment avoir une idée de la véritable valeur de  $\mu$  ?

**Estimation.** Évidemment, la moyenne empirique

$$\bar{x}_n = \frac{x_1 + \dots + x_n}{n}$$

vient naturellement à l'esprit. En faisant le calcul, on trouve  $\bar{x}_n \approx 21,6$ . Cette valeur est une *estimation* du CA moyen  $\mu$ . Ce chiffre peut être utilisé par l'enseigne, par exemple pour jauger la rentabilité d'un possible plan d'ouverture de nouveaux magasins.

**Précision.** On pourrait se demander à quel point cette estimation est précise ou, disons, essayer de quantifier l'erreur possible qu'on fait si l'on dit que  $\mu$  est égal à 21,6 millions d'euros. Cela nécessite de faire quelques hypothèses sur le hasard qui génère les fluctuations des  $x_i$  autour de  $\mu$ . Ces fluctuations observées au cours de l'année proviennent de l'agrégation de toutes les fluctuations quotidiennes, lesquelles sont à peu près indépendantes, et pour cette raison on peut supposer (pour commencer) que ces fluctuations sont gaussiennes et ont à peu près la même variance, disons  $\sigma^2 = 1$ . Comme on a supposé que les  $x_i$  sont des réalisations d'une loi gaussienne  $N(\mu, 1)$ , alors on sait que  $\bar{x}_n$  est la réalisation d'une loi  $N(\mu, 1/n)$ , ou encore que  $\bar{x}_n - \mu$  est la réalisation d'une gaussienne centrée de variance  $1/n$ . Les lois gaussiennes sont bien connues ; par exemple, avec probabilité supérieure à 99%, une gaussienne  $N(0, \sigma^2)$  est comprise entre les valeurs  $-2,96\sigma$  et  $2,96\sigma$ . Autrement dit, il y a 99% de chances pour que le nombre  $|\bar{x} - \mu|$ , qui représente l'erreur d'estimation, soit plus petite que  $2,96/\sqrt{n} = 2,96/10 \approx 0,3$ .

Ce dernier raisonnement peut être vu d'une autre façon. Dire que  $\bar{x}_n$  et  $\mu$  ne diffèrent pas de plus de 0,3, c'est équivalent à dire que  $\mu$  appartient à l'intervalle  $[\bar{x} - 0,3, \bar{x} + 0,3]$ . En d'autres termes, avec une probabilité supérieure à 99%, le vrai CA  $\mu$  de chaque magasin se situe entre 21,3 et 21,9. Cela laisse tout de même une chance de 1% que le paramètre  $\mu$  ne soit pas dans cette région.

**Tests.** Il existe encore un autre point de vue sur ce problème. Par exemple, le conseil d'administration de la firme veut s'assurer que le dirigeant a bien tenu sa promesse selon laquelle le CA de chaque magasin était supérieur à 21 millions d'euros. La valeur exacte de  $\mu$  n'est pas le plus important : ce qui nous intéresse maintenant, c'est plutôt d'être sûrs que  $\mu$  n'est pas inférieur au seuil de 21. Le dirigeant, fin statisticien, effectue alors un raisonnement par l'absurde *en probabilité*. Supposons que le CA  $\mu$  soit effectivement égal à 21 (ou même, inférieur). Alors, par les mêmes calculs que ci-dessus, cela voudrait dire qu'avec 99% de chances,  $\bar{x}_n$  et 21 ne devraient pas différer de plus de 0,3 ; autrement dit, que  $\bar{x}_n$  devrait se situer entre 20,7 et 21,3. Ce n'est pas le cas, puisque  $\bar{x}_n = 21,6$ . Si  $\mu$  est réellement plus petit que 21, alors ce qu'on a observé est extrêmement peu probable. Par contraposée probabiliste, il est raisonnable de rejeter l'hypothèse selon laquelle  $\mu$  est inférieur à 21.

---

Les trois points de vue donnés ci-dessus sont en quelque sorte les piliers de l'analyse statistique. L'estimation consiste à deviner une valeur cachée dans du bruit ; les intervalles de confiance consistent à donner une région dans laquelle se trouve cette valeur ; les tests d'hypothèse permettent de raisonner de façon logique sur cette valeur.

L'objectif du cours de statistiques de quantifier l'incertitude liée au hasard dans chacun de ces objectifs. Comme dans les exemples donnés ci-dessus, c'est un ensemble de méthodes scientifiques qui s'appuient sur la théorie des probabilités ; dans ce cours, on fera des *hypothèses* sur le hasard qui est en jeu, et on en tirera des conséquences *probables* sur le modèle sous-jacent. En théorie des probabilités, le jeu est plutôt inverse : partant d'un modèle probabiliste fixé, on essaie de déterminer quel sera le comportement des réalisations de ce modèle. Il semble difficile de faire l'un sans l'autre.

## 1.2 Qu'est-ce qu'un problème statistique ?

Il n'y aurait pas de statistiques s'il n'y avait pas de monde réel, et comme chacun sait, le monde réel est principalement composé de quantités aléatoires.

Un problème statistique tire donc toujours sa source d'un ensemble d'observations, disons  $n$  observations notées  $x_1, \dots, x_n$  ; cet ensemble d'observations est appelé un *échantillon*. L'hypothèse de base de tout travail statistique consiste à supposer que cet échantillon suit une certaine loi de probabilité ; l'objectif est de trouver laquelle. Évidemment, on ne va pas partir de rien : il faut bien faire des hypothèses minimales sur cette loi. Ce qu'on appelle un *modèle statistique* est le choix d'une famille de lois de probabilités que l'on suppose pertinentes.

**Définition 1.1.** Formellement, choisir un modèle statistique revient à choisir trois choses :

- $\mathcal{X}$ , l'espace dans lequel vit notre échantillon ;
- $\mathcal{F}$ , une tribu sur  $\mathcal{X}$ , pour donner du sens à ce qui est observable ou non ;
- $(P_\theta)_{\theta \in \Theta}$ , une famille de mesures de probabilités sur  $\mathcal{X}$  indexée par  $\theta \in \Theta$ , où  $\Theta$  est appelé espace des paramètres. On écrira fréquemment  $\mathbb{E}_\theta$  ou  $\text{Var}_\theta$  pour désigner des espérances, variances, etc., calculées avec la loi  $P_\theta$ .

En pratique, dans ce cours, on aura toujours un échantillon  $(x_1, \dots, x_n)$  où les  $x_i$  vivent dans un même espace, disons  $\mathbb{R}^d$  pour simplifier. On devrait donc écrire  $\mathcal{X} = \mathbb{R}^{d \times n}$  ; et l'on fera toujours l'hypothèse que ces observations sont indépendantes les unes des autres, et que ces observations ont la même loi de probabilité. Autrement dit, on se donnera toujours une mesure  $p_\theta$  sur  $\mathbb{R}^d$  et on supposera que la loi de notre échantillon est  $P_\theta = p_\theta^{\otimes n}$ . Dans ce cadre, les observations  $x_i$  sont des *réalisations* de variables aléatoires  $X_i$  iid de loi  $p_\theta$ .

Il faut prendre garde à distinguer les variables aléatoires  $X_i$ , qui sont des objets théoriques, de leurs réalisations  $x_i$ , qui, elles, sont bel et bien observées.

**Définition 1.2.** On dit qu'un modèle statistique est identifiable si  $\theta \neq \theta'$  entraîne  $P_\theta \neq P_{\theta'}$ .

Si l'on a bien choisi notre modèle statistique, alors il existe un « vrai » paramètre, disons  $\theta_*$ , tel que les observations  $x_1, \dots, x_n$  sont des réalisations de loi  $p_{\theta_*}$ . L'objectif est alors de trouver  $\theta_*$  ou quelque information que ce soit le concernant.

Dans un modèle identifiable, la statistique inférentielle (classique) permet de faire trois choses :

- Trouver une valeur approchée du vrai paramètre  $\theta_*$  (estimation ponctuelle).
- Donner une zone de  $\Theta$  dans laquelle le vrai paramètre  $\theta_*$  a des chances de se trouver (intervalle de confiance).
- Répondre à des questions binaires sur  $\theta_*$ , par exemple «  $\theta_*$  est-il positif ? ».

### 1.3 Qu'est-ce qu'un estimateur ?

**Définition 1.3.** Une *statistique* est une fonction mesurable des observations. Plus formellement, si le modèle statistique fixé est  $(\mathcal{X}, \mathcal{F}, P)$ , alors une statistique est n'importe quelle fonction mesurable de  $(\mathcal{X}, \mathcal{F})$ .

- 1) Le premier point important est qu'une statistique ne peut pas prendre  $\theta$  en argument. Ses valeurs ne doivent dépendre du paramètre  $\theta$  qu'au travers de  $P_\theta$ .
- 2) Le second point important est que, si  $X$  est une variable aléatoire et  $T$  une statistique, alors  $T(X)$  est une variable aléatoire. On peut donc définir des quantités théoriques liées à  $T$ : typiquement, si  $X$  a pour loi  $P_\theta$ , on peut définir la valeur moyenne de  $T$  sous le modèle  $P_\theta$  comme

$$\mathbb{E}_\theta[T(X)] = \int_{\mathcal{X}} T(x) P_\theta(dx)$$

ou encore sa variance  $\mathbb{E}_\theta[T(X)^2] - (\mathbb{E}_\theta[T(X)])^2$ , etc. On peut aussi calculer la valeur de cette statistique sur l'échantillon dont on dispose, c'est-à-dire  $T(x_1, \dots, x_n)$ . Par exemple, la moyenne empirique d'un  $n$ -échantillon réel est la fonction  $T : (a_1, \dots, a_n) \rightarrow n^{-1}(a_1 + \dots + a_n)$ . Si les  $x_i$  sont des réalisations des variables aléatoires  $X_i$ , alors  $T(x_1, \dots, x_n)$  est une réalisation de la variable aléatoire  $T(X_1, \dots, X_n)$ .

- 3) Ce qui ne se voit pas dans la définition, c'est qu'une bonne statistique devrait être facilement calculable ; à la place de *statistique*, on peut penser à *algorithme* : une bonne statistique doit pouvoir être calculée facilement par un algorithme ne prenant en entrée que les échantillons  $x_i$ .



Si le but est de deviner la valeur de  $\theta$  à partir des observations, il est naturel de considérer des statistiques à valeurs dans  $\Theta$ . C'est précisément la définition d'un estimateur.

**Définition 1.4.** Dans le modèle  $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$ , un estimateur de  $\theta$  est une statistique à valeurs dans  $\Theta$ .

En fait, on n'est pas obligés de vouloir estimer précisément  $\theta$ . Peut-être qu'on veut estimer quelque chose qui dépend de  $\theta$ , mais qui n'est pas  $\theta$  ; disons, une fonction  $\varphi(\theta)$ . Dans ce cas, un estimateur de  $\varphi(\theta)$  sera simplement une statistique à valeurs dans l'espace où vit  $\varphi(\theta)$ .

## 1.4 Points de vue

**Inférence paramétrique.** La plupart des expériences/modèles statistiques que nous rencontrerons dans ce cours, seront de nature dite *paramétrique*, autrement dit indexés par des parties de  $\mathbb{R}^d$ . Le mot “paramètre” est en lui-même trompeur : on parle souvent de paramètre d'une distribution pour désigner ce qui devrait plutôt s'appeler une fonctionnelle. Par exemple, la moyenne, la covariance d'une distribution sur  $\mathbb{R}^d$  sont des paramètres de cette distribution. Les quantiles, l'asymétrie, la kurtosis sont d'autres paramètres.

**Statistique non paramétrique.** Tous les modèles ne sont pas paramétriques au sens ci-dessus : dans de nombreux développements des statistiques, par exemple en estimation de densité, on travaille sur des modèles plus riches qui n'admettent pas de paramétrisation naturelle par une partie d'un espace euclidien de dimension finie. C'est ce qu'on appelle l' *estimation non-paramétrique*. Nous y reviendrons au dernier chapitre.

**Statistique bayésienne.** En statistique paramétrique, les paramètres  $\theta$  déterminent le hasard qui génère les observations  $x_i$ . La statistique bayésienne consiste à renverser le point de vue, et à rendre le paramètre  $\theta$  lui-même aléatoire ; sa loi, appelée *prior*, mesure “le degré de connaissance a priori” qu'on en a. La règle de Bayes explique comment cette loi est modifiée par les observations. C'est un point de vue qui ne sera pas abordé dans ce cours.

## 2 Estimation de paramètre

On fixe un modèle statistique  $(\mathcal{X}, \mathcal{F}, (P_\theta))$ , et l'on cherche à estimer le paramètre  $\theta$  ou un autre paramètre qui dépend de  $\theta$ . Dans ce chapitre, on explique comment juger la qualité d'un estimateur, et l'on donne une technique générale pour construire de bons estimateurs dans des situations assez naturelles.

### 2.1 Précision d'un estimateur

**Définition 2.1** (Biais, risque quadratique).

- Le biais de  $\hat{\theta}$  est la quantité  $\mathbb{E}_\theta[\hat{\theta} - \theta]$ . L'estimateur est dit *sans biais* s'il est de biais nul.
- Le risque quadratique de  $\hat{\theta}$  est la quantité  $\mathbb{E}_\theta[|\hat{\theta} - \theta|^2]$ .

En pratique, on peut vouloir estimer non pas  $\theta$  lui-même, mais un paramètre  $\psi = \psi_\theta$  qui dépend de  $\theta$ , comme  $\cos(\theta)$  ou  $|\theta|$  par exemple. Dans ce cas, si  $\hat{\psi}$  est un estimateur de  $\psi$  alors le biais est défini par  $\mathbb{E}_\theta[\hat{\psi} - \psi_\theta]$  et le risque quadratique par  $\mathbb{E}_\theta[|\hat{\psi} - \psi_\theta|^2]$ .

La dépendance du risque quadratique vis à vis de la taille de l'échantillon est une question importante : pour une suite d'expériences donnée, quelles sont les meilleures vitesses envisageables, et comment les obtenir ?

**Théorème 2.1** (Décomposition biais-variance).

$$\mathbb{E}_\theta[|\hat{\theta} - \theta|^2] = \underbrace{\text{Var}_\theta(\hat{\theta})}_{\text{variance}} + \underbrace{\mathbb{E}_\theta[\hat{\theta} - \theta]^2}_{\text{carré du biais}}.$$

### 2.2 Convergence

Rappelons brièvement deux notions de convergence des variables aléatoires. Une suite de variables aléatoires  $X_n$  à valeurs dans  $\mathbb{R}^d$  converge en probabilité vers une variable aléatoire  $X$  à valeurs dans  $\mathbb{R}^d$  si pour tout  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

**Définition 2.2** (consistance d'un estimateur). Une suite d'estimateurs  $(\hat{\theta}_n)$  est convergente pour l'estimation de  $\theta$  lorsque, pour tout  $\theta \in \Theta$ , sous  $P_\theta$ , la suite  $(\hat{\theta}_n)$  converge en probabilité vers  $\theta$  ; autrement dit, lorsque

$$\forall \varepsilon > 0, \quad \lim_n P_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

La suite est fortement convergente si, pour tout  $\theta$ , la convergence a lieu  $P_\theta$ -presque sûrement. \$\$

On voit parfois le mot *consistent* utilisé au lieu de *convergent*. Je pense que c'est un anglicisme.

## 2.3 Normalité asymptotique

Lorsqu'un estimateur est convergent, on peut se demander à quoi ressemblent ses fluctuations autour de sa valeur limite. Le théorème central limite indique que le comportement asymptotique gaussien est relativement fréquent, et beaucoup d'estimateurs sont des sommes de réalisations de variables indépendantes.

**Définition 2.3** (normalité asymptotique). Soit  $\theta$  un paramètre à estimer, et  $\hat{\theta}_n$  une suite d'estimateurs de  $\theta$ . On dit que ces estimateurs sont *asymptotiquement gaussiens* (ou *normaux*) si, après les avoir renormalisés convenablement, ils convergent en loi vers une loi gaussienne. Autrement dit, s'il existe une suite  $a_n$  de nombres réels tels que

$$a_n(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{\text{loi}} N(0, \Sigma)$$

où  $\Sigma$  est une matrice de covariance qui dépend peut-être de  $\theta$  — pour éviter les cas dégénérés, on demande à ce que  $\Sigma$  soit non-nulle.

## 2.4 Trois outils sur la normalité asymptotique

La normalité asymptotique n'est pas intéressante en elle-même ; l'idée est plutôt de chercher le comportement asymptotique de la statistique recentrée pour pouvoir en déduire des garanties en terme de risque asymptotique ou d'intervalle de confiance. Nous utiliserons cela de nombreuses fois dans la suite ; la normalité asymptotique sera par exemple utilisée dans la construction des intervalles de confiance. Aussi, prouver que des estimateurs sont asymptotiquement normaux est une tâche importante, qui est grandement simplifiée par les deux outils suivants. On commence par rappeler le théorème centra-limite.

**Théorème 2.2** (Théorème Central-Limite). Soit  $(X_i)$  une suite de variables aléatoires réelles, indépendantes et identiquement distribuées. On suppose que ces variables ont une variance  $\sigma^2$  finie. Alors, la variable aléatoire

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] \right)$$

converge en loi vers une loi  $N(0, \sigma^2)$ .

Le Lemme de Slutsky sera fréquemment utilisé pour combiner convergence en loi et convergence en probabilité.

**Théorème 2.3** (Lemme de Slutsky). Soit  $(X_n)$  une suite de variables aléatoire qui converge en loi vers  $X$  et  $(Y_n)$  une suite de variables aléatoires qui converge en probabilité (ou en loi) vers une constante  $c$ . Alors, le couple  $(X_n, Y_n)$  converge en loi vers  $(X, c)$  ; autrement dit, pour toute fonction continue bornée  $\varphi$ ,

$$\mathbb{E}[\varphi(X_n, Y_n)] \rightarrow \mathbb{E}[\varphi(X, c)].$$

*Démonstration.* Fixons une fonction test  $\varphi$  continue à support compact, donc bornée par un certain  $M$ . Il faut montrer que  $\mathbb{E}[\varphi(X_n, Y_n) - \varphi(X, c)]$  tend vers zéro. L'intégrande est égal à la somme de  $A = \varphi(X_n, Y_n) - \varphi(X_n, c)$  et de  $B = \varphi(X_n, c) - \varphi(X, c)$ .

Comme  $X_n$  tend en loi vers  $X$  et que  $t \rightarrow \varphi(t, c)$  est continue bornée, l'espérance de  $B$  tend vers zéro. Il faut donc montrer que l'espérance de  $A$  tend vers zéro. On fixe un  $\varepsilon > 0$ .

- Par le [théorème de Heine](#),  $\varphi$  est uniformément continue : il existe  $\delta > 0$  tel que  $|(x, y) - (x', y')| < \delta$  entraîne que  $|\varphi(x, y) - \varphi(x', y')| < \varepsilon/2$ .
- On introduit l'événement  $\{|Y_n - c| \leq \delta\}$ . Par le point précédent, sur cet événement on a  $|A| < \varepsilon/2$ . Hors de cet événement, on peut toujours borner  $|A|$  par  $2M$ . On a donc

$$|\mathbb{E}A| \leq \mathbb{P}(|Y_n - c| \leq \delta)\varepsilon/2 + \mathbb{P}(|Y_n - c| > \delta)2M.$$

- Comme  $Y_n$  converge en probabilité vers  $c$ , lorsque  $n$  est assez grand on a  $\mathbb{P}(|Y_n - c| > \delta) < \varepsilon/4M$ .
- En regroupant tout ce qui a été dit, on obtient bien  $|\mathbb{E}A| \leq \varepsilon$  dès que  $n$  est assez grand, ce qui montre bien que  $\mathbb{E}A \rightarrow 0$ .

□

**Théorème 2.4** (Delta-méthode). *Soit  $(X_n)$  une suite de variables aléatoires réelles telle que  $\sqrt{n}(X_n - \alpha)$  converge en loi vers  $N(0, \sigma^2)$ . Pour toute fonction  $g : \mathbb{R} \rightarrow \mathbb{R}$  dérivable en  $\alpha$  (de dérivée non nulle en  $\alpha$ ), on a*

$$\sqrt{n}(g(X_n) - g(\alpha)) \xrightarrow[n \rightarrow \infty]{loi} N(0, g'(\alpha)^2 \sigma^2).$$

Plus généralement, si les  $X_n$  sont à valeurs dans  $\mathbb{R}^d$  et que  $\sqrt{n}(X_n - \alpha) \rightarrow N(0, \Sigma)$ , alors pour toute application  $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ , la suite  $\sqrt{n}(g(X_n) - g(\alpha))$  converge en loi vers

$$N(0, Dg(\alpha)\Sigma Dg(\alpha)^\top)$$

où  $Dg(x)$  est la [matrice jacobienne](#) de  $g$  en  $x$ .

*Démonstration.* À écrire.

□

## 2.5 Deux estimateurs importants

Deux estimateurs sont omniprésents en statistique : la moyenne empirique et la variance empirique. Ils sont pertinents dans n'importe quel modèle où les observations sont des réalisations de variables iid possédant une moyenne  $\mu$  et une variance  $\sigma^2$ .

La moyenne empirique est définie par

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

Il est évident que  $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X] = \mu$ . Cet estimateur est donc toujours sans biais, et son risque quadratique est égal à sa variance, c'est-à-dire  $\frac{\sigma^2}{n}$ .

L'estimateur de la variance empirique est défini comme

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

**Théorème 2.5.** *L'estimateur  $\hat{\sigma}_n^2$  est sans biais.*

*Démonstration.* À écrire.



## 3 La méthode des moments

Il existe plusieurs techniques générales pour *construire* des estimateurs. Deux se démarquent : la méthode des moments, et la méthode du maximum de vraisemblance. La méthode des moments est naturelle et donne beaucoup des estimateurs avec de bonnes propriétés, mais elle est moins puissante que la méthode du maximum de vraisemblance que nous verrons plus tard.

### 3.1 Qu'est-ce qu'un *moment* ?

Dans un modèle statistique, supposons qu'on dispose d'une statistique intégrable  $T$  (pas forcément réelle), dont la moyenne n'est pas le paramètre  $\theta$  lui-même, mais plutôt une *fonction* de  $\theta$  :

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

C'est cette fonction  $\varphi$  qu'on appelle *moment*. Typiquement,

- la moyenne d'une loi  $\mathcal{E}(\theta)$  n'est pas  $\theta$  mais  $1/\theta$ .
- la moyenne d'une loi log-normale de paramètres  $(0, \sigma^2)$  est  $e^{\sigma^2/2}$ .

Prenons la moyenne empirique associée à cet estimateur,  $\bar{T}_n$ . Par la loi des grands nombres,

$$\bar{T}_n = \frac{1}{n} \sum_{i=1}^n T(X_i) \rightarrow \varphi(\theta) \quad P_\theta - ps,$$

ce qui permet d'estimer  $\varphi(\theta)$ . Peut-on alors estimer  $\theta$  ?

### 3.2 Estimateur des moments

Si la fonction  $\varphi$  est inversible et si  $\bar{T}_n$  appartient presque sûrement à l'ensemble de définition de  $\varphi^{-1}$ , alors  $\varphi^{-1}(\bar{T}_n)$  est bien définie. Pour qu'en plus cette quantité converge presque sûrement vers  $\theta$ , il faut s'assurer que  $\varphi^{-1}$  est continue. C'est par exemple le cas lorsque l'ensemble des paramètres  $\Theta$  est un ouvert, et que  $\varphi$  est un difféomorphisme sur son image — une situation si fréquente qu'elle mérite son propre théorème, et si agréable qu'elle garantit que l'estimateur associé est asymptotiquement normal.

**Théorème 3.1** (Estimation par moments). *Sous l'hypothèse mentionnée ci-dessus (la fonction  $\varphi$  est un difféomorphisme), l'estimateur*

$$\hat{\theta}_n = \varphi^{-1}(\bar{T}_n)$$

*est presque sûrement bien défini pour tout  $n$  suffisamment grand ; il est également consistant pour l'estimation de  $\theta$ . En outre, si  $T$  est de carré intégrable, cet estimateur est asymptotiquement normal, au sens où  $\sqrt{n}(\hat{\theta}_n - \theta)$  converge en loi vers une gaussienne centrée de matrice de covariance*

$$(D\varphi(\theta))^{-1} \text{Cov}_\theta(T) (D\varphi(\theta)^\top)^{-1}.$$

*Démonstration.* La première partie a essentiellement été démontrée un peu plus haut. Pour la seconde, il faut d'abord remarquer que si  $T$  est de carré intégrable, alors  $\sqrt{n}(\bar{T}_n - \varphi(\theta))$  converge vers une loi  $N(0, \text{Cov}_\theta(T))$  par le TCL. Une simple application de la delta-méthode (Théorème 2.4) donne alors le résultat, puisque la matrice jacobienne de  $\varphi^{-1}$  en  $\varphi(\theta)$  n'est autre que l'inverse de la matrice jacobienne de  $\varphi$  en  $\theta$ .

□

## 4 Exercices

### 4.1 Questions

1. Montrer que la convergence en loi vers une constante implique la convergence en probabilité.
2. Montrer que, si un modèle statistique n'est pas identifiable, alors il ne peut exister aucun estimateur convergent.
3. Trouver un couple de variables aléatoires  $(X_n, Y_n)$  tel que  $X_n$  converge en loi et  $Y_n$  converge en loi, mais le couple ne converge pas en loi.
4. On observe un échantillon de lois de Poisson de paramètre  $\lambda$ , que l'on estime par la moyenne empirique. Calculer le risque quadratique de cet estimateur.
5. Quelle est la loi d'une somme de lois de Bernoulli indépendantes ? L'écart-type ?

### 4.2 Exercices

**Exercice 4.1** (Variance empirique). On se donne  $Y_1, \dots, Y_n$ , i.i.d de moyenne  $\mu$  et variance  $\sigma^2$ .

1. On suppose  $\mu$  connu. Donner un estimateur non biaisé de  $\sigma^2$ .
2. On suppose  $\mu$  inconnu. Calculer l'espérance de  $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ . En déduire un estimateur non biaisé de  $\sigma^2$ .

**Exercice 4.2** (Estimation de masse). Au cours de la seconde guerre mondiale, l'armée alliée notait les numéros de série  $X_1, \dots, X_n$  de tous les tanks nazis capturés ou détruits, afin d'obtenir un estimateur du nombre total  $N$  de tanks produits.

1. Proposer un modèle pour le tirage de  $X_1, \dots, X_n$ .
2. Calculer l'espérance de  $\bar{X}_n$ . En déduire un estimateur non biaisé de  $N$ . Indication: la loi de  $n$  tirages sans remise est échangeable.
3. Étudier la loi de  $X_{(n)}$  et en déduire un estimateur non biaisé de  $N$ .
4. Proposer deux intervalles de confiance de niveau  $1 - \alpha$  de la forme  $[aS, bS]$  avec  $a, b \in \mathbb{R}$  et  $S$  une statistique. On pourra utiliser le fait que l'inégalité de Hoeffding s'applique également aux tirages sans remise.

Selon Ruggles et Broodie (1947, JASA), la méthode statistique a fourni comme estimation une production moyenne de 246 tanks/mois entre juin 1940 et septembre 1942. Des méthodes d'espionnage traditionnelles donnaient une estimation de 1400 tanks/mois. Les chiffres officiels du ministère nazi des Armements ont montré après la guerre que la production moyenne était de 245 tanks/mois.

**Exercice 4.3** (Lois uniformes (1)). On considère  $(X_1, \dots, X_n)$  un échantillon de loi uniforme sur  $]\theta, \theta + 1[$ .



1. Donner la densité de la loi de la variable  $R_n = X_{(n)} - X_{(1)}$ , où  $X_{(1)} = \min(X_1, \dots, X_n)$  et  $X_{(n)} = \max(X_1, \dots, X_n)$ .
2. Étudier les différents modes de convergence de  $R_n$  quand  $n \rightarrow \infty$ .
3. Étudier le comportement en loi de  $n(1 - R_n)$  quand  $n \rightarrow \infty$ .

**Exercice 4.4** (Lois uniformes (2)). Soit  $X_1, \dots, X_n$  un échantillon de loi  $\mathcal{U}([0, \theta])$ , avec  $\theta > 0$ . On veut estimer  $\theta$ .

1. Déterminer un estimateur de  $\theta$  à partir de  $\bar{X}_n$ .
2. On considère l'estimateur  $X_{(n)} = \max_{1 \leq i \leq n} X_i$ . Déterminer les propriétés asymptotiques de ces estimateurs.
3. Comparer les performances des deux estimateurs.

**Exercice 4.5** (Lois Gamma). La loi Gamma  $\Gamma(\alpha, \beta)$  de paramètres  $\alpha, \beta > 0$  a pour densité

$$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x > 0.$$

On se donne un échantillon  $(X_1, \dots, X_n)$  de loi  $\Gamma(\alpha, \beta)$  et on cherche à estimer les paramètres.

1. On suppose le paramètre  $\beta$  connu. Proposer un estimateur de  $\alpha$  par la méthode des moments.
2. On suppose à présent que les deux paramètres  $\alpha, \beta$  sont inconnus. Proposer un estimateur de  $(\alpha, \beta)$  par la méthode des moments.

**Exercice 4.6** (Lois de Gumbel). La loi de Gumbel (centrale) de paramètre  $\beta$  a pour fonction de répartition  $F(x) = e^{-e^{-x/\beta}}$ . On observe un échantillon de lois de Gumbel et l'on cherche à estimer  $\beta$ .

1. Calculer la densité des lois de Gumbel, ainsi que leur moyenne et variance [indice : 0.57721...]
2. En déduire un estimateur convergent dont on calculera le risque quadratique et les propriétés asymptotiques.

**Exercice 4.7** (Lois de Yule-Simon). Une variable aléatoire  $X$  suit la loi de Yule-Simon de paramètre  $\rho > 0$  lorsque  $\mathbb{P}(X = n) = \rho B(n, 1 + \rho)$ , où  $n \geq 1$  et  $B$  est la [fonction beta](#).

1. Montrer que si  $\rho > 1$ , alors  $\mathbb{E}[X] = \rho/(\rho - 1)$ .
2. Trouver un estimateur de  $\rho$  et donner ses propriétés.

# 5 Intervalles de confiance

## 5.1 Principe

Dans un modèle statistique, l'estimation du paramètre d'intérêt  $\theta$  par intervalles de confiance consiste à spécifier un intervalle calculable à partir des données, et qui contient  $\theta$  avec grande probabilité : en d'autres termes, une *région de confiance* pour  $\theta$ .

Pour simplifier, on supposera d'abord que  $\theta$  est un paramètre réel.

**Définition 5.1** (intervalle de confiance). Un intervalle de confiance de niveau  $1 - \alpha$  est un intervalle  $I = [A, B]$  dont les bornes  $A, B$  sont des statistiques, et tel que pour tout  $\theta$ ,

$$P_{\theta}(\theta \in I) \geq 1 - \alpha.$$

Un intervalle de confiance de niveau asymptotique  $1 - \alpha$  est une *suite* d'intervalles  $I_n = [A_n, B_n]$  dont les bornes  $A_n, B_n$  sont des statistiques, et tels que pour tout  $n$ ,

$$P_{\theta}(\theta \in I_n) \geq 1 - \alpha.$$

Le terme « niveau » désigne  $1 - \alpha$  ; la vocation de ce nombre est d'être proche de 1, typiquement 99%. Le nombre  $\alpha$  est parfois appelé « erreur », « marge d'erreur » ou encore « niveau de risque » ; la vocation de ce nombre est d'être proche de zéro, typiquement 1%.

Il n'y a rien d'autre à savoir sur les intervalles de confiance ; tout l'art de la chose consiste à savoir les construire. Commençons par des exemples essentiels à plusieurs titres : le cas d'un échantillon gaussien, et le cas de lois de Bernoulli.

## 5.2 Exemples gaussiens

On dispose de variables aléatoires  $X_1, \dots, X_n$  de loi  $N(\mu, \sigma^2)$ . On va donner des intervalles de confiance pour l'estimation des paramètres  $\mu$  et  $\sigma$  dans plusieurs cas de figure.

### 5.2.1 Estimation de $\mu$

**Lorsque  $\sigma$  est connue.**

Nous avons déjà vu que la moyenne empirique  $\bar{X}_n$  est un estimateur sans biais de  $\mu$ . Or, nous savons aussi la loi *exacte* de  $\bar{X}_n$ , qui est  $N(\mu, \sigma^2/n)$ . Autrement dit,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \sim N(0, 1). \quad (5.1)$$

Dans cette équation, on a trouvé une variable aléatoire dont la loi ne dépend plus de  $\mu$ . Il est donc possible de déterminer un intervalle dans lequel elle fluctue à l'aide des quantiles de la loi normale, qui sont étudiés dans Section 6.1. Si l'on se donne une marge d'erreur  $\alpha = 1\%$ , alors

$$\mathbb{P}((\sqrt{n}/\sigma)|\bar{X}_n - \mu| > z_{0.99}) = 1\%$$

où  $z_{0.99} \approx 2.32$ . Or,

$$\frac{\sqrt{n}}{\sigma}|\bar{X}_n - \mu| > z_{0.99} \quad (5.2)$$

est équivalent à

$$\bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}. \quad (5.3)$$

Le passage de Équation 5.2 à Équation 5.3 est souvent appelé *pivot* et sert à passer d'un intervalle de fluctuation à un intervalle de confiance.

Nous avons donc les deux bornes de notre intervalle de confiance :

$$A = \bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}}$$

$$B = \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}.$$

Ces deux quantités sont bien des statistiques, car  $\sigma$  est connu. De plus, nous venons de montrer que  $P_\mu(\mu \in [A, B]) = 99\%$ . Ici, le choix de la marge d'erreur  $\alpha = 1\%$  ne jouait aucun rôle particulier ; ainsi, un intervalle de confiance de niveau  $1 - \alpha$  pour l'estimation de  $\mu$  est donné par

$$\left[ \bar{X}_n - \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \ ; \ \bar{X}_n + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \right]. \quad (5.4)$$

### Lorsque $\sigma$ est inconnue.

Lorsque  $\sigma$  n'est pas connu, les bornes  $A, B$  ci-dessus ne sont pas des statistiques, car elles dépendent de  $\sigma$ . Heureusement, on peut estimer  $\sigma$  sans biais via l'estimateur

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

Que se passe-t-il si, dans la statistique Équation 5.1, on remplace  $\sigma$  par son estimation  $\hat{\sigma}_n^2$  ? On obtient la statistique dite *de Student*,

$$T_n = \frac{\sqrt{n}}{\sqrt{\hat{\sigma}_n^2}}(\bar{X}_n - \mu). \quad (5.5)$$

Sa loi n'est plus une loi gaussienne, mais une loi de Student à  $n - 1$  paramètres de liberté  $\mathcal{T}(n - 1)$ : le calcul de la densité est fait en détails dans Section 6.2.3 - Section 6.2.4. Les quantiles des lois de Student ont été calculés avec précision. On notera  $t_{k,\alpha}$  le quantile symétrique de niveau  $\alpha$  de  $\mathcal{T}(k)$ . Alors,

$$P_{\mu,\sigma^2}(|T_n| > t_{n-1,\alpha}) \leq \alpha.$$

Par le même raisonnement que tout à l'heure, l'inégalité

$$\left| \frac{\sqrt{n}}{\sqrt{\hat{\sigma}_n^2}}(\bar{X}_n - \mu) \right| > t_{n-1,\alpha}$$

est équivalente à

$$\bar{X}_n - \frac{t_{n-1,\alpha}\hat{\sigma}_n^2}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{t_{n-1,\alpha}\hat{\sigma}_n^2}{\sqrt{n}}.$$

et les deux côtés de ces inégalités sont des statistiques; en les notant  $A, B$ , on a bien trouvé un intervalle de confiance de niveau  $\alpha$ , c'est-à-dire tel que  $P_{\mu,\sigma^2}(\mu \in [A, B]) = \alpha$ . Cet intervalle de confiance est d'une grande importance en pratique et mérite son propre théorème. Il est dû à [William Gosset](#).

**Théorème 5.1** (Intervalle de Student). *Un intervalle de confiance de niveau  $1 - \alpha$  pour l'estimation de  $\mu$  lorsque  $\sigma$  n'est pas connue est donné par*

$$\left[ \bar{X}_n - \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}} \ ; \ \bar{X}_n + \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}} \right].$$

### 5.2.2 Estimation de $\sigma$

Supposons maintenant qu'on désire estimer la variance  $\sigma^2$ .

**Lorsque  $\mu$  est connue.**

En supposant que  $\mu$  est connue, l'estimateur des moments le plus naturel pour estimer  $\sigma^2$  est évidemment

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Comme les  $(X_i - \mu)/\sigma$  sont des variables aléatoires gaussiennes centrées réduites, l'estimateur  $\tilde{\sigma}_n^2 \times (n/\sigma^2)$  est une somme de  $n$  gaussiennes standard indépendantes. La loi de cette statistique est connue : c'est une [loi du chi-deux](#) à  $n$  paramètres de liberté comme démontré dans Section 6.2.2. Cette loi n'est pas symétrique, puisqu'elle est supportée sur  $[0, \infty[$ . On note souvent  $k_{n,\alpha}^-$  et  $k_{n,\alpha}^+$  les nombres les plus éloignées possible (ils existent) tels que  $\mathbb{P}(k_{n,\alpha}^- < \chi^2(n) < k_{n,\alpha}^+) = 1 - \alpha$ . Ainsi,

$$P_{\sigma^2}(k_{n,\alpha}^- < \frac{n\tilde{\sigma}_n^2}{\sigma^2} < k_{n,\alpha}^+) = \alpha.$$

En pivotant comme dans les exemples précédents, on obtient que l'intervalle

$$\left[ \frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^+} \ ; \ \frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^-} \right]$$

est un intervalle de confiance de niveau  $\alpha$  pour  $\sigma^2$ .

**Lorsque  $\mu$  est inconnue.**

Cette fois, on utilise l'estimateur déjà évoqué plus tôt, à savoir

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

La loi de  $(n-1)\hat{\sigma}_n^2/\sigma^2$  est encore une loi du chi-deux, mais à  $n-1$  paramètres de liberté. Ainsi, le même raisonnement que ci-dessus donne l'intervalle de confiance de niveau  $\alpha$  suivant :

$$\left[ \frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^+} \ ; \ \frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^-} \right].$$

## 5.3 Exemples asymptotiques

### 5.3.1 Estimation du paramètre $p$ dans un modèle de Bernoulli.

Soient  $X_1, \dots, X_n$  des variables indépendantes de loi  $\mathcal{B}(p)$ , dont on cherche à estimer le paramètre  $p \in ]0, 1[$ . Un estimateur naturel est donné par la moyenne empirique,  $\hat{p}_n = (X_1 + \dots + X_n)/n$ . Cet estimateur est non biaisé et son risque quadratique est égal à  $p(1-p)/n$ . De plus, la loi de  $\hat{p}_n$  est connue :  $n\hat{p}_n \sim \text{Bin}(n, p)$ . Par conséquent, si l'on connaît les quantiles de  $\text{Bin}(n, p) - p$ , on pourra construire des intervalles de confiance de niveau  $1 - \alpha$ . Ces quantiles peuvent être calculés par des méthodes numériques, mais il existe des façons plus simples de faire.

**Inégalité BT.** L'inégalité de Bienaymé-Tchebychev dit que

$$P_p(|\hat{p}_n - p| > t) \leq \frac{p(1-p)}{nt^2}. \quad (5.6)$$

Si l'on choisit

$$t = \sqrt{\frac{p(1-p)}{n\alpha}},$$

cette probabilité est plus petite que  $\alpha$ . En pivotant, on en déduit que l'intervalle  $[\hat{p}_n \pm \sqrt{p(1-p)/n\alpha}]$  contient  $p$  avec une probabilité supérieure à  $1 - \alpha$ . Mais les bornes de cet intervalle ne sont pas des statistiques, car elles dépendent de  $p$  ! Fort heureusement, on sait que  $p$  est entre 0 et 1, ce qui entraîne que  $p(1-p)$  est plus petit que  $1/4$ , donc l'intervalle ci-dessus est contenu dans l'intervalle plus grand

$$\left[ \hat{p}_n \pm \frac{1}{2\sqrt{n\alpha}} \right].$$

Ce dernier est bien un intervalle de confiance de niveau  $1 - \alpha$  pour l'estimation de  $p$ .

**TCL.** On a mentionné que les quantiles des lois binomiales pourraient être calculés ; or, ils peuvent également être approchés grâce au théorème central-limite. Celui-ci dit que

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \rightarrow N(0, 1). \quad (5.7)$$

Si  $z_\alpha$  est le quantile symétrique d'ordre  $\alpha$  de  $N(0, 1)$ , alors on en déduit que

$$\mathbb{P} \left( \left| \frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \right| > z_\alpha \right) \rightarrow \alpha.$$

En pivotant, on voit alors que l'intervalle

$$\left[ \hat{p}_n \pm z_\alpha \sqrt{p(1-p)/n} \right]$$

contient  $p$  avec une probabilité *qui tend lorsque  $n \rightarrow \infty$  vers  $1 - \alpha$* . Là encore, cet intervalle n'est pas un intervalle de confiance. On pourrait utiliser deux techniques.

1. Comme tout à l'heure, l'intervalle ci-dessus est contenu dans l'intervalle plus grand  $[\hat{p}_n \pm z_\alpha/2\sqrt{n}]$  qui est un intervalle de confiance *asymptotique* de niveau  $1 - \alpha$ .

2. Il y a plus fin. Nous savons par la loi des grands nombres que  $\hat{p}_n \rightarrow p$  en probabilité. Ainsi,  $\sqrt{\hat{p}_n(1-\hat{p}_n)} \rightarrow \sqrt{p(1-p)}$  en probabilité. Le lemme de Slutsky nous assure alors que dans Équation 5.8, on peut remplacer le dénominateur par  $\sqrt{\hat{p}_n(1-\hat{p}_n)}$  pour obtenir

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \rightarrow N(0, 1). \quad (5.8)$$

Le reste du raisonnement est identique, et l'on obtient l'intervalle de confiance asymptotique de niveau  $1 - \alpha$  suivant :

$$\left[ \hat{p}_n \pm z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right]$$

**Hoeffding.** L'inégalité de Bienaymé-Tchebychev n'est pas très fine. Il existe de nombreuses autres inégalités de concentration : l'inégalité de Hoeffding (Théorème 6.4) concerne les variables bornées, comme ici où les  $X_i$  sont dans  $[0, 1]$ . Cette inégalité dit que

$$\mathbb{P}(|\hat{p}_n - p| > t) \leq 2e^{-2nt^2}.$$

Le choix

$$t = \sqrt{\frac{1}{2n} \ln \left( \frac{1}{\alpha} \right)}$$

donne une probabilité inférieure à  $\alpha$ , et fournit donc l'intervalle de confiance **non-asymptotique** de niveau  $1 - \alpha$  suivant :

$$\left[ \bar{X}_n \pm \frac{\ln(1/\alpha)}{\sqrt{2n}} \right].$$

### 5.3.2 Estimation de moyenne dans un modèle non-gaussien.

Les deux techniques ci-dessus n'ont rien de spécifique au cas de variables de Bernoulli. En fait, elles s'appliquent à tout modèle statistique iid dont on cherche à estimer la moyenne  $\mu$ , pourvu que la variance existe.

La première méthode utilisant Bienaymé-Tchebychev nécessite de borner la variance. Cela peut se faire dans certains cas, mais pas dans tous.

La seconde méthode s'applique systématiquement en utilisant l'estimateur de la variance empirique  $\hat{\sigma}_n^2$ . En effet, la convergence

$$\frac{\sqrt{n}}{\hat{\sigma}_n}(\bar{X}_n - \mu) \rightarrow N(0, 1)$$

est toujours vraie d'après le théorème de Slutsky.

**Théorème 5.2.** Soient  $X_1, \dots, X_n$  des variables iid possédant une variance. L'intervalle

$$\left[ \bar{X}_n \pm \frac{z_\alpha \hat{\sigma}_n}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique de niveau  $\alpha$  pour l'estimation de la moyenne des  $X_i$ .

# 6 Outils pour les IC

## 6.1 Quantiles

Si  $X$  est une variable aléatoire sur  $\mathbb{R}$ , un quantile d'ordre  $\beta \in ]0, 1[$ , noté  $q_\beta$ , est un nombre tel que  $\mathbb{P}(X \leq q_\beta) = \beta$ . Lorsque  $X$  est continue, un tel nombre existe forcément, car la fonction de répartition  $F(x) = \mathbb{P}(X \leq x)$  est une surjection continue. Les quantiles symétriques  $z_\beta$  sont, eux, définis par  $\mathbb{P}(|X| \leq z_\beta) = \beta$ .

Si la loi de  $X$  est de surcroît symétrique, les quantiles symétriques s'expriment facilement en fonction des quantiles classiques. En effet,  $\mathbb{P}(|X| \leq z)$  est égal à  $\mathbb{P}(X \leq z) - \mathbb{P}(X \leq -z)$ . Or, si la loi de  $X$  est symétrique, alors  $\mathbb{P}(X \leq -z) = 1 - \mathbb{P}(X \leq z)$ , et donc

$$\mathbb{P}(|X| \leq z) = 2\mathbb{P}(X \leq z) - 1.$$

Il suffit alors de choisir pour  $z$  le quantile  $q_{\frac{1+\beta}{2}}$  pour obtenir  $\mathbb{P}(|X| \leq z) = \beta$ . Lorsque  $\beta$  est de la forme  $1 - \alpha$  avec  $\alpha$  petit (comme les niveaux des intervalles de confiance), on trouve alors  $z_{1-\alpha} = q_{1-\alpha/2}$ .

Les quantiles s'obtiennent en inversant la fonction de répartition : lorsque celle-ci est une bijection sur  $]0, 1[$ , alors  $q_\beta = F^{-1}(\beta)$ . En règle générale, il n'y a pas de forme fermée. Par exemple, pour une loi gaussienne standard,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du$$

qui elle-même n'a pas d'écriture plus simple. Fort heureusement, les outils de calcul numérique permettent d'effectuer ces calculs avec une grande précision. La table suivante donne les quantiles symétriques de la gaussienne.

$\beta$	90%	95%	98%	99%	99.9%	99.99999%
$z_\beta$	1.64	1.96	2.32	2.57	3.2	5.32

Voir aussi la [règle 1-2-3](#). Il existe de nombreuses tables de quantiles pour les lois usuelles.

**Théorème 6.1** (Queues de distribution de la gaussienne). *Si  $x$  est plus grand que 1,*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \leq \mathbb{P}(X > x) \leq \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

*En particulier, si  $x$  est grand,  $\mathbb{P}(X \geq x) \sim e^{-x^2/2}/x\sqrt{2\pi}$  avec une erreur d'ordre  $O(e^{-x^2/2}/x^3)$ .*

À titre d'exemple, pour  $x = 2.32$  cette approximation donne 98.83%, ce qui est remarquablement proche de 98%. Pour  $x = 2.57$  on trouve 99.42%.

Démonstration. À écrire.

□

## 6.2 Calculs de lois

### 6.2.1 Lois Gamma

Une variable aléatoire suit une loi Gamma de paramètres  $\lambda > 0, \alpha > 0$  lorsque sa densité est donnée par

$$\gamma_{r,\alpha}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \mathbf{1}_{x>0}.$$

Les lois Gamma rassemblent les lois exponentielles ( $\Gamma(\lambda, 1) = \mathcal{E}(\lambda)$ ) et les lois du chi-deux qu'on verra ci-dessous ( $\Gamma(1/2, n/2) = \chi_2(n)$ ). La transformée de Fourier  $\varphi_{\lambda,\alpha}$  d'une loi  $\Gamma(\lambda, \alpha)$  se calcule facilement par un changement de variables :

$$\varphi_{\lambda,\alpha}(t) = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}. \quad (6.1)$$

Cette identité montre également que si  $X_1, \dots, X_n$  sont des variables indépendantes de loi  $\Gamma(\lambda, \alpha_i)$ , alors leur somme est une variable de loi  $\Gamma(\lambda, \alpha_1 + \dots + \alpha_n)$ .

### 6.2.2 Loi du chi-deux

Soit  $X$  une loi gaussienne standard. Calculons la densité de  $X^2$  ; pour toute fonction-test  $\varphi$ ,  $\mathbb{E}[\varphi(X^2)]$  est donné par

$$\frac{1}{\sqrt{2\pi}} \int e^{-x^2/2} \varphi(x^2) dx.$$

Cette intégrale est symétrique, donc on peut ajouter un facteur 2 et intégrer sur  $[0, \infty[$ . En posant  $u = x^2$ , on obtient alors la valeur

$$\frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-u/2} \varphi(u) \frac{1}{2\sqrt{u}} du.$$

On reconnaît la densité d'une [loi Gamma](#) de paramètres  $(1/2, 1/2)$ . Cette loi s'appelle *loi du chi-deux* et on la note  $\chi_2(1)$ . Sa transformée de Fourier est donnée par

$$\mathbb{E}[e^{itX^2}] = \frac{1}{\sqrt{1-2it}}.$$

Soient maintenant  $X_1, \dots, X_n$  des variables de loi  $N(0, 1)$  indépendantes. Chaque  $X_i^2$  est une  $\chi_2(1)$  ; leur somme a pour loi la convolée  $n$  fois de  $\chi_2(1)$ . Calculons sa transformée de Fourier :

$$\mathbb{E}[e^{it(X_1^2 + \dots + X_n^2)}] = \mathbb{E}[e^{itX_1^2}]^n \quad (6.2)$$

$$= (1 - 2it)^{-\frac{n}{2}}. \quad (6.3)$$

On reconnaît la transformée de Fourier d'une loi  $\Gamma(n/2, 1/2)$  ; cette loi s'appelle *loi du chi-deux à  $n$  paramètres de liberté* et elle est notée  $\chi_2(n)$ . Sa densité est donnée par

$$\frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} x^{n/2-1} \mathbf{1}_{x>0}. \quad (6.4)$$



### 6.2.3 Loi de Student

Soit  $X$  une variable de loi  $N(0, 1)$  et  $Y_n$  une variable de loi  $\chi_2(n)$  indépendante de  $X$ . On va calculer la loi de  $T_n = X/\sqrt{Y_n/n}$ . Soit  $\varphi$  une fonction test ; l'espérance  $\mathbb{E}[\varphi(T_n)]$  est égale à

$$\frac{1}{Z_n \sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi\left(\frac{x}{\sqrt{y/n}}\right) e^{-\frac{x^2}{2}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} dx dy$$

où  $Z_n = 2^{n/2} \Gamma(n/2)$ . Dans l'intégrale en  $x$ , on effectue le changement de variable  $u = x/\sqrt{y/n}$  afin d'obtenir

$$\frac{1}{Z_n \sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi(u) e^{-\frac{yu^2}{2n}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} \sqrt{\frac{y}{n}} dx dy.$$

La densité de  $T_n$  est donc donnée par

$$t_n(u) = \frac{1}{Z_n \sqrt{2\pi n}} \int_0^\infty e^{-\frac{yu^2}{2n} - \frac{y}{2}} y^{\frac{n+1}{2}-1} dy.$$

Le changement de variables  $z = y(1 + u^2/n)/2$  nous ramène à

$$t_n(u) = \frac{1}{Z_n \sqrt{2\pi n}} \left( \frac{2}{1 + \frac{u^2}{n}} \right)^{\frac{n+1}{2}} \int_0^\infty e^{-z} z^{\frac{n+1}{2}-1} dz.$$

On reconnaît  $\Gamma((n+1)/2)$  à droite. La densité  $t_n(x)$  est donc

$$t_n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left( \frac{1}{1 + \frac{x^2}{n}} \right)^{\frac{n+1}{2}}.$$

Cette loi s'appelle *loi de Student* de paramètre  $n$  ; on dit parfois à  $n$  *degrés de liberté*. Elle est notée  $\mathcal{T}(n)$ . La loi de Student de paramètre  $n = 1$  est tout simplement une loi de Cauchy.

### 6.2.4 Loi de la statistique de Student

Soient  $X_1, \dots, X_n$  des variables gaussiennes  $N(\mu, \sigma^2)$  indépendantes, et soit  $T_n = (\bar{X}_n - \mu)/\sqrt{\hat{\sigma}_n^2}$ , où

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

**Théorème 6.2.**

$$T_n \sim \mathcal{T}(n-1).$$

*Démonstration.* On va montrer 1° que  $\bar{X}_n$  et  $\sqrt{\hat{\sigma}_n^2/\sigma^2}$  sont indépendantes, et 2° que  $\sqrt{\hat{\sigma}_n^2/\sigma^2}$  a bien la même loi que  $\sqrt{Y_{n-1}/(n-1)}$  où  $Y_{n-1}$  est une  $\chi_2(n-1)$ . Dans la suite, on supposera que  $\mu = 0$  et  $\sigma = 1$ , ce qui n'enlève rien en généralité.

*Premier point.* Le vecteur  $X = (X_1, \dots, X_n)$  est gaussien. Posons  $Z = (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ . Le couple  $(\bar{X}_n, Z_n)$  est linéaire en  $X$ , donc ce couple est aussi un vecteur gaussien. Or, la covariance de ses deux éléments est nulle. Par exemple,  $\text{Cov}(\bar{X}_n, Z_1)$  est égale à  $\text{Cov}(\bar{X}_n, X_1) - \text{Var}(\bar{X}_n)$ , ce qui par linéarité donne  $1/n - 1/n = 0$ . Ainsi,  $\bar{X}_n$  et  $Z$  sont deux variables conjointement gaussiennes et décorréées : elles sont donc indépendantes. Comme  $\hat{\sigma}_n$  est une fonction de  $Z$ , elle est aussi indépendante de  $\bar{X}_n$ .

*Second point.*  $Z$  est la projection orthogonale de  $X$  sur le sous-espace vectoriel  $\mathcal{V} = \{x \in \mathbb{R}^n : x_1 + \dots + x_n = 0\}$ . Soit  $(f_i)_{i=2, \dots, n}$  une base orthonormale de  $\mathcal{V}$ , de sorte que  $Z = \sum_{i=2}^n \langle f_i, X \rangle f_i$ . Par l'identité de Parseval,

$$|Z|^2 = \sum_{i=2}^n |\langle f_i, X \rangle|^2.$$

Or, les  $n-1$  variables aléatoires  $G_i = \langle f_i, X \rangle$  sont des gaussiennes standard iid. En effet, on vérifie facilement que  $\text{Cov}(G_i, G_j) = \langle f_i, f_j \rangle = \delta_{i,j}$ . On en déduit donc que  $|Z|^2$  suit une loi  $\chi_2(n-1)$ .

□

La seconde partie de la démonstration est un cas particulier du théorème de Cochran, que nous verrons dans le chapitre sur la régression linéaire.

## 6.3 Inégalités de concentration

Les outils de base pour construire des intervalles de confiance dans des circonstances générales (non gaussiennes) sont les inégalités de concentration. Une inégalité de concentration pour une variable aléatoire intégrable  $X$  consiste à borner  $\mathbb{P}(|X - \mathbb{E}[X]| > x)$  par quelque chose de petit quand  $x$  est grand : on cherche à contrôler la probabilité pour que les réalisations de la variable aléatoire  $X$  soient éloignées de leur valeur moyenne  $\mathbb{E}[X]$  de plus de  $x$ .

## 6.4 Inégalité de Bienaymé-Tchebychev

**Théorème 6.3.** Soit  $X$  une variable aléatoire de carré intégrable. Alors,

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

*Démonstration.* Élever au carré les deux membres de l'inégalité dans  $\mathbb{P}$ , puis appliquer l'inégalité de Markov à la variable aléatoire positive  $|X - \mathbb{E}[X]|^2$  dont l'espérance est  $\text{Var}(X)$ .

□

## 6.5 Inégalité de Hoeffding

**Théorème 6.4** (Inégalité de Hoeffding). *Soient  $X_1, \dots, X_n$  des variables aléatoires indépendantes, pas forcément de même loi. On suppose que chaque  $X_i$  est à valeurs dans un intervalle borné  $[a_i, b_i]$  et on pose  $S_n = X_1 + \dots + X_n$ . Pour tout  $t > 0$ ,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \quad (6.5)$$

et

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}. \quad (6.6)$$

La démonstration se fonde sur le lemme suivant.

**Lemme 6.1** (lemme de Hoeffding). *Soit  $X$  une variable aléatoire à valeurs dans  $[a, b]$ . Pour tout  $t$ ,*

$$\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leq e^{\frac{t^2(b-a)^2}{8}}. \quad (6.7)$$

*Démonstration.* Soit  $X$  une variable aléatoire, que par simplicité on supposera centrée et à valeurs dans l'intervalle  $[a, b]$  ( $a$  est forcément négatif). En écrivant

$$x = a \times \frac{b-x}{b-a} + b \times \left(1 - \frac{b-x}{b-a}\right)$$

et en utilisant la convexité de la fonction  $x \mapsto e^{tx}$ , on obtient  $e^{tX} \leq (b-X)e^{ta}/(b-a) + (1 - (b-X)/(b-a))e^{bt}$ , puis en prenant l'espérance et le fait que  $X$  est centrée et en simplifiant,

$$\mathbb{E}[e^{tX}] \leq \frac{be^{ta} - ae^{tb}}{b-a}.$$

Notons  $f(t)$  le terme à droite ; pour montrer l'équation 6.7, il suffit de montrer que  $\ln f(t) \leq t^2(b-a)^2/8$ . La formule de Taylor dit que

$$\ln f(t) = \ln f(0) + t(\ln f)'(0) + \frac{t^2}{2}(\ln f)''(\xi)$$

pour un certain  $\xi$ . Or,  $\ln f(0) = \ln 1 = 0$ ,  $(\ln f)'(0) = f'(0)/f(0) = 0$ , et il suffit donc de montrer que  $(\ln f)''(t)$  est toujours plus petit que  $(b-a)^2/4$  pour conclure. Un simple calcul montre que  $\ln f(t) = \ln(b/(b-a)) + ta + \ln(1 - ae^{t(b-a)}/b)$ , et donc

$$(\ln f)''(t) = \frac{(a/b)(b-a)e^{t(b-a)}}{(1 - ae^{t(b-a)}/b)^2}.$$

L'inégalité  $uv/(u-v)^2 \leq 1/4$  appliquée à  $u = a/b$  et  $v = e^{t(b-a)}$  permet alors de conclure.

□

*Preuve de l'inégalité de Hoeffding.* En remplaçant  $X_k$  par  $X_k - \mathbb{E}[X_k]$ , on peut supposer que tous les  $X_i$  sont centrés et étudier seulement  $\mathbb{P}(S_n > t)$ . Écrivons  $\mathbb{P}(S_n > t) = \mathbb{P}(e^{\lambda S_n} > e^{\lambda t})$ , où  $\lambda$  est un nombre positif que l'on choisira plus tard. L'inégalité de Markov borne cette probabilité par  $\mathbb{E}[e^{\lambda S_n}]e^{-\lambda t}$ . Comme les  $X_i$  sont indépendantes,  $\mathbb{E}[e^{tS_n}]$  est le produit des  $e^{\varphi_k(\lambda)}$  où  $\varphi_k(t) = \ln \mathbb{E}[e^{itX_k}]$ . En appliquant le lemme de Hoeffding à chaque  $\varphi_k$ , on borne  $\mathbb{P}(S_n > t)$  par

$$\exp \left( \sum_{i=1}^n \frac{(b_i - a_i)^2 \lambda^2}{8} - t\lambda \right).$$

Le minimum en  $\lambda$  du terme dans l'exponentielle est atteint au point  $4t / \sum (a_i - b_i)^2$  et la valeur du minimum est le terme dans l'exponentielle de Équation 6.5. On déduit Équation 6.6 par une simple borne de l'union.

La démonstration de l'inégalité de Hoeffding ne dépend pas directement du fait que  $X$  est bornée, mais plutôt de Équation 6.7. Toutes les variables aléatoires qui vérifient une inégalité de type  $\mathbb{E}[e^{tX}] \leq e^{ct^2}$  pour une constante  $c$  peuvent donc avoir leur propre inégalité de Hoeffding.

## 7 Exercices

### 7.1 Questions

1. Soit  $X_n$  une variable aléatoire de loi de Student de paramètre  $n$ . Montrer que  $X_n$  converge en loi vers  $N(0, 1)$ .
2. Soit  $X_n \sim \chi_2(n)$ . La suite  $(X_n)$  est-elle asymptotiquement normale ?
3. Donner un intervalle de confiance de la forme  $[A, +\infty[$  pour la moyenne d'un échantillon gaussien.
4. Même question pour la variance dans un modèle gaussien centré.
5. Dans l'estimation de la moyenne  $\mu$  d'un modèle gaussien où la variance  $\sigma^2$  est connue, montrer que l'intervalle de confiance obtenu (Équation 5.4) est le plus grand possible de niveau  $1 - \alpha$ .
6. Démontrer le théorème Théorème 6.1 sur l'asymptotique des queues de distribution de la loi gaussienne.
7. Montrer la borne suivante sur les quantiles de loi gaussienne standard:  $q_\beta < \sqrt{\ln \frac{1}{\beta\sqrt{2\pi}}}$  (pour tout  $1/2 < \beta < 1$ ).
8. Comparer les queues de distribution des lois  $N(0, 1)$ ,  $\chi_2(n)$  et  $\mathcal{T}(n)$ .
9. Expliquer à votre grand-mère la différence entre un intervalle de fluctuation et un intervalle de confiance.
10. L'intervalle de confiance de niveau  $1 - \alpha$  pour la moyenne d'un modèle  $N(\mu, 1)$  avec  $n$  observations est  $I_n = [\bar{X}_n \pm z_\alpha / \sqrt{n}]$ . Supposons qu'on obtienne une nouvelle observation indépendante des autres, disons  $Z$ . La probabilité  $\mathbb{P}(Z \in I_n)$  est-elle plus grande ou plus petite que  $1 - \alpha$  ?
11. Comparer la longueur des intervalles de confiance obtenus par les différentes méthodes de la section Section 5.3.1.

### 7.2 Exercices

**Exercice 7.1** (Lois de Poisson). On suppose que l'on observe  $X_1, \dots, X_n$  i.i.d de loi  $\mathcal{P}(\theta)$ .

1. Étudier  $\bar{X}_n$ .
2. Montrer que  $\sqrt{\bar{X}_n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \sqrt{\theta}$ .
3. Donner deux intervalles de confiance au niveau 98% pour  $\sqrt{\theta}$ , et les comparer.

**Exercice 7.2** (Lois uniformes). Soit  $X_1, \dots, X_n$  des variables aléatoires iid de loi  $\mathcal{U}[0, \theta]$ . Donner un intervalle de confiance non asymptotique pour  $\theta$  en utilisant l'estimateur  $\hat{\theta}_n = \max_{i=1, \dots, n} X_i$ .

**Exercice 7.3** (Lois exponentielles décalées). Soit  $X_1, \dots, X_n$  des variables aléatoires iid de densité  $e^{\theta-x} \mathbf{1}_{x>\theta}$ , où  $\theta > 0$ .

1. Calculer  $\mathbb{E}_\theta[X_1]$  et en déduire un estimateur de  $\theta$  que l'on notera  $\hat{\theta}_n$ . Étudier ses propriétés (risque quadratique, convergence) et l'utiliser pour construire un premier intervalle de confiance  $I_1(\alpha)$  non-asymptotique pour  $\theta$  de niveau  $1 - \alpha$ .
2. Construire un intervalle de confiance asymptotique  $I_2(\alpha)$  pour  $\theta$  à partir de  $\hat{\theta}_n$ .
3. Montrer que l'estimateur  $\theta_n^* := \min_{1 \leq i \leq n} X_i$  est meilleur que  $\hat{\theta}_n$  au sens du risque quadratique, puis l'utiliser pour construire un intervalle de confiance  $I_3(\alpha)$  de niveau  $1 - \alpha$ .
4. Comparer les longueurs de tous ces différents intervalles de confiance.

**Exercice 7.4** (Lois exponentielles). Soit  $X_1, \dots, X_n$  des variables aléatoires iid exponentielles de paramètre  $\lambda > 0$ .

1. Quelle est la loi de  $S_n = X_1 + \dots + X_n$  ?
2. Construire un intervalle de confiance de niveau  $1 - \alpha$  pour  $\lambda$ .

**Exercice 7.5** (Inégalité d'Azuma). Montrer que l'inégalité de Hoeffding reste valable lorsque les  $X_i$  ne sont plus supposés indépendants, mais que la suite  $S_k = X_1 + \dots + X_k$  est une martingale. Indice :  $\mathbb{E}[e^{\lambda S_{n+1}}] = \mathbb{E}[e^{\lambda S_n} \mathbb{E}[e^{\lambda X_{n+1}} | S_n]]$ .

Ce raffinement s'appelle *inégalité de Hoeffding-Azuma*. C'est celui que nous avons utilisé dans l'exercice (ex-tanks?), lorsque les  $X_1, \dots, X_n$  sont des tirages *sans remise* dans une urne à  $N$  éléments.

## 8 Test d'hypothèses

Si l'on essaie d'estimer le rendement  $\mu$  d'un actif financier, on cherche implicitement à savoir si l'on va investir ou pas. Cette décision dépendra de notre estimation : pour faire simple, on peut considérer que si nous estimons que le rendement est positif ( $\hat{\mu} > 0$ ), alors il faut investir. Sinon, on n'investira pas.

Les tests d'hypothèses visent à formaliser cela. Faire une *hypothèse* dans un modèle statistique  $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta_0})$ , c'est supposer que  $\theta$  appartient à une certaine région de  $H_0 \subset \Theta$ . Les *tests* visent à construire des procédures pour tester une hypothèse nulle, que l'on notera  $H_0$ , contre une hypothèse alternative, notée  $H_1$ .

Dans le cadre ci-dessus, on peut se placer dans un modèle où les rendements sont  $\mathcal{N}(\mu, \sigma^2)$ . On veut tester l'hypothèse nulle  $H_0 : \mu \in ]-\infty, 0]$  contre l'hypothèse alternative  $H_1 : \mu \in ]0, +\infty[$ .

**Définition 8.1.** Un test est un événement qui, s'il survient, nous incite à rejeter l'hypothèse nulle. Cet événement sera noté *rejeter* et son complémentaire sera noté *accepter*.

- L'erreur de première espèce est la probabilité de rejeter l'hypothèse nulle à tort :  $\alpha = \sup_{\theta \in H_0} P_\theta(\text{rejeter})$ . Le **niveau d'un test** est  $1 - \alpha$ . C'est la probabilité d'accepter l'hypothèse nulle à raison.
- L'erreur de seconde espèce est la probabilité de ne pas rejeter l'hypothèse nulle, à tort :  $\beta = \sup_{\theta \in H_1} P_\theta(\text{accepter})$ . La **puissance d'un test** est  $1 - \beta$ . C'est la probabilité de « détecter » l'hypothèse alternative à raison.
- L'affinité d'un test est la somme des erreurs de première et seconde espèce. On parle aussi de *l'erreur totale*.

Par « événement », on veut bien dire « un élément de  $\mathcal{F}$  », c'est-à-dire qui n'est déterminé que par les observations et pas par  $\theta$ . Formellement on écrit souvent qu'un test est une statistique, disons  $T$ , à valeurs dans  $\{0, 1\}$ . L'événement  $\{T = 0\}$  est *rejeter*, l'événement  $\{T = 1\}$  est *accepter*.

Un des grands objectifs de la statistique mathématique est de construire des familles de tests qui, pour un niveau de confiance  $1 - \alpha$  fixé, ont la plus grande puissance possible ; autrement dit, **trouver un événement hautement improbable sous l'hypothèse nulle, et hautement probable sous l'hypothèse alternative**.

Comme on verra dans les exemples, le rôle des deux hypothèses n'est pas interchangeable. Maximiser le niveau et la puissance ne reviennent pas au même. Le choix des hypothèses  $H_0$  et  $H_1$  n'est pas anodin : l'hypothèse  $H_0$  est une hypothèse que l'on cherche implicitement à réfuter.

1. Si  $\theta \in H_0$  quel qu'il soit, les probabilités pour qu'un certain événement *rejeter* sont infimes – disons, 1%.
2. Si cet événement arrive, par contraposée, on est amenés à rejeter l'hypothèse selon laquelle  $\theta$  est dans  $H_0$ .

C'est pour cela que les tests sont une forme de logique statistique. Le raisonnement de base une contraposée : en logique,  $A \Rightarrow B$  est équivalent à  $\neg B \Rightarrow \neg A$ . En statistiques, on pourrait écrire  $\theta \in H_0 \Rightarrow \text{accepter}$  (avec grande probabilité), donc  $\text{rejeter} \Rightarrow \theta \notin H_0$  (probablement).

## 8.1 Exemples de tests gaussiens

On se place dans un modèle où  $X_1, \dots, X_n$  sont des gaussiennes  $N(\mu, \sigma^2)$ . Nous avons déjà vu plusieurs fois que  $\bar{X}_n \sim N(\mu, \sigma^2/n)$ .

### 8.1.1 Construction du test

On cherche à réfuter l'hypothèse selon laquelle ces variables aléatoires sont centrées ; autrement dit, on posera  $H_0 = \{0\}$ . Sous cette hypothèse, nos variables aléatoires sont donc des variables  $N(0, \sigma^2)$ .

**Supposons dans un premier temps que  $\sigma^2$  est connue.** Sous  $H_0$ , on a donc

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} \sim N(0, 1)$$

et par conséquent,  $P_0(|\bar{X}_n| < z_{1-\alpha}\sigma/\sqrt{n}) = 1 - \alpha$ . Autrement dit, sous l'hypothèse  $\mu = 0$ , on devrait observer l'événement

$$\bar{X}_n \in \left[ \pm \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \right]$$

avec probabilité élevée  $1 - \alpha$ . Si cet événement n'est pas observé, il est alors très douteux que  $\mu$  soit effectivement égal à zéro ! On pose donc

$$\text{rejeter}_\alpha = \{|\bar{X}_n| > z_{1-\alpha}\sigma/\sqrt{n}\}.$$

Le niveau de ce test est bien  $1 - \alpha$  : nous l'avons construit pour cela.

**Supposons maintenant que  $\sigma$  n'est pas connue.** En l'estimant via  $\hat{\sigma}_n$ , nous savons que (toujours sous l'hypothèse selon laquelle  $\mu = 0$ )

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} \sim \mathcal{T}(n-1).$$

On reproduit alors le raisonnement ci-dessus : comme  $\mathbb{P}(|\bar{X}_n| < t_{n-1, 1-\alpha}\sqrt{\hat{\sigma}_n}/\sqrt{n}) = \alpha$  où  $t_{n-1, 1-\alpha}$  est le quantile symétrique de  $\mathcal{T}(n-1)$ , on voit que l'événement

$$\text{rejeter}_\alpha = \{|\bar{X}_n| > t_{n-1, 1-\alpha}\hat{\sigma}_n/\sqrt{n}\}$$

est bien un test de niveau  $1 - \alpha$ .



### 8.1.2 Calcul de la puissance et hypothèse alternative

Nous n'avons pas encore eu besoin de spécifier une hypothèse alternative, mais nous allons en avoir besoin pour calculer la puissance du test. Pour commencer, on va supposer que, si  $\mu$  n'est pas nulle, alors elle ne peut être égale qu'à 1. Autrement dit,  $H_1 = \{1\}$ . Ce genre d'hypothèse alternative ne peut évidemment avoir de pertinence qu'en fonction du problème réel sous-jacent !

Sous l'hypothèse alternative, donc, nous savons que  $\bar{X}_n \sim N(1, \sigma^2)$ . La puissance du test est définie par  $1 - \beta$  où  $\beta = P_1(\text{accepter}_\alpha)$  c'est-à-dire

$$\beta = P_1(|\bar{X}_n| \leq z_{1-\alpha}\sigma/\sqrt{n}) \quad (8.1)$$

$$= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right) \quad (8.2)$$

$$= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1 \leq \bar{X}_n - 1 \leq \frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1\right) \quad (8.3)$$

$$= \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}). \quad (8.4)$$

où  $\Phi(x) = \mathbb{P}(N(0, 1) \leq x)$ . Cette expression ne peut pas plus se simplifier, mais on peut quand même la borner par  $F(-\sqrt{n}/\sigma + z_{1-\alpha})$ . Lorsque  $x$  est grand, nous avons vu (Théorème 6.1) que  $F(x) < e^{-x^2}/|x|\sqrt{2\pi}$ . Ainsi, l'erreur de première espèce est bornée par  $O(e^{-n/\sigma^2}/\sqrt{n})$ . Cela tend extrêmement vite vers 0 ; en fait, dès que  $n$  est plus grand que 10 et  $\sigma = 1$ , cette erreur est inférieure à 0.1%, donc dans ce cas le test aura une puissance supérieure à 99.9%.

Que se serait-il passé si notre hypothèse alternative n'avait pas été  $\mu = 1$  mais  $\mu = m$  pour n'importe quel  $m \neq 0$  ? Dans ce cas, on aurait eu  $H_1 = \mathbb{R} \setminus \{0\}$ . L'erreur de première espèce aurait alors été  $\beta = \sup_{m \neq 0} \beta_m$  où

$$\beta_m = P_m(\text{accepter}_\alpha).$$

On revoyant les calculs ci-dessus, on voit que

$$\beta_m = \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}).$$

En particulier,

$$\lim_{m \rightarrow 0} \beta_m = \Phi(-z_{1-\alpha}) - \Phi(-z_{1-\alpha}) = 1 - \alpha$$

par continuité de  $\Phi$  et par définition de  $z_{1-\alpha}$ . Ainsi,  $1 - \beta = \alpha$  : pour cette seconde hypothèse alternative, la puissance de notre test... est extrêmement faible.

**Cela vient du fait que notre hypothèse alternative contient des situations quasiment indiscernables de notre hypothèse nulle.** Par exemple, il est quasiment impossible de distinguer  $\mu = 0$  de  $\mu = 10^{-100}$  par exemple. Cet exemple illustre la dissymétrie entre  $H_0$  et  $H_1$ .

## 8.2 La notion de $p$ -valeur

La construction d'un test dépend du niveau de risque  $\alpha$ . Si le niveau de risque acceptable est de plus en petit, alors l'événement  $\text{rejeter}_\alpha$  devrait être de moins en moins probable. D'ailleurs,  $\text{rejeter}_0 = \emptyset$  et  $\text{accepter}_0 = \Omega$  : si l'on ne tolère aucun niveau de risque de première espèce, c'est qu'on ne veut pas rejeter l'hypothèse nulle.

Très souvent, si  $\alpha < \beta$ , on a même

$$\text{rejeter}_\alpha \subset \text{rejeter}_\beta.$$

**Définition 8.2.** La  $p$ -valeur d'une famille croissante de tests est le plus petit niveau de risque qui nous amène à rejeter l'hypothèse nulle compte tenu des observations. Formellement,

$$p_\star = \inf\{\alpha > 0 : \text{rejeter}_\alpha\} = \sup\{\alpha > 0 : \text{accepter}_\alpha\}.$$

**La  $p$ -valeur dépend des observations.** C'est une observation cruciale : la  $p$ -valeur n'est pas une propriété intrinsèque d'un test. Sur deux ensembles différents d'observations, la  $p$ -valeur ne sera pas la même en général.

**Calcul de  $p$ -valeur.** Dans de nombreux tests, la construction d'un test se fonde sur une statistique, disons  $S$ , qui sous l'hypothèse nulle suit une loi particulière (par exemple,  $\sqrt{n}\bar{X}_n/\hat{\sigma}_n \sim \mathcal{T}(n-1)$  sous l'hypothèse  $X_i \sim N(\mu, \sigma^2)$  avec  $\mu = 0$  dans le cas d'un test de Student). Si le test est de la forme  $S < q_{1-\alpha}$ , ce qui équivaut à  $F(S) < 1 - \alpha$ . La  $p$ -valeur est donnée par

$$p_\star = \sup\{\alpha > 0 : S < q_{1-\alpha}\} = \sup\{\alpha : F(S) < 1 - \alpha\} = 1 - F(S).$$

# 9 Théorie des tests simples

## 9.1 La distance en variation totale

Lorsqu'on cherche à tester une hypothèse de type  $\text{loi} = P$  contre une hypothèse de type  $\text{loi} = Q$  (c'est-à-dire, deux hypothèses simples), on en revient à chercher un événement très improbable sous la loi  $P$ , et très probable sous la loi  $Q$ . On peut se demander en toute généralité quels sont les événements pour lesquels ces probabilités diffèrent le plus, c'est-à-dire les événements  $A$  qui maximisent  $P(A) - Q(A)$ . Cela mène directement à la définition de la *variation totale*.

**Définition 9.1** (distance en variation totale). Soient  $P, Q$  deux mesures de probabilité sur un même espace  $(\mathcal{X}, \mathcal{F})$ . Leur distance en variation totale est

$$d_{\text{TV}}(P, Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A).$$

La distance en variation totale est un objet important en probabilités, qui possède de nombreuses propriétés. Parmi elles, voici les plus importantes.

1. C'est une distance sur l'espace des mesures de probabilité.
2. Elle génère une topologie plus fine que celle de la convergence en loi ; autrement dit, si  $d_{\text{TV}}(P_n, Q) \rightarrow 0$  alors  $P_n$  converge en loi vers  $Q$  mais l'inverse n'est pas vrai.

**Proposition 9.1.** Soit  $\nu$  une mesure telle que  $P$  et  $Q$  sont absolument continues<sup>1</sup> par rapport à  $\nu$ , de densités respectives  $p$  et  $q$  par rapport à  $\nu$ . Alors,  $d_{\text{TV}}(P, Q)$  est égale à chacune des quantités suivantes :

$$\begin{aligned} & \int_{\mathcal{X}} (p(x) - q(x))_+ d\nu \\ & \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| d\nu. \end{aligned} \tag{9.1}$$

De plus, notons  $E$  l'ensemble mesurable  $\{x \in \mathcal{X} : p(x) > q(x)\}$ . Alors,

$$d_{\text{TV}}(P, Q) = P(E) - Q(E). \tag{9.2}$$

L'hypothèse selon laquelle  $P, Q$  sont a.c. par rapport à  $\nu$  est toujours vérifiée pour  $\nu = (P + Q)/2$ , et n'est donc pas restrictive.

---

<sup>1</sup>Celui qui est biaisé, contrairement à  $\hat{\sigma}_n^2$ .

*Démonstration.* Pour tout événement  $A \in \mathcal{F}$ , la différence  $P(A) - Q(A)$  est égale à  $\int_A p(x) - q(x) d\nu$ , qui peut elle-même s'écrire sous la forme

$$\int_{A \cap E} (p - q) d\nu + \int_{A \cap \bar{E}} (p - q) d\nu.$$

Le second terme est négatif, puisque si  $x \notin E$  alors  $p(x) \leq q(x)$ . Ainsi,  $P(A) - Q(A)$  est plus petit que le premier terme, lequel est à son tour plus petit que  $\int_E (p - q) d\nu = P(E) - Q(E)$ . Cela montre directement l'Équation 9.2. Au passage, il est évident que

$$\int_E (p(x) - q(x)) d\nu = \int_X (p(x) - q(x))_+ d\nu,$$

ce qui montre la première égalité de l'Équation 9.1. La seconde égalité résulte de la première, puisque comme  $p$  et  $q$  sont des densités de probabilité, on a forcément  $\int (p - q)_+ = \int (p - q)_-$ .

□

Dans la suite, on supposera toujours que les diverses lois possèdent toutes une densité par rapport à une mesure de référence  $\nu$ . C'est le cas dans de très nombreux modèles — pas tous, hélas. Les lettres majuscules désigneront les mesures, tandis que les lettres minuscules désigneront leurs densités.

## 9.2 Test optimal au sens de l'affinité

L'affinité d'un test est la somme de ses erreurs de première et seconde espèce : c'est la probabilité de « se tromper » en général, quelle que soit l'hypothèse.

**Théorème 9.1.** *Soit  $\mathfrak{T}$  l'ensemble des tests possibles de l'hypothèse  $H_0 : P = P_0$  contre l'hypothèse alternative  $H_1 : P = P_1$ . Alors, le test possédant la meilleure affinité possible parmi tous les tests possibles vérifie*

$$\inf_{T \in \mathfrak{T}} \{\alpha_T + \beta_T\} = 1 - d_{TV}(P_0, P_1).$$

*En particulier, le test optimal pour l'affinité est donné par la région de rejet*

$$\text{rejeter}_\star = \{p_0(x) < p_1(x)\}.$$

*Démonstration.* Soit  $T$  n'importe quel test. Son affinité est  $P_1(\{T = 0\}) + P_0(\{T = 1\})$ . En passant au complémentaire dans le second terme, on obtient

$$1 - (P_0(\{T = 0\}) - P_1(\{T = 0\})).$$

Cette quantité est forcément plus petite que  $1 - d_{TV}(P_0, P_1)$  par la définition même de la variation totale. De plus, cette borne est atteinte en choisissant le test  $T$  donné dans l'énoncé, d'où l'égalité.

□

**Commentaire.** Le théorème précédent semble donner au problème de la construction de tests une réponse définitive : il donne le test optimal au sens de l'affinité, test qui est élémentaire et intuitif. En effet, si  $P_0, P_1$  sont les deux lois et si  $(x_1, \dots, x_n)$  est l'échantillon observé, alors on rejette l'hypothèse nulle si la probabilité de cette observation est plus grande sous  $P_1$  que sous  $P_0$  : autrement dit, si

$$\frac{p_1(x_1, \dots, x_n)}{p_0(x_1, \dots, x_n)} > 1.$$

Le terme de droite s'appelle **rapport de vraisemblance**. Pourtant, ce test ne permet pas de contrôler l'erreur de première espèce. Il peut tout à fait exister d'autres tests qui ont un niveau plus élevé. Il est donc naturel de se demander si, parmi les tests ayant un niveau fixé  $1 - \alpha$ , il existe un autre critère d'optimalité.

### 9.3 Théorème de Neyman-Pearson

On se place toujours dans un cadre où les deux lois  $P_0$  et  $P_1$  possèdent deux densités  $p_0, p_1$  par rapport à une mesure commune  $\nu$ .

**Définition 9.2.** Un *test du rapport de vraisemblance* est un test dont la région de rejet est de la forme

$$\text{rejeter} = \left\{ \frac{p_1(x)}{p_0(x)} > z \right\} \quad (9.3)$$

pour un certain  $z > 0$ .

Le test optimal au sens de l'affinité est un test de rapport de vraisemblance ( $z = 1$ ).

**Théorème 9.2** (Théorème de Neyman-Pearson). *Tout test de même niveau qu'un test du rapport de vraisemblance est moins puissant que celui-ci.*

*Démonstration.* On suppose que la région de rejet de  $T_*$  est de la forme Équation 9.3. Soit  $T$  un autre test de même niveau que  $T_*$ . La quantité

$$\int_{\mathcal{X}} (T(x) - T_*(x))(p_1(x) - zp_0(x)) d\nu$$

est forcément négative ou nulle : en effet, si  $T_*(x) = 1$ , alors  $T(x) - T_*(x) = T(x) - 1 \leq 0$ , mais  $p_1(x)$  est plus grand que  $zp_0(x)$ , donc  $(p_1(x) - zp_0(x)) \geq 0$ . De même, si  $T(x) = 0$ , alors cette fois ce terme est négatif. Dans les deux cas, la fonction dans l'intégrale est toujours le produit de deux nombres de signes opposés : elle est donc négative. Or, en développant cette intégrale, on constate qu'elle vaut aussi

$$P_1(T = 1) - P_1(T_* = 1) - zP_0(T = 1) + zP_0(T_* = 1).$$

Tout ceci n'est rien d'autre que  $\beta_* - \beta - z(\alpha - \alpha_*)$ , où  $\alpha, \beta$  désignent les deux types d'erreurs du test  $T$  et  $\alpha_*, \beta_*$  celles de  $T_*$ . Mais nous avons supposé que  $\alpha = \alpha_*$  : des deux termes ci-dessus, ne reste que le premier, à savoir  $\beta_* - \beta$ , qui est bien négatif comme demandé.

□

## 9.4 Bornes sur la variation totale

La construction du test optimal au sens de l'affinité nécessite le calcul de la distance en variation totale, laquelle peut être notoirement difficile : - d'abord, parce que la formule Équation 9.1 peut être impossible à calculer même si  $P$  et  $Q$  sont connues ; - ensuite, parce que  $Q$  elle-même peut parfois être très difficile à calculer (le calcul peut être de complexité exponentielle).

En pratique, on peut chercher à *borner* cette distance par d'autres quantités plus faciles à calculer. Parmi ces quantités, la *divergence de Kullback-Leibler* joue un rôle extrêmement important, notamment pour son lien avec le maximum de vraisemblance que nous verrons plus tard.

**Définition 9.3.** Soient  $P$  et  $Q$  deux mesures,  $P$  étant absolument continue par rapport à  $Q$ . Alors,

$$d_{\text{KL}}(P \mid Q) = \int \ln \left( \frac{dP}{dQ} \right) dP.$$

Si  $P$  n'est pas absolument continue par rapport à  $Q$ , on pose simplement  $d_{\text{KL}}(P \mid Q) = +\infty$ .

La notation  $dP/dQ$  désigne la densité de  $P$  par rapport à  $Q$ . Formellement, c'est la dérivée de Radon-Nikodym. Dans le cas de variables aléatoires continues sur  $\mathbb{R}^d$ , c'est le rapport des densités de  $P$  et de  $Q$ .

La divergence  $d_{\text{KL}}$  n'est pas une distance, et c'est pour cela qu'on l'appelle *divergence* et qu'on la note avec une barre plutôt qu'une virgule : elle n'est pas symétrique en général. Cependant, elle est toujours positive (éventuellement égale à  $+\infty$  même si  $P \ll Q$ ), et n'est nulle que si  $P = Q$ .

**Théorème 9.3** (Borne de Bretagnole-Huber-Pinsker).

$$d_{\text{TV}}(P, Q) \leq \sqrt{1 - e^{-d_{\text{KL}}(P \mid Q)}}. \quad (9.4)$$

**Remarque.** Il est facile de vérifier que  $\sqrt{1 - e^{-x}} \leq \sqrt{x}$  lorsque  $x > 0$ . Ainsi, Équation 9.4 entraîne la borne plus simple  $d_{\text{TV}} \leq \sqrt{d_{\text{KL}}}$ . La borne *classique* de Pinsker améliore légèrement ce résultat, puisqu'elle dit que  $d_{\text{TV}} \leq \sqrt{d_{\text{KL}}/2}$ .

*Démonstration.* Si  $P$  n'est pas absolument continue par rapport à  $Q$ , alors  $d_{\text{KL}}(P \mid Q) = +\infty$  et la borne demandée est vraie. Sinon, on note  $\rho$  la densité de  $P$  par rapport à  $Q$ , de sorte que  $d_{\text{KL}}(P \mid Q) = -\int \ln \rho(x) dP$ . On définit ensuite  $v = (\rho - 1)_+$  et  $w = (\rho - 1)_-$ , de sorte que  $vw$  vaut toujours 0, et donc  $(1 + v)(1 - w) = 1 - w + v = \rho$ . En particulier,  $d_{\text{KL}}(P \mid Q)$  vaut

$$\int (-\ln(1 + v)) dP + \int (-\ln(1 - w)) dP.$$

Or, les deux fonctions  $x \mapsto -\ln(1 + x)$  et  $x \mapsto -\ln(1 - x)$  sont concaves sur leurs ensembles de définition. Ainsi, l'inégalité de Jensen entraîne d'une part

$$\int (-\ln(1 + v)) dP \leq -\ln \left( 1 + \int v dP \right)$$

et d'autre part

$$\int (-\ln(1-w))dP \leq -\ln\left(1 - \int wdP\right).$$

Or, la formule Équation 9.1 montre que  $\int vdP = d_{TV}(P, Q)$ , et de même pour  $\int wdP$ . En additionnant les deux inégalités ci-dessus, on obtient Alors

$$-d_{KL} \leq -\ln((1 + d_{TV})(1 - d_{TV}))$$

soit  $-d_{KL} \leq -\ln(1 - d_{TV}^2)$ , c'est-à-dire Équation 9.4.

□

## 10 Tests du $\chi_2$

Les tests du  $\chi_2$  sont une vaste famille de tests qui visent, pour la plupart, à tester si un échantillon (souvent discret) a été généré par une loi précise ; on parle parfois de test d'ajustement.

### 10.1 Loi multinomiale

Soit  $\Omega$  un ensemble fini à  $k$  éléments, disons pour simplifier  $\{1, \dots, k\}$ . On notera  $S_k$  l'ensemble des lois de probabilités sur cet ensemble, c'est-à-dire les  $k$ -uplets  $\mathbf{p} = (p_1, \dots, p_k)$  de nombres positifs dont la somme vaut 1. On observe  $n$  tirages indépendants et identiquement distribués selon une même loi sur  $\Omega$ . Formellement, le modèle statistique est donné par  $(\mathbf{p}^{\otimes n} : \mathbf{p} \in S_k)$ .

On note  $N_j$  le nombre d'observations égales à  $j$ . Le vecteur  $N = (N_1, \dots, N_k)$  suit alors une loi multinomiale de paramètres  $n$  et  $\mathbf{p}$ , donnée par

$$\mathbb{P}(N = (n_1, \dots, n_k)) = \frac{n!}{n_1! \dots n_k!} \prod_{j=1}^k p_j^{n_j},$$

où  $\sum_{j=1}^k n_j = n$ . Cette loi sera notée  $\text{Mult}(n, \mathbf{p})$ .

**Théorème 10.1.** *Soit  $N \sim \text{Mult}(n, \mathbf{p})$ . Alors,  $\sqrt{n}(\frac{N}{n} - \mathbf{p})$  converge en loi lorsque  $n \rightarrow \infty$  vers  $\mathcal{N}(0, \Sigma)$ , où*

$$\Sigma = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top. \quad (10.1)$$

*Démonstration.* On commence par remarquer que  $N = \sum_{i=1}^n Z_i$ , où  $Z_i = (\mathbf{1}_{X_i=1}, \dots, \mathbf{1}_{X_i=k})$ . Les  $Z_i$  sont iid de moyenne  $\mathbf{p}$ . Les covariances des entrées  $i$  et  $j$  de  $Z_k$  sont données par

$$\mathbb{E}[\mathbf{1}_{X_k=i} \mathbf{1}_{X_k=j}] - p_i p_j = \delta_{i,j} p_i - p_i p_j,$$

ce qui montre que la matrice de covariance des  $Z_k$  est Équation 10.1. Comme  $\mathbb{E}[|Z_1|^2] = 1$ , il suffit d'appliquer le TCL. □

**Remarque.** On considère que cette approximation normale est correcte dès que  $\mathbb{E}[N_j]$  est plus grand que 5 pour tout  $j$ .



## 10.2 Test d'adéquation

Le test du  $\chi^2$  d'adéquation consiste à tester l'hypothèse nulle

$$H_0 : \mathbf{p} = \mathbf{p}_0 \quad (10.2)$$

contre l'hypothèse alternative

$$H_1 : \mathbf{p} \neq \mathbf{p}_0, \quad (10.3)$$

pour une valeur de  $\mathbf{p}_0$  fixée au préalable.

**Exemple 10.1.** On peut se demander si, dans la langue courante, les 21 lettres de l'alphabet ont à peu près la même probabilité d'apparaître comme première lettre d'un mot. Cela revient à tester si  $\mathbf{p}_0 = (1/26, \dots, 1/26)$ , hypothèse qui est évidemment fausse.

Qu'en est-il des 9 chiffres ? On peut vouloir tester si, dans n'importe quel document (journal, site internet, article scientifique), ces 9 chiffres apparaissent à peu près uniformément en tant que premier chiffre d'un nombre. Cela reviendrait à tester  $\mathbf{p}_0 = (1/9, \dots, 1/9)$ .

Ce n'est pas le cas et cette hypothèse est très fréquemment réfutée : le premier chiffre significatif d'un nombre est bien plus souvent 1 ( $\approx 30\%$  des cas) que 9 ( $\approx 5\%$  cas). Ce phénomène s'appelle *loi de Benford*.

Le théorème précédent dit que  $\sqrt{n}(\frac{N}{n} - \mathbf{p}) \approx N(0, \Sigma)$ . Ainsi, sous  $H_0$ ,  $\text{diag}(1/\sqrt{\mathbf{p}_0})\sqrt{n}(\frac{N}{n} - \mathbf{p}_0)$  converge en loi vers  $N(0, I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T)$  (en supposant tout de même que pour tout  $p_{j,0} > 0$  pour chaque  $j$ ). En manipulant légèrement cette expression, on obtient la forme classique donnée dans la définition suivante.

**Définition 10.1** (Contraste du  $\chi_2$ ). Dans le contexte ci-dessus, le *contraste* du  $\chi_2$  associé à la loi  $\mathbf{p}$  est la statistique

$$D_n(\mathbf{p}) = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}.$$

Pour faire des tests, il suffit donc de trouver la loi asymptotique de cette statistique.

**Théorème 10.2.** Sous l'hypothèse nulle Équation 10.2, la statistique  $D_n$  converge en loi vers  $\chi_2(k-1)$ . De plus, sous l'hypothèse alternative Équation 10.3,  $D_n$  tend vers  $+\infty$  presque sûrement.

*Démonstration.* La matrice  $\pi_0 = I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T$  est la matrice de projection sur l'orthogonal du vecteur  $\sqrt{\mathbf{p}_0}$ . Le théorème de Cochran (Théorème 14.3) implique alors que la statistique  $D_n$ , qui est égale à

$$\left| \text{diag}(1/\sqrt{\mathbf{p}_0})\sqrt{n} \left( \frac{N}{n} - \mathbf{p}_0 \right) \right|^2, \quad (10.4)$$

converge en loi vers la norme de la projection d'une gaussienne  $N(0, I_k)$  sur un sous-espace de dimension  $k-1$ , c'est-à-dire une loi  $\chi_2(k-1)$ . Sous l'hypothèse alternative, il y a au moins un  $p_i$  non nul tel que  $p_i \neq (p_0)_i$ . Ainsi, Équation 10.4 est plus grand que  $n(N_i/n - (p_0)_i)^2/p_i$ , mais  $N_i$  suit une loi  $\text{Bin}(n, p_i)$  et donc  $N_i/n$  converge en probabilité vers  $p_i$ . Il est alors clair que  $n(N_i/n - (p_0)_i)$  converge vers  $+\infty$ .

□

Un test de niveau  $1 - \alpha$  pour l'hypothèse Équation 10.2 est alors donné par la région de rejet

$$\{D_n(\mathbf{p}_0) > \kappa_{k-1, 1-\alpha}\}$$

où  $\kappa_{k-1, 1-\alpha}$  est le quantile d'ordre  $1 - \alpha$  d'une  $\chi^2(k-1)$ . Si  $\mathbf{p}$  n'est pas égal à  $\mathbf{p}_0$ , le contraste  $D_n$  tend vers l'infini, donc le test sera forcément dans la zone de rejet : si l'hypothèse alternative est simple, la puissance du test tend donc vers 1.

### 10.3 Test d'indépendance

Les tests du  $\chi^2$  d'indépendance sont omniprésents en sciences humaines. Dans ces tests, on observe des variables aléatoires qui sont des *couples* à valeur dans deux espaces discrets ; disons, pour simplifier, que cet espace est  $\Omega = \{1, \dots, k\} \times \{1, \dots, h\}$ . Les observations  $(x_i, y_i)$  sont des réalisations d'une variable aléatoire  $(X, Y)$ . Ici, le modèle statistique sera donc  $(\mathbf{p}^{\otimes n} : \mathbf{p} \in S_{k,h})$ , où  $S_{k,h}$  est l'ensemble des  $\mathbf{p} = (p_{i,j}, i \in \{1, \dots, k\}, j \in \{1, \dots, h\})$  qui sont des lois de probabilité.

Si  $\mathbf{p}$  est la loi de  $(X, Y)$ , alors  $X$  et  $Y$  sont indépendantes si et seulement si  $\mathbf{p}$  peut s'écrire sous la forme  $p_{i,j} = p_i^x p_j^y$ , où  $\mathbf{p}^x \in S_k$  et  $\mathbf{p}^y \in S_h$ . L'ensemble de ces lois sera noté  $I_{k,h}$  (« I » pour « Indépendant »). Les tests d'indépendance visent à tester l'hypothèse nulle

$$H_0 : \mathbf{p} \in I_{k,h} \tag{10.5}$$

contre l'hypothèse alternative

$$H_1 : \mathbf{p} \notin I_{k,h}.$$

**Exemple 10.2.** On récolte des données sur le groupe socio-professionnel (GSP) et le genre. Chaque observation correspond à une personne, possédant deux attributs : **genre**, valant 0 ou 1, et **GSP**, valant l'une des 6 groupes définis par l'INSEE (Agriculteur, artisan, cadre, etc.). Le test ci-dessus vise à déterminer si les deux modalités sont indépendantes, c'est-à-dire si la proportion d'hommes et de femmes dans chaque groupe ne diffère pas significativement en fonction du groupe.

La procédure pour effectuer un tel test nécessite plusieurs étapes.

Si  $\mathbf{p}$  était effectivement la loi de deux variables indépendantes  $\mathbf{p}^x$  et  $\mathbf{p}^y$ , alors ses marginales seraient précisément  $\mathbf{p}^x$  et  $\mathbf{p}^y$ , que l'on pourrait facilement estimer. Pour chaque  $i$  et chaque  $j$ , les estimateurs  $\hat{\mathbf{p}}^x$  et  $\hat{\mathbf{p}}^y$  définis par

$$\hat{p}_i^x = \frac{\sum_{j=1}^h N_{i,j}}{n}$$

et

$$\hat{p}_j^y = \frac{\sum_{i=1}^k N_{i,j}}{n}$$

sont effectivement des estimateurs sans biais et convergents des quantités  $p_i^x, p_j^y$ . De plus, sous l'hypothèse nulle,  $\hat{p}_i^x \hat{p}_j^y$  serait effectivement un estimateur convergent de  $p_{i,j}$ .

De plus, si  $\mathbf{p}$  était effectivement de la forme  $\mathbf{p}^x \mathbf{p}^y$ , alors la moyenne théorique des éléments de classe  $(i, j)$  serait  $n \hat{p}_i^x \hat{p}_j^y$ . Cette quantité, notée  $\tilde{N}_{i,j}$ , s'appelle *effectif théorique*. Nous pouvons maintenant construire la statistique qui nous servira à tester tout cela.

**Définition 10.2** (Statistique de Pearson). La statistique de Pearson est définie par

$$C_n = \sum_{i=1}^k \sum_{j=1}^k \frac{(N_{i,j} - \check{N}_{i,j})^2}{\hat{N}_{i,j}}.$$

Cette statistique possède une loi limite connue, encore en vertu du théorème de Cochran. Noter que la statistique de Pearson possède une expression alternative,

$$C_n = \sum \sum \frac{n(\hat{p}_{i,j} - \hat{p}_i^x \hat{p}_j^y)^2}{\hat{p}_i^x \hat{p}_j^y}.$$

**Théorème 10.3** (Loi de la statistique de Pearson). *Sous l'hypothèse nulle Équation 10.5,  $C_n$  converge en loi vers*

$$\chi_2((k-1)(h-1)).$$

*De plus, pour n'importe quelle loi  $\mathbf{p}_1$  qui n'est pas dans  $I_{k,h}$ ,  $C_n \rightarrow +\infty$  presque sûrement.*

*Démonstration.* C'est une conséquence un peu plus technique du théorème de Cochran.

□

Tout cela permet encore une fois d'obtenir des tests très efficacement : en abrégant  $\kappa_{1-\alpha} = \kappa_{(k-1)(h-1), 1-\alpha}$ , on obtient que  $\mathbb{P}(C_n > \kappa_{1-\alpha}) \rightarrow \alpha$ . Ainsi, la région de rejet

$$\{C_n > \kappa_{1-\alpha}\}$$

fournit un test de niveau asymptotique  $1-\alpha$ . La seconde partie du théorème dit que si la véritable loi sous-jacente n'est effectivement pas la loi de deux variables indépendantes, alors ce test sera systématiquement rejeté — autrement dit, si l'hypothèse alternative est simple, la puissance de ce test tend vers 1.

# 11 Exercices

## 11.1 Questions

- Quelles sont les erreurs du test consistant à toujours accepter l'hypothèse nulle ?
- Quelles sont les erreurs du test consistant à toujours refuser l'hypothèse nulle ?
- Montrer que la distance en variation totale entre deux mesures de densités  $p, q$  peut aussi s'écrire  $\int (p/q - 1)_+ dp$ .
- Montrer que si  $d_{\text{KL}}(P_n | Q) \rightarrow 0$ , alors  $P_n$  converge en loi vers  $Q$ .
- Calculer la distance en variation totale entre deux lois de Bernoulli de paramètres respectifs  $p$  et  $q$ .
- Calculer la distance en variation totale entre une loi  $\text{Bin}(n, p)$  et une loi  $N(\mu, \sigma^2)$ .
- Soient  $P, Q$  deux mesures. Montrer que  $d_{\text{KL}}(P^{\otimes n} | Q^{\otimes n}) = n d_{\text{KL}}(P | Q)$ .

## 11.2 Tests élémentaires

Pour tous les cas suivants, il faut savoir réaliser rapidement un test puissant, voire même optimal au sens qu'il vous plaira.

- Tester  $\mu = \mu_0$  contre  $\mu = \mu_1$  dans un échantillon  $N(\mu, \sigma^2)$  lorsque  $\mu$  est connu.
- Même question lorsque  $\mu$  est inconnu.
- Soient  $X_1, \dots, X_n$  un échantillon iid  $N(\mu_1, \sigma_1^2)$  et  $Y_1, \dots, Y_m$  ( $m$  et  $n$  ne sont pas forcément égaux) un échantillon iid de loi  $N(\mu_2, \sigma_2^2)$ . Tester  $\sigma_1 = \sigma_2$  lorsque  $\mu_1$  et  $\mu_2$  sont connues.
- Même question lorsque  $\mu_1$  et  $\mu_2$  ne sont pas connues.
- Donner la forme d'un test sur la valeur de  $p$  pour une réalisation d'une loi  $\text{Bin}(n, p)$  et calculer son niveau asymptotique quand  $n \rightarrow \infty$ .
- Donner la forme d'un test sur la valeur de  $\lambda$  dans un échantillon de  $n$  variables aléatoires de Poisson de paramètre  $\lambda$ .

## 11.3 Exercices

**Exercice 11.1.** Soient  $X_1, \dots, X_n$  des variables indépendantes de loi  $\chi_2(p)$ . On cherche à tester l'hypothèse nulle  $p = 1$  contre l'hypothèse alternative  $p = 2$ .

1. Écrire le test ayant la plus grande affinité possible.
2. Essayer de calculer l'affinité de ce test ; si ce n'est pas possible, essayer de la borner.

**Exercice 11.2.** Soient  $X_1, \dots, X_n$  des variables indépendantes de loi  $N(0, \sigma^2)$ . Proposer un test de niveau  $\alpha$  de l'hypothèse  $\sigma^2 = 1$  contre l'hypothèse  $\sigma^2 = 1 + \varepsilon$ , et estimer sa puissance. Comment varie-t-elle en fonction de  $n$  et de  $\varepsilon$  ?

**Exercice 11.3.** Soient  $X_1, \dots, X_n$  des variables indépendantes de même loi  $P$ . On cherche à tester l'hypothèse nulle  $P = N(0, 1)$  contre l'hypothèse alternative  $P = \mathcal{T}(n)$ .

1. Donner le test optimal au sens de l'affinité.
2. Donner un autre test, de niveau  $1 - \alpha$ , et calculer sa puissance.
3. Comparer ces deux tests, en particulier dans le régime où  $n$  est grand.

**Exercice 11.4.** Montrer que le nombre de lancers nécessaire pour distinguer une pièce équilibrée ( $p = 1/2$ ) d'une pièce légèrement déséquilibrée ( $p_1 = 1/2 + \varepsilon$ ) est d'ordre  $1/\varepsilon^2$ .

**Exercice 11.5.** On note  $p$  la probabilité qu'un enfant né vivant soit un garçon. On suppose que les enfants sont de sexe indépendants, et que cette probabilité est la même pour toutes les grossesses.

1. Il y a eu en France métropolitaine en 2015  $n = 760\,421$  naissances, dont 389 181 garçons. Tester l'hypothèse  $p = \frac{1}{2}$  contre l'alternative pertinente.
2. En 1920, il y a eu 838 137 naissances dont 432 044 garçons. Tester l'hypothèse  $p_{2015} = p_{1920}$ .

**Exercice 11.6.** Soient  $X_1, \dots, X_n$  i.i.d de loi  $N(\theta, 1)$ , où  $\theta$  est un paramètre réel.

1. Donner un intervalle de confiance pour  $\theta$  au niveau de risque 5% de la forme  $[\hat{\theta}_n, +\infty[$ .
2. En déduire un test de niveau 5% pour les hypothèses  $H_0 : \theta = 0$  et  $H_1 : \theta > 0$ .
3. Donner le modèle de l'expérience statistique. Donner l'expression du test de rapport de vraisemblance  $T$  pour les hypothèses  $H_0 : \theta = 0$  et  $H_1 : \theta = \mu$ , où  $\mu > 0$ . Quel test retrouve-t-on?
4. Construire le test de rapport de vraisemblance au niveau 5% pour les hypothèses  $H_0 : \theta = 0$  et  $H_1 : \theta > 0$ .

**Exercice 11.7** (Test sur des lois uniformes). On se donne  $X_1, \dots, X_n$  iid de loi  $\mathcal{U}(0, \theta)$ , et on note  $M_n = \max_{j=1, \dots, n} X_j$ .

1. Écrire la fonction de répartition de  $M_n$ , puis en déduire un test  $T$  de niveau  $1 - \alpha$  pour les hypothèses  $H_0 : \theta = 1$  contre  $H_1 : \theta < 1$ .
2. Donner le test du rapport de vraisemblance pour les hypothèses  $H_0 : \theta = 1$  contre  $H_1 : \theta = \theta_0$ , où  $\theta_0 < 1$ . Calculer sa puissance.
3. On cherche à tester  $H_0 : \theta = 1$  contre  $H_1 : \theta < 1$ . Comme la seconde hypothèse est composite, on ne peut pas directement appliquer le test du rapport de vraisemblance ; à la place, on utilise un test du *maximum* de vraisemblance, qui est de la forme

$$\frac{\sup_{\theta < 1} \rho_\theta(x_1, \dots, x_n)}{\rho_1(x_1, \dots, x_n)} > z$$

où  $\rho_\theta$  est la densité d'un échantillon iid de lois  $\mathcal{U}[0, \theta]$ . Calculer le supremum dans cette expression, et en déduire la région de rejet.

4. Montrer que la puissance de  $T$  vaut  $1 - \alpha$ .
5. En utilisant la même technique, construire le test du rapport de maximum de vraisemblance pour les hypothèses  $H_0 : \theta = 1$  contre  $H_1 : \theta > 1$ , noté  $T'$ , au niveau  $1 - \alpha$ . Calculer sa puissance.
6. Donner un test de niveau  $1 - \alpha$  pour  $H_0 : \theta = 1$  contre  $H_1 : \theta > 1$ , plus puissant que  $T'$  pour n'importe quel  $\theta > 1$ .

**Exercice 11.8.** Une réalisation d'une variable aléatoire  $X \sim \text{Bin}(20, p)$  donne  $X = 8$ .

1. Proposer un test du rapport de vraisemblance de l'hypothèse nulle  $p = p_0 = 1/2$  contre l'hypothèse alternative  $p = p_1 = 1/3$ . Donner l'expression de la  $p$ -valeur du test.
2. On tire des variables aléatoires iid de Bernoulli jusqu'à obtenir 8 succès. Écrire la loi de probabilité du nombre de lancers  $N$ .
3. Il se trouve que le nombre de lancers nécessaires pour cela était  $N = 20$ . Proposer un test du rapport de vraisemblance de l'hypothèse nulle  $p = p_0 = 1/2$  contre l'hypothèse alternative  $p = p_1 = 1/3$ . Donner l'expression de la  $p$ -valeur du test.
4. Pourquoi les deux  $p$ -valeurs sont-elles différentes, alors que les deux tests sont identiques ?

**Exercice 11.9** (Test d'adéquation du  $\chi^2$ ). On lance 60 fois un dé et on obtient les résultats suivants :

Face $k$	1	2	3	4	5	6
Effectif $N_k$	10	13	8	12	9	8

Le dé est-il bien équilibré ? À titre indicatif, le quantile d'une loi  $\chi^2(5)$  d'ordre 95% est 11.07.

**Exercice 11.10** (Test d'indépendance du  $\chi^2$ ). On cherche à savoir si les variables « être riche » et « être heureux » sont indépendantes. On interroge un grand échantillon de personnes à ce sujet, et l'on récolte les données suivantes :

	riche	pauvre
heureux	344	700
triste	257	705

L'argent fait-il le bonheur ?

## 12 Moindres carrés

Les modèles *linéaires* sont les modèles les plus simples dans lesquels on raisonne en termes d'*entrées* et de *sorties*. Dans ces modèles, on dispose de variables  $x_i$ , dites *explicatives*, et de variables  $y_i$ , dites à *expliquer*, et l'on suppose qu'il existe une fonction inconnue  $f$  telle que

$$y_i \approx f(x_i),$$

et que l'on voudrait estimer. Les modèles linéaires consistent à supposer que  $f$  est affine. Les modèles plus complexes, comme les réseaux de neurones, placent  $f$  dans des classes plus riches. L'objectif de la méthode des moindres carrés est de trouver la meilleure approximation de  $f$  possible dans la classe des fonctions affines.

### 12.1 Ajustement affine en une dimension.

On suppose qu'il existe entre les données  $x_i$  et  $y_i$  une relation de la forme  $y_i = \alpha + \beta x_i$  où  $\alpha, \beta$  sont deux nombres réels. Ici,  $\approx$  signifie que la relation n'est pas parfaite : peut-être par exemple que les sorties sont bien égales à  $\alpha + \beta x_i$ , mais que les observations  $y_i$  ont été polluées par du bruit ou des erreurs. Nous verrons cela plus tard.

Pour l'instant, nous voulons chercher les meilleurs  $\alpha, \beta$  possibles. On calcule la distance entre le nuage de points  $(x_i, y_i)$  et la droite d'équation  $y = \alpha + \beta x$ . Cette distance au carré est donnée par

$$L(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

On cherchera donc les  $(\hat{\alpha}, \hat{\beta})$  qui minimisent cette distance. La fonction  $L$  est manifestement une fonction quadratique qui tend vers  $+\infty$  lorsque  $(\alpha, \beta) \rightarrow \infty$ , par conséquent cette fonction possède un unique minimiseur  $(\hat{\alpha}, \hat{\beta})$ , et ce minimiseur est le seul point en lequel les dérivées partielles s'annulent (*conditions de premier ordre*) :  $\partial_\alpha L(\hat{\alpha}, \hat{\beta}) = 0$  et  $\partial_\beta L(\hat{\alpha}, \hat{\beta}) = 0$ . Or,

$$\partial_\alpha L(\alpha, \beta) = \sum_{i=1}^n (\alpha + \beta x_i - y_i)$$

$$\partial_\beta L(\alpha, \beta) = \sum_{i=1}^n x_i (\alpha + \beta x_i - y_i).$$

Les conditions de premier ordre deviennent donc  $n\alpha + \beta(x_1 + \dots + x_n) - (y_1 + \dots + y_n) = 0$  soit encore  $\alpha + \beta\bar{x} - \bar{y} = 0$ , et d'autre part  $\alpha(x_1 + \dots + x_n) + \beta(x_1^2 + \dots + x_n^2) - (x_1 y_1 + \dots + x_n y_n) = 0$ , soit  $\alpha\bar{x} + \beta\overline{x^2} - \overline{xy} = 0$ , où  $\bar{x}$  est la moyenne des carrés des  $x_i$  et  $\overline{xy}$  la moyenne des  $x_i y_i$ . En résolvant ces équations, on trouve d'abord  $\alpha$  puis  $\beta$  :

$$\beta = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \quad \alpha = \bar{y} - \beta\bar{x}.$$

Le coefficient  $\beta$  n'est rien d'autre que la covariance empirique des  $x_i$  et des  $y_i$ , normalisé par la variance empirique des  $x_i$ .

L'inégalité de Cauchy-Schwartz dit que  $|\overline{xy} - \bar{x}\bar{y}| \leq \tilde{\sigma}_x \tilde{\sigma}_y$ , où l'on a noté

$$\tilde{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2$$

l'estimateur naïf de la variance<sup>1</sup>. L'inégalité n'est une égalité que si  $x$  et  $y$  sont effectivement colinéaires, c'est-à-dire si  $y_i = \hat{\alpha} + x_i \hat{\beta}$  pour tous les  $i$ . La qualité de l'ajustement affine est donc bien mesurée par la quantité

$$r^2 = \frac{\overline{xy} - \bar{x}\bar{y}}{\tilde{\sigma}_x \tilde{\sigma}_y}.$$

## 12.2 Moindres carrés généraux

Dans le cadre général, les variables explicatives ne sont pas de dimension 1 mais  $d$ . On notera  $x = (x_1, \dots, x_d)$  un élément de  $\mathbb{R}^d$ ; les variables explicatives seront alors  $x_1, \dots, x_n$ . On cherchera des nombres  $\theta_i$  tels que  $y_i$  est aussi proche que possible de

$$\theta_1 x_{i,1} + \dots + \theta_d x_{i,d} = \langle \theta, x_i \rangle = \theta^\top x_i.$$

**Remarque : où est passée la constante ?** Dans l'équation ci-dessus, on a l'impression que le terme constant, qui correspondait à  $\alpha$  dans l'exemple en dimension 1, a disparu. Ce n'est pas le cas : intégrer la constante au modèle revient à considérer que la variable constante égale à 1 fait partie des variables explicatives. En pratique, cela revient à poser, par exemple,  $x_{i,1} = 1$  pour tout  $i$ .

On pose  $X$  la matrice  $n \times d$  dont la  $i$ -ème ligne est  $x_i$ . La matrice dont les lignes sont composées des  $\theta^\top x_i$  n'est autre que la matrice  $X\theta$ . La distance entre le nuage de points  $(X, Y)$  et la droite d'équation  $Y = X\theta$  est alors  $|Y - X\theta|$ . On pourrait reproduire la méthode analytique ci-dessus pour trouver les paramètres optimaux, à savoir

$$\hat{\theta} = \arg \min_{\theta} |Y - X\theta|^2. \quad (12.1)$$

Cependant, une interprétation géométrique simplifie la tâche : le  $\hat{\theta}$  qui minimise Équation 12.1 est précisément celui qui garantit que  $X\hat{\theta}$  est la projection orthogonale de  $Y$  sur le sous-espace vectoriel  $\mathcal{V}_X = \{X\theta : \theta \in \mathbb{R}^d\}$ .

**Théorème 12.1.** Si  $d \leq n$  et si  $X$  est de rang  $d$ , alors

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y. \quad (12.2)$$

*Démonstration.* La projection orthogonale sur le sous-espace vectoriel engendré par les colonnes d'une matrice  $X$  est la matrice  $X(X^\top X)^{-1} X^\top$ . Ainsi, la projection de  $Y$  sur ce sous-espace est  $X(X^\top X)^{-1} X^\top Y$ , et c'est aussi (par définition de l'argmin)  $X\hat{\theta}$ . Comme  $X$  est injective en vertu du théorème du rang, on en déduit le résultat. □

---

<sup>1</sup>Celui qui est biaisé, contrairement à  $\hat{\sigma}_n^2$ .



Le vecteur  $\hat{Y} = X\hat{\theta}$  est appelé *vecteur des prédictions*. Le vecteur  $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\theta}$  est appelé *vecteur des résidus*. Si ce dernier est nul ou très petit, cela veut dire que les  $Y$  sont presque parfaitement des fonctions linéaires des  $X$ .

L'expression Équation 12.2 possède de nombreuses expressions alternatives. Parmi elles, on pourra noter que

$$\hat{\theta} = \theta + \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i. \quad (12.3)$$

# 13 Modèles linéaires

## 13.1 Modèle gaussien

À ce stade, nous n'avons fait aucune hypothèse statistique ni probabiliste sur le modèle : les  $x_i, y_i$  étaient donnés tels quels. Le *modèle linéaire gaussien* avec variables explicatives  $x_1, \dots, x_n$  exogènes consiste à supposer que  $Y = X\theta + \varepsilon$ , où  $\varepsilon = N(0, \sigma^2 I_n)$ . Formellement, le modèle est indexé par  $\theta$  et  $\sigma^2$ , et donné par

$$P_{\theta, \sigma^2} = N(X\theta, \sigma^2 I_d).$$

Dans ce modèle, la loi de l'estimateur Équation 12.2 est connue. Par simplicité, je note  $H = X(X^\top X)^{-1}X^\top$  la matrice de projection orthogonale sur l'espace vectoriel engendré par les colonnes de  $X$ , qui est de dimension  $d$ .

**Théorème 13.1.** *Sous le modèle linéaire gaussien  $P_{\theta, \sigma^2}$ ,*

$$\hat{\theta} \sim N(\theta, \sigma^2 (X^\top X)^{-1}),$$

$$\frac{|\hat{\varepsilon}|^2}{\sigma^2} \sim \chi_2(n - d),$$

*et ces deux variables aléatoires sont indépendantes.*

*Démonstration.* Ce n'est rien de plus que le théorème de Cochran appliqué à notre problème : en effet, le vecteur des résidus est la projection orthogonale de  $Y$  sur le sous-espace orthogonal à l'espace des colonnes de  $X$ .

□

La variable aléatoire  $|\hat{\varepsilon}|^2$  est souvent appelée *Somme des Carrés des Résidus* (SCR). Le théorème précédent implique que

$$\hat{\sigma}_n^2 = \frac{|\hat{\varepsilon}|^2}{n - d}$$

est un estimateur sans biais de  $\sigma^2$ . et ces deux variables aléatoires sont indépendantes. En particulier,  $(n - d)\hat{\sigma}_n^2/\sigma^2 \sim \chi_2(n - d)$ .

## 13.2 Modèle linéaire général

Il est possible de ne pas faire d'hypothèses gaussiennes sur le modèle. Dans ce cadre plus général, on supposera que  $Y = X\theta + \varepsilon$ , où les  $\varepsilon_i$  sont iid, centrés, et de même variance  $\sigma^2$  — sous cette dernière hypothèse, on parle de modèle *homoscédastique*.

Sous ces hypothèses,  $\hat{\theta}$  est toujours un estimateur sans biais de  $\theta$  : cela se voit directement en prenant l'espérance de l'Équation 12.3. De plus, la loi de  $\theta$  n'est plus gaussienne, mais  $\theta$  est asymptotiquement normal sous des hypothèses supplémentaires sur  $X$ . Ces hypothèses sont les suivantes.

On suppose que les variables explicatives  $x_i$  vérifient la propriété suivante :

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \Sigma_x, \quad (13.1)$$

où  $\Sigma_x$  est inversible. Cette propriété s'écrit aussi  $X^\top X/n \rightarrow \Sigma_x$ .

**Théorème 13.2.** *Sous les hypothèses précédentes,  $\sqrt{n}(\hat{\theta} - \theta)$  converge en loi lorsque  $n \rightarrow \infty$  vers  $N(0, \sigma^2 \Sigma_x^{-1})$ .*

*Démonstration.* Rappelons que  $\hat{\theta}$  peut s'écrire  $\theta + (X^\top X/n)^{-1} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i$ . Pour montrer que  $\sqrt{n}(\hat{\theta} - \theta)$  converge, il suffit donc de démontrer que

$$\sqrt{n} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i \quad (13.2)$$

converge en loi vers  $N(0, \Sigma_x^2)$  : comme le terme  $(X^\top X/n)^{-1}$  converge vers  $\Sigma_x^{-1}$  par hypothèse, la limite de  $\sqrt{n}(\hat{\theta} - \theta)$  sera bien  $N(0, \Sigma_x^{-1} \Sigma_x \Sigma_x^{-1}) = N(0, \Sigma_x^{-1})$ . Malheureusement, on ne peut pas directement appliquer le TCL classique à l'Équation 13.2 : en effet, les variables aléatoires  $X_i = x_i \varepsilon_i$  ne sont pas identiquement distribuées. On doit pour cela appliquer une version plus générale du TCL, dite *de Lindeberg*, que je détaillerai plus tard.

□

# 14 Outils gaussiens

## 14.1 Vecteurs gaussiens

Un vecteur aléatoire  $X$  à valeurs dans  $\mathbb{R}^n$  est un vecteur gaussien de loi  $N(\mu, \Sigma)$  si sa densité est donnée par

$$\frac{1}{(2\pi \det(\Sigma))^{n/2}} \exp \left\{ -\frac{1}{2} \langle x - \mu, \Sigma^{-1}(x - \mu) \rangle \right\}.$$

Ici, le vecteur  $\mu \in \mathbb{R}^n$  est appelé *moyenne* de  $X$  parce que

$$\mathbb{E}[X] = \mu.$$

La matrice  $\Sigma$ , qui est toujours supposée symétrique et à valeurs propres strictement positives (on dit *définie positive*), est appelée *matrice de covariance*, parce que

$$\mathbb{E}[(X - \mu)(X - \mu)^\top] = \Sigma.$$

De même que la transformée de Fourier d'une variable gaussienne réelle  $N(m, \sigma^2)$  est égale à  $e^{imt - \frac{t^2 \sigma^2}{2}}$ , la transformée de Fourier d'un vecteur gaussien  $N(\mu, \Sigma)$  est

$$\mathbb{E}[e^{i\langle t, X \rangle}] = \exp \left\{ i\langle t, \mu \rangle - \frac{\langle (t - \mu), \Sigma(t - \mu) \rangle}{2} \right\}.$$

### Théorème 14.1.

1. Toute fonction linéaire d'un vecteur gaussien est encore un vecteur gaussien. Si  $M$  est une matrice et  $X \sim N(\mu, \Sigma)$ ,

$$MX \sim N(M\mu, M\Sigma M^\top).$$

2. Si le couple  $(X, Y)$  forme un vecteur gaussien, alors  $X$  et  $Y$  sont indépendants si et seulement si leur covariance  $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top]$  est la matrice nulle.

## 14.2 Conditionnement gaussien

Soit  $(X, Y)$  un vecteur gaussien de dimension  $n + m$ , avec  $X \in \mathbb{R}^n$  et  $Y \in \mathbb{R}^m$ . On peut écrire sa moyenne  $\mu$  en deux blocs

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

et sa covariance  $\Sigma$  en quatre blocs

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}$$

où, par symétrie,  $\Sigma_{2,1} = \Sigma_{1,2}^\top$ .

**Théorème 14.2.** *La loi de  $X$  conditionnellement à  $Y$  est une loi gaussienne de moyenne*

$$\mu_1 + \Sigma_{1,2} \Sigma_{2,2}^{-1} (X_2 - \mu_2)$$

*et de covariance*

$$\Sigma_{1,1} - \Sigma_{2,1} \Sigma_{2,2}^{-1} \Sigma_{1,2}.$$

L'expression *loi conditionnelle* signifie ici que, pour toute fonction test  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ , l'espérance conditionnelle  $\mathbb{E}[\varphi(X) | Y]$ , qui est une variable aléatoire  $Y$ -mesurable, vaut

$$\frac{1}{(2\pi \det(S^{-1}))^{n/2}} \int_{\mathbb{R}^n} \varphi(x) e^{-\frac{\langle X - (\mu_1 + \Sigma_{1,2} \Sigma_{2,2}^{-1} (X_2 - \mu_2)), S^{-1} (X_1 - (\mu_1 + \Sigma_{1,2} \Sigma_{2,2}^{-1} (X_2 - \mu_2))) \rangle}{2}} dx$$

où  $S = \Sigma_{1,1} - \Sigma_{2,1} \Sigma_{2,2}^{-1} \Sigma_{1,2}$ .

### 14.3 Théorème de Cochran

**Théorème 14.3** (Théorème de Cochran). *Soit  $X \sim N(0, I_d)$  et soient  $E_1, \dots, E_k$  des sous-espaces orthogonaux de  $\mathbb{R}^n$  tels que  $\mathbb{R}^n = \oplus_{j=1}^k E_j$ . On note  $\pi_j(X)$  la projection orthogonale de  $X$  sur  $E_j$ . Alors, la famille  $(\pi_j(X))_{j=1, \dots, k}$  est une famille de vecteurs gaussiens indépendants. De plus,*

$$|\pi_j(X)|^2 \sim \chi_2(\dim E_j).$$

*Démonstration.* Pour chaque  $E_i$ , notons  $d_i$  sa dimension et choisissons-lui une base orthonormale  $e_1^i, \dots, e_{d_i}^i$ . La projection orthogonale de  $X$  sur  $E_i$  est  $\pi_i(X) = \sum_{t=1}^{d_i} \langle X, e_t^i \rangle e_t^i$ . Notons  $X_t^i = \langle X, e_t^i \rangle$ . Le vecteur  $(X_t^i)$ , qui contient bien  $d_1 + \dots + d_k = n$  éléments, est une fonction linéaire du vecteur gaussien centré  $X$ , donc est lui-même un vecteur gaussien centré. Calculons sa covariance : de façon générale, si  $e, f$  sont deux vecteurs fixés,

$$\mathbb{E}[\langle X, e \rangle \langle X, f \rangle] = \sum_{i,j} e_i f_j \text{Cov}(X_i, X_j) = \langle e, f \rangle.$$

Il est alors immédiat que la matrice de covariance du vecteur gaussien  $(X_t^i)$  n'est autre que la matrice  $(\langle e_t^i, e_s^j \rangle)$ , c'est-à-dire l'identité puisque les  $(e_t^i)$  forment une base orthonormale de  $\mathbb{R}^n$ . Il en résulte les deux points de l'énoncé.

1. Les  $\pi_i(X)$  sont des variables indépendantes, puisque fonctions linéaires de variables indépendantes entre elles.
2. La formule de Parseval dit que

$$|\pi_i(X)|^2 = \sum_{t=1}^{d_i} |X_t^i|^2$$

ce qui est bien une somme de  $d_i$  gaussiennes  $N(0, 1)$  indépendantes, donc une  $\chi_2(d_i)$ .

□

## 14.4 Loi de Fisher

Si  $N$  est un vecteur et  $X, Y$  sont les projections de  $N$  sur deux sous-espaces vectoriels orthogonaux, le théorème de Cochran dit que  $X$  et  $Y$  sont des lois du  $\chi_2$  indépendantes de paramètres  $p = \dim E, q = \dim F$ . La loi de leur rapport  $X/Y$  est connue et fréquemment utilisée en statistiques.

**Théorème 14.4.** *Soient  $X, Y$  deux variables aléatoires indépendantes, de lois respectives  $\chi_2(p)$  et  $\chi_2(q)$ . La loi du rapport  $(X/p)/(Y/q)$  s'appelle loi de Fisher de paramètres  $p, q$ . Sa densité est donnée par*

$$f_{p,q}(x) = \frac{\mathbf{1}_{x>0}}{Z_{p,q}} \frac{\left(\frac{px}{px+q}\right)^{\frac{p}{2}} \left(1 - \frac{px}{px+q}\right)^{\frac{q}{2}}}{x} \quad (14.1)$$

où la constante  $Z_{p,q}$  est  $B(p/2, q/2)$ , c'est-à-dire

$$Z_{p,q} = \int_0^1 u^{\frac{p}{2}-1} (1-u)^{\frac{q}{2}-1} du.$$

Le calcul est facile, puisque les lois du  $\chi_2$  ont une densité connue donnée par Équation 6.4. Soit  $\varphi$  une fonction test et soit  $F = (X/p)/(Y/q)$ . Alors,  $\mathbb{E}[\varphi(F)]$  vaut

$$\frac{1}{C_p C_q} \int_0^\infty \int_0^\infty \varphi\left(\frac{uq}{vp}\right) e^{-\frac{u}{2} - \frac{v}{2}} u^{\frac{p}{2}-1} v^{\frac{q}{2}-1} du dv$$

avec  $C_n = 2^{n/2} \Gamma(n/2)$ . Dans l'intégrale en  $v$ , on pose  $x = uq/vp$ , de sorte que l'intégrale ci-dessus devient

$$\frac{(p/q)^{\frac{p}{2}}}{C_p C_q} \int_0^\infty \varphi(x) x^{\frac{p}{2}-1} \int_0^\infty e^{-\frac{vpx}{2q} - \frac{v}{2}} v^{\frac{p}{2}-1} v^{\frac{q}{2}} dv dx.$$

On reconnaît dans l'intégrale en  $v$  une fonction Gamma, égale à

$$\frac{\Gamma(p/2 + q/2)}{\left(\frac{px+q}{2q}\right)^{\frac{p+q}{2}}}.$$

L'espérance  $\mathbb{E}[\varphi(F)]$  vaut donc

$$\frac{(p/q)^{p/2} \Gamma\left(\frac{p+q}{2}\right)}{C_p C_q (2q)^{\frac{p+q}{2}}} \int_0^\infty \varphi(x) \frac{x^{\frac{p}{2}-1}}{(px+q)^{\frac{p+q}{2}}} dx.$$

En simplifiant, on trouve exactement la densité donnée par Équation 14.1.

# **15 Tests linéaires**

## **15.1 Significativité d'un coefficient**

## **15.2 Significativité de la régression**

## **15.3 Test d'une contrainte**

## 16 Exercices

### 16.1 Questions

auinesanea



# 17 Modèles exponentiels

Cours 8-9-10

## **18 Maximum de vraisemblance**

## **19 Information et entropie**

# 20 Estimation de densité

Cours 11-12

**Et après ?**

# Références

- [Statistics done Wrong](#)
- [The earth is round \(p<.05\)](#)
- [Statistiques mathématiques en action](#), pour ceux qui vont passer l'agrégation.
- [Introduction à l'économétrie](#) de Brigitte Dormont est un excellent livre, écrit en français, sur les modèles linéaires.
- En anglais, la référence sur les modèles linéaires est [Econometric analysis](#) de Greene.
- [Méthodes statistiques](#) de Philippe Tassi est un bon livre général.
- [All of statistics](#) de Larry Wasserman est un ouvrage de référence.