ENS-Cachan 1D2/3
2016-2017

Hugo Harari-Kermadec
hugo.harari@ens-cachan.fr

Statistique Appliquée

Notes partielles du Cours

12 décembre 2016

Avertissement préliminaire

Ces notes de cours ne sont pas exhaustives et ne se substituent en aucun cas au cours.
La rédaction de ces notes doit beaucoup à Léonard Moulin, Lucie Le Roland et Paul-Antoine Chevalier.

Plateforme moodle

Les documents sont disponibles sur
`https://elearn.ens-cachan.fr/course/view.php?id=532`
Les D3 doivent s'inscrire sur
`https://elearn.ens-cachan.fr/course/view.php?id=604`

Bibliographie

Anderson, David Ray et al. (2015). *Statistiques pour l'économie et la gestion*. De Boeck. bib 31 STA.

Armatte, Michel (2010). « Statut de la Dispersion : de l'erreur à la variabilité ». In : *Journ@l électronique d'Histoire des Probabilités et de la Statistique*.

Bourdieu, Pierre (1973). « L'opinion publique n'existe pas ». In : *Temps modernes* 318, p. 1292–1309.

Bruno, Isabelle, Julien Prévieux & Emmanuel Didier (2014). *Statactivisme, comment lutter avec des nombres*. Zones. bib 31 STA.

Denglos, Grégory (2016). *Statistiques et probabilités appliquées*. Presses universitaires de France. bib 519.3 DEN.

Desrosières, Alain (2008). *L'argument statistique ; 1 : Pour une sociologie historique de la quantification ; 2 : Gouverner par les nombres*. français. Paris, France : Mines ParisTech-Les Presses. bib 31 DES.

Gould, S. J. (1983). *La mal-mesure de l'homme*. Le livre de Poche. bib 57 GOU.

MESR (2012). *Les étudiants en classes préparatoires aux grandes écoles. Rentrée 2011*. Note d'information 12.02, p. 7.

Moulin, L. (2015). « Polycopié de Statistique ».

Quetelet, Adolphe (1835). *Sur l'homme et le développement de ses facultés ou essai de physique sociale*. Bachelier.

Réau, Jean-Philippe & Gérard Chauvat (2006). *Probabilités & statistiques : résumé des cours, exercices et problèmes corrigés, QCM*. Paris : A. Colin. Cote bibliothèque cachan 519.2 REA.

Table des matières

1 Introduction **9**

A La statistique pré-moderne . . . 9
1 La préhistoire de la statistique . . . 9
2 Formalisation et probabilités . . . 9

B Anthropométrie . . . 10
1 Des sciences naturelles… . . . 10
2 … à une physique sociale . . . 10
3 Les exemples controversés de la « race » et du quotient intellectuel (QI) . . . 11

C Les statistiques, quantifier pour objectiver ? . . . 13
1 La statistique publique et la statistique industrielle . . . 13
2 Deux visions des statistiques . . . 13
3 Quantification instituante : auto-réalisation et performativité 14

D Notations . . . 15

E Probabilités : définitions . . . 16
1 Variables aléatoires réelles . . . 16
2 Fonction de répartition . . . 16
3 Fonction de densité . . . 17
4 Quantile . . . 18

F Moments . . . 18
1 Espérance . . . 18
2 Moments simples, moments centrés . . . 18

G Couple de variables aléatoires . . . 20
1 Distribution jointe . . . 20
2 Distributions marginales . . . 20
3 Distributions conditionnelles . . . 21
4 Indépendance . . . 21
5 Covariance, corrélation . . . 21

H Probabilités : théorèmes . . . 22
1 Modes de convergences . . . 22
2 Lois des grands nombres et théorème central limite . . . 23
3 Extensions . . . 24
4 Lois associées à la loi Gaussienne . . . 25

2 Estimation par substitution **27**
A Estimation de l'espérance 29
B Estimation de la variance 33

3 Intervalles de confiance et tests **37**
A Intervalles de confiance 37
1 Définition 37
2 Espérance d'un échantillon gaussien, variance connue 38
3 Espérance d'un échantillon gaussien, variance inconnue 40
4 Espérance d'un échantillon non gaussien 41
5 Taille d'échantillon et précision 42
B Premiers tests 43
1 Définitions 43
2 Test bilatéral pour l'espérance d'une gaussienne 46
2.1 Cas où la variance est connue 47
2.2 Cas où la variance est inconnue 48
2.3 Cas où la loi est inconnue 48
3 Test bilatéral de la variance d'une loi normale 49
4 Tests de comparaison de deux espérances 50
4.1 Cas où les variances sont connues 50
4.2 Cas où les variances sont inconnues, mais supposées égales (test de Student) 50
5 Tests de comparaisons de deux variances (test de Fisher) . . . 51

4 Contraste du χ^2 **55**
A Test d'adéquation du χ^2 55
1 Adéquation à une loi 55
2 Adéquation à une famille de lois 58
3 Test d'indépendance du χ^2 60

5 Statistique mathématique **65**
A Cadre formel 65
1 Vraisemblance 65
2 Statistique exhaustive 68
B Ordres sur les estimateurs 69
1 Comparaison d'estimateurs : biais et variance 69
2 Estimateur efficace : l'inégalité de Fréchet, Darmois, Cramer, Rao 70
3 Modèles de forme exponentielle 76

6 Maximum de vraisemblance **79**
A Maximum de vraisemblance unidimensionnel 79
1 Définitions 79
2 Propriétés à distance finie 81
3 Propriétés asymptotiques 82
B Maximum de vraisemblance multidimensionnel 82

1 Cadre multidimensionnel 82
2 Borne de Cramer-Rao 83
3 Maximum de vraisemblance multidimensionnelle 83
C Rapport de vraisemblance 84
1 Intervalles de confiance 84
2 La méthode de Neyman et Pearson 84
3 Test de la moyenne d'une loi normale 85

Chapitre 1

Introduction

A RELIRE

Les outils statistiques sont inventés dans un contexte historique qu'ils impactent en retour. Si dans leur cadre moderne de branche des mathématiques, la recherche en statistiques jouit d'une forte autonomie, elle est aussi guidée et marquée par ses applications et ses conséquences sociales. Théorie et applications sont donc intimement liées dans le champ des statistiques.

A La statistique pré-moderne

1 La préhistoire de la statistique

Dès la période égyptienne, on note la présence de nombreux **recensements de populations**. À partir du XIII^e^ siècle se développe l'**estimation des risques** maritimes à Venise. Ces premiers balbutiements de la théorie du risque donnent naissance aux premières assurances, dont l'activité se fonde sur des théories probabilistes. Le mot « statistique » n'est cependant véritablement consacré qu'au XVII^e^ siècle en Allemagne, où il désigne la science de l'État (*Staat* en allemand) : c'est alors un outil pour l'**armée** et les **impôts**. Par la suite, l'essor de la statistique a permis de nombreuses découvertes et a trouvé de multiples applications. Par exemple, l'outil statistique permet de déterminer qu'il y a plus de bébés garçons que de bébés filles (107 pour 100), il permet d'estimer le prix d'une rente viagère… Au XVIII^e^ siècle apparaît la première méthodologie d'inférence, à travers l'estimation d'une population par « **coefficient multiplicateur** » (on multiplie le nombre de naissance annuelle par 27,5 pour obtenir la population à partir des registres paroissiaux).

2 Formalisation et probabilités

La statistique se formalise à l'aune de la **théorie des probabilités**. Dès 1713, Bernoulli formule la **loi des grands nombres**. C'est aussi à cette époque que l'on formalise la **loi normale**, à travers le problème des moindres carrés de Gauss et le théorème central limite de Laplace. Bayes et Laplace développent ensuite respectivement la statistique dite bayésienne et la statistique dite inférentielle : on parle alors de « **probabilité inverse** ». De nos jours, la statistique bayésienne trouve

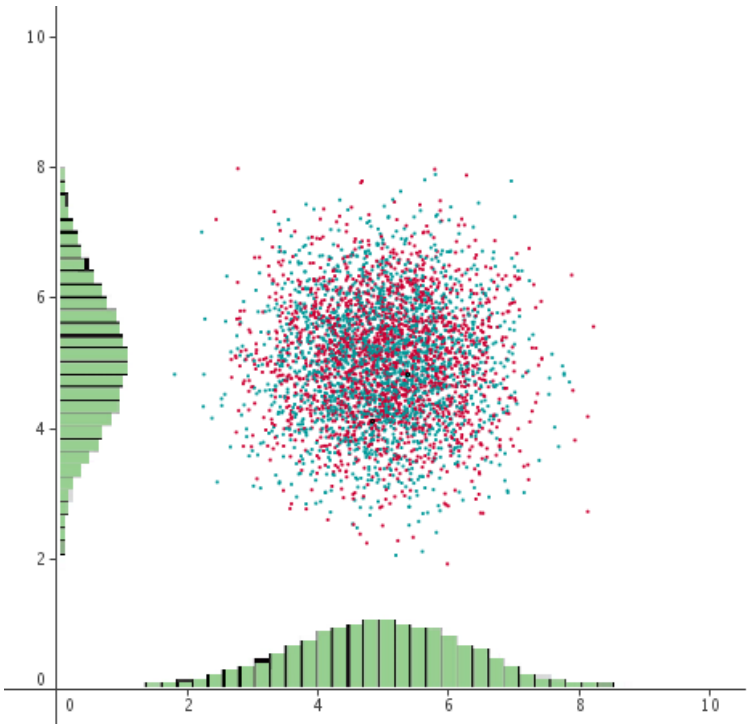


FIGURE 1.1 – Les différentes positions observées d'une étoile se répartissent selon une loi normale

notamment une application dans les boîtes mail, à travers la gestion des filtres anti-spam.

B Anthropométrie

1 Des sciences naturelles...

C'est l'**astronomie** qui constitue le point de départ du développement des statistiques en sciences. En effet, lors d'observations astronomiques différentes, on observe qu'un astre n'apparaît pas au même point. Alors, l'étoile bouge-t-elle ? Une telle hypothèse semble farfelue. La seule hypothèse crédible est donc d'admettre qu'il existe des **erreurs de mesure**. Dès lors, prendre la **position moyenne** permet, en vertu de la loi des grands nombres, d'estimer au mieux la véritable position de l'étoile. Et, si les données observées sont « normales », on observe alors empiriquement qu'elles **se répartissent selon une loi de Gauss (courbes en cloche ci-contre)**.

2 ... à une physique sociale

QUATELET (1796-1874) développe un observatoire de la population et cherche à appliquer les techniques statistiques de l'astronomie en démographie et en physique sociale. Pour QUATELET, « *un être nouveau mais bien réel lui aussi, l'homme moyen, existe en amont des individus tous différents les uns des autres. Il constitue la cause*

constante de la distribution des tailles observées », écrit Alain Desrosières. Quetelet défend fermement ce concept d'**homme moyen**, esquissant déjà de l'idée d'un patrimoine génétique commun.

Galton (1822-1911), fondateur de la statistique mathématique à travers des outils comme la régression, la médiane et les quartiles, va lui plus loin avec sa **pensée eugéniste**. Du quantifiable, Galton dévie ainsi vers un usage héréditaire appliqué au contrôle des naissances et à l'analyse factorielle comme réification du QI. Galton a en outre fondé la première revue de statistiques, *Biometrika*. La pensée statistique de Galton a essaimé : **coefficients de corrélation, écart-type, ajustement du χ^2** (Pearson, père), **vraisemblance, plan d'expérience** (Fisher), **tests et intervalles de confiance** (Pearson, fils) deviennent rapidement des outils statistiques utilisées couramment en physique sociale.

3 Les exemples controversés de la « race » et du quotient intellectuel (QI)

Prenons ainsi l'exemple de la « race » comme variable. L'expérience STAR (Tennessee) de 1995 visait à déterminer quelle classe parmi une classe normale, une petite classe et une classe aidée réussissait le mieux. Anticipant sur les résultats, certains parents auraient pu **fausser l'expérience** en choisissant de placer leur enfant dans une classe spécifique, ruinant la part d'aléatoire dans l'expérience. Il fallait donc s'assurer de la bonne répartition des élèves au sein des trois groupes. Aux États-Unis, le critère « neutre » pour assurer la bonne répartition fut… la race des élèves (chose qui aurait été bien évidemment impossible en France, la loi interdisant les **statistiques ethniques**). La répartition des élèves au sein des classes est celle de la table 1.1.

	Blanc	**Noir**	**Autre**
Classe normale	1 354	636	10
Petite classe	1 183	540	10
Classe aidée	1 332	676	7

Table 1.1 – Répartition des élèves par race et type de classe, expérience STAR (Tennessee), Mosteller, 1995

La question qui se pose alors est la suivante : **les races existent-elles vraiment ?** Selon la décision n° 2007-557 DC du Conseil constitutionnel du 15 novembre 2007, « *si les traitements nécessaires à la conduite d'études sur la mesure de la diversité des origines des personnes, de la discrimination et de l'intégration peuvent porter sur des données objectives, ils ne sauraient, sans méconnaître le principe énoncé par l'article 1^er^ de la Constitution, reposer sur l'origine ethnique ou la race* ». La question n'est donc pas tranchée.

L'expérience de la table 1.2 prétend avoir démontré la **supériorité de l'homme blanc**. Au delà du fait que le volume crânien ne détermine pas l'intelligence, des erreurs de mesures ont été commises par Morton et ses assistants, ce qui a poussé Gould à retravailler ses résultats dans la table 1.3, en utilisant des grenailles de plomb, moins déformables que les grains de poivre :

Race	Volume moyen (cm³)
Caucasien (blanc)	1 426
Mongol (asiatique)	1 360
Américain (indien)	1 344
Malaise (océanien)	1 327
Éthiopien (noir)	1 278

TABLE 1.2 – Mesures de volumes crâniens par graines de poivre par MORTON dans *Crania Americana*, 1839

Race	Volume originel (cm³)	Volume corrigé (cm³)
Caucasien (blanc)	1 426	1 401
Mongol (asiatique)	1 360	1 426
Américain (indien)	1 344	1 409
Malaise (océanien)	1 327	1 393
Éthiopien (noir)	1 278	1 360

TABLE 1.3 – Mesure de volumes crâniens par grenaille de plomb corrigées des effets de taille et sexe par GOULD dans *La mal-mesure de l'homme*, 1983

Tout aussi controversés que les statistiques ethniques, les **tests de quotient intellectuel** (QI) présentent eux aussi des biais. Ainsi, dans une population donnée, le QI moyen doit s'élever à 100. Or les tests de QI par pays sont de nos jours mal conçus, car ils ne sont **pas adaptés aux populations du pays**. Les tests de Richard LYNN (voir figure 1.2) sont dès lors biaisés. Peut-on alors mesurer la race et le QI par pays avec des intentions pures ? Telles sont les questions éthiques que soulèvent les problèmes statistiques.

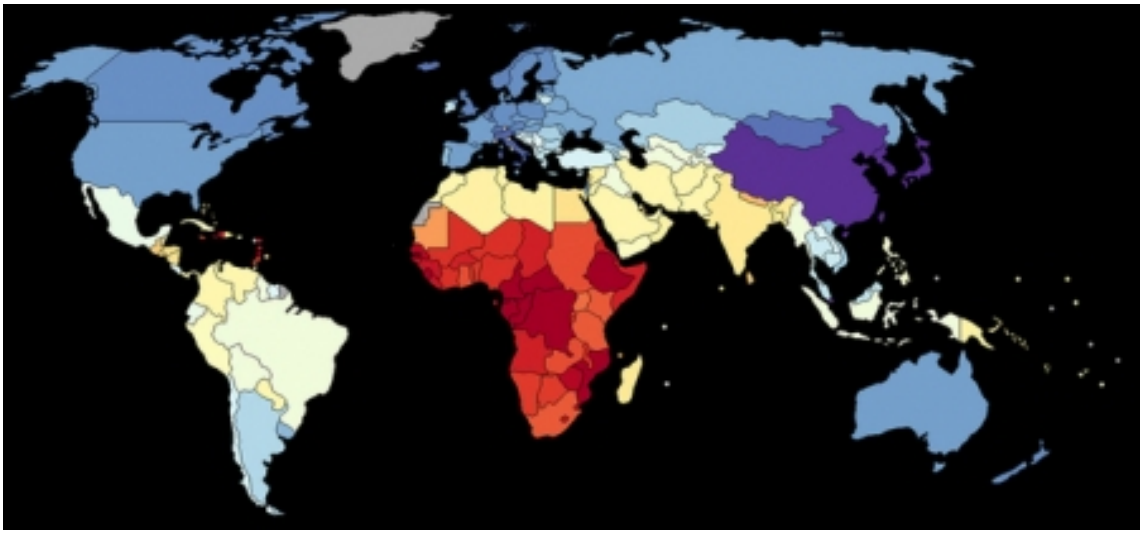


FIGURE 1.2 – Évalution du quotient intellectuel moyen par pays, Richard LYNN, *IQ and global inequality*, 2006. *Légende* – Rouge : QI ≤ 65 ; fraise : QI $\in]65, 70]$; orange : QI $\in]75, 80]$; jaune pâle : QI $\in]80, 85]$; jaune : QI $\in]85, 90]$, bleu ciel : QI $\in]90, 95]$; bleu : QI $\in]95, 100]$; bleu foncé : QI $\in]100, 105]$; violet : QI $\in]105, 110]$ gris : données non disponibles.

C Les statistiques, quantifier pour objectiver ?

Les données ne sont *in fine* que des **constructions sociales** ; elles ne vont jamais d'elles-mêmes. En particulier, elles peuvent être produites par l'État ou un autre pouvoir à des **fins de gouvernement ou de gestion**. La statistique n'est donc jamais neutre. Les données statistiques peuvent également être récoltées et travaillées par des **chercheurs**, pour éclairer un questionnement. Dans les deux cas, une **intention** est à l'origine d'une construction.

1 La statistique publique et la statistique industrielle

Alain Desrosières a étudié la typologie de la **statistique publique**, qu'on traduit dans la table 1.4.

Type d'État	Vision de l'économie	Mode d'action	Type de statistique
Ingénieur, XVII^e siècle	Hiérarchisée et rationnelle	Planification et optimisation	Échanges interindustriels
Libéral, XVIII^e siècle	Marché libre	Lois anti-truts	Information économique sur les parts de marchés
Providence, fin XIX^e siècle	Spécificité du travail	Législation socio-économique	Statistique de l'emploi et des inégalités
Keynésien, années 1940	Régulation anti-crise	Monnaie et budget	Comptabilité nationale et macroéconomique
Néolibéral, années 1990	Marché libre, finance	Incitations anti-truts	Benchmarking et évalution

Table 1.4 – Une typologie de la statistique publique d'après Alain Desrosières

Au XX^e siècle se développe la **statistique industrielle**. Dès les années 1900 à 1940, Fisher s'intéresse aux **rendements agricoles**. Par la suite, Student alias William Gosset, brasseur chez Guinness, étudie des **phénomènes statistiques en disposant de peu de données**. Entre les années 1940 et 1970, Shewart, Deming et Weibull permettent le développement de l'**industrie militaire** puis de l'**industrie civile**. Les applications pratiques des résultats établis par ces derniers pontes de la statistique vont peut être converger avec le nouveau management public. En effet, la statistique publique version néolibérale (Eurostat en Europe) et la statistique industrielle vont sans doute être utilisées conjointement dans le cadre du **management des établissements publics**.

2 Deux visions des statistiques

Il existe deux visions des statistiques. Tout d'abord, avec la **vision réaliste**, on « *postule qu'il existe pour chaque individu des "variables" de valeur unique réelles,*

indépendantes de tout jugement et des modalités de questionnement » (tome II, chapitre 6, page 138). C'est le postulat adopté par QUATELET lorsqu'on mesure le tour de poitrine d'un conscrit. En revanche, l'approche statistique peut également être **conventionnelle** : « *la trace de l'acte initial de codage reste visible et importance, soit dans une perspective de dénonciation, soit parce que l'agrégat est directement articulé sur une forme d'action collective. Dans ce cas, l'intention de l'agrégation et de l'addition reste présente dans l'usage de la statistique présentée* » (tome I, chapitre 9, page 161). On peut donc faire dire différentes choses aux statistiques en fonction des conventions adoptées, mais il ne s'agit pas de faire dire aux statistiques n'importe quoi non plus ! La statistique procède certes d'une construction, mais le fait par exemple, pour un étudiant, d'être boursier ou non, est une convention claire et distincte.

Réalisme et conventionnalisme ne s'opposent cependant pas totalement. Ces deux visions s'opposent au **trucage**, à travers le rôle des **indicateurs**. Ainsi, en statistiques, « *la "réalité" n'est pas rejetée comme métaphysique, comme elle l'est par* PEARSON*, mais elle n'est pas non plus directement mesurable comme l'est le réel des sciences de la nature* »(tome II, chapitre 11, page 191). Ces indicateurs peuvent donc fixer des critères réels à l'aune desquels effectuer un travail statistique. Ainsi, le **seuil de pauvreté** est déterminé par rapport à la distribution des niveaux de vie de l'ensemble de la population. Par exemple, la France, à travers les travaux de l'**INSEE**, privilégie ainsi un seuil de pauvreté à 60 % de la médiane des niveaux de vie, mais publie des taux de pauvreté selon d'autres seuils (40 %, 50 % ou 70 %). Cependant, la multiplication des indicateurs pose parfois un **risque de réification**, comme en témoigne l'exemple du QI.

Le risque de réification n'est pas l'unique danger auquel sont soumises les variables statistiques. En effet, dans leur deux approches, celles-ci sont aussi exposées aux risques de la **double conscience**, c'est-à-dire un glissement sémantique dû à un changement de rôle social ou d'interlocuteur. Ainsi, CAHUE et CARCILLO dans la *Revue économique* de 2007 répondent à la question « Que peut-on attendre de l'interdiction de licencier pour améliorer la compétitivité des entreprises ? » Dans la partie empirique, le chômage est considéré comme recherche d'emploi tandis que, dans la partie théorique, le chômage est pris comme temps de loisirs : un même mot désigne alors deux réalités bien différentes.

3 Quantification instituante : auto-réalisation et performativité

Les statistiques, par leur travail même, produisent des effets. Ces effets peuvent être de deux genres. Le premier, c'est l'**auto-réalisation**. Ainsi, « *les classements reproduisent et renforcent la stratification qu'ils prétendent mesurer* » (ESPELAND et SAUDER, 2007). Dès lors, comme le constate Valérie PÉCRESSE en 2013, « *le classement de Shanghaï [est] certes critiquable, mais [...] on ne peut s'en abstraire et nous devons donc gagner des places* ». Le second effet des statistiques peut être **performatif**. Ainsi, les classements internationaux comme celui de Shanghaï produisent une séparation entre « universités mondiales » et établissements locaux.

D Notations

Alphabet grec

symbol	Symbol	prononciation	Signification usuelle
θ	Θ	théta	paramètre
ω	Ω	oméga	événement
μ		mu	espérance
σ		sigma	écart-type
α		alpha	niveau de probabilité
β		béta	autre niveau de probabilité
δ		delta	masse de Dirac
ρ		rho	corrélation linéaire
ϵ		epsilon	petite valeur, erreur
φ	Φ	fi	f_X et F_X pour $X \sim \mathcal{N}(0;1)$
λ	Λ	lambda	paramètre d'une loi de Poisson
	$\sum_{i=1}^{n}$	Sigma	somme pour i allant de 1 à n
	$\prod_{i=1}^{n}$	Pi	produit pour i allant de 1 à n

Pour X un vecteur ou une matrice, X' est sa transposée

$\xrightarrow[n\to+\infty]{\mathbb{P}}$: convergence en probabilité

$\xrightarrow[n\to+\infty]{\mathcal{L}}$: convergence en loi

$\xrightarrow[n\to+\infty]{p.s.}$: convergence presque sûre

E Probabilités : définitions

1 Variables aléatoires réelles

Définition 1 (Variable aléatoire) *Une variable aléatoire réelle X est une façon d'assigner un nombre réel à chaque élément ω de l'univers Ω. Formellement, c'est une fonction X de Ω dans $\mathbb{R}$ qui, à chaque $\omega \in \Omega$ associe une valeur $X(\omega) \in \mathbb{R}$.*

$$\begin{aligned} X : \ & \Omega \to \mathbb{R} \\ & \omega \mapsto X(\omega) = x \end{aligned}$$

On note $\mathcal{X}$ l'image de Ω par X, c'est-à-dire l'ensemble des valeurs atteintes (c'est donc une partie de $\mathbb{R}$).

Exemple Les exemples canoniques sont le dé ou la pièce, pour lesquels Ω est l'ensemble des conditions physiques qui décrivent le lancer et déterminent le résultat. Elles sont inaccessibles, et donc toute l'idée des probabilités est de s'intéresser à ce qu'on peut quand même dire sur le résultat sans savoir ce qui se passe dans Ω. $\mathcal{X}$ est $[|1;6|]$ pour le dé, {Pile, Face} ou $\{0;1\}$ pour la pièce.

Définition 2 (Mesure de probabilité) *La mesure de probabilité $\mathbb{P}_X$ associée à la variable aléatoire est X associe à chaque partie de $\mathbb{R}$ (en fait pas nécessairement toute, mais toutes celles qui sont utiles) un réel entre 0 et 1 :*

$$\begin{aligned} \mathbb{P}_X : \ & \mathcal{P}(\mathbb{R}) \to [0;1] \\ & B \mapsto \mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega, X(\omega) \in B\}) = \mathbb{P}(X^-(B)) \end{aligned}$$

où X^- est une forme de fonction inverse.

2 Fonction de répartition

Définition 3 (Fonction de répartition) *La fonction de répartition d'une variable aléatoire X est l'application F de $\mathcal{X}$ dans $[0,1]$ définie par :*

$$F_X(x) = \Pr(X \leq x)$$

La fonction de répartition $F_X(x)$ d'une variable aléatoire X représente la probabilité que X soit inférieure ou égale à x.

Définition 4 (Variable aléatoire discrète, variable aléatoire continue) *Une variable aléatoire X est continue si sa fonction de répartition $F_X(x)$ est continue. Elle sera discrète si $F_X(x)$ est une fonction en escalier.*

Proposition 1

- F_X est croissante.
- $\Pr(a < X \leq b) = F_X(b) - F_X(a)$.
- F_X est continue à droite.
- F_X a une limite à gauche. Elle est continue à gauche dans le cas des variable aléatoire continues.

- $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$.

Démonstration *La croissance provient directement du fait que si $a < b$, l'inclusion $]-\infty; a] \subset]-\infty; b]$ entraine par croissance des probabilités*

$$F_X(a) = \mathbb{P}\left(X \in]-\infty; a]\right) \leq \mathbb{P}\left(X \in]-\infty; b]\right) = F_X(b)$$

La relation de Chasles provient de l'additivité des probabilités sur les ensembles disjoints suivants :

$$]-\infty; a] \bigcup]a : b] =]-\infty; b]$$

La continuité à droite vient du fait que pour tout a,

$$F_X(a + 1/n) - F_X(a) = \Pr(a < X \leq a + 1/n) \xrightarrow[n\to\infty]{} \Pr(\emptyset) = 0$$

puisque $a \notin]a; a + 1/n]$ et que pour tout $x > a$ $\exists N,\ n \geq N \Rightarrow a + 1/n < x$ et donc la limite des $]a; a + 1/n]$ ne contient pas non plus x. Donc elle ne contient rien.

Á gauche par contre, on n'a pas exactement une situation symétrique à cause de la définition de F où l'intervalle est semi-fermé à droite :

$$F_X(a) - F_X(a - 1/n) = \Pr(a - 1/n < X \leq a) \xrightarrow[n\to\infty]{} \Pr(\{a\})$$

En toute généralité, cette probabilité n'est pas nécessairement nulle, mais elle est finie (inférieure à 1) et donc $F_X(a - 1/n)$ a bien une limite :

$$F_X(a - 1/n) \xrightarrow[n\to\infty]{} F_X(a) - \Pr(\{a\})$$

Il y a continuité pour tout a dans le cas où la variable est continue puisque la probabilité correspondante est alors nulle pour tout singleton.

3 Fonction de densité

Définition 5 (Fonction de densité d'une variable aléatoire continue) *Soit une variable aléatoire continue X. La fonction de densité $f_X(x)$ est telle que :*

$$F(x) = \int_{-\infty}^{x} f_X(u)du$$

en d'autres termes :

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

Proposition 2

- $\forall x, f_X(x) \geq 0$.
- $P(X \in]x_1, x_2[) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(u)du$.
- $P(X = x_0) = \int_{x_0}^{x_0} f_X(u)du = 0$.
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$.

4 Quantile

Définition 6 (Quantile) *Le quantile d'ordre $q \in [0,1]$ de la variable aléatoire X est la valeur x_q telle que $\Pr(X \leq x_q) = q$, c'est-à-dire telle que $F_X(x_q) = q$.*

F Moments

1 Espérance

Définition 7 (Espérance) *L'espérance de la variable aléatoire discrète X, notée $\mathbb{E}[X]$ est définie par :*

$$\mathbb{E}[X] = \sum_{k=1}^{K} x_k \Pr(X = x_k)$$

Dans le cas d'une variable aléatoire continue, on a :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx$$

Définition 8 (Espérance d'une transformation de X) *Pour une fonction g « simple », on note*

$$\mathbb{E}[g(X)] = \sum_{k=1}^{K} g(x_k) \Pr(X = x_k) \quad \textit{pour une v.a. discrète}$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx \quad \textit{pour une v.a. continue}$$

Proposition 3

- $\mathbb{E}(a) = a$.
- $\mathbb{E}(aX) = a\mathbb{E}(X)$.
- $\mathbb{E}(X + a) = \mathbb{E}(X) + a$.

2 Moments simples, moments centrés

Définition 9 (Moments simples) *Pour tout entier $r \geq 1$, le moment simple d'ordre r de la variable aléatoire X est donné par :*

$$M_X^r = \mathbb{E}[X^r]$$

Définition 10 (Moments centrés) *Pour tout entier $r \geq 2$, le moment centré d'ordre r de la variable aléatoire X est donné par :*

$$M_X'^r = \mathbb{E}[(X - \mathbb{E}[X])^r]$$

Le moment centré d'ordre 3 est appelé coefficient de dissymétrie (skewness) : s'il est nulle, c'est que la distribution est symétrique. Celui d'ordre 4 est appelé coefficient d'aplatissement (kurtosis) ; on le compare usuellement à celui de la gaussienne et on parle d'excès de kurtosis lorsque la différence avec la gaussienne est positive. La distribution a alors des queues épaisses (la densité décroit doucement lorsqu'on s'éloigne de la médiane).

Définition 11 (Variance) *La variance* $\mathrm{Var}(X)$ *d'une variable aléatoire* X *est son moment centré d'ordre 2 :*

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

Remarque 1 Dans la suite, nous noterons souvent celle-ci σ_X^2.

Démonstration

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\right] \\
&= \mathbb{E}\left[X^2\right] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

La variance est une mesure servant à caractériser la dispersion d'une distribution. Elle indique de quelle manière la variable aléatoire se disperse autour de son espérance. Une variance de zéro signale que tous les tirages seront identiques. Une petite variance est signe que ces tirages seront proches les uns des autres alors qu'une variance élevée est signe que'ils seront très écartés.

Proposition 4 $\mathrm{Var}(aX + b) = a^2 V(X)$.

Définition 12 (Écart-type) *L'écart-type de la variable aléatoire* X *est la racine carrée de sa variance :*

$$\sigma_X = \sqrt{\sigma_X^2}$$

Il permet de mettre à l'échelle une variable : soit $Y = \dfrac{X}{\sigma_X}$, on a $Var(Y) = 1$.

Définition 13 (Fonction indicatrice) *Soit* A *une partie de* $\mathbb{R}$*, on note* $\mathbb{1}_A$ *la fonction qui vaut 1 lorsque* $x \in A$ *et 0 sinon.*

$$\begin{aligned}
\mathbb{1}_A : \quad & \mathbb{R} \to [0;1] \\
& x \mapsto \mathbb{1}_A(x) = \begin{cases} 1 & si\ x \in A \\ 0 & sinon. \end{cases}
\end{aligned}$$

La fonction indicatrice est très utile grâce à la proposition suivante, qui permet de passer de l'espérance à la probabilité et réciproquement :

Proposition 5 Soit A une partie de $\mathbb{R}$, on a

$$\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$$

Démonstration *Il suffit de passer par l'intégrale (ou la somme pour une variable* X *discrète) :*

$$\int_{\mathcal{X}} \mathbb{1}_A(X) f_X(x)dx = \int_A 1 * f_X(x)dx + \int_{A^c} 0 * f_X(x)dx = \mathbb{P}(X \in A)$$

G Couple de variables aléatoires

1 Distribution jointe

Définition 14 (Distribution jointe, cas discret) *Pour un couple de variables aléatoires discrètes X et Y, la distribution jointe correspondante est donnée par :*

$$P_{X,Y}(x,y) = \Pr(X = x, Y = y)$$

la fonction de répartition correspondante est donnée par :

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \sum_{x_k \leq x,\ y_l \leq y} \Pr(X = x_k, Y = y_l)$$

Définition 15 (Distribution jointe, cas continu) *Pour un couple de variables aléatoires continues X et Y, la densité jointe est la fonction $f_{X,Y}(x,y)$ telle que :*

$$\int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy = \Pr(X \leq x, Y \leq y)$$

la fonction de répartition correspondante est donnée par :

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y)$$

de plus, on a :

$$\frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = f_{X,Y}(x,y)$$

2 Distributions marginales

Dans un couple de variables aléatoires, la distribution marginale d'une des variables est la distribution de cette dernière, quelles que soient les valeurs prises par l'autre. On l'obtient en sommant les probabilités au travers de toutes la valeurs possibles de l'autre membre du couple, ou en intégrant la densité jointe par rapport aux valeurs de l'autre membre du couple.

Définition 16 (Distribution marginale, cas discret) *Soit X, Y un couple de variables aléatoires discrètes. Les distributions marginales de X et Y sont donnée par :*

$$\Pr_X(x) = \sum_{k=1}^{K} \Pr_{X,Y}(x, y_k) \ \text{ et } \ \Pr_Y(y) = \sum_{k=1}^{K} \Pr_{X,Y}(x_k, y)$$

Définition 17 (Distribution marginale, cas continu) *Soit X, Y un couple de variables aléatoires continues. Les distributions marginales de X et Y sont donnée par :*

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy \ \text{ et } \ f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx$$

3 Distributions conditionnelles

Dans un couple de variables aléatoires X, Y, la distribution conditionnelle de l'une sachant la valeur de l'autre correspond à la distribution de l'une des variables aléatoires pour une réalisation donnée de l'autre.

Définition 18 (Distribution conditionnelle, cas discret) *Dans un couple de variables aléatoires discrètes X, Y, la distribution conditionnelle de X sachant $Y = y$ est donnée par :*

$$\Pr_{X|Y=y}(x\,|Y=y\,) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

Définition 19 (Distribution conditionnelle, cas continu) *Dans un couple de variables aléatoires continues X, Y, la distribution conditionnelle de X sachant $Y = y$ est donnée par :*

$$f_{X|Y=y}\,(x\,|Y=y\,) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

4 Indépendance

Définition 20 (Indépendance) *Deux variables aléatoires X et Y sont indépendantes si pour tous les événements $\{X \in A\}$ et $\{Y \in B\}$, on a $\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$. Si X et Y sont indépendantes, alors $F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \Pr(X \leq x)\ \Pr(Y \leq y) = F_X(x)F_Y(y)$. Cette condition s'étend aux fonctions de densité : $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.*

5 Covariance, corrélation

Dans un couple de variables aléatoires, on mesure la façon dont les deux variables (co)varient à l'aide de la covariance.

Définition 21 (Covariance) *La covariance entre deux variables X et Y s'écrit :*

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

Démonstration

$$\begin{aligned}
\mathrm{Cov}\,(X,Y) &= \mathbb{E}\,[(X - \mathbb{E}\,[X])\,(Y - \mathbb{E}\,[Y])] \\
&= \mathbb{E}\,[XY - Y\mathbb{E}\,[X] - X\mathbb{E}\,[Y] + \mathbb{E}\,[X]\,\mathbb{E}\,[Y]] \\
&= \mathbb{E}\,[XY] - \mathbb{E}\,[Y\mathbb{E}\,[X]] - \mathbb{E}\,[X\mathbb{E}\,[Y]] + \mathbb{E}\,[\mathbb{E}\,[X]\,\mathbb{E}\,[Y]] \\
&= \mathbb{E}\,[XY] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y] + \mathbb{E}\,[X]\,\mathbb{E}\,[Y] \\
&= \mathbb{E}\,[XY] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y]
\end{aligned}$$

Proposition 6

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + E(Y)$.
- Si X et Y sont deux variable aléatoire indépendantes, alors $\mathbb{E}\,[XY] = \mathbb{E}\,[X]\,\mathbb{E}\,[Y]$, et par conséquence $\mathrm{Cov}\,(X,Y) = 0$. Mais la réciproque n'est pas nécessairement vraie : deux variables décorrélées peuvent être dépendantes.
- $\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) + 2\mathrm{Cov}\,(X,Y)$.

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

Définition 22 (Coefficient de corrélation)

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \text{ et } |\rho_{X,Y}| \leq 1.$$

H Probabilités : théorèmes

1 Modes de convergences

En probabilité, il y a plusieurs modes de convergence.

Définition 23 (Convergence presque sûre) *On dit qu'une suite de variable aléatoire $X_1, \ldots, X_n \ldots$, converge presque sûrement vers la variable aléatoire Y, et on note $X_n \xrightarrow{p.s.} Y$, si :*

$$\forall \varepsilon > 0,\ \Pr\left(\lim_{n\to\infty} |X_n - Y| \geq \varepsilon\right) = 0.$$

La convergence presque sûre peut se comprendre ainsi : si $X_n \xrightarrow{p.s.} Y$, alors il n'y a aucune chance d'observer une réalisation de la suite X_n qui ne tende pas vers Y.

Définition 24 (Convergence en probabilité) *On dit qu'une suite de variable aléatoire $X_1, \ldots, X_n \ldots$, converge en probabilité vers la variable aléatoire Y, et on note $X_n \xrightarrow{\mathbb{P}} Y$, si :*

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} \Pr(|X_n - Y| \geq \varepsilon) = 0.$$

ou de manière équivalente, si :

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} \Pr(|X_n - Y| < \varepsilon) = 1.$$

La convergence en probabilité indique que la probabilité d'observer un écart donné, disons 0,01 , entre X_n et Y tend vers 0.

Définition 25 (Convergence en loi) *Une suite de variable aléatoire $X_1, \ldots, X_n \ldots$, de fonction de répartition $F_n(X)$ converge en loi vers la variable aléatoire Y de fonction de répartition $F_Y(x)$, et on note $X_n \xrightarrow{\mathcal{L}} Y$, si :*

$$\lim_{n\to\infty} F_n(x) = F_Y(x)$$

La convergence en loi (ou convergence en distribution) concerne le comportement globale d'une variable aléatoire, c'est-à-dire sa loi de probabilité. Ainsi, X_n peut tendre en loi vers Y sans que les réalisations ne soient jamais proches.

Par exemple, si Y est la réalisation d'un dé, la suite constante égale à $X_n = 7 - Y$ converge en loi vers Y, puisque sa loi est constante et égale à la loi uniforme discrète sur $[|1;6|]$, comme Y. Mais pour tout tirage, $|X_n - Y| = |7 - 2Y| \geq 1$.

En ce sens il est courant de dire qu'une suite de variables aléatoires converge en loi vers une loi particulière (et non pas vers une autre variable aléatoire qui elle-même suit une loi donnée). Nous utiliserons ce type de notation dans le théorème central limite.

C'est convergences sont liées entre elles :

Théorème 1 (Ordre sur les convergences) *Soit une suite de variable aléatoire* $X_1, \ldots, X_n \ldots$, *et une variable aléatoire* Y.
Si (X_n) *convergence presque sûrement vers* Y *alors* (X_n) *converge en probabilité vers* Y.

$$\left(X_n \xrightarrow{p.s.} Y\right) \Rightarrow \left(X_n \xrightarrow{\mathbb{P}} Y\right)$$

Si (X_n) *convergence en probabilité vers* Y *alors* (X_n) *converge en loi vers* Y.

$$\left(X_n \xrightarrow{\mathbb{P}} Y\right) \Rightarrow \left(X_n \xrightarrow{\mathcal{L}} Y\right)$$

2 Lois des grands nombres et théorème central limite

Définition 26 (Échantillon i.i.d.) *On appelle échantillon i.i.d. une suite de variables aléatoires* $(X_1, \ldots, X_n)$ *indépendantes deux à deux et toutes de même loi.*

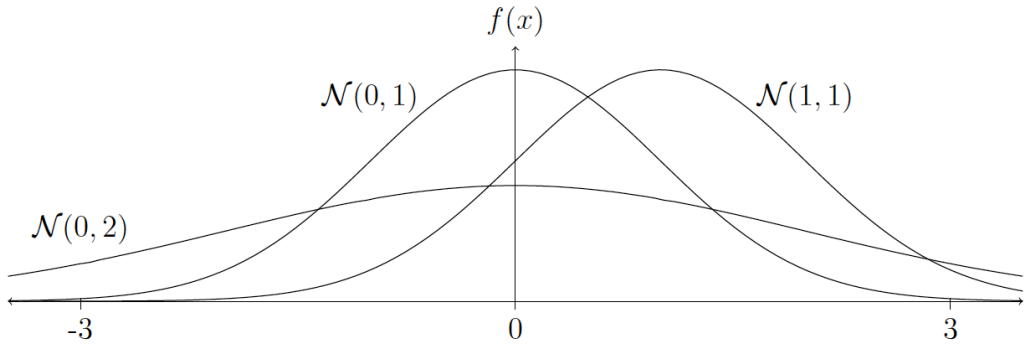
Théorème 2 (Loi (faible) des Grands Nombres) *Soient* $(X_1, \ldots, X_n)$ *une suite de* n *vecteurs aléatoires i.i.d. de* $\mathbb{R}^d$ *avec* $\mathbb{E}[|X|] < \infty$, *alors l'espérance* $\mathbb{E}[X_1] = \mu$ *esiste et :*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathbb{P}} \mu.$$

Définition 27 *La loi normale (ou loi gaussienne) d'espérance* μ *et de variance* σ^2 *est notée* $\mathcal{N}\left(\mu, \sigma^2\right)$. *Sa fonction de densité s'écrit :*

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

On utilisera régulièrement dans ce cours la loi normale centrée-réduite : $\mathcal{N}(0;1)$, c'est-à-dire la loi normale d'espérance $\mu = 0$ et de variance $\sigma^2 = 1$. On note $z_{(p)}$ le quantile d'ordre p de la $\mathcal{N}\ (0,1)$. En d'autres termes, $z_{(p)}$ est la valeur telle que, pour une variable aléatoire $Z \sim \mathcal{N}\ (0,1)$, $\Pr\left(Z \leq z_{(p)}\right) = \Phi\left(z_{(p)}\right) = p$.



Proposition 7

- La loi normale est stable par addition, *i.e.* si $X \sim \mathcal{N}\ \left(\mu_X, \sigma_X^2\right)$ et $Y \sim \mathcal{N}\ \left(\mu_Y, \sigma_Y^2\right)$ alors

$$(X+Y) \sim \mathcal{N}\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\mathrm{Cov}\,(X,Y)\right).$$

- La loi normale est la seule loi pour laquelle absence de corrélation et indépendance sont équivalentes : si $X \sim \mathcal{N}\ \left(\mu_X, \sigma_X^2\right)$ et $Y \sim \mathcal{N}\ \left(\mu_Y, \sigma_Y^2\right)$, alors

$$\mathrm{Cov}\,(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y.$$

Théorème 3 (Théorème central limite) *Soient $(X_1, \ldots, X_n)$ une suite de n variables aléatoires i.i.d. d'espérance $\mu = \mathbb{E}(X_1)$, de variance $\sigma^2 = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^2\right]$ finie, alors :*

$$\sqrt{n}\left(\overline{X} - \mu\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)\,.$$

La différence entre le théorème central-limite et la loi des grands nombres tient au facteur $\sqrt{n}$ qui vient dilater la suite des $\dfrac{\overline{X}_n - \mu}{\sigma}$. Sans ce facteur, la suite $\overline{X}_n - \mu$ convergerait vers 0 par la loi des grands nombres. Le facteur multiplicatif $\sqrt{n}$ permet d'empêcher la suite de s'effondrer en un point unique au fur et à mesure que n grandit.

Remarque 2 Le théorème central limite est d'une importance capitale en statistiques. En effet, il énonce que, sous des conditions très générales, dès lors que l'on manipule des moyennes d'échantillon de taille assez grande, alors on peut raisonnablement approximer la distribution de cette moyenne empirique par une loi Normale, et ce quelque soit la loi initiale ayant généré les observations.

Cette autre formulation du théorème peut en aider la compréhension :

$$\overline{X} = \mu + \frac{\sigma}{\sqrt{n}} Z_n \text{ où } Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0; 1)$$

3 Extensions

Théorème 4 (Théorème central limite multidimensionnel) *Soient $(X_1, \ldots, X_n)$ une suite de n vecteurs aléatoires i.i.d. de $\mathbb{R}^d$, d'espérance $\mu = \mathbb{E}(X_1)$, de matrice de covariance $V = \mathbb{E}(X_1 X_1') - \mathbb{E}(X_1)\mathbb{E}(X_1)'$ finie, alors :*

$$\sqrt{n}\left(\overline{X} - \mu\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, V)\,.$$

Théorème 5 (Slutsky) *Soit $(X_1, \ldots, X_n)$ une suite de variables aléatoires qui converge en loi vers une variable aléatoire X. Soit $(Y_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires définies sur le même espace que les v.a. X_n, qui converge presque sûrement ou en probabilité ou en loi vers la constante a. Alors la suite $((X_n, Y_n), n \in \mathbb{N}^*)$ converge en loi vers le couple (X, a).*

Corollaire *Soit $X_1, \ldots, X_n$ une suite de variable aléatoire i.i.d. telles que $\mathbb{E}[X_i] = \mu < \infty$ et $\mathrm{Var}\,(X_i) = \sigma^2 < \infty$; soit $\widehat{\sigma^2}_n$ une suite de variable aléatoire tel que $\widehat{\sigma^2}_n \xrightarrow{P} \sigma^2$ alors*

$$\sqrt{\frac{n}{\widehat{\sigma^2}_n}}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

Théorème 6 (Delta-Méthode) *Soit $(X_1, \ldots, X_n)$ une suite de n vecteurs aléatoires. i.i.d. de $\mathbb{R}^d$, d'espérance μ, de matrice de covariance Σ finie, telle que :*

$$\frac{1}{\sqrt{n}}(X_n - \mu) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)\,,$$

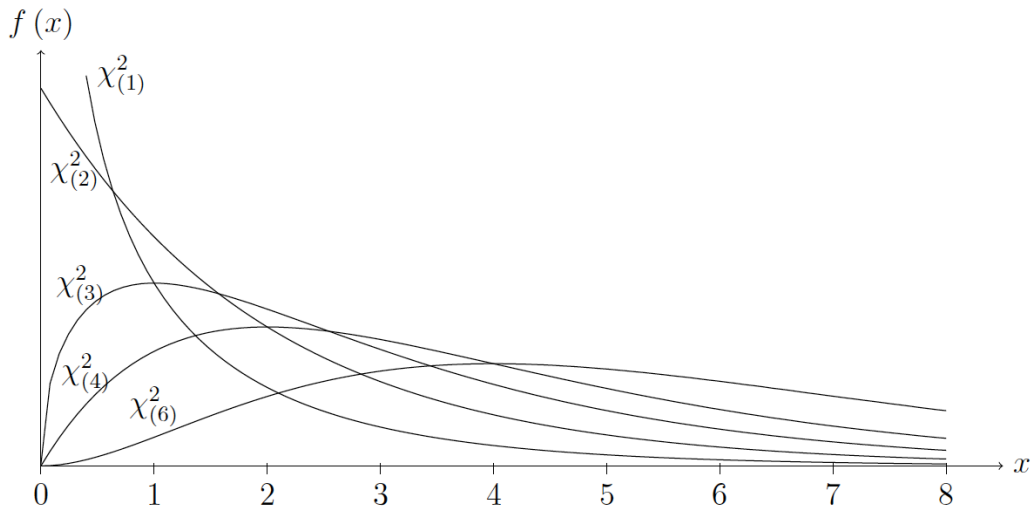
et soit une fonction φ de $\mathbb{R}^k$ dans $\mathbb{R}^l$, de classe C^1, telle que : $\dfrac{\partial \varphi}{\partial x}(\mu) \neq 0$, alors, on a :

$$\frac{1}{\sqrt{n}}(\varphi(X_n) - \varphi(\mu)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\partial \varphi}{\partial x^t}(\mu)\Sigma\frac{\partial \varphi'}{\partial x}(\mu)\right)\,,$$

4 Lois associées à la loi Gaussienne

Définition 28 *La loi du Khi-deux à n degrés de liberté, notée $\chi^2_{(n)}$ est définie comme la somme des carrés de n variables aléatoires indépendantes distribuées selon une loi normale centrée-réduite : si $X_1, \ldots, X_n$ sont telles que $X_i \sim \mathcal{N}(0,1)\,\forall i = 1, \ldots, n$; alors $\sum_{i=1}^{n} X_i^2 = Z \sim \chi^2_{(n)}$.*

On note $k_{(p,n)}$ le quantile d'ordre p d'une $\chi^2_{(n)}$; *i.e.* la valeur telle que si $Z \sim \chi^2_{(n)}$, alors $\Pr\left(Z \leq k_{(p,n)}\right) = p$.



Proposition 8

- Si $Z \sim \chi^2_{(n)}$ alors $\mathbb{E}[Z] = n$ et $\mathrm{Var}(Z) = 2n$.
- Si $Z_1 \sim \chi^2_{(n_1)}$ et $Z_2 \sim \chi^2_{(n_2)}$ alors $(Z_1 + Z_2) \sim \chi^2_{(n_1+n_2)}$.
- Si $Z = Z_1 + Z_2, Z \sim \chi^2_{(n)}$ et $Z_1 \sim \chi^2_{(p)}$ avec $p < n$; alors $Z_2 \sim \chi^2_{(n-p)}$ et Z_1 et Z_2 sont indépendantes.

Définition 29 (Lois de Student et de Fisher)

1. *Si X suit une loi normale réelle standard, Y suit une loi χ^2_k, et que X et Y sont indépendants, on dit que $T = \dfrac{X}{\sqrt{(Y/k)}}$ suit une loi de Student à k degrés de libertés, notée $\mathcal{T}_k$.*
2. *Si p et q sont des entiers, si X suit une loi de χ^2_p, X indépendante de Y qui suit une loi de χ^2_q, alors $F = \dfrac{X/p}{Y/q}$ suit une loi de Fisher-Snedecor à p et q degrés de liberté, notée $\mathcal{F}(p,q)$.*

Proposition 9 1. On a $\mathcal{T}_k \xrightarrow[k\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$

2. $Z \sim \mathcal{F}(p,q) \Longrightarrow \dfrac{1}{Z} \sim \mathcal{F}(q,p)$,
3. $X \sim \mathcal{T}_n \Longrightarrow X^2 \sim \mathcal{F}(1,n)$.

Théorème 7 (Student) *Soit $(X_1,\ldots,X_n)$ un n-échantillon i.i.d. de loi $\mathcal{N}(\mu,\sigma^2)$, alors :*

1. $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ *suit une loi* $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.
2. $\dfrac{R_n}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ *suit une loi* χ^2_{n-1}.
3. $\overline{X}_n$ *et* R_n *sont indépendants.*
4. *Si* $S_n = \sqrt{\dfrac{R_n}{n-1}}$, *alors* $T_n = \sqrt{n}\,\dfrac{\overline{X}_n - \mu}{S_n}$ *suit une loi* $\mathcal{T}_{n-1}$.

Chapitre 2

Estimation par substitution

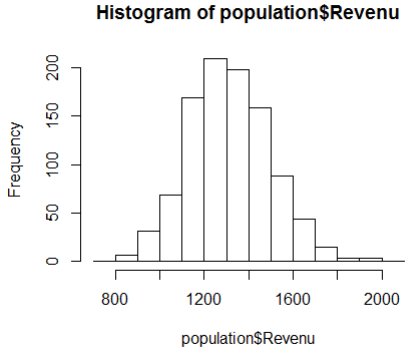
La statistique mathématique s'est surtout développée aux XIX^e et XX^e siècles pour le traitement de données issues de mesures en sciences expérimentales, avec des applications en agronomie et dans l'industrie. L'idée de base est qu'on peut répliquer une expérience dans des conditions pratiquement identiques, ce qui conduit à des mesures répétées du même phénomène, à un aléas résiduel prés.

Une première grande difficulté de l'usage de la statistique mathématiques en sciences sociales est que ce cadre ne tient pas. Un jeu de données ne correspond en général pas à une expérience répétée dans des conditions identiques, mais plutôt à une caractéristique similaire chez des personnes distinctes.

Lire « Statut de la Dispersion » d'Armatte sur le passage des sciences dures aux sciences sociales à propos de la dispersion chez Quételet (l'auteur justement de *Sur l'homme et le développement de ses facultés ou essai de physique sociale* et père du concept de l'« homme moyen »).

On dispose d'un fichier *Population* qui contient l'identifiant, le revenu et le genre d'une population de $N = 1\,000$ individus, les salarié-es de l'ENS Cachan.

Le salaire moyen de cette population est de 1 316€, avec une variance de 37 796 €^{2}. La proportion de femmes est de 40,1%. On obtient l'histogramme suivant pour les salaires



Dans la suite, on suppose ces informations inconnues.

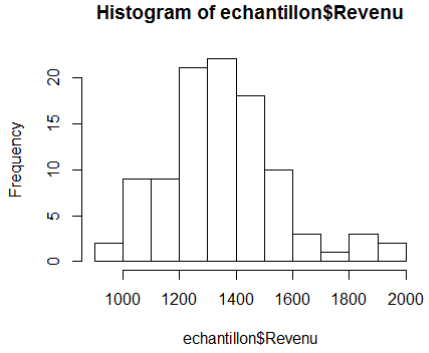
Échantillonnage et sondage sont des synonymes dans le langage courant : on extrait une petite partie au sein d'une population. En statistique mathématique, l'échantillonnage est un modèle théorique de sondage aléatoire simple, qui insiste le

caractère répétitif de la méthode de sélection de l'individu interrogé : on suppose pouvoir obtenir des observations répétées issues de la même loi. Le terme met en avant l'homogénéité de la production des données de l'échantillon. On pense davantage à la répétition avec l'échantillonnage, quand le sondage renvoie plus à l'idée d'extrait au sein d'une population large.

On considère un échantillon aléatoire i.i.d. $(X_1, \ldots, X_n)$ de taille n, d'une loi $\mathcal{L}$ d'espérance et de variance finies ($\mathbb{E}[X] = \mu_X < \infty$ et $\mathrm{Var}(X) = \sigma_X^2 < \infty$).

Exemple Dans notre exemple, on veut obtenir des estimateurs du revenu moyen dans la population. On note R le revenu d'un individu tiré uniformément dans la population, et on cherche $\mathbb{E}[R] = \mu_R$ =1 316€ et $\mathrm{Var}(R) = \sigma_R^2$ =37 796 €^{2}, qui sont donc supposées inconnues.

On tire aléatoirement avec remise un échantillon de taille $n = 100$, parmi les identifiants des $N = 1\ 000$ individus. On obtient l'histogramme suivant pour une réalisation d'un tel échantillon de salaires



On note $I_1, \ldots, I_{100}$ ces identifiants et pour chaque individu d'identifiant I_k, on note R_{I_k} son revenu.

On a alors, pour n'importe quel $k \in [|1; 100|]$, que l'espérance de R_{I_k} est la moyenne des revenus dans la population :

$\mathbb{E}[R_{I_k}]$ est l'espérance d'une loi discrète uniforme parmi les N revenus des individus de la population. C'est donc la moyenne pondérée sur l'échantillon, avec une pondération uniforme :

$$\mu_R = \mathbb{E}[R_{I_k}] = \frac{1}{N} \sum_{j=1}^{N} R_j.$$

Définition 30 (Paramètre) *On appelle paramètre une caractéristique d'une loi de probabilité décrite par un réel : son espérance, sa médiane ou sa variance par exemple. L'objet de l'estimation statistique est de donner une approximation de ce paramètre à partir de réalisations issues de cette loi.*

Définition 31 (Statistique) *On appelle statistique toute valeur calculable à partir d'un échantillon $(X_1, \ldots, X_n)$, c'est-à-dire toute fonction de $X_1, \ldots, X_n$.*

Il convient de remarquer qu'une statistique est une variable aléatoire car sa valeur est déterminée par une épreuve aléatoire, c'est-à-dire par le tirage aléatoire d'un

échantillon. Sur un échantillon particulier, on observe une valeur particulière de la statistique, la statistique observée.

En tant que variable aléatoire, toute statistique est distribuée selon une loi de probabilité, et on peut définir son espérance, sa variance, etc.

Définition 32 (Estimateur) *Soit θ un paramètre appartenant à un espace Θ. On appelle estimateur de θ une statistique qui prend ses valeurs dans Θ. On le note généralement $\hat{\theta}_n$:*

$$\begin{array}{rccc} \hat{\theta}_n & : & \mathcal{X}^n & \to & \Theta \\ & & (X_1, \ldots, X_n) & \mapsto & \hat{\theta}_n(X_1, \ldots, X_n) \end{array}$$

On omet généralement l'échantillon en notant indifféremment $\hat{\theta}_n$ pour la fonction ou pour sa valeur sur l'échantillon. On peut également sous-entendre le n.

Définition 33 (Biais) *Le biais est l'espérance de la différence entre l'estimateur et le paramètre qu'il estime : $\mathbb{E}\left[\hat{\theta}_n - \theta\right]$. S'il est nul, on dit que l'estimateur est sans biais.*

A Estimation de l'espérance

Dans cette section, on cherche à estimer l'espérance en supposant connue la variance.

Définition 34 (Moyenne empirique) *On appelle moyenne empirique la statistique :*

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

Exemple Ici, on estime, on utilise donc l'échantillon pour dire quelque chose sur la population : la moyenne empirique sur l'échantillon donne une information sur l'espérance μ_R de chaque tirage, qui est égale à la moyenne sur toute la population comme on l'a montré ci-dessus.

La moyenne empirique n'a que très peu de chances d'être égale à la moyenne sur la population :

$$\overline{R}_n = \frac{1}{n}\sum_{k=1}^{n} R_{I_k} = 1\ 330€ \neq 1\ 316€ = \mu_R = \frac{1}{N}\sum_{j=1}^{N} R_j.$$

La moyenne empirique étant une variable aléatoire, elle possède une espérance et une variance :

Proposition 10 La moyenne empirique d'un échantillon i.i.d. est un estimateur sans biais de l'espérance ; elle a même espérance que chacune des variables :

$$Esp\left[\overline{X}_n\right] = \mu_X = \mathbb{E}[X].$$

Sa variance est inversement proportionnelle à la taille de l'échantillon :

$$\mathrm{Var}\left(\overline{X}_n\right) = \frac{1}{n}\sigma_X^2 = \frac{1}{n}\mathrm{Var}(X).$$

Démonstration

$$\begin{aligned}
\mathbb{E}\left[\overline{X}_n\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] \ \textit{par linéarité de l'espérance} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_X \ \textit{car } \mathbb{E}\left[X_i\right] = \mu_X \ \forall i \\
&= \mu_X
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left(\overline{X}_n\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) \ \textit{car les } X_i \ \textit{sont indépendants} \\
&= \frac{1}{n}\sigma_X^2
\end{aligned}$$

Remarque 3 Ces résultats indiquent d'une part que le calcul de la moyenne empirique sur un échantillon nous fournit une approximation de μ_X qui a autant de chance de sur-estimer que de sous-estimer ce paramètre. Et d'autre part, que la distribution de la moyenne empirique devient de plus en plus concentrée autour de la vraie valeur μ_X au fur et à mesure que la taille de l'échantillon grandit (car $\mathrm{Var}\left(\overline{X}_n\right)$ diminue avec n).

Exemple On peut donc appliquer cette proposition aux revenus des salarié-es de Cachan et on obtient pour un échantillon théorique

$$\mathbb{E}\left[\overline{R}_n\right] = \mu_R = 1\ 316€ \ \text{ et } \ \mathrm{Var}\left(\overline{R}_n\right) = \frac{1}{n}\sigma_R^2 = \frac{37\ 796}{100}€^2$$

Qu'il ne faut pas confondre avec la réalisation de $\overline{R}_n$ =1 330€ et la variance empirique sur l'échantillon observé, qu'on va étudier dans la section suivante.

Définition 35 (Estimateur convergent)
Un estimateur est convergent s'il converge ***en probabilité*** *vers le paramètre qu'il estime :*

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

.

Proposition 11 (Caractérisation de la convergence en probabilité)
Un estimateur dont le biais et la variance tendent vers 0 est convergent.

Démonstration *On doit montrer que pour tout $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

On utilise la fonction indicatrice et en particulier $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$. À n fixé, on a donc

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) = \mathbb{E}\left[\mathbb{1}_{|\hat{\theta}_n - \theta| \geq \varepsilon}\right]$$

Lorsque l'indicatrice vaut 1, $\dfrac{|\hat{\theta}_n - \theta|}{\varepsilon} \geq 1$. Par croissance de la fonction carrée, on a aussi $\dfrac{\left(\hat{\theta}_n - \theta\right)^2}{\varepsilon^2} \geq 1$. On a donc la majoration suivante :

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq \mathbb{E}\left[\frac{\left(\hat{\theta}_n - \theta\right)^2}{\varepsilon^2}\mathbb{1}_{|\hat{\theta}_n - \theta| \geq \varepsilon}\right]$$

Puisqu'un carré est toujours positif, on majore en ajoutant $\dfrac{\left(\hat{\theta}_n - \theta\right)^2}{\varepsilon^2}\mathbb{1}_{|\hat{\theta}_n - \theta| \geq \varepsilon}$. Comme les deux indicatrices indiquent des événements complémentaires, on peut les remplacer par 1 ($\mathbb{1}_A + \mathbb{1}_{A^c} = 1$).

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq \mathbb{E}\left[\frac{\left(\hat{\theta}_n - \theta\right)^2}{\varepsilon^2}\right] = \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$$

On remarque alors que

$$\begin{aligned}\left(\hat{\theta}_n - \theta\right)^2 &= \left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta\right)^2 \\ &= \left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\right)^2 + \left(\mathbb{E}[\hat{\theta}_n] - \theta\right)^2 + 2\left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\right)\left(\mathbb{E}[\hat{\theta}_n] - \theta\right).\end{aligned}$$

L'espérance du dernier terme est nulle puisque $\left(\mathbb{E}[\hat{\theta}_n] - \theta\right)$ est une constante qui sort de l'espérance avec le 2 et par linéarité de l'espérance

$$\mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\right)\right] = \mathbb{E}[\hat{\theta}_n] - \mathbb{E}[\hat{\theta}_n] = 0.$$

On a alors

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\right)^2 + \left(\mathbb{E}[\hat{\theta}_n] - \theta\right)^2\right] = \mathrm{Var}(\hat{\theta}_n) + \mathit{biais}^2$$

et chacun de ces termes tend vers 0 par hypothèse.

Corollaire *La moyenne empirique d'un échantillon i.i.d. est un estimateur convergent de l'espérance commune μ_X.*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu_X$$

Exemple Dans notre exemple, ca ne veut par dire que la moyenne empirique égale à 1 330€ sur l'échantillon de 100 individus tiré converge vers l'espérance (la moyenne sur toute la population) égale à 1 316€. Ca n'aurait aucun sens. Pour avoir convergence, il faut d'une part répéter le processus en tirant de plus en plus d'individus : $n \to \infty$. D'autre part, c'est une convergence en probabilité, ce qui veut dire que pour tout ε, il existe un n à partir duquel, à chaque tirage, on a une probabilité plus grande que $1 - \varepsilon$ que la moyenne empirique soit égale à 1 316€ à ε près.

Définition 36 (Estimateur asymptotiquement normal) *Un estimateur est asymptotiquement normal ssi il existe un réel positif v tel que :*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v)$$

Proposition 12 La moyenne empirique d'un échantillon i.i.d. est un estimateur asymptotiquement normal de l'espérance commune μ_X :

$$\sqrt{n}\left(\overline{X}_n - \mu_X\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma_X^2)$$

On peut aussi écrire

$$\overline{X}_n = \mu_X + \frac{\sigma_X}{\sqrt{n}} Z_n \text{ où } Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0; 1).$$

L'estimateur est donc égal au paramètre avec une erreur asymptotiquement normale, proportionnelle à l'écart-type et à l'inverse de la racine carrée du nombre d'observations.

Démonstration *On applique toujours le Théorème Central Limite pour obtenir la normalité asymptotique. Il s'applique puisque $\overline{X}_n$ est une moyenne de variables aléatoires i.i.d. de variance finie (σ_X^2). On a alors*

$$\frac{\sqrt{n}}{\sigma_X}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mu_X\right) = \frac{\sqrt{n}}{\sigma_X}(\overline{X}_n - \mu_X) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0; 1).$$

Proposition 13 (Moyenne d'un échantillon gaussien)
La moyenne empirique d'un échantillon gaussien suit une loi normale :

$$\overline{X}_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

Remarque 4 Nous considérons un échantillon i.i.d. $(X_1, \ldots, X_n)$ de taille n d'une loi $\mathcal{N}\left(\mu, \sigma^2\right)$. Cette hypothèse est beaucoup plus forte que celles de la section précédente : c'est une hypothèse de dimension infinie, sur l'ensemble de la distribution des variables aléatoires, par opposition aux hypothèses de dimension 1 quantifiant la position centrale (l'espérance) ou la dispersion (la variance). De plus, en sciences sociales, il n'y a aucun cas où cette hypothèse est réaliste.

La démonstration est immédiate en utilisant la proposition 7 précédente aux $\dfrac{X_i}{n}$. On a alors à distance finie (c'est-à-dire pour tout n) le comportement donné par la normalité asymptotique

$$\overline{X}_n = \mu_X + \frac{\sigma_X}{\sqrt{n}} Z_n \text{ où } Z_n \sim \mathcal{N}(0; 1).$$

B Estimation de la variance

Le cas le plus simple, mais peu réaliste, d'estimation de la variance est symétrique du précédent : on cherche à estimer la variance en supposant connue l'espérance. C'est moins réaliste que le cas précédent au sens où la variance est une information de second ordre par rapport à l'espérance : en général on s'intéresse à la variance après avoir estimé l'espérance. On viendra rapidement à cette configuration.

Définition 37 (Variance empirique, espérance connue) *Lorsque l'espérance est connue, la variance empirique est la statistique :*

$$\hat{V}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_X\right)^2$$

Il s'agit évidemment d'un estimateur de la variance σ_X^2 : la statistique $\hat{V}_n^2$ prend ses valeurs dans l'espace $\mathbb{R}^+$ des valeurs possibles pour une variance.

Exemple Dans notre exemple, il faut donc supposer qu'on connait l'espérance, c'est-à-dire la moyenne su la population : $\mu_R = 1\ 316$. On obtient alors $V_n^2 = 85\ 164$ c'est-à-dire un écart-type de 292€.

Proposition 14 Dans le cas d'un échantillon i.i.d. d'espérance connue $\mu_X < \infty$ et de variance σ_X^2 finie mais inconnue, l'estimateur $\hat{V}_n^2$ est sans biais.

Démonstration *Calculons l'espérance de cet estimateur.*

$$\mathbb{E}\left[\hat{V}_n^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_X\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu_X\right)^2\right]$$

par linéarité de l'espérance. Comme les variables sont identiquement distribuées, tous les termes de la somme sont égaux (on obtient un facteur n qui se simplifie avec le $1/n$). On obtient donc

$$\mathbb{E}\left[\hat{V}_n^2\right] = \mathbb{E}\left[\left(X_1 - \mu_X\right)^2\right] = \sigma_X^2.$$

L'estimateur est donc sans biais.

Dans le cas plus commun où l'espérance est inconnue, il faut donc l'estimer, par exemple par substitution par la moyenne empirique $\overline{X}_n$.

Définition 38 (Variance empirique non corrigée) *On appelle variance empirique non corrigée la statistique :*

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

Exemple Cette fois-ci on centre sur la moyenne sur l'échantillon. On obtient alors $\tilde{S}_n^2 = 51\ 465$ c'est-à-dire un écart-type de 227€.

Proposition 15 Dans le cas d'un échantillon i.i.d. d'espérance inconnue $\mu_X < \infty$ et de variance σ_X^2 finie et inconnue, $\tilde{S}_n^2$ est un estimateur biaisé : son espérance est différente de la variance des variables aléatoires, σ_X^2

$$\mathbb{E}\left[\tilde{S}_n^2\right] = \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2$$

Le biais tend néanmoins vers 0, $\tilde{S}_n^2$ est donc un estimateur de σ_X^2 **asymptotiquement sans biais**.

Démonstration

$$\begin{aligned}
\mathbb{E}\left[\tilde{S}_n^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 + \overline{X}_n^2 - 2X_i\overline{X}_n\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) + \overline{X}_n^2 - 2\overline{X}_n^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - \overline{X}_n^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}_n^2\right] \\
&= \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}_n^2\right] \\
&= \mathrm{Var}\left(X_i\right) + \mathbb{E}\left[X_i\right]^2 - \mathrm{Var}\left(\overline{X}_n\right) - \mathbb{E}\left[\overline{X}_n\right]^2 \\
&= \sigma_X^2 + \mu_X^2 - \frac{1}{n}\sigma_X^2 - \mu_X^2 \\
&= \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2
\end{aligned}$$

On peut alors utiliser la linéarité de l'espérance pour corriger ce biais.

Définition 39 (Variance empirique corrigée) *On appelle variance empirique corrigée la statistique :*

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

Exemple La correction est minime : on passe de $\tilde{S}_n^2 = 51\,465$ à $S_n^2 = 51\,984$ et pour les écart-types de 227€ à 228€.

Proposition 16 Dans le cas d'un échantillon i.i.d. d'espérance inconnue $\mu_X < \infty$ et de variance σ_X^2 finie et inconnue, S_n^2 est un estimateur sans biais de la variance σ_X^2 :

$$\mathbb{E}\left[S_n^2\right] = \sigma_X^2$$

Démonstration

$$\begin{aligned}
\mathbb{E}\left[S_n^2\right] &= \mathbb{E}\left[\frac{n}{n-1}\tilde{S}_n^2\right] = \frac{n}{n-1}\mathbb{E}\left[\tilde{S}_n^2\right] \\
&= \frac{n}{n-1}\frac{n-1}{n}\sigma_X^2 = \sigma_X^2
\end{aligned}$$

Proposition 17 (Variance d'un échantillon gaussien)
Si l'on suppose l'échantillon gaussien, c'est-à-dire que chaque X_i suit une loi normale $\mathcal{N}(\mu_X, \sigma_X^2$, alors l'estimateur par substitution de la variance suit une loi du χ^2 (à un facteur près).

Dans le cas où l'espérance est connue :

$$n\frac{V_n^2}{\sigma^2} \sim \chi^2(n), \text{ avec } V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Dans le cas où l'espérance est inconnue :

$$\frac{(n-1)\,S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}, \text{ avec } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

.

Chapitre 3

Intervalles de confiance et tests

A Intervalles de confiance

On va utiliser les résultats précédents pour ajouter à l'estimation une information sur la précision de cette estimation : au-lieu de ne proposer qu'une valeur (l'estimateur), on donne un intervalle et une probabilité (la confiance) que cet intervalle contienne la vraie valeur du paramètre.

Exemple On peut prendre l'exemple du sondage électoral : au lendemain du premier tour de l'élection présidentielle de 2012, l'Ifop interroge 1 004 personnes sur leur intention de vote au second tour et obtient 457 pour Sarkozy et 557 pour Hollande (on peut remarquer qu'il n'est pas fait mention d'intentions de vote blanc ou nul, ni de non réponses). Ces données permettent-elles de départager les deux candidats ?

Si l'on en reste à l'estimation, on peut estimer le résultat de Sarkozy par la moyenne des intentions de vote sur l'échantillon : $\overline{X} = 0,455$ et donc une défaite. Mais quelle est la précision de cet estimateur ? Est-il suffisamment différent de 50% pour conclure ? Est-ce que cette précision dépend du nombre de données, du nombre d'électeurs ?

1 Définition

Définition 40 *Soit $X_1 \dots X_n$ un échantillon, θ_0 un paramètre réel et $1-\alpha \in [0;1]$ un niveau de confiance. On appelle intervalle de confiance au niveau $1-\alpha$ la donnée de deux suites (a_n) et (b_n) telles que pour tout n*

$$\mathbb{P}([a_n; b_n] \ni \theta_0) = \mathbb{P}(\theta_0 \in [a_n; b_n]) = 1-\alpha$$

Lorsqu'on ne peut obtenir un intervalle exact, on construit un intervalle par excès :

$$\mathbb{P}([a_n; b_n] \ni \theta_0) \geq 1-\alpha$$

ou un intervalle asymptotique

$$\mathbb{P}([a_n; b_n] \ni \theta_0) \xrightarrow{n\to\infty} 1-\alpha$$

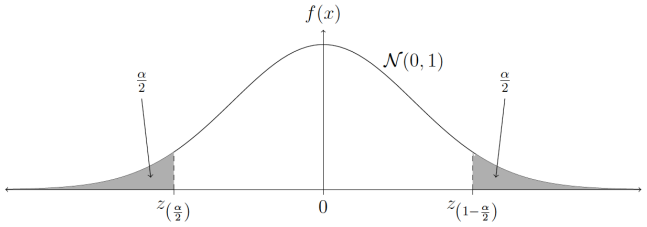


FIGURE 3.1 – Densité de la gaussienne centrée réduite et intervalle central

Remarque 5 L'existence même des intervalles de confiance rappel que l'estimation n'est jamais exacte, et qu'y compris en prenant un intervalle, il faut s'attendre régulièrement à ce que la vraie valeur du paramètre n'appartienne pas à l'intervalle. Le niveau de confiance le plus couramment utilisé est $1-\alpha = 0,95$, avec donc une erreur attendue de 5%.

Ajoutons que les intervalles les plus courants sont asymptotiques, ce qui signifie que le niveau de confiance lui-même n'est qu'approximatif.

Statistique pivotale

Pour construire un intervalle de confiance, il faut donc être capable de mesurer la précision d'un estimateur. On utilise en général un théorème donnant la loi (ou la loi asymptotique) de l'estimateur, ou plus souvent d'une expression contenant à la fois l'estimateur et le paramètre, du type :

$$\sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0;1)$$

pour l'espérance d'un échantillon gaussien. Il reste alors à faire « pivoter » cette expression pour en tirer un intervalle de confiance, c'est-à-dire à la traduire en une probabilité qu'un intervalle contiennent μ.

2 Espérance d'un échantillon gaussien, variance connue

Par simplicité, on suppose donc que les données sont i.i.d. de loi normale. Il suffit de penser à l'exemple du sondage électoral (dont la loi est de Bernoulli) pour comprendre que c'est rarement plausible en sciences sociales. On suppose de plus ici que la variance est connue.

L'idée est d'utiliser la statistique pivotale précédente, en gardant le centre de la distribution normale :

On identifie alors un intervalle central dans la distribution de la gaussienne centrée réduite, dont la probabilité vaut $1-\alpha$. Pour avoir un intervalle centré en 0 et symétrique, on a pour $Z \sim \mathcal{N}(0;1)$:

$$\mathbb{P}(|Z| \leq t) = \mathbb{P}(-t \leq Z \leq t) = 1 - \mathbb{P}(Z \leq -t) - \mathbb{P}(Z \geq t) = 1 - 2 * \mathbb{P}(Z \leq -t)$$

Pour obtenir $1-\alpha$, il faut donc choisir t tel que $\mathbb{P}(Z \leq -t) = \dfrac{\alpha}{2}$.

$$\frac{\alpha}{2} = \mathbb{P}(Z \leq -t) = F_{\mathcal{N}(0;1)}(-t) \Leftrightarrow t = -F^{-1}_{\mathcal{N}(0;1)}(\alpha/2) = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$$

On note z_α ce quantile $F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$.

En remplaçant la gaussienne neutre Z par la statistique pivotale, on obtient

$$\begin{aligned}1-\alpha &= \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}-\mu}{\sigma}\right| \leq z_\alpha\right)\\ &= \mathbb{P}\left(\left|\overline{X}-\mu\right| \leq \frac{\sigma}{\sqrt{n}}z_\alpha\right)\\ &= \mathbb{P}\left(\overline{X}-\frac{\sigma}{\sqrt{n}}z_\alpha \leq \mu \leq \overline{X}+\frac{\sigma}{\sqrt{n}}z_\alpha\right)\end{aligned}$$

On obtient alors le théorème suivant :

Théorème 8 (IC pour l'espérance d'un échantillon gaussien, σ^2 connue) *Si $X_1,\ldots,X_n$ est un échantillon gaussien d'espérance inconnue μ et de variance connue $\sigma^2>0$, alors*

$$\left[\overline{X}-\frac{\sigma}{\sqrt{n}}z_\alpha;\overline{X}+\frac{\sigma}{\sqrt{n}}z_\alpha\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$

Démonstration *Exercice : remettre les calculs précédents dans l'ordre pour démontrer le théorème.*

Exemple On peut appliquer ce théorème à notre sondage électoral, en supposant abusivement les données gaussiennes. Il faut donc d'une part choisir α, par exemple comme c'est l'usage à 5% : $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(0.975) = 1,96$.

Et il faut d'autre part, connaître σ. Puisque les données sont en réalité Bernoulli, la variance vaut $\sigma^2 = p(1-p)$ pour un certain p (qui est justement l'espérance que l'on cherche) entre 0 et 1. On a donc un intervalle de confiance exact mais dépendant du paramètre inconnu :

$$\left[\overline{X}-\frac{\sigma}{\sqrt{n}}z_\alpha;\overline{X}+\frac{\sigma}{\sqrt{n}}z_\alpha\right] = [0,455-\frac{\sqrt{p(1-p)}}{\sqrt{1014}}*1,96;0,455-\frac{\sqrt{p(1-p)}}{\sqrt{1014}}*1,96]$$

On peut majorer $p(1-p)$ sur $[0;1]$ par $1/4$ et donc obtenir un intervalle de confiance à 95% par excès (ce qui veut dire qu'on a une confiance de plus que 95%, mais qu'on ne la connait pas exactement :

$$\begin{aligned}&\left[0,455-\frac{\sqrt{p(1-p)}}{\sqrt{1014}}*1,96;0,455-\frac{\sqrt{p(1-p)}}{\sqrt{1014}}*1,96\right]\\ &\qquad\subset\left[0,455-\frac{\sqrt{1/4}}{\sqrt{1014}}*1,96;0,455-\frac{\sqrt{1/4}}{\sqrt{1014}}*1,96\right] = [0,424;0,486]\end{aligned}$$

On peut donc conclure que Sarkozy est bien donné perdant.

Remarque 6 La forme de l'intervalle de confiance mérite d'être commentée. L'intervalle est centré sur l'estimateur, $\overline{X}_n$. La largeur croît avec l'écart-type (plus la distribution est dispersée, plus il faut un intervalle large) et de la confiance (plus on veut de confiance, plus α doit être petit et donc z_α grand) et décroît avec n (plus on a de données, plus l'intervalle se réduit).

3 Espérance d'un échantillon gaussien, variance inconnue

Dans le cas où la variance est inconnue, il faut l'estimer, par exemple par l'estimateur par substitution de la variance. Mais du coup la statistique pivotale change, il faut donc s'appuyer sur une autre théorème :

$$\sqrt{n}\frac{\overline{X}-\mu}{S_n} \rightsquigarrow \mathcal{N}(0;1)$$

où $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Le théorème 7 de Student (c'est-à-dire Gosset, celui de Guinness) sert justement pour cette situation : pour un échantillon gaussien

$$\sqrt{n}\frac{\overline{X}-\mu}{S_n} \sim \mathcal{T}_{n-1}$$

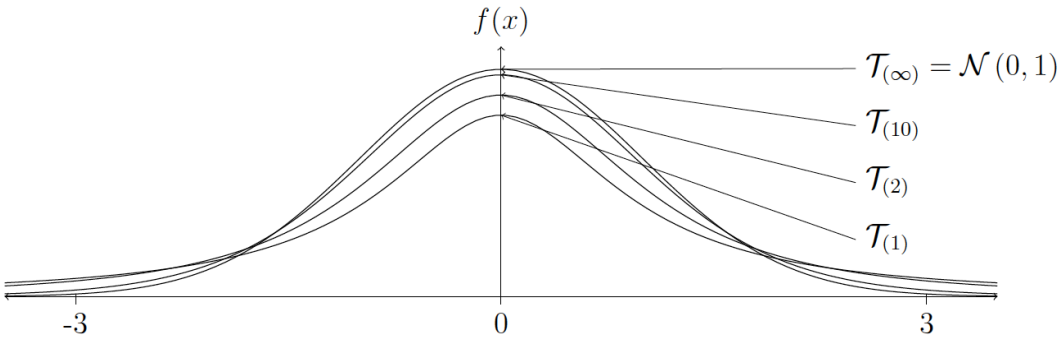
Il faut donc modifier l'intervalle de confiance en conséquence.

Théorème 9 (Espérance d'un échantillon gaussien, variance inconnue) *Si $X_1, \ldots, X_n$ est un échantillon gaussien d'espérance inconnue μ et de variance inconnue $\sigma^2 > 0$, alors*

$$\left[\overline{X} - \frac{S_n}{\sqrt{n}}t_\alpha ; \overline{X} + \frac{S_n}{\sqrt{n}}t_\alpha\right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour $t_\alpha = F^{-1}_{\mathcal{T}_{n-1}}(1-\alpha/2)$

En pratique, l'intervalle est un peu plus large, ce qui est logique puisque à moins d'information que dans le cas où la variance est supposée connue.



Proposition 18

- Si $X \sim \mathcal{T}_n$, alors $\mathbb{E}[X] = 0$ et $\mathrm{Var}(X) = \dfrac{n}{n-2}$.
- Si $\forall n \in \mathbb{N},\ X_n \sim \mathcal{T}_n$ alors $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$.

Exemple Dans le cas des variables de Bernoulli du sondage, on a $S_n^2 = \overline{X}_n(1-\overline{X}_n)$ (il suffit de développer le carré $\sum(X_i - \overline{X}_n)^2$ et remarque que $X_i^2 = X_i$ pour une variable valant 0 ou 1.

L'intervalle de confiance prend donc la forme précédente, mais avec un quantile qui prend réellement en compte qu'on ne connait pas la variance mais qu'on l'estime. Du coup ce n'est plus un intervalle par excès mais bien un intervalle exact.

$$\left[0,455 - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{1014}} * 1,96; 0,455 - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{1014}} * 1,96\right]$$
$$\subset \left[0,455 - \frac{\sqrt{0,455 * (1-0,455)}}{\sqrt{1014}} * 1,96; 0,455 - \frac{\sqrt{0,455 * (1-0,455)}}{\sqrt{1014}} * 1,96\right]$$
$$= [0,424; 0,486]$$

Le résultat est le même 0.1% près.

Si l'intervalle est exact, c'est encore sous l'hypothèse gaussienne, dont on a dit qu'elle ne correspondait pas aux données du sondage (Bernoulli). Il reste donc à lever cette hypothèse.

4 Espérance d'un échantillon non gaussien

On lève cette fois-ci l'hypothèse gaussienne. On utilise alors le théorème centrale limite pour se ramener à la situation gaussienne, mais on obtient un intervalle de confiance asymptotique.

Soit un échantillon i.i.d. d'espérance μ et de variance connue σ^2, on à d'après le TCL :

$$\sqrt{n}\frac{\overline{X}-\mu}{\sigma} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0;1)$$

Si la variance est inconnue, il faut l'estimer par S_n^2 et utiliser le Théorème 5 de Slutsky :

$$\sqrt{n}\frac{\overline{X}-\mu}{S_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0;1)$$

Dans les deux cas, on revient aux quantiles de la loi normale centrée réduite.

Théorème 10 (Espérance d'une loi inconnue) *Si $X_1, \ldots, X_n$ est un échantillon i.i.d. d'espérance inconnue μ.*

1. *Si la variance $\sigma^2 > 0$ est connue, alors*

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}}z_\alpha; \overline{X} + \frac{\sigma}{\sqrt{n}}z_\alpha\right]$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$:

$$\mathbb{P}\left(\left[\overline{X} - \frac{\sigma}{\sqrt{n}}z_\alpha; \overline{X} + \frac{\sigma}{\sqrt{n}}z_\alpha\right] \ni \mu\right) \xrightarrow{n\to\infty} 1-\alpha$$

2. Si la variance est inconnue, alors

$$\left[\overline{X} - \frac{S_n}{\sqrt{n}} z_\alpha ; \overline{X} + \frac{S_n}{\sqrt{n}} z_\alpha\right]$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$ *pour* $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$ *:*

$$\mathbb{P}\left(\left[\overline{X} - \frac{S_n}{\sqrt{n}} z_\alpha ; \overline{X} + \frac{S_n}{\sqrt{n}} z_\alpha\right] \ni \mu\right) \xrightarrow{n\to\infty} 1-\alpha$$

5 Taille d'échantillon et précision

Une autre façon d'utiliser les intervalles de confiance est d'anticiper la question de la précision nécessaire. Avant de faire un sondage, peut-on savoir combien il faudra de réponses pour obtenir une précision, par exemple de 1% ? Il suffit alors de s'assurer que la largeur de l'intervalle de confiance sera inférieur à 2*1% (puisqu'il est centré sur l'estimateur).

Théorème 11 (Taille d'un sondage électoral) *Soit un échantillon i.i.d. d'espérance inconnue et de variance* σ^2 *(connue ou inconnue), alors pour obtenir un estimateur de précision* h *avec une confiance asymptotique* $1-\alpha$*, il faut*

$$n \geq \frac{\sigma^2 t_\alpha^2}{h^2}$$

La taille de l'échantillon nécessaire est inversement proportionnelle au carré de la précision recherchée.

Démonstration *On utilise le Théorème 10 et on a comme intervalle :*

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_\alpha ; \overline{X} + \frac{\sigma}{\sqrt{n}} z_\alpha\right]$$

La précision est donc $h = \frac{\sigma}{\sqrt{n}} z_\alpha$*, d'où on tire* n *en fonction de* h *:*

$$n = \frac{\sigma^2 t_\alpha^2}{h^2}.$$

Or n *doit être entier, il faut donc arrondir au-dessus.*

Exemple Dans le cas du sondage électoral, on a vu que la variance était au pire égale à 1/4 (le pire étant logiquement le cas où les deux candidat·es sont impossibles à départager, à 50% chacun·e). Il faut donc, pour une précision de un point de pourcentage et une confiance asymptotique de 95%,

$$n \geq \frac{1/4\ (1,96)^2}{(0,01)^2} = 9\ 604.$$

Pour une précision de 3 points de pourcentage

$$n \geq \frac{1/4\ (1,96)^2}{(0,03)^2} = 1\ 067.$$

J'invite le·a lecteur·trice à se référer à Bourdieu (1973), *L'opinion publique n'existe pas*, disponible sur `http://www.acrimed.org/L-opinion-publique-n-existe-pas` .

B Premiers tests

Étant donné un échantillon aléatoire $X = (X_1, \ldots, X_n)$ issu d'une loi $\mathcal{L}$ de paramètre inconnu, les données observées $x = (x_1, \ldots, x_n)$ nous permettent-elles de contredire une hypothèse a priori sur la valeur de θ ? En d'autres termes, les données observées sont-elles compatibles ou incompatibles avec notre hypothèse ? Afin de répondre à la question et décider de rejeter ou de ne pas rejeter notre hypothèse, on doit se fixer une règle de décision. Cette règle de décision sera appelée un test de notre hypothèse. Cette règle de décision ne pourra jamais être fiable à 100 %, et il existera toujours un risque d'erreur. on doit donc déterminer notre règle de décision de façon à contrôler pour ce risque d'erreur.

Notre règle de décision peut aboutir à deux conclusions : le rejet ou le non-rejet de l'hypothèse. Remarquez que l'on ne parle pas d'acceptation de l'hypothèse. En effet, le fait que nos données soient compatibles avec notre hypothèse (ou plutôt n'y soit pas incompatible) ne veut pas dire que cette dernière est exacte. Il se peut que la réalité soit différente de ce que notre hypothèse annonce, bien que nos données ne permettent pas de la rejeter. Pensez à la situation d'un procès criminel. On se demande si les informations recueillies par l'instruction permettent de rejeter l'hypothèse d'innocence de l'accusé. Si on trouve des preuves de l'implication de l'accusé, alors on peut rejeter l'hypothèse de son innocence (car les données observées – les preuves – sont incompatibles avec son innocence). Par contre, en l'absence de preuve on ne peut pas rejeter l'hypothèse d'innocence. Cela ne veut pas dire que l'accusé est effectivement innocent, simplement que les données ne permettent pas de rejeter l'hypothèse d'innocence.

Pour prendre un exemple plus économique, formulons l'hypothèse suivante : « Le salaire moyen des femmes est le même que celui des hommes parmi la population des salariés de l'ENS Cachan ». Si l'on note μ_F le salaire moyen des femmes de l'ENS Cachan et μ_H celui des hommes de l'établissement, cela signifie donc mathématiquement qu'on fait l'hypothèse $\mu_H = \mu_F$ ou $\mu_F - \mu_H = 0$. On appelle alors *hypothèse nulle*, que l'on note usuellement (H_0), l'hypothèse que l'on souhaite tester. On nomme *hypothèse alternative*, que l'on note usuellement (H_1), ce qui arrive lorsque H_0 est fausse. Attention, dans le cas général, $H_1 \subset (H_0)^c$ mais il n'y a pas nécessairement égalité. En effet, on peut considérer comme alternative à (H_0) : $\mu_H = \mu_F$, soit (H_1) : $\mu_H \neq \mu_F$, soit (H_1') : $\mu_H > \mu_F$ qui permet d'exclure a priori du domaine du possible des salaires féminins en moyenne supérieur à ceux des hommes.

1 Définitions

Les tests statistiques paramétriques peuvent être vus comme un problème d'estimation. En effet, tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ (i.e. prendre une décision parmi deux hypothèses H_0 et H_1), revient à estimer la fonction qui vaut 1 si $\theta \in \Theta_0$ et 0 sinon : $\mathbb{1}_{\Theta_0}(\theta)$.

Définition 41 (Test) *La fonction :* $\Phi : \mathcal{X} \to \{0, 1\}$ *est un test de niveau* $1 - \alpha$ *de l'hypothèse nulle* H_0 *contre une alternative* H_1*. Si* $\Phi = 0$*, l'hypothèse nulle* H_0 *est acceptée et l'alternative rejetée alors que si* $\Phi = 1$ *l'hypothèse nulle* H_0 *est rejetée et l'alternative* H_1 *est acceptée.*

Définition 42 (Hypothèses simples et composites) *Une hypothèse est dite simple quand l'ensemble des valeurs considérées pour θ se réduit à un singleton. Si $\Theta_0 = \{\theta_0\}$ alors l'hypothèse nulle est une hypothèse simple. Sinon, l'hypothèse est composite : plusieurs valeurs sont possibles pour θ.*

Chaque règle de décision comporte un risque d'erreur. On ne sait pas quel est le vrai état de la nature (on ne sait pas si c'est H_0 ou H_1 qui est vraie) ; mais on va prendre une décision de rejet ou de non-rejet de H_0 sur la base d'informations issues d'une expérience aléatoire (le tirage de l'échantillon). L'échantillon observé ne va donc pas nécessairement refléter exactement la population mère. En appliquant notre règle de décision, il y a donc deux façons de se tromper :

Définition 43 (Erreur de première espèce et niveau d'un test) *L'erreur de première espèce α d'un test Φ est la probabilité de commettre une erreur de type 1, c'est-à-dire de décider de rejeter H_0 alors qu'elle est vraie :*

$$\alpha = \mathbb{P}_{H_0}(\Phi = 1)$$

Lorsque l'hypothèse nulle H_0 est composite, on se place dans le pire des cas :

$$\alpha = \sup\ \mathbb{P}_{H_0}(\Phi = 1) = \sup_{\theta \in \Theta_0} \mathbb{P}(\textit{rejet de } H_0 \,|H_0\ \ \textit{est vraie})$$

On appelle niveau, et on note $1-\alpha$, la probabilité d'accepter à raison l'hypothèse nulle :

$$1 - \alpha = \mathbb{P}_{H_0}(\Phi = 1)$$

Si l'hypothèse nulle est composite :

$$1 - \alpha = \inf_{H_0} \mathbb{P}(\Phi = 0) = \inf_{\Theta_0} \mathbb{P}(\textit{acceptation de } H_0 \,|H_0\ \ \textit{est vraie})$$

Remarque 7 Dans le cas où H_0 est une hypothèse simple alors son niveau est égal à son risque de première espèce.

Définition 44 (Erreur de seconde espèce et puissance du test) *L'erreur de seconde espèce β d'un test Φ est la probabilité de commettre une erreur de type 2, c'est-à-dire de décider d'accepter H_0 alors qu'elle est fausse :*

$$\beta = \mathbb{P}_{H_1}(\Phi = 0)$$

Lorsque l'hypothèse alternative H_1 est composite, on se place dans le pire des cas :

$$\beta = \sup\ \mathbb{P}_{H_1}(\Phi = 0) = \sup_{\theta \in \Theta_1} \mathbb{P}(\textit{acceptation de } H_0 \,|H_0\ \ \textit{est faussse})$$

On appelle puissance, et on note $1-\beta$, la probabilité de rejeter à raison l'hypothèse nulle :

$$1 - \beta = \mathbb{P}_{H_1}(\Phi = 0)$$

Si l'hypothèse alternative est composite :

$$1 - \beta = \inf_{H_1} \mathbb{P}(\Phi = 1) = \inf_{\Theta_1} \mathbb{P}(\textit{rejet de } H_0 \,|H_0\ \ \textit{est fausse})$$

	$\Phi = 0$	$\Phi = 1$
H_0	On accepte a raison $\mathbb{P}_{H_0}(\phi = 0) = 1 - \alpha$ Niveau du test	On rejette à tord $\mathbb{P}_{H_0}(\phi = 1) = \alpha$ Erreur de type I
H_1	On accepte à tord $\mathbb{P}_{H_1}(\phi = 0) = \beta$ Erreur de type II	On rejette à raison $\mathbb{P}_{H_1}(\phi = 1) = 1 - \beta$ Puissance du test

TABLE 3.1 – Principales situations d'un test. Lire le tableau en ligne.

Définition 45 (Région critique) *La région critique (ou de rejet) W associée au test Φ, appelée aussi région de rejet de H_0, est définie par : $W = \Phi^{-1}(\{1\})$. C'est la région de $\mathcal{X}^n$ constituée des données qui conduisent à rejeter H_0. On la note :*

$$W = \{(X_1, \ldots, X_n) \in \mathbb{R}^n : \text{ au vu de } (X_1, \ldots, X_n) \text{ on rejette } H_0\}$$

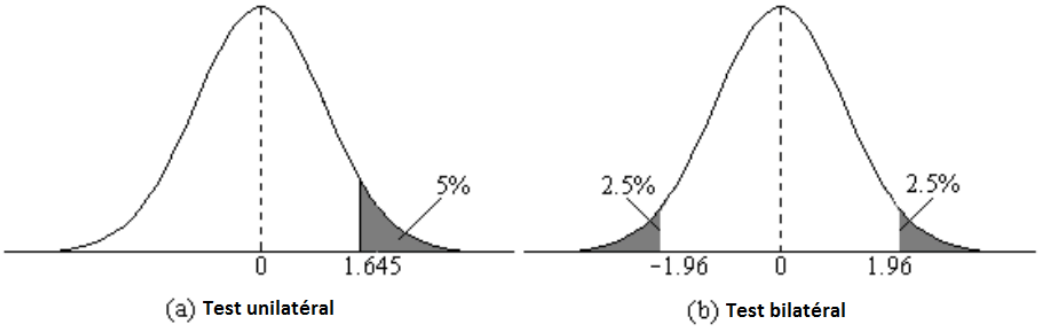
En pratique, le test dépend en général des données à travers l'estimateur du paramètre sur lequel porte le test. Pour un test sur une espérance, on construit une fonction Φ dépendant de la moyenne $\overline{X}$. La région de rejet W peut alors s'écrire comme une fonction de $\overline{X}$ plutôt que comme fonction de chacun des X_i.

La taille de la région W dépend de l'erreur de première espèce α. Plus α est grand, plus l'on rejette souvent et plus la région de rejet est grande. Si $\alpha = 0,05$, cela signifie que si H_0 est vraie, la région de rejet contient l'estimateur avec une probabilité 5% :

$$\mathbb{P}_{H_0}(\Phi = 1) = \mathbb{P}_{H_0}(\overline{X} \in W) = \int_W d\mathbb{P}_{\overline{X}|H_0}$$

Autrement dit, si l'on représente la densité de l'estimateur, l'aire au dessus de W est égale à 0,05. Ceci peut se représenter à l'aide de la densité de l'estimateur.

La forme de la zone de rejet dépend de l'alternative : si l'alternative est symétrique ($H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$) alors la zone est en deux morceaux, elle est dite bilatérale ; si l'alternative est asymétrique ($H_0 : \theta = 0$ contre $H_1 : \theta > 0$) alors la région de rejet est d'un seul morceau, elle est dite unilatérale. Les régions de rejet unilatérales et bilatérales avec une erreur de première espèce $\alpha = 0,05$ sont illustrées sur la figure ci-dessous. Les régions diffèrent par leur emplacement, mais leur taille totale est la même.



(a) Test unilatéral (b) Test bilatéral

Si le test statistique donne une valeur qui se trouve dans la région de rejet, il faut rejeter H_0. Ce processus décisionnel repose sur un raisonnement très simple. Si, sous l'hypothèse nulle, la probabilité de se trouver dans la zone de rejet est très faible, l'apparition effective d'une réalisation de l'échantillon dans cette zone peut s'expliquer de deux manières : premièrement en décidant que l'hypothèse nulle est fausse, et deuxièmement, en décidant qu'un événement rare et improbable s'est produit. Dans le processus de décision, nous choisissons la première de ces explications. De temps en temps, c'est bien entendu la seconde qui est la bonne. De fait, la probabilité que la deuxième explication soit la bonne est donnée par α, car le fait de rejeter l'hypothèse H_0 alors qu'elle est vraie est une erreur de type I.

Dans la pratique, on se rend compte que certaines réalisations de l'échantillon se situent très loin de la zone de rejet, ou à la frontière, ou au contraire en plein dans la zone de rejet. Dans le premier cas, on accepte « solidement » H_0 et dans le dernier cas, on la rejette avec force. Mais dans le cas intermédiaire, il peut être insatisfaisant d'en rester au résultat du test (acceptation ou rejet) alors qu'il correspond à une décision « de justesse ». La p-value sert à exprimer quantitativement si l'on accepte fortement ou de justesse :

Définition 46 (p-value) *soit Φ_α une famille de test de H_0 contre H_1 indexé par leur erreur de première espèce :*

$$\forall \alpha \in [0;1],\ \mathbb{P}_{H_0}\left(\Phi_\alpha(X_1,\ldots,X_n)=0\right)=1-\alpha.$$

La p-value de cette famille de test est une variable aléatoire à valeur dans $[0;1]$ définie par :

$$p=\inf_{\alpha\in[0;1]}\left\{\alpha,\ \Phi_\alpha(X_1,\ldots,X_n)=1\right\}$$

Plus α est petite, plus on a de chances d'accepter H_0 (pour $\alpha=0$, on accepte toujours H_0) et donc moins on a de chance de rejeter. Au cours de cette réduction de α vers 0, p est donc la dernière valeur de α pour laquelle on rejette, donc la plus petite.

Si on rejette jusqu'à une α très petite, par exemple 1%, c'est qu'il faut un niveau très élevé (99%) pour accepter H_0, donc que cette hypothèse est finalement peu crédible. Si l'on avait fait un test à 95%, la réalisation de l'échantillon serait tombée en plein dans la zone de rejet.

Si la p-value vaut à peu près 5%, le test à 95% tranche (on accepte ou on rejette H_0) mais la situation est en réalité tangente.

Enfin, si la p-value est très grande, par exemple 50%, c'est qu'on peut accepter avec un niveau très réduit, H_0 est donc une hypothèse solide. La réalisation de l'échantillon est loin à l'extérieur de la zone de rejet du test de niveau 95%.

2 Test bilatéral pour l'espérance d'une gaussienne

Soit $X=(X_1,\ldots,X_n)$ un échantillon i.i.d. de taille n issu d'une loi normale $\mathcal{N}\left(\mu,\sigma^2\right)$. Le test :

$$\begin{cases} H_0 & : \mu=\mu_0 \\ H_1 & : \mu\neq\mu_0 \end{cases}$$

est un test bilatéral.

On va se ramener à la construction d'un intervalle de confiance pour μ : tester au niveau $1-\alpha$ si μ_0 est plausible au vu des données revient à vérifier si l'intervalle de confiance de niveau $1-\alpha$ contient μ_0.

Exemple Si on reprend les salarié·es de l'ENS Cachan, on peut se demander si leur salaire net est équivalent à celui de la fonction publique au niveau national. Selon l'INSEE, « en 2014, un salarié de la fonction publique d'État perçoit en moyenne un salaire net de 2 477 euros par mois en équivalent temps plein ». On pose donc $\mu_0 = 2\,477$.

2.1 Cas où la variance est connue

Le théorème 8 nous donne comme intervalle de confiance pour l'espérance d'un échantillon gaussien de variance connue :

$$\mathbb{P}\left(\mu \in \left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_\alpha ; \overline{X} + \frac{\sigma}{\sqrt{n}} z_\alpha\right]\right) = 1-\alpha$$

pour $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$, à partir de la connaissance de la loi de

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

Pour se placer dans le contexte du test, il faut supposer H_0 vraie et construire une fonction test Φ telle que

$$\mathbb{P}_{H_0}\left(\Phi = 0\right) = 1-\alpha.$$

Lorsque l'on suppose que $\mu = \mu_0$, on voit qu'il suffit de définit Φ telle que $\Phi = 0$ si et seulement si μ_0 est dans l'intervalle de confiance :

$$\Phi = \mathbb{1}_{\mu_0 \notin \left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_\alpha ; \overline{X} + \frac{\sigma}{\sqrt{n}} z_\alpha\right]}$$

On réécrit cette fonction plus simplement pour obtenir le théorème suivant :

Théorème 12 (Espérance une gaussienne, σ^2 connue) *Soit $X_1, \dots, X_n$ un échantillon i.i.d. gaussien d'espérance μ et de variance σ^2 connue. On peut tester H_0 : $\mu = \mu_0$ contre H_1 : $\mu \neq \mu_0$ au niveau $1-\alpha$ à l'aide du test*

$$\Phi = \mathbb{1}_{\sqrt{n}\frac{|\overline{X}-\mu_0|}{\sigma} \geq z_\alpha}$$

où $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$.

Démonstration *Pour vérifier que c'est bien un test de niveau $1-\alpha$, il suffit de calculer $\mathbb{P}_{H_0}(\Phi = 0)$. Or $\Phi = 0$ si l'événement dans l'indicatrice n'est pas vérifié :*

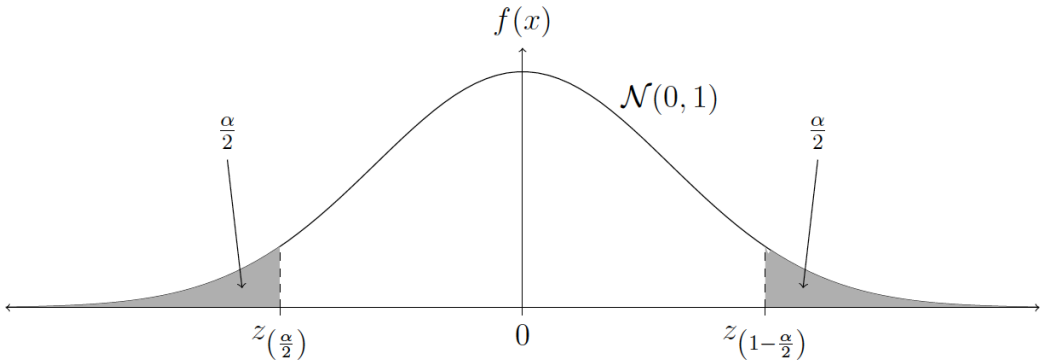
$$Z_n = \sqrt{n}\frac{|\overline{X} - \mu_0|}{\sigma} < z_\alpha.$$

Sous l'hypothèse nulle, cette statistique pivotale Z_n est $\mathcal{N}(0;1)$ et donc

$$\begin{aligned}\mathbb{P}_{H_0}(\Phi = 0) &= \mathbb{P}_{H_0}(|Z_n| < z_\alpha) = \mathbb{P}_{H_0}(Z_n < z_\alpha) - (1 - \mathbb{P}_{H_0}(Z_n < z_\alpha)) \\ &= 2 * \mathbb{P}_{H_0}(Z_n < z_\alpha) - 1 = 2 * F_{\mathcal{N}(0;1)}\left(F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)\right) - 1 \\ &= 2 * (1-\alpha/2) - 1 = 1 - \alpha.\end{aligned}$$

La région de rejet s'écrit donc :

$$W = \left\{\overline{X} \in \mathbb{R} : \left|\sqrt{n}\frac{\overline{X} - \mu_0}{\sigma}\right| \geq z_\alpha\right\}.$$



2.2 Cas où la variance est inconnue

Dans le cas où la variance σ^2 est inconnue, elle est estimée par S_n^2 et ce n'est plus un quantile de gaussienne mais de loi de Student qui intervient dans l'intervalle de confiance.

On retrouve les mêmes modifications pour le test.

Théorème 13 (Espérance une gaussienne, σ^2 inconnue) *Soit $X_1, \ldots, X_n$ un échantillon i.i.d. gaussien d'espérance μ et de variance σ^2 inconnue. On peut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ au niveau $1-\alpha$ à l'aide du test*

$$\Phi = \mathbb{1}_{\sqrt{n}\frac{|\overline{X}-\mu_0|}{S_n} \geq t_\alpha}$$

où $t_\alpha = F^{-1}_{\mathcal{T}_{n-1}}(1-\alpha/2)$.

Démonstration *La démonstration est la même, avec $T_n \sim \mathcal{T}_{n-1}$ d'après le théorème de Student.*

2.3 Cas où la loi est inconnue

Si la loi est inconnue, on utilise le théorème central limite et on obtient un test de confiance asymptotique $1-\alpha$, qui fait intervenir un quantile de gaussienne que la variance soit connue ou non.

Théorème 14 (Espérance une loi inconnue) *Soit $X_1, \dots, X_n$ un échantillon i.i.d. d'espérance μ et de variance σ^2 inconnue. On peut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ au niveau asymptotique $1-\alpha$ à l'aide des tests*

$$\Phi = \mathbb{1}_{\sqrt{n}\frac{|\overline{X}-\mu_0|}{\sigma} \geq z_\alpha} \qquad \text{si } \sigma^2 \text{ est connue ;}$$

$$\Phi = \mathbb{1}_{\sqrt{n}\frac{|\overline{X}-\mu_0|}{S_n} \geq z_\alpha} \qquad \text{si } \sigma^2 \text{ est inconnue.}$$

Dans les deux cas $z_\alpha = F^{-1}_{\mathcal{N}(0;1)}(1-\alpha/2)$.

Démonstration *On reprend exactement les démonstrations précédentes, avec $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0;1)$ par le théorème central limite (et le lemme de Slutsky lorsque la variance est inconnue).*

3 Test bilatéral de la variance d'une loi normale

Soit $X = (X_1, \dots, X_n)$ un échantillon i.i.d. de taille n issu d'une loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$. Au vu d'un échantillon de n réalisations indépendantes X_i, on veut choisir entre les deux hypothèses :

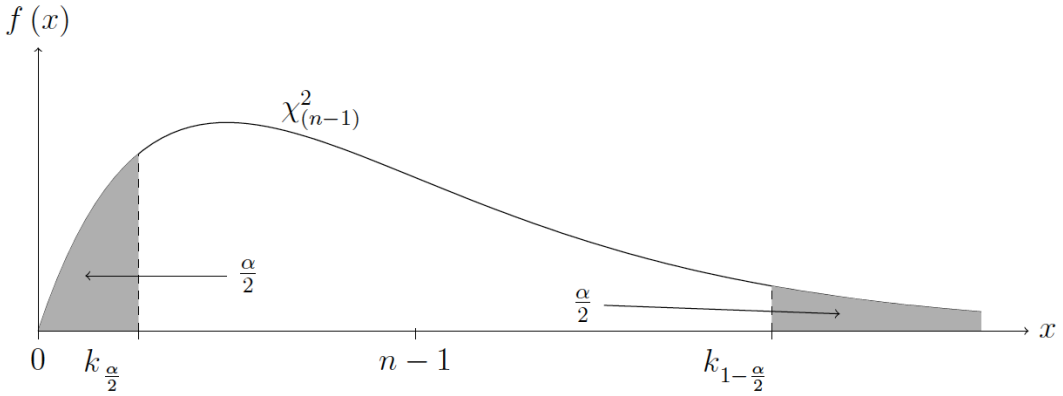
$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

On sait alors que $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$ est une statistique pivotale pour σ^2. On peut alors écrire :

$$\Pr\left(k_{\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma_0^2} \leq k_{1-\alpha/2} \middle| \, \sigma^2 = \sigma_0^2\right) = 1-\alpha$$

La zone de rejet de notre test s'écrit alors :

$$W = \left\{(X_1, \dots, X_n) \in \mathbb{R}^n : \frac{(n-1)S_n^2}{\sigma_0^2} \notin \left[k_{\alpha/2}, k_{1-\alpha/2}\right]\right\}$$



4 Tests de comparaison de deux espérances

On dispose de deux échantillons indépendants $X = (X_1, \ldots, X_n)$ et $Y = (Y_1, \ldots, Y_n)$ tels que $X_i \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ et $Y_i \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$. On veut tester :

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

4.1 Cas où les variances sont connues

Dans ce cas, la différence deux moyennes empiriques $\overline{X}_n$ et $\overline{Y}_n$ va suivre une loi normale :

$$\left(\overline{X}_n - \overline{Y}_n\right) \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

On a donc :

$$\frac{\overline{X}_n - \overline{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1)$$

qui est une statistique pivotale pour $\mu_X - \mu_Y$. Si H_0 est vraie, alors :

$$\frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$$

On doit donc avoir :

$$\Pr_{\mu_X = \mu_Y}\left(\left|\frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}\right| \leq z_\alpha\right) = 1 - \alpha.$$

Notre région de rejet s'écrit :

$$W = \left\{(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in \mathbb{R}^{n_X + n_Y} : \left|\frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}\right| \geq z_\alpha\right\}$$

4.2 Cas où les variances sont inconnues, mais supposées égales (test de Student)

Dans ce cas, on peut estimer le σ^2 inconnu mais commun aux deux échantillons comme une moyenne pondérée des S_n^2 des deux échantillons :

$$S_{n_X + n_Y}^2 = \frac{(n_X - 1)\, S_{n_X}^2 + (n_Y - 1)\, S_{n_Y}^2}{n_X + n_Y - 2}$$

On peut montrer dans ce cas que :

$$\frac{\overline{X}_n - \overline{Y}_n - (\mu_X - \mu_Y)}{S_{n_X+n_Y}\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}} \sim \mathcal{T}_{(n_X+n_Y-2)}$$

qui est donc une statistique pivotale pour $\mu_X - \mu_Y$. Notre région de rejet s'écrit alors :

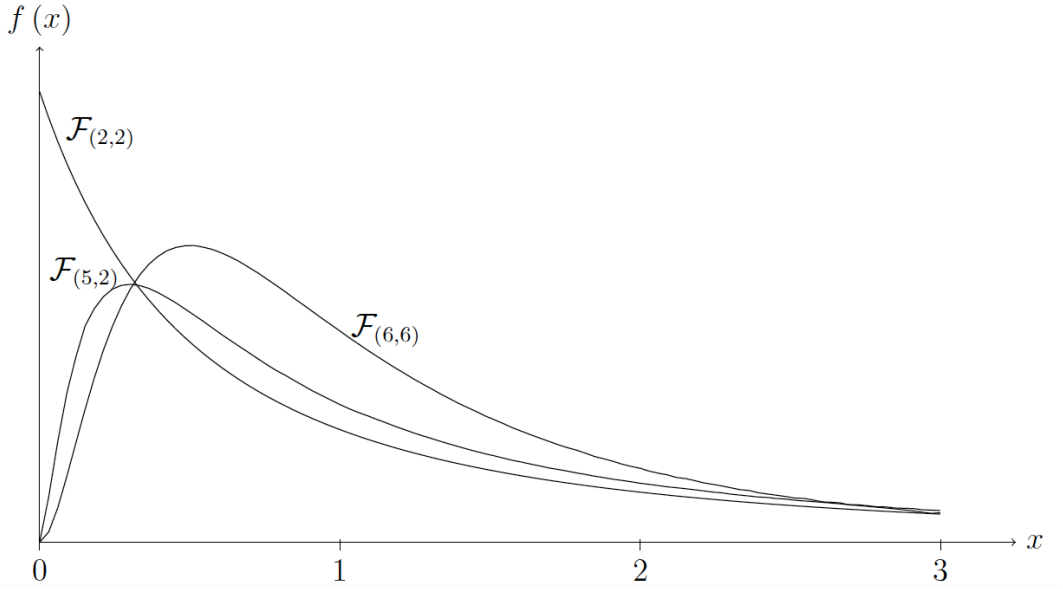
$$W = \left\{ (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in \mathbb{R}^{n_X+n_Y} : \left| \frac{\overline{X}_n - \overline{Y}_n}{S_{n_X+n_Y}\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}} \right| \geq t_\alpha \right\}$$

où t_α est le quantile d'une loi de Student à $(n_X + n_Y - 2)$ degrés de liberté.

5 Tests de comparaisons de deux variances (test de Fisher)

Définition 47 *La loi de Fisher à (n_1, n_2) degrés de liberté, notée $\mathcal{F}_{(n_1,n_2)}$ peut se définir comme le ratio de deux variables aléatoires indépendantes suivant une loi χ^2, divisées chacune par leurs degrés de liberté. Plus précisément : si $X_1 \sim \chi^2_{(n1)}, X_2 \sim \chi^2_{(n2)}$ et que X_1 et X_2 sont indépendantes, alors $\dfrac{X_1/n_1}{X_2/n_2} \sim \mathcal{F}_{(n_1,n_2)}$.*

On note f_p le quantile d'ordre p d'une $\mathcal{F}_{(n_1,n_2)}$, c'est-à-dire la valeur telle que si $X \sim \mathcal{F}_{(n_1,n_2)}$, alors $\Pr(X \leq f_p) = p$.



Théorème 15 (Loi de Fisher-Snedecor (admis)) *Si $X \sim \chi^2_{(n)}$ et $Y \sim \chi^2_{(m)}$, et si X et Y sont indépendantes, alors :*

$$\frac{X/n}{Y/m} \sim \mathcal{F}_{(n,m)}$$

Remarque 8 Cette loi nous sera utile dans le test de comparaison de variances.

On cherche à savoir si deux échantillons indépendants $X = (X_1, \ldots, X_n)$ et $Y = (Y_1, \ldots, Y_n)$ tels que $X_i \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ et $Y_i \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ proviennent de loi de même variance. On teste :

$$\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$$

Le théorème de Fisher-Cochran nous dit que, dans le cas d'échantillon gaussiens de taille n, on a :

$$\frac{(n-1)\, S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Nos deux échantillons étant normaux, on a donc :

$$\begin{cases} \dfrac{(n_X - 1)\, S_{n_X}^2}{\sigma_X^2} \sim \chi^2_{(n_X-1)} \\ \dfrac{(n_Y - 1)\, S_{n_Y}^2}{\sigma_Y^2} \sim \chi^2_{(n_Y-1)} \end{cases}$$

Ces deux variables aléatoires étant indépendantes (car les échantillons sont indépendants), on sait (toujours par le théorème de Fischer-Cochran) que :

$$\frac{\frac{1}{n_X-1}\dfrac{(n_X - 1)\, S_{n_X}^2}{\sigma_X^2}}{\frac{1}{n_Y-1}\dfrac{(n_Y - 1)\, S_{n_Y}^2}{\sigma_Y^2}} = \frac{S_{n_X}^2/\sigma_X^2}{S_{n_Y}^2/\sigma_Y^2} \sim \mathcal{F}_{(n_X-1, n_Y-1)}$$

Sous$H_0 : \sigma_X^2 = \sigma_Y^2$, et on peut alors écrire :

$$\frac{S_{n_X}^2}{S_{n_Y}^2} \underset{H_0}{\sim} \mathcal{F}_{(n_X-1, n_Y-1)}$$

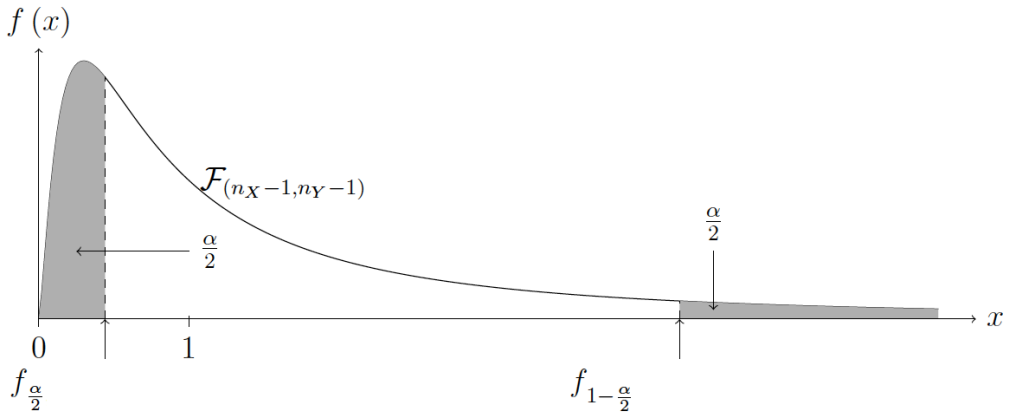
On doit donc avoir :

$$\Pr\left(f_{\alpha/2} \leq \frac{S_{n_X}^2}{S_{n_Y}^2} \leq f_{1-\alpha/2} \middle| \sigma_X^2 = \sigma_Y^2 \right) = 1 - \alpha$$

On va donc rejeter H_0 si $\dfrac{S_{n_X}^2}{S_{n_Y}^2} \leq f_{\alpha/2}$ ou si $\dfrac{S_{n_X}^2}{S_{n_Y}^2} \geq f_{1-\alpha/2}$:

$$W = \left\{ (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in \mathbb{R}^{n_X+n_Y} : \frac{S_{n_X}^2}{S_{n_Y}^2} \notin \left[f_{\alpha/2}, f_{1-\alpha/2} \right] \right\}$$

Remarque 9 Cette fois-ci la loi considérée n'étant plus symétrique notre région de rejet doit prendre en compte les deux bornes.



Remarque 10 Les tables de la loi de Fisher ne donnent en général que les valeurs des bornes de la partie droite de la région de rejet $\left(f_{\left(1-\frac{\alpha}{2},n_X-1,n_Y-1\right)}\right)$. Pour trouver la borne de la zone de gauche $\left(f_{\left(\frac{\alpha}{2},n_X-1,n_Y-1\right)}\right)$ il faut utiliser l'égalité :

$$f_{\left(\frac{\alpha}{2},n_X-1,n_Y-1\right)} = \frac{1}{f_{\left(1-\frac{\alpha}{2},n_Y-1,n_X-1\right)}}$$

Chapitre 4

Contraste du χ^2

Le contraste du χ^2 permet de mesurer la dissimilarité entre deux distributions. En ce sens, elle répond à une logique assez distincte de celle de la substitution. Dans le cas de la substitution, on se concentre sur le paramètre (espérance ou variance) en cherchant une méthode d'estimation (et de construction d'intervalle de confiance ou de test) qui dépende le moins possible d'hypothèses sur la loi des données. Le contraste du χ^2 s'intéresse au contraire directement aux distributions : la distribution empirique des données est-elle proche d'une certaine loi ? d'une famille de loi ? Les distributions empiriques de deux échantillons sont-elles identiques ?

Le contraste du χ^2 peut servir à construire des estimateurs et des intervalles de confiance, mais c'est surtout dans le cadre des tests qu'il est utilisé.

A Test d'adéquation du χ^2

1 Adéquation à une loi

Le test d'adéquation vise à comparer la distribution empirique des données, notée $\mathbb{P}_n$, à une loi donnée $\mathbb{P}_0$, c'est-à-dire à tester s'il est crédible que les observations soient tirées suivant cette loi $\mathbb{P}_0$. $\mathbb{P}_n$ est la loi uniforme discrète sur l'échantillon, c'est-à-dire la loi d'un dé à n faces numérotées $x_1, \ldots, x_n$.

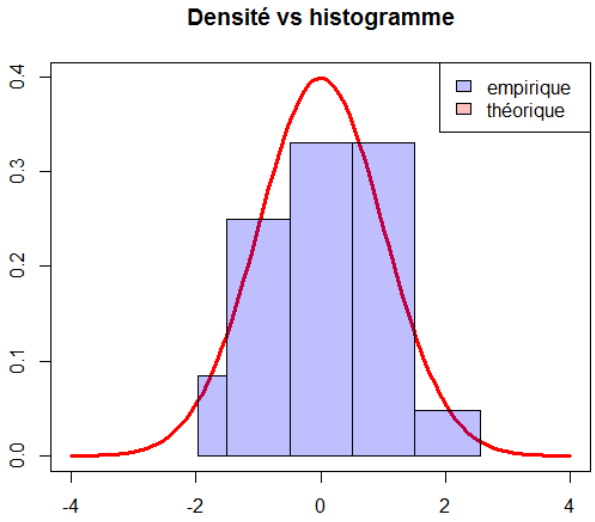


FIGURE 4.1 – Histogramme d'un échantillon de 100 gaussiennes centrées réduites i.i.d. et densité.

Exemple (Genre des salarié·es) Dans le cas des salarié·s de Cachan, on peut tester si le genre suit une loi $Ber(1/2)$, c'est-à-dire si la représentation des genres est significativement distincte de 50% - 50%.

Parmi les $n = 100$ individus de l'échantillon, on compte 42 femmes. $\mathbb{P}_n$ est donc la loi d'un dé à 100 faces avec 42 faces « femme » et 58 faces « homme ». C'est évidement la même loi qu'une pièce déséquilibrée avec une probabilité 0,48 de tirer « femme » et 0,58 de tirer « homme » : une loi $Ber(0,48)$.

Il s'agit donc de dire si une loi $Ber(0,48)$ diffère suffisamment d'une $Ber(0,5)$, pour $n = 100$, pour rejeter l'hypothèse nulle.

Formellement, on observe un échantillon $(X_1, \dots, X_n)$ i.i.d. de loi $\mathbb{P}$ à valeurs dans $\mathcal{X}$. On cherche à tester :

$$\begin{cases} H_0 : & \mathbb{P} = \mathbb{P}_0 \\ H_1 : & \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

La méthode consiste à comparer l'histogramme observée des fréquences empiriques à l'histogramme théorique de la loi $\mathbb{P}_0$. On doit donc découper l'ensemble $\mathcal{X}$ des valeurs possibles en K classes.

Définition 48 *On se donne une partition de $\mathcal{X}$ en K classes :*

$$\mathcal{X} = \bigcup_{k=1}^{K} C_k$$

On note

- $\hat{p}_k = \mathbb{P}_n(C_k)$ *la fréquence observée de la classe* C_k.

- $p_k^0 = \mathbb{P}_0(C_k)$ *la probabilité théorique (ou fréquence espérée) de la classe* C_k *sous* H_0.
- $\hat{n}_k = np_k$ *l'effectif observé dans la classe* C_k..
- $n_k^0 = np_k^0$ *l'effectif théorique (espéré) de la classe* C_k.

Le contraste du χ^2 permet ensuite de mesurer l'écart entre les deux histogrammes, donc entre les $\hat{n}_k$ et les n_k^0 ou de façon équivalente entre les $\hat{p}_k$ et les p_k^0.

Définition 49 *Le contraste du* χ^2*, noté* D_n*, entre la distribution observée* $\mathbb{P}_n$ *et la distribution théorique* $\mathbb{P}_0$ *est défini de façon équivalente par :*

$$D_n = D_n(\mathbb{P}_n, \mathbb{P}_0) = \sum_{k=1}^{K} \frac{(\hat{n}_k - n_k^0)^2}{n_k^0} = n \sum_{k=1}^{K} \frac{(\hat{p}_k - p_k^0)^2}{p_k^0}$$

Attention, il y a un n en facteur dans la définition en fréquences. Dans les deux cas, la somme porte sur les K classes et non sur le nombre n d'observations.

Théorème 16 *Sous* $H_0 : \mathbb{P} = \mathbb{P}_0$*, la statistique de test* D_n *converge en loi vers un* χ^2 *à* $(K-1)$ *degrés de liberté. La fonction*

$$\Phi = \mathbb{1}_{D_n > k_\alpha}$$

est un test de niveau asymptotique $1-\alpha$ *pour* $k_\alpha = F^{-1}_{\chi^2_{K-1}}(1-\alpha)$*. La région critique est donc la suivante :*

$$W = \left\{ D_n > F^{-1}_{\chi^2_{K-1}}(1-\alpha) \right\}$$

Remarque 11

- Le choix et le nombre de classes est arbitraire. Cependant pour que la méthode fonctionne bien, il faut pouvoir calculer facilement les fréquences théoriques et avoir des effectifs théoriques dans chacune des classes pas trop faibles, typiquement $n_k^0 \geq 5$. Si ce n'est pas le cas, on peut regrouper les classes contiguës afin d'avoir un effectif suffisant. La valeur de K intervenant dans le quantile est celle obtenue après les éventuels regroupements.
- On met toujours les effectifs ou les fréquences théoriques au dénominateur pour éviter de diviser par 0.
- Le -1 intervenant dans le nombre de degrés de liberté du χ^2 s'explique par le fait que lorsqu'on partitionne $\mathcal{X}$ en K classes, on n'en choisit en fait que $K-1$, la dernière étant donnée par ce qu'il reste. Ou, si l'on préfère, on choisit K classes sous la contrainte de dimension 1 $\bigcup_{k=1}^{K} C_k = \mathcal{X}$.
- La probabilité de rejeter le modèle théorique à tort est connue (α) ; par contre on ne peut calculer la probabilité d'accepter le modèle théorique à tort. D'autre part, un test, faute de preuves expérimentales suffisantes, se rabat sur l'hypothèse H_0. On se gardera d'utiliser les facilités des logiciels pour tester à tort et à travers une multitude de modèles théoriques. Il faut au préalable avoir de bonnes raisons pour soupçonner tel ou tel modèle théorique.

Exemple (Genre des salarié·es) Dans cet exemple, on n'a pas le choix des classes : $C_1 = \{0\}$ (hommes) et $C_2 = \{1\}$ (femmes).

On observe 42 femmes parmi les 100 individus de l'échantillon donc $\hat{n}_1 = 58$ et $\hat{n}_2 = 42$.

Sous l'hypothèse nulle, la répartition est uniforme : $p_1^0 = 0,5$ et donc $n_1^0 = n \times 0,5 = 50$; $p_2^0 = 0,5$ et $n_2^0 = 50$.

On calcule

$$D_n = \frac{(58-50)^2}{50} + \frac{(42-50)^2}{50} = 2,56.$$

et on compare au quantile à 95% d'un χ^2_{2-1} : $k_{0,05} = 3,84$. On a alors

$$\Phi = \mathbb{1}_{D_n > k_\alpha} = \mathbb{1}_{2,56>3,84} = 0.$$

On accepte l'adéquation : l'échantillon ne permet pas d'exclure la parité de la population des salarié·es.

2 Adéquation à une famille de lois

On observe toujours un échantillon $(X_1, \dots, X_n)$ i.i.d. de loi $\mathbb{P}$ à valeurs dans $\mathcal{X}$ et on élargie l'hypothèse nulle à une famille de lois indexée par r paramètres :

$$\begin{cases} H_0 : & \exists \theta \in \Theta,\ \mathbb{P} = \mathbb{P}_\theta \\ H_1 : & \forall \theta \in \Theta,\ \mathbb{P} \neq \mathbb{P}_\theta \end{cases}$$

Exemple (Salaires log-normaux) On peut tester si le log des salaires des cachanais·es suit bien une loi log-normale. Il faudra donc comparer la répartition des $\log(R)$ à une répartition gaussienne.

On a donc toute une famille de lois théoriques possibles. On calcule alors le contraste entre la distribution empirique et le meilleur représentant $\mathbb{P}_{\hat{\theta}}$ au sein de la famille $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$. Ce $\mathbb{P}_{\hat{\theta}}$ pourrait être choisi en optimisant en θ D_n, mais il est bien plus simple (et asymptotiquement équivalent) d'estimer θ par substitution.

Une fois estimé θ par $\hat{\theta}$, on reproduit le test d'adéquation à une loi, avec simplement une modification dans le nombre de degré de liberté de la loi du χ^2 intervenant dans le quantile. Il faut en effet prendre en compte le fait qu'il est plus facile que les données soient proches, par exemple, de l'ensemble des gaussiennes $\{\mathcal{N}(\mu; \sigma^2),\ (\mu, \sigma^2) \in (\mathbb{R} \times \mathbb{R}_*^+)\}$ que des gaussiennes de variance 1 $\{\mathcal{N}(\mu; 1),\ \mu \in \mathbb{R}\}$ et a fortiori de la gaussienne centrée réduite $\{\mathcal{N}(0; 1)$.

Définition 50 *On adapte les définitions pour les effectifs et fréquences théoriques :*

- $p_k = p_k(\hat{\theta}) = \mathbb{P}_{\hat{\theta}}(C_k)$ *la probabilité théorique (ou fréquence espérée) de la classe* C_k *sous* $H_{\hat{\theta}}$.
- $n_k = n_k(\hat{\theta}) = np_k(\hat{\theta})$ *l'effectif théorique (espéré) de la classe* C_k.
- *Le contraste est donc*

$$D_n = D_n(\mathbb{P}_n, \mathbb{P}_{\hat{\theta}}) = \sum_{k=1}^{K} \frac{\left(\hat{n}_k - n_k(\hat{\theta})\right)^2}{n_k} = n \sum_{k=1}^{K} \frac{\left(\hat{p}_k - p_k(\hat{\theta})\right)^2}{p_k(\hat{\theta})}$$

Théorème 17 *Sous* $H_0 : \exists \theta \in \Theta,\ \mathbb{P} = \mathbb{P}_\theta$, *la statistique de test* D_n *converge en loi vers un* χ^2 *à* $(K-1-r)$ *degrés de liberté, où* r *donne le nombre de paramètres laissés libres sous* H_0, $r = dim(\Theta)$. *La fonction*

$$\Phi = \mathbb{1}_{D_n > k_\alpha}$$

est un test de niveau asymptotique $1-\alpha$ *pour* $k_\alpha = F^{-1}_{\chi^2_{K-1-r}}(1-\alpha)$. *La région critique est donc la suivante :*

$$W = \left\{ D_n > F^{-1}_{\chi^2_{K-1-r}}(1-\alpha) \right\}$$

Ce théorème s'applique également dans le cas de l'adéquation à une unique loi, avec $r = 0$.

Remarque 12 On ne peut pas tester l'adéquation à la famille des lois de Bernoulli $Ber(p),\ p \in [0;1]$. En effet, on a deux classes pour ces variables, et on obtiendrait un quantile de $\chi^2_{2-1-1} = \chi^2_0$ ce qui n'existe pas.

C'est simplement parce qu'il existe toujours une loi de Bernoulli qui a une adéquation parfaite avec n'importe quel échantillon, il suffit de prendre p égal à la fréquence observée de la catégorie correspondant à $X = 1$. Le fait de suivre une loi de Bernoulli quelconque n'impose finalement aucune contrainte sur la distribution d'une variable à deux modalités.

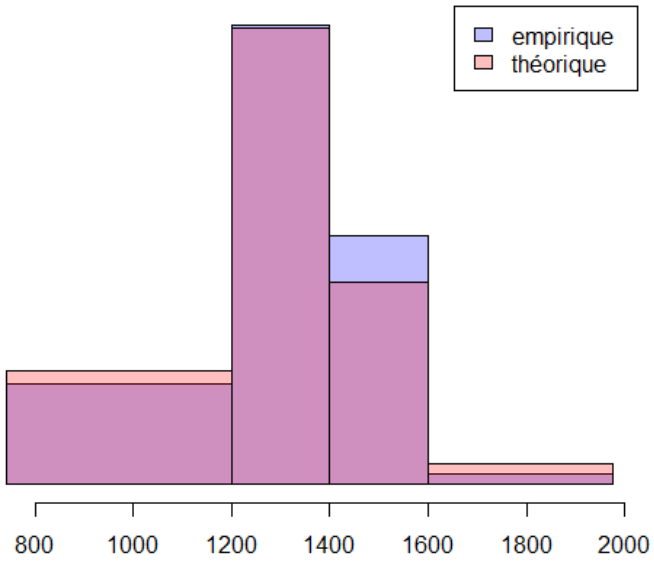
Exemple (Salaires log-normaux)
On considère donc les $\log(R)$ et on partitionne $\mathcal{X} = \mathbb{R}^+_*$. La famille des lois log-normales à $r = 2$ degrés de liberté, on obtiendra donc un quantile de χ^2_{K-1-2}, il faut donc se donner au moins $K = 4$ classes.

Il faut maintenant estimer ces deux paramètres pour choisir la meilleure loi dans la famille pour modéliser la distribution des données.

On estime par substitution l'espérance et la variance de $\log(R)$: $\hat{\mu} = 3,1$ et $\hat{\sigma^2} = 0,0037$. On calcule ensuite les effectifs espérés $n_k(\hat{\theta})$ en centrant par $\hat{\mu}$ et réduisant par $\sqrt{\hat{\sigma^2}}$ les bornes de classes pour $\log(R)$ et en calculant la probabilité qu'une gaussienne centrée réduite se trouve entre ces bornes. Par exemple :

$$n_1(\hat{\theta}) = n\mathbb{P}(Z < \frac{3,08 - 3,1}{\sqrt{0,0037}}) = nF_{\mathcal{N}(0;1)}(-0,63) = 100 \times 0,27 = 27.$$

Classe pour R	]0 ;1 200]	]1 200 ;1 400]	]1 400 ;1 600]	]1 600 ;+∞[
Classe pour $\log(R)$	]−∞ ;3,08]	]3,08 ;3,15]	]3,15 ;3,2]	]3,2 ;+∞[
$\hat{n}_k$	20	45	28	9
n_k	27	41	24	8



On obtient alors

$$D_n = \frac{(20-27)^2}{27} + \frac{(45-41)^2}{41} + \frac{(28-24)^2}{24} + \frac{(9-8)^2}{8} = 3$$

et finalement

$$\Phi = \mathbb{1}_{D_n > F_{\chi^2_{4-1-2}}} = \mathbb{1}_{3>3,84} = 0$$

On accepte donc l'adéquation à la famille des lois log-normales.

3 Test d'indépendance du χ^2

On suppose qu'on observe $((X_1, Y_1), \ldots, (X_1, Y_1))$ un échantillon i.i.d. de couples de variables aléatoires. Les observations de deux variables faites sur un même échantillon permettent-elles de conclure à l'indépendance de ces variables ?

Les exemples les plus naturels correspondent à des variables qualitatives (c'est-à-dire discrètes). Par exemple, X désigne une catégorie de population (salarié, employé, agriculteur, cadre supérieur, chômeur, ...) et Y un critère géographique (rubain/rural, ou la région). L'hypothèse à tester est l'indépendance entre la situation professionnelle X de l'individu et sa situation géographique Y. L'hypothèse affirme donc que le fait de connaître la situation géographique d'un individu ne donnerait aucune information sur la situation professionnelle, et réciproquement.

Dans le cas des variables continues, on se ramène à une situation discrète comme pour l'adéquation, en créant des classes C_k.

Exemple (Étudiant·es boursier·es)
On reporte dans le tableau ci-dessous la répartition du nombre (en milliers) de boursier·es sur critères sociaux en CPGE et à l'université, par échelon de bourse (en 2005-2006).

	Université	CPGE	Total
non boursier	748	57	805
échelon 0	30	1,6	31
échelon 1	56	3,2	59
échelon 2	31	1,5	32
échelon 3	31	1,4	32
échelon 4	31	1,3	33
échelon 5	127	4,2	131
Total	1 054	71	1 125

La méthode consiste à comparer les effectifs réels des croisements des modalités des deux variables qualitatives avec les effectifs théoriques qu'on devrait obtenir dans le cas d'indépendance de ces deux variables. C'est à nouveau le contraste du χ^2 qui permet de quantifier l'écart entre les effectifs réels et les effectifs théoriques.

On cherche à tester :

$$\begin{cases} H_0 : & \text{les deux variables sont indépendantes} \\ H_1 : & \text{les deux variables sont dépendantes} \end{cases}$$

Définition 51 *On suppose qu'on dispose de I classes $C_{i,.}$ pour X et J classes $C_{.,j}$ pour Y. On pose*

- *n est l'effectif total observé.*
- *$n_{i,j}$ est l'effectif observé de la classe $C_{i,j}$ des individus possédant à la fois la modalité i de la première variable et la modalité j de la deuxième variable.*
- *$p_{i,j} = n_{i,j}/n$ est la fréquence observée de $C_{i,j}$.*
- *$p^0_{i,j}$ est la probabilité théorique sous l'hypothèse d'indépendance d'obtenir une observation possédant la modalité i de la première variable et la modalité j de la deuxième variable, c'est-à-dire :*

$$p^0_{i,j} = \mathbb{P}_n(X \in C_{i,.}) \times \mathbb{P}_n(Y \in C_{.,j}) = \left(\sum_{j=1}^{I} n_{i,j}\right)/n \times \left(\sum_{i=1}^{I} n_{i,j}\right)/n = \frac{n_{i,.}}{n} \times \frac{n_{.,j}}{n}$$

- *$n^0_{i,j} = np^0_{i,j}$ est l'effectif théorique de la classe $C_{i,j}$.*

Soit D_n la statistique de test définie par :

$$D_n = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{i,j} - n^0_{i,j}\right)^2}{n^0_{i,j}}$$

où :

Théorème 18 *Sous H_0, la statistique de test D_n converge en loi vers une loi de χ^2 à $(I-1)(J-1)$ degrés de liberté. On obtient donc un test de niveau asymptotique $1-\alpha$ avec la région critique suivante :*

$$W = \left\{ D > F^{-1}_{\chi^2_{(I-1)(J-1)}}(1-\alpha) \right\}$$

Exemple (Étudiant·es boursier·es)
On teste ici l'indépendance entre le niveau de bourse et le type d'établissement. Ici l'espace $\mathcal{X} \times \mathcal{Y}$ est déjà complètement partitionné en 2×7 classes.

On calcule donc les effectifs théoriques sous l'hypothèse d'indépendance entre le type d'établissement et l'échelon de bourse. Pour les boursiers non boursier·es de l'Université, on devrait donc avoir un effectif égal au nombre total d'étudiant·es multiplié par le taux de non boursier·es et par la part de l'Université dans le total des inscriptions (attention tout est en milliers) :

$$1\ 125 \times \frac{805}{1\ 125} \times \frac{1\ 054}{1\ 125} = 755$$

On répète l'opération pour toutes les cases, puis on calcule pour chaque case la différence au carré divisé par l'effectif théorique. Pour la case en haut à gauche : $\frac{(748-755)^2}{755} = 0,058$. On additionne les 14 résultats et on obtient un statistique du χ^2 égale 3,692 milliers soit 3 692, ce qui est énorme. Il faut comparer au quantile à 95% d'une loi du χ^2 à $(2-1)\times(7-1) = 6$ degrés de liberté : 12,59. On est donc très très au-dessus du seuil de significativité : les répartitions sont distinctes.

On peut remarquer que les deux cases qui contribuent le plus à la statistique du χ^2 et donc à la différence entre les répartitions sont le nombre de non boursiers en prépa (trop élevé) et celui des boursiers échelon 5 en prépa (trop faible). A l'inverse, sur les échelons 0 et 1 les classes préparatoires recrutent en proportion similaires à l'Université.

Exemple (Le taux de boursier est-il le bon indicateur d'ouverture sociale ?)
Le taux de boursier dans le supérieur et en particulier dans les grandes écoles et leurs classes préparatoires cristallise le débat sur la reproduction des classes sociales. C'est en effet le principal indicateur retenu dans le débat, plutôt que la profession et catégorie sociale des parents (15% d'enfants d'ouvriers ou d'employés en CPGE, alors qu'ils représentent 30% de la population active, selon les chiffres du MESR, 2012). Prenons par exemple le rapport d'information au Sénat n° 441 du 12 septembre 2007, qui regrette la fermeture sociale marquée en CPGE :

> « Cette "surreprésentation" des étudiants d'origine sociale favorisée ne se retrouve pas de façon si prononcée dans les autres filières de l'enseignement supérieur [...]
> A l'inverse, la part d'étudiants boursiers dans les classes préparatoires est globalement plus faible que dans l'ensemble de l'enseignement supérieur. »

Opposer ainsi origine sociale favorisée (c'est-à-dire selon l'INSEE avoir un père chef d'entreprise de plus de 10 salariés, cadre, de profession libérale ou professeur) et boursier n'est qu'en partie légitime : on peut par exemple prétendre à une bourse sur critères sociaux avec un père professeur des écoles et d'une mère sans emploi.

Mais surtout, l'évolution du taux de boursiers, sans distinction d'échelon, peut-être trompeuse. On l'a vu, dans les échelons les plus faibles, la disparité entre CPGEs et universités est négligeable. Or le taux de boursier dépend de la population étudiante, mais aussi de la politique sociale de l'Etat. Si l'on élargi les conditions d'attribution des bourses de bas échelon (ou que l'on créé de nouveaux échelons en bas de

l'échelle), on transforme des étudiant·es non boursier·ères en boursier·ères, sans rien modifier de la composition sociale de cette population. Les conditions d'études (et les dépenses) sont-elles mêmes très marginalement modifiées puisque l'échelon zéro n'ouvre droit à aucun versement mais uniquement à des exonérations de frais d'inscription (eux-mêmes nuls en classes préparatoires) et de sécurité sociales étudiantes.

Chapitre 5

Statistique mathématique

A Cadre formel

Dans ce chapitre, nous définissons un cadre formel pour les statistiques, qui permet d'en faire une branche des mathématiques à part entière, reliée à l'analyse, à la théorie de la mesure et aux probabilités. Comme pour la substitution, nous intéressons à l'information qui est contenue dans un échantillon et nous nous demandons comment choisir au mieux une loi parmi une famille de distributions indexée par un paramètre $\theta \in \Theta$ (Θ étant une partie de $\mathbb{R}$ ou, plus rarement, de $\mathbb{R}^2$).

On va donc chercher une manière d'estimer la vraie valeur θ_0, inconnue, de $\theta \in \Theta$ à partir d'un échantillon $(X_1, \ldots, X_n)$ de loi marginale $\mathbb{P}_{\theta_0}$, dont on observe une réalisation *i.e.* nos observations. On va voir qu'il n'est même pas toujours possible d'estimer θ_0 : dans certaines situations, quelques soient les réalisations de l'échantillon observées, on ne peut pas dire si une valeur de θ est plus « crédible » qu'une autre, ou plutôt plus vraisemblable :

1 Vraisemblance

Définition 52 (Vraisemblance) *Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d., on appelle vraisemblance du paramètre θ l'application qui associe à θ la densité de probabilité jointe de l'échantillon . Lorsque la loi est discrète, la probabilité atomique $\Pr_\theta(X = x)$ fait office de densité.*

$$L_{X_1,\ldots,X_n} : \begin{cases} \Theta \to \mathbb{R}^+ \\ \theta \mapsto L_{X_1,\ldots,X_n}(\theta) = f_\theta\left(X_1, \ldots, X_n\right) \end{cases}$$

Lorsque l'échantillon étant i.i.d., on a :

$$L_{X_1,\ldots,X_n}(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$$

C'est la densité jointe des observations, du point de vue de la dépendance en θ. C'est donc une variable ***aléatoire*** *parce qu'elle dépend des X_i, mais comme fonction c'est sa dépendance en θ qui est principale. On note souvent L_n pour simplifier l'écriture*

Exemple Dans le cas iid gaussien de variance 1, la vraisemblance s'écrit

$$L_n(\mu) = \prod_{i=1}^{n} f_\mu(X_i) = (\sqrt{2\Pi})^{-n} \prod_{i=1}^{n} e^{-\frac{1}{2}(X_i-\mu)^2} = (\sqrt{2\Pi})^{-n} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

Dans le cas d'un échantillon iid Bernoulli, on a

$$L_n(p) = \prod_{i=1}^{n} (p\mathbb{1}_{X_i=1} + (1-p)\mathbb{1}_{X_i=0})$$

Il est beaucoup plus commode d'écrire la probabilité atomique (« densité ») sous une forme factorisée du fait du produit correspond au caractère iid de l'échantillon :

$$L_n(p) = \prod_{i=1}^{n} \left(p^{X_i}(1-p)^{1-X_i}\right) = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

En effet, $p^{X_i}(1-p)^{1-X_i} = p$ si $X_i = 1$ et $1-p$ sinon, comme $p\mathbb{1}_{X_i=1} + (1-p)\mathbb{1}_{X_i=0}$.

Définition 53 (Modèle statistique) *On appelle modèle statistique, le quintuplet*

$$\left((X_1, \ldots, X_n) \in \mathcal{X}^n, L_n(\theta), \theta \in \Theta\right).$$

Définition 54 (Identifiabilité) *Un modèle est dit identifiable si la fonction vraisemblance $L_{X_1,\ldots,X_n}$ est injective, c'est-à-dire si*

$$\forall \theta_1 \neq \theta_2,\ \exists (x_1, \ldots, x_n) \in \mathcal{X}^n,\ L_{x_1,\ldots,x_n}(\theta_1) \neq L_{x_1,\ldots,x_n}(\theta_2)$$

$$\Leftrightarrow \forall \theta_1 \neq \theta_2,\ \left((X_1, \ldots, X_n) \mapsto \prod_{i=1}^{n} f_{\theta_1}(X_i)\right) \neq \left((X_1, \ldots, X_n) \mapsto \prod_{i=1}^{n} f_{\theta_2}(X_i)\right)$$

Cette notation un peu lourde souligne que c'est en tant que fonctions des observations que les vraisemblances doivent être différentes. Il suffit donc de trouver une valeur de $(X_1, \ldots, X_n)$ pour laquelle $L_n(\theta_1) \neq L_n(\theta_2)$.

Exemple (Identifiabilité des lois de Bernoulli)
Le modèle de Bernoulli est par exemple identifiable dès qu'on a $n = 1$, il suffit de choisir $X_1 = 1$ et on a $L_1(p) = p \neq p' = L_1(p')$ dès que p et p' sont différents.

Exemple (Identifiabilité des lois exponentielles)
Soit $(X_1, ..., X_n)$ un échantillon i.i.d. suivant une loi exponentielle de paramètre λ. Le modèle statistique est donc $\left((X_1, ..., X_n) \in \mathbb{R}_+, \prod_{i=1}^{n} \lambda e^{-\lambda X_i}, \lambda \in \mathbb{R}_+^*\right)$. La fonction de vraisemblance est $L_{X_1,\ldots,X_n}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$.

Pour montrer que le modèle est identifiable, il faut choisir des X_i tels que $L_{X_1,..,X_n}(\lambda) \neq L_{X_1,..,X_n}(\lambda')$. Soit $(\lambda, \lambda') \in \left(\mathbb{R}_+^*\right)^2$, avec $\lambda \neq \lambda'$. Si on prend $(X_i)_{i\in[i;n]}$ tels que $\forall i \in [1;n], X_i = 0$, alors

$$\begin{aligned} L_{X_1,..,X_n}(\lambda) \neq L_{X_1,..,X_n}(\lambda') &\Leftrightarrow \lambda^n \exp^{-\lambda \sum_{i=1}^{n} X_i} \neq \lambda'^n \exp^{-\lambda' \sum_{i=1}^{n} X_i} \\ &\Leftrightarrow \lambda^n \neq \lambda'^n \\ &\Leftrightarrow \lambda \neq \lambda' \text{ car } \lambda, \lambda' \in \mathbb{R}_+^* \end{aligned}$$

Or, cette dernière inégalité est vraie, donc toutes les équivalences le sont, donc $L_{X_1,..,X_n}(\lambda) \neq L_{X_1,..,X_n}(\lambda')$: le modèle est identifiable.

Remarque : on ne démontre pas que les deux fonctions sont distinctes pour toutes valeurs de l'échantillon : il suffit de trouver une valeur de l'échantillon pour laquelle elles diffèrent pour démontrer que le modèle est identifiable.

Exemple (Contre-exemple)
Pour bien comprendre l'importance de l'identifiabilité, il vaut mieux prendre un contre-exemple. L'usage des boursiers comme intermédiaire pour analyser l'ouverture sociale de l'enseignement supérieur peut en faire office.

Notons p la proportion d'étudiant·es d'origine populaire (enfants d'ouvrier·ères ou d'employé·es) et parmi ceux-ci, t le taux de boursier·ères. Notons enfin r la proportion d'étudiant·es d'origine non populaire boursier·ères Le taux de boursier·ères dans l'enseignement supérieur est donné par $pt + r$. Soit enfin un échantillon iid qui indique si l'étudiant·e est boursier·ère, on a pour vraisemblance

$$L_n(p, t, r) = \prod_i (pt + r)^{\sum X_i} (1 - pt - r)^{n - \sum X_i}$$

qui n'est clairement pas identifiable pour p (ni pour t ou r) : on a par exemple

$$L_n(0,15; 1; 0, 1) = L_n(0, 25; 1; 0)$$

On ne peut donc pas estimer si l'accès des classes populaires au supérieur s'améliore ou non à partir de la seule observation du taux de boursier·ères ;

Exemple (Contre-exemple II)
On observe $(X_1, ..., X_n)$ des variables i.i.d. qui sont les résultats du candidat i aux écrits d'un concours, avec une phase d'admissibilité et d'admission. On veut estimer la barre d'admissibilité.

On se donne le modèle suivant, avec $p \in]0; 1[$,

- i a une probabilité $1 - p$ d'être admissible ;
- si i est admissible, alors $X_i \sim \mathcal{U}(20p, p)$;
- si i n'est pas admissible, alors $X_i \sim \mathcal{U}(0, 20p)$

Le modèle statistique correspondant est $((X_1, ..., X_n) \in [0; 20]^n, L_{X_1,...,X_n}(p), p \in]0; 1[)$. Les variables sont indépendantes et identiquement distribuées, donc

$$\begin{aligned} L_{X_1,...,X_n}(p) &= \prod_{i=1}^{n} f_p(X_i) \\ &= \prod_{i=1}^{n} p^{\text{non admis}} \left(\frac{1}{20p}\mathbb{1}_{X\in[0;20p]}\right) (1-p)^{\text{admis}} \left(\frac{1}{20-20p}\mathbb{1}_{X\in[20p;20]}\right) \end{aligned}$$

Autrement dit, on a

$$f_p(X_I) = \begin{cases} p\dfrac{1}{20p}\mathbb{1}_{X\in[0;20p]}(X_i) = \dfrac{1}{20}\mathbb{1}_{X\in[0;20p]}(X_i) & \text{si non admis} \\ (1-p)\left(\dfrac{1}{20-20p}\mathbb{1}_{X\in[20p;20]}\right) = \dfrac{1}{20}\mathbb{1}_{X\in[0;20p]}(X_i) & \text{si admis} \end{cases}$$

En d'autres termes, $f_p(X_i) = \dfrac{1}{20} f_{\mathcal{U}([0;20])}(X_i)$. Donc $L_{(X_1,...,X_n)}(p) = \dfrac{1}{20^n}$. Si on prend deux valeurs différentes, on a par exemple $L_{(X_1,...,X_n)}\left(\dfrac{1}{4}\right) = L_{(X_1,...,X_n)}\left(\dfrac{1}{2}\right)$, mais $\dfrac{1}{4} \neq \dfrac{1}{2}$. Le modèle n'est donc pas identifiable.

2 Statistique exhaustive

Une fois que l'on est assuré de disposer d'assez d'informations pour identifier θ, on cherche à réduire l'information utilisée à celle strictement nécessaire.

Définition 55 (Exhaustivité) *Une statistique T est exhaustive si et seulement si à T fixé, la vraisemblance ne dépend pas de θ. Autrement dit, $\exists T : \mathcal{X}^n \to \mathcal{T}, L_{X_1,...,X_n|T(X_1,...,X_n)}(\theta)$ ne dépend pas de θ.*

Proposition 19 (Théorème de factorisation)
$T : \mathcal{X}^n \to \mathcal{T}$ est exhaustive si $\exists g : \mathcal{X}^n \to \mathbb{R}_+$ et $\exists h : \mathcal{T} \times \Theta \to \mathbb{R}_+$ telles que

$$L_{X_1,...,X_n}(\theta) = g(X_1, ..., X_n)h(T(X_1, ..., X_n), \theta)$$

La démonstration demande une formalisation plus simple dans le cas discret :

$$\begin{aligned} L_{X_1,...,X_n|T(X_1,...,X_n)=t}(\theta) &= \frac{\mathbb{P}_{\theta,\ T(X_1,...,X_n)=t}(X_1, ..., X_n)}{\sum_{(X_1,...,X_n),\ T(X_1,...,X_n)=t} \mathbb{P}_\theta(X_1, ..., X_n)} \\ &= \frac{g(X_1, ..., X_n)h(t, \theta)}{\sum_{(X_1,...,X_n),\ T(X_1,...,X_n)=t} g(X_1, ..., X_n)h(t, \theta)} \end{aligned}$$

Au dénominateur, on peut sortir le $h(t, \theta)$ de la somme, et donc le simplifier avec celui du numérateur. On obtient donc

$$L_{X_1,...,X_n|T(X_1,...,X_n)=t}(\theta) = \frac{g(X_1, ..., X_n)}{\sum_{(X_1,...,X_n)\in\mathcal{X},\ T(X_1,...,X_n)=t} g(X_1, ..., X_n)}.$$

La loi jointe de l'échantillon conditionnellement à T ne dépend plus de θ. On peut remarque que le dénominateur ne fait que normaliser celle la loi jointe conditionnelle de façon à ce que la somme des probabilités fasse bien 1.

Dans le cas continu, on obtient une écriture équivalente avec une intégrale multiple à la place de la somme.

Exemple (Statistique exhaustive pour la famille des lois uniformes) Soit $(X_1, ..., X_n)$ i.i.d. de loi $\mathcal{U}([0;\theta])$. On pose $L_{X_1,...X_n}(\theta) = L_n(\theta)$.

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_{\mathcal{U}([0;\theta])}(X_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0;\theta]}(X_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{0\leq X_i\leq\theta}(X_i) \\ &= \left(\frac{1}{\theta}\right)^n \mathbb{1}_{\forall i\in]1;n[,0\leq X_i\leq\theta}(X_i) \\ L_n &= \left(\frac{1}{\theta}\right)^n \mathbb{1}_{\min X_i\geq 0} \mathbb{1}_{\max X_i\leq\theta} \end{aligned}$$

Notons que le produit de variables indicatrices est égal à l'indicatrice de l'intersection. On a donc factorisé notre fonction de vraisemblance, avec

$$g(X_1, ..., X_n) = \mathbb{1}_{\min X_i\geq 0}, h(T(X_1, ..., X_n), \theta) = \mathbb{1}_{\max X_i\leq\theta}, T(X_1, ..., X_n) = \max X_i$$

B Ordres sur les estimateurs

1 Comparaison d'estimateurs : biais et variance

Lorsqu'on dispose de plusieurs estimateurs pour un même paramètre θ, il est utile de pouvoir les comparer afin de choisir le meilleur de ces estimateurs. Pour cela, il nous faut pour cela trouver un critère permettant d'arbitrer entre le biais et la variance : un estimateur légèrement biaisé, mais dont la variance est faible pourrait s'avérer meilleur qu'un estimateur sans biais mais moins précis. Un critère classique de comparaison est l'erreur quadratique moyenne, appelé aussi risque quadratique espéré, d'un estimateur :

Définition 56 (Erreur quadratique moyenne) *Soit $\widehat{\theta}$ un estimateur de θ. La dis-*

tance de $\widehat{\theta}$ à θ peut être mesuré par l'erreur quadratique moyenne :

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\theta}-\theta\right)^2\right] &= \mathbb{E}\left[\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]+\mathbb{E}\left[\widehat{\theta}\right]-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]\right)^2+\left(\mathbb{E}\left[\widehat{\theta}\right]-\theta\right)^2+2\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]\right)\left(\mathbb{E}\left[\widehat{\theta}\right]-\theta\right)\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta}-\mathbb{E}\left[\widehat{\theta}\right]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}\right]-\theta\right)^2\right] \\
&= \mathrm{Var}\left(\widehat{\theta}\right)+\left(Biais\left(\widehat{\theta}\right)\right)^2
\end{aligned}$$

Proposition 20 Pour un estimateur sans biais, l'erreur quadratique moyenne est égale à la variance.

L'objectif est de minimiser le biais et la variance pour minimiser l'erreur quadratique moyenne. En faisant intervenir le biais et la variance, l'erreur quadratique moyenne permet donc de trancher dans une situation où il existe un estimateur sans biais et un autre biaisé mais de variance plus petite.

Exemple Comparons les deux estimateurs de σ^2

$$S_n^2=\frac{1}{n-1}\sum(X_i-\overline{X})^2 \text{ et } \tilde{S}_n^2=\frac{n-1}{n}S_n^2$$

- $\mathbb{E}[S_n^2]=\sigma^2 \quad \Rightarrow \quad Biais=0.$
- $\mathrm{Var}\left(S_n^2\right)=\dfrac{2\sigma^4}{n-1}.$
- $\mathbb{E}[\tilde{S}_n^2]=\dfrac{n-1}{n}\mathbb{E}[S_n^2]=\dfrac{(n-1)}{n}\sigma^2 \quad \Rightarrow \quad Biais=\dfrac{-\sigma^2}{n}.$
- $\mathrm{Var}\left(\tilde{S}_n^2\right)=\dfrac{(n-1)^2}{n^2}\mathrm{Var}\left(S_n^2\right)=\dfrac{2(n-1)}{n^2}\sigma^4.$

L'estimateur S_n^2 est sans biais mais a une plus forte variance que l'estimateur $\tilde{S}_n^2$. La comparaison des erreurs quadratiques moyennes donne :

$$\mathrm{EQM}(\tilde{S}_n^2)-\mathrm{EQM}(S_n^2)=\frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1}+\left(\frac{-\sigma^2}{n}\right)^2-\frac{2\sigma^4}{n-1}=\frac{3n-1}{n^2(n-1)}\sigma^4<0$$

L'estimateur biaisé $\tilde{S}_n^2$ est donc plus précis en termes d'erreur quadratique moyenne.

Remarque 13 Malheureusement, dans le cas général, on ne sait pas résoudre le problème de minimisation de l'EQM (il dépend de manière complexe du paramètre). Un « second best » plus simple consiste à rechercher un estimateur sans biais de variance minimale.

2 Estimateur efficace : l'inégalité de Fréchet, Darmois, Cramer, Rao

L'erreur quadratique moyenne mesure l'efficacité d'un estimateur, c'est à dire sa capacité à utiliser toute l'information disponible dans les données. On va donc mesurer cette information pour les estimateurs sans biais :

Définition 57 (Score) *Soit un modèle statistique pour lequel la densité $f_\theta(x)$ est dérivable en θ et strictement positive sur $\mathcal{X} \times \Theta$. Pour une donnée, on appelle score la quantité :*

$$S(X_1, \theta) = \frac{\partial \log L_{X_1}(\theta)}{\partial \theta} = \frac{\partial \log f_\theta}{\partial \theta} = \frac{1}{f_\theta} \frac{\partial f}{\partial \theta}.$$

Pour un échantillon i.i.d., la vraisemblance est donnée par le produit des densités et donc

$$\begin{aligned} S(X_1, \ldots, X_n, \theta) &= \frac{\partial \log L_n(\theta)}{\partial \theta} = \frac{\partial \log \prod_i f_\theta(X_i)}{\partial \theta} \\ &= \sum_{i=1}^{n} \frac{\partial \log f_\theta(X_i)}{\partial \theta} = \sum_{i=1}^{n} S(X_i, \theta) \end{aligned}$$

Remarque 14 Dans la suite nous remplacerons souvent les observations $(X_1, \ldots, X_n)$ par un indice n afin d'allèger les notations : $S_n(\theta)$.

Le score mesure la sensibilité de la vraisemblance à θ, c'est-à-dire la facilité avec laquelle elle parvient à discriminer entre les θ possibles, au vu d'un échantillon $X = (X_1, \ldots, X_n)$. Lorsqu'il est de grand en valeur absolue, cela signifie que des petites modifications de θ font fortement varier la log-vraisemblance des observations ; au contraire un score presque nul signifie que les variations de θ n'affectent pratiquement pas la log vraisemblance $\log L_n(\theta)$.

Définition 58 (Information de Fisher) *Sous les conditions d'existence du score, l'information de Fisher est définie comme l'espérance du carré du score. Pour une donnée :*

$$I_1(\theta) = \mathbb{E}[S(X_1, \theta)^2] = \mathbb{E}\left[\left(\frac{\partial \log f_\theta}{\partial \theta}(X_1)\right)^2\right]$$

et pour un échantillon :

$$I_n(\theta) = \mathbb{E}[S_n(\theta)^2] = \mathbb{E}\left[\left(\frac{\partial \log L_n(\theta)}{\partial \theta}\right)^2\right]$$

Remarque 15 L'information de Fisher est toujours **positive** !

Exemple (Information de Fisher pour une loi $\mathcal{U}([0; \theta])$)

Pour une donnée, la vraisemblance du modèle sur $\mathcal{X} \times \theta = [0; \theta] \times \mathbb{R}_+^*$ est : $L_{X_1}(\theta) = f_\theta(X) = \dfrac{1}{\theta}$ d'où :

$$L_n(\theta) = \prod_i \frac{1}{\theta} \mathbb{1}_{X_i \leq \theta} = \frac{1}{\theta^n} \mathbb{1}_{\max_i\{X_i\} \leq \theta}$$

le score s'écrit alors pour les X_i dans $\mathcal{X}$, c'est-à-dire $\max\limits_i\{X_i\} \leq \theta$:

$$S_n(\theta) = \frac{\partial \ln L_n(\theta)}{\partial \theta} = -\frac{n}{\theta}.$$

On obtient donc :

$$I_n(\theta) = \mathbb{E}\left[S_n(\theta)^2\right] = \mathbb{E}\left[\left(-\frac{n}{\theta}\right)^2\right] = \frac{n^2}{\theta^2}$$

Exemple (Information de Fisher pour la famille des lois exponentielles)

Soit $(X_1, ..., X_n)$ i.i.d. de loi exponentielle de paramètre λ, avec $\mathcal{X} = \mathbb{R}_+$ et $\lambda \in \Lambda = \mathbb{R}_+^*$. On a donc $f_\lambda(x) = e^{-\lambda x}$.

Calculons le score :

$$S(X, \lambda) = \frac{\log f_\lambda(x)}{\partial\lambda} = \frac{\partial\left(\log\lambda - \lambda x\right)}{\partial\lambda} = \frac{1}{\lambda} - x$$

Et donc

$$S_n(X, \lambda) = \sum_{i=1}^{n} S(X_i, \lambda) = \sum_{i=1}^{n}\left(\frac{1}{\lambda} - x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = n\left(\frac{1}{\lambda} - \bar{X}\right)$$

On peut maintenant calculer l'information de Fisher :

$$\begin{aligned} I_X(\lambda) &= \mathbb{E}[S^2(X, \lambda)] \\ &= \mathbb{E}\left[\left(\frac{1}{\lambda} - X\right)^2\right] \\ &= \int_0^{+\infty}\left(\frac{1}{\lambda} - x\right)^2 \lambda e^{-\lambda x} dx \text{ par théorème du transfert} \end{aligned}$$

Il faut procéder par intégration par parties pour continuer, en intégrant $\lambda e^{-\lambda x}$ et en dérivant $\left(\frac{1}{\lambda} - x\right)^2$.

$$\begin{aligned} I_X(\lambda) &= \left[\left(\frac{1}{\lambda} - x\right)^2\left(-e^{-\lambda x}\right)\right]_0^{+\infty} - \int_0^{+\infty}\left(-2\left(\frac{1}{\lambda} - x\right)\right)\left(-e^{-\lambda x}\right) dx \\ &= \frac{1}{\lambda^2} - 2\int_0^{+\infty}\left(\frac{1}{\lambda} - x\right) e^{-\lambda x} dx \end{aligned}$$

À nouveau, on procède par intégration par parties, en intégrant $e^{-\lambda x}$ et en dérivant $\frac{1}{\lambda} - x$.

$$\begin{aligned} I_X(\lambda) &= -2\left(\left[\left(\frac{1}{\lambda} - x\right)\frac{e^{-\lambda x}}{-\lambda}\right]_0^{+\infty} - \int_0^{+\infty}(-1)\frac{e^{-\lambda x}}{-\lambda} dx\right) + \frac{1}{\lambda^2} \\ &= -2\int_0^{+\infty}(-1)\frac{e^{-\lambda x}}{-\lambda} dx - 2\left(\frac{1}{\lambda^2}\right) + \frac{1}{\lambda^2} \end{aligned}$$

Reste alors à intégrer la dernière intégrale, dont on connait la primitive :

$$I_X(\lambda) = -2\left[\frac{e^{-\lambda x}}{\lambda^2}\right]_0^{+\infty} - \frac{1}{\lambda^2} = -2\left(-\frac{1}{\lambda^2}\right) - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Ou alors, on sait que si $X \sim \mathcal{E}(\lambda), \mathbb{E}(X) = \dfrac{1}{\lambda}$, et on reconnaît alors la variance dans l'expression initiale :

$$\begin{aligned} I_X(\lambda) &= \mathbb{E}[S^2(X,\lambda)] \\ &= \mathbb{E}\left[\left(\frac{1}{\lambda} - X\right)^2\right] \\ &= \mathbb{E}(X - \mathbb{E}(X))^2] \\ &= \mathrm{Var}(X) \\ I_X(\lambda) &= \frac{1}{\lambda^2} \end{aligned}$$

Définition 59 (Modèle régulier)
Un modèle est dit régulier si et seulement si :

(H1) L'ensemble $\mathcal{X}$ des valeurs possible pour X ne dépend pas de la valeur de θ.

(H2) La fonction de vraisemblance $L_n(\theta)$ est de classe C^2 : $\dfrac{\partial L_n(\theta)}{\partial \theta}$ et $\dfrac{\partial^2 L_n(\theta)}{\partial \theta^2}$ existent $\forall (X_1, \ldots, X_n) \in \mathcal{X}^n$ et $\forall \theta \in \Theta$.

(H3) Les fonctions $\dfrac{\partial \ln L_n}{\partial \theta}$ et $\dfrac{\partial^2 \ln L_n}{\partial \theta^2}$ sont d'espérance finie $\forall \theta \in \Theta$, et que les dérivées de $L_n(\theta)$ peuvent s'effectuer sous le signe somme. Ainsi $\forall \theta \in \Theta$, on a :

$$\begin{cases} \dfrac{\partial}{\partial \theta} \displaystyle\int_{\mathcal{X}^n} L_n(\theta)\, dX = \int_{\mathcal{X}^n} \frac{\partial}{\partial \theta} L_n(\theta)\, dX\,, \\ \dfrac{\partial^2}{\partial \theta^2} \displaystyle\int_{\mathcal{X}^n} L_n(\theta)\, dX = \int_{\mathcal{X}^n} \frac{\partial^2}{\partial \theta^2} L_n(\theta)\, dX \end{cases}$$

(H4) Le score est de carré intégrable et donc l'information de Fisher existe.

En pratique dans le cadre de ce cours, on ne vérifiera que les deux premières hypothèses, la troisième étant admise (la quatrième se vérifie d'elle-même au cours du calcul de I).

Théorème 19 (Seconde forme de l'information de Fisher) *Si le modèle est régulier alors*

- *le score est une variable aléatoire centrée, i.e. $\mathbb{E}[S_n(\theta)] = 0$; l'information de Fisher est alors sa variance ;*
- *l'information de Fisher pour une donnée est égale à :*

$$I_1(\theta) = -\mathbb{E}\left[\frac{\partial S(X_1,\theta)}{\partial \theta}\right] = -\mathbb{E}\left[\frac{\partial^2 \ln L(X_1,\theta)}{\partial \theta^2}\right]$$

Attention, il y a un signe $-$ qui permet de rendre I_1 positive !

- *l'information de Fisher pour un échantillon i.i.d. vaut :*

$$I_n(\theta) = nI_1(\theta)$$

L'exemple précédent de la loi uniforme n'est pas un modèle régulier, on ne peut donc pas calculer l'information de Fisher sous sa deuxième forme, cela conduirait à un résultat faux :

$$-\mathbb{E}\left[\frac{\partial^2 \ln L_n(\theta)}{\partial\theta^2}\right] = -\mathbb{E}\left[\frac{\partial S_n(\theta)}{\partial\theta}\right] = -\mathbb{E}\left[\frac{n}{\theta^2}\right] = -\frac{n}{\theta^2} \neq \frac{n^2}{\theta^2}$$

En effet, l'ensemble $\mathcal{X}$ tel que $f_\theta(x) > 0$ est $[0;\theta]$ et dépend donc de θ. Ceci contredit l'hypothèse $H1$. L'information de Fisher est définie mais on ne peut pas utiliser la seconde forme ni $I_n = n \times I_1$.

Exemple Calculez l'information de Fisher pour une loi de Poisson :

$$\Pr_\lambda X = k = e^{-\lambda}\frac{\lambda^k}{k!} \text{ pour } k \in \mathbb{N},\ \lambda \in \mathbb{R}_+^*$$

Montrons que la loi de Poisson conduit à un modèle régulier. Le modèle s'écrit :

$$\left((X_1,\ldots,X_n), \mathcal{X}^n = \mathbb{N}^n, \mathcal{P}_\lambda^{\otimes n}, \lambda \in \mathbb{R}_*^+\right)$$

On a bien les deux hypothèses nécessaires : $\mathcal{X}$ ne dépend pas de λ et $\lambda \mapsto f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ est C^∞ donc dérivable deux fois.

Comme le modèle est régulier, on peut calculer l'information de Fisher sous sa deuxième forme et I_n égale à n fois l'information de Fisher pour une observation (I_1). On se contente donc de calculer l'information de Fisher dans le cas $n = 1$.

$$\begin{aligned} f_\lambda(k) &= e^{-\lambda}\frac{\lambda^k}{k!} \\ \log f_\lambda(k) &= -\lambda + k\ln\lambda - \ln(k!) \\ S(k,\lambda) = \frac{\partial \log f_\lambda(k}{\partial\lambda} &= -1 + \frac{k}{\lambda} \\ \frac{\partial^2 \log f_\lambda(k}{\partial\lambda^2} &= -\frac{k}{\lambda^2} \end{aligned}$$

$$\text{On a } I_1(\lambda) = -\mathbb{E}\left[\frac{\partial^2 \ln L(X_1,\lambda)}{\partial\lambda^2}\right] = \mathbb{E}\left[\frac{k}{\lambda^2}\right]$$

Or $\mathbb{E}[X] = \lambda$, donc $I_1(\lambda) = \frac{1}{\lambda}$ et $I_n(\lambda) = \frac{n}{\lambda}$.

On peut aussi utiliser la première forme de l'information de Fisher $I_n = \mathbb{E}\left[S_n^2(\lambda)\right]$. Comme l'échantillon est i.i.d. :

$$S_n(\lambda) = \sum_i S(X_i,\lambda) = \sum_i -1 + \frac{X_i}{\lambda} = n\frac{\overline{X} - \lambda}{\lambda}.$$

On obtient

$$I_n = \mathbb{E}\left[\left(n\frac{\overline{X}-\lambda}{\lambda}\right)^2\right] = \frac{n^2}{\lambda^2}\mathrm{Var}(\overline{X}) = \frac{n^2}{\lambda^2}\frac{\lambda}{n} = \frac{n}{\lambda}$$

parce que la variance d'une loi de poisson est égale à son espérance. On a donc bien égalité des deux formes de l'information de Fisher.

Théorème 20 (Borne de Fréchet-Darmois-Cramer-Rao ou FDCR)
Soit un modèle régulier et $\widehat{\theta}$ un estimateur sans biais de θ, alors

$$Var(\widehat{\theta}) \geq I_n(\theta)^{-1}.$$

Démonstration *On fait la démonstration dans le cas $n = 1$ et on note $\widehat{\theta} = T(X)$. $\widehat{\theta}$ est un estimateur sans biais :*

$$\mathbb{E}[T(X)] = \int_{\mathcal{X}} T(X) f_\theta(X) dX = \theta$$

On a donc :

$$\frac{\partial \mathbb{E}[T(X)]}{\partial \theta} = \int_{\mathcal{X}} T(X) \frac{\partial f_\theta(X)}{\partial \theta} dX = 1$$

Comme $S_X(\theta) = \dfrac{1}{f_\theta(X)}\dfrac{\partial f_\theta(X)}{\partial \theta}$, on a :

$$\int_{\mathcal{X}} T(X) S_X(\theta) f_\theta(X) dX = 1 \Leftrightarrow \mathbb{E}[T(X) SS_X(\theta)] = 1$$

En outre $\mathbb{E}[S_X(\theta)] = 0$, donc $\mathbb{E}[T(X) S_X(\theta)] = \mathrm{Cov}[T(X) S_X(\theta)]$
En effet, nous rappelons que :

$$\begin{aligned}\mathrm{Cov}[T(X) S_X(\theta)] &= \mathbb{E}[(T(X) - \mathbb{E}[T(X)])(S_X(\theta) - \mathbb{E}[S_X(\theta)])] \\ &= \mathbb{E}[T(X) S_X(\theta)] - \mathbb{E}[T(X)]\mathbb{E}[S_X(\theta)].\end{aligned}$$

Il s'en suit :

$$\begin{aligned}1 = \frac{\partial \mathbb{E}[T(X)]}{\partial \theta} &= \mathbb{E}[T(X) S_X(\theta)] \\ &= \mathrm{Cov}[T(X) S_X(\theta)]\end{aligned}$$

Le théorème de Cauchy-Schwartz donne

$$\mathrm{Cov}^2[T(X) S_X(\theta)] \leq \mathrm{Var}(T(X)) \mathrm{Var}(S_X(\theta))$$

Comme le modèle est régulier, la variance du score est égale à l'information de Fisher. On a donc :

$$\mathrm{Var}(T(X)) = \mathrm{Var}\left(\widehat{\theta}\right) \geq \frac{\mathrm{Cov}^2[T(X) S_X(\theta)]}{\mathrm{Var}(S_X(\theta))} = \frac{1}{I_n(\theta)}$$

La variance de tout estimateur sans biais (dans un modèle régulier) est donc minorée par $I_n(\theta)^{-1}$: c'est la complexité intrinsèque au modèle, qui ne dépend pas de l'estimateur choisi, mais seulement du paramètre que l'on cherche à estimer et du modèle statistique. Si l'information de Fisher est grande, cette complexité est petite et on peut espérer trouver de bons estimateurs. Au contraire, si l'information de Fisher est petite, alors on n'aura que des estimateurs de grande variance (c'est-à-dire peu précis).

Nous venons de démontrer qu'aucun estimateur sans biais ne peut avoir une variance inférieure à la borne de Cramer-Rao. Ceci nous permet de définir un estimateur efficace et un estimateur asymptotiquement efficace :

Définition 60 (Efficacité) *Un estimateur sans biais est efficace si sa variance est égale à la borne de Cramer-Rao :*

$$\widehat{\theta} \text{ est efficace si } \mathbb{E}[\widehat{\theta}] = \theta \text{ et } \mathrm{Var}\left(\widehat{\theta}\right) = \frac{1}{I_n(\theta)}$$

Définition 61 (Efficacité asymptotique) *Une suite d'estimateurs $\widehat{\theta}$ est asymptotiquement efficace si :*

$$\lim_{n\to\infty} \mathbb{E}[\widehat{\theta}] = \theta \text{ et } \lim_{n\to\infty} n\mathrm{Var}\left(\widehat{\theta}\right) = \frac{1}{I_1(\theta)}$$

Si on dispose d'un estimateur efficace, il est inutile d'essayer d'en trouver un qui soit plus performant, le théorème FDCR nous dit qu'il n'en existe pas (parmi les estimateurs sans biais).

Remarque 16 Un estimateur biaisé ne peut pas être efficace, mais il peut éventuellement être asymptotiquement efficace.

Proposition 21 L'efficacité (asymptotique ou non) est conservée par transformation affine :
si $\hat{\theta}$ est un estimateur efficace de θ_0 alors $a\hat{\theta} + b$ estime efficacement $a\theta + b$.

La démonstration est l'application immédiate des propriétés de l'espérance et la variance.

3 Modèles de forme exponentielle

Pour la plupart des familles de lois classiques, on peut construire un modèle où l'estimation est efficace. C'est une similitude dans la forme des densités de ces lois qui permet de conclure. On définit alors cette classe de modèles :

Définition 62 (Modèle de forme exponentielle)
Un modèle $\left((X_1,\ldots,X_n) \in \mathcal{X}^n, \prod_i f_\theta(X_i), \theta \in \Theta\right)$ *est de forme exponentielle ssi il*

existe trois fonctions $c : \Theta \to \mathbb{R}^+$, $g : \mathcal{X}^n \to \mathbb{R}^+$ *et* $b : \Theta \to \mathbb{R}$ *avec* $\forall \theta \in \Theta$, $\dfrac{\partial b}{\partial \theta} \neq 0$ *et une statistique* $T\mathcal{X}^n \to \mathbb{R}$ *telles que*

$$L_n(\theta) = c(\theta) g(X_1, \ldots, X_n) \exp\left(b(\theta) T(X_1, \ldots, X_n) \right)$$

Dans certain cas, en particulier lorsque Θ *est multidimensionnel,* b *et* T *peuvent être à valeurs dans* $\mathbb{R}^p$ *avec* $p > 1$*. Leur produit est alors un produit vectoriel :*

$$L_n(\theta) = c(\theta) g(X_1, \ldots, X_n) \exp\left(\sum_{k=1}^{p} b_k(\theta) T_k(X_1, \ldots, X_n) \right)$$

Proposition 22 Si le modèle est de forme exponentielle, alors T est une statistique exhaustive.

La démonstration est immédiate par le théorème de factorisation.

Théorème 21 *Soit un modèle régulier.*

- *S'il existe un estimateur efficace* T*, alors le modèle est de forme exponentielle,* T *est la statistique exhaustive qui intervient dans l'exponentielle.*
- *Réciproquement, si le modèle est de forme exponentielle alors* T *est un estimateur efficace de* θ*.*

On a vu que l'efficacité était conservée par les transformations affines, on peut donc transformer le modèle en posant $\theta' = a\theta + c$ et on obtient un nouvel estimateur efficace $T' = aT + b$. La forme exponentielle est-elle même stable par transformation affine de b et T.

On peut également remarquer qu'il est très simple de vérifier qu'un modèle de forme exponentielle est régulier :

- le lien entre θ et l'échantillon ne peut avoir lieu que sous la forme d'un produit dans l'exponentielle, donc $\mathcal{X}$ ne peut pas dépendre de θ ;
- la double dérivabilité ne conserve que c et b (qui est déjà dérivable une fois).

Exemple Soit un échantillon gaussien de variance σ^2 connue. Montrons que le modèle est de forme exponentielle. $\overline{X}$ est alors automatiquement une statistique exhaustive et un estimateur efficace de μ. La vraisemblance s'écrit :

$$\begin{aligned} L_n(\mu) &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} e^{\frac{-\sum_i X_i^2}{2\sigma^2}} \times e^{\frac{-n\mu^2}{2\sigma^2}} \times \exp\left(\frac{n\overline{X}}{\sigma^2} \right) \\ &= c(\mu) g(X_1, \ldots, X_n) \exp\left(b(\mu) T(X_1, \ldots, X_n) \right) \end{aligned}$$

avec $b(\mu) = \dfrac{n\mu}{\sigma^2}$ et $T(X_1, \ldots, X_n) = \overline{X}$.

On voit que le modèle est régulier. On peut faire passer le $\frac{n}{\sigma^2}$ du côté de b ou de T, à cause de la stabilité par transformation affine de l'efficacité. Le calcul de l'espérance de T permet de vérifier que c'est bien μ qu'on estime, et non $n\mu$ par exemple. Ici c'est évident.

On obtient par l'efficacité

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} = I_n^{-1}(\mu)$$

On peut également le vérifiant en calculant l'information de Fisher (par la seconde forme puisque le modèle est régulier).

Chapitre 6

Maximum de vraisemblance

A Maximum de vraisemblance unidimensionnel

1 Définitions

Définition 63 (Estimateur du maximum de vraisemblance) *Soit un modèle statistique. Alors on appelle* estimateur du maximum de vraisemblance $\widehat{\theta}$ *toute statistique (s'il en existe) vérifiant :*

$$L_n\left(\widehat{\theta}\right) = \max_{\theta\in\Theta} L_n(\theta)$$

On note alors :

$$\hat{\theta}_{MV} = \arg\max_{\theta\in\Theta} L_n(\theta)$$

Théorème 22 *Si le modèle est régulier, alors $\widehat{\theta}$ est donné par :*

$$\begin{cases} \dfrac{\partial L}{\partial\theta}(\widehat{\theta}) = 0 \\ \dfrac{\partial^2 L}{\partial\theta^2}(\widehat{\theta}) < 0 \end{cases}$$

On note souvent

$$\frac{\partial L_n}{\partial\theta}(\hat{\theta}_{MV}) = \left.\frac{\partial L_n(\theta)}{\partial\theta}\right|_{\theta=\widehat{\theta}_{MV}}$$

pour insister sur le fait qu'on dérive par rapport à θ l'expression générale de la vraisemblance avant de prendre sa valeur en $\widehat{\theta}_{MV}$. Évidement, si on commence par remplacer θ par $\widehat{\theta}_{MV}$, on ne peut plus dériver en θ !

Proposition 23 (Équation de score)
Si le modèle est régulier, alors $\hat{\theta}_{MV}$ est aussi donné par

$$\begin{cases} S_n(\hat{\theta}_{MV}) &= 0 \\ \dfrac{\partial S_n}{\partial\theta}(\hat{\theta}_{MV}) &< 0 \end{cases}$$

Remarque 17 Pour les modèles réguliers, le score est centré : $\mathbb{E}(S_n(\theta)) = 0$. C'est l'intuition qui justifie de choisir un estimateur qui annule le score empirique.

Démonstration *On remarque simplement que* $S_n(\theta) = \dfrac{1}{L_n(\theta)} \dfrac{\partial L_n}{\partial \theta}(\theta)$. *Le premier facteur étant toujours non nul,* S_n *s'annule si et seulement si* $\dfrac{\partial L_n}{\partial \theta}(\theta)$ *s'annule.*

Pour la condition de second ordre, on a

$$\frac{\partial S_n}{\partial \theta}(\theta) = \frac{\partial}{\partial \theta} \left(\frac{\dfrac{\partial L_n}{\partial \theta}(\theta)}{L_n(\theta)} \right)$$

On retrouve la dérivée du quotient de la démonstration de la borne de Cramer-Rao :

$$\frac{\partial S_n}{\partial \theta}(\theta) = \frac{\dfrac{\partial^2 L_n}{\partial \theta^2}(\theta)}{L_n(\theta)} - \left(\frac{\dfrac{\partial L_n}{\partial \theta}}{L_n(\theta)} \right)^2$$

Le second terme, le score au carré, est nul en $\hat{\theta}_{MV}$ *et le premier est du même signe que son numérateur,* $\dfrac{\partial^2 L_n}{\partial \theta^2}(\theta)$, *on retrouve donc la condition du second ordre sur* L_n.

Exemple On dispose d'un échantillon $(X_1, \ldots, X_n)$ issu d'une $\mathcal{N}\left(\mu, \sigma^2\right)$ avec σ^2 connue. La densité de la $\mathcal{N}\left(\mu, \sigma^2\right)$ s'écrit :

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

et donc la vraisemblance de notre échantillon de n observations est :

$$L_n\left(\mu, \sigma^2\right) = \prod_{i=1}^{n} f_\mu(X_i) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2} \right)$$

et la log-vraisemblance :

$$\log L_n\left(\mu\right) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

On doit donc trouver la valeur de μ qui maximise cette fonction. Écrivons la condition du premier ordre en dérivant la log-vraisemblance en μ puis en les égalisant à 0.

$$S_n(\mu) = \frac{\partial \log L_n\left(\mu\right)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \mu\right)$$

On a donc en $\widehat{\theta}_{MV}$:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\widehat{\theta}_{MV}\right)=0 \quad \Rightarrow \quad \sum_{i=1}^{n}X_i=n\widehat{\theta}_{MV} \quad \Rightarrow \quad \widehat{\theta}_{MV}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}_n$$

il reste juste à vérifier que la condition de second ordre :

$$\frac{\partial S_n}{\partial \mu}(\mu)=\frac{-n}{\sigma^2}$$

c'est donc négatif pour toute valeur de μ et donc en particulier en $\mu=\overline{X}_n$.

On sait que $\overline{X}_n$ estime μ sans biais et que sa variance est $\dfrac{\sigma^2}{n}$. On peut alors vérifier qu'il est efficace (ce que l'on savait déjà à cause de la forme exponentielle du modèle gaussien). Il suffit de calculer l'information de Fisher :

$$I_n(\mu)=-\mathbb{E}\left[\frac{\partial}{\partial \mu}S_n(\mu)\right]=\frac{n}{\sigma^2}$$

d'après le calcul précédent de la condition de second ordre.

On a bien $V\left(\widehat{\theta}_{MV}\right)=I_n^{-1}(\mu)$.

2 Propriétés à distance finie

Théorème 23 *Soit un modèle statistique. S'il existe une statistique exhaustive T, alors l'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV}$ est une fonction de T.*

Démonstration *Comme on dispose d'une statistique exhaustive T, en vertu du théorème de factorisation, il existe donc deux fonctions positives g et h telles qu'on ait :*

$$L_n(\theta)=g(X_1,\ldots,X_n)h(\theta,T(X_1,\ldots,X_n))$$

Donc $\widehat{\theta}_{MV}=\underset{\theta\in\Theta}{argmax}\, L_n(\theta)=\underset{\theta\in\Theta}{argmax}\, h\left(\theta,T(X_1,\ldots,X_n)\right)$, qui est bien une fonction de T.

Théorème 24 *Si le modèle est régulier et qu'il existe un estimateur efficace $\widehat{\theta}^*$, alors $\widehat{\theta}^*=\widehat{\theta}_{MV}$. L'existence de $\widehat{\theta}^*$ entraîne que le modèle est de forme exponentielle et $\widehat{\theta}^*=T$.*

Démonstration *Pour avoir un estimateur efficace, il faut qu'il y ait égalité dans la borne de Cramer-Rao : $I_n^{-1}(\theta)=Var\left(\widehat{\theta}^*\right)$, c'est-à-dire égalité dans l'inégalité de Cauchy-Schwarz : $\mathbb{E}\left[(\widehat{\theta}^*-\theta)S_n(\theta)\right]^2=Var(\widehat{\theta}^*)\,Var\left(S_n(\theta)\right)$. Donc $\widehat{\theta}^*-\theta$ et $S_n(\theta)$ sont colinéaires. Or en $\widehat{\theta}_{MV}$, on a $S_n\left(\widehat{\theta}_{MV}\right)=0$. Donc on a bien $\widehat{\theta}^*-\widehat{\theta}_{MV}=0$.*

3 Propriétés asymptotiques

On s'intéresse aux propriétés d'une suite d'estimateurs $\widehat{\theta}_{\mathrm{MV}}$ comme correspondant à une suite de modèles $((X_1, \ldots, X_n) \in \mathcal{X}^n, L_n(\theta), \theta \in \Theta)$.

Théorème 25 *Si, pour tout n, le modèle est régulier, alors $\widehat{\theta}_{MV}$ converge presque sûrement vers θ et* a fortiori *$\widehat{\theta}_{MV}$ est convergent. On a de plus un estimateur asymptotique normal $\dfrac{\widehat{\theta}_{MV} - \theta}{\sqrt{(I_n(\theta))^{-1}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0;1)$. Donc $\widehat{\theta}_{MV}$ est asymptotiquement sans biais, efficace et normal.*

Théorème 26 *Soit h une fonction $\mathcal{C}^1$ de Θ vers $\mathbb{R}$. On peut alors estimer $h(\theta)$ par $h\left(\widehat{\theta}_{MV}\right)$ avec la Δ-méthode :*

$$\sqrt{n}\left(h\left(\widehat{\theta}_{MV}\right) - h(\theta)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{h'(\theta)^2}{I_1(\theta)}\right)$$

Par exemple, si $\theta := \sigma^2$, la fonction racine carrée permet de passer de l'estimateur de la variance à celle de l'écart-type.

B Maximum de vraisemblance multidimensionnel

Prenons l'exemple d'un modèle gaussien avec $\theta := (\mu, \sigma^2)$. Par substitution, on estime θ par $\widehat{\theta} := \left(\overline{X}, \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)$.

1 Cadre multidimensionnel

Soit $X := (X_1, \ldots, X_n)$ un vecteur aléatoire. Alors on a pour espérance le vecteur $\mathbb{E}(X) := (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))$; le biais est alors $\left(\mathbb{E}(\widehat{\theta}_1 - \theta_1), \ldots, \mathbb{E}(\widehat{\theta}_n - \theta_n)\right)$ et la variance est en fait une matrice de variance-covariance

$$\mathrm{Var}(X) := (\mathrm{Cov}(X_i, X_j))_{1\le i,j\le p} = \mathbb{E}\left((X - \mathbb{E}(X))(X - \mathbb{E}(X))^t\right).$$

Les matrices de variance-covariance sont symétriques réelles.

Définition 64 *Sur l'espace des matrices symétriques, on définit la relation d'ordre partielle $\succeq$ par $M \succeq N$ si, et seulement si, pour tout X de $\mathbb{R}^p$, on a $X^t(M-N)X \ge 0$.*

On définit la convergence des estimateurs et la normalité asymptotique comme en unidimensionnel. La loi des grands nombres et le théorème de la limite centrée existent aussi en multidimensionnel.

2 Borne de Cramer-Rao

Définition 65 (Score) *Si f_θ est $\mathcal{C}^1$ en chacune des coordonnées de θ, alors* $S(x,\theta) = \dfrac{\partial}{\partial\theta}\log(f_\theta(x)) = \begin{pmatrix} \dfrac{\partial}{\partial_1\theta}\log(f_\theta(x)) \\ \vdots \\ \dfrac{\partial}{\partial_p\theta}\log(f_\theta(x)) \end{pmatrix}$.

Définition 66 (Information de Fischer)

$$I_1(\theta) = \mathbb{E}\left[S(X,\theta)^t S(X,\theta)\right]$$

Si l'échantillon est i.i.d. alors on a $S_n(\theta) = \sum_{i=1}^{n} S(X_i,\theta)$.

Si le modèle est régulier, alors $\mathcal{X}$ ne dépend pas de θ, f_θ est $\mathcal{C}^2$ en θ. On a dès lors $I_1(\theta) = -\mathbb{E}\left(\dfrac{\partial}{\partial\theta}S(x,\theta)\right)$ *et* $I_n(\theta) = -n\mathbb{E}\left(\dfrac{\partial}{\partial\theta}S(x,\theta)\right)$ *et*

$$I_n(\theta) = -n\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(f_\theta(x))\right)$$

Exemple pour la gaussienne :

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\theta_2}}\exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right).$$

Après calcul, on en déduit

$$I_n(\theta) = -\mathbb{E}\begin{pmatrix} -\dfrac{n}{\theta_2} & -\dfrac{\sum_{i=1}^n (X_i-\theta_i)}{\theta_2^2} \\ -\dfrac{X_1-\theta_1-X_2+\theta_2}{\theta_2^2} & -\dfrac{n}{2\theta_2^2}\dfrac{\sum_{i=1}^n (X_i-\theta_i)^2}{\theta_2^3} \end{pmatrix} = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{pmatrix}$$

Théorème 27 (de Cramer-Rao) *Si le modèle est régulier et que $\widehat{\theta}$ est sans biais, alors $Var(\widehat{\theta}) \geq I_n^{-1}(\theta)$. Lorsqu'il y a égalité, on dit que $\widehat{\theta}$ est efficace.*

3 Maximum de vraisemblance multidimensionnelle

Définition 67

$$\widehat{\theta}_{MV} = \underset{\theta\in\Theta\subset\mathbb{R}^p}{argmax}\, L_n(\theta)$$

Théorème 28 *Si le modèle est régulier, alors $\widehat{\theta}_{MV}$ vérifie :*

$$\begin{cases} S_n\left(\widehat{\theta}_{MV}\right) = 0_p \\ Hess\left(log\left(L_n(\theta)\right)\right)\Big|_{\theta=\widehat{\theta}_{MV}} \prec 0 \end{cases}$$

Définition 68 *Une matrice M de $\mathcal{M}_p(\mathbb{R})$ est définie négative si, et seulement si, pour tout X de $(R^*)^p$, on a $X^t MX < 0$.*

Théorème 29 *S'il existe une statistique T exhaustive, alors $\widehat{\theta}_{\boldsymbol{MV}} = h(T)$.*

Dans un mod-le régulier, s'il existe $\widehat{\theta}^*$ efficace, c'est $\widehat{\theta}_{\boldsymbol{MV}}$.

Sans condition, $\widehat{\theta}_{\boldsymbol{MV}} = h(T)$ est convergent, asymptotiquement sans biais, efficace et normal.

C Rapport de vraisemblance

1 Intervalles de confiance

On dispose d'un estimateur $\widehat{\theta}_{\mathrm{MV}} = \underset{\theta \in \Theta \subset \mathbb{R}^p}{\operatorname{argmax}} L_n(\theta)$.

Définition 69 (Rapport de vraisemblance)

$$\begin{array}{rccc} R: & \Theta & \to & \mathbb{R}_+^* \\ & \theta & \mapsto & \dfrac{L_n(\theta)}{\underset{\theta \in \Theta \subset \mathbb{R}^p}{argmax}\, L_n(\theta)} \end{array}$$

Le rapport de vraisemblance est donc une vraisemblance « relative », qui exprime la vraisemblance d'une valeur du paramètre comme proportion de la vraisemblance maximale.

Théorème 30 (Intervalle de confiance)
Soit $X_1, \ldots, X_n$ i.i.d. de variance finie (inconnue) de densité f_{θ_0} alors $-2\log\left(R\left(\theta_0\right)\right) \xrightarrow{\mathcal{L}} \chi^2(p)$, où $p = dim(\Theta)$.

$$C_n := \left\{ \theta \in \Theta;\ -2\log\left(R\left(\theta\right)\right) \leq F^{-1}_{\chi^2(p)}(1-\alpha) \right\}$$

est donc un intervalle de confiance au niveau de confiance asymptotique $(1-\alpha)$.

Théorème 31 (Test)
Le test $\Phi = \mathbb{1}_{\theta_0 \in C_n}$ est un test asymptotique de niveau $1-\alpha$ de $H_0 : (\theta = \theta_0)$ contre $(\theta \neq \theta_0)$.

2 La méthode de Neyman et Pearson

Le test précédent est très général, mais n'est valide qu'asymptotiquement, et on ne sait rien de sa puissance $(1-\beta)$: comment se comporte-t-il lorsque l'hypothèse nulle est fausse ?

Idéalement, on souhaiterait obtenir un test qui minimiserait à la fois les deux types d'erreurs, c'est-à-dire qui minimiserait à la fois α et β. La méthode de Neyman et Pearson, basée sur le rapport de vraisemblance, permet de minimiser β pour un niveau $1-\alpha$ fixé. On fixe généralement α à des valeurs faibles comme 10 %, 5 %, 1 %, etc. En contrôlant en priorité l'erreur de première espèce, on considère implicitement

que l'erreur la plus grave est de rejeter H_0 à tort. La méthode de Neyman-Pearson dit qu'une fois qu'on a limité à une valeur acceptable le risque de rejeter H_0 à tort, on cherche ensuite la procédure qui minimise le risque de ne pas rejeter H_0 alors qu'elle est fausse.

Théorème 32 (Lemme de Neyman-Pearson) *Pour tout $\alpha \in]0,1[$, le test le plus puissant de niveau $1-\alpha$ pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ est de la forme*

$$\Phi\left(X_1, \ldots, X_n\right) = \mathbb{1}_{\frac{L_n\left(\theta_0\right)}{L_n\left(\theta_1\right)} \leq k}$$

Autrement dit :

$$\Phi\left(X_1, \ldots, X_n\right) = \left\{ \begin{array}{lll} 1 & si & L_n(\theta_0) \leq k\, L_n(\theta_1) \\ 0 & si & L_n(\theta_0) > k\, L_n(\theta_1) \end{array} \right.$$

Le seuil k est alors défini par :

$$\Pr\left(\frac{L_n(\theta_0)}{L_n(\theta_1)} \leq k\right) = \alpha$$

La région de rejet du test de Neyman est :

$$W = \left\{ (X_1, \ldots, X_n) \in \mathbb{R}^n : \frac{L_n(\theta_0)}{L_n(\theta_1)} \leq k \right\}$$

3 Test de la moyenne d'une loi normale

Soit un échantillon $X = (X_1, \ldots, X_n)$ i.i.d de taille n issu d'une loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$ où la variance σ^2 est connue. On va appliquer le Lemme de Neyman-Pearson afin de construire le test.

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1, \text{ avec } \mu_1 > \mu_0 \end{array} \right.$$

Le Lemme de Neyman-Pearson nous dit que la région de rejet du test UPP de ce jeu d'hypothèse est donnée par :

$$W = \left\{ (X_1, \ldots, X_n) \in \mathbb{R}^n : \frac{L_n(\mu_0)}{L_n(\mu_1)} \leq k \right\}$$

On va transformer l'expression $\dfrac{L_n(\mu_0)}{L_n(\mu_1)} \leq k$ afin d'aboutir à une inégalité comparant une statistique $S\left(X\right)$ à une quantité ne dépendant pas des données :

$$
\begin{aligned}
\frac{L_n(\mu_0)}{L_n(\mu_1)} &\leq k \\
\frac{\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{(X_i-\mu_0)}{\sigma}\right)^2\right)}{\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{(X_i-\mu_1)}{\sigma}\right)^2\right)} &\leq k \\
\frac{\prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{(X_i-\mu_0)}{\sigma}\right)^2\right)}{\prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{(X_i-\mu_1)}{\sigma}\right)^2\right)} &\leq k \\
\sum_{i=1}^{n}\left(-\frac{1}{2}\left(\frac{(X_i-\mu_0)}{\sigma}\right)^2\right) - \sum_{i=1}^{n}\left(-\frac{1}{2}\left(\frac{(X_i-\mu_1)}{\sigma}\right)^2\right) &\leq \log(k) \\
\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_1)^2 - \sum_{i=1}^{n}(X_i-\mu_0)^2\right) &\leq \log(k) \\
\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(X_i^2 - 2X_i\mu_1 + \mu_1^2\right) - \sum_{i=1}^{n}\left(X_i^2 - 2X_i\mu_0 + \mu_0^2\right)\right) &\leq \log(k) \\
\frac{1}{2\sigma^2}\left(-2\overline{X}_n\mu_1 + n\mu_1^2 + 2\overline{X}_n\mu_0 - n\mu_0^2\right) &\leq \log(k) \\
\frac{1}{2\sigma^2}\left(\overline{X}_n\left(2n\left(\mu_0-\mu_1\right)\right) + n\left(\mu_1^2-\mu_0^2\right)\right) &\leq \log(k) \\
\overline{X}_n &\geq k' = \frac{2\sigma^2\log(k) - n\left(\mu_1^2-\mu_0^2\right)}{2n\left(\mu_0-\mu_1\right)}
\end{aligned}
$$

où la dernière inégalité est inversée car dans notre cas $\mu_0 - \mu_1 < 0$.

On aboutit donc à la région de rejet suivante :

$$W = \left\{(X_1, \ldots, X_n) \in \mathbb{R}^n : \overline{X}_n \geq k'\right\}$$

On rejette si la moyenne observée sur notre échantillon est supérieure à une valeur seuil k' ne dépendant pas des données. Afin de finaliser la construction du test, il nous faut déterminer la valeur de k'. Pour ce faire, on doit fixer le niveau α du test. On doit avoir :

$$\Pr\left(\overline{X}_n \geq k' \,|\mu = \mu_0\right) = \alpha$$

La distribution de $\overline{X}_n$ est la $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$; simplifions cette expression en faisant dépendre de la $\mathcal{N}(0,1)$. L'événement $\overline{X}_n \geq k'$ est identique à l'événement $\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq k''$ avec $k'' = \frac{k'-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On peut donc réécrire la région de rejet du test comme :

$$W = \left\{ (X_1, \ldots, X_n) \in \mathbb{R}^n : \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq k'' \right\}$$

et la valeur de k'' est déterminée par :

$$\Pr\left(\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq k'' \middle| \mu = \mu_0 \right) = \alpha$$

Cette fois-ci on a alors :

$$\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$$

Cette probabilité est facilement calculable grâce aux tables de la $\mathcal{N}(0,1)$. La valeur de k'' est donc le quantile de niveau $1-\alpha$ de la $\mathcal{N}(0,1)$, c'est-à-dire $z_{(1-\alpha)}$.

La région de rejet de notre test pour un niveau α est donnée par :

$$W = \left\{ (X_1, \ldots, X_n) \in \mathbb{R}^n : \overline{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{(1-\alpha)} \right\}$$

Notons que cette région de rejet est complètement définie : n, $\overline{X}_n$ et σ sont directement calculables à partir de l'échantillon observé, nous fixons μ_0, et $z_{(1-\alpha)}$ se lit dans une table de la $\mathcal{N}(0,1)$. La règle de décision du test est donc : si $\bar{X}_n > k'$, alors décider H_1, sinon accepter H_0.

La puissance $1-\beta$ de notre test va s'écrire :

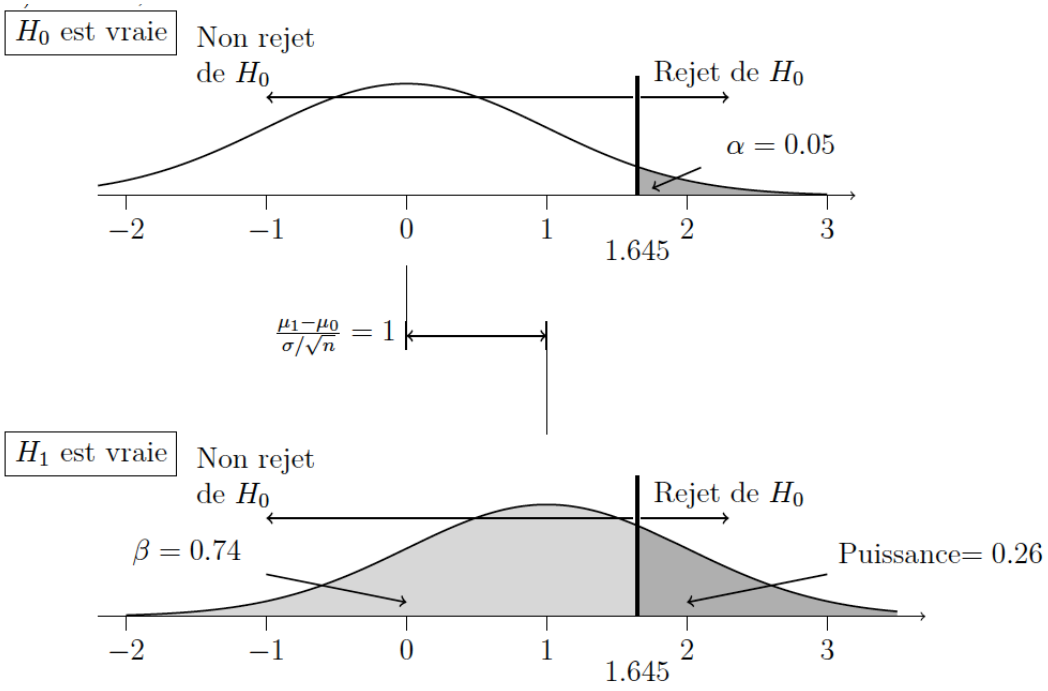
$$\Pr\left(\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{(1-\alpha)} \middle| \mu = \mu_1 \right) = 1 - \beta$$

Lorsque $\mu = \mu_1$, $\dfrac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}\left(\dfrac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}, 1 \right)$. Pour les besoins du calcul, on se ramène à une $\mathcal{N}(0,1)$:

$$\Pr\left(\frac{\overline{X}_n - \mu_1}{\frac{\sigma}{\sqrt{n}}} \geq z_{(1-\alpha)} + \frac{\mu_0 - \mu_1}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_1 \right) = 1 - \beta$$

La puissance du test sera alors donnée par $1 - \Phi\left(z_{(1-\alpha)} + \dfrac{\mu_0 - \mu_1}{\frac{\sigma}{\sqrt{n}}} \right)$.

La figure ci-dessous donne une représentation des distributions de la statistique de test $\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu_0\right)$ sous H_0 et sous H_1, de la région critique ainsi que de la puissance pour un test de $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1 > \mu_0$ avec $\frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}} = 1$ et $\alpha = 0,05$ et donc $z_{(1-\alpha)} = 1,645$.



Exemple Deux sondages menés auprès des français pour les résultats du premier tour de l'élection présidentielle de 2017 dans l'optique de la candidature de François Fillon pour l'UMP donnent les résultats suivants pour la candidate du Front National :

- 31% (sondage Ifop-Fiducial pour i>Télé et Sud Radio du 4 novembre 2014).
- 32% (sondage Ifop pour le Figaro des 3 et 4 septembre 2014).

En jeune statisticien.ne vous avez mené le même sondage sur 100 personnes et vous avez obtenu le résultat suivant : 29 % pour Marine Le Pen (dans l'hypothèse de la candidature de François Fillon pour l'UMP). On admet que les intentions de vote suivent une loit normale de moyenne μ et d'écart-type 2 %. On désire tester l'estimation des deux sondages au seuil de 5 %.

On effectue le test :

$$\begin{cases} H_0 : \mu = 31 \\ H_1 : \mu = 32, \text{ avec } \mu_1 > \mu_0 \end{cases}$$

Notre zone de rejet s'écrit alors :

$$W = \left\{ \overline{x}_n \geq 31 + \frac{2}{\sqrt{100}} 1,645 = 31,329 \right\}$$

Comme $\overline{x}_n = 29 < 31,329$, on accepte H_0.

Annexes