SY02
Cours de Statistique

T. Denœux, G. Govaert, J.-B. Leger, B. Quost et S. Rousseau

Version courante :
- Base : Printemps 2022
- Date : 2023-01-05 11:47:58
- Version : P2022-7-gbe81c6c

Table des matières

1 Introduction **9**
- 1.1 Concepts fondamentaux . . . 9
- 1.2 Variables, distributions . . . 9
- 1.3 Modes d'étude d'une population . . . 10
- 1.4 Objectifs d'une étude statistique . . . 10

2 Éléments de Statistique descriptive **11**
- 2.1 Tableaux et graphiques . . . 11
  - 2.1.1 Variable quantitative discrète ou qualitative ordinale . . . 11
  - 2.1.2 Variable continue . . . 12
  - 2.1.3 Fonction de répartition empirique . . . 14
- 2.2 Résumés numériques . . . 15
  - 2.2.1 Indicateurs de tendance centrale . . . 15
  - 2.2.2 Indicateurs de dispersion . . . 18
  - 2.2.3 Diagramme en boîte . . . 19
  - 2.2.4 Coefficients de corrélation . . . 21

3 Rappels de Probabilités **23**
- 3.1 Introduction . . . 23
- 3.2 Expérience aléatoire . . . 23
- 3.3 Variable aléatoire . . . 24
  - 3.3.1 Définition . . . 24
  - 3.3.2 Loi de probabilité d'une v.a. . . . 25
  - 3.3.3 Exemples de lois de probabilité . . . 26
- 3.4 Résumés numériques d'une loi de probabilité . . . 36
  - 3.4.1 Espérance mathématique . . . 36
  - 3.4.2 Variance . . . 36
  - 3.4.3 Généralisation : moments d'ordre k . . . 37
  - 3.4.4 Fractiles (ou quantiles) d'une loi continue . . . 38
- 3.5 Vecteurs aléatoires . . . 38
  - 3.5.1 Définition . . . 38
  - 3.5.2 Loi de probabilité d'un vecteur aléatoire . . . 38
  - 3.5.3 Exemples de lois multidimensionnelles . . . 40
  - 3.5.4 Moments d'un vecteur aléatoire . . . 40
  - 3.5.5 Indépendance de variables aléatoires . . . 42
  - 3.5.6 Loi conditionnelle . . . 43
  - 3.5.7 Somme de variables aléatoires . . . 43
- 3.6 Notions de convergence stochastique . . . 43

3.6.1 Convergence en probabilité . . . 43
3.6.2 Convergence en loi . . . 44
3.7 Lois dérivées de la loi normale . . . 45
3.7.1 Loi du χ^2 . . . 45
3.7.2 Loi de Student . . . 47
3.7.3 Loi de Fisher . . . 49

4 Échantillonnage **51**
4.1 Notion d'échantillon aléatoire . . . 51
4.1.1 Définition . . . 51
4.1.2 Notion de statistique . . . 52
4.2 Moyenne empirique . . . 52
4.2.1 Propriétés à distance finie (n fixé) . . . 52
4.2.2 Propriétés asymptotiques ($n \to \infty$) . . . 53
4.3 Variance empirique . . . 55
4.3.1 Propriétés à distance finie . . . 55
4.3.2 Propriétés asymptotiques . . . 56
4.4 Moments empiriques . . . 56
4.5 Cas d'un échantillon gaussien . . . 57
4.6 Fonction de répartition empirique . . . 57
4.7 Fractiles (ou quantiles) empiriques . . . 58
4.8 Échantillonnage stratifié . . . 59
4.8.1 Notation . . . 59
4.8.2 Définition . . . 59
4.8.3 Propriétés . . . 59
4.8.4 Échantillonnage stratifié optimal . . . 60
4.8.5 Utilisation pratique . . . 60
4.9 Démonstrations . . . 60
4.9.1 Preuve de la Proposition 4.3 . . . 60
4.9.2 Éléments de preuve du Théorème 4.3 . . . 61

5 Estimation ponctuelle **63**
5.1 Notion d'estimateur . . . 63
5.2 Propriétés élémentaires d'un estimateur . . . 63
5.2.1 Estimateur sans biais ou asymptotiquement sans biais . . . 63
5.2.2 Estimateur convergent . . . 64
5.2.3 Estimation de l'espérance et de la variance . . . 64
5.2.4 Précision . . . 64
5.3 Moments . . . 65
5.4 Maximum de vraisemblance . . . 66
5.4.1 Fonction de vraisemblance . . . 66
5.4.2 Estimateur du maximum de vraisemblance . . . 66
5.4.3 Équation de vraisemblance . . . 67
5.4.4 Invariance fonctionnelle . . . 67
5.4.5 Convergence . . . 67
5.4.6 Normalité asymptotique . . . 67
5.4.7 Cas d'un paramètre vectoriel . . . 68
5.5 Efficacité . . . 69
5.5.1 Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR) . . . 69
5.5.2 Estimateur efficace . . . 69

5.5.3 Estimateur sans biais de variance minimale . . . 70
5.6 Démonstrations . . . 70
5.6.1 Preuve du Théorème 5.2 . . . 70
5.6.2 Preuve du Théorème 5.3 . . . 71
5.6.3 Preuve du Théorème 5.4 . . . 72
5.6.4 Preuve du Théorème 5.5 . . . 73

6 Intervalle de confiance 75
6.1 Introduction . . . 75
6.2 Définitions . . . 75
6.3 Construction pratique . . . 76
6.3.1 Notion de fonction pivotale . . . 76
6.3.2 Utilisation . . . 76
6.4 Espérance . . . 78
6.4.1 Cas gaussien, variance connue . . . 78
6.4.2 Cas gaussien, variance inconnue . . . 79
6.4.3 Cas général . . . 79
6.5 Variance . . . 80
6.5.1 Cas où l'espérance est connue . . . 80
6.5.2 Cas où l'espérance est inconnue . . . 81
6.6 Proportion . . . 81
6.6.1 Avec application du théorème de Slutsky . . . 81
6.6.2 Sans l'application du théorème de Slutsky . . . 82
6.7 Construction à partir d'un EMV . . . 83

7 Régression linéaire 85
7.1 Exemple introductif . . . 85
7.2 Le modèle . . . 86
7.3 Estimation des paramètres . . . 86
7.3.1 Estimation de a et b . . . 87
7.3.2 Estimation de σ^2 . . . 88
7.4 Principales propriétés . . . 88
7.4.1 Propriétés des estimateurs $\hat{a}$ et $\hat{b}$ . . . 88
7.4.2 Analyse de la variance . . . 88
7.4.3 Estimateur sans biais de σ^2 . . . 89
7.5 Pratique de la régression . . . 90
7.5.1 Effet de levier et identification des abscisses extrêmes . . . 90
7.5.2 Analyse des résidus . . . 93
7.5.3 Tableau d'analyse de la variance . . . 97
7.5.4 Mesure de l'ajustement . . . 98
7.5.5 Intervalles de confiance sur a et b . . . 99
7.5.6 Prévision . . . 99
7.6 Exemple . . . 101
7.6.1 Détermination de la droite de régression estimée . . . 101
7.6.2 Analyse de la variance . . . 101
7.6.3 Prévision . . . 101
7.7 Démonstrations . . . 102
7.7.1 Preuve de la Proposition 7.1 . . . 102
7.7.2 Preuve de la Proposition 7.3 . . . 103

8 Tests d'hypothèses **105**
8.1 Exemples introductifs . . . 105
8.2 Principes de résolution . . . 106
8.3 Théorème de Neyman-Pearson . . . 109
8.3.1 Principe . . . 109
8.3.2 Application : test sur l'espérance d'une loi normale (σ^2 connu) 110
8.4 Test UPP . . . 111
8.4.1 Principe . . . 111
8.4.2 Test sur l'espérance d'une loi normale (σ^2 connu) . . . 114
8.5 Test du rapport de vraisemblance . . . 114
8.5.1 Principe . . . 114
8.5.2 Application : test sur l'espérance d'une loi normale (σ^2 connu) 115
8.5.3 Test approché . . . 117
8.6 Stratégie empirique . . . 117
8.7 Preuve du Théorème de Neyman-Pearson . . . 117

9 Tests de conformité **119**
9.1 Espérance . . . 119
9.1.1 Variance connue . . . 119
9.1.2 Variance inconnue (test de Student) . . . 121
9.2 Variance . . . 124
9.3 Lien avec les intervalles de confiance . . . 125
9.4 Test sur une proportion . . . 125
9.5 Démonstrations . . . 128
9.5.1 Preuve de la Proposition 9.3 . . . 128
9.5.2 Preuve de la Proposition 9.4 . . . 129
9.5.3 Preuve de la Proposition 9.5 . . . 130

10 Tests d'homogénéité **133**
10.1 Introduction et notations . . . 133
10.2 Espérances . . . 133
10.2.1 Notations . . . 133
10.2.2 Variances connues . . . 133
10.2.3 Variances inconnues mais égales : test de Student . . . 135
10.2.4 Variances inconnues et différentes . . . 136
10.2.5 Cas de deux échantillons appariés . . . 138
10.3 Variances . . . 139
10.4 Proportions . . . 140
10.5 Tests non paramétriques . . . 141
10.5.1 Test de permutation . . . 141
10.5.2 Test de Wilcoxon-Mann-Whitney . . . 143
10.5.3 Cas de deux échantillons appariés . . . 145
10.6 Preuve de la Proposition 10.1 . . . 148

11 Tests d'adéquation **151**
11.1 Le test du χ^2 . . . 151
11.2 Diagramme quantile–quantile (Q-Q plot) . . . 154
11.2.1 Diagramme . . . 154
11.2.2 Interprétation du diagramme . . . 154
11.3 Shapiro-Wilk . . . 156

11.4 Kolmogorov-Smirnov 157

12 Tests d'indépendance **161**
12.1 χ^2 de contingence 161
12.2 Coefficient de corrélation de Pearson 164
12.3 Coefficient de corrélation de Spearman 165

13 Analyse de la variance **167**
13.1 Le problème 167
13.2 Test du rapport de vraisemblance 167
13.3 Mise en œuvre du test 169
13.3.1 Vérification des hypothèses du modèle 169
13.3.2 Calculs et présentation des résultats 169
13.3.3 Étude de l'effet du facteur : comparaisons multiples 170
13.4 Exemple 170
13.4.1 Vérifications des hypothèses du modèle 170
13.4.2 Tableau d'analyse de la variance 171
13.4.3 Comparaisons multiples 171
13.5 Le test de Kruskal-Wallis 171
13.6 Preuve de la Proposition 13.1 172

Références bibliographiques **175**

Index **177**

Chapitre 1

Introduction

1.1 Concepts fondamentaux

On peut définir la Statistique comme l'activité qui consiste dans le recueil, le traitement et l'interprétation de données d'observation. Ces observations portent, de manière générale, sur des *individus*, définis comme les éléments d'une certaine *population*. C'est la population, en tant qu'ensemble d'entités, qui est l'objet de l'investigation statistique, et non telle ou telle entité particulière.

Dans certains cas, la population de référence est finie et ses éléments peuvent être explicitement dénombrés (par exemple, l'ensemble des étudiants inscrits à l'UTC au semestre d'automne 2000). Mais la notion de population revêt parfois une signification plus abstraite et moins bien définie. Par exemple, dans une étude statistique sur la mise au point d'un vaccin contre le sida, la population de référence est l'ensemble des malades du sida présents et à venir : on parle alors de population *hypothétique*. Parfois, la notion de population s'identifie avec celle de procédure de génération de données. La description précise de conditions expérimentales (par exemple, mesure de la température au sommet de la tour Saint-Jacques le 1er septembre à 17 heures) définit la population des mesures obtenues dans ces conditions. L'hypothèse fondamentale est ici que l'expérience est répétable indéfiniment.

1.2 Variables, distributions

Chaque individu d'une population est typiquement décrit par un ensemble de caractéristiques appelées *caractères* ou *variables*. Un caractère peut être soit *qualitatif* (par exemple : le sexe, la nationalité, l'état matrimonial d'une personne), soit *quantitatif* ou numérique (taille, poids, etc.).

Dans le premier cas, les valeurs prises par le caractère (appelées *modalités*) ne sont pas de nature numérique et ne pourront donc pas être combinées par des opérations arithmétiques, même si elles sont codées par des nombres (codage d'une variable binaire par les nombres 0 et 1 par exemple). Parmi les variables qualitatives, on distingue les variables *ordinales*, dont les modalités sont intrinsèquement ordonnées (par exemple : le grade pour une population de militaires), et les variables *nominales* pour lesquelles il n'existe pas de structure d'ordre particulière.

Une variable quantitative peut être *discrète* ou *continue*, selon que son domaine de définition est dénombrable ou non. Remarquons qu'en réalité n'importe quelle obser-

vation est toujours réalisée avec une précision finie : il n'y a pas de véritable grandeur continue. La notion de variable continue est une abstraction commode pour modéliser des grandeurs mesurées sur des échelles possédant un très grand nombre de valeurs.

La *distribution* d'un caractère X quantitatif dans une population P peut être décrite par la fonction $F_X : \mathbb{R} \to [0, 1]$ qui à tout réel x associe la proportion dans P d'individus pour lesquels on a $X \leq x$. Cette fonction (appelée *fonction de répartition*) est parfaitement définie et observable dans le cas d'une population finie. On suppose encore son existence dans le cas d'une population hypothétique, même si elle n'est plus observable dans ce cas.

1.3 Modes d'étude d'une population

L'étude exhaustive (dite *par recensement*) d'une population de grande taille est souvent difficile, voire impossible. Une méthode consiste alors à n'étudier qu'un sous-ensemble de la population totale, appelé *échantillon* (c'est la seule méthode envisageable dans le cas d'une population infinie). Le processus de sélection d'un échantillon est appelé *échantillonnage*.

Un ensemble de n valeurs observées $x_1, \ldots, x_n$ d'une variable (scalaire ou vectorielle) sur un échantillon de taille n est appelé une *distribution empirique* (ou également un *échantillon*, par abus de langage). Une distribution empirique peut être considérée comme apportant une information sur la distribution de la variable correspondante dans la population totale, appelée *distribution théorique*. On appelle *inférence statistique* le processus visant à déduire des conclusions générales relatives à la population totale, à partir d'une connaissance partielle relative à un nombre fini de cas particuliers. Remarquons que l'inférence statistique ne conduit jamais à des certitudes, mais à des conclusions possédant un certain degré de vraisemblance, que l'on cherche à quantifier.

1.4 Objectifs d'une étude statistique

L'étude d'une distribution empirique peut être menée dans deux buts différents, bien que souvent complémentaires :

1. synthétiser, résumer, structurer l'information contenue dans les données, à l'aide de tableaux, de graphiques et de résumés numériques ; c'est l'objet de la statistique *descriptive*, ou *exploratoire* ;
2. formuler et valider des hypothèses relatives à la population totale ; cette branche de la statistique, qui fait largement appel à la théorie des probabilités, est appelée statistique *inférentielle*.

Le second chapitre de ce cours sera consacré à la présentation de techniques simples de statistique descriptive, applicables à la description de distributions empiriques unidimensionnelles (une seule variable numérique scalaire). La statistique inférentielle, qui joue un rôle fondamental dans de nombreux domaines scientifiques et industriels, sera étudiée plus longuement dans les autres chapitres.

Chapitre 2

Éléments de Statistique descriptive

Dans ce chapitre, on considère n observations $x_1, \dots, x_n$ d'une variable quantitative (ou, dans certains cas, ordinale) X, relatives à n individus d'une population. Le problème posé consiste à représenter ou à résumer ces n observations sous forme de tableaux, graphiques et indicateurs numériques.

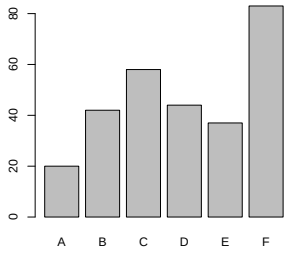
2.1 Tableaux et graphiques

2.1.1 Variable quantitative discrète ou qualitative ordinale

Considérons tout d'abord le cas d'une variable quantitative discrète ou qualitative ordinale, à valeurs ordonnées dans $V_x = \{\xi_1, \dots, \xi_K\}$ avec $\xi_1 < \dots < \xi_K$. Une distribution de n observations de x peut être présentée sous forme d'un *tableau de fréquences* où figurent pour chaque modalité ξ_k de x, le nombre n_k (appelé effectif ou fréquence) d'observations ayant la valeur ξ_k, la fréquence relative $f_k = n_k/n$ correspondante et la fréquence relative cumulée $F_k = \sum_{j=1}^{k} f_j$.

note ξ_k	fréquence n_k	fréquence relative f_k (%)	fréquence cumulée f_k (%)
A	20	7.04	7.04
B	42	14.79	21.83
C	58	20.42	42.25
D	44	15.49	57.75
E	37	13.03	70.77
F	83	29.23	100

(a) Tableau de fréquences



(b) Diagramme en bandes

FIGURE 2.1 – Tableau de fréquence et diagramme en bandes correspondant

Cette information peut être représentée graphiquement sous forme d'un *diagramme à bandes* dans lequel chaque modalité est associée à une bande de longueur proportionnelle à la fréquence de cette modalité dans l'échantillon. Remarquons que plusieurs modalités sont parfois regroupées pour obtenir une description plus concise

des données (au prix d'une perte d'information).

Exemple 2.1. *Le tableau 2.1a de la figure 2.1 résume les notes (sur l'échelle ECTS à 6 valeurs littérales) obtenues en SY02 au cours du semestre de printemps 2016. Le diagramme en bandes correspondant est représenté sur la figure 2.1b.*

Les variables qualitatives sont déclarées dans GNU R comme facteurs. Pour un vecteur de facteurs `data` représentant les données, on utilisera les fonctions `table` et `barplot` pour obtenir le tableau de fréquences et le diagramme à bandes.

```
donnees <- read.csv('donnees.csv')
data <- donnees[['colA']]
table(data)
barplot(table(data))
```

La fonction générale `summary` peut être également utilisée pour obtenir le tableau de fréquences.

Les variables qualitatives doivent être déclarées dans `pandas` avec le type `'category'`. Le tableau de fréquence peut être obtenu avec la méthode `.value_count()` et à partir de celui-ci un barplot peut être obtenu.

```
import pandas as pd

donnees = pd.read_csv('donnees.csv')
data = donnees['colA']
data.value_counts()
data.value_counts().plot.bar()
```

2.1.2 Variable continue

Tableau de fréquences, histogramme

Dans le cas où la variable x est continue, la réalisation d'un tableau de fréquences et sa représentation graphique nécessitent un partitionnement préalable du domaine de définition de la variable en K classes de largeur constante ou variable. Le plus souvent, ce découpage se fait en divisant l'étendue des données (intervalles entre la plus petite valeur et la plus grande) en un certain nombre d'intervalles de même longueur. On regroupe ensuite éventuellement certaines classes d'effectif trop faible (< 5). On obtient ainsi K intervalles $[a_k, a_{k+1}[$, non nécessairement de même longueur.

Remarquons que le choix du nombre de classes résulte d'un compromis entre deux objectifs antagonistes : résumer les données (ce qui nécessite que K ne soit pas trop grand) sans perdre l'information pertinente. La règle empirique de Sturges (qui n'est qu'indicative) consiste à prendre $K = 1 + \frac{10}{3} \log_{10} n$.

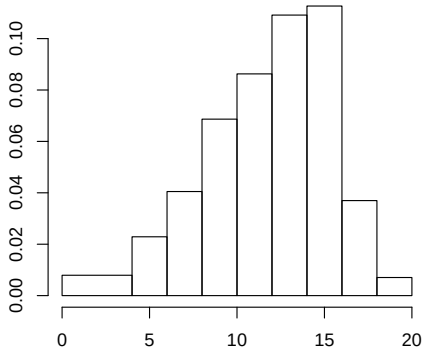
On représente graphiquement le tableau de fréquence sous forme d'un *histogramme*, qui associe à chaque classe k un rectangle dont la base est délimitée par les limites de la classe et dont l'aire est proportionnelle à l'effectif n_k correspondant. Il convient d'insister sur le fait que c'est l'*aire* et non la *hauteur* de chaque rectangle qui doit

être proportionnelle à l'effectif, faute de quoi un regroupement de plusieurs classes se traduirait par un changement radical de la forme de l'histogramme.

Exemple 2.2. *Le tableau de fréquences de la figure 2.2a décrit la distribution des notes générales (échelle de 0 à 20) obtenues en SY02 au semestre de printemps 2016. L'histogramme correspondant est représenté sur la figure 2.2b.*

$[a_k, a_{k+1}[$	n_k	f_k (%)	F_k (%)
[0; 4]	9	3.17	3.17
[4; 6]	13	4.58	7.75
[6; 8]	23	8.1	15.85
[8; 10]	39	13.73	29.58
[10; 12]	49	17.25	46.83
[12; 14]	62	21.83	68.66
[14; 16]	64	22.54	91.2
[16; 18]	21	7.39	98.59
[18; 20]	4	1.41	100

(a) Tableau de fréquences



(b) Histogramme

FIGURE 2.2 – Notes générales (échelle de 0 à 20) obtenues en SY02 au semestre de printemps 2016

Un type d'histogramme particulier, inventé par J. W. Tuckey, est le *diagramme en tige et feuilles*. Il s'agit d'un histogramme horizontal dans lequel les « rectangles » sont formés à l'aide des représentations décimales des valeurs observées. La suite de chiffres composant chaque nombre est séparée en deux parties : les chiffres de gauche servent à définir les classes et forment la tige, tandis que les chiffres de droite forment les feuilles. Différentes variantes existent, permettant le regroupement ou la scission de plusieurs classes. Le diagramme en tige et feuilles met clairement en évidence à la fois la « forme » générale de la distribution et la présence d'observations *atypiques* (valeurs 4.07 et, dans une moindre mesure, 4.88 dans l'exemple ci-dessous).

Exemple 2.3. *Données de Cavendish (1798) relatives à 29 mesures de densité du globe terrestre obtenues par une balance à torsion (voir figure 2.3a). Le diagramme en tige et feuilles est représenté sur la figure 2.3b.*

Les histogrammes et les diagrammes en tige et feuilles peuvent être réalisés avec `GNU R` au moyen des fonctions `hist` et `stem`; supposons qu'un échantillon soit stocké dans le vecteur `data` :

```
donnees <- read.csv('donnees.csv')
data <- donnees[['colA']]
hist(data)
stem(data)
```

Les histogrammes peuvent être réalisés avec `pandas` au moyen de la méthode `.plot.hist()`.

5.50	5.57	5.42	5.61	5.53
5.47	4.88	5.62	5.63	4.07
5.29	5.34	5.26	5.44	5.46
5.55	5.34	5.30	5.36	5.79
5.75	5.29	5.10	5.86	5.58
5.27	5.85	5.65	5.39	

(a) Les 29 mesures

```
4.0|7
4.1|
4.2|
4.3|
4.4|
4.5|
4.6|
4.7|
4.8|8
4.9|
5.0|
5.1|0
5.2|6799
5.3|04469
5.4|2467
5.5|03578
5.6|1235
5.7|59
5.8|56
```

(b) Diagramme tige et feuilles.

FIGURE 2.3 – Données de Cavendish

```
import pandas as pd

donnees = pd.read_csv('donnees.csv')
data = donnees['colA']
data.plot.hist()
```

Dans le cas de valeurs stockées dans un vecteur `numpy`, il est possible de le faire directement avec `matplotlib` :

```
import numpy as np
import matplotlib as plt

data = np.array([12.1, 14.1, 11, 10.8, 17.4, 1.4, 14.3, 7.8,
    12.5, 17.2, 19.0, 1.5, 18.1, 14.1, 7.9, 9.1, 7.0, 5.1])
plt.hist(data)
plt.show()
```

2.1.3 Fonction de répartition empirique

La notion de *fonction de répartition empirique* s'applique aux distributions de variables quantitatives continues et discrètes. Elle se définit de la façon suivante :

$$\begin{aligned} \widehat{F} : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto \frac{1}{n}\,\mathrm{card}\{i : x_i \leq x\} = \frac{1}{n}\sum_{i=1}^{n} 1_{[x_i,+\infty[}(x), \end{aligned}$$

où $1_{[x_i,+\infty[}(\cdot)$ représente la fonction indicatrice de l'intervalle $[x_i, +\infty[$. Pour tout réel x, $\widehat{F}(x)$ est donc la proportion d'observations inférieures ou égales à x. La fonction $\widehat{F}$ est une fonction en escalier, continue à droite.

Le graphe de la fonction de répartition empirique est appelé diagramme cumulatif.

La fonction de répartition empirique peut être calculée avec GNU R au moyen de la méthode `ecdf` (*empirical cumulative distribution function*), qui retourne une fonction constante par morceaux qui peut alors être utilisée ou tracée. Par exemple, avec des données quantitatives stockées dans le vecteur `data` :

```
FrepEmpirique <- ecdf(data)
FrepEmpirique(12.1)
plot(FrepEmpirique)
```

La fonction de répartition empirique peut être calculée est manipulée facilement avec `statsmodels`. La fonction retournée est une fonction constante par morceau, et elle peut être tracée ou manipulée. Par exemple, en supposant les données quantitatives stockées dans le vecteur `data` :

```
import numpy as np
import matplotlib as plt
from statsmodels.distributions.empirical_distribution \
    import ECDF

FrepEmpirique = ECDF(data)
print(FrepEmpirique(12.1))

span = data.max()-data.min()
xplotmin = data.min()-.1*span
xplotmax = data.max()+.1*span
xplot = np.linspace(xplotmin, xplotmax, 10000)
plt.plot(xplot, FrepEmpirique(xplot))
plt.show()
```

2.2 Résumés numériques

Une approche, complémentaire de la précédente, pour résumer un ensemble de données consiste à définir un petit nombre d'indicateurs numériques qui fournissent une description concise des données, permettant ainsi d'appréhender rapidement certaines caractéristiques essentielles de la distribution, notamment la *tendance centrale* et la *dispersion*. On définit également des indicateurs permettant de mesurer le degré de liaison entre deux variables, appelés *coefficients de corrélation*.

2.2.1 Indicateurs de tendance centrale

Il s'agit de résumer les données par une valeur « typique » ou « centrale ». L'indicateur de tendance centrale (ou *de position*) le plus courant est la *moyenne empirique*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dans le cas de données catégorisées (réparties en classes), on a

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{K} n_k c_k,$$

où c_k est le centre de la classe k, n_k son effectif et K le nombre de classes.

La moyenne empirique s'interprète géométriquement comme le centre de gravité d'un nuage de points situés sur une droite, aux abscisses x_i, $i = 1, \dots, n$. Elle possède les deux propriétés suivantes.

Proposition 2.1. *La somme des écarts à la moyenne empirique est nulle.*

Preuve. $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = 0.$ □

Proposition 2.2. *La moyenne empirique est la valeur la plus proche des observations, au sens de la somme des carrés des écarts.*

Preuve. Soit a une valeur quelconque. Définissons la distance de a aux observations comme la somme des carrés des écarts :

$$Q(a) = \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n} x_i^2 - 2a \sum_{i=1}^{n} x_i + na^2 = \sum_{i=1}^{n} x_i^2 - 2an\overline{x} + na^2.$$

Cette fonction atteint son minimum lorsque la dérivée s'annule, c'est-à-dire lorsque :

$$Q'(a) = -2n\overline{x} + 2na = 0 \Leftrightarrow a = \overline{x}.$$

□

L'inconvénient principal de la moyenne empirique comme indicateur de tendance centrale est d'être assez sensible à la présence de valeurs « aberrantes ». Deux indicateurs de tendance centrale plus robustes sont la *moyenne tronquée d'ordre k* et la *médiane*. La moyenne tronquée d'ordre k s'obtient en supprimant les k plus petites et les k plus grandes observations :

$$M_k = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)},$$

$x_{(i)}$ désignant l'observation de rang i dans l'échantillon. Une médiane — il peut en exister plusieurs pour un même échantillon — est une valeur qui sépare l'échantillon en deux parties égales. Il s'agit donc de déterminer une valeur M vérifiant $\widehat{F}(M) = 0.5$. La fonction $\widehat{F}$ n'étant pas bijective, elle n'a pas d'inverse ; nous nous baserons sur la définition d'une *fonction inverse généralisée* :

$$\forall \alpha \in]0,1[, \quad \widehat{F}^{-1}(\alpha) = \inf\{x \in \mathbb{R} \mid \widehat{F}(x) \geq \alpha\}.$$

On pose alors $M = \widehat{F}^{-1}(0.5)$. On montre facilement que

$$M = x_{(\lceil n/2 \rceil)},$$

où $\lceil n/2 \rceil$ représente l'arrondi de $n/2$ au plus grand entier supérieur. Notons qu'il existe d'autres conventions de calcul de la médiane. Ainsi, plutôt que d'utiliser la fonction

inverse généralisée définie ci-dessus, on pourrait associer à tout α l'ensemble des valeurs vérifiant $\widehat{F}(M) = \alpha$, puis choisir arbitrairement l'une d'entre elles. La convention adoptée précédemment consiste à choisir l'infimum de cet ensemble de valeurs ; une autre stratégie consisterait par exemple à utiliser le milieu de l'intervalle.

Plus généralement, on peut définir une valeur qui sépare l'échantillon en deux parties de tailles approximativement égales à αn et $(1-\alpha)n$, pour tout $\alpha \in]0,1[$ (comme dans le cas de la médiane, il n'existe généralement pas de valeur dans l'échantillon considéré telle qu'*exactement* αn valeurs de l'échantillon lui sont inférieures et les $(1-\alpha)n$ restantes lui sont supérieures). Une telle valeur est appelée *fractile (ou quantile) empirique d'ordre* α. Comme pour la médiane, il peut exister plusieurs fractiles empiriques de même ordre α pour un échantillon donné : il est nécessaire d'adopter une convention pour le calcul des fractiles. Nous utiliserons ici encore l'inverse généralisée de la fonction de répartition empirique, et définirons $\widehat{f}_\alpha = \widehat{F}^{-1}(\alpha)$, soit :

$$\widehat{f}_\alpha = x_{([n\alpha])}.$$

On a donc, avec ces notations, $M = \widehat{f}_{0.5}$.

Enfin, un indicateur de position souvent utilisé dans le cas d'un caractère discret est le *mode*, défini comme la valeur la plus fréquente dans la série d'observations (cette valeur n'est pas nécessairement unique). Dans le cas d'un caractère continu, cette notion ne s'applique pas directement, mais on peut définir une *classe modale*, lorsque les données ont été préalablement catégorisées, comme une classe pour laquelle l'histogramme présente un maximum.

Ces indicateurs peuvent être calculés dans `GNU R` à l'aide des fonctions `mean`, `median` et `quantile` qui calculent respectivement la moyenne empirique, la médiane empirique et les quantiles empiriques. Les fonctions `median` et `quantile` gèrent différentes conventions de calcul des quantiles.

```
mean(data)
median(data)
quantile(data, 0.4)
```

Pour utiliser la convention choisie dans le poly (qui n'est pas souvent utilisée d'un point de vue pratique), il faut spécifier l'argument `type=1` lors de l'appel.

Ces indicateurs peuvent être calculés avec `numpy` à l'aide des fonctions ou méthodes `mean`, `median` et `quantile`.

```
import numpy as np

data.mean()
np.mean(data)
np.median(data)
np.quantile(data, 0.4)
```

Pour utiliser la convention choisie dans le poly (qui n'est pas souvent utilisée

d'un point de vue pratique), il faut utiliser la méthode `quantile` (même pour calculer la médiane), avec l'argument `interpolation='higher'` lors de l'appel.

2.2.2 Indicateurs de dispersion

Les indicateurs de dispersion traduisent dans quelle mesure les observations ont tendance à s'écarter de leur valeur centrale. Le plus courant est la *variance empirique* :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

On montre facilement que s^2 s'exprime également comme la différence entre la moyenne des carrés des observations et le carré de la moyenne :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2.$$

Dans le cas de données catégorisées, on a, avec les mêmes notations que précédemment :

$$s^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(c_k - \overline{x})^2 = \frac{1}{n}\sum_{k=1}^{K} n_k c_k^2 - \overline{x}^2.$$

La racine carrée de la variance empirique est appelée *écart-type empirique*. La *variance empirique corrigée* s'obtient en divisant la somme des carrés des écarts à la moyenne par $n-1$:

$$s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}s^2.$$

Un autre indicateur de dispersion dont le calcul est très facile est l'*étendue*

$$W = \max_i x_i - \min_i x_i = x_{(n)} - x_{(1)}.$$

Cet indicateur est malheureusement très peu robuste. On lui préférera souvent l'*étendue interquartile* H, définie comme la différence entre les fractiles empiriques d'ordre 0.25 et 0.75 (appelés, respectivement, premier et troisième quartile de la distribution, le deuxième quartile étant la médiane). L'étendue interquartile est donc la longueur d'un intervalle « central » contenant 50 % des données.

Les indicateurs de dispersion peuvent être calculés dans GNU R : `sd` et `var` calculent respectivement l'écart type empirique corrigé et la variance empirique corrigée. La fonction `IQR` (*interquartile range*) calcule l'étendue interquartile.

```
sd(data)
var(data)
IQR(data)
```

Les indicateurs de dispersion peuvent être calculés avec `numpy` et `pandas` avec les méthodes `std()` et `var()` qui calculent respectivement les ecart-types et variances empiriques. La correction se règle avec `ddof`.

— avec `numpy`, par défaut `ddof=0`, les valeurs retournées sont non corrigées,

il faut passer `ddof=1` pour avoir les versions corrigés.
— avec `pandas`, par défaut `ddof=1`, les valeurs retournées sont corrigées, il faut passer `ddof=0` pour avoir les versions corrigés.

Conseil : toujours spécifier `ddof`, quel que soit l'outil qu'on utilise. `ddof=0` pour la version non corrigée et `ddof=1` pour la version corrigée.

```
data.std(ddof=0) # ecart-type empirique non corrigé
data.std(ddof=1) # ecart-type empirique corrigé
data.var(ddof=0) # variance empirique non corrigée
data.var(ddof=1) # variance empirique corrigée
```

Pour le calcul de l'étendue interquartile, il est possible de faire la différence entre les quartiles.

```
q1, q3 = data.quantile((.25,.75))
print(q3-q1)
```

2.2.3 Diagramme en boîte

Il est commode de représenter sur un même graphique plusieurs indicateurs décrivant conjointement la « forme » de la distribution. Un exemple d'une telle représentation est le *diagramme en boîte*, encore appelé *boîte à moustaches* ou *boxplot*. Il s'agit d'un graphique formé d'une boîte délimitée par le premier et le troisième quartile, sur laquelle on fait également figurer la médiane. Les moustaches sont des segments de droite qui s'étendent de part et d'autre de la boîte jusqu'aux points les plus extrêmes situés à une distance inférieure à $1.5H$ des extrémités de la boîte. Les points situés au-delà des extrémités des moustaches sont représentés individuellement.

Une boîte à moustaches décrivant les données de Cavendish est représentée sur la Figure 2.4. Les boîtes à moustaches sont couramment utilisées pour comparer plusieurs distributions. Une telle représentation est adoptée dans la figure 2.5, qui juxtapose les distributions des notes de SY02 obtenues au semestre de printemps 2016 par les étudiants de différentes branches.

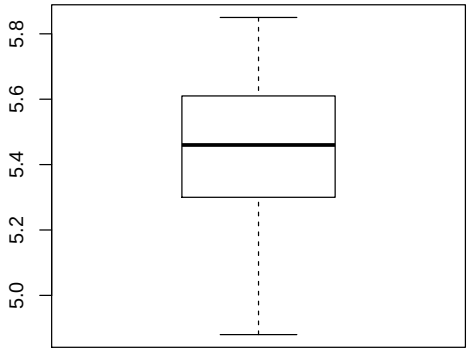


FIGURE 2.4 – Données de Cavendish : boîte à moustaches.

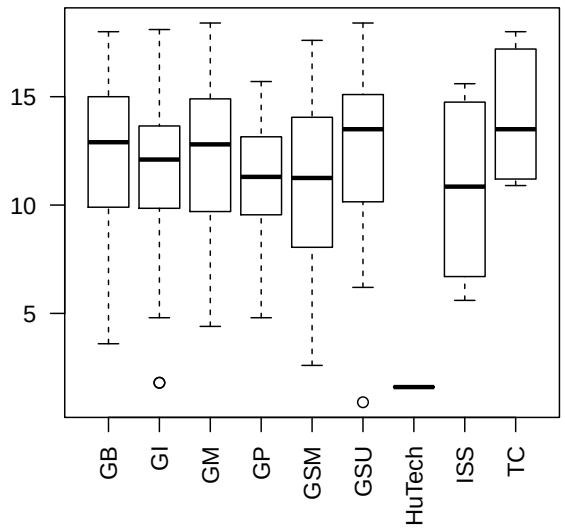


FIGURE 2.5 – Distributions des notes obtenues en SY02 au semestre de printemps 2016, en fonction de la branche.

Les diagrammes en boîte peuvent être facilement tracés dans GNU R à l'aide de la fonction `boxplot`. Pour faire un boxplot à partir d'un vecteur de données on pourra utiliser directement la fonction :

```
boxplot(data)
```

Si l'on dispose d'un tableau, par exemple `sy02` où deux colonnes, nommées `notes` et `branche`, renseignent la note et la branche des étudiants (chaque ligne correspondant à un étudiant), on utilisera :

```
boxplot(notes~branche, data=sy02)
```

Les diagrammes en boîte peuvent être facilement tracés avec `pandas` ou `matplotlib` à l'aide de la fonction `boxplot`. Pour faire un boxplot à partir d'un vecteur de données on pourra utiliser directement la fonction :

```
import matplotlib.pyplot as plt

plt.boxplot(data)
```

Avec `pandas`, si l'on dispose d'un `DataFrame`, par exemple `sy02` où deux colonnes, nommées `notes` et `branche`, renseignent la note et la branche des étudiants (chaque ligne correspondant à un étudiant), on utilisera :

```
sy02.boxplot('notes', 'branche')
```

2.2.4 Coefficients de corrélation

Il est souvent utile de mesurer le degré de liaison entre deux variables quantitatives. Supposons que l'on ait observé les valeurs de deux variables X et Y pour n individus, notées $(x_1, y_1), \ldots, (x_n, y_n)$. On appelle *coefficient de corrélation (de Pearson)* l'indicateur suivant :

$$r = \frac{n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{s_x s_y},$$

où s_x et s_y sont les écarts-types des x_i et des y_i. Il est facile de voir que r est le cosinus de l'angle formé par les vecteurs $\vec{u} = (x_1 - \overline{x}, \ldots, x_n - \overline{x})^T$ et $\vec{v} = (y_1 - \overline{y}, \ldots, y_n - \overline{y})^T$ dans l'espace $\mathbb{R}^n$. On en déduit les propriétés suivantes :

$$-1 \leq r \leq 1$$

et

$$|r| = 1 \Leftrightarrow \exists (a, b, c) \in \mathbb{R}^3, \forall i \in \{1, \ldots, n\}, ax_i + by_i + c = 0.$$

Le coefficient de corrélation de Pearson mesure donc la linéarité de la liaison entre deux variables. Il est égal à $+1$ lorsque les points (x_i, y_i) sont alignés sur une droite de pente positive, et il vaut moins -1 lorsque les points sont alignés sur une droite de coefficient négatif. Le coefficient r se calcule plus facilement par la formule suivante :

$$r = \frac{n^{-1} \sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{s_x s_y}.$$

Pour mesurer le degré de dépendance de deux variables, indépendamment du caractère linéaire ou non de la relation, Charles Spearman a proposé en 1904 de calculer le coefficient de corrélation sur les rangs. On obtient alors un nouvel indice r_S, appelé *coefficient de corrélation de Spearman*, qui mesure le degré de dépendance monotone (et non plus linéaire) entre deux variables. En particulier, l'indice r_S est invariant pour toute transformation monotone des deux variables. Soient r_i le rang (entre 1 et n) de x_i dans la série $x_1, \ldots, x_n$, et s_i le rang de y_i dans la série $y_1, \ldots, y_n$. Les moyennes des r_i et des s_i sont égales à $(n + 1)/2$, et les écarts-types à $(n^2 - 1)/12$. On en déduit donc l'expression de r_S :

$$r_S = \frac{n^{-1} \sum_{i=1}^{n} r_i s_i - (n + 1)^2/4}{(n^2 - 1)/12}.$$

En posant $d_i = r_i - s_i$, on obtient l'expression simplifiée suivante :

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}.$$

Le calcul des coefficients de corrélation peut s'effectuer avec GNU R au moyen de la fonction `cor`. L'argument `method` permet de spécifier quelle méthode est utilisée ; par exemple :

```
cor(x, y, method='pearson')
cor(x, y, method='spearman')
```

Avec `scipy`, le calcul des coefficients de corrélation peut s'effectuer au moyen des fonctions `pearsonr` et `spearmanr`. Ces fonctions retournent deux valeurs, le coeffiscient de corrélation et la p-value associée. *Cf.* chapitre 12 pour l'interprétation de la p-value.

```
from scipy import stats

correlation, pval = stats.pearsonr(x, y)
print("Correlation (Pearson): ', correlation)

correlation, pval = stats.spearmanr(x, y)
print("Correlation (Spearman): ', correlation)
```

Chapitre 3

Rappels de Probabilités

3.1 Introduction

Les méthodes étudiées au chapitre précédent visent à décrire de manière synthétique un ensemble d'observations relatives à n individus d'une population. Très souvent, cependant, ces individus ne représentent pas la totalité de la population, mais un sous-ensemble, appelé échantillon, à partir duquel on cherche à tirer des conclusions relatives à la population entière.

Les conclusions d'une telle étude dépendent évidemment de la façon dont est constitué l'échantillon. Par exemple, une étude statistique sur des habitudes de consommation donnera des résultats différents selon l'âge et le milieu social des personnes sondées. La méthode d'échantillonnage qui, à l'usage, s'est révélée offrir le maximum de garantie d'objectivité et de représentativité des résultats est l'*échantillonnage aléatoire simple*. Cette méthode consiste à choisir *au hasard* des éléments dans une population, de telle sorte que chaque individu ait autant de chance d'être sélectionné [1].

Les expressions « choix au hasard » et « autant de chances » renvoient aux notions d'expérience aléatoire et de probabilités, qui constituent les notions premières de la théorie des Probabilités, dont l'utilisation est fondamentale en statistique inférentielle. Les définitions et résultats essentiels en seront rappelés dans ce chapitre. Le lecteur est renvoyé au polycopié de SY01 ou aux ouvrages référencés à la fin de ce texte pour des présentations plus complètes et mathématiquement plus rigoureuses.

3.2 Expérience aléatoire

On appelle *expérience aléatoire* une expérience qui, répétée plusieurs fois dans des conditions opératoires identiques, produit des résultats qui peuvent être différents. Les exemples classiques sont issus des jeux de hasard : lancer d'un ou plusieurs dés, tirage dans une urne, etc. Mathématiquement, la notion d'expérience aléatoire $\mathscr{E}$ se formalise en définissant :

1. un *ensemble fondamental* Ω définissant l'ensemble des résultats possibles de $\mathscr{E}$, appelés *événements élémentaires* ;

1. Cette définition ne s'applique en toute rigueur qu'à une population finie ; nous admettrons qu'elle peut être étendue au cas d'une population infinie, ou même hypothétique.

2. un ensemble A de parties de Ω, appelées *événements*. Un événement aléatoire correspond à une affirmation qui peut être vraie ou fausse suivant le résultat de l'expérience aléatoire. Dans le cas où Ω est fini ou infini dénombrable, on prend en général $\mathscr{A} = \mathscr{P}(\Omega)$. Dans le cas infini non dénombrable, on se limite à un ensemble de parties de Ω possédant une structure particulière (*tribu* ou *σ-algèbre*), c'est-à-dire ayant les propriétés suivantes :
   — $\Omega \in A$;
   — $A \in A \Rightarrow \overline{A} \in A$;
   — si $(A_i)_{i\in I}$ est une famille dénombrable d'éléments de A, alors $\cup_{i\in I} A_i \in A$.
3. une fonction $\mathbb{P} : \mathscr{A} \to [0,1]$, appelée *mesure* ou *distribution de probabilité*, qui à tout événement A associe un nombre $\mathbb{P}(A)$ appelé *probabilité* de cet événement ; on impose à $\mathbb{P}$ les propriétés suivantes :
   — $\mathbb{P}(\Omega) = 1$ (Ω est appelé *événement certain*) ;
   — Pour toute famille $(A_i)_{i\in I}$ finie ou dénombrable d'événements deux à deux disjoints ($A_i \cap A_j = \emptyset$, $\forall i, j \in I$), $\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$. Cette dernière propriété est appelée propriété *d'additivité*.

La quantité $\mathbb{P}(A)$ s'interprète (dans la théorie classique des Probabilités) comme la limite de la fréquence de réalisation de l'événement A, lorsque le nombre n de répétitions de l'expérience aléatoire tend vers l'infini :

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

n_A étant le nombre de réalisations de l'événement A obtenu au cours de n répétitions de $\mathscr{E}$. La structure $(\Omega, \mathscr{A})$ est appelée *espace probabilisable*, et $(\Omega, \mathscr{A}, \mathbb{P})$ *espace probabilisé*.

3.3 Variable aléatoire

3.3.1 Définition

La notion de *variable aléatoire réelle*, ou simplement variable aléatoire (v.a.), modélise une grandeur numérique dont la valeur est fonction du résultat d'une expérience aléatoire. Formellement, on définit une v.a. réelle X comme une application

$$\begin{array}{rl} X : & \Omega \to \mathbb{R} \\ & \omega \mapsto X(\omega). \end{array}$$

Cette application doit en outre vérifier la propriété de *mesurabilité* suivante,

$$\forall B \in \mathscr{B}(\mathbb{R}), \quad X^{-1}(B) \in \mathscr{A},$$

où $\mathscr{B}(\mathbb{R})$ est la plus petite tribu contenant les intervalles de $\mathbb{R}$, appelée *tribu borélienne* sur $\mathbb{R}$.

Pour éviter de confondre une variable aléatoire, qui est une fonction de Ω dans $\mathbb{R}$, avec la valeur prise par cette variable, qui est un réel, on prendra dans la suite de ce texte la convention habituelle : les variables aléatoires seront notées par une lettre majuscule (X) et les valeurs prises par ces variables aléatoires par la lettre minuscule correspondante (x).

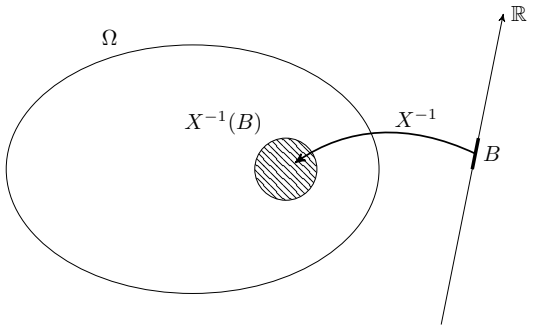


FIGURE 3.1 – Principe de définition de la loi d'une variable aléatoire.

3.3.2 Loi de probabilité d'une v.a.

Pour tout élément B de $\mathscr{B}(\mathbb{R})$, on peut définir la probabilité que la v.a. X prenne sa valeur dans B comme

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)),$$

quantité notée simplement $\mathbb{P}(X \in B)$ (cf. figure 3.1).

L'application $\mathbb{P}_X$ de $\mathscr{B}(\mathbb{R})$ dans $\mathbb{R}$ est une mesure de probabilité sur $\mathbb{R}$, appelée *loi (ou distribution) de probabilité* de X. On dit que la variable X *transporte* la mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathscr{A})$ en une mesure $\mathbb{P}_X$ sur $(\mathbb{R}; \mathscr{B}(\mathbb{R}))$.

Pour décrire complètement $\mathbb{P}_X$, il suffit de donner les probabilités pour des intervalles de la forme $]-\infty, x]$ pour tout $x \in \mathbb{R}$. On appelle *fonction de répartition* de X la fonction

$$\begin{array}{rll} F_X : & \mathbb{R} \to [0,1] \\ & x \mapsto \mathbb{P}_X(]-\infty, x]), \end{array}$$

ce que l'on note $F_X(x) = \mathbb{P}(X \leq x)$. Toute fonction de répartition possède les propriétés suivantes :

1. croissance au sens large ;
2. continuité à droite : $\lim_{\epsilon \to 0, \epsilon > 0} F_X(x + \epsilon) = F_X(x)$;
3. conditions aux limites : $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.

Soit $V_X = X(\Omega)$ l'ensemble des valeurs prises par la v.a. X, appelé *domaine* de X. Si V_X est fini ou dénombrable, la v.a. X est dite *discrète*. Sa loi de probabilité peut alors également être décrite par sa *fonction de masse de probabilité*, définie comme la fonction

$$\begin{array}{rll} p_X : & \mathbb{R} \to [0,1] \\ & x \mapsto \begin{cases} \mathbb{P}_X(\{x\}) & \text{si } x \in V_X \\ 0 & \text{sinon.} \end{cases} \end{array}$$

D'après la propriété d'additivité, on a :

$$\forall B \in \mathscr{B}(\mathbb{R}), \quad \mathbb{P}_X(B) = \sum_{x \in B \cap V_X} p_X(x)$$

et

$$\forall x \in \mathbb{R}, \quad F_X(x) = \sum_{x' \in]-\infty, x] \cap V_X} p_X(x').$$

On a en outre la propriété :

$$\sum_{x \in V_X} p_X(x) = \mathbb{P}(V_X) = 1.$$

Lorsque V_X est infini non dénombrable (typiquement, un intervalle de $\mathbb{R}$), la v.a. X est généralement dite *continue*. Plus précisément, une v.a. continue X est dite *absolument continue* s'il existe une fonction $f_X : \mathbb{R} \to \mathbb{R}_+$, appelée fonction de *densité de probabilité*, telle que

$$\forall B \in \mathscr{B}(\mathbb{R}), \quad \mathbb{P}_X(B) = \int_B f_X(t)dt.$$

On a les propriétés suivantes :

$$\int_{-\infty}^{+\infty} f_X(t)dt = \mathbb{P}_X(\mathbb{R}) = 1 \quad \text{et} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

Réciproquement, $F'_X(x) = f_X(x)$ en tout point x où F_X est dérivable.

Remarque 3.1. *La quantité $f_X(x)$ n'est pas une probabilité; en particulier, on peut avoir $f_X(x) > 1$. C'est l'intégrale de f_X sur un intervalle qui est une probabilité. Notons que l'on a, pour une v.a. continue, $\mathbb{P}_X(\{x\}) = 0$ pour tout $x \in \mathbb{R}$.*

3.3.3 Exemples de lois de probabilité

Loi de Bernoulli

Soit une expérience aléatoire $\mathscr{E}$, et A un événement associé à $\mathscr{E}$, de probabilité p. Soit X la fonction indicatrice de A, définie pour tout $\omega \in \Omega$ par :

$$X(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{sinon.} \end{cases}$$

On a de façon évidente :

$$p_X(1) = \mathbb{P}(A) = p, \quad p_X(0) = \mathbb{P}(\overline{A}) = 1 - p$$

et

$$p_X(x) = 0, \quad \forall x \in \mathbb{R} \setminus \{0, 1\}.$$

Par définition, on dit que X suit une *loi de Bernoulli* de paramètre p, ce que l'on note $X \sim \mathscr{B}(p)$. La fonction de répartition de X est :

$$F_{\mathscr{B}(p)}(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1. \end{cases}$$

Loi binomiale

Supposons que l'on répète n fois l'expérience précédente, de manière indépendante (c'est-à-dire en procédant de telle sorte que la mesure de probabilité associée $\mathbb{P}$ ne change pas d'une fois sur l'autre). On définit ainsi n v.a. indépendantes de Bernoulli $X_i \sim \mathscr{B}(p)$, $i = 1, \ldots, n$. Soit

$$Y = \sum_{i=1}^{n} X_i,$$

qui n'est autre que le nombre de réalisations de A sur les n expériences. Par définition, $Y \sim \mathscr{B}(n, p)$. Le domaine de Y est $V_Y = \{0, \dots, n\}$. Un simple calcul de combinatoire [2] permet de calculer la fonction de masse de probabilité de Y :

$$p_Y(y) = C_n^y p^y (1-p)^{n-y}, \quad \forall y \in V_Y.$$

Une illustration de la fonction de masse est donnée figure 3.2. Par définition, $\mathscr{B}(1, p) = \mathscr{B}(p)$. La loi binomiale est donc une généralisation de la loi de Bernoulli.

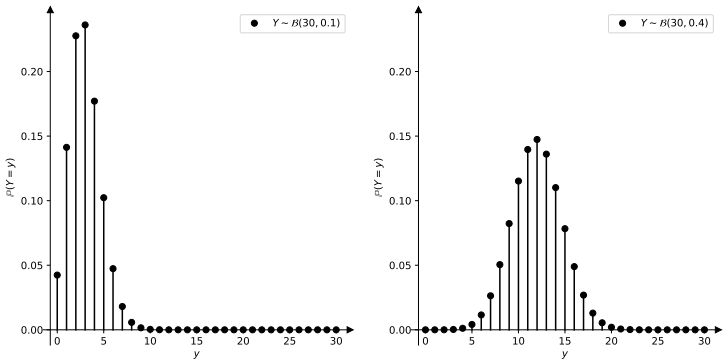


FIGURE 3.2 – Fonction de masse de probabilité de la loi binomiale $\mathscr{B}(n, p)$ pour $n = 30$ et pour deux valeurs de p.

La loi binomiale est manipulable dans GNU R au moyen des fonctions :
- `pbinom`, sa fonction de répartition,
- `dbinom`, sa fonction de masse de probabilité,
- `qbinom`, sa fonction fractile,
- `rbinom`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{B}(30,0.4)}(12)$, c'est à dire la fonction de répartition de la loi $\mathscr{B}(30, 0.4)$ évaluée en 12 :

```
pbinom(12, size=30, prob=0.4)
```

Pour tracer la fonction de masse $k \mapsto \mathbb{P}(X = k)$ pour $X \sim \mathscr{B}(30, 0.4)$:

```
points <- 0:30
plot(points, dbinom(points, size=30, prob=0.4))
```

La loi binomiale est manipulable avec `scipy` au moyen de la classe `binom` du module `stats`. Chaque instance de classe possède en particulier les méthodes :
- `.cdf()`, sa fonction de répartition,
- `.pmf()`, sa fonction de masse de probabilité,
- `.ppf()`, sa fonction fractile,

2. La probabilité d'obtenir y succès et $n - y$ échecs *dans un certain ordre* est $p^y(1-p)^{n-y}$. Par ailleurs, il y a C_n^y façons de choisir y succès parmi n épreuves, d'où le résultat.

— `.rvs()`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{B}(30,0.4)}(12)$, c'est à dire la fonction de répartition de la loi $\mathscr{B}(30, 0.4)$ évaluée en 12 :

```
from scipy import stats

my_law = stats.binom(n=30, p=0.4)
my_law.cdf(12)
```

Pour tracer la fonction de masse $k \mapsto \mathbb{P}(X = k)$ pour $X \sim \mathscr{B}(30, 0.4)$:

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

my_law = stats.binom(n=30, p=0.4)
points = np.arange(0,31)
plt.plot(points, my_law.pmf(points), 'o')
plt.show()
```

Loi de Poisson

La v.a. X suit une loi de Poisson de paramètre λ sur $\mathbb{N}$ si sa fonction de masse s'exprime comme :

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

On notera $X \sim \mathscr{P}(\lambda)$. La fonction de masse de la loi de Poisson est illustrée figure 3.3.

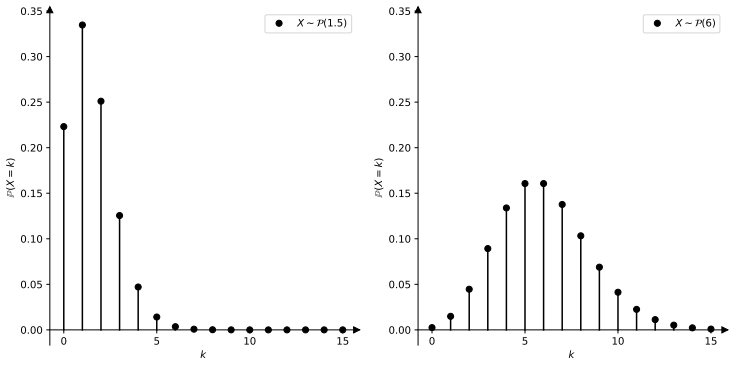


FIGURE 3.3 – Fonction de masse de probabilité de la loi de Poisson $\mathscr{P}(\lambda)$ pour deux valeurs de λ.

Cette loi apparaît comme une limite de loi binomiale :

Théorème 3.1. *Si :*

— $X_n \sim \mathscr{B}(n, p_n)$
— $np_n \to \lambda$
Alors (voir section 3.6.2 pour la convergence en loi) :

$$X_n \xrightarrow{\mathscr{L}} \mathscr{P}(\lambda)$$

Intuitivement, lorsqu'on étudie un phénomène qui correspond à une loi binomiale $\mathscr{B}(n, p)$ avec n très grand et p très faible on voit donc apparaître une loi de poisson $\mathscr{P}(np)$. C'est pour ces raisons qu'elle est souvent utilisée pour modéliser un comptage. Ce phénomène est illustré à la figure 3.4.

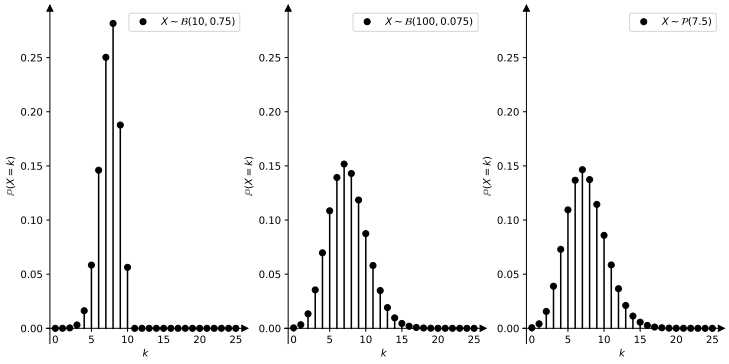


FIGURE 3.4 – Fonctions de masse de probabilité des lois binomiales $\mathscr{B}(10, 0.75)$ et $\mathscr{B}(100, 0.075)$ et de la loi de Poisson $\mathscr{P}(7.5)$. Dans le cas des deux lois binomiales, on a $np = 7.5$, et on remarque que quand n est grand et que p est faible, alors la loi $\mathscr{B}(n, p)$ est approchable par une loi $\mathscr{P}(np)$.

La loi de Poisson est manipulable dans GNU R au moyen des fonctions :
— `ppois`, sa fonction de répartition,
— `dpois`, sa fonction de masse de probabilité,
— `qpois`, sa fonction fractile,
— `rpois`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{P}(15)}(12)$, c'est à dire la fonction de répartition de la loi $\mathscr{P}(15)$ évaluée en 12 :

```
ppois(12, lambda=15)
```

Pour tracer la fonction de masse $k \mapsto \mathbb{P}(X = k)$ pour $X \sim \mathscr{P}(15)$:

```
points <- 0:40
plot(points, dpois(points, lambda=15))
```

La loi de Poisson est manipulable avec `scipy` au moyen de la classe `poisson` du module `stats`. Chaque instance de classe possède en particulier les méthodes :

- `.cdf()`, sa fonction de répartition,
- `.pmf()`, sa fonction de masse de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{P}(15)}(12)$, c'est à dire la fonction de répartition de la loi $\mathscr{P}(15)$ évaluée en 12 :

```
from scipy import stats

my_law = stats.poisson(mu=15)
my_law.cdf(12)
```

Pour tracer la fonction de masse $k \mapsto \mathbb{P}(X = k)$ pour $X \sim \mathscr{P}(15)$:

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

my_law = stats.poisson(mu=15)
points = np.arange(0,41)
plt.plot(points, my_law.pmf(points), 'o')
plt.show()
```

Loi continue uniforme

La v.a. X suit une loi continue uniforme sur l'intervalle $[a, b]$ (ce que l'on note $X \sim \mathscr{U}_{[a,b]}$) si sa fonction de densité est définie par :

$$f_{\mathscr{U}([a,b])}(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x), \forall x \in \mathbb{R},$$

où $1_{[a,b]}$ est la fonction indicatrice de l'intervalle $[a, b]$. La fonction de densité est illustrée figure 3.5.

La fonction de répartition correspondante est :

$$F_{\mathscr{U}([a,b])}(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x \leq b \\ 1 & x > b. \end{cases}$$

La loi uniforme est manipulable dans `GNU R` au moyen des fonctions :

- `punif`, sa fonction de répartition,
- `dunif`, sa fonction de densité de probabilité,
- `qunif`, sa fonction fractile,
- `runif`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{U}([4,9])}(6)$, c'est à dire la fonction de répartition de la loi $\mathscr{U}([4, 9])$ évaluée en 6 :

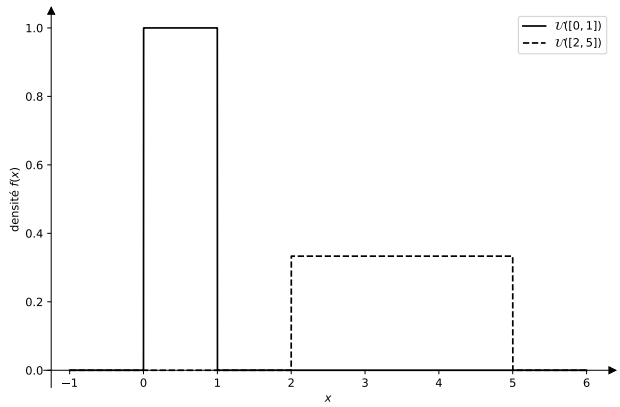


FIGURE 3.5 – Fonction de densité de probabilité de la loi uniforme $\mathcal{U}([a,b])$ pour deux intervalles.

```
punif(6, min=4, max=9)
```

Pour tracer la fonction de densité $f_{\mathcal{U}([4,9])}$:

```
curve(dunif(x, min=4, max=9), from=0, to=20, n=10**4)
```

La loi uniforme est manipulable avec `scipy` au moyen de la classe `uniform` du module `stats`. Chaque instance de classe possède en particulier les méthodes :

- `.cdf()`, sa fonction de répartition,
- `.pdf()`, sa fonction de densité de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Attention, dans `scipy`, la loi uniforme se parametrise avec le premier point de l'intervalle noté `loc`, et avec la longueur de l'intervalle, notée `scale`. Donc pour une loi $\mathcal{U}([a,b])$, on a `loc` $= a$ et `scale` $= b - a$.

Par exemple, pour obtenir $F_{\mathcal{U}([4,9])}(6)$, c'est à dire la fonction de répartition de la loi $\mathcal{U}([4,9])$ évaluée en 6 :

```
from scipy import stats

my_law = stats.uniform(loc=4, scale=5)
my_law.cdf(6)
```

Pour tracer la fonction de densité $f_{\mathcal{U}([4,9])}$:

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

my_law = stats.uniform(loc=4, scale=5)
points = np.linspace(0, 20, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

Loi normale

La loi normale $\mathcal{N}(\mu, \sigma^2)$ (encore appelée loi de Gauss) est définie par la fonction de densité de probabilité :

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

dont le graphe a l'allure d'une « courbe en cloche » centrée sur μ (figure 3.6). Lorsque $\mu = 0$ et $\sigma^2 = 1$, on dit que la loi normale est centrée réduite et on la note $\varphi = f_{\mathcal{N}(0,1)}$.

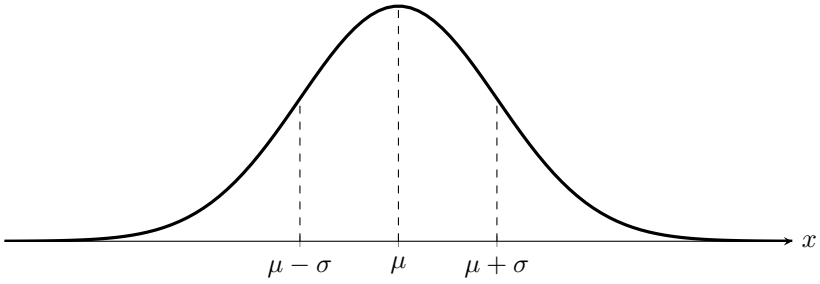


FIGURE 3.6 – Fonction de densité de probabilité de la loi normale $\mathcal{N}(\mu, \sigma^2)$.

La fonction de répartition correspondante n'a pas d'expression analytique. On l'exprime communément à l'aide de la fonction Φ suivante :

$$F_{\mathcal{N}(0,1)} = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt,$$

qui est la fonction de répartition de la loi $\mathcal{N}(0, 1)$. On a en effet la propriété suivante :

Proposition 3.1.

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \forall x \in \mathbb{R}.$$

Cette proposition peut être reformulée en :

$$F_{\mathcal{N}(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \ \forall x \in \mathbb{R}$$

Preuve. Il suffit de faire le changement de variable $y = (t - \mu)/\sigma$ dans l'intégrale définissant $F_{\mathcal{N}(\mu,\sigma^2)}$:

$$F_{\mathcal{N}(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt.$$

Par suite de la parité de la fonction de densité de la loi $\mathcal{N}(0,1)$, la fonction Φ vérifie par ailleurs la propriété suivante : $\Phi(-x) = 1 - \Phi(x)$ pour tout $x \in \mathbb{R}$. □

La loi normale est manipulable dans `GNU R` au moyen des fonctions :
- `pnorm`, sa fonction de répartition,
- `dnorm`, sa fonction de densité de probabilité,
- `qnorm`, sa fonction fractile,
- `rnorm`, un générateur aléatoire suivant cette loi.

Attention, dans `GNU R`, la loi uniforme se paramétrise avec la moyenne notée `mean`, et avec l'ecart-type, noté `sd`. Donc pour une loi $\mathcal{N}(a,b)$, on a `mean` $= a$ et `sd` $= \sqrt{b}$.

Par exemple, pour obtenir $F_{\mathcal{N}(1,3)}(6)$, c'est à dire la fonction de répartition de la loi $\mathcal{N}(1,3)$ évaluée en 6 :

```
pnorm(6, mean=1, sd=sqrt(3))
```

Pour tracer la fonction de densité $f_{\mathcal{N}(1,3)}$:

```
curve(dnorm(x, mean=1, sd=sqrt(3)), from=-6, to=8, n=10**4)
```

Par défaut, la moyenne vaut 0 et l'écart-type vaut 1, donc les fonctions *phi* et *Phi* sont directement accessibles avec `dnorm` et `pnorm` sans préciser les arguments `mean` et `sd`.

La loi de normale est manipulable avec `scipy` au moyen de la classe `norm` du module `stats`. Chaque instance de classe possède en particulier les méthodes :
- `.cdf()`, sa fonction de répartition,
- `.pdf()`, sa fonction de densité de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Attention, dans `scipy`, la loi uniforme se paramétrise avec la moyenne notée `loc`, et l'écart-type noté `scale`. Donc pour une loi $\mathcal{N}(a,b)$, on a `loc` $= a$ et `scale` $= \sqrt{b}$.

Par exemple, pour obtenir $F_{\mathcal{N}(1,3)}(6)$, c'est à dire la fonction de répartition de la loi $\mathcal{N}(1,3)$ évaluée en 6 :

```
from scipy import stats
import numpy as np

my_law = stats.norm(loc=1, scale=np.sqrt(3))
my_law.cdf(6)
```

Pour tracer la fonction de densité $f_{\mathcal{N}(1,3)}$:

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

my_law = stats.norm(loc=1, scale=np.sqrt(3))
points = np.linspace(-6, 8, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

Par défaut, la moyenne vaut 0 et l'écart-type vaut 1, donc les fonctions *phi* et *Phi* sont directement accessibles avec `stats.norm.pdf` et `stats.norm.cdf`.

Loi exponentielle

La v.a. X suit une loi exponentielle de paramètre d'intensité θ (ce que l'on note $X \sim \mathscr{E}(\theta)$) sa fonction de densité est définie par :

$$f_{\mathscr{E}(\theta)}(x) = \theta \exp(-\theta x) \mathbb{1}_{\mathbb{R}_+}(x), \forall x \in \mathbb{R},$$

où $\mathbb{1}_{\mathbb{R}_+}$ est la fonction indicatrice de $\mathbb{R}_+$. La fonction de densité est illustrée figure 3.7.

La fonction de répartition correspondante est :

$$F_{\mathscr{E}(\theta)}(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & x \leq 0 \\ 1 - \exp(-\theta x) & x > 0 \end{cases}$$

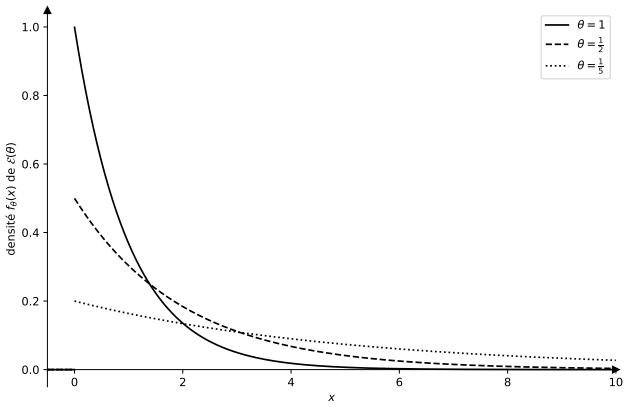


FIGURE 3.7 – Fonction de densité de probabilité de la loi exponentielle $\mathscr{E}(\theta)$ pour différentes valeurs de θ.

Cette loi est souvent utilisée pour représenter des durées de vies de phénomènes sans vieillissement. On peut montrer que la probabilité de survivre jusqu'à $x_0 + x_1$ sachant qu'on a déjà survécu jusqu'à x_0 est la même que la probabilité de survivre jusqu'à x_1 : $\mathbb{P}(X > x_1) = \mathbb{P}(X > x_0 + x_1 | X > x_0)$. Cette identité est caractéristique de durée de vie sans phénomène de vieillissement.

Dans ce document, et dans le cadre de ce cours, la paramétrisation en intensité est la seule utilisée. Il existe également une paramétrisation faisant intervenir un paramètre d'échelle $\tau = 1/\theta$. Dans ce cas, la densité s'écrit comme $t \mapsto \frac{1}{\tau} \exp(-t/\tau) \mathbb{1}_{\mathbb{R}_+}(t)$. Lors de l'utilisation d'outils informatique (GNU R, Python…), il conviendra de vérifier si un paramètre d'intensité (*rate* en anglais) ou un paramètre d'échelle (*scale* en anglais) est utilisé.

La loi exponentielle est manipulable dans GNU R au moyen des fonctions :
- `pexp`, sa fonction de répartition,
- `dexp`, sa fonction de densité de probabilité,
- `qexp`, sa fonction fractile,
- `rexp`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathscr{E}(3)}(6)$, c'est à dire la fonction de répartition de la loi $\mathscr{E}(3)$ évaluée en 6 :

```
pexp(6, rate=3)
```

Pour tracer la fonction de densité $f_{\mathscr{E}(3)}$:

```
curve(dexp(x, rate=3), from=0, to=2, n=10**4)
```

La loi exponentielle est manipulable avec `scipy` au moyen de la classe `expon` du module `stats`. Chaque instance de classe possède en particulier les méthodes :
- `.cdf()`, sa fonction de répartition,
- `.pdf()`, sa fonction de densité de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Attention, dans `scipy`, la loi exponentielle se paramétrise avec un paramètre d'échelle noté `scale` qui vaut l'inverse du paramètre d'intensité. Donc pour une loi $\mathscr{E}(\theta)$, on a `scale` $= 1/\theta$.

Par exemple, pour obtenir $F_{\mathscr{E}(3)}(6)$, c'est à dire la fonction de répartition de la loi $\mathscr{E}(3)$ évaluée en 6 :

```
from scipy import stats
import numpy as np

my_law = stats.expon(scale=1/3)
my_law.cdf(6)
```

Pour tracer la fonction de densité $f_{\mathscr{E}(3)}$:

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

my_law = stats.expon(scale=1/3)
points = np.linspace(0, 2, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

3.4 Résumés numériques d'une loi de probabilité

3.4.1 Espérance mathématique

L'*espérance mathématique*, que nous appellerons simplement *espérance*, d'une variable aléatoire réelle représente la valeur moyenne prise par cette variable aléatoire.

Définition 3.1. *L'espérance mathématique d'une v.a. X est définie par :*

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in V_X} x p_X(x) & \text{si } X \text{ est une v.a. discrète,} \\ \int_{\mathbb{R}} x f_X(x) dx & \text{si } X \text{ est une v.a. continue} \end{cases}$$

si ces quantités existent. Dans le contraire, X n'a pas d'espérance mathématique.

Exemple 3.1. *Si $X \sim \mathcal{B}(p)$, on a $\mathbb{E}(X) = 0\ p_X(0) + 1\ p_X(1) = p$.*

Exemple 3.2. *Si $X \sim \mathcal{U}_{[a,b]}$, on a $\mathbb{E}(X) = \frac{1}{b-a} \int_a^b x dx = \frac{a+b}{2}$.*

D'une manière plus générale, si le graphe de la fonction de densité est symétrique par rapport à une valeur a (c'est-à-dire si $f_X(a-x) = f_X(a+x)$ pour tout x), alors $\mathbb{E}(X) = a$. Il en résulte que l'espérance de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est égale à μ. L'espérance mathématique vérifie par ailleurs les propriétés suivantes.

Proposition 3.2. $\forall \alpha, \beta \in \mathbb{R},\ \mathbb{E}(\alpha X + \beta) = \alpha \mathbb{E}(X) + \beta$.

Théorème 3.2 (Théorème de transfert)**.** *Soit X une v.a. et φ une fonction $\mathbb{R} \to \mathbb{R}$. On a :*

$$\mathbb{E}(\varphi(X)) = \begin{cases} \int_{\mathbb{R}} \varphi(x) f_X(x) dx & \text{si } X \text{ est une v.a. continue,} \\ \sum_{x \in V_X} \varphi(x) p_X(x) & \text{si } X \text{ est une v.a. discrète.} \end{cases}$$

Ce résultat très important permet de calculer l'espérance d'une v.a. $\varphi(X)$ sans avoir besoin de calculer sa loi. Nous l'utiliserons souvent par la suite.

3.4.2 Variance

C'est une mesure de dispersion de la v.a. autour de son espérance.

Définition 3.2.

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

La racine carrée de la variance est appelée *écart-type* de la v.a. X et notée σ. La variance, étant une espérance, peut ne pas être définie.

On a les propriétés suivantes :

Proposition 3.3. $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Proposition 3.4. $\forall \alpha, \beta \in \mathbb{R}, \mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X)$.

Proposition 3.5. *Inégalité de Bienaymé-Tchebycheff :*

$$\forall \varepsilon > 0, \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \mathrm{Var}(X).$$

Cette dernière propriété montre que plus la variance est petite, plus faible est la probabilité que X s'écarte de son espérance d'une valeur donnée.

Exemple 3.3. *Si $X \sim \mathscr{B}(p)$, on a $\mathbb{E}(X^2) = 0^2 \cdot p_X(0) + 1^2 \cdot p_X(1) = p$, d'où* $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p - p^2 = p(1-p)$.

Exemple 3.4. *Loi continue uniforme : on a*

$$\mathbb{E}(X^2) = \frac{1}{b-a} \int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)},$$

d'où

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{b^3 - a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Exemple 3.5. *Pour une variable normale $X \sim \mathscr{N}(\mu, \sigma^2)$, on a $\mathrm{Var}(X) = \sigma^2$ (admis).*

3.4.3 Généralisation : moments d'ordre k

Les notions d'espérance mathématique et de variance se rattachent aux notions plus générales de *moments d'ordre k*. On appelle *moment (non centré) d'ordre k* ($k \in \mathbb{N}^*$) la quantité suivante, si elle existe :

$$m_k = \mathbb{E}(X^k).$$

L'espérance est donc le moment d'ordre 1. De la même façon, on appelle *moment centré d'ordre k* la quantité :

$$\mu_k = \mathbb{E}\left[(X - \mathbb{E}(X))^k\right].$$

La variance est donc le moment centré d'ordre 2. On montre qu'une v.a. qui n'admet pas de moment à l'ordre k n'admet pas non plus de moment d'ordre supérieur.

Les moments (centrés ou non) d'une loi de probabilité jusqu'à un certain ordre constituent une information partielle sur la loi, information d'autant plus riche que le nombre de moments est élevé. On peut définir à partir de certains moments des coefficients qui mettent en évidence certains aspects de la distribution. Par exemple, on définit les *coefficients de Fisher* γ_1 et γ_2 par :

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}, \qquad \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3.$$

Le premier de ces coefficients caractérise la symétrie de la distribution, le second son aplatissement (c'est-à-dire essentiellement la vitesse à laquelle la fonction de probabilité ou de densité de probabilité tend vers 0 en $-\infty$ et $+\infty$). On a, pour la loi normale, $\gamma_1 = \gamma_2 = 0$.

3.4.4 Fractiles (ou quantiles) d'une loi continue

Soit X une v.a. continue de fonction de répartition F_X.

Définition 3.3. *On appelle fractile d'ordre α de X ($\alpha \in]0,1[$) la quantité $f_\alpha = F_X^{-1}(\alpha)$.*

On a donc, par définition, $F_X(f_\alpha) = \alpha$, ou encore $\int_{-\infty}^{f_\alpha} f_X(x)dx = \alpha$.

Exemple 3.6. *Soit $X \sim \mathcal{U}_{[a,b]}$. On a $F_X(f_\alpha) = \alpha \Leftrightarrow \frac{f_\alpha - a}{b-a} = \alpha \Leftrightarrow f_\alpha = a + \alpha(b-a)$.*

Exemple 3.7. *Le fractile d'ordre α de la loi normale centrée-réduite est noté $u_\alpha = \Phi^{-1}(\alpha)$. On peut facilement exprimer le fractile f_α de la loi $\mathcal{N}(\mu, \sigma^2)$ en fonction de u_α. En effet,*

$$\Phi\left(\frac{f_\alpha - \mu}{\sigma}\right) = \alpha \Leftrightarrow \frac{f_\alpha - \mu}{\sigma} = \Phi^{-1}(\alpha) \Leftrightarrow f_\alpha = \mu + \sigma u_\alpha.$$

On a par ailleurs la propriété : $u_{1-\alpha} = -u_\alpha$.

3.5 Vecteurs aléatoires

3.5.1 Définition

La notion de *vecteur aléatoire (réel)*, ou *variable aléatoire vectorielle*, généralise celle de variable aléatoire présentée au paragraphe précédent.

Définition 3.4. *On appelle vecteur aléatoire (réel) un vecteur de $\mathbb{R}^n$ dont les composants sont fonctions du résultat d'une expérience aléatoire $\mathscr{E}$. Si $\mathscr{E}$ est modélisée par un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$, il s'agit donc d'une application mesurable :*

$$\begin{aligned} X \;:\; \Omega &\to \mathbb{R}^n \\ \omega &\mapsto X(\omega) = (X_1(\omega), \dots, X_n(\omega)). \end{aligned}$$

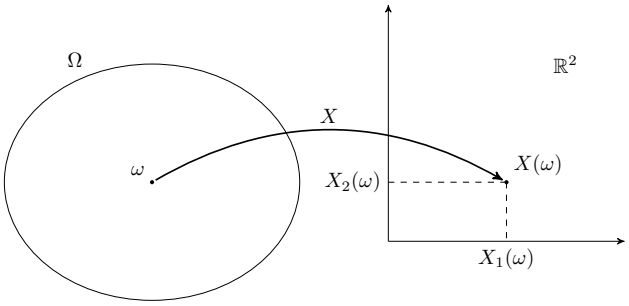


FIGURE 3.8 – Notion de vecteur aléatoire (cas $n = 2$).

3.5.2 Loi de probabilité d'un vecteur aléatoire

La loi de probabilité de X se définit comme une application $\mathbb{P}_X : \mathscr{B}(\mathbb{R}^n) \to [0,1]$ (où $\mathscr{B}(\mathbb{R}^n)$ est la tribu borélienne sur $\mathbb{R}^n$, c'est-à-dire la plus petite tribu contenant les pavés de $\mathbb{R}^n$) telle que :

$$\forall B \in \mathscr{B}(\mathbb{R}^n), \quad \mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)).$$

$\mathbb{P}_X$ est appelée *loi jointe* du vecteur aléatoire X (cf. Figure 3.9). La loi $\mathbb{P}_{X_i}$ de chaque composante X_i est appelée *loi marginale* de X_i. En général, la loi de X ne se déduit pas des lois marginales de ses composants.

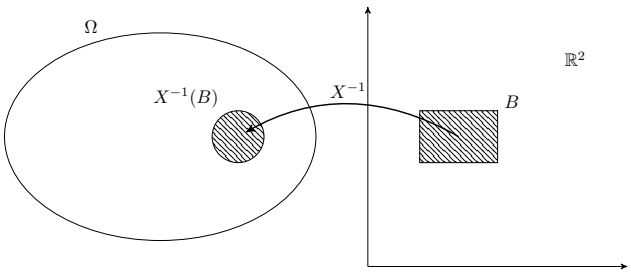


FIGURE 3.9 – Principe de définition de la loi d'un vecteur aléatoire (cas $n = 2$).

Comme dans le cas monodimensionnel, on peut décrire $\mathbb{P}_X$ de deux façons :

1. Par la fonction de répartition de X $F_X : \mathbb{R}^n \to [0,1]$, définie par

$$(x_1, \dots, x_n) \mapsto \mathbb{P}_X(]-\infty, x_1] \times]-\infty, x_2] \times \cdots \times]-\infty, x_n]),$$

   ce que l'on note $F_X(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1; X_2 \leq x_2; \dots; X_n \leq x_n)$.
2. Par la fonction de masse de probabilité (cas discret) ou de densité de probabilité (cas continu) de X.

Définition 3.5. *Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire discret (dont chaque composante est une v.a. discrète). On appelle* fonction de probabilité *de X la fonction*

$$\begin{array}{rccc} p_X : & \mathbb{R}^n & \longrightarrow & [0,1] \\ & (x_1, \dots, x_n) & \longmapsto & \mathbb{P}_X(\{x_1, x_2, \dots, x_n\}), \end{array}$$

ce que l'on note $p_X(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1; X_2 = x_2; \dots; X_n = x_n)$.

Proposition 3.6. *$\forall B \in \mathscr{B}(\mathbb{R}^n)$, on a*

$$\mathbb{P}_X(B) = \sum_{(x_1,\dots,x_n) \in B} p_X(x_1, \dots, x_n).$$

Définition 3.6. *Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire continu (dont chaque composante est une v.a. continue). On appelle* fonction de densité de probabilité de X la fonction $f_X : \mathbb{R}^n \to \mathbb{R}_+$ telle que

$$\forall B \in \mathscr{B}(\mathbb{R}^n), \quad \mathbb{P}_X(B) = \int_B f_X(x_1, \dots, x_n)\, dx_1 \cdots dx_n.$$

Proposition 3.7. *On a donc en particulier*

$$F_X(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f_X(t_1, \dots, t_n) dt_1 \cdots dt_n$$

et, réciproquement,

$$f_X(x_1, \dots, x_n) = \frac{\partial^n F_X}{\partial x_1 \cdots \partial x_n}(x_1, \dots, x_n).$$

Proposition 3.8. *Soit $X = (X_1, X_2)$ un vecteur aléatoire continu de dimension 2. La densité de X_1 s'obtient en intégrant la densité jointe f_X par rapport à x_2 :*

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_X(x_1, x_2) dx_2.$$

Cette propriété se généralise de manière évidente au cas discret (l'intégrale étant remplacée par une somme), et au cas d'un vecteur aléatoire de dimension n quelconque.

3.5.3 Exemples de lois multidimensionnelles

Loi multinomiale

Soient $\mathscr{E}$ une expérience aléatoire modélisée par un espace probabilisé noté $(\Omega, \mathscr{A}, \mathbb{P})$, et $A_1, \dots, A_K$ un *système complet d'événements*, c'est-à-dire une famille d'événements vérifiant $\bigcup_{k=1}^{K} A_k = \Omega$ et $A_i \cap A_j = \emptyset$ $\forall i, j \in \{1, \dots, K\}$. Soit $p_k = \mathbb{P}(A_k)$. On a donc $\sum_{k=1}^{K} p_k = 1$

Supposons que l'on répète n fois l'expérience, et notons N_k le nombre de réalisation de l'événement A_k. Par définition, le vecteur aléatoire $N = (N_1, \dots, N_K)$ suit une *loi multinomiale* de paramètre $(n, p_1, \dots, p_K)$, ce que l'on note

$$N \sim \mathscr{M}(n, p_1, \dots, p_K).$$

On a, pour tout $(n_1, \dots, n_k) \in \{0, \dots, n\}^K$:

$$p_N(n_1, \dots, n_K) = \begin{cases} \dfrac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_K^{n_K} & \text{si } \sum_{k=1}^{K} n_k = n, \\ 0 & \text{sinon.} \end{cases}$$

On a $N_k \sim \mathscr{B}(n, p_k)$, $k = 1, \dots, K$.

Loi normale bidimensionnelle

Soit $X = (X_1, X_2)$ le vecteur aléatoire bidimensionnel de fonction de densité

$$f_X(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right]$$

avec $\sigma_1, \sigma_2 > 0$ et $\rho \in [-1, 1]$.

Par définition, X suit une *loi normale bidimensionnelle*. On montre que $X_1 \sim \mathscr{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathscr{N}(\mu_2, \sigma_2^2)$.

3.5.4 Moments d'un vecteur aléatoire

Espérance

Définition 3.7. *L'espérance du vecteur $X = (X_1, \dots, X_n)$ est le vecteur des espérances de chaque composante :*

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_n)).$$

Espérance d'une fonction $\varphi(X_1, \dots, X_n)$ d'un vecteur aléatoire

Théorème 3.3 (Théorème de transfert – cas vectoriel)**.** *Si φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, et $X = (X_1, \dots, X_n)$ un vecteur aléatoire,*

$$\mathbb{E}(\varphi(X_1 \dots X_n)) = \begin{cases} \displaystyle\int_{\mathbb{R}^n} \varphi(x_1 \dots x_n) f_X(x_1 \dots x_n)\, dx_1 \cdots dx_n & \text{si } X \text{ continu,} \\ \displaystyle\sum_{(x_1 \dots x_n) \in V_X} \varphi(x_1 \dots x_n) p_X(x_1 \dots x_n) & \text{si } X \text{ discret.} \end{cases}$$

Comme pour les variables aléatoires (théorème 3.2), ce résultat très important permet de calculer l'espérance d'une v.a. $\varphi(X_1, \dots, X_n)$ sans avoir besoin de calculer sa loi.

Covariance

Définition 3.8. *Étant données deux variables aléatoires X et Y, on appelle covariance entre X et Y la quantité*

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

En appliquant le résultat du paragraphe précédent, on obtient

$$\mathrm{Cov}(X, Y) = \begin{cases} \displaystyle\int_{\mathbb{R}^2} (x - \mathbb{E}(X))(y - \mathbb{E}(Y)) f_{X,Y}(x, y) dx dy & \text{si } X \text{ est continu,} \\ \displaystyle\sum_{x \in V_X} \sum_{y \in V_Y} (x - \mathbb{E}(X))(y - \mathbb{E}(Y)) p(x, y) & \text{si } X \text{ est discret.} \end{cases}$$

Nous donnons ci-dessous quelques propriétés de la covariance.

Proposition 3.9. $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Preuve.

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

□

Proposition 3.10. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.

Proposition 3.11. $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$.

Proposition 3.12 (Inégalité de Cauchy-Schwarz)**.**

$$[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y),$$

l'égalité n'étant atteinte que ssi $X - \mathbb{E}(X) = k(Y - \mathbb{E}(Y))$.

Définition 3.9. *On appelle* matrice de variance *du vecteur aléatoire $X = (X_1, \dots, X_n)$ la matrice $\mathrm{Var}(X) = \Sigma$ de dimension (n, n) et de terme général $\mathrm{Cov}(X_i, X_j)$.*

Définition 3.10. *On appelle* coefficient de corrélation (théorique) ρ *entre* X *et* Y *la covariance* $\mathrm{Cov}(X,Y)$ *divisée par le produit des écarts-types de* X *et de* Y *:*

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Le coefficient de corrélation $\rho(X,Y)$ entre deux v.a. X et Y est toujours compris entre -1 et $+1$. Il vaut -1 ou $+1$ si et seulement les v.a. X et Y sont liées par une relation linéaire $aX + bY + c = 0$ pour des constantes réelles a, b et c.

3.5.5 Indépendance de variables aléatoires

Définition 3.11. *Les v.a.* $X_1, \dots, X_n$ *sont* indépendantes *si la loi jointe* $\mathbb{P}_X$ *du vecteur aléatoire* $X = (X_1, \dots, X_n)$ *s'exprime comme le produit des lois marginales* $\mathbb{P}_{X_i}$ *c'est-à-dire si et seulement si :*

$$F_X(x_1, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$

ou encore si et seulement si :

$$p_X(x_1, \dots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \qquad \text{(cas discret)}$$

ou

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \text{(cas continu)}$$

Intuitivement, la notion d'indépendance entre v.a. correspond à l'absence de relation entre ces variables.

Exemple 3.8. *Soit* $X = (X_1, X_2)$ *suivant une loi normale bidimensionnelle de paramètres* μ_1,μ_2, σ_1, σ_2 *et* $\rho = 0$. *On a* $f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$. *Les v.a.* X_1 *et* X_2 *sont donc indépendantes.*

Proposition 3.13. *Fonctions de variables aléatoires :*

Si :

- $X_1, \dots, X_n$ *sont indépendantes,*
- $U = g(X_1, \dots, X_k)$,
- $V = h(X_{k+1}, \dots, X_n)$,

Alors U *et* V *sont indépendantes.*

On dira que deux fonctions de variables aléatoires indépendantes qui utilisent des variables ***différentes*** *sont indépendantes.*

Proposition 3.14. *Liens entre indépendance et covariance :*

- X *et* Y *indépendantes* $\Rightarrow \mathrm{Cov}(X,Y) = 0$ *;*
- *La réciproque est fausse ;*
- *Si* (X,Y) *est gaussien alors* X *et* Y *indépendantes* $\Leftrightarrow \mathrm{Cov}(X,Y) = 0$.

3.5.6 Loi conditionnelle

En considérant la loi jointe du vecteur $\mathbb{P}_X$ du vecteur $X = (X_1, X_2)$, les probabilités conditionnelles sont introduites telle que suit :

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_X(x_1, x_2)}{p_{X_2}(x_2)} \qquad \text{(cas discret)}$$

ou

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)} \qquad \text{(cas continu)}$$

Proposition 3.15. *Loi conditionnelle et indépendance : Les variables aléatoires X_1 et X_2 sont indépendantes, si et seulement si pour toute valeur de x_2 alors la loi conditionnelle de X_1 est égale à la loi marginale.*

$$X_1 \text{ indép } X_2 \Leftrightarrow \begin{cases} \forall x_1, x_2 \ p_{X_1|X_2}(x_1|x_2) = p_{X_1}(x_1) & \textit{(cas discret)} \\ \forall x_1, x_2 \ f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1) & \textit{(cas continu)} \end{cases}$$

3.5.7 Somme de variables aléatoires

Soit $(X_1, \dots, X_n)$ un vecteur aléatoire.

Proposition 3.16.

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i).$$

Proposition 3.17. *Si les v.a. $X_1, \dots, X_n$ sont indépendantes, alors*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Exemple 3.9. *Soit $Y \sim \mathscr{B}(n, p)$. On a vu que Y peut s'écrire comme la somme de n v.a. $X_i \sim \mathscr{B}(p)$ indépendantes. On en déduit :* $\mathbb{E}(Y) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = np$ *et* $\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = np(1-p)$.

Proposition 3.18. *Soient $X_1, \dots, X_n$ des v.a. normales indépendantes, avec $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Alors, pour tous réels $\alpha_1, \dots, \alpha_n$, on a :*

$$\sum_{i=1}^{n} \alpha_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2\right).$$

3.6 Notions de convergence stochastique

3.6.1 Convergence en probabilité

Définition 3.12. *La suite (X_n) converge en probabilité vers la constante $a \in \mathbb{R}$ (noté $(X_n) \xrightarrow{\mathbb{P}} a$) si et seulement si*

$$\forall \epsilon > 0 \ \lim_{n\to\infty} \mathbb{P}\left[|X_n - a| > \epsilon\right] = 0;$$

soit, de manière équivalente, si et seulement si

$$\forall \epsilon, \eta > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geq n_0 \quad \mathbb{P}\left[|X_n - a| > \epsilon\right] < \eta.$$

Théorème 3.4. *Si les v.a. X_n $(n = 1, \dots, \infty)$ ont pour espérance a et si leur variance tend vers 0 quand $n \to \infty$, alors $(X_n) \xrightarrow{\mathbb{P}} a$.*

Preuve. On sait tout d'abord que $\forall \epsilon,\ \mathbb{P}\left(|X_i - a| \geq \epsilon\right) \geq 0$

En utilisant l'inégalité de Bienaymé-Tchebycheff on a :

$$\forall \epsilon > 0,\ \mathbb{P}(|X_n - a| \geq \epsilon) \leq \frac{1}{\epsilon^2} \operatorname{Var}(X_n).$$

Sachant que $\lim_{n\to\infty} 0 = 0$ et que $\forall \epsilon > 0,\ \lim_{n\to\infty} \left(\frac{1}{\epsilon^2} \operatorname{Var} X_n\right) = 0$, et en utilisant le théorème d'encadrement des limites (aussi connu sous le nom de théorème des gendarmes) on obtient que

$$\forall \epsilon > 0,\ \lim_{n\to\infty} \mathbb{P}(|X_n - a| \geq \epsilon) = 0$$

□

Remarque 3.2. *Il suffit en fait pour assurer la convergence en probabilité de (X_n) vers a que $\lim_{n\to\infty} \mathbb{E}(X_n) = a$ et $\lim_{n\to\infty} \operatorname{Var}(X_n) = 0$. Cela se démontre en utilisant une forme plus générale de l'inégalité de Bienaymé-Tchebycheff.*

Théorème 3.5. *Si $(X_n) \xrightarrow{\mathbb{P}} a$ et si g est une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$, alors $g(X_n) \xrightarrow{\mathbb{P}} g(a)$.*

3.6.2 Convergence en loi

Définition 3.13. *La suite aléatoire (X_n) converge en loi vers une v.a. X de fonction de répartition F si, en tout point x de continuité de F, on a*

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

où, pour tout n, F_n est la fonction de répartition de X_n. On note $(X_n) \xrightarrow{\mathscr{L}} X$.

Théorème 3.6 (Continuous mapping theorem)**.** *Si $(X_n) \xrightarrow{\mathscr{L}} X$ et si g est une application continue de $\mathbb{R}$ dans $\mathbb{R}$, alors $g(X_n) \xrightarrow{\mathscr{L}} g(X)$.*

Théorème 3.7 (Théorème de Slutsky)**.**

$$\left.\begin{array}{l} X_n \xrightarrow{\mathscr{L}} X \\ Y_n \xrightarrow{\mathbb{P}} a \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} X_n + Y_n \xrightarrow{\mathscr{L}} X + a \\ X_n Y_n \xrightarrow{\mathscr{L}} aX \\ \dfrac{X_n}{Y_n} \xrightarrow{\mathscr{L}} \dfrac{X}{a} \quad \text{si } a \neq 0. \end{array}\right.$$

3.7 Lois dérivées de la loi normale

Certaines lois dérivées de la loi normale jouent un rôle important en statistique. Nous nous contenterons ici de mentionner les principales d'entre elles, en énonçant leurs propriétés les plus importantes.

3.7.1 Loi du χ^2

Définition 3.14. *Soient $U_1, \dots, U_n$ n v.a. indépendantes de loi $\mathcal{N}(0,1)$. Alors la v.a. $T_n = \sum_{i=1}^{n} U_i^2$ suit par définition une loi du χ^2 à n degrés de liberté. On note $T_n \sim \chi^2_n$.*

Proposition 3.19. *La loi χ^2_n a les propriétés suivantes :*

— *Fonction de densité :*

$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} e^{-x/2} x^{n/2-1} 1_{\mathbb{R}^+}(x),$$

où Γ est la fonction Gamma, définie par $\Gamma(r) = \int_0^\infty e^{-x} x^{r-1} dx$. Cette fonction de densité est illustrée figure 3.10.

— *Espérance et variance : $\mathbb{E}(T_n) = n$ et $\mathrm{Var}(T_n) = 2n$*

— *Approximation quand $n \to \infty$:*

$$\frac{T_n - n}{\sqrt{2n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

— *Additivité : d'après la définition,*

$$\left.\begin{array}{l} T_n \sim \chi^2_n \\ T_m \sim \chi^2_m \\ T_n \text{ et } T_m \text{ indépendantes} \end{array}\right\} \Longrightarrow T_n + T_m \sim \chi^2_{n+m}.$$

On notera $\chi^2_{n;\alpha}$ le fractile d'ordre α de la loi χ^2_n.

La loi du χ^2 est manipulable dans `GNU R` au moyen des fonctions :

— `pchisq`, sa fonction de répartition,
— `dchisq`, sa fonction de densité de probabilité,
— `qchisq`, sa fonction fractile,
— `rchisq`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\chi^2_4}(3)$, c'est à dire la fonction de répartition de la loi χ^2_4 évaluée en 3 :

```
pchisq(3, df=4)
```

Pour tracer la fonction de densité $f_{\chi^2_4}$:

```
curve(dchisq(x, df=4), from=0, to=15, n=10**4)
```

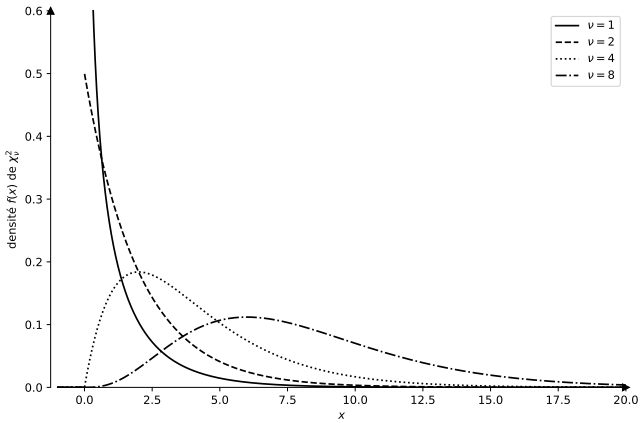


FIGURE 3.10 – Fonction de densité de probabilité de la loi χ^2_ν pour différentes valeurs de ν.

La loi du χ^2 est manipulable avec `scipy` au moyen de la classe `chi2` du module `stats`. Chaque instance de classe possède en particulier les méthodes :

- `.cdf()`, sa fonction de répartition,
- `.pdf()`, sa fonction de densité de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\chi^2_4}(3)$, c'est à dire la fonction de répartition de la loi χ^2_4 évaluée en 3 :

```
from scipy import stats
import numpy as np

my_law = stats.chi2(df=4)
my_law.cdf(3)
```

Pour tracer la fonction de densité $f_{\chi^2_4}$:

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

my_law = stats.chi2(df=4)
points = np.linspace(0, 15, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

3.7.2 Loi de Student

Définition 3.15. *Soient deux v.a. indépendantes* $U \sim \mathcal{N}(0,1)$ *et* $Y \sim \chi^2_\nu$. *La v.a.* $T = U/\sqrt{\frac{Y}{\nu}}$ *suit une loi de Student à* ν *degrés de liberté (d.d.l.), notée* T_ν.

Proposition 3.20. *La loi de Student a les propriétés suivantes :*

— *Fonction de densité :*

$$f_{\mathcal{T}_\nu}(t) = \frac{1}{\sqrt{n}B(\frac{1}{2},\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2},$$

avec

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Cette fonction de densité est paire. Une illustration est donnée de cette fonction de densité figure 3.11.

— *Espérance et variance :* $\mathbb{E}[T_\nu] = 0$ *et* $\mathrm{Var}(T_\nu) = \frac{\nu}{\nu-2} \quad (\nu > 2)$

— $T_\nu \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$ *quand* $\nu \to \infty$. *En pratique, on peut approcher la loi* T_ν *par* $\mathcal{N}(0,1)$ *quand* ν *est grand. Plus les valeurs auxquelles on s'intéresse sont des valeurs extrêmes de la loi (situées en* queue de distribution*), plus* ν *devra être grand pour que l'approximation soit valide. Soit* $t_{\nu,\alpha}$ *le fractile d'ordre* α *de la loi* T_ν. *Par exemple, si l'on tolère une erreur sur la valeur du fractile de 2% au maximum, on aura* $t_{\nu,\alpha} \simeq u_\alpha$ *lorsque*
  — $\nu \geq 50$ *pour tout* $0.05 \leq \alpha \leq 0.95$,
  — $\nu \geq 60$ *pour tout* $0.025 \leq \alpha \leq 0.975$,
  — $\nu \geq 100$ *pour tout* $0.01 \leq \alpha \leq 0.99$.

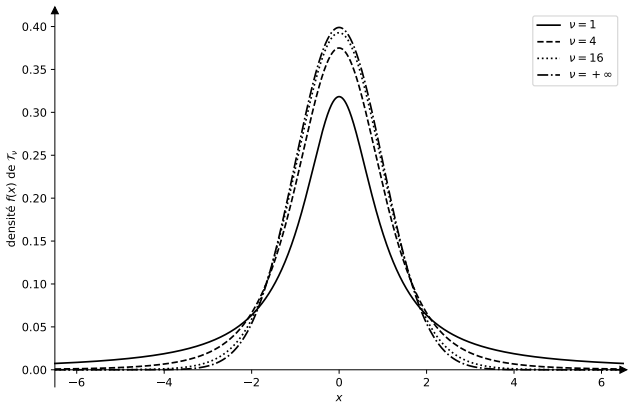


FIGURE 3.11 – Fonction de densité de probabilité de la loi $\mathcal{T}_\nu$) pour différentes valeurs de ν. On peut observer que plus ν est faible, plus les queues de distributions sont lourdes. On peut également observer que pour ν très grand, on se rapproche de la loi normale centrée réduite.

Définition 3.16 (Loi de Student décentrée). *Soient deux v.a. indépendantes* $U \sim \mathcal{N}(\delta, 1)$ *et* $Y \sim \chi^2_\nu$. *La v.a.* $\frac{U}{\sqrt{Y/\nu}}$ *suit une loi de Student décentrée, de paramètre de décentrage* δ, *notée* $\mathcal{T}_\nu(\delta)$.

La loi de Student est manipulable dans `GNU R` au moyen des fonctions :
- `pt`, sa fonction de répartition,
- `dt`, sa fonction de densité de probabilité,
- `qt`, sa fonction fractile,
- `rt`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathcal{T}_4}(3)$, c'est à dire la fonction de répartition de la loi $\mathcal{T}_4$ évaluée en 3 :

```
pt(3, df=4)
```

Pour tracer la fonction de densité $f_{\mathcal{T}_4}$:

```
curve(dt(x, df=4), from=-6, to=6, n=10**4)
```

La loi de Student décentrée est également manipulable au moyen de ces fonctions, en utilisant donnant une valeur non nulle à l'argument `ncp` (qui correspond au δ de la définition précédente.

La loi de Student est manipulable avec `scipy` au moyen de la classe `t` du module `stats`. Chaque instance de classe possède en particulier les méthodes :
- `.cdf()`, sa fonction de répartition,
- `.pdf()`, sa fonction de densité de probabilité,
- `.ppf()`, sa fonction fractile,
- `.rvs()`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathcal{T}_4}(3)$, c'est à dire la fonction de répartition de la loi $\mathcal{T}_4$ évaluée en 3 :

```
from scipy import stats
import numpy as np

my_law = stats.t(df=4)
my_law.cdf(3)
```

Pour tracer la fonction de densité $f_{\mathcal{T}_4}$:

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

my_law = stats.t(df=4)
points = np.linspace(-6, 6, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

La loi de Student décentrée est également manipulable en utilisant la classe `nct` de module `stats` de `scipy`. Elle nécessite en plus l'argument `nc` qui correspond à δ dans la définition.

3.7.3 Loi de Fisher

Définition 3.17. *Soient deux variables indépendantes $Z_1 \sim \chi^2_{\nu_1}$ et $Z_2 \sim \chi^2_{\nu_2}$. La v.a. $\frac{Z_1/\nu_1}{Z_2/\nu_2}$ suit une loi de Fisher à ν_1 et ν_2 degrés de liberté, notée $\mathcal{F}_{\nu_1,\nu_2}$.*

Une illustration de la fonction de densité est donnée figure 3.12 pour différentes valeurs des degrés de libertés.

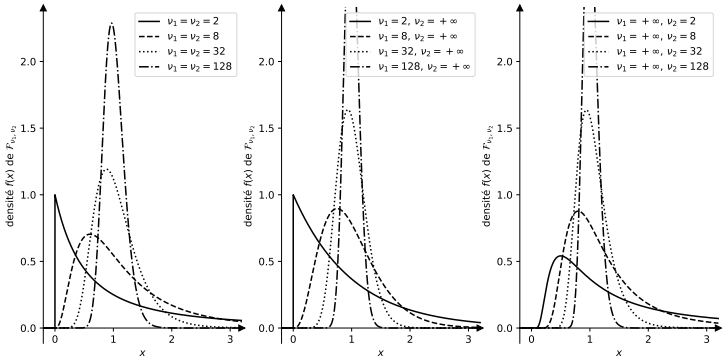


FIGURE 3.12 – Fonction de densité de probabilité de la loi uniforme $\mathcal{F}_{\nu_1,\nu_2}$ pour différentes valeurs de ν_1 et de ν_2. On peut observer que plus les degrés de libertés ν_1 et ν_2 sont élevés, plus la distribution se concentre autour de 1. On observe qu'il faut que les deux degrés de libertés ν_1 et ν_2 soient élevés pour que la distribution se concentre, et que un seul ne suffit pas.

Proposition 3.21. *Soit $f_{\nu_1,\nu_2;\alpha}$ le fractile d'ordre α de la loi de Fisher à ν_1 et ν_2 degrés de liberté. On a :*

$$f_{\nu_2,\nu_1;1-\alpha} = \frac{1}{f_{\nu_1,\nu_2;\alpha}}.$$

Preuve.

$$\begin{aligned}\mathbb{P}(F \leq f_{\nu_1,\nu_2;\alpha}) = \alpha &\Leftrightarrow \mathbb{P}\left(\frac{Z_1/\nu_1}{Z_2/\nu_2} \leq f_{\nu_1,\nu_2;\alpha}\right) = \alpha \\ &\Leftrightarrow \mathbb{P}\left(\frac{Z_2/\nu_2}{Z_1/\nu_1} \geq \frac{1}{f_{\nu_1,\nu_2;\alpha}}\right) = \alpha,\end{aligned}$$

d'où $f_{\nu_2,\nu_1;1-\alpha} = f^{-1}_{\nu_1,\nu_2;\alpha}$. □

La loi de Fisher est manipulable dans `GNU R` au moyen des fonctions :
- `pf`, sa fonction de répartition,
- `df`, sa fonction de densité de probabilité,

— `qf`, sa fonction fractile,
— `rf`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathcal{F}_{5,9}}(2)$, c'est à dire la fonction de répartition de la loi $\mathcal{F}_{5,9}$ évaluée en 2 :

```
pf(2, df1=5, df2=9)
```

Pour tracer la fonction de densité $f_{\mathcal{F}_{5,9}}$:

```
curve(df(x, df1=5, df2=9), from=0, to=5, n=10**4)
```

La loi de Fisher est manipulable avec `scipy` au moyen de la classe `f` du module `stats`. Chaque instance de classe possède en particulier les méthodes :

— `.cdf()`, sa fonction de répartition,
— `.pdf()`, sa fonction de densité de probabilité,
— `.ppf()`, sa fonction fractile,
— `.rvs()`, un générateur aléatoire suivant cette loi.

Par exemple, pour obtenir $F_{\mathcal{F}_{5,9}}(2)$, c'est à dire la fonction de répartition de la loi $\mathcal{F}_{5,9}$ évaluée en 2 :

```
from scipy import stats
import numpy as np

my_law = stats.f(dfn=5, dfd=9)
my_law.cdf(3)
```

Pour tracer la fonction de densité $f_{\mathcal{F}_{5,9}}$:

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

my_law = stats.f(dfn=5, dfd=9)
points = np.linspace(0, 5, 10**4)
plt.plot(points, my_law.pdf(points))
plt.show()
```

Chapitre 4

Échantillonnage

4.1 Notion d'échantillon aléatoire

4.1.1 Définition

Soit $\mathscr{P}$ une population et F_X la fonction de répartition d'un caractère X dans la population (Rappel : $F_X(x)$ est la proportion d'individus dans $\mathscr{P}$ pour lesquels $X \leq x$).

Soit $\mathscr{E}$ l'expérience aléatoire consistant à prélever au hasard un individu de $\mathscr{P}$ avec hypothèse d'équiprobabilité (chaque individu a la même chance d'être choisi). À cette expérience aléatoire correspond un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$ et la variable X devient une variable aléatoire, dont la loi est décrite par la fonction de répartition F_X :

$$\mathbb{P}(X \leq x) = F_X(x).$$

Il y a donc identité entre la distribution des fréquences du caractère X dans la population et la distribution de probabilité de la v.a. X.

Supposons maintenant que l'on répète n fois l'expérience $\mathscr{E}$, de manière indépendante (c'est-à-dire en procédant de telle sorte que la mesure de probabilité associée $\mathbb{P}$ reste constante). On parle alors d'*échantillonnage aléatoire simple*. Ayant effectué n tirages indépendants, on obtient n v.a. indépendantes $X_1, \ldots, X_n$ de même loi que X. Un tel vecteur aléatoire est appelé *échantillon indépendant, identiquement distribué* (iid) de variable parente X. Le nombre n est appelée *taille* de l'échantillon.

Remarque 4.1. *Dans le cas d'une population finie, la procédure d'échantillonnage aléatoire simple postulée ici suppose en toute rigueur que les n tirages d'un individu dans $\mathscr{P}$ se fassent avec remise. Cependant, si $\mathscr{P}$ est de taille importante par rapport à n, les probabilités changent peu d'une fois sur l'autre et l'hypothèse d'indépendance reste valable comme approximation, même si les n tirages se font sans remise (la probabilité pour un individu d'être sélectionné est alors de $1/N$ au premier tirage, $1/(N-1)$ au second et $1/(N-n+1)$ au n^e tirage, N étant le cardinal de $\mathscr{P}$).*

Remarque 4.2. *On prendra soin de distinguer dans les notations l'échantillon aléatoire $X_1, \ldots, X_n$, qui est un vecteur aléatoire et une réalisation de l'échantillon $x_1, \ldots, x_n$, qui est un vecteur de réels.*

4.1.2 Notion de statistique

Soit $t = g(x_1, \dots, x_n)$ un indicateur numérique calculé à partir d'une distribution empirique $x_1, \dots, x_n$ de n observations (par exemple : $t = \overline{x}$, $t = s^2$, $t = \widehat{f_\alpha}$, etc.).

Si les observations $x_1, \dots, x_n$ sont considérées comme des réalisations d'un échantillon $X_1, \dots, X_n$, t est une réalisation d'une v.a. :

$$T = g(X_1, \dots, X_n).$$

Une telle v.a. est appelée une *statistique*. Une statistique apporte une information sur la population de référence. Pour exploiter cette information, il faut connaître certains aspects de la loi de probabilité de T. Dans ce chapitre, nous allons étudier les propriétés de quelques-unes des statistiques les plus courantes associées à un échantillon iid $X_1, \dots, X_n$ de v.a. parente X ayant une espérance μ et une variance σ^2 et poser ainsi les bases des chapitres ultérieurs.

4.2 Moyenne empirique

Rappelons que la *moyenne empirique de l'échantillon* est définie par

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

4.2.1 Propriétés à distance finie (n fixé)

Proposition 4.1.

$$\mathbb{E}(\overline{X}) = \mu \quad \textit{et} \quad \mathrm{Var}(\overline{X}) = \sigma^2/n$$

Preuve. On a $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu$. Donc

$$\mathbb{E}(\overline{X}) = \frac{1}{n}\, \mathbb{E}\left(\sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}\, n\mu = \mu.$$

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \frac{1}{n^2}\, \mathrm{Var}\left(\sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) && \text{(v.a. } X_i \text{ indépendantes)} \\ &= \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

□

Remarque 4.3. *La moyenne empirique de n observations X_i a donc la même espérance (« valeur moyenne ») que la v.a. parente X, mais une plus faible variabilité. Ceci explique, par exemple, que l'on considère une moyenne de plusieurs notes comme plus fiable qu'une seule note pour évaluer un étudiant.*

4.2.2 Propriétés asymptotiques ($n \to \infty$)

Loi faible des grands nombres

Théorème 4.1. *Si $X_1, \dots, X_n$ est une suite de v.a. indépendantes de même loi ayant une espérance μ et une variance σ^2, alors la suite $(\overline{X}_n)$ définie par $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ vérifie :*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Preuve. Résulte directement de la propriété 4.1 et du théorème 3.4. □

Ce résultat signifie que, si on extrait un nombre de plus en plus grand d'individus d'une population (tirage avec remise pour une population finie, tirage avec ou sans remise pour une population infinie), la moyenne de l'échantillon devient « de plus en plus proche » de la moyenne de la population totale.

Un cas particulier important d'un point de vue à la fois pratique et théorique est le Théorème de de Moivre-Laplace. Supposons que l'on effectue n tirages avec remise dans une urne contenant une proportion p de boules blanches. Soit $X_i = 1$ si on a obtenu une boule blanche au i^e tirage, 0 sinon. On a $X_i \sim \mathscr{B}(p)$, $\mathbb{E}(X_i) = p$ et $\mathrm{Var}(X_i) = p(1-p)$. Dans ce cas, la moyenne $\overline{X}_n$ est la proportion de boules blanches parmi les n boules tirées. D'après la loi des grands nombres, on a donc $(\overline{X}_n) \xrightarrow{\mathbb{P}} p$. Ce résultat justifie l'interprétation des probabilités comme limites des fréquences observées, donnée au chapitre précédent (section 3.2).

Théorème Central Limite (TCL)

Théorème 4.2 (admis)**.** *Soit (X_n) une suite de v.a. iid, d'espérance μ et de variance σ^2, et $(\overline{X}_n)$ la suite de terme général $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On a*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$$

ou, de manière équivalente,

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

D'un point de vue pratique, ce théorème a pour conséquence le fait que la loi de $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ peut être approchée par la loi $\mathcal{N}(0,1)$ quand n est suffisamment grand. On a donc asymptotiquement

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

et ce quelle que soit la loi de X, pourvu que $\mathbb{E}(X)$ et $\mathrm{Var}(X)$ existent.

La convergence vers la loi $\mathcal{N}(0,1)$ peut être plus ou moins rapide, mais le théorème fonctionne dans tous les cas tant que $\mathbb{E}(X)$ et $\mathrm{Var}(X)$ existent. Intuitivement, plus la loi de X est proche d'une loi normale, plus la convergence est rapide, en particulier quand la loi de X est une loi symétrique. Par exemple figure 4.1, est représentée la convergence de $\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$ pour X suivant une loi $\mathscr{U}([0,1])$, donc avec $\mu = 1/2$ et

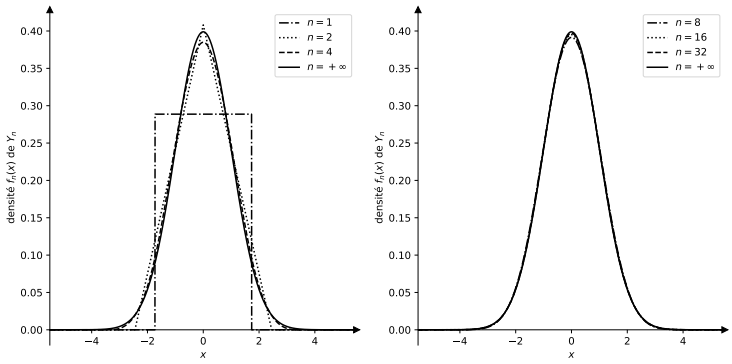


FIGURE 4.1 – Fonction de densité de probabilité de $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$, avec $X \sim \mathcal{U}([0,1])$, $\mu = 1/2$, $\sigma = \sqrt{1/12}$. La loi $\mathcal{U}([0,1])$ étant symétrique, la convergence de la somme renormalisée vers la loi normale est rapide.

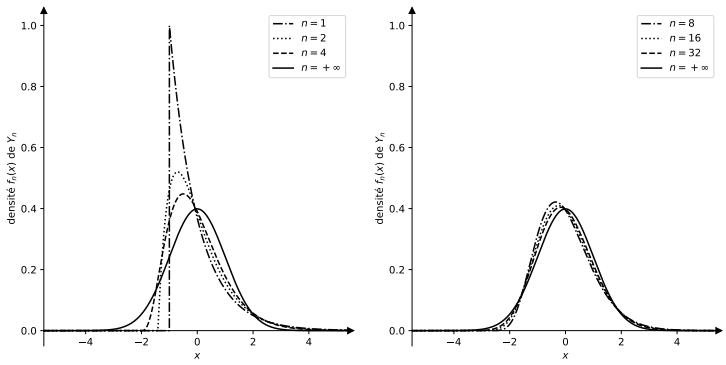


FIGURE 4.2 – Fonction de densité de probabilité de $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$, avec $X \sim \mathcal{E}(1)$, $\mu = 1$, $\sigma = 1$. La loi $\mathcal{E}(1)$ étant asymétrique, la convergence de la somme renormalisée vers la loi normale est moins rapide que dans le cas où la loi de départ est symétrique.

$\sigma = \sqrt{1/12}$. Cette loi est symétrique, et on observe que la convergence vers la loi $\mathcal{N}(0,1)$ est très rapide.

Dans un autre cas, on considère pour X une loi $\mathcal{E}(1)$, donc avec $\mu = 1$ et $\sigma = 1$, et on illustre figure 4.2 la loi de $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$. On observe toujours la convergence vers $\mathcal{N}(0,1)$, mais la loi de $\mathcal{E}(1)$ étant non symétrique, la convergence est plus lente.

Une conséquence immédiate du TCL est la possibilité d'approcher la loi binomiale par la loi normale. En effet, soit $Y \sim \mathcal{B}(n,p)$. On a vu que Y peut s'écrire comme la somme de n v.a. indépendantes suivant une loi $\mathcal{B}(p)$: $Y = \sum_{i=1}^{n} X_i$. On a donc, d'après le TCL,

$$\frac{Y - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Pour n assez grand, on aura donc approximativement

$$\frac{Y - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1),$$

c'est-à-dire $Y \sim \mathcal{N}(np, np(1-p))$. En pratique, on admet que cette approximation est valide quand $np \geq 5$ et $n(1-p) \geq 5$. On a alors

$$\mathbb{P}(Y \leq y) \simeq \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right).$$

On obtient cependant une meilleure approximation par la *correction de continuité* :

$$\mathbb{P}(Y \leq y) \simeq \Phi\left(\frac{y - np + 0.5}{\sqrt{np(1-p)}}\right).$$

Ceci s'explique par le fait que la fonction de probabilité de Y est proche de la fonction de densité de $\mathcal{N}(np, np(1-p))$. Donc $\mathbb{P}(Y = y)$ peut être approché par l'aire de la surface sous la courbe de densité, entre les abscisses $x - 0.5$ et $x + 0.5$:

$$\mathbb{P}(Y = y) \simeq \Phi\left(\frac{y + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{y - 0.5 - np}{\sqrt{np(1-p)}}\right),$$

d'où l'approximation.

4.3 Variance empirique

Rappelons que la *variance empirique de l'échantillon* est définie par

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

4.3.1 Propriétés à distance finie

Proposition 4.2.

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2 \quad \textit{et} \quad \mathrm{Var}(S^2) = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right)$$

μ_4 étant le moment centré d'ordre 4 de X : $\mu_4 = \mathbb{E}[(X-\mu)^4]$.

Preuve. On rappelle que S^2 peut s'exprimer sous la forme : $S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$. On a donc :

$$\mathbb{E}(S^2) = \frac{1}{n}\,\mathbb{E}(\sum_{i=1}^{n} X_i^2) - \mathbb{E}(\overline{X}^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\overline{X}^2) = \mathbb{E}(X^2) - \mathbb{E}(\overline{X}^2).$$

Or, $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, donc $\mathbb{E}(X^2) = \mathrm{Var}(X) + \mathbb{E}(X)^2 = \sigma^2 + \mu^2$. De même, $\mathbb{E}(\overline{X}^2) = \mathrm{Var}(\overline{X}) + \mathbb{E}(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$. On en déduit :

$$\mathbb{E}(S^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

On dit que S^2 est une statistique *biaisée* pour σ^2. Soit la variance empirique corrigée :

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

On a

$$\mathbb{E}(S^{*2}) = \frac{n}{n-1}\mathbb{E}(S^2) = \sigma^2.$$

On dit que S^{*2} est une statistique *sans biais* pour σ^2.

L'expression de la variance de S^2 sera admise. Pour n assez grand, on a l'approximation suivante, en négligeant les termes au-delà du premier ordre en $1/n$:

$$\mathrm{Var}(S^2) \simeq \frac{\mu_4 - \sigma^4}{n}.$$

□

4.3.2 Propriétés asymptotiques

Il existe pour la variance empirique des propriétés analogues à la loi des grands nombres et au TCL.

Proposition 4.3.

$$S^2 \xrightarrow{\mathbb{P}} \sigma^2$$

$$\sqrt{n}\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1) \quad ou \quad S^2 \underset{app.}{\sim} \mathcal{N}(\sigma^2, (\mu_4 - \sigma^4)/n).$$

Preuve. cf. Section 4.9.1. □

Remarque 4.4. *On a les mêmes propriétés asymptotiques pour la variance empirique corrigée S^{*2}.*

4.4 Moments empiriques

La notion de moment empirique généralise celles de moyenne et de variance empiriques.

Définition 4.1. *On appelle moment empirique non centré d'ordre k la statistique :*

$$\widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Définition 4.2. *On appelle moment empirique centré d'ordre k la statistique :*

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

On a donc $\widehat{m}_1 = \overline{X}$ et $\hat{\mu}_2 = S^2$. La propriété suivante généralise à la fois la loi des grands nombres et la première partie de la Proposition 4.3.

Proposition 4.4.

$$\widehat{m}_k \xrightarrow{\mathbb{P}} m_k$$
$$\hat{\mu}_k \xrightarrow{\mathbb{P}} \mu_k.$$

4.5 Cas d'un échantillon gaussien

Dans le cas où X suit une loi normale, on peut déterminer la loi exacte du vecteur aléatoire $(\overline{X}, S^2)$.

Théorème 4.3 (Fisher). *Si $X_1, \dots, X_n$ est un échantillon iid de variable parente X de loi $\mathcal{N}(\mu, \sigma^2)$, alors :*

1. *$\overline{X}$ et S^2 sont indépendants ;*
2. $\overline{X} \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n})$;
3. $\dfrac{nS^2}{\sigma^2} = \dfrac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$.

Éléments de preuve : cf. Section 4.9.2 et Section 3.7.1 pour la définition de la loi du χ^2.

Remarque 4.5. *On en déduit la variance de S^{*2} dans le cas gaussien :*

$$\mathrm{Var}\left[\frac{(n-1)}{\sigma^2}S^{*2}\right] = 2(n-1) \Longrightarrow \frac{(n-1)^2}{\sigma^4}\mathrm{Var}(S^{*2}) = 2(n-1)$$
$$\Longrightarrow \mathrm{Var}(S^{*2}) = \frac{2(n-1)}{(n-1)^2}\sigma^4 = \frac{2\sigma^4}{n-1}.$$

4.6 Fonction de répartition empirique

Rappelons que la *fonction de répartition empirique de l'échantillon* est définie par

$$\widehat{F}(x) = \frac{1}{n}\,\mathrm{card}\{i \in \{1, \dots, n\} \mid X_i \leq x\}, \quad \forall x \in \mathbb{R}\,.$$

Nous précisons maintenant le lien entre cette fonction de répartition empirique et la fonction de répartition théorique de la variable parente X que nous noterons ici F et qui est définie, rappelons-le, par $F(x) = \mathbb{P}(X \leq x)$. Considérant un nombre $x \in \mathbb{R}$ fixé, on a les propriétés suivantes :

Proposition 4.5.

$$\mathbb{E}(\widehat{F}(x)) = F(x)$$

$$\widehat{F}(x) \xrightarrow{\mathbb{P}} F(x)$$

$$\sqrt{n}\frac{\widehat{F}(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1) \quad ou \quad \widehat{F}(x) \underset{app.}{\sim} \mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Preuve. Introduisons les v.a. $Y_1, \dots, Y_n$ définies par :

$$Y_i = \begin{cases} 1 & \text{si } X_i \leq x, \\ 0 & \text{sinon.} \end{cases}$$

Les v.a. $Y_1, \dots, Y_n$ constituent un échantillon iid de variable parente $Y \sim \mathscr{B}(p)$, avec $p = \mathbb{P}(X \leq x) = F(x)$. Par ailleurs, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \widehat{F}(x)$. On peut donc en déduire la première relation :

$$\mathbb{E}(\overline{Y}) = \mathbb{E}(Y) = p = F(x).$$

D'après la loi des grands nombres, on a $\overline{Y} \xrightarrow{\mathbb{P}} \mathbb{E}(Y)$ et on peut en déduire la seconde relation que l'on peut aussi exprimer, par définition, de la manière suivante :

$$\forall x \in \mathbb{R},\ \forall \epsilon, \eta > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geq n_0 \quad \mathbb{P}\left(\left|\widehat{F}_n(x) - F(x)\right| > \epsilon\right) < \eta.$$

La troisième relation est une conséquence du TCL. □

Ce résultat important indique que, pour une réalisation $x_1, \dots, x_n$ d'un échantillon iid $X_1, \dots, X_n$ de v.a. parente X, la fonction de répartition empirique $\widehat{F}$ peut être considérée, pour n assez grand, comme une approximation de la fonction de répartition F de X.

4.7 Fractiles (ou quantiles) empiriques

Rappelons que le fractile empirique d'ordre α d'un échantillon $X_1, \dots, X_n$ est $X_{([n\alpha])}$ avec $\alpha \in]0,1[$ (voir paragraphe 2.2.1), tandis que le fractile théorique d'ordre α d'une v.a. continue X est défini par $f_\alpha = F^{-1}(\alpha)$. Il existe, entre fractiles empiriques et fractiles théoriques de même ordre α, le même type de lien qu'entre la moyenne empirique d'un échantillon et l'espérance mathématique de la variable parente, comme le montre la proposition suivante (que nous ne démontrerons pas ici) :

Proposition 4.6. *Pour tout $x \in \mathbb{R}$, on a*

$$\widehat{f_\alpha} \xrightarrow{\mathbb{P}} f_\alpha$$

$$\sqrt{n}\frac{\widehat{f_\alpha} - f_\alpha}{\sqrt{\frac{\alpha(1-\alpha)}{f_X^2(f_\alpha)}}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1) \quad ou \quad \widehat{f_\alpha} \underset{app.}{\sim} \mathcal{N}\left(f_\alpha, \frac{\alpha(1-\alpha)}{n f_X^2(f_\alpha)}\right).$$

En particulier, pour la médiane :

$$\sqrt{n}(M - f_{0,5}) \xrightarrow{\mathscr{L}} \mathcal{N}\left(0, \frac{1}{4 f_X^2(f_{0.5})}\right).$$

4.8 Échantillonnage stratifié

Dans certains cas, nous avons des informations sur la population, et nous pouvons la diviser en sous-populations. Dans le cas où les sous-populations sont connues, et que nous avons une disparité entre les populations, il peut être judicieux de réaliser un échantillonnage stratifié.

4.8.1 Notation

Soit $\mathscr{P}_1, \dots, \mathscr{P}_p$ les sous-populations d'une population $\mathscr{P}$, dont les tailles respectives sont connues :

$$w_k = \frac{\operatorname{card} \mathscr{P}_k}{\operatorname{card} \mathscr{P}}.$$

4.8.2 Définition

L'échantillonnage stratifié consiste à prendre un échantillon de chaque sous-population, et de combiner les échantillons obtenus.

Un échantillon stratifié s'exprime donc comme un ensemble d'échantillons sur chaque sous-population :

$$\left\{\{X_{1,1}, \dots, X_{1,n_1}\}, \dots, \left\{X_{p,1}, \dots, X_{p,n_p}\right\}\right\}$$

Où $X_{k,i}$ désigne le i^e individu de la sous-population $\mathscr{P}_k$.

Toutes les variables sont supposés indépendantes.

La moyenne empirique sur la sous-population $\mathscr{P}_k$, s'exprime comme :

$$\overline{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{k,i}$$

La moyenne empirique de la population s'exprime comme une moyenne pondérée des sous-populations :

$$\overline{X} = \sum_{k=1}^{p} w_k \overline{X}_k$$

4.8.3 Propriétés

Cas identiquement distribué Lorsque l'on a une loi commune entre les sous-populations, la moyenne empirique converge vers la moyenne théorique :

$$\overline{X} \underset{\forall k, n_k \to +\infty}{\longrightarrow} \mathbb{E}(X)$$

Avec la variance :

$$\operatorname{Var}\left(\overline{X}\right) = \left(\sum_{k=1}^{p} \frac{w_k^2}{n_k}\right) \operatorname{Var}(X)$$

Cas non identiquement distribué Lorsqu'il existe une loi commune au sein de chaque sous-population, mais qu'il n'existe pas de loi commune au sein de la population, en notant X_k la variable parente de chaque population, on a :

$$\overline{X} \underset{\forall k, n_k \to +\infty}{\longrightarrow} \sum_{k=1}^{p} w_k \mathbb{E}(X_k)$$

C'est à dire que l'on a bien convergence vers ce qui est attendu, la moyenne des moyennes.

Avec la variance :

$$\operatorname{Var}\left(\overline{X}\right) = \sum_{k=1}^{p} \frac{w_k^2}{n_k} \operatorname{Var}\left(X_k\right)$$

4.8.4 Échantillonnage stratifié optimal

Pour une taille n de l'échantillon complet, il est possible de choisir n_k de telle manière à minimiser $\operatorname{Var}\left(\overline{X}\right)$.

Cas identiquement distribué La variance minimale est obtenue pour

$$n_k = n w_k$$

Ceci est nommé échantillonnage stratifié proportionnel, puisque la taille de chaque sous-echantillon est proportionnelle à la taille de la population correspondante : $n_k \propto w_k \propto \operatorname{card} \mathscr{P}_k$.

Cas non identiquement distribué La variance minimale est obtenue pour

$$n_k = n \frac{w_k \sqrt{\operatorname{Var} X_k}}{\sum_{l=1}^{p} w_l \sqrt{\operatorname{Var} X_l}}$$

4.8.5 Utilisation pratique

On choisira un échantillonnage stratifié dès que l'on a des informations sur la répartition de la population en sous-populations ; d'autant plus si on présuppose une influence de la sous-population sur le paramètre que l'on veut mesurer.

On fera souvent dans un premier temps un échantillonnage stratifié proportionnel, et dans un second temps, seulement si cela semble nécessaire, en estimant les variances, ou pourra être amené à réaliser un échantillonnage stratifié optimal dans le cas de non-identique distribution.

4.9 Démonstrations

4.9.1 Preuve de la Proposition 4.3

On utilise la seconde condition suffisante de convergence en probabilité :

$$\left.\begin{array}{l} \lim\limits_{n \longrightarrow \infty} \mathbb{E}(S^2) = \sigma^2 \\ \lim\limits_{n \longrightarrow \infty} \operatorname{Var}(S^2) = 0 \end{array}\right\} \Longrightarrow S^2 \xrightarrow{\mathbb{P}} \sigma^2$$

La variance empirique étant invariante par translation (elle ne change pas si l'on ajoute une constante à X), on peut se limiter au cas où $\mathbb{E}(X) = 0$. On a :

$$\sqrt{n}\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} = \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}}\right)$$
$$= \underbrace{\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}}\right)}_{Z_n} - \underbrace{\frac{\sqrt{n}\,\overline{X}^2}{\sqrt{\mu_4 - \sigma^4}}}_{T_n}.$$

D'après le TCL, $\sqrt{n}\,\frac{\overline{X}}{\sigma} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1) \iff \sqrt{n}\,\overline{X} \xrightarrow{\mathscr{L}} \mathcal{N}(0,\sigma^2)$ et d'après la loi des grands nombres $\overline{X} \xrightarrow{\mathbb{P}} 0$. Le théorème de Slutsky 3.7 permet d'en déduire

$$\sqrt{n}\,\overline{X}^2 \xrightarrow{\mathscr{L}} 0 \Longrightarrow T_n \xrightarrow{\mathscr{L}} 0.$$

Considérons maintenant le premier terme Z_n. D'après le TCL appliqué à l'échantillon $X_1^2, \ldots, X_n^2$:

$$\sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mathbb{E}(X^2)}{\sqrt{\mathrm{Var}(X^2)}}\right) \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Or, $\mathbb{E}(X^2) = \mathrm{Var}(X) + \mathbb{E}(X)^2 = \sigma^2$ et $\mathrm{Var}(X^2) = \mathbb{E}(X^4) - \mathbb{E}(X^2)^2 = \mu_4 - \sigma^4$. Donc

$$Z_n \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

En appliquant une nouvelle fois le théorème 3.7, on obtient finalement :

$$\sqrt{n}\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

4.9.2 Éléments de preuve du Théorème 4.3

1. Pour tout i, on a

$$\mathrm{Cov}(\overline{X}, X_i - \overline{X}) = \mathrm{Cov}(\frac{1}{n}\sum_j X_j, X_i - \frac{1}{n}\sum_k X_k)$$
$$= \frac{1}{n}\sum_j \mathrm{Cov}(X_i, X_j) - \frac{1}{n^2}\sum_{j,k}\mathrm{Cov}(X_j, X_k)$$
$$= \frac{1}{n}\mathrm{Cov}(X_i, X_i) - \frac{1}{n^2}\sum_j \mathrm{Cov}(X_j, X_j)$$
$$= \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_j \sigma^2 = 0.$$

Les variables $\overline{X}$ et $(X_i - \overline{X})$ formant un vecteur gaussien, la covariance nulle entraîne l'indépendance. Les variables $\overline{X}$ et $(X_i - \overline{X})^2$ sont aussi indépendantes, ce qui montre que $\overline{X}$ et S^2 sont indépendants.

2. Résulte du fait que toute combinaison linéaire de v.a. normales indépendantes suit une loi normale (Propriété 3.18). Ce qui est une approximation dans le cas général (TCL) est donc vrai rigoureusement dans le cas gaussien.

3. On a

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu)^2 \\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2 + 2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\overline{X}).\end{aligned}$$

Le dernier terme du membre de droite étant nul, on a donc :

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2.$$

En divisant par σ^2, on obtient

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2,$$

c'est-à-dire :

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{nS^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

Par suite de l'indépendance des X_i et du fait que $(X_i-\mu)/\sigma \sim \mathcal{N}(0,1)$, le membre de gauche suit une loi du χ^2 à n degrés de liberté et le second terme du membre de droite suit une loi du χ^2 à 1 degré de liberté. Par ailleurs, les deux termes du membre de droite sont indépendants, par suite de l'indépendance de $\overline{X}$ et S^2, on peut montrer que $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$.

Chapitre 5

Estimation ponctuelle

5.1 Notion d'estimateur

Une manière habituelle de poser un problème en statistique est la suivante. On s'intéresse à un caractère X dont la distribution dans une population $\mathscr{P}$ est inconnue. Si l'on prélève aléatoirement un individu dans $\mathscr{P}$, la loi de probabilité de X est donc inconnue. On suppose cependant que l'on peut formuler un modèle sous la forme d'une famille de lois de probabilité, indicée par un paramètre, scalaire ou vectoriel, $\theta \in \Theta$. On a donc, en convenant de représenter une loi de probabilité par sa fonction de répartition :

$$F_X \in (F_\theta)_{\theta\in\Theta}.$$

Une fois posé le modèle $(F_\theta)_{\theta\in\Theta}$, il reste donc à *estimer* le paramètre θ, à partir de l'information disponible constituée par un échantillon iid de X noté $X_1, \dots, X_n$: c'est le problème de l'*estimation statistique*. Son objectif est de déterminer une statistique, appelée *estimateur* du paramètre θ , dont chaque réalisation, appelée *estimation* de θ, peut être considérée comme une approximation du paramètre θ. On notera souvent $\hat{\theta}$ une telle statistique.

Dans ce chapitre, nous nous limiterons en général au cas d'un paramètre réel $\theta \in \Theta \subseteq \mathbb{R}$.

Remarque 5.1. *Dans cette modélisation, il existe une valeur $\theta_0 \in \Theta$ du paramètre telle que $F_X = F_{\theta_0}$. En toute rigueur, il faudrait donc prendre soin de distinguer les notations θ (une valeur quelconque du paramètre) et θ_0 (la « vraie » valeur du paramètre). Le plus souvent, le contexte permet d'éviter toute confusion et nous ne ferons donc généralement pas cette distinction.*

5.2 Propriétés élémentaires d'un estimateur

5.2.1 Estimateur sans biais ou asymptotiquement sans biais

Soit T un estimateur de θ. On peut toujours écrire $\mathbb{E}(T) = \theta + b(n, \theta)$. La quantité $b(n, \theta)$ est appelée *biais* de T.

Définition 5.1. *Une statistique T est un* estimateur sans biais *de θ si $b(n,\theta) = 0$, c'est-à-dire si*

$$\mathbb{E}(T) = \theta.$$

Définition 5.2. *Une statistique T est un* estimateur asymptotiquement sans biais *de θ si*

$$\lim_{n\to\infty} b(n,\theta) = 0, \quad \textit{c'est-à-dire si} \quad \lim_{n\to\infty} \mathbb{E}(T) = \theta.$$

5.2.2 Estimateur convergent

Définition 5.3. *Une statistique T est un* estimateur convergent *de θ si $T \xrightarrow{\mathbb{P}} \theta$, c'est-à-dire si*

$$\forall \epsilon, \eta > 0,\ \exists n_0,\quad \forall n \geq n_0\ \ \mathbb{P}(|T-\theta| > \epsilon) < \eta,$$

ou encore :

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|T-\theta| > \epsilon) = 0.$$

Proposition 5.1. *Si T est un estimateur sans biais ou asymptotiquement sans biais et si*

$$\lim_{n\to\infty} \mathrm{Var}(T) = 0,$$

alors T est convergent.

Preuve. Cf. Remarque 3.2. □

5.2.3 Estimation de l'espérance et de la variance

Soit μ l'espérance de X. On a vu que $\mathbb{E}(\overline{X}) = \mu$ et $\overline{X} \xrightarrow{\mathbb{P}} \mu$, donc $\overline{X}$ est un estimateur sans biais et convergent de μ.

Soient σ^2 la variance de X. Sachant que $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n}$ et $S^2 \xrightarrow{\mathbb{P}} \sigma^2$, la statistique S^2 est un estimateur asymptotiquement sans biais et convergent de σ^2. Son biais est égal à $b(n,\sigma^2) = -\sigma^2/n$. En outre, sachant que $\mathbb{E}(S^{*2}) = \sigma^2$ et $S^{*2} \xrightarrow{\mathbb{P}} \sigma^2$, la statistique S^{*2} est un estimateur sans biais et convergent de σ^2.

En conclusion, si le paramètre θ est l'espérance $\mathbb{E}(X)$ ou la variance $\mathrm{Var}(X)$, les résultats précédents permettent donc dans un premier temps de proposer comme estimateurs, respectivement, la moyenne empirique $\overline{X}$ et la variance empirique S^2, qui sont des estimateurs convergents de l'espérance et de la variance. Puis, dans un second temps après correction, la moyenne empirique $\overline{X}$ et la variance empirique corrigée S^{*2} sont proposés, ce sont des estimateurs sans biais et convergents de l'espérance et de la variance.

La méthode des moments, qui fait l'objet du paragraphe 5.3, généralise la première étape de cette procédure de construction d'un estimateur.

5.2.4 Précision

Définition 5.4. *On appelle* risque quadratique *d'un estimateur T la fonction*

$$R(T,\cdot)\ :\ \theta \mapsto R(T,\theta) = \mathbb{E}[(T-\theta)^2].$$

$R(T,\theta)$ représente donc l'erreur quadratique moyenne commise lorsqu'on estime θ par T.

Proposition 5.2.

$$R(T,\theta) = \mathrm{Var}(T) + b(n,\theta)^2.$$

Preuve.

$$\begin{aligned} R(T,\theta) &= \mathbb{E}[(T-\theta)^2] \\ &= \mathbb{E}[(T-\mathbb{E}(T)+\mathbb{E}(T)-\theta)^2] \\ &= \mathbb{E}[(T-\mathbb{E}(T))^2] + (\mathbb{E}(T)-\theta)^2 + 2(\mathbb{E}(T)-\theta)\mathbb{E}(T-\mathbb{E}(T)) \\ &= \mathrm{Var}(T) + b(n,\theta)^2 + \underbrace{2(\mathbb{E}(T)-\theta)[\mathbb{E}(T)-\mathbb{E}(T)]}_{0}. \end{aligned}$$

□

Définition 5.5. *On dit qu'un estimateur T_1 est* plus précis *qu'un estimateur T_2 si*

$$R(T_1,\theta) < R(T_2,\theta), \quad \forall\theta \in \Theta.$$

Si T_1 et T_2 sont tous deux sans biais, le plus précis est donc celui qui a la plus petite variance. Mais un estimateur biaisé peut être plus précis qu'un estimateur sans biais.

5.3 Méthode des moments

Cette méthode repose sur l'utilisation des moments (centrés ou non) de la variable X (cf. Section 3.4.3).

Soient p la dimension de θ et $mt_1, \dots, mt_p$ p moments de X, centrés ou non (la notation mt se substitue donc ici aux notations m et μ introduites dans la Section 3.4.3). Comme la loi de X dépend de θ, ces moments dépendent également de θ, ce qu'on peut écrire :

$$(mt_1, \dots, mt_p) = g(\theta).$$

Si la fonction g est inversible, on en déduit

$$\theta = g^{-1}(mt_1, \dots, mt_p).$$

La méthode des moments consiste à remplacer dans cette expression les moments théoriques par les moments empiriques correspondants (cf. Section 4.4) :

$$\hat{\theta}_m = g^{-1}(\widehat{mt}_1, \dots, \widehat{mt}_p).$$

Remarque 5.2. *Le choix des p moments est arbitraire et cette méthode peut donc conduire à plusieurs estimateurs. On utilisera en général les premiers moments, à condition que ceux-ci conduisent à une fonction g inversible.*

Proposition 5.3. *Comme $\widehat{mt}_k \xrightarrow{\mathbb{P}} mt_k$ pour tout k (cf. Proposition 4.4), en supposant la fonction g continue, on en déduit $\hat{\theta}_m \xrightarrow{\mathbb{P}} \theta$ en utilisant le continuous mapping theorem (3.6). La méthode des moments conduit donc à des estimateurs convergents si la fonction liant les moments aux paramètres est continue.*

Proposition 5.4.

$$\frac{\sqrt{n}(\hat{\theta}_m - \theta)}{\sigma_m} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1),$$

où σ_m est une constante dépendant de la loi de X et des moments utilisés.

5.4 Méthode du maximum de vraisemblance

Il s'agit d'une des méthodes les plus utilisées. Elle fournit dans les cas les plus simples les estimateurs classiques (au moins asymptotiquement), mais cette méthode se révèle surtout utile dans les situations plus complexes. En outre, elle conduit à des estimateurs dont on connaît la distribution asymptotique, ce qui sera très pratique pour l'estimation par intervalle de confiance et pour les tests d'hypothèses. Nous allons au préalable définit la notion de fonction de vraisemblance.

5.4.1 Fonction de vraisemblance

Soient X une v.a. dont la loi dépend d'un paramètre θ inconnu : $F_X \in (F_\theta)_{\theta\in\Theta}$, $X_1, \dots, X_n$ un échantillon iid de v.a. parente X et $x_1, \dots, x_n$ une réalisation de $X_1, \dots, X_n$. On note $f(x_1, \dots, x_n; \theta)$ la fonction de densité de $X_1, \dots, X_n$ si X est une v.a. continue et $p(x_1, \dots, x_n; \theta)$ la fonction de probabilité de l'échantillon si X est une v.a. discrète. L'échantillon étant indépendant, on a suivant le cas :

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

ou

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p_X(x_i; \theta).$$

Définition 5.6. *On appelle* fonction de vraisemblance *la fonction :*

$$\begin{aligned} L : \quad & \Theta \to \mathbb{R}_+ \\ & \theta \mapsto L(\theta; x_1, \dots, x_n) = \begin{cases} f(x_1, \dots, x_n; \theta) \text{ si } X \text{ continue,} \\ p(x_1, \dots, x_n; \theta) \text{ si } X \text{ discrète.} \end{cases} \end{aligned}$$

Ayant observé la réalisation $x_1, \dots, x_n$ la fonction de vraisemblance permet de comparer différentes valeurs possibles pour le paramètre θ du point de vue de leur plausibilité, ou *vraisemblance*. Cette fonction joue un rôle fondamental en inférence statistique.

5.4.2 Estimateur du maximum de vraisemblance

Définition 5.7. *On appelle* estimateur du maximum de vraisemblance *(EMV) pour θ toute statistique $\hat{\theta}_{MV} = t(X_1, \dots, X_n)$ telle que :*

$$L(\hat{\theta}_{MV}; X_1, \dots, X_n) \geq L(\theta; X_1, \dots, X_n), \quad \forall \theta \in \Theta.$$

La méthode du maximum de vraisemblance consiste à rechercher le maximum de la fonction de vraisemblance $\theta \mapsto L(\theta; X_1, \dots, X_n)$. Dans la suite de ce chapitre (sauf dans la partie 5.4.7), nous supposerons que le paramètre θ est scalaire.

5.4.3 Équation de vraisemblance

La fonction $\ln L(\theta; x_1, \dots, x_n)$, notée $\ell(\theta; x_1, \dots, x_n)$ et appelée *fonction de log-vraisemblance*, a les mêmes variations que L, et conserve donc notamment le lieu du maximum. De plus elle a souvent une expression plus simple. En pratique, on étudiera donc souvent ℓ plutôt que L. Si le support de X ne dépend pas de θ et si ℓ est deux fois dérivable par rapport à θ, le maximum (ou les maxima) se trouvent en résolvant l'*équation de vraisemblance* :

$$\ell'(\theta; x_1, \dots, x_n) = 0.$$

Toute solution de cette équation vérifiant la condition

$$\ell''(\theta; x_1, \dots, x_n) < 0$$

correspond alors à un maximum de ℓ, donc de L.

5.4.4 Invariance fonctionnelle

Dans certains cas, on ne s'intéresse non pas au paramètre θ lui même, mais à une fonction $u(\theta)$.

Proposition 5.5. *Si $\hat{\theta}$ est un estimateur du maximum de vraisemblance de θ et $u : \mathbb{R} \to \mathbb{R}$ une application, alors $u(\hat{\theta})$ est un estimateur du maximum de vraisemblance de $u(\theta)$.*

5.4.5 Convergence

Théorème 5.1. *Sous certaines conditions, il existe une suite $(\hat{\theta}_n)_{n\geq 1}$ de solutions de l'équation de vraisemblance qui converge en probabilité vers la vraie valeur θ_0 :*

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0.$$

On peut résumer cette proposition en disant que l'estimateur du maximum de vraisemblance est convergent.

5.4.6 Normalité asymptotique

Cette partie traite le cas où θ est un scalaire.
Il est nécessaire au préalable de définir la notion d'information de Fisher.

Définition 5.8. *On appelle* information de Fisher *apportée par un échantillon iid $X_1, \dots, X_n$, relativement à un paramètre θ, la quantité positive ou nulle (si elle existe) :*

$$I_n(\theta) = \mathbb{E}\left[\left(\ell'(\theta; X_1, \dots, X_n)\right)^2\right].$$

La quantité $\ell'(\theta; X_1, \dots, X_n)$ est appelée fonction score.

Théorème 5.2. *Si le support de X ne dépend pas de θ (ensemble des valeurs x de probabilité ou de densité strictement positive), alors, sous certaines conditions de régularité,*

$$\mathbb{E}\left[\ell'(\theta; X_1, \dots, X_n)\right] = 0.$$

et

$$I_n(\theta) = -\mathbb{E}\left[\ell''(\theta; X_1, \dots, X_n)\right].$$

Preuve. cf. Section 5.6.1 □

Proposition 5.6. *Si le support de X ne dépend pas de θ, alors $I_n(\theta) = nI_1(\theta)$.*

Preuve. Cette propriété découle du théorème précédent et de l'additivité de la log-vraisemblance :

$$\ell(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \ell(\theta; x_i).$$

□

Remarque 5.3. *La propriété d'additivité énoncée dans la proposition ci-dessus confirme l'interprétation de $I_n(\theta)$ comme une quantité d'information : la quantité d'information apportée par un échantillon est la somme des quantités d'information apportées par chaque observation. Le théorème présenté ci-après renforcera encore cette interprétation.*

On peut alors établir le résultat suivant.

Théorème 5.3. *Sous certaines conditions de régularité, pour toute suite de solutions de l'équation de vraisemblance $(\hat{\theta}_n)_{n\geq 1}$ t.q. $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, on a*

$$\frac{\hat{\theta}_n - \theta_0}{\sqrt{1/I_n(\theta_0)}} \xrightarrow{\mathscr{L}} \mathcal{N}(0, 1).$$

Preuve. cf. Section 5.6.2 □

En pratique, ce résultat permet d'approcher, pour n assez grand, la loi de $\hat{\theta}$ par $\mathcal{N}(\theta, 1/I_n(\theta))$.

Remarque 5.4. *L'estimateur du maximum de vraisemblance n'est pas nécessairement unique.*

5.4.7 Cas d'un paramètre vectoriel

On a supposé jusqu'à présent que θ était un scalaire, mais la loi de X peut dépendre de plusieurs paramètres scalaires, c'est-à-dire d'un paramètre vectoriel.

Exemple 5.1. *Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, on peut noter $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$.*

Dans ce cas la fonction de vraisemblance L et son logarithme ℓ sont des fonctions de plusieurs variables réelles :

$$\begin{array}{rccc} \ell : & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \theta = (\theta_1, \dots, \theta_p) & \longmapsto & \ell(\theta; x_1, \dots, x_n). \end{array}$$

Lorsque ℓ est deux fois dérivable, son maximum se trouve en résolvant le système :

$$\frac{\partial \ell}{\partial \theta_1} = 0, \dots, \frac{\partial \ell}{\partial \theta_p} = 0,$$

appelé *système des équations de vraisemblance.* Toute solution de ce système est un maximum de ℓ à condition que la matrice $H = \left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right)$ soit définie négative.

5.5 Efficacité

5.5.1 Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR)

Le théorème suivant indique que la variance (et donc la précision) d'un estimateur sans biais de θ (ou plus généralement d'une fonction $u(\theta)$) ne peut, en général, descendre en dessous d'une certaine valeur, qui dépend de la l'information de Fisher.

Théorème 5.4 (Théorème de Fréchet-Darmois-Cramer-Rao)**.** *Soit X une v.a. de loi $F_X \in (F_\theta)_{\theta\in\Theta}$, $\Theta \subseteq \mathbb{R}$, u une application continue et dérivable de $\mathbb{R} \to \mathbb{R}$ et $\hat{u}$ un estimateur sans biais de $u(\theta)$. Si le support ne dépend pas de θ alors, sous certaines conditions de régularité portant sur F_X, on a :*

$$\mathrm{Var}(\hat{u}) \geq \frac{u'(\theta)^2}{I_n(\theta)}.$$

Preuve. cf. Section 5.6.3. □

Remarque 5.5. *La quantité $\dfrac{u'(\theta)^2}{I_n(\theta)}$ est notée $B_F[u(\theta)]$ et appelée* borne de Fréchet relative à $u(\theta)$. *Les conditions d'application du théorème (support de X ne dépendant pas de θ et conditions de régularité) sont appelées* conditions de Cramer-Rao.

Remarque 5.6. *Dans le cas particulier où $u(\theta) = \theta$, on a donc :*

$$B_F(\theta) = \frac{1}{I_n(\theta)}.$$

et $\mathrm{Var}(\hat{\theta}) \geq B_F(\theta)$ pour tout estimateur sans biais $\hat{\theta}$ de θ.

5.5.2 Estimateur efficace

Définition 5.9. *Les conditions de Cramer-Rao étant vérifiées, soit $\hat{u}$ un estimateur sans biais de $u(\theta)$, où u est une fonction dérivable quelconque. $\hat{u}$ est un estimateur* efficace *de $u(\theta)$ si et seulement si :*

$$\mathrm{Var}(\hat{u}) = \frac{u'(\theta)^2}{I_n(\theta)} = B_F[u(\theta)].$$

$\hat{u}$ est un estimateur asymptotiquement efficace *de $u(\theta)$ si et seulement si :*

$$\lim_{n\to\infty} \frac{B_F[u(\theta)]}{\mathrm{Var}(\hat{u})} = 1.$$

Remarque 5.7. *La notion d'efficacité n'est définie que lorsque les conditions de Cramer-Rao sont vérifiées. Par exemple, si $X \sim \mathcal{U}_{[0,\theta]}$ alors le support de X est $[0, \theta]$ et dépend de θ, il n'y a donc pas d'estimateur efficace de θ.*

Théorème 5.5. *Les conditions de Cramer-Rao étant vérifiées, $\hat{u}$ est un estimateur efficace de $u(\theta)$ si et seulement si*

$$\ell'(\theta; X_1, \dots, X_n) = A(n, \theta)(\hat{u} - u(\theta)), \tag{5.1}$$

où $A(n, \theta)$ est un terme ne dépendant pas des x_i. On a alors :

$$\mathrm{Var}(\hat{u}) = \frac{u'(\theta)}{A(n, \theta)}.$$

Preuve. cf. Section 5.6.4. □

Théorème 5.6. *S'il existe un estimateur efficace de θ, il est identique à l'unique EMV.*

Preuve. Si $\hat{\theta}$ est efficace,

$$\ell'(\theta; X_1, \dots, X_n) = A(n, \theta)(\hat{\theta} - \theta).$$

Donc $\hat{\theta} = \theta$ est l'unique solution de l'équation de vraisemblance. □

5.5.3 Estimateur sans biais de variance minimale

Définition 5.10. *L'estimateur $\hat{u}$ est dit* sans biais de variance minimale (ou optimal) *si et seulement si $\mathbb{E}(\hat{u}) = u(\theta)$ et*

$$\mathrm{Var}(\hat{u}) \leq \mathrm{Var}(\hat{u}')$$

pour tout estimateur $\hat{u}'$ sans biais pour $u(\theta)$.

Donc, si $\hat{u}$ est efficace alors $\hat{u}$ est optimal. Mais la réciproque est fausse : la borne de Fréchet n'est pas forcément définie ou atteinte. L'optimalité est donc une propriété plus faible que l'efficacité.

Proposition 5.7. *S'il existe un estimateur optimal, il est unique presque sûrement : si $\hat{u}$ et $\widehat{u'}$ sont deux estimateurs sans biais de variance minimale, on a $\mathbb{P}(\hat{u} = \widehat{u'}) = 1$.*

5.6 Démonstrations

5.6.1 Preuve du Théorème 5.2

Nous nous contenterons de faire la preuve dans le cas où X est une v.a. continue, la preuve dans le cas discret étant similaire. Par ailleurs, nous ne détaillerons pas les « conditions de régularité » qui concernent l'existence des dérivées et intégrales manipulées et qui sont toujours vérifiées dans les cas rencontrés en pratique à ce niveau d'exposition.

La fonction $(x_1, \dots, x_n) \mapsto L(\theta; x_1, \dots, x_n)$ étant la fonction de densité par rapport à l'échantillon, on a

$$\int_{\mathbb{R}^n} L(\theta; x_1, \dots, x_n)\, dx_1 \cdots dx_n = 1.$$

Le domaine de X ne dépendant pas de θ, on peut dériver par rapport à θ sous le signe d'intégration ; on obtient donc :

$$\int_{\mathbb{R}^n} L'(\theta; x_1, \dots, x_n)\, dx_1 \cdots dx_n = 0.$$

Par ailleurs, en utilisant la dérivée logarithmique, on a

$$\ell'(\theta; x_1, \dots, x_n) = \frac{L'(\theta; x_1, \dots, x_n)}{L(\theta; x_1, \dots, x_n)},$$

d'où

$$\int_{\mathbb{R}^n} \ell'(\theta; x_1, \dots, x_n) L(\theta; x_1, \dots, x_n)\, dx_1 \cdots dx_n = 0, \tag{5.2}$$

c'est-à-dire

$$\mathbb{E}\left[\ell'(\theta; X_1, \dots, X_n)\right] = 0,$$

ce qui démontre la première partie du théorème. En dérivant une seconde fois (5.2), on a :

$$\int_{\mathbb{R}^n} \ell'' L \, dx_1 \cdots dx_n + \int_{\mathbb{R}} \ell' L' \, dx_1 \cdots dx_n = 0.$$

En utilisant une nouvelle fois la dérivée logarithmique, on peut en déduire

$$\int_{\mathbb{R}} \ell'' L \, dx_1 \cdots dx_n + \int_{\mathbb{R}} (\ell')^2 L \, dx_1 \cdots dx_n = 0,$$

c'est-à-dire

$$\mathbb{E}\left[\ell''(\theta; X_1, \dots, X_n)\right] + \mathbb{E}\left[(\ell'(\theta; X_1, \dots, X_n))^2\right] = 0,$$

d'où le résultat.

5.6.2 Preuve du Théorème 5.3

Soit la fonction

$$\phi_n(\mathbf{X}, \theta) = \frac{1}{n}\ell'(\theta; X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(X_i; \theta)}{\partial \theta},$$

où $\mathbf{X} = (X_1, \dots, X_n)$. D'après le théorème des accroissements finis, pour tout θ, il existe $\theta' \in]\min(\theta_0, \theta), \max(\theta_0, \theta)[$ tel que

$$\phi_n(\mathbf{X}, \theta) - \phi_n(\mathbf{X}, \theta_0) = (\theta - \theta_0)\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta').$$

Pour $\theta = \hat{\theta}_n$ on a donc

$$\underbrace{\phi_n(\mathbf{X}, \hat{\theta}_n)}_{0} - \phi_n(\mathbf{X}, \theta_0) = (\hat{\theta}_n - \theta_0)\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'')$$

avec $\min(\hat{\theta}_n, \theta_0) < \theta'' < \max(\hat{\theta}_n, \theta_0)$, soit

$$\phi_n(\mathbf{X}, \theta_0) + (\hat{\theta}_n - \theta_0)\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'') = 0.$$

En multipliant par $\sqrt{n}$, on obtient

$$\sqrt{n}\ \phi_n(\mathbf{X}, \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'') = 0,$$

d'où

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\phi_n(\mathbf{X}, \theta_0)}{-\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'')}. \tag{5.3}$$

Étude du dénominateur :

$$-\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'') = -\frac{1}{n}\ell''(\theta''; X_1, \dots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \ell''(\theta''; X_i).$$

Sachant que la moyenne empirique converge en probabilité vers l'espérance (loi faible des grand nombre), et que θ'' converge en probabilité vers θ_0 (étant encadré entre θ_0 et $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$), on peut en déduire

$$-\frac{\partial \phi_n}{\partial \theta}(\mathbf{X}, \theta'') = -\frac{1}{n}\sum_{i=1}^{n} \ell''(\theta''; X_i) \xrightarrow{\mathbb{P}} -\mathbb{E}(\ell''(\theta_0; X_1)) = I(\theta_0), \tag{5.4}$$

où $I(\theta_0)$ est l'information de Fisher en θ_0 pour un échantillon de taille 1.

Étude du numérateur :

$$\phi_n(\mathbf{X}, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(X_i; \theta_0)}{\partial \theta} = \overline{Z}, \quad \text{avec} \quad Z_i = \frac{\partial \ln f(X_i; \theta_0)}{\partial \theta}.$$

D'après le TCL :

$$\frac{\sqrt{n}(\phi_n(X, \theta_0) - \mathbb{E}(Z))}{\sqrt{\mathrm{Var}(Z)}} \xrightarrow{\mathscr{L}} \mathcal{N}(0, 1).$$

Or $\mathbb{E}(Z) = 0$ (Théorème 5.2) et $\mathrm{Var}(Z) = I(\theta_0)$, donc

$$\sqrt{n}\ \phi_n(X, \theta_0) \xrightarrow{\mathscr{L}} \mathcal{N}(0, I(\theta_0)) \tag{5.5}$$

D'après le théorème de Slutsky, les équations (5.3), (5.4) et (5.5) impliquent :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathscr{L}} \mathcal{N}(0, I(\theta_0)) \times \frac{1}{I(\theta_0)} = \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Comme $I_n(\theta_0) = nI(\theta_0)$ (proposition 5.6), on a finalement

$$\frac{\hat{\theta}_n - \theta_0}{\sqrt{1/I_n(\theta_0)}} \xrightarrow{\mathscr{L}} \mathcal{N}(0, 1).$$

5.6.3 Preuve du Théorème 5.4

Notons $\hat{u} = \hat{u}(X_1, \dots, X_n)$. En utilisant le théorème 5.2 et la proposition 3.9, on a

$$\mathrm{Cov}\left[\hat{u}, \ell'(\theta; X_1, \dots, X_n)\right] = \mathbb{E}\left[\hat{u}\ \ell'(\theta; X_1, \dots, X_n)\right].$$

On a donc

$$\mathrm{Cov}(\hat{u}, \ell'(\theta; X_1, \dots, X_n)) = \int_{\mathbb{R}^n} \hat{u}(x_1, \dots, x_n)\ell'(\theta; x_1, \dots, x_n)L(\theta; x_1, \dots, x_n)dx_1 \dots dx_n$$

$$= \int \hat{u}(x_1, \dots, x_n)L'(\theta; x_1, \dots, x_n)dx_1 \dots dx_n$$

$$= \frac{d}{d\theta} \int \hat{u}(x_1, \dots, x_n) L(\theta; x_1, \dots, x_n) dx_1 \dots dx_n$$

$$= \frac{d}{d\theta} \underbrace{E(\hat{u})}_{u(\theta)} = u'(\theta). \tag{5.6}$$

D'autre part, l'inégalité de Cauchy-Schwarz (proposition 3.12) entraîne

$$[\mathrm{Cov}(\hat{u}, \ell'(\theta; X_1, \dots, X_n))]^2 \leq \mathrm{Var}(\hat{u}) \mathrm{Var}\left[\ell'(\theta; X_1, \dots, X_n).\right] \tag{5.7}$$

De (5.6) et (5.7), on déduit

$$(u'(\theta))^2 \leq \mathrm{Var}(\hat{u}) \mathrm{Var}\left(\ell'(\theta; X_1, \dots, X_n)\right) = \mathrm{Var}(\hat{u}) \mathbb{E}\left[\left(\ell'(\theta; X_1, \dots, X_n)\right)^2\right].$$

5.6.4 Preuve du Théorème 5.5

Tout d'abord, montrons que la condition (5.1) implique que l'estimateur est sans biais. On a

$$\mathbb{E}\left[\ell'(\theta; X_1, \dots, X_n)\right] = A(n, \theta) \mathbb{E}[\hat{u} - u(\theta)]$$

d'où $A(n, \theta)(\mathbb{E}(\hat{u}) - u(\theta)) = 0$ et $\mathbb{E}(\hat{u}) = u(\theta)$. Dans la démonstration de l'inégalité de FDCR, la borne est atteinte si et seulement si l'inégalité de Cauchy-Schwarz (5.7) est une égalité, c'est-à-dire si $\hat{u} - u(\theta)$ et $\ell'(\theta; X_1, \dots, X_n)$ sont deux v.a. proportionnelles :

$$\ell'(\theta; X_1, \dots, X_n) = A(n, \theta)(\hat{u} - u(\theta)).$$

Chapitre 6

Estimation par intervalle de confiance

6.1 Introduction

Soit X une v.a. de loi $F_X \in (F_\theta)_{\theta\in\Theta}$. Nous avons défini un estimateur de θ comme une statistique $T = \varphi(X_1, \dots, X_n)$, fonction d'un échantillon $X_1, \dots, X_n$ de v.a. parente X, dont toute réalisation $t = \varphi(x_1, \dots, x_n)$ peut être considérée comme une approximation de θ. Or une telle approximation, appelée *estimation ponctuelle de θ*, est évidemment entachée d'incertitude : plusieurs réalisations de l'échantillon conduiront en général à des résultats différents ; mais on ne dispose généralement que d'une seule réalisation. L'incertitude relative à une estimation dépend par ailleurs de la taille de l'échantillon. Par exemple, une estimation de 55% d'électeurs en faveur d'un candidat à une élection n'a pas même fiabilité selon que le nombre d'électeurs sondés est de 100 ou 10000.

Il semble donc souvent plus satisfaisant de donner une estimation non pas sous la forme d'un seul nombre, mais sous la forme d'un *ensemble de valeurs plausibles* pour le paramètre θ. Lorsqu'il s'agit d'un paramètre scalaire (seul cas considéré dans ce chapitre), cet ensemble est choisi parmi les intervalles réels : on parle d'*intervalle de confiance sur le paramètre θ*. Les bornes de cet intervalle sont déterminées en fonction de l'échantillon : ce sont donc des variables aléatoires, qui seront définies de manière à encadrer la vraie valeur du paramètre θ avec une probabilité $1 - \alpha$ (appelée *niveau de confiance*) choisie par l'utilisateur.

6.2 Définitions

Définition 6.1. *Soit $\alpha \in]0; 1[$. On appelle :*

— *intervalle de confiance bilatéral pour le paramètre θ, au niveau $1 - \alpha$, tout intervalle $[T_1, T_2]$, où T_1 et T_2 sont des statistiques de l'échantillon, $T_1 = \varphi_1(X_1, \dots, X_n)$ et $T_2 = \varphi_2(X_1, \dots, X_n)$, vérifiant*

$$\mathbb{P}(T_1 \leq \theta \leq T_2) = \mathbb{P}([T_1, T_2] \ni \theta) = 1 - \alpha;$$

— *intervalle de confiance unilatéral pour θ, au niveau $1-\alpha$, tout intervalle de la forme $[T, +\infty[$ ou $]-\infty, T]$, vérifiant :*

$$\mathbb{P}(T \leq \theta) = \mathbb{P}([T, +\infty[\ni \theta) = 1-\alpha,$$

ou

$$\mathbb{P}(T \geq \theta) = \mathbb{P}(]-\infty, T] \ni \theta) = 1-\alpha.$$

Remarque 6.1. *Soient $[T_1, T_2]$ un intervalle de confiance de niveau $1-\alpha$, t_1 et t_2 des réalisations de T_1 et T_2. L'intervalle constant $[t_1, t_2]$ est appelé* réalisation de l'intervalle $[T_1, T_2]$ *(ou intervalle de confiance par abus de langage). Il est incorrect d'écrire $\mathbb{P}(\theta \in [t_1, t_2]) = 1-\alpha$, car θ, t_1, t_2 sont des constantes, et non des variables aléatoires. La proposition $\theta \in [t_1, t_2]$ est vraie ou fausse : sa probabilité ne peut donc valoir que 1 ou 0. On sait seulement que l'intervalle $[t_1, t_2]$ a été obtenu par une procédure qui fournit dans $100(1-\alpha)$ % des cas un intervalle contenant la vraie valeur de θ.*

Remarque 6.2. *Il n'y a pas unicité de l'intervalle de confiance. Le choix peut se faire en cherchant l'intervalle de plus petite longueur, c'est-à-dire le plus précis, ou plus souvent pour des raisons de simplicité en répartissant de manière symétrique la valeur α. Sous certaines hypothèses, comme par exemple la symétrie de la densité de probabilité, il est possible de montrer que ces deux critères se rejoignent.*

6.3 Construction pratique

6.3.1 Notion de fonction pivotale

Il est souvent pratique, pour construire un intervalle de confiance, de passer par la notion de *fonction pivotale.*

Définition 6.2. *On appelle fonction pivotale pour le paramètre θ une v.a. $\pi(X_1, \dots, X_n; \theta)$, fonction de l'échantillon et de θ, dont la loi ne dépend pas de θ, ni d'aucun autre paramètre inconnu.*

Exemple 6.1. *$X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 connue. Soit*

$$\pi(X_1, \dots, X_n; \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

La loi de $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ne dépend pas de μ : c'est une fonction pivotale pour ce paramètre.

Supposons maintenant que le paramètre σ^2 est inconnu. On sait d'après le théorème de Fisher que

$$\frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2_{n-1}.$$

La statistique $(n-1)S^{*2}/\sigma^2$ est donc une fonction pivotale pour σ^2.

6.3.2 Utilisation

La méthode de construction d'un intervalle de confiance à partir d'une fonction pivotale est illustrée par la figure 6.1. Elle peut être décomposée en deux étapes.

1. On commence par définir un intervalle de probabilité $1-\alpha$ pour la fonction pivotale : on détermine donc deux nombres a et b tels que

$$\mathbb{P}(a \leq \pi(X_1, \dots, X_n, \theta) \leq b) = 1 - \alpha.$$

   Il existe en général une infinité de façons de choisir a et b. La plus simple consiste à prendre les fractiles $\pi_{\alpha/2}$ et $\pi_{1-\alpha/2}$ de la loi de π :

$$a = \pi_{\alpha/2}, \quad b = \pi_{1-\alpha/2}.$$

   En particulier, on montre que ce choix fournit l'intervalle de longueur minimale, lorsque la loi de π est symétrique et unimodale.
2. On détermine ensuite l'ensemble des valeurs de θ tel que

$$a \leq \pi(x_1, \dots, x_n; \theta) \leq b, \tag{6.1}$$

   où $x_1, \dots, x_n$ désigne la réalisation observée de l'échantillon. Si π est fonction monotone de θ, la résolution du système de deux inéquations (6.1) donne un intervalle de la forme

$$\varphi_1(x_1, \dots, x_n) \leq \theta \leq \varphi_2(x_1, \dots, x_n).$$

   Par construction, on a bien

$$\begin{aligned}\mathbb{P}([\varphi_1(X_1, \dots, X_n), \varphi_2(X_1, \dots, X_n)] \ni \theta) &= \mathbb{P}(a \leq \pi(X_1, \dots, X_n; \theta) \leq b) \\ &= 1 - \alpha.\end{aligned}$$

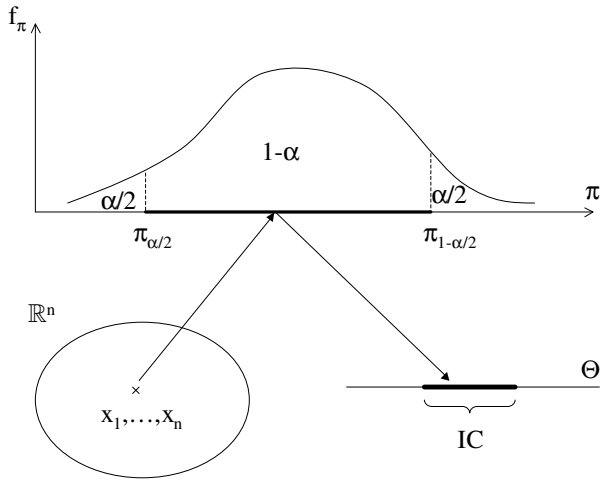


FIGURE 6.1 – Principe de construction d'un intervalle de confiance à partir d'une fonction pivotale.

Remarque 6.3. *La construction d'un intervalle de confiance unilatéral se fait de manière similaire : on part de l'égalité*

$$\mathbb{P}(\pi(X_1, \dots, X_n; \theta) \geq \pi_\alpha) = 1 - \alpha$$

ou

$$\mathbb{P}(\pi(X_1, \dots, X_n; \theta) \leq \pi_{1-\alpha}) = 1 - \alpha,$$

et on résout l'inéquation en θ (voir les exemples ci-dessous).

6.4 Intervalle de confiance sur une espérance

6.4.1 Cas gaussien, variance connue

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 connu. On a vu que $\sqrt{n}(\overline{X}-\mu)/\sigma \sim \mathcal{N}(0,1)$ est une fonction pivotale pour μ. Supposons que l'on cherche à construire un intervalle de confiance bilatéral pour μ. En appliquant la technique précédente, on a :

$$\mathbb{P}\left(u_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \overline{X}-\mu \leq u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\overline{X}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Appliquons maintenant la méthode à la construction d'un intervalle de confiance unilatéral. On a, par exemple,

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \geq u_{\alpha}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\overline{X}-\mu \geq u_{\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\mu \leq \overline{X}-u_{\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\mu \leq \overline{X}+u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

d'où l'intervalle de confiance unilatéral : $\left]-\infty, \overline{X}+u_{1-\alpha}\sigma/\sqrt{n}\right]$. Pour obtenir l'IC unilatéral dans l'autre sens, il suffit d'écrire :

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\overline{X}-\mu \leq u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\mu \geq \overline{X}-u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

d'où l'intervalle de confiance unilatéral $\left[\overline{X}-u_{1-\alpha}\sigma/\sqrt{n}, +\infty\right[$.

Remarque 6.4. *La longueur de l'intervalle de confiance bilatéral de niveau $1-\alpha$ est $2u_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$. Elle est donc proportionnelle à σ, inversement proportionnelle à $\sqrt{n}$ et elle est fonction croissante du niveau de confiance $1-\alpha$: quand on augmente le niveau de confiance, la longueur de l'intervalle de confiance (et donc l'imprécision de l'estimation) augmente.*

6.4.2 Cas gaussien, variance inconnue

Considérons maintenant la situation plus réaliste où $X \sim \mathcal{N}(\mu, \sigma^2)$, les paramètres μ et σ^2 étant tous deux inconnus. On a toujours $\sqrt{n}(\overline{X} - \mu)/\sigma \sim \mathcal{N}(0,1)$, mais on ne peut plus utiliser cette fonction pivotale car σ est inconnu.

Soit la v.a. $\sqrt{n}(\overline{X} - \mu)/S^*$, obtenue en remplaçant σ par S^* au dénominateur. Pour déterminer la loi de cette v.a., on fait apparaître $\sqrt{n}(\overline{X} - \mu)/\sigma$ et $(n-1)S^{*2}/\sigma^2$ dont les lois sont connues. Il vient

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S^*} = \frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{S^*/\sigma} = \frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\sqrt{\frac{(n-1)S^{*2}/\sigma^2}{n-1}}}.$$

La v.a. $\sqrt{n}(\overline{X} - \mu)/S^*$ est donc le rapport d'une v.a. normale centrée réduite, et de la racine carrée d'une v.a. suivant une loi du χ^2, divisée par son nombre de degrés de liberté. De plus, le numérateur et le dénominateur sont indépendants (théorème 4.3 de Fisher). Par définition, $\sqrt{n}(\overline{X} - \mu)/S^*$ suit une *loi de Student à $n-1$ degrés de liberté* (cf. Section 3.7.2).

On peut donc à présent caractériser la loi de la fonction pivotale $\sqrt{n}(\overline{X} - \mu)/S^*$:

$$\frac{\overline{X} - \mu}{S^*/\sqrt{n}} \sim \mathcal{T}_{n-1}.$$

Celle-ci est de la même forme que $\sqrt{n}(\overline{X} - \mu)/\sigma$. Les intervalles de confiance obtenus auront donc la même forme que dans le cas où la variance est connue : il suffira de remplacer σ par S^* et les fractiles de la loi normale par ceux de la loi de Student à $n-1$ d.d.l. On a donc :

— intervalle de confiance bilatéral :

$$IC = \left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\right];$$

— intervalles de confiance unilatéraux :

$$IC_1 = \left[\overline{X} - t_{n-1,1-\alpha}\frac{S^*}{\sqrt{n}}, +\infty\right[\quad \text{et} \quad IC_2 = \left]-\infty, \overline{X} + t_{n-1,1-\alpha}\frac{S^*}{\sqrt{n}}\right].$$

6.4.3 Cas général

Supposons maintenant que X suit une loi quelconque, d'espérance μ et de variance σ^2. Considérons tout d'abord le cas où la variance σ^2 est connue. Le TCL permet d'affirmer que : $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ quand $n \to \infty$. La suite $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ converge donc en loi vers une loi ne dépendant pas de μ : on dit qu'il s'agit d'une *fonction asymptotiquement pivotale pour μ*. On a donc, pour n assez grand :

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq u_{1-\frac{\alpha}{2}}\right) \longrightarrow 1 - \alpha$$

$$\Leftrightarrow \mathbb{P}\left(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \longrightarrow 1 - \alpha.$$

L'intervalle de confiance de niveau $1-\alpha$ établi rigoureusement dans le cas gaussien peut donc être utilisé comme approximation dans le cas général (on parle alors d'intervalle de confiance approché). En pratique, on admet que l'approximation est valide dès que $n \geq 5$.

Considérons maintenant le cas où la variance est inconnue. On a toujours

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$$

et

$$S^{*2} \xrightarrow{\mathbb{P}} \sigma^2 \;\Rightarrow\; \frac{S^{*2}}{\sigma^2} \xrightarrow{\mathbb{P}} 1.$$

On en déduit, d'après le théorème de Slutsky (théorème 3.7) :

$$\frac{\overline{X}-\mu}{S^*/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\frac{\sigma}{S^*} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$$

On a donc :

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{S^*/\sqrt{n}} \leq u_{1-\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha,$$

d'où l'on déduit l'intervalle de confiance bilatéral approché suivant :

$$IC = \left[\overline{X} - u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}, \overline{X} + u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\right].$$

On admet que cette approximation peut être utilisée dès que $n \geq 30$.

6.5 Intervalle de confiance sur la variance d'une loi normale

6.5.1 Cas où l'espérance est connue

Dans ce cas, on peut montrer que $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ est l'EMV de σ^2. On a $\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$. C'est une fonction pivotale pour σ^2. On a donc :

$$\mathbb{P}\left(\chi^2_{n;\frac{\alpha}{2}} \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n;1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{\chi^2_{n;\frac{\alpha}{2}}}{n\hat{\sigma}^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n;1-\frac{\alpha}{2}}}{n\hat{\sigma}^2}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\chi^2_{n;1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi^2_{n;\frac{\alpha}{2}}}\right) = 1-\alpha,$$

d'où l'intervalle de confiance bilatéral :

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n;1-\frac{\alpha}{2}}}, \frac{n\hat{\sigma}^2}{\chi^2_{n;\frac{\alpha}{2}}}\right].$$

On obtient de manière similaire les intervalles de confiance unilatéraux suivants :

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n;1-\alpha}}, +\infty\right[\quad \text{et} \quad \left]0, \frac{n\hat{\sigma}^2}{\chi^2_{n;\alpha}}\right].$$

6.5.2 Cas où l'espérance est inconnue

D'après le théorème de Fisher, $\frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$. Cette fonction pivotale a la même forme que la précédente : il suffit de remplacer n par $n-1$ et σ^2 par S^{*2}. On en déduit les intervalles de confiance suivants :

— intervalle de confiance bilatéral :

$$\left[\frac{(n-1)S^{*2}}{\chi^2_{n-1;1-\frac{\alpha}{2}}}, \frac{(n-1)S^{*2}}{\chi^2_{n-1;\frac{\alpha}{2}}}\right];$$

— intervalles de confiance unilatéraux :

$$\left[\frac{(n-1)S^{*2}}{\chi^2_{n-1;1-\alpha}}, +\infty\right[\quad \text{et} \quad \left]0, \frac{(n-1)S^{*2}}{\chi^2_{n-1;\alpha}}\right].$$

6.6 Intervalles de confiance sur une proportion

Supposons que, dans un sondage réalisé sur un échantillon de 100 personnes, 60 électeurs se soient prononcés en faveur d'un candidat. Comment en déduire un intervalle de confiance sur la proportion p d'électeurs favorables au candidat dans la population totale ?

Ce problème peut être modélisé de la manière suivante. Soit X le nombre de personnes favorables au candidat parmi les n personnes interrogées. Si n est très inférieur à la taille de la population totale, on peut admettre que $X \sim \mathscr{B}(n, p)$. L'estimateur de maximum de vraisemblance de p est $\hat{p} = X/n$. On a vu que, par suite du TCL :

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1), \quad \text{c'est-à-dire} \quad \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

C'est une fonction asymptotiquement pivotale pour p. Néanmoins, il n'est pas facile d'en déduire un intervalle de confiance sur p, car p apparaît sous le radical, au dénominateur.

6.6.1 Avec application du théorème de Slutsky

À partir de la fonction asymptotiquement pivotale, on peut obtenir une autre fonction pivotale pour p en remarquant que $\hat{p} \xrightarrow{\mathbb{P}} p$, d'où :

$$\sqrt{\hat{p}(1-\hat{p})} \xrightarrow{\mathbb{P}} \sqrt{p(1-p)} \quad \text{c'est-à-dire} \quad \sqrt{\frac{p(1-p)}{\hat{p}(1-\hat{p})}} \xrightarrow{\mathbb{P}} 1.$$

D'après le théorème de Slutsky (Théorème 3.7), on a donc :

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sqrt{\frac{p(1-p)}{\hat{p}(1-\hat{p})}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

On en déduit :

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq u_{1-\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\hat{p}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \longrightarrow 1-\alpha.$$

Remarque 6.5. *Il est également possible de déterminer la taille minimale n de l'échantillon telle que l'intervalle de confiance au niveau $1-\alpha$ soit de la forme $\hat{p} \pm \Delta\hat{p}$, avec $\Delta\hat{p} \leq c$, où c est une constante fixée. En effet, on a :*

$$\Delta\hat{p} = u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq c \Leftrightarrow \frac{\hat{p}(1-\hat{p})}{n} \leq \frac{c^2}{u^2_{1-\alpha/2}} \Leftrightarrow n \geq \frac{\hat{p}(1-\hat{p})u^2_{1-\alpha/2}}{c^2}.$$

On ne connaît pas $\hat{p}$, mais on peut remarquer que $\hat{p}(1-\hat{p}) \leq 1/4, \forall \hat{p} \in [0,1]$. Donc :

$$n \geq \frac{u^2_{1-\frac{\alpha}{2}}}{4c^2} \Rightarrow n \geq \frac{\hat{p}(1-\hat{p})u^2_{1-\alpha/2}}{c^2}, \quad \forall \hat{p} \in [0,1] \Rightarrow \Delta\hat{p} \leq c, \quad \forall \hat{p} \in [0,1].$$

Application numérique :

$$c = 0.05, \alpha = 0.05 \quad \Rightarrow \quad n \geq \frac{(1.96)^2}{4(0.05)^2} = 384.16,$$

$$c = 0.01, \alpha = 0.05 \quad \Rightarrow \quad n \geq \frac{(1.96)^2}{4(0.01)^2} = 9604.$$

6.6.2 Sans l'application du théorème de Slutsky

À partir de la fonction asymptotiquement pivotale on peut déduire un intervalle de confiance sur p avec un peu de calcul. En commençant déduire de la loi de la fonction pivotale

$$\mathbb{P}\left(\left|\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| < u_{1-\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha,$$

on en déduit en élevant au carré que

$$\mathbb{P}\left(\frac{(\hat{p}-p)^2}{\frac{p(1-p)}{n}} < u^2_{1-\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha,$$

puis en développant

$$\mathbb{P}\left(\left(1+\frac{u^2_{1-\frac{\alpha}{2}}}{n}\right)p^2 - \left(2\hat{p}+\frac{u^2_{1-\frac{\alpha}{2}}}{n}\right)p + \hat{p}^2 < 0\right) \longrightarrow 1-\alpha.$$

On remarque un polynôme de degré deux en p, à branches asymptotiques vers $+\infty$ car $1 + u^2_{1-\frac{\alpha}{2}}/n$ est positif, donc ce polynôme est négatif entre ses racines. En notant les racines $\hat{p}_1$ et $\hat{p}_2$ les racines qui sont des fonctions de $\hat{p}$:

$$\hat{p}_{1,2} = \frac{2n\hat{p} + u^2_{1-\frac{\alpha}{2}} \pm u_{1-\frac{\alpha}{2}}\sqrt{u^2_{1-\frac{\alpha}{2}} + 4n\hat{p}(1-\hat{p})}}{2n + 2u^2_{1-\frac{\alpha}{2}}},$$

on obtient

$$\mathbb{P}\left(\hat{p}_1 < p < \hat{p}_2\right) \longrightarrow 1 - \alpha,$$

ce qui termine le raisonnement.

Remarque 6.6. *Quand n devient grand les intervalles de confiance obtenus avec et sans l'application du théorème de Slutsky deviennent équivalents, toutefois pour des petites valeurs de pn ou* $(1-p)n$ *on préfèrera utiliser la version sans l'application du théorème de Slutsky.*

Remarque 6.7. *Dans le cas où p est éloigné de 1/2, cette version sans l'application du théorème de Slutsky continue de donner des bornes cohérentes (à l'intérieur de* $[0,1]$*), à la différence de la version utilisant le théorème de Slutsky. Ceci s'explique par le fait que la fonction variance est ici* $p(1-p)$*, qui a une tangente nulle en 1/2 donc très peu sensible à une erreur sur p au voisinage de 1/2. Bien que l'application du théorème de Slutsky reste vraie asymptotiquement, l'erreur effectuée à n fixé est d'autant plus importante que p s'éloigne de 1/2.*

6.7 Construction à partir d'un EMV

Soit X une v.a. de loi $F_X \in (F_\theta)_{\theta\in\Theta}$ et $\hat{\theta}$ un estimateur de maximum de vraisemblance de θ. On a vu que, sous certaines conditions de régularité :

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{I_n(\theta)}}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Ce résultat fournit donc une fonction asymptotiquement pivotale pour θ, à partir de laquelle on peut construire un intervalle de confiance approché pour θ.

Chapitre 7

Régression linéaire

7.1 Exemple introductif

Au cours d'une enquête effectuée en 1975, on a relevé la surface en m^2 et le prix en milliers d'euros de 24 appartements situés dans deux arrondissements de Paris et on a obtenu les données suivantes :

surface	28	50	196	55	190	110	60	48	90	35	86	65
prix	130	280	800	268	790	500	320	250	378	250	350	300
surface	32	52	40	70	28	30	105	52	80	60	20	100
prix	155	245	200	325	85	78	375	200	270	295	85	495

On cherche à savoir si le prix est lié, et de quelle façon, à la surface. Une représentation graphique des données sous forme d'un diagramme de corrélation (figure 7.1) permet déjà d'avoir une idée de cette relation mais il est nécessaire d'aller plus loin pour la mesurer.

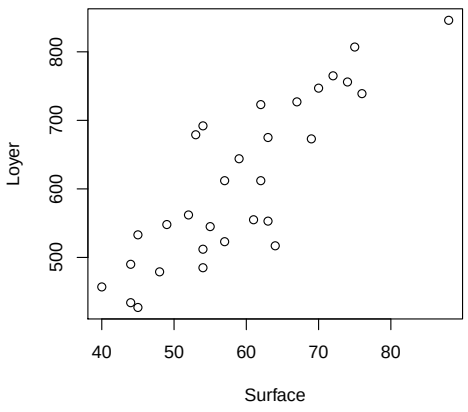


FIGURE 7.1 – Représentation graphique des données.

Pour cela, on peut considérer que le prix d'un appartement est une v.a. Y, dont l'espérance dépend de la surface x. Dans ce modèle, la variable Y, appelée *variable à expliquer* ou *variable dépendante*, et donc une variable aléatoire, tandis que la variable x, appelée *variable explicative* ou *variable indépendante*, est supposée connue et non aléatoire.

Les problèmes posés sont multiples :
— spécifier le modèle ;
— estimer les paramètres du modèle ;
— vérifier qu'il y a bien une relation entre les deux variables ;
— vérifier la validité du modèle retenu ;
— prédire le prix d'un nouvel appartement en fonction de sa surface, etc.

7.2 Le modèle

Le modèle de la régression linéaire simple ("simple" fait référence à une seule variable explicative) considère que l'espérance de la variable aléatoire Y est une fonction linéaire de la variable x

$$\mathbb{E}(Y_i) = a + bx_i,$$

ce qui peut aussi s'écrire

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \dots, n,$$

avec $\mathbb{E}(\varepsilon_i) = 0$. On supposera en outre que les ε_i sont indépendants et suivant la même loi normale $\mathcal{N}(0, \sigma^2)$. Le modèle peut alors s'écrire

$$Y_i \sim \mathcal{N}(a + bx_i, \sigma^2), \quad i = 1, \dots, n.$$

Remarquons que de nombreux modèles peuvent, par changement de variables, se ramener à ce modèle linéaire. Par exemple, le modèle

$$Y_i = a + bx_i^2 + \varepsilon_i$$

se ramène au modèle linéaire par le changement de variable $x' = x^2$.

La droite d'équation $y = a + bx$ est appelée *droite de régression de Y en x*.

Quatre hypothèses fondamentales sont effectuées lors de l'utilisation d'un modèle linéaire, elles sont ici détaillées :

linéarité : c'est l'hypothèse qui est la plus naturelle, dans le cas général, cela consiste à dire que $\mathbb{E}(Y_i)$ est une forme linéaire des paramètres. En régression linéaire simple cela s'écrit $\mathbb{E}(Y_i) = a + bx_i$.

normalité : les erreurs sont supposées suivre une loi normale. La loi normale est une loi à queue légère, il est donc très improbable d'avoir des données extrêmes, et des données aberrantes ont une grande influence (souvent néfaste) sur les résultats.

homoscédasticité : les erreurs ont même dispersion. Cela se traduit par le fait que la variance des erreurs est la même pour toute les erreurs.

indépendance : les erreurs sont supposées indépendantes.

7.3 Estimation des paramètres

Pour estimer les paramètres du modèle, nous utilisons la méthode du maximum de vraisemblance. On a

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right), \quad i = 1, \dots, n,$$

d'où l'on déduit la fonction de vraisemblance de l'échantillon (indépendant mais non identiquement distribué) :

$$L(a,b,\sigma^2;y_1,\dots,y_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-a-bx_i)^2}{2\sigma^2}\right),$$

et la fonction de log-vraisemblance :

$$\ell(a,b,\sigma^2;y_1,\dots,y_n)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2.$$

7.3.1 Estimation de a et b

La maximisation de la vraisemblance par rapport à a et b est donc obtenue en minimisant la quantité

$$Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2,$$

appelée *critère des moindres carrés*. La solution est obtenue en annulant les dérivées partielles :

$$\begin{cases}\dfrac{\partial Q}{\partial a}(a,b)=-2\displaystyle\sum_{i=1}^{n}(y_i-(a+bx_i))=0\\[2ex]\dfrac{\partial Q}{\partial b}(a,b)=-2\displaystyle\sum_{i=1}^{n}x_i(y_i-(a+bx_i))=0\end{cases}$$

$$\Leftrightarrow\begin{cases}\overline{y}=a+b\overline{x}\\[2ex]\displaystyle\sum_{i=1}^{n}x_i(y_i-a-bx_i)=0\end{cases}$$

$$\Leftrightarrow\begin{cases}a=\overline{y}-b\overline{x}\\[2ex]\displaystyle\sum_{i=1}^{n}x_iy_i-(\overline{y}-b\overline{x})n\overline{x}-b\sum_{i=1}^{n}x_i^2=0\end{cases}$$

$$\Leftrightarrow\begin{cases}a=\overline{y}-b\overline{x}\\[2ex]\displaystyle\sum_{i=1}^{n}x_iy_i-n\overline{y}\,\overline{x}+nb\overline{x}^2-b\sum_{i=1}^{n}x_i^2=0\end{cases}$$

$$\Leftrightarrow\begin{cases}a=\overline{y}-b\overline{x}\\[2ex]b=\dfrac{\sum_{i=1}^{n}x_iy_i-n\overline{y}\,\overline{x}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\end{cases}$$

En notant

$$S_{xY}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})=\frac{1}{n}\sum_{i=1}^{n}x_iY_i-\overline{x}\overline{Y}$$

la *covariance empirique* des x_i et des Y_i, et s_x^2 la variance empirique des x_i, on obtient les estimateurs suivants :

$$\hat{b}=\frac{S_{xY}}{s_x^2}\quad\text{et}\quad\hat{a}=\overline{Y}-\frac{S_{xY}}{s_x^2}\overline{x},$$

appelés *estimateurs des moindres carrés* de a et de b. La droite d'équation $y = \hat{a} + \hat{b}x$ est la *droite des moindres carrés de Y en x*. L'expression de l'estimateur de a montre que point $(\overline{x}, \overline{Y})$ appartient à la droite des moindres carrés.

On introduit par ailleurs classiquement les notations suivantes :

$$\widehat{Y}_i = \hat{a} + \hat{b}x_i, \quad i = 1, \dots, n \quad \text{et} \quad \hat{\varepsilon}_i = Y_i - \widehat{Y}_i, \quad i = 1, \dots, n.$$

Les statistiques $\hat{\varepsilon}_i$ sont appelées *résidus*. Remarquons que nous avons utilisé la notation $\widehat{Y}_i$ bien que $\widehat{Y}_i$ ne soit pas un estimateur de Y_i, mais de $\mathbb{E}(Y_i)$. Il s'agit d'une notation couramment utilisée en régression.

7.3.2 Estimation de σ^2

L'estimateur du maximum de vraisemblance de σ^2 est obtenu en résolvant l'équation :

$$\frac{\partial \ell}{\partial \sigma^2}(a, b, \sigma^2; y_1, \dots, y_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0,$$

qui a pour unique solution :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}(y_i - a - bx_i)^2.$$

On obtient donc

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}(Y_i - \hat{a} - \hat{b}x_i)^2 = \frac{1}{n}\sum_{i=1}\hat{\varepsilon}_i^2,$$

c'est-à-dire la moyenne des carrés des résidus.

7.4 Principales propriétés

7.4.1 Propriétés des estimateurs $\hat{a}$ et $\hat{b}$

Proposition 7.1.

$$\hat{a} \sim \mathcal{N}\left(a, \frac{\sigma^2}{n}(1 + \frac{\overline{x}^2}{s_x^2})\right), \quad \hat{b} \sim \mathcal{N}\left(b, \frac{\sigma^2}{ns_x^2}\right), \quad \mathrm{Cov}(\hat{a}, \hat{b}) = -\frac{\overline{x}\sigma^2}{ns_x^2}.$$

Preuve. cf. Section 7.7.1. □

7.4.2 Analyse de la variance

La régression linéaire et l'analyse de la variance — une méthodologie statistique permettant de comparer les valeurs moyennes de différents échantillons qui fera l'objet du chapitre 13 — sont deux cas particuliers du modèle linéaire, et présentent donc des liens étroits. Les résultats, donnés ci-dessous dans le cas de la régression linéaire, feront l'objet d'une étude plus détaillée dans le chapitre 13 consacré à l'analyse de la variance.

Proposition 7.2 (Équation d'analyse de la variance)**.**

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{\textit{Variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2}_{\substack{\textit{Variance expliquée}\\ \textit{par la régression}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2}_{\textit{Variance résiduelle}}$$

que l'on notera

$$S_Y^2 = S_{reg} + S_{res}.$$

Preuve. En effectuant la décomposition

$$(Y_i-\overline{Y}) = (\widehat{Y}_i-\overline{Y}) + (Y_i-\widehat{Y}_i),$$

il reste à montrer que $\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})(Y_i-\widehat{Y}_i) = 0$: sachant que

$$\widehat{Y}_i-\overline{Y} = \frac{S_{xY}}{s_x^2}(x_i-\overline{x}),$$

on a

$$\begin{aligned}\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)(Y_i-\widehat{Y}_i) &= \sum_{i=1}^{n}\frac{S_{xY}}{s_x^2}(x_i-\overline{x})\left(Y_i-\overline{Y}-\frac{S_{xY}}{s_x^2}(x_i-\overline{x})\right)\\ &= \frac{S_{xY}}{s_x^2}\left(S_{xY}-\frac{S_{xY}}{s_x^2}s_x^2\right)\\ &= 0.\end{aligned}$$

□

Proposition 7.3 (Expression et propriétés de S_{reg} et S_{res})**.**

$$S_{reg} = \hat{b}^2 s_x^2, \quad S_{res} = \hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \quad \textit{et} \quad \mathbb{E}(S_{res}) = \frac{n-2}{n}\sigma^2$$

Preuve. cf. Section 7.7.2 □

Remarque 7.1. *Il est préférable, pour éviter trop d'erreurs d'arrondi, de calculer S_{res} à l'aide de la relation*

$$S_{res} = S_Y^2 - S_{reg} = S_Y^2 - \hat{b}^2 s_x^2,$$

plutôt que par le calcul direct de la somme des $\hat{\varepsilon}_i^2$.

7.4.3 Estimateur sans biais de σ^2

À partir des résultats précédents, on montre sans difficulté que $\frac{n}{n-2}S_{res}$ est un estimateur sans biais de σ^2. Nous utiliserons désormais cet estimateur qui sera noté simplement $\hat{\sigma}^2$. On peut aussi montrer les propositions suivantes :

Proposition 7.4.

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Proposition 7.5. *$\hat{\sigma}^2$ est une v.a. indépendante de $\overline{Y}$, de $\hat{a}$ et de $\hat{b}$.*

7.5 Pratique de la régression

La pratique de la regression recouvre un ensemble de techniques. Cette section présente un sous-ensemble réduit de l'ensemble de ces techniques [1].

La régression linéaire est effectuée dans `GNU R` grâce à la fonction `lm` (*linear model*) : l'objet de type `lm` retourné contient toutes les informations utiles. Par exemple, avec un `data.frame` nommé `D`, qui contient les colonnes `x` et `y` on pourra utiliser :

```
regression <- lm(y~1+x, data=D)
regression
```

`y~1+x` est une *formule*. Le `1` représente l'ordonnée à l'origine (nommé *intercept* dans `GNU R`). Le `x` représente la partie proportionnelle à `x` correspondant au paramètre de pente.

La régression linéaire peut être effectuée au moyen de `statsmodels`, une des manière de faire est d'utiliser l'API `formula`. Par exemple avec un `D` un `DataFrame` de `pandas` qui contient les colonnes `x` et `y`, on pourra utiliser :

```
import statsmodels.formula.api as smf

model = smf.ols(formula="y~1+x", data=D)
res = model.fit()
res.summary()
```

`"y~1+x"` est une *formule*. Le `1` représente l'ordonnée à l'origine (nommé *intercept* dans `statsmodels`). Le `x` représente la partie proportionnelle à `x` correspondant au paramètre de pente.

7.5.1 Effet de levier et identification des abscisses extrêmes

En utilisant sa définition, $\widehat{Y}_i$ peut s'écrire en fonction des Y_j :

1. Le lecteur interessé pour aller consulter par exemple le cours STAT462 (https://online.stat.psu.edu/stat462/) et en particulier les cours 9, 10 et 11, ou un ensemble beaucoup plus complet de ces techniques est présenté.

$$\begin{aligned}\widehat{Y}_i &= \hat{a} + \hat{b}x_i \\ &= \overline{Y} + \hat{b}(x_i - \overline{x}) \\ &= \overline{Y} + \frac{\sum_j \left(Y_j - \overline{Y}\right)\left(x_j - \overline{x}\right)}{\sum_j \left(x_j - \overline{x}\right)^2}(x_i - \overline{x}) \\ &= \overline{Y} + \frac{\sum_j \left(Y_j\right)\left(x_j - \overline{x}\right)}{\sum_j \left(x_j - \overline{x}\right)^2}(x_i - \overline{x}) \\ &= \sum_j h_{ij} Y_j\end{aligned}$$

Avec :

$$h_{ij} = \frac{1}{n} + \frac{\left(x_i - \overline{x}\right)\left(x_j - \overline{x}\right)}{\sum_k \left(x_k - \overline{x}\right)^2}$$

h_{ij} représente l'influence du points Y_j sur la prédiction $\widehat{Y}_i$.

Remarque 7.2. *On utilise la lettre h, car sous forme matricielle on écrit $\widehat{Y} = HY$, où H est la matrice de terme général h_{ij}. On nomme H the hat matrix en référence au fait que multiplier par elle revient à ajouter un chapeau sur Y pour avoir $\widehat{Y}$.*

Les termes qui nous intéressent particulièrement sont les termes de la forme h_{ii}, c'est à dire les termes correspondant à l'influence du point Y_i sur la prédiction au même endroit : $\widehat{Y}_i$. On nomme ces valeurs h_{ii} les leviers.

$$h_{ii} = \frac{1}{n} + \frac{\left(x_i - \overline{x}\right)^2}{\sum_k \left(x_k - \overline{x}\right)^2}$$

Proposition 7.6. *En régression linéaire simple on a :*

$$\sum_i h_{ii} = 2$$

En régression linéaire dans le cas général (plusieurs variables explicatives), avec p le nombre de paramètres de la forme linéaire (ordonnée à l'origine comprise), on a $\sum_i h_{ii} = p$.

Les leviers des points dont les abscisses sont proches de $\overline{x}$ sont faibles, et les points les plus éloignés de $\overline{x}$ sont ceux avec les leviers les plus important. La figure 7.2 représente deux cas de régression avec les leviers associés.

Diagnostic On sait que $\sum_i h_{ii} = 2$, donc le levier moyen est de $2/n$. On considèrera qu'un point a un levier extrême quand son levier est 3 fois supérieur au levier moyen, c'est à dire quand $h_{ii} > 3 \times 2/n$. Dans l'exemple de la figure 7.2, $n = 10$, donc on considère qu'un levier est extrême à partir de $6/10$, soit le point extrême de l'exemple de droite.

Avoir un levier extrême n'est pas un problème en soi dans la régression, il faut uniquement avoir conscience que les résultats seront très influencés par d'éventuelles

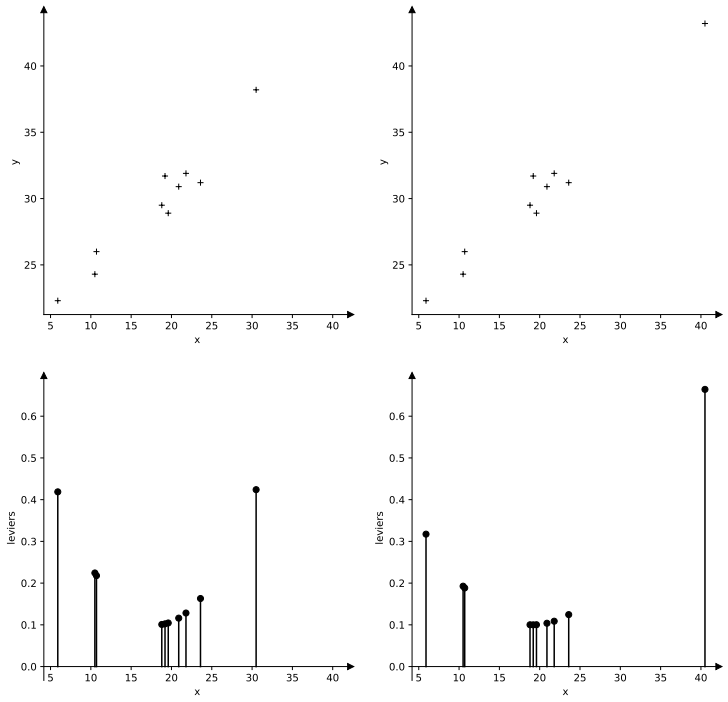


FIGURE 7.2 – Illustration des leviers dans le cas de deux exemples. Dans la colonne de gauche, les points sont plutôt bien répartis sur l'axe des abscisses, et on observe que les leviers ne sont pas très importants. Dans l'exemple de droite, il y a un point qui est loin dans l'axe des abscisses, et on observe que le levier associé est très grand, ce qui implique une forte sensibilité du résultat de la régression sur une erreur de ce point.

erreurs de ce point, et qu'il sera difficile d'obtenir des variances d'estimateurs petites ou des intervalles de confiance petits (ou des tests statistiques puissants) à cause de ces points.

Lorsqu'on a la maîtrise du plan d'expérience, on prendra soin de calculer les leviers avant de faire l'expérience (on a uniquement besoin des x_i, on peut utiliser de faux y_i pour duper les outils informatiques au besoin), puis d'ajouter des points de sorte que tous les points ne soient pas loin les uns des autres afin d'avoir les plus faibles leviers possibles.

Les leviers s'obtiennent sous GNU R avec la fonction `hatvalues`.

```
regression <- lm(y~1+x, data=D)
hatvalues(regression)
```

On peut les tracer facilement :

```
plot(d[['x']], hatvalues(regression))
```

Les leviers s'obitennent avec `statsmodels` au sein de l'influence obtenue à partir du résultat. Puis on peut tracer ce résultat.

```
import statsmodels.formula.api as smf

model = smf.ols(formula="y~1+x", data=D)
res = model.fit()
influence = res.get_influence()
infl.hat_matrix_diag
```

On peut les tracer facilement :

```
import matplotlib.pyplot as plt

plt.plot(D['x'], infl.hat_matrix_diag, 'o')
plt.show()
```

7.5.2 Analyse des résidus

Validation des hypothèses du modèle

Pour vérifier la validité de la régression, on peut étudier la distribution des résidus et vérifier que leur distribution est voisine d'une distribution normale centrée, hypothèse qui est faite dans la régression.

$$\hat{\varepsilon}_i = Y_i - \widehat{Y}_i, \quad i = 1, \dots, n$$

Les résidus ne sont pas homoscédastiques, ainsi, au lieu de travailler directement avec les résidus, on travaillera souvent avec les résidus standardisés $\tilde{\varepsilon}_i$ (aussi nommés résidus studentisés en interne, *internally studentized residuals*).

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}, \quad i = 1, \dots, n$$

avec h_{ii} le levier de i^e point introduit dans la section précédente, et avec $\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \varepsilon_i^2$ tel qu'introduit précédement.

Ainsi les résidus standardisés seront étudiés pour vérifier l'homoscédasticité et la normalité.

Si les hypothèses du modèle sont vérifiées alors tous les résidus doivent suivre la même loi (en particulier la dispersion ne doit pas changer), et celle-ci doit être une loi $\mathcal{N}(0,1)$.

Remarque 7.3. *Formellement les résidus standardisés doivent suivre une loi $\mathcal{T}_{n-2}$ puisqu'on utilise $\hat{\sigma}^2$. Toutefois on observe que n points. Quand n est faible, le nombre de points est faible et il est très difficile de faire la différence entre une loi de Student et une loi $\mathcal{N}(0,1)$; quand n est elevé, la loi $\mathcal{T}_{n-2}$ converge vers une loi normale. En pratique on comparera les résidus standardisés à une loi normale.*

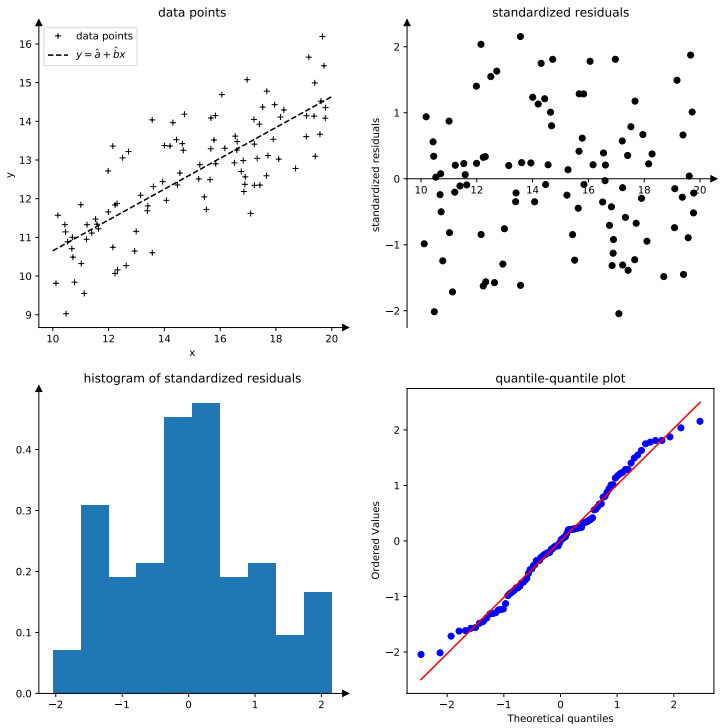


FIGURE 7.3 – Exemple de regression linéaire et de l'analyse des résidus standardisés. Au nord-est on observe que la dispersion de semble pas varier selon x, ce qui confirme l'hypothèse d'homoscédasticité. Au sud-ouest on observe que les résidus standardisés suivent une loi qui pourrait être une loi normale. Au sud-est on observe qu'il peut être acceptable de considérer que les résidus suivent une loi normale.

Diagnostic On vérifiera :

l'homoscédasticité en traçant les résidus standardisés en fonction de l'abscisse des données. On ne doit pas observer de variation dans la dispersion. Le graphique au nord-est de la figure 7.3 est un exemple de ce tracé.

la normalité en traçant soit un histogramme des résidus standardisés ou de manière non exclusive diagramme quantile–quantile (*cf.* 11.2. Le graphique au sud-ouest de la figure 7.3 représente l'histogramme des résidus standardisés, et le graphique au sud-est représente le diagramme quantile-quantile. On préférera utiliser la majorité du temps un diagramme quantile-quantile pour un diagnostic graphique de la normalité. On pourra également utiliser un test statistique pour cela (le test de Shapiro-Wilk, abordé en section 11.3, page 156).

Avec `GNU R`, on peut extraire les résidus standardisés de l'objet `regression` avec `rstandard` et tracer ces graphiques :

```r
regression <- lm(y~1+x, data=D)
qqnorm(rstandard(regression))
plot(D[['x']], rstandard(regression))
```

Avec `statsmodels`, on peut extraire les résidus standardisés à partir de l'attribut `.resid_studentized_internal` de l'influence obtenue à partir du résultat. Puis on peut tracer ce résultat.

```python
import statsmodels.formula.api as smf
from scipy import stats
import matplotlib.pyplot as plt

model = smf.ols(formula="y~1+x", data=D)
res = model.fit()
infl = res.get_influence()
standard_residuals = infl.resid_studentized_internal

stats.probplot(standard_residuals, dist='norm', plot=plt)
plt.show()

plt.plot(D['x'], stardard_residuals, '+')
plt.show()
```

Détection des points anormaux

Pour détecter les points anormaux, on pourrait utiliser le résidu standardisé. Toutefois, si un point est anormal, l'estimateur $\hat{\sigma}^2$ est faussé par le point anormal. Pour quantifier les points anormaux, on va utiliser les résidus studentisés $\tilde{\varepsilon}_i^\star$ (aussi nommés résidus studentisés en externe, *externally studentized residuals*) :

$$\tilde{\varepsilon}_i^\star = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2_{[i]}(1-h_{ii})}}, \quad i = 1, \dots, n$$

où h_{ii} est le levier du point i et où $\hat{\sigma}^2_{[i]}$ est l'estimateur de la variance obtenu à partir des résidus sauf le i^{e} résidus :

$$\hat{\sigma}^2_{[i]} = \frac{1}{n-3} \sum_{k=1}^{n} \mathbb{1}_{k \neq i}\, \varepsilon_k^2$$

La prise en compte d'une donnée aberrante aussi bien dans le résidu que dans la variance estimée, peut conduire à un résidu standardisé non extrême, ce qui ne permettra pas la détection de cette donnée aberrante. Les résidus studentisés, eux, ne souffrent pas de ce défaut, et c'est donc eux qui doivent être utilisés pour la détection de données aberrantes.

Si toutes les hypothèses du modèle sont vérifiées et qu'il n'y a pas de données aberrantes alors les résidus studentisés suivent une loi $\mathcal{T}_{n-3}$.

On peut donc dire que :

$$\mathbb{P}\left(\exists i,\ \left|\tilde{\varepsilon}_i^\star\right| > t_{n-3,1-\alpha/2n}\right) \leq \alpha$$

(Ce résultat est une conséquence directe de l'inégalité de Boole.)

Pour un α faible, on peut donc dire qu'il est improbable qu'il y ait un point tel que $\left|\tilde{\varepsilon}_i^\star\right| > t_{n-3,1-\alpha/2n}$. On utilisera donc le seuil $t_{n-3,1-\alpha/2n}$ pour détecter les points aberrants (souvent avec $\alpha = 5\%$).

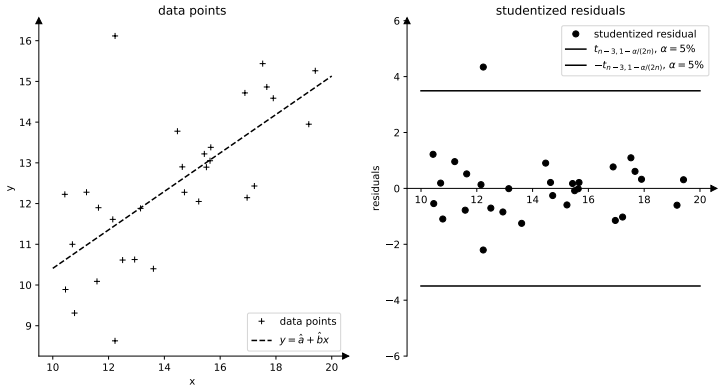


FIGURE 7.4 – Exemple de régression linéaire et de l'analyse des résidus studentisés. À gauche on observe les points originaux et la droite de régression, et un point semble suspect. À droite, on observe les résidus studentisés et les barres placées à $t_{n-3,1-\alpha/2n}$ et $-t_{n-3,1-\alpha/2n}$ pour $\alpha = 5\%$. On observe que selon le critère défini, un point peut être considéré comme aberrant.

Diagnostic On tracera les résidus studentisés et on cherchera les résidus studentisés qui ne sont pas dans l'intervalle $\left[-t_{n-3,1-\alpha/2n}, t_{n-3,1-\alpha/2n}\right]$, ces résidus hors de cet intervalle seront considérés comme un indicateur que le point correspondant est possiblement aberrant. Une illustration de diagnostique est présenté à la figure 7.4.

Remarque 7.4. *Certaines approches plus pratiques proposent de comparer les résidus studentisés à une même valeur de référence souvent* 3 *ou* 3.5. *On préfèrera calculer le seuil* $t_{n-3,1-\alpha/2n}$ *quand c'est possible.*

Avec `GNU R`, on peut extraire les résidus studentisés de l'objet `regression` avec `rstandard` et tracer le graphique avec les seuils :

```r
regression <- lm(y~1+x, data=D)

seuil <- qt(1-.05/2/nrow(D), df=nrow(D)-3)
residuals <- rstudent(regression)

plot_ylim <- range(c(residuals, seuil, -seuil))
plot(D[['x']], residuals, ylim=plot_ylim)
abline(h=seuil, col='red')
abline(h=-seuil, col='red')
```

Avec `statsmodels`, on peut extraire les résidus studentisés à partir de l'attribut `.resid_studentized_external` de l'influence obtenue à partir du résultat. Puis on peut tracer ce résultat.

```python
import statsmodels.formula.api as smf
from scipy import stats
import matplotlib.pyplot as plt

model = smf.ols(formula="y~1+x", data=D)
res = model.fit()
infl = res.get_influence()
studentized_residuals = infl.resid_studentized_external

seuil = stats.t(len(D['x'])-3).ppf(1-.05/2/len(D['x']))

plt.plot(D['x'], studentized_residuals, 'b+')
spanx = np.min(D['x']),np.max(D['x'])
plt.plot(spanx, seuil*np.ones(2), 'r-')
plt.plot(spanx, -seuil*np.ones(2), 'r-')
plt.show()
```

7.5.3 Tableau d'analyse de la variance

La plupart des logiciels de statistique présentent les résultats de la régression à l'aide d'un *tableau d'analyse de la variance* (voir table 7.1).

TABLE 7.1 – Tableau d'analyse de la variance de la régression linéaire

Source de variation	degrés de liberté	Somme des carrés	Moyenne des carrés	
Régression	1	SSR	$MSR = SSR$	$\frac{MSR}{MSE}$
Résiduelle	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
total	$n-1$	SST		

Les éléments présentés dans le tableau sont les suivants :
- SST (*Sum of Squares Total*) est la dispersion totale : $SST = nS_Y^2$;
- SSR (*Sum of Squares Regression*) est la dispersion due à la régression : $SSR = nS_{reg}$;
- SSE (*Sum of Squares Error*) est la dispersion résiduelle : $SSE = nS_{res}$;

— Le rapport $\frac{MSR}{MSE}$ est la statistique qui sera utilisée pour *tester la non nullité* de la pente b : il s'agit alors de mettre en œuvre un *test statistique* (voir chapitre 8), c'est-à-dire une stratégie permettant de déterminer, au vu des données observées, si l'on peut ou non considérer que b est nul. Le test de significativité des coefficients de la régression sera abordé au chapitre 9, paragraphe 9.1.2.

Les informations sur les sommes des carrés (et sur les tests effectués à partir d'icelles qui seront traités dans un chapitre ultérieur) sont accessibles dans `GNU R` à partir de l'objet `regression` via la méthode `anova` ou la méthode générale `summary` :

```
regression <- lm(y~1+x, data=D)
anova(regression)
summary(regression)
```

Les informations sur les sommes des carrés (et sur les tests effectués à partir d'icelles qui seront traités dans un chapitre ultérieur) sont accessibles avec `statsmodels` à partir de l'objet `res` via la méthode `anova_lm` (il est à noter que la méthode `.summary()` donne déjà beaucoup d'informations) :

```
import statsmodels.formula.api as smf
import statsmodels.api as sm

model = smf.ols(formula="y~1+x", data=D)
res = model.fit()
res.summary()

sm.stats.anova_lm(res)
```

7.5.4 Mesure de l'ajustement

On mesure la qualité de l'ajustement de la régression par le *coefficient de détermination*

$$R^2 = \frac{S_{reg}}{S_Y^2},$$

qui est une quantité variant entre 0 et 1. Une valeur proche de 1 indiquera que la proportion de variance non expliquée par le modèle est faible et que la régression est donc très explicative.

Le coefficient de détermination R^2 ne mesure pas la qualité de la modélisation. Un faible R^2 signifie uniquement que la regression linéaire explique une faible part du phénomène en terme de variance.

La fonction `summary` permet d'obtenir le coefficient de détermination :

```
summary(regression)[['r.squared']]
```

La méthode `.summary()` permet d'obtenir le coefficient de détermination (entre autres informations) ou celui-ci peut être obtenu directement au moyen de l'attribut `.rsquared`

```
res.summary()
res.rsquared
```

7.5.5 Intervalles de confiance sur a et b

À partir des résultats précédents, on peut obtenir les deux fonctions pivotales pour a et b qui vont permettre de déterminer des intervalles de confiance :

Proposition 7.7.

$$\frac{\hat{a}-a}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_x^2}}} \sim \mathcal{T}_{n-2} \quad et \quad \frac{\hat{b}-b}{\frac{\hat{\sigma}}{\sqrt{ns_x^2}}} \sim \mathcal{T}_{n-2}.$$

7.5.6 Prévision

Si le modèle de régression semble valide, il est possible de l'utiliser pour « prédire » la valeur de la variable $Y_0 = a + bx_0$ correspondant à une nouvelle valeur x_0 de la variable indépendante x. Cette prédiction peut prendre la forme

1. d'une estimation ponctuelle de $\mathbb{E}(Y_0)$;
2. d'un intervalle de confiance sur $\mathbb{E}(Y_0)$;
3. d'un intervalle de prédiction, défini comme un intervalle contenant Y_0 avec une probabilité donnée $1-\alpha$.

Estimation ponctuelle de $\mathbb{E}(Y_0)$. Soit

$$\widehat{Y}_0 = \hat{a} + \hat{b}x_0.$$

On a

$$\mathbb{E}(\widehat{Y}_0) = \mathbb{E}(\hat{a}) + \mathbb{E}(\hat{b})x_0 = a + bx_0 = \mathbb{E}(Y_0).$$

$\widehat{Y}_0$ est donc un estimateur sans biais de $\mathbb{E}(Y_0)$.

Intervalle de confiance sur $\mathbb{E}(Y_0)$. Il peut être obtenu en cherchant une fonction pivotale. Tout d'abord, $\widehat{Y}_0$, combinaison linéaire de v.a. gaussiennes ($\hat{a}$ et $\hat{b}$) suit une loi gaussienne. De plus, on a vu que $\mathbb{E}(\widehat{Y}_0) = \mathbb{E}(Y_0)$, et

$$\begin{aligned}
\operatorname{Var}(\widehat{Y}_0) &= \operatorname{Var}(\hat{a} + \hat{b}x_0) = \operatorname{Var}(\hat{a}) + x_0^2 \operatorname{Var}(\hat{b}) + 2x_0 \operatorname{Cov}(\hat{a}, \hat{b}) \\
&= \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right) + x_0^2 \frac{\sigma^2}{ns_x^2} - 2x_0 \frac{\overline{x}\sigma^2}{ns_x^2} \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2 - 2\overline{x}x_0 + x_0^2}{ns_x^2}\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}\right).
\end{aligned}$$

On a donc :

$$\frac{\widehat{Y}_0 - \mathbb{E}(Y_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}} \sim \mathcal{N}(0,1), \quad \text{et} \quad \frac{\widehat{Y}_0 - \mathbb{E}(Y_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}} \sim \mathcal{T}_{n-2},$$

ce qui est une fonction pivotale pour $\mathbb{E}(Y_0)$. On en déduit un intervalle de confiance bilatéral de niveau $1-\alpha$:

$$\left[\widehat{Y}_0 - t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}, \widehat{Y}_0 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}\right]. \tag{7.1}$$

Remarque 7.5. *Les bornes de cet intervalle de confiance sont situées sur des branches d'hyperboles de part et d'autre de la droite des moindres carrés. La longueur de l'IC est minimale pour $x_0 = \overline{x}$, et croît avec $|x_0 - \overline{x}|$.*

Un intervalle de confiance pour de nouvelles données peut être obtenu dans GNU R à partir de l'objet `regression`. On construira tout d'abord un `data.frame` contenant les nouvelles données (une seule valeur x_0 dans ce qui vient d'être traité), et on utilisera la fonction `predict` :

```
new_data <- data.frame(x = 24)
predict(regression, new_data, interval='confidence')
```

Intervalle de prédiction. On cherche deux bornes A et B telles que $\mathbb{P}(A \leq Y_0 \leq B) = 1-\alpha$. Pour cela, calculons l'espérance et la variance de $\widehat{Y}_0 - Y_0$. On a

$$\mathbb{E}(\widehat{Y}_0 - Y_0) = \mathbb{E}(\widehat{Y}_0) - \mathbb{E}(Y_0) = 0,$$

et, de par l'indépendance de $\widehat{Y}_0$ et Y_0 :

$$\begin{aligned}\mathrm{Var}(\widehat{Y}_0 - Y_0) &= \mathrm{Var}(\widehat{Y}_0) + \mathrm{Var}(Y_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}\right) + \sigma^2 \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}\right).\end{aligned}$$

Notons σ_p^2 cette variance. On a donc

$$\frac{\widehat{Y}_0 - Y_0}{\sigma_p} \sim \mathcal{N}(0,1).$$

En remplaçant σ_p par $\hat{\sigma}_p = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{ns_x^2}}$, on obtient

$$\frac{\widehat{Y}_0 - Y_0}{\hat{\sigma}_p} \sim \mathcal{T}_{n-2}.$$

On peut alors en déduire l'intervalle de prédiction pour la valeur de Y_0 :

$$\left[\widehat{Y}_0 - t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_p, \widehat{Y}_0 + t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}_p\right].$$

Remarque 7.6. *Cet intervalle est strictement plus large que l'intervalle de confiance de l'équation (7.1). En effet, il tient compte non seulement de l'incertitude de l'estimation des coefficients de la droite de régression, mais aussi de la variabilité de Y_0 autour de son espérance.*

Un intervalle de prédiction pour de nouvelles données peut être obtenu dans `GNU R` à partir de l'objet `regression`. On construira tout d'abord un `data.frame` contenant les nouvelles données (une seule valeur x_0 dans ce qui vient d'être traité), et on utilisera la fonction `predict` :

```
new_data <- data.frame(x = 24)
predict(regression, new_data, interval='prediction')
```

7.6 Exemple

Nous reprenons maintenant l'exemple initial.

7.6.1 Détermination de la droite de régression estimée

On obtient $\overline{x} = 70.08$, $\overline{y} = 309.33$, $s_x^2 = 1997$, $S_Y^2 = 33471$, $S_{xY} = 7958$, $\hat{a} = 30.09$ et $\hat{b} = 3.984$. L'équation de la droite des moindres carrés est donc $\hat{y} = 30.09 + 3.984x$ et le coefficient de détermination est $R^2 = 0.947$, ce qui indique que le prix d'un appartement est en grande partie expliqué par sa surface.

7.6.2 Analyse de la variance

Source de variation	degrés de liberté	Somme des carrés	Moyenne des carrés	
Régression	1	761009	761009	396
Résiduelle	22	42292	1922	
total	23	803301		

7.6.3 Prévision

Cherchons maintenant l'intervalle de prédiction à 95 % de la valeur y_0 correspondant à $x_0 = 100$:

$$\hat{\sigma} = \frac{nS_{res}}{n-2} = 43.845$$

$$\widehat{Y_0} = 428.53$$

$$\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_x^2}} = 1.03$$

$$t_{n-2,0.975} = 2.074$$

d'où l'intervalle de prédiction $[335, 522]$, ce qui peut sembler assez large, bien que le coefficient de détermination soit relativement proche de 1.

7.7 Démonstrations

7.7.1 Preuve de la Proposition 7.1

Remarquons tout d'abord que S_{xY} étant fonction linéaire des Y_i, il en est de même pour $\hat{a}$ et $\hat{b}$. Par conséquent, $\hat{a}$ et $\hat{b}$ suivent des distributions normales (d'après la Proposition 3.18).

Montrons maintenant que $\hat{a}$ et $\hat{b}$ sont des estimateurs sans biais de a et b. On a :

$$\begin{aligned}\mathbb{E}(\hat{b}) = \mathbb{E}\left(\frac{S_{xY}}{s_x^2}\right) &= \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})\mathbb{E}(Y_i - \overline{Y})}{s_x^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(a + bx_i - a - b\overline{x})}{s_x^2} \\ &= \frac{b\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2}{s_x^2} = b.\end{aligned}$$

et

$$\mathbb{E}(\hat{a}) = \mathbb{E}(\overline{Y}) - \mathbb{E}(\hat{b})\overline{x} = a + b\overline{x} - b\overline{x} = a.$$

Avant de calculer les moments d'ordre 2 de $\hat{a}$ et $\hat{b}$, établissons le lemme suivant.

Lemme 7.1.

$$\mathrm{Cov}(\hat{b}, \overline{Y}) = 0.$$

Preuve. On a

$$\begin{aligned}S_{xY} &= \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) \\ &= \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})Y_i - \frac{\overline{Y}}{n}\sum_{i=1}^n (x_i - \overline{x}) \\ &= \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})Y_i,\end{aligned}$$

d'où

$$\begin{aligned}\mathrm{Cov}\left(\hat{b}, \overline{Y}\right) &= \mathrm{Cov}\left(\frac{S_{xY}}{s_x^2}, \overline{Y}\right) \\ &= \frac{1}{ns_x^2}\,\mathrm{Cov}\left(\sum_{i=1}^n (x_i - \overline{x})Y_i, \overline{Y}\right) \\ &= \frac{1}{ns_x^2}\sum_{i=1}^n (x_i - \overline{x})\,\mathrm{Cov}(Y_i, \overline{Y}).\end{aligned}$$

Sachant que

$$\mathrm{Cov}(Y_i, \overline{Y}) = \frac{1}{n}\sum_{j=1}^n \mathrm{Cov}(Y_i, Y_j) = \frac{1}{n}\,\mathrm{Var}(Y_i) = \frac{\sigma^2}{n}$$

on en déduit finalement :

$$\mathrm{Cov}(\hat{b}, \overline{Y}) = \frac{1}{n s_x^2} \sum_{i=1}^{n} (x_i - \overline{x}) \frac{\sigma^2}{n} = 0.$$

□

Nous pouvons maintenant calculer les moments d'ordre 2 du couple aléatoire $(\hat{a}, \hat{b})$:

$$\begin{aligned}\mathrm{Var}(\hat{b}) = \mathrm{Var}\left(\frac{S_{xY}}{s_x^2}\right) = \mathrm{Var}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{s_x^2}\right) = \frac{\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 \mathrm{Var}(Y_i)}{(s_x^2)^2} \\ = \frac{\sigma^2 s_x^2}{n(s_x^2)^2} = \frac{\sigma^2}{n s_x^2}.\end{aligned}$$

D'autre part, en utilisant le Lemme 7.1, on a :

$$\mathrm{Var}\,(\hat{a}) = \mathrm{Var}\left(\overline{Y} - \hat{b}\overline{x}\right) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{n s_x^2} = \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right).$$

Enfin,

$$\begin{aligned}\mathrm{Cov}\left(\hat{a}, \hat{b}\right) &= \mathrm{Cov}\left(\overline{Y} - \hat{b}\overline{x}, \hat{b}\right) \\ &= \mathrm{Cov}\left(\overline{Y}, \hat{b}\right) - \overline{x}\,\mathrm{Cov}\left(\hat{b}, \hat{b}\right) \\ &= 0 - \overline{x}\,\mathrm{Var}(\hat{b}) \\ &= -\frac{\overline{x}\sigma^2}{n s_x^2}.\end{aligned}$$

7.7.2 Preuve de la Proposition 7.3

$$S_{reg} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\overline{x})^2 = \frac{1}{n}\hat{b}^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 = \hat{b}^2 s_x^2$$

$$S_{res} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2 = \hat{\sigma}_{MV}^2$$

$$(Y_i - \overline{Y}) = (a + bx_i + \varepsilon_i - a - b\overline{x} - \overline{\varepsilon}) = b(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = nb^2 s_x^2 + \sum_i \varepsilon_i^2 - n\overline{\varepsilon}^2 + 2b\sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}).$$

Sachant que

$$\mathbb{E}\left(\sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})\right) = \sum_{i=1}^{n}(x_i - \overline{x})\mathbb{E}(\varepsilon_i - \overline{\varepsilon}) = 0$$

on obtient

$$\mathbb{E}(S_Y^2) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) = \frac{1}{n}b^2 s_x^2 + \sigma^2 - \frac{\sigma^2}{n} = \frac{1}{n}(nb^2 s_x^2 + (n-1)\sigma^2)$$

et

$$\mathbb{E}(S_{reg}) = \mathbb{E}\left(\hat{b}^2 s_x^2\right) = s_x^2(\mathrm{Var}(\hat{b}) + \mathbb{E}(\hat{b})^2) = s_x^2\left(\frac{\sigma^2}{ns_x^2} + b^2\right) = \frac{1}{n}(\sigma^2 + ns_x^2 b^2).$$

On en déduit :

$$\mathbb{E}(S_{res}) = \frac{1}{n}\left(nb^2 s_x^2 + (n-1)\sigma^2 - \sigma^2 - ns_x^2 b^2\right) = \frac{n-2}{n}\sigma^2.$$

Chapitre 8

Tests d'hypothèses

8.1 Exemples introductifs

De nombreux problèmes dans les domaines scientifiques et techniques se ramènent à vérifier une hypothèse à propos d'une certaine population. À titre d'exemples, considérons les problèmes suivants :

Exemple 8.1. *Pour une certaine maladie, le pourcentage de guérison spontanée en l'absence de traitement est de 50%. On soumet 100 malades à un nouveau traitement et on observe 60 guérisons. Peut-on conclure à l'efficacité du traitement ?*

Exemple 8.2. *Des études climatologiques sur une longue période ont montré que la température moyenne à Paris au mois d'août peut être modélisée par une v.a. normale X d'espérance 20°C et d'écart-type 1°C. On a observé au cours des 10 dernières années les valeurs suivantes :*

$$22 \quad 19 \quad 21 \quad 23 \quad 20 \quad 22 \quad 21 \quad 18 \quad 20 \quad 22.$$

Ces observations rendent-elles vraisemblable l'hypothèse d'un changement climatique ?

Exemple 8.3. *Dans une étude de fiabilité, on a mesuré la durée de vie en heures d'un échantillon de 100 matériels identiques. On cherche à vérifier si cette durée de vie suit une loi exponentielle.*

Qu'y a-t-il de commun entre ces exemples ? On peut observer que, dans les trois cas, on s'interroge sur la distribution d'une v.a. X pour laquelle on a observé une réalisation d'un échantillon iid $X_1, \dots, X_n$ et on cherche à vérifier une hypothèse relative à cette distribution :

— dans l'exemple 8.1, la v.a. définie de la manière suivante

$$x = \begin{cases} 1 & \text{si le patient guérit} \\ 0 & \text{sinon.} \end{cases}$$

suit une loi de Bernoulli de paramètre θ, taux de guérison après traitement ; l'hypothèse peut alors s'écrire $\theta = 0.5$ (le traitement n'a pas d'effet) ;

— dans l'exemple 8.2, la v.a. X représente la température moyenne à paris au mois d'août, $X \sim \mathcal{N}(\theta, \sigma^2)$ où θ est la température à Paris au mois d'août, moyennée sur une longue période et $\sigma = 1$; l'hypothèse peut alors s'écrire $\theta = 20$ (absence de changement climatique) ;

— enfin, dans l'exemple 8.3, la v.a. X représente la durée de vie et l'hypothèse qui porte cette fois sur la forme elle-même de la distribution peut s'écrire $X \sim \mathscr{E}$.

8.2 Principes de résolution

Notion d'hypothèse Formellement, nous définirons une hypothèse comme une partie de l'espace des distributions sur X. Une hypothèse sera dite *simple* si elle est réduite à un seul élément, c'est-à-dire si $\mathrm{card}(H) = 1$. Dans le cas contraire ($\mathrm{card}(H) > 1$), H est appelée *hypothèse composite* ou *multiple*. Si la loi F_X de X ne dépend que du paramètre θ (test paramétrique) : $F_X \in (F_\theta)_{\theta\in\Theta}$, une hypothèse correspondra à une partie de l'espace Θ du paramètre : $H \subset \Theta$.

Problème de test L'objectif d'un problème de test est de vérifier la cohérence d'une hypothèse relative à la loi d'une v.a. X avec la réalisation d'un échantillon de X. Cette hypothèse sera appelée *hypothèse nulle* et notée H_0.

Hypothèse alternative Il est souvent nécessaire de prendre en compte de quelle façon est envisagé le rejet de l'hypothèse nulle et de préciser une hypothèse alternative que l'on notera H_1 ; dans l'exemple climatique, l'hypothèse alternative peut correspondre à une augmentation de la température. Les deux hypothèses H_0 et H_1 sont *complémentaires* et il est implicitement supposé qu'elles sont les seules envisageables. Dans le cas paramétrique, H_0 et H_1 forment donc une partition de l'espace Θ des paramètres. La spécification de ces hypothèses peut donc amener à écarter certaines valeurs de paramètre.

Région critique, région d'acceptation et statistique de test Un test est une procédure de choix entre deux actions : acceptation ou rejet de l'hypothèse H_0 en fonction des observations de la v.a. X. Cette règle de décision revient donc à partitionner l'espace des observations (ensemble des réalisations possibles de l'échantillon $X_1, \dots, X_n$) en 2 régions : l'ensemble W des observations pour lesquelles on refuse H_0, appelée *région critique* (RC) ; et l'ensemble $\overline{W}$ des observations pour lesquelles on accepte H_0, appelée *région d'acceptation*. Par abus de notation et en fonction du contexte, nous ne distinguerons pas la région critique et sa version aléatoire ; d'ailleurs nous utiliserons la même lettre W pour désigner ces deux quantités.

Critères de performance Ayant pris une décision concernant l'acceptation ou le rejet de H_0, quatre situations peuvent se présenter, selon que l'hypothèse H_0 est vraie ou fausse : ces quatre situations sont résumées dans le tableau 8.1.

TABLE 8.1 – Définition des erreurs de première et de seconde espèce.

vérité / décision	H_0	H_1
H_0	bonne décision	erreur de 2^{e} espèce
H_1	erreur de 1re espèce	bonne décision

Un « bon test » doit maximiser la probabilité de prendre une bonne décision, que l'on se trouve en réalité sous l'hypothèse H_0 ou sous l'hypothèse H_1. Pour fixer les idées, supposons tout d'abord que $H_0 = \{\theta_0\}$ et $H_1 = \{\theta_1\}$ sont des hypothèses simples. La probabilité $\alpha_{\theta_0}(W)$ de commettre une erreur, lorsque H_0 est vraie, est appelée *risque de 1re espèce*. Elle se définit formellement par :

$$\alpha_{\theta_0}(W) = \mathbb{P}_{\theta_0}((X_1, \dots, X_n) \in W) \quad \text{notée} \quad \mathbb{P}_{\theta_0}(W),$$

la notation $\mathbb{P}_{\theta_0}$ signifiant que la probabilité est calculée pour la valeur θ_0 du paramètre. De même, la probabilité $\beta_{\theta_1}(W)$ de commettre une erreur, lorsque H_1 est vraie, est appelée *risque de seconde espèce*; elle se définit par :

$$\beta_{\theta_1}(W) = \mathbb{P}_{\theta_1}((X_1, \dots, X_n) \in \overline{W}) \quad \text{notée} \quad \mathbb{P}_{\theta_1}(\overline{W}).$$

De manière équivalente, on s'intéresse souvent à la probabilité $\pi_{\theta_1}(W)$ d'accepter H_1 si H_1 est vraie, appelée *puissance du test*. On a :

$$\pi_{\theta_1}(W) = \mathbb{P}_{\theta_1}(W) = 1 - \beta_{\theta_1}(W).$$

On notera de manière simplifiée α, β et π ces différentes quantités, lorsqu'aucune confusion ne sera possible. Un « bon » test doit donc minimiser α et β ou, de manière équivalente, minimiser α et maximiser π.

Dans le cas où H_0 est une hypothèse composite, le risque de 1re espèce devient fonction de $\theta \in H_0$. On peut alors caractériser la performance du test sous H_0 par la borne supérieure de $\alpha_\theta(W)$, pour $\theta \in H_0$. Cette borne $\alpha(W)$ est appelée *niveau* du test :

$$\alpha(W) = \sup_{\theta \in H_0} \alpha_\theta(W).$$

De même, si H_1 est une hypothèse composite, le risque de 2e espèce et la puissance deviennent des fonctions de $\theta \in H_1$. On appelle *fonction de puissance* la fonction :

$$\begin{aligned} H_1 &\to [0,1] \\ \theta &\mapsto \pi_\theta(W). \end{aligned}$$

Choix d'un test Nous avons défini au paragraphe précédent non pas un critère d'erreur, mais deux : les risques de 1re espèce et de 2e espèce. Étant donnés deux tests définis par les régions critiques W et W', le test défini par W est *préférable au sens large* à celui défini par W', si :

$$\alpha_\theta(W) \leq \alpha_\theta(W') \quad \forall \theta \in H_0 \quad \text{et} \quad \beta_\theta(W) \leq \beta_\theta(W') \quad \forall \theta \in H_1.$$

Cette approche permet de comparer deux tests, mais elle ne permet pas en général de trouver un test optimal : il n'y a aucune raison, en général, pour que le test qui minimise le risque de première espèce, soit identique au test qui minimise le risque de seconde espèce. Les deux objectifs sont même antagonistes : pour une taille d'échantillon donnée, β varie en sens inverse de α et aucune règle ne pourra minimiser les deux risques. Par exemple, si on prend la décision de ne jamais rejeter l'hypothèse H_0, le risque de première espèce est nul alors que si on prend la décision de toujours rejeter l'hypothèse H_0, c'est le risque de seconde espèce qui devient nul.

Une approche plus pertinente, proposée par Neyman et Pearson en 1933, consiste à traiter les deux risques de façon dissymétrique, en se limitant à la classe des tests dont le risque de 1re espèce est au plus égal à un seuil α^* fixé au préalable, appelé niveau de signification. Cette classe $C(\alpha^*)$ se définit donc formellement comme :

$$C(\alpha^*) = \{W \ / \ \alpha_\theta(W) \leq \alpha^*, \forall \theta \in H_0\}.$$

À l'intérieur de cette famille, on cherche un test de risque de 2e espèce le plus faible possible (ou, de manière équivalente, de puissance la plus élevée possible), idéalement pour toute valeur $\theta \in H_1$. Si un tel test existe, il est dit *uniformément plus puissant* (UPP).

Définition 8.1. *Un test défini par la région critique W^* est dit uniformément plus puissant (UPP) au niveau α^* si :*

- *le test est de niveau α^*, c'est-à-dire $\sup_{\theta \in H_0} \alpha_\theta(W^*) = \alpha^*$;*
- *$\pi_\theta(W^*) \geq \pi_\theta(W), \forall W, \forall \theta \in H_1$.*

Remarque 8.1. *Cette approche introduit une dissymétrie entre les hypothèses H_0 et H_1. En imposant $\alpha_\theta(W) \leq \alpha^*, \forall \theta \in H_0$, on se prémunit contre un abandon trop hâtif de H_0, qui prend donc le statut d'hypothèse privilégiée. Dans la formalisation d'un problème de test, on choisit donc comme hypothèse nulle l'hypothèse « par défaut », que l'on décide de n'abandonner que si les observations sont réellement « concluantes ».*

La notion de test UPP est illustrée aux figures 8.2 et 8.3.

Remarque 8.2. *L'existence d'un test UPP est un cas de figure idéal, qui ne se rencontre que dans les cas les plus simples. Dans beaucoup de situations rencontrées en pratique, il n'existe pas de test UPP. On se contente alors de trouver un élément de $C(\alpha^*)$.*

Statistique de test La région critique s'appuiera souvent sur une fonction de l'échantillon ne dépendant que de l'hypothèse H_0 et, par exemple dans le cas paramétrique, non pas des valeurs particulières que peut prendre le paramètre θ dans H_0. Sous l'hypothèse H_0 cette fonction est donc une statistique qui sera appelée *statistique de test*. Remarquons que dans cette situation, le risque de première espèce ne dépend que de l'hypothèse H_0 et est donc égal au niveau du test.

***p-value* (degré de signification)** L'usage s'est imposé de donner le plus souvent à α^* une valeur fixée arbitrairement à 1‰, 1%, ou 5%. Pour atténuer l'arbitraire lié au choix d'une valeur unique, il est courant de déterminer le résultat du test pour plusieurs niveaux de signification différents, correspondant à des régions critiques emboîtées. Plus on augmente α^*, plus on tend à rejeter l'hypothèse H_0, que l'on finit nécessairement par rejeter au delà d'une d'une certaine valeur $\hat{\alpha}(x_1, \ldots, x_n)$, fonction des observations, appelée *degré de signification du test* ou *p-value* en anglais. La *p-value*, notée $\hat{\alpha}$, est donc le plus petit niveau de signification pour lequel l'hypothèse nulle H_0 est rejetée. On rejettera donc H_0 au niveau α^* si la *p-value* du test est inférieur à α^*. Plus généralement, la *p-value* fournit une mesure de la confiance avec laquelle on peut rejeter H_0 (plus la *p-value* est faible, plus cette confiance est élevée).

Pour calculer la *p-value*, on détaillera quatre cas, dans tous les cas T est la statistique de test, t_{obs} est l'observation de la statistique de test et F_{H_0} est la fonction de répartition sous H_0 :

Cas unilatéral à gauche la région critique prend la forme $\{t < k\}$, on a :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(T \leq t_{\text{obs}}) \\ &= F_{H_0}(t_{\text{obs}})\end{aligned}$$

Cas unilatéral à droite la région critique prend la forme $\{t > k\}$, on a :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(T \geq t_{\text{obs}}) \\ &= 1 - F_{H_0}(t_{\text{obs}}) \qquad \text{(si } T \text{ continue)}\end{aligned}$$

Cas bilatéral symétrique la région critique s'écrit $\{|t| > k\}$, on a :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(|T| \geq |t_{\text{obs}}|) \\ &= 2F_{H_0}(-|t_{\text{obs}}|) \\ &= 2\Big(1 - F_{H_0}(|t_{\text{obs}}|)\Big) \qquad \text{(si } T \text{ continue)}\end{aligned}$$

Cas bilatéral général la région critique prend la forme $W = \{t < k_1\} \cup \{t > k_2\}$ avec k_1 et k_2 de telle manière à rejeter sous H_0 avec la même probabilité de chaque coté [1], on a :

$$\begin{aligned}\hat{\alpha} &= 2\min\Big(\mathbb{P}_{H_0}(T \leq t_{\text{obs}}), \mathbb{P}_{H_0}(T \geq t_{\text{obs}})\Big) \\ &= 2\min\Big(F_{H_0}(t_{\text{obs}}), 1 - F_{H_0}(t_{\text{obs}})\Big) \qquad \text{(si } T \text{ continue)}\end{aligned}$$

.

8.3 Théorème de Neyman-Pearson

8.3.1 Principe

Soit X une v.a. de loi $F_X \in (F_\theta)_{\theta\in\Theta}$, avec $\Theta = \{\theta_0, \theta_1\}$. On considère le problème de test :

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta = \theta_1 \end{cases}.$$

Il s'agit donc d'un problème de comparaison entre deux hypothèses simples. Soit $(X_1, \dots, X_n)$ un échantillon iid de X. Les risques de première et de seconde espèce associés à un test défini par la région critique W sont :

$$\begin{cases} \alpha(W) = \mathbb{P}_{\theta_0}((X_1, \dots, X_n) \in W) = \mathbb{P}_{\theta_0}(W) \\ \beta(W) = \mathbb{P}_{\theta_1}((X_1, \dots, X_n) \in \overline{W}) = \mathbb{P}_{\theta_1}(\overline{W}) \end{cases}.$$

Le théorème de Neyman-Pearson permet de trouver pour ce problème le test le plus puissant (c'est-à-dire de risque de 2e espèce minimum) pour un niveau donné.

Théorème 8.1. *Pour le problème H_0 contre H_1, ayant observé l'échantillon $X_1, \dots, X_n$, la région critique du test le plus puissant au niveau α^* vérifie les conditions suivantes :*

1. $W = \{\frac{L(\theta_1; x_1, \dots, x_n)}{L(\theta_0; x_1, \dots, x_n)} > k\}$ *pour une constante k,*

1. formellement $W = \Big\{t < F^{-1}_{H_0}\Big(\frac{\alpha^*}{2}\Big)\Big\} \cup \Big\{t > F^{-1}_{H_0}\Big(\frac{1-\alpha^*}{2}\Big)\Big\}$,

2. $\alpha(W) = \mathbb{P}_{\theta_0}(W) = \alpha^*$.

Preuve. cf. Section 8.7 □

Remarque 8.3. *Le théorème de Neyman-Pearson permet ainsi de définir une statistique de test* $\frac{L(\theta_1;X_1,\dots,X_n)}{L(\theta_0;X_1,\dots,X_n)}$ *garantissant l'optimalité.*

8.3.2 Application : test sur l'espérance d'une loi normale (σ^2 connu)

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu = \mu_1 \end{cases}$$

Région critique

$$\begin{aligned} \frac{L_1}{L_0} = \frac{L(\mu_1; x_1, \dots, x_n)}{L(\mu_0; x_1, \dots, x_n)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2]}{(2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2]} \\ &= \exp\left[\frac{1}{2\sigma^2}(\sum x_i^2 - 2\mu_0 \sum x_i + n\mu_0^2 - \sum x_i^2 + 2\mu_1 \sum x_i - n\mu_1^2)\right] \\ &= \exp\left[\frac{n}{2\sigma^2}(2\overline{x}(\mu_1 - \mu_0) + \mu_0^2 - \mu_1^2)\right]. \end{aligned}$$

On constate que le rapport des vraisemblances L_1/L_0 ne dépend de l'échantillon qu'au travers de $\overline{x}$. Supposons tout d'abord que $\mu_1 > \mu_0$. Dans ce cas, L_1/L_0 est fonction croissante de $\overline{x}$. Donc la condition $L_1/L_0 > k$ est équivalente à la condition $\overline{x} > c$ pour une certaine constante c.

Les risques de première et de seconde espèce pour cet exemple sont représentés sur la figure 8.1.

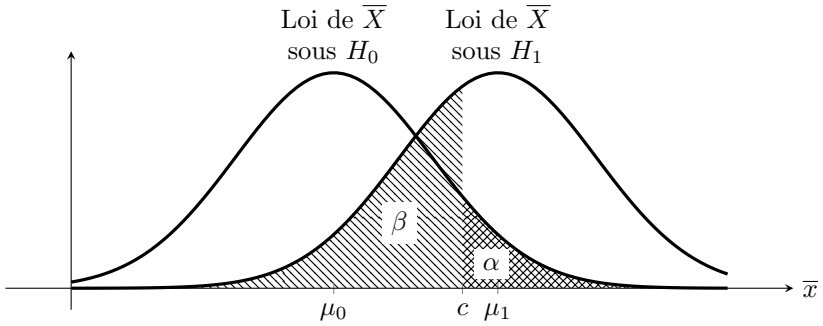


FIGURE 8.1 – Représentation des risques de première et de seconde espèce.

Cherchons quelle doit être la valeur de c pour que le test soit de niveau α^* :

$$\alpha^* = \mathbb{P}_{\mu_0}\left(\overline{X} > c\right) = \mathbb{P}_{\mu_0}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$

$$\Leftrightarrow \frac{c - \mu_0}{\sigma/\sqrt{n}} = \Phi^{-1}(1 - \alpha^*) = u_{1-\alpha^*} \Leftrightarrow c = \mu_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha^*}$$

On en déduit que le test vérifiant :

$$W = \left\{ \overline{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha^*} \right\},$$

est le plus puissant au niveau α^*.

Nous avons obtenu ce résultat en supposant $\mu_1 > \mu_0$. Envisageons maintenant le cas $\mu_1 < \mu_0$. Le rapport de vraisemblance L_1/L_0 est alors fonction décroissante de $\overline{x}$. La condition $L_1/L_0 > k$ est dans ce cas équivalente à la condition $\overline{x} < c$ pour une certaine constante c. On a maintenant :

$$\alpha^* = \mathbb{P}_{\mu_0}\left(\overline{X} < c\right) = \mathbb{P}_{\mu_0}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$\Leftrightarrow c = \mu_0 + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(\alpha^*) \Leftrightarrow c = \mu_0 - \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha^*}.$$

On en déduit cette fois que le test vérifiant :

$$W = \left\{ \overline{x} < \mu_0 - \frac{\sigma}{\sqrt{n}} u_{1-\alpha^*} \right\},$$

est le plus puissant au niveau α^*.

Puissance du test ($\mu_1 > \mu_0$) En utilisant le seuil déjà calculé dans ce cas $c = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha^*}$, on a :

$$\pi_{\mu_1}(W) = \mathbb{P}_{\mu_1}(W) = \mathbb{P}_{\mu_1}(\overline{X} > c) = \mathbb{P}_{\mu_1}\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{c - \mu_1}{\sigma/\sqrt{n}}\right)$$
$$= 1 - \Phi\left(\frac{c - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} - u_{1-\alpha^*}\right).$$

C'est une fonction croissante de μ_1, définie pour $\mu_1 > \mu_0$.

8.4 Test UPP

8.4.1 Principe

Soit un problème de test d'une hypothèse simple $H_0 = \{\theta_0\}$ contre une hypothèse composite H_1 et W la région critique d'un test associé à ce problème. H_1 étant une hypothèse composite, la puissance $\pi_{\theta_1}(W)$ est fonction de θ_1, $\theta_1 \in H_1$.

La notion de test UPP est illustrée aux figures 8.2 et 8.3.

Pour rechercher s'il existe un test UPP pour le problème $H_0 = \{\theta_0\}$ contre $H_1 \subset \Theta$, il suffit de considérer le problème suivant :

$$\begin{cases} H_0 : & \theta = \theta_0 \\ h_1 : & \theta = \theta_1 \end{cases}.$$

avec $\theta_1 \in H_1$. S'agissant d'un problème de comparaison entre deux hypothèses simples, le théorème de Neyman-Pearson permet en général de trouver le test le plus puissant pour ce problème. Soit W_{θ_1} la région critique de ce test ; on a par conséquent $\pi_{\theta_1}(W_{\theta_1}) \geq \pi_{\theta_1}(W)$ pour tout $W \in \mathscr{C}(\alpha^*)$. Lorsque $W_{\theta_1} = W^*$ ne dépend pas de θ_1, on

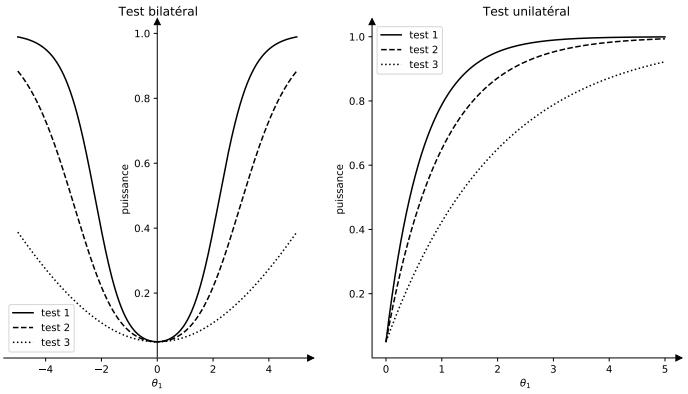


FIGURE 8.2 – Représentation de la puissance dans le cas de trois tests $H_0 = \{\theta = 0\}$ contre à gauche $H_1 = \{\theta \neq 0\}$ et contre à droite $H_1 = \{\theta > 0\}$. En supposant qu'il n'existe que ces trois tests. On remarque que dans le cas bilatéral, le test 1 est plus puissant que tous les autres tests pour toutes les valeurs de θ_1, le test 1 est donc uniformément plus puissant. On obtient la même conclusion dans le cas unilatéral.

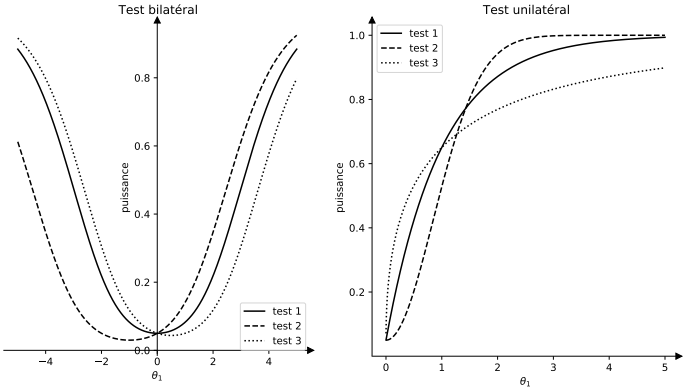


FIGURE 8.3 – Représentation de la puissance dans le cas de trois tests $H_0 = \{\theta = 0\}$ contre à gauche $H_1 = \{\theta \neq 0\}$ et contre à droite $H_1 = \{\theta > 0\}$. En supposant qu'il n'existe que ces trois tests. On remarque que dans le cas bilatéral, que le meilleur test pour certaines valeurs de θ_1 ne l'est pas pour d'autres valeurs de θ_1, on en conclut qu'aucun test est meilleur pour toutes les valeurs de θ_1, il n'existe donc pas de test uniformément plus puissant. On obtient la même conclusion dans le cas unilatéral.

a par conséquent $\pi_{\theta_1}(W^*) \geq \pi_{\theta_1}(W)$ pour tout $W \in \mathscr{C}(\alpha^*)$ et pour tout $\theta_1 \in H_1$. Le test de région critique W^* est alors un test UPP pour le problème H_0-H_1. Au contraire, si W_{θ_1} dépend de θ_1, il n'existe pas un test unique qui soit le plus puissant pour toute valeur $\theta_1 \in H_1$: il n'y a donc pas de test UPP.

8.4.2 Test sur l'espérance d'une loi normale (σ^2 connu)

Test unilatéral

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 connu. Considérons les hypothèses

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

Appliquons la méthode précédente. Soit le problème :

$$\begin{cases} H_0 : & \mu = \mu_0 \\ h_1 : & \mu = \mu_1 \quad \text{avec } \mu_1 > \mu_0 \end{cases}.$$

Nous avons vu précédemment qu'il existe un test optimal pour ce problème, de région critique $W = \{\overline{x} > \mu_0 + (\sigma/\sqrt{n})u_{1-\alpha^*}\}$. Ce test ne dépend pas de μ_1. Par conséquent, il est UPP pour le problème H_0-H_1.

Partant des résultats du paragraphe 8.3.2, la puissance du test s'écrit

$$\pi_\mu(W) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - u_{1-\alpha^*}\right).$$

Test bilatéral

Avec la même loi pour X, considérons maintenant les hypothèses :

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

Cette fois, le test optimal pour le problème $H_0 = \{\mu_0\}$ contre $h_1 = \{\mu_1\}$ avec $\mu_1 \in H_1$ a pour région critique

$$W = \{\overline{x} > \mu_0 + \frac{\sigma}{\sqrt{n}}u_{1-\alpha^*}\},$$

si $\mu_1 > \mu_0$ et

$$W = \{\overline{x} < \mu_0 - \frac{\sigma}{\sqrt{n}}u_{1-\alpha^*}\},$$

si $\mu_1 < \mu_0$. Le test optimal dépend de μ_1. Par conséquent, il n'existe pas de test UPP pour le problème H_0-H_1.

8.5 Test du rapport de vraisemblance

8.5.1 Principe

Il s'agit d'une méthode générale permettant de construire un test ayant généralement de bonnes propriétés, mais non nécessairement UPP.

Définition 8.2. *Soit X une v.a. de loi $F_X \in (F_\theta)_{\theta\in\Theta}$, avec $\Theta \subseteq \mathbb{R}^p$ et soit $X_1, \dots, X_n$ un échantillon iid de X. Soit un problème de test H_0 contre H_1, avec $H_0, H_1 \subset \Theta$ et $\Lambda(x_1, \dots, x_n)$ la statistique suivante appelée* rapport de vraisemblance *:*

$$\Lambda(X_1, \dots, X_n) = \frac{\sup_{\theta\in H_0} L(\theta; X_1, \dots, X_n)}{\sup_{\theta\in\Theta} L(\theta; X_1, \dots, X_n)}.$$

On appelle test du rapport de vraisemblance (RV) *pour les hypothèses H_0-H_1 au niveau α^* le test défini par :*

$$W = \{\Lambda(x_1, \dots, x_n) < c\},$$

où c est une constante déterminée de manière à avoir :

$$\alpha(W) = \alpha^*.$$

Remarque 8.4. *On a nécessairement $\Lambda(x_1, \dots, x_n) \in [0, 1]$.*

Remarque 8.5. *Ce test généralise le test de Neyman-Pearson. En effet, dans le cas où H_0 et H_1 sont des hypothèses simples, on a :*

$$\begin{aligned}\Lambda(x_1, \dots, x_n) &= \frac{L(\theta_0; x_1, \dots, x_n)}{\max(L(\theta_0; x_1, \dots, x_n), L(\theta_1; x_1, \dots, x_n))} \\ &= \frac{1}{\max(1, L(\theta_1; x_1, \dots, x_n)/L(\theta_0; x_1, \dots, x_n))} \\ &= \frac{1}{\max(1, L_1/L_0)} = \min(1, L_0/L_1),\end{aligned}$$

avec $L_1 = L(\theta_1; x_1, \dots, x_n)$ et $L_0 = L(\theta_0; x_1, \dots, x_n)$. La statistique Λ est donc bien fonction décroissante de L_1/L_0.

Remarque 8.6. *Soit $\hat{\theta}$ un EMV de θ et $\hat{\theta}_0$ un EMV de θ pour l'espace des paramètres réduit à H_0 ($\hat{\theta}_0$ est appelé* estimateur du maximum de vraisemblance restreint*). On a :*

$$\Lambda(x_1, \dots, x_n) = \frac{L(\hat{\theta}_0; x_1, \dots, x_n)}{L(\hat{\theta}; x_1, \dots, x_n)}.$$

8.5.2 Application : test sur l'espérance d'une loi normale (σ^2 connu)

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 connu. Considérons les hypothèses

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

On a :

$$\begin{aligned}\Lambda(x_1, \dots, x_n) &= \frac{\sup_{\mu \in H_0} L(\mu; x_1, \dots, x_n)}{\sup_{\mu \in \Theta} L(\mu; x_1, \dots, x_n)} \\ &= \frac{L(\mu_0; x_1, \dots, x_n)}{L(\overline{x}; x_1, \dots, x_n)},\end{aligned}$$

puisque $H_0 = \{\mu_0\}$ est une hypothèse simple et $\overline{X}$ est l'EMV de μ. On a donc :

$$\begin{aligned}
\Lambda &= \frac{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2)}{(2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\overline{x})^2)} \\
&= \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2)}{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\overline{x})^2)} \\
&= \exp\left(\frac{1}{2\sigma^2}\Big(\sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 - \sum x_i^2 + 2\mu_0\sum x_i - n\mu_0^2\Big)\right) \\
&= \exp\left(-\frac{n}{2\sigma^2}(\overline{x}^2 - 2\mu_0\overline{x} + \mu_0^2)\right) \\
&= \exp\left(-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2\right).
\end{aligned}$$

Sachant que sous H_0, $\overline{X} - \mu_0 \sim \mathcal{N}(0, \sigma^2/n)$, on pose

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}.$$

On a donc : $\Lambda(x_1, \dots, x_n) < c \Leftrightarrow |z| > k$ pour une constante k. Cherchons la valeur de cette constante pour avoir un test de niveau α^* :

$$\mathbb{P}_{\mu_0}(|Z| > k) = \alpha^*$$

$$\Leftrightarrow \mathbb{P}_{\mu_0}(Z > k) + \mathbb{P}_{\mu_0}(Z < -k) = \alpha^*.$$

Or, sous H_0, $Z \sim \mathcal{N}(0,1)$: les deux probabilités du membre de gauche sont donc égales. On a par conséquent :

$$\mathbb{P}_{\mu_0}(Z > k) = \frac{\alpha^*}{2} \Leftrightarrow 1 - \Phi(k) = \frac{\alpha^*}{2} \Leftrightarrow k = u_{1-\frac{\alpha^*}{2}}.$$

On obtient donc la région critique :

$$W = \{|z| > u_{1-\frac{\alpha^*}{2}}\}.$$

La fonction puissance peut être calculée de la façon suivante :

$$\begin{aligned}
\pi_\mu(W) &= \mathbb{P}_\mu(W) = \mathbb{P}_\mu(|Z| > u_{1-\frac{\alpha^*}{2}}) \\
&= \mathbb{P}_\mu\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > u_{1-\frac{\alpha^*}{2}}\right) + \mathbb{P}_\mu\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -u_{1-\frac{\alpha^*}{2}}\right) \\
&= \mathbb{P}_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + u_{1-\frac{\alpha^*}{2}}\right) + \mathbb{P}_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha^*}{2}}\right) \\
&= \Phi\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha^*}{2}}\right) + \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha^*}{2}}\right).
\end{aligned}$$

8.5.3 Test approché

Lorsqu'il n'est pas possible d'obtenir la région critique de manière exacte, comme dans l'exemple précédent, le théorème suivant (théorème de Wilks) permet d'obtenir une région critique approchée.

Théorème 8.2. *Soit l'hypothèse H_0 suivante :*

$$H_0 \ : \ (\theta_1, \dots, \theta_r) = (\theta_{10}, \dots, \theta_{r0}),$$

avec $r \leq p$. Sous certaines conditions de régularité, la statistique $-2 \ln \Lambda$ est asymptotiquement pivotale pour les paramètres $\theta_1, \dots, \theta_r$ et suit asymptotiquement, sous H_0, une loi du χ^2_r.

La région critique du test du RV est donc $W = \{-2 \ln \Lambda \geq c\}$ avec $c \simeq \chi^2_{r;1-\alpha^*}$. Nous verrons par la suite des applications de ce théorème.

8.6 Stratégie empirique

Lorsque les techniques optimales précédentes ne permettent pas de conclure, il est souvent possible de construire un test utile en s'appuyant sur une statistique de test T vérifiant les propriétés suivantes :

- La distribution de T doit être connue, au moins approximativement, sous l'hypothèse H_0 ;
- La valeur t de cette statistique doit permettre de mesurer l'écart entre l'échantillon et l'hypothèse H_0 ; par exemple, en prenant une statistique qui sera d'autant plus grande que l'échantillon s'écartera de l'hypothèse H_0.

Dans les tests paramétriques, une telle statistique pourra être obtenue à partir d'une fonction pivotale. Cette statistique de test permet alors de définir la forme de la région critique et la région critique elle-même est déterminée de manière à avoir un test de niveau de signification donné.

Cette stratégie ne garantit pas l'optimalité des tests mais peut conduire, si le choix de la statistique de test est judicieux, à de bonnes solutions. En particulier, dans les cas simples traités précédemment, on peut retrouver par cette approche les tests optimaux.

8.7 Preuve du Théorème de Neyman-Pearson

Nous nous contenterons de démontrer le théorème de Neyman-Pearson dans le cas où X est une v.a. absolument continue de densité $f(x, \theta)$. Le cas discret se démontre de manière similaire, en remplaçant les intégrales par des sommes.

Soit W la région critique associée au test le plus puissant et W' la région critique d'un autre test avec $\alpha(W') = \alpha^*$. Montrons que $\pi(W') < \pi(W)$. Posons $U = W \cap W'$, $V = W \setminus U$, $V' = W' \setminus U$.

1. On a

$$\mathbb{P}_{\theta_0}(W) = \mathbb{P}_{\theta_0}(W') = \alpha^*$$

$$\Rightarrow \mathbb{P}_{\theta_0}(V) + \mathbb{P}_{\theta_0}(U) = \mathbb{P}_{\theta_0}(V') + \mathbb{P}_{\theta_0}(U) \Rightarrow \mathbb{P}_{\theta_0}(V) = \mathbb{P}_{\theta_0}(V')$$

$$\Leftrightarrow \int_V f(x_1,\dots,x_n;\theta_0)\,dx_1\cdots dx_n = \int_{V'} f(x_1,\dots,x_n;\theta_0)\,dx_1\cdots dx_n$$
$$\Leftrightarrow \int_V L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n = \int_{V'} L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n.$$

2. Exprimons la puissance de chacun des deux tests :

$$\begin{array}{ccccc} \pi(W) & = & \mathbb{P}_{\theta_1}(W) & = & \mathbb{P}_{\theta_1}(V)+\mathbb{P}_{\theta_1}(U) \\ \pi(W') & = & \mathbb{P}_{\theta_1}(W') & = & \mathbb{P}_{\theta_1}(V')+\mathbb{P}_{\theta_1}(U). \end{array}$$

On a donc

$$\begin{aligned} \Delta\pi = \pi(\varphi') - \pi(\varphi) &= \mathbb{P}_{\theta_1}(V') - \mathbb{P}_{\theta_1}(V) \\ &= \int_{V'} f(x_1,\dots,x_n;\theta_1)\,dx_1\cdots dx_n - \int_V f(x_1,\dots,x_n;\theta_1)\,dx_1\cdots dx_n \\ &= \int_{V'} L(\theta_1;x_1,\dots,x_n)\,dx_1\cdots dx_n - \int_V L(\theta_1;x_1,\dots,x_n)\,dx_1\cdots dx_n; \end{aligned}$$

or $\forall (x_1,\dots,x_n)\in V \quad \frac{L(\theta_1;x_1,\dots,x_n)}{L(\theta_0;x_1,\dots,x_n)} > k$, donc

$$\int_V L(\theta_1;x_1,\dots,x_n)\,dx_1\cdots dx_n > k\times\int_V L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n.$$

Par ailleurs, $\forall (x_1,\dots,x_n)\in V' \quad \frac{L(\theta_1;x_1,\dots,x_n)}{L(\theta_0;x_1,\dots,x_n)} \leq k$ car $V'\subset \overline{W}$, donc

$$\int_{V'} L(\theta_1;x_1,\dots,x_n)\,dx_1\cdots dx_n \leq k\times\int_{V'} L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n.$$

En utilisant les deux derniers résultats du second point dans la définition de $\Delta\pi$, puis en utilisant le premier point de la démonstration, on obtient :

$$\Delta\pi \leq k\times\int_{V'} L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n - k\times\int_V L(\theta_0;x_1,\dots,x_n)\,dx_1\cdots dx_n \leq 0.$$

Le test de région critique W' est donc bien moins puissant que le test associé à la région critique W.

Chapitre 9

Tests de conformité

Un *flowchart* est présenté figure 9.1 permettant d'appréhender les cas d'usages des différents tests de conformité.

9.1 Tests relatifs à l'espérance d'une loi normale

On cherche à tester si l'espérance μ de la v.a. X, que l'on suppose normale, est égale à une certaine constante μ_0.

9.1.1 Variance connue

Test bilatéral

Proposition 9.1. *Soit $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$ d'espérance μ inconnue et de variance σ^2 connue. Le test du rapport de vraisemblance pour les hypothèses :*

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

a pour région critique :

$$W = \left\{ \frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} > u_{1-\frac{\alpha^*}{2}} \right\}.$$

La fonction puissance s'écrit

$$\pi_\mu(W) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha^*}{2}} \right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha^*}{2}} \right).$$

Ce résultat a été obtenu dans le chapitre précédent.

Tests unilatéraux

Proposition 9.2. *Soit $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$ d'espérance μ inconnue et de variance σ^2 connue. Le test UPP pour les hypothèses :*

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

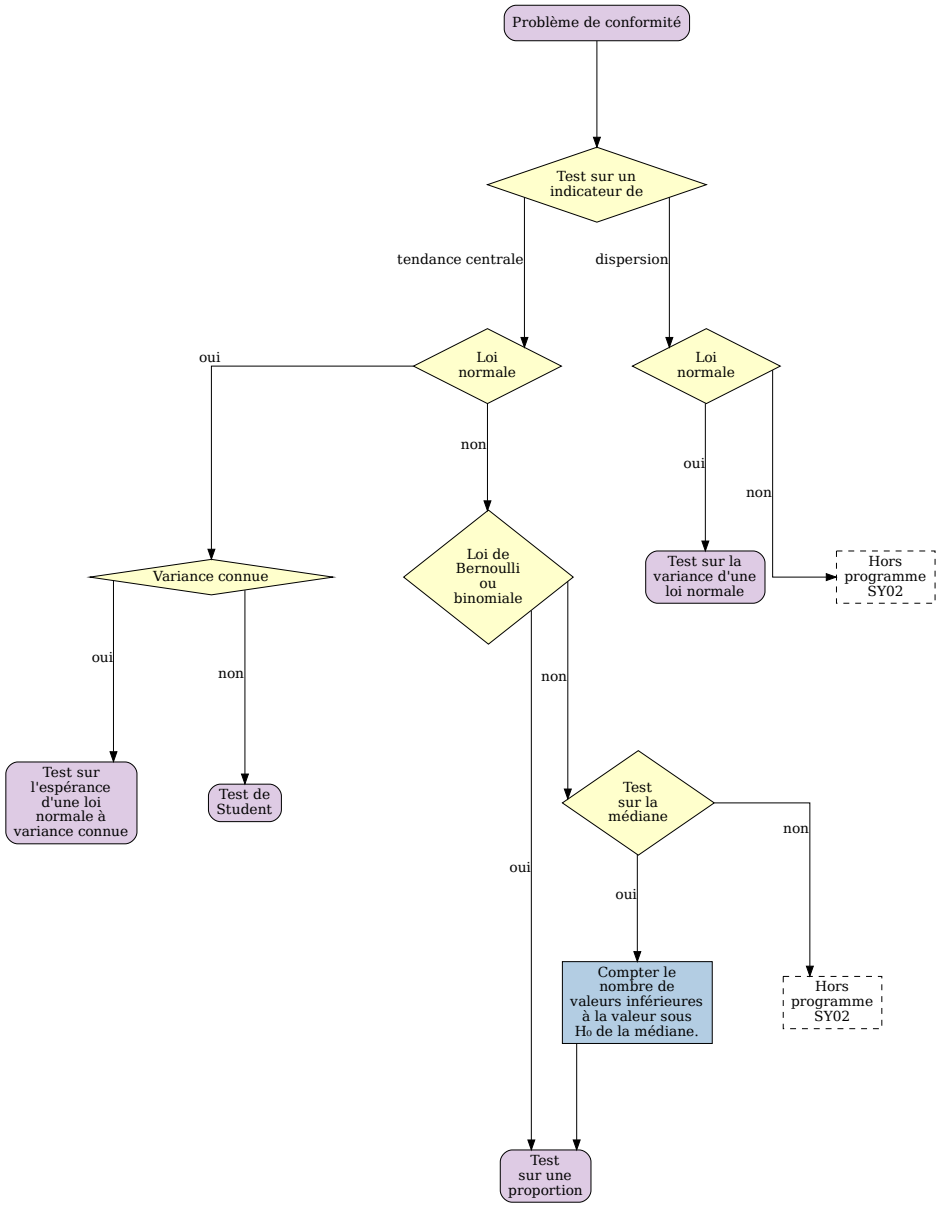


FIGURE 9.1 – *Flowchart* représentant les différents tests de conformité et les cas d'usages. Ce diagramme est donné à titre informatif, il ne remplace pas une lecture précise des conditions d'application de chaque test.

a pour région critique :

$$W = \left\{ \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > u_{1-\alpha^*} \right\}$$

et la fonction puissance s'écrit

$$\pi_\mu(W) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - u_{1-\alpha^*} \right).$$

Ce résultat a été obtenu dans le chapitre précédent.

Pour le problème :

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

on obtient de manière similaire la région critique :

$$W = \left\{ \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -u_{1-\alpha^*} \right\}.$$

9.1.2 Variance inconnue (test de Student)

Test bilatéral

Proposition 9.3. *Soit $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 inconnues. Le test du rapport de vraisemblance pour les hypothèses :*

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

a pour région critique :

$$W = \left\{ \frac{|\overline{x} - \mu_0|}{s^*/\sqrt{n}} > t_{n-1,1-\frac{\alpha^*}{2}} \right\}.$$

Preuve. cf. Section 9.5.1 □

La fonction puissance de ce test, égale à :

$$\pi_\mu(\varphi) = \mathbb{P}_\mu\left(\frac{|\overline{x} - \mu_0|}{s^*/\sqrt{n}} > t_{n-1,1-\frac{\alpha^*}{2}} \right)$$

peut être calculé à partir des fractiles de la loi de Student décentrée (définition 3.16). En effet, lorsque l'espérance de X est égale à μ, on a

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right) \Rightarrow \overline{X} - \mu_0 \sim \mathcal{N}\left(\mu_1 - \mu_0, \frac{\sigma^2}{n}\right),$$

d'où

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right),$$

et

$$\frac{\overline{X} - \mu_0}{s^*/\sqrt{n}} = \frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)s^{*2}}{\sigma^2}}{n-1}}} \sim \mathcal{T}_{n-1}(\delta), \quad \text{avec } \delta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}.$$

Il existe des abaques permettant de déterminer la puissance de ce test, en fonction de n et de δ. Ce calcul peut être également réalisé à l'aide d'un logiciel comme GNU R.

Le test de Student est présent dans GNU R dans la fonction `t.test`. Par exemple, avec un échantillon contenu dans `ech` et $\mu_0 = 12$ on utilisera :

```
t.test(ech, mu=12)
```

Le test du Student est présent dans `scipy` avec la fonction `ttest_1samp` du module `stats`. Par exemple avec un échantillon contenu dans `ech`, et avec $\mu_0 = 12$ on utilisera :

```
from scipy import stats

stats.ttest_1samp(ech, popmean=12)
```

Tests unilatéraux

Proposition 9.4. *Soit $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 inconnues. Le test du rapport de vraisemblance pour les hypothèses :*

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

a pour région critique :

$$W = \left\{ \frac{\overline{x} - \mu_0}{s^*/\sqrt{n}} > t_{n-1,1-\alpha^*} \right\}.$$

Preuve. cf. Section 9.5.2. □

Pour le problème :

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

on obtient de manière similaire la région critique :

$$W = \left\{ \frac{\overline{x} - \mu_0}{s^*/\sqrt{n}} < -t_{n-1,1-\alpha^*} \right\}.$$

La fonction `t.test` de GNU R gère les tests unilatéraux via l'argument `alternative`. Par exemple dans le cas où l'on teste l'hypothèse nulle « $\mu = 12$ » contre l'hypothèse alternative « $\mu < 12$ », on utilisera :

```
t.test(ech, mu=12, alternative='less')
```

La fonction `ttest_1samp` du module `stats` de `scipy` gère les tests unilatéraux via l'argument `alternative`. Par exemple dans le cas où l'on teste l'hypothèse nulle « $\mu = 12$ » contre l'hypothèse alternative « $\mu < 12$ », on utilisera :

```
from scipy import stats

stats.ttest_1samp(ech, popmean=12, alternative='less')
```

Application : tests de significativité des coefficients de la droite de régression linéaire

Test sur la pente b À partir de la fonction pivotale sur b (proposition 7.7), il est possible de faire différents tests sur ce paramètre ; en particulier, le test de dépendance linéaire permet de déterminer s'il y a une relation linéaire ou non entre la variable explicative x et la variable à expliquer Y, en testant si la pente b est nulle :

$$\begin{cases} H_0 : b = 0 \\ H_1 : b \neq 0. \end{cases}$$

À partir de la fonction pivotale sur b, on obtient

$$\frac{\hat{b}}{\hat{\sigma}/\sqrt{ns_x^2}} \underset{H_0}{\sim} \mathcal{T}_{n-2}.$$

Il est logique de rejeter H_0 lorsque cette quantité en valeur absolue dépasse un certain seuil. On en déduit la RC au niveau α^* :

$$W = \left\{ \frac{|\hat{b}|}{\hat{\sigma}/\sqrt{ns_x^2}} > t_{n-2,1-\frac{\alpha^*}{2}} \right\}.$$

On peut également montrer que, sous l'hypothèse H_0,

$$\frac{S_{reg}}{S_{res}/(n-2)} \sim \mathcal{F}_{1,n-2}.$$

On rejettera H_0 lorsque le rapport de la variance expliquée à la variance résiduelle dépassera un certain seuil, ce qui conduit à la région critique

$$W' = \left\{ \frac{S_{reg}}{S_{res}/(n-2)} > f_{1,n-2;1-\frac{\alpha^*}{2}} \right\}. \tag{9.1}$$

On montre que ce test est strictement équivalent au test sur le coefficient de régression b. Si l'on reprend les données du chapitre 7, on obtient les régions critiques suivantes, pour un niveau de signification de 0.05 :

$$W = \left\{ \frac{|\hat{b}|}{\hat{\sigma}/\sqrt{ns_x^2}} > t_{22,0.975} = 2.07 \right\}$$
$$W' = \left\{ \frac{S_{reg}}{S_{res}/(n-2)} > f_{1,22;0.95} = 4.30 \right\}.$$

Comme nous avons ici

$$\frac{|\hat{b}|}{\hat{\sigma}/\sqrt{ns_x^2}} = 19.89 \text{ et } \frac{S_{reg}}{S_{res}/(n-2)} = 396,$$

nous pouvons conclure que le test est très significatif.

Test sur l'ordonnée à l'origine a Là encore, il est possible de faire différents tests sur a en utilisant sa fonction pivotale ; en particulier, on peut tester si la droite de régression passe à l'origine :

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0. \end{cases}$$

À partir de la fonction pivotale sur a, on obtient

$$\frac{\hat{a}}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_x^2}}} \underset{H_0}{\sim} \mathcal{T}_{n-2}.$$

On en déduit la RC au niveau α^* :

$$W = \left\{ \frac{|\hat{a}|}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\overline{x}^2}{s_x^2}}} > t_{n-2,1-\frac{\alpha^*}{2}} \right\}.$$

9.2 Tests sur la variance d'une loi normale d'espérance inconnue

Test bilatéral

Proposition 9.5. *Soit $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 inconnues. Le test du rapport de vraisemblance pour les hypothèses :*

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

a pour région critique :

$$W = \left\{ \frac{(n-1)s^{*2}}{\sigma_0^2} < \chi^2_{n-1;\frac{\alpha^*}{2}} \text{ ou } \frac{(n-1)s^{*2}}{\sigma_0^2} > \chi^2_{n-1;1-\frac{\alpha^*}{2}} \right\}.$$

Preuve. cf. section 9.5.3 □

Tests unilatéraux

Pour le problème de test : $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 > \sigma_0^2$, on obtient la région critique :

$$W = \left\{ \frac{(n-1)s^{*2}}{\sigma_0^2} > \chi^2_{n-1;1-\alpha^*} \right\}.$$

De même, la région critique pour le problème : $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 < \sigma_0^2$ est :

$$W = \left\{ \frac{(n-1)s^{*2}}{\sigma_0^2} < \chi^2_{n-1;\alpha^*} \right\}.$$

9.3 Lien avec les intervalles de confiance

Considérons à titre d'exemple le cas de la loi normale, de variance et d'espérance inconnues. Nous avons vu qu'un intervalle de confiance bilatéral pour μ au niveau $1-\alpha$ est

$$IC_{1-\alpha} = \left[\overline{X} - \frac{s^*}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}; \overline{X} + \frac{s^*}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]. \tag{9.2}$$

On a

$$\mu_0 \in IC_{1-\alpha} \Leftrightarrow |\overline{X} - \mu_0| < \frac{s^*}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \Leftrightarrow \frac{|\overline{X} - \mu_0|}{s^*/\sqrt{n}} < t_{n-1,1-\frac{\alpha}{2}},$$

ce qui correspond à la région d'acceptation du test de Student bilatéral, pour les hypothèses $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$. On peut donc définir un intervalle de confiance comme l'ensemble des valeurs μ_0 pour lesquelles l'hypothèse $H_0 : \mu = \mu_0$ serait acceptée au vu de l'échantillon.

Réciproquement, on peut définir le test de Student à partir de l'intervalle de confiance de l'équation (9.2) de la façon suivante :

$$\begin{cases} \text{Rejet de } H_0 & \text{si} \quad \mu_0 \notin IC_{1-\alpha} \\ \text{Acceptation de } H_0 & \text{si} \quad \mu_0 \in IC_{1-\alpha}. \end{cases}$$

Ce test est de niveau α. En effet

$$\mathbb{P}_{H_0}(W) = \mathbb{P}_{H_0}(\mu_0 \notin IC_{1-\alpha}) = \alpha.$$

Cet exemple montre la dualité qui existe entre les notions d'intervalle de confiance et de test. On peut de la même façon relier les tests étudiés dans ce chapitre aux intervalles de confiance unilatéraux ou bilatéraux introduits au chapitre 6.

9.4 Test sur une proportion

Considérons une unique variable aléatoire $X \sim \mathcal{B}(n, p)$. On veut tester la conformité de la proportion p avec une proportion de référence p_0.

Cas unilatéral inférieur On a $H_0 = \{p = p_0\}$ vs $H_1 = \{p < p_0\}$, le test UPP obtenu à la forme :

$$W = \{X < k\}$$

Test exact La détermination de k (sans approximation asymptotique) est délicate, mais le calcul de la *p-value* est simple :

$$\begin{aligned} \hat{\alpha} &= \mathbb{P}_{H_0}(X \leq x_{\text{obs}}) \\ &= F_{\mathcal{B}(n,p_0)}(x_{\text{obs}}) \end{aligned}$$

Test asymptotique On peut utiliser le TCL, on obtient que

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0,1)$$

On obtient avec la correction de continuité :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(X \leq x_{\text{obs}}) \\ &= F_{\mathscr{B}(n,p_0)}(x_{\text{obs}}) \\ &\approx \Phi\left(\frac{x + 1/2 - np_0}{\sqrt{np_0(1-p_0)}}\right)\end{aligned}$$

On peut également calculer la région critique, et on obtient en utilisant la correction de continuité :

$$W = \left\{X < np_0 + 1/2 - u_{1-\alpha^*}\sqrt{np_0(1-p_0)}\right\}$$

Cas unilatéral supérieur On a $H_0 = \{p = p_0\}$ vs $H_1 = \{p > p_0\}$, le test UPP obtenu à la forme :

$$W = \{X > k\}$$

Test exact La détermination de k (sans approximation asymptotique) est délicate, mais le calcul de la *p-value* est simple :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(X \geq x_{\text{obs}}) \\ &= 1 - \mathbb{P}_{H_0}(X < x_{\text{obs}}) \\ &= 1 - \mathbb{P}_{H_0}(X \leq x_{\text{obs}} - 1) \\ &= 1 - F_{\mathscr{B}(n,p_0)}(x_{\text{obs}} - 1)\end{aligned}$$

Test asymptotique On peut utiliser le TCL, on obtient que

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow[H_0]{\mathscr{L}} \mathcal{N}(0,1)$$

On obtient avec la correction de continuité :

$$\begin{aligned}\hat{\alpha} &= \mathbb{P}_{H_0}(X \geq x_{\text{obs}}) \\ &= 1 - F_{\mathscr{B}(n,p_0)}(x_{\text{obs}} - 1) \\ &\approx 1 - \Phi\left(\frac{x - 1/2 - np_0}{\sqrt{np_0(1-p_0)}}\right)\end{aligned}$$

On peut également calculer la région critique, et on obtient en utilisant la correction de continuité :

$$W = \left\{X > np_0 - 1/2 + u_{1-\alpha^*}\sqrt{np_0(1-p_0)}\right\}$$

Cas bilatéral On a $H_0 = \{p = p_0\}$ vs $H_1 = \{p \neq p_0\}$, il n'existe pas de test UPP, nous partons du test de la forme :

$$W = \{X < k_1\} \cup \{X > k_2\}$$

avec k_1 et k_2 de telle manière à rejeter avec la même probabilité sous H_0.

Test exact La détermination de k_1 et k_2 (sans approximation asymptotique) est délicate, mais le calcul de la *p-value* est simple :

$$\begin{aligned}\hat{\alpha} &= 2\min\left(\mathbb{P}_{H_0}(X \leq x_{\text{obs}}), \mathbb{P}_{H_0}(X \geq x_{\text{obs}})\right)\\ &= 2\min\left(\mathbb{P}_{H_0}(X \leq x_{\text{obs}}), 1 - \mathbb{P}_{H_0}(X < x_{\text{obs}})\right)\\ &= 2\min\left(\mathbb{P}_{H_0}(X \leq x_{\text{obs}}), 1 - \mathbb{P}_{H_0}(X \leq x_{\text{obs}} - 1)\right)\\ &= 2\min\left(F_{\mathscr{B}(n,p_0)}(x_{\text{obs}}), 1 - F_{\mathscr{B}(n,p_0)}(x_{\text{obs}} - 1)\right)\end{aligned}$$

Test asymptotique On peut utiliser le TCL, on obtient que

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow[H_0]{\mathscr{L}} \mathcal{N}(0,1)$$

On obtient avec la correction de continuité :

$$\begin{aligned}\hat{\alpha} &\approx 2\,\mathbb{P}_{H_0}(|X - np_0| \geq |x_{\text{obs}} - np_0|)\\ &\approx 2\left(1 - \Phi\left(\frac{|x - np_0| - 1/2}{\sqrt{np_0(1-p_0)}}\right)\right)\end{aligned}$$

On peut également calculer la région critique, et on obtient en utilisant la correction de continuité :

$$W = \left\{|X - np_0| > -1/2 + u_{1-\alpha^*/2}\sqrt{np_0(1-p_0)}\right\}$$

Le test sur une proportion est disponible dans `GNU R` avec la fonction `prop.test`. Les tests unilatéraux sont gérés via l'argument `alternative`.

Par exemple, si $x_{\text{obs}} = 12$, $n = 100$, $p_0 = 0.2$, dans le cas d'un test unilatéral inférieur, on utilisera :

```
prop.test(x=12, n=100, p=0.2, alternative='less')
```

À noter : le test effectué est le test asymptotique avec correction de continuité.

Le test sur une proportion est disponible dans `GNU R` au sein du sous-module `statsmodels.stats.proportion` du module `statsmodels` au sein de la fonction `binom_test` qui ne retourne que la *p-value*. Les tests unilatéraux sont gérés via l'argument `alternative`.

Par exemple, si $x_{\text{obs}} = 12$, $n = 100$, $p_0 = 0.2$, dans le cas d'un test unilatéral inférieur, on utilisera :

```
from statsmodels.stats.proportion import binom_test
binom_test(count=12, nobs=100, prop=0.2, alternative='smaller')
# seulement la pvalue est retournée
```

À noter : le test effectué est le test exact.

9.5 Démonstrations

9.5.1 Preuve de la Proposition 9.3

Ici, H_0 et H_1 sont toutes deux des hypothèses multiples car σ^2 est inconnu. Déterminons la RC du test du RV. On a :

$$\Lambda(x_1,\dots,x_n) = \frac{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}_+} L(\mu,\sigma^2;x_1,\dots,x_n)}{\sup_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} L(\mu,\sigma^2;x_1,\dots,x_n)}.$$

Le numérateur est égal à $L(\mu_0,\hat{\sigma}^2;x_1,\dots,x_n)$, où $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2$ est l'EMV de σ^2 quand $\mu=\mu_0$ est connu. Le dénominateur vaut $L(\overline{x},s^2;x_1,\dots,x_n)$, $\overline{x}$ et s^2 étant les EMV de μ et σ^2. On a donc :

$$\begin{aligned}
\Lambda &= \frac{(2\pi\hat{\sigma}^2)^{-n/2}\exp(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i-\mu_0)^2)}{(2\pi s^2)^{-n/2}\exp(-\frac{1}{2s^2}\sum_{i=1}^n (x_i-\overline{x})^2)} \\
&= \frac{(2\pi\hat{\sigma}^2)^{-n/2}\exp(-\frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2)}{(2\pi s^2)^{-n/2}\exp(-\frac{1}{2s^2}ns^2)} \\
&= \left(\frac{s^2}{\hat{\sigma}^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sum_{i=1}^n (x_i-\mu_0)^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x}+\overline{x}-\mu_0)^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu_0)^2 + 2(\overline{x}-\mu_0)\sum(x_i-\overline{x})}\right)^{n/2} \\
\Rightarrow \Lambda^{\frac{2}{n}} &= \frac{1}{1+\frac{n(\overline{x}-\mu_0)^2}{\sum(x_i-\overline{x})^2}} = \frac{1}{1+\frac{n(\overline{x}-\mu_0)^2}{(n-1)s^{*2}}} = \frac{1}{1+\frac{t^2}{n-1}},
\end{aligned}$$

avec

$$t = \frac{\overline{x}-\mu_0}{s^*/\sqrt{n}}.$$

Donc la condition $\Lambda < c$ équivaut à la condition $|t| > k$ pour une constante k. Or, sous l'hypothèse H_0 :

$$T = \frac{\overline{X}-\mu_0}{s^*/\sqrt{n}} \sim \mathcal{T}_{n-1}.$$

On a donc, sous l'hypothèse H_0 :

$$\begin{aligned}
&\mathbb{P}(|T|>k) = \alpha^* \\
\Leftrightarrow\ &\mathbb{P}(T>k) + \mathbb{P}(T<-k) = \alpha^* \\
\Leftrightarrow\ &\mathbb{P}(T>k) = \frac{\alpha^*}{2} \\
\Leftrightarrow\ &k = t_{n-1,1-\frac{\alpha^*}{2}}.
\end{aligned}$$

D'où la région critique :

$$W = \left\{ \frac{|\overline{x} - \mu_0|}{s^*/\sqrt{n}} > t_{n-1,1-\frac{\alpha^*}{2}} \right\}.$$

9.5.2 Preuve de la Proposition 9.4

On a :

$$\Lambda = \frac{\sup_{(\mu,\sigma^2)\in\{\mu_0\}\times\mathbb{R}_+} L(\mu, \sigma^2; x_1, \dots, x_n)}{\sup_{(\mu,\sigma^2)\in[\mu_0;+\infty[\times\mathbb{R}_+} L(\mu, \sigma^2; x_1, \dots, x_n)}.$$

Comme précédemment, le numérateur vaut $L(\mu_0, \hat{\sigma}^2; x_1, \dots, x_n)$ avec

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2.$$

Pour le dénominateur, il faut distinguer deux cas :

1. $\overline{x} > \mu_0$. Dans ce cas, le maximum de vraisemblance est obtenu comme précédemment pour $\mu = \overline{x}$ et $\sigma^2 = s^2$. On a donc :

   $$\Lambda = \left(1 + \frac{t^2}{n-1}\right)^{-n/2},$$

   d'où

   $$\Lambda < c \Leftrightarrow |t| > k \Leftrightarrow t > k,$$

   t étant une constante positive.

2. $\overline{x} < \mu_0$. Dans ce cas, la fonction de vraisemblance étant strictement décroissante par rapport à μ sur le domaine $[\mu_0; +\infty[$ et ce quel que soit σ^2, on a forcément au maximum $\mu = \mu_0$. Le maximum correspond donc dans ce cas à $\mu = \mu_0$ et $\sigma^2 = \hat{\sigma}^2$, d'où $\Lambda = 1$.

On a donc finalement, $\forall c < 1$:

$$\Lambda < c \Leftrightarrow \frac{\overline{x} - \mu_0}{s^*/\sqrt{n}} > k \text{ et } \overline{x} > \mu_0,$$

pour une constante k positive, c'est-à-dire

$$\Lambda < c \Leftrightarrow \frac{\overline{x} - \mu_0}{s^*/\sqrt{n}} > k,$$

k étant solution de l'équation :

$$\mathbb{P}_{H_0}\left(\frac{\overline{X} - \mu_0}{s^*/\sqrt{n}} > k\right) = \alpha^* \Leftrightarrow k = t_{n-1,1-\alpha^*}.$$

On a donc finalement la région critique :

$$W = \left\{ \frac{\overline{x} - \mu_0}{s^*/\sqrt{n}} > t_{n-1,1-\alpha^*} \right\}.$$

9.5.3 Preuve de la Proposition 9.5

On considère les hypothèses :

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

On a

$$\Lambda(x_1, \dots, x_n) = \frac{\sup_{(\mu,\sigma^2)\in\{\sigma_0^2\}*\mathbb{R}_+} L(\mu, \sigma^2; x_1, \dots, x_n)}{\sup_{(\mu,\sigma^2)\in\mathbb{R}*\mathbb{R}_+} L(\mu, \sigma^2; x_1, \dots, x_n)}.$$

Le numérateur est égal à $L(\overline{x}, \sigma_0^2; x_1, \dots, x_n)$ et le dénominateur est égal à $L(\overline{x}, s^2; x_1, \dots, x_n)$. On a donc :

$$\Lambda = \frac{(2\pi\sigma_0^2)^{-n/2} \exp(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \overline{x})^2)}{(2\pi s^2)^{-n/2} \exp(-\frac{1}{2s^2}\sum_{i=1}^n (x_i - \overline{x})^2)} = \left(\frac{s^2}{\sigma_0^2}\right)^{n/2} \exp\left[\frac{n}{2}(1 - \frac{s^2}{\sigma_0^2})\right].$$

En posant $y = \frac{s^2}{\sigma_0^2}$, on obtient $\ln\Lambda = \frac{n}{2}(\ln y + 1 - y)$ et $\frac{\partial \ln\Lambda}{\partial y} = \frac{n}{2}(\frac{1}{y} - 1)$. La dérivée s'annule donc pour $y = 1$: la fonction $\ln\Lambda$ (et donc Λ) est croissante, puis décroissante. Par ailleurs, on a $\lim_{y\to 0}\Lambda = 0$ et $\lim_{y\to+\infty}\Lambda = 0$, d'où le tableau de variation 9.1.

TABLE 9.1 – Tableau de variation de Λ en fonction de y.

y	0		1		$+\infty$
Λ	0	$\nearrow$		$\searrow$	0

On a par conséquent :

$$\Lambda < c \Leftrightarrow \frac{(n-1)s^{*2}}{\sigma_0^2} < k_1 \text{ ou } \frac{(n-1)s^{*2}}{\sigma_0^2} > k_2.$$

Pour obtenir un test de niveau α^*, les constantes k_1 et k_2 doivent satisfaire l'équation :

$$\mathbb{P}_{H_0}\left(\frac{(n-1)s^{*2}}{\sigma_0^2} < k_1 \text{ ou } \frac{(n-1)s^{*2}}{\sigma_0^2} > k_2\right) = \alpha^*$$
$$\Leftrightarrow \mathbb{P}_{H_0}\left(\frac{(n-1)s^{*2}}{\sigma_0^2} < k_1\right) + \mathbb{P}_{H_0}\left(\frac{(n-1)s^{*2}}{\sigma_0^2} > k_2\right) = \alpha^*$$

La résolution de cette équation n'est pas évidente. En pratique, on résout le problème suivant, plus simple :

$$\begin{cases} \mathbb{P}_{H_0}\left(\frac{(n-1)s^{*2}}{\sigma_0^2} < k_1\right) = \frac{\alpha^*}{2} \\ \mathbb{P}_{H_0}\left(\frac{(n-1)s^{*2}}{\sigma_0^2} > k_2\right) = \frac{\alpha^*}{2} \end{cases}.$$

Comme sous H_0 $\frac{(n-1)s^{*2}}{\sigma_0^2} \sim \chi^2_{n-1}$, on a

$$k_1 = \chi^2_{n-1;\frac{\alpha^*}{2}} \text{ et } k_1 = \chi^2_{n-1;1-\frac{\alpha^*}{2}},$$

d'où la région critique :

$$W = \left\{ \frac{(n-1)s^{*2}}{\sigma_0^2} < \chi^2_{n-1;\frac{\alpha^*}{2}} \text{ ou } \frac{(n-1)s^{*2}}{\sigma_0^2} > \chi^2_{n-1;1-\frac{\alpha^*}{2}} \right\}.$$

Chapitre 10

Tests d'homogénéité (deux populations)

10.1 Introduction et notations

On se place dans le cadre général suivant : deux variables aléatoires X et Y correspondent à une même caractéristique observée dans deux populations différentes. Peut-on considérer que les variables X et Y ont la même loi ?

Un *flowchart* est présenté figure 10.1 permettant d'appréhender les cas d'usages des différents tests d'homogénéité.

Dans ce chapitre, nous étudierons deux cas :

1. Cas gaussien : $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Le problème se décompose alors en deux parties :
   — le test de l'égalité des espérances ;
   — le test de l'égalité des variances.
2. Cas binomial : $X \sim \mathcal{B}(n, p_1)$, $Y \sim \mathcal{B}(m, p_2)$. Il s'agit ici de tester l'égalité des probabilités p_1 et p_2.

Nous décrirons ensuite des tests *non paramétriques*, qui ne se basent pas sur une famille paramétrée de lois définie a priori.

10.2 Égalité de deux espérances (cas gaussien)

10.2.1 Notations

Soient $X_1, \dots, X_n$ un échantillon iid de variable parente $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y_1, \dots, Y_m$ un échantillon iid de variable parente $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. On suppose en outre que les $N = n + m$ v.a. $X_1, \dots, X_n, Y_1, \dots, Y_m$ sont indépendantes.

10.2.2 Variances connues

Soit le problème de test :

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

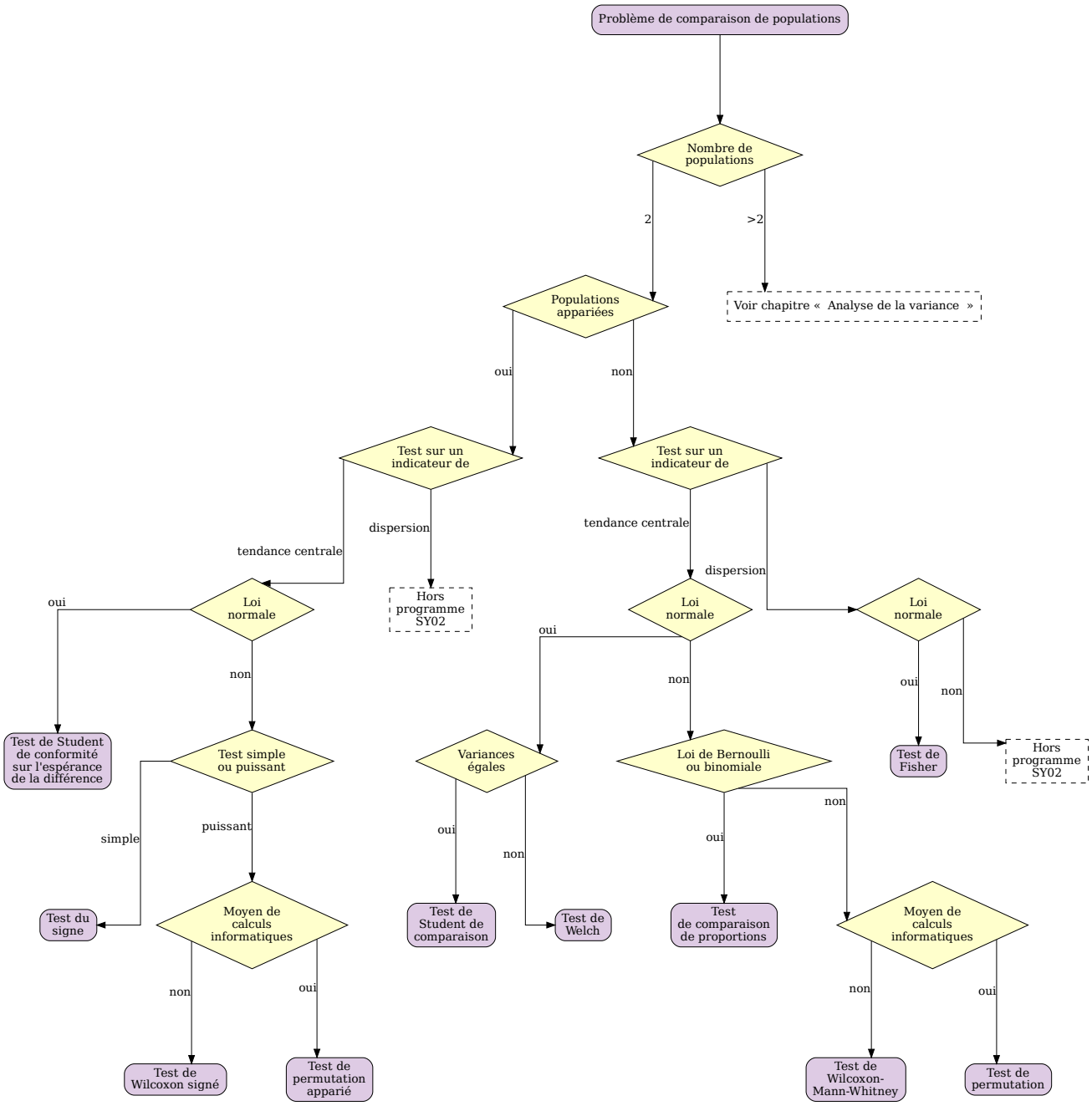


FIGURE 10.1 – *Flowchart* représentant les différents tests d'homogénéité et les cas d'usages. Ce diagramme est donné à titre informatif, il ne remplace pas une lecture précise des conditions d'application de chaque test.

On peut se ramener à un problème de test étudié précédemment en considérant la différence des moyennes empiriques $\overline{X}-\overline{Y}$. On a $\overline{X} \sim \mathcal{N}(\mu_X, \sigma_X^2/n)$, $\overline{Y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2/m)$ et $\overline{X}$ et $\overline{Y}$ indépendantes, donc :

$$D = \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

Posons $\mu_D = \mathbb{E}(D) = \mu_X - \mu_Y$. Les hypothèses H_0 et H_1 peuvent être reformulées de la manière suivante :

$$\begin{cases} H_0 : & \mu_D = 0 \\ H_1 : & \mu_D \neq 0. \end{cases}$$

Il s'agit d'un problème de test bilatéral sur l'espérance d'une v.a. de variance $\sigma_D^2 = \sigma_X^2/n + \sigma_Y^2/m$ connue, l'échantillon de la variable aléatoire D étant ici de taille 1. Nous avons vu au chapitre précédente que ce problème a pour RC :

$$W = \left\{|d| > u_{1-\frac{\alpha^*}{2}}\sigma_D\right\}.$$

Dans le cas d'une hypothèse alternative $H_1 : \mu_X > \mu_Y$, le même raisonnement conduit à la RC :

$$W = \{d > u_{1-\alpha^*}\sigma_D\}.$$

Pour l'hypothèse alternative $H_1 : \mu_X < \mu_Y$, on obtient

$$W = \{d < -u_{1-\alpha^*}\sigma_D\}.$$

10.2.3 Variances inconnues mais égales : test de Student

On suppose maintenant $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, ce paramètre étant inconnu et on considère le problème de test :

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

Proposition 10.1. *Le test du rapport de vraisemblance pour ce problème a pour région critique :*

$$W = \left\{\frac{|\overline{x} - \overline{y}|}{s^*\sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{N-2,1-\frac{\alpha^*}{2}}\right\},$$

où

$$s^{*2} = \frac{1}{N-2}\left((n-1)s_X^{*2} + (m-1)s_Y^{*2}\right).$$

Preuve. cf. Section 10.6. □

Le test de Student est présent dans GNU R sous la fonction `t.test`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
t.test(x, y, var.equal=TRUE)
```

Le test de Student est présent dans le module `stats` de `scipy` sous la fonction `ttest_ind`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
from scipy import stats

stats.ttest_ind(x, y)
```

Tests unilatéraux

Dans le cas d'un problème de test unilatéral, on montre que l'on obtient les résultats suivants :

$$W = \left\{ \frac{\overline{x} - \overline{y}}{S^* \sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{N-2,1-\alpha^*} \right\} \text{pour } H_1 \ : \ \mu_X > \mu_Y$$

$$W = \left\{ \frac{\overline{x} - \overline{y}}{S^* \sqrt{\frac{1}{n} + \frac{1}{m}}} < -t_{N-2,1-\alpha^*} \right\} \text{pour } H_1 \ : \ \mu_X < \mu_Y.$$

10.2.4 Variances inconnues et différentes

Dans ce cas, la méthode du rapport de vraisemblance ne permet pas de construire un test exact et on a recours à des solutions approchées. Pour cela, on s'appuiera sur la différence des moyennes :

$$D = \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_D, \sigma_D^2\right),$$

avec $\mu_D = \mu_X - \mu_Y$ et $\sigma_D^2 = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ et sur la statistique de test,

$$\frac{D}{\sqrt{S_D^{*2}}},$$

avec

$$S_D^{*2} = \frac{S_X^{*2}}{n} + \frac{S_Y^{*2}}{m}.$$

Approximation de Satterthwaite

Sous l'hypothèse d'égalité des variances, le carré du dénominateur de la statistique de test suit une loi du χ^2 (à un facteur multiplicatif près). L'approximation de Satterthwaite permet d'aboutir au même résultat de manière approximative sans l'hypothèse d'égalité des variances.

Proposition 10.2 (Approximation de Satterthwaite). *Celle-ci est ici présentée dans le cas général, pour k_1 et k_2 quelconques.*

Pour $k_1, k_2 > 0$, si on a

$$\begin{cases} \nu_1 \frac{S_1^{*2}}{\sigma_1^2} \sim \chi_{\nu_1}^2 \\ \\ \nu_2 \frac{S_2^{*2}}{\sigma_2^2} \sim \chi_{\nu_2}^2, \end{cases}$$

alors on peut déterminer la distribution approchée suivante :

$$\nu \frac{k_1 S_1^{*2} + k_2 S_2^{*2}}{k_1\sigma_1^2 + k_2\sigma_2^2} \overset{approx}{\sim} \chi^2_\nu,$$

avec

$$\nu = \frac{\left(k_1 s_1^{*2} + k_2 s_2^{*2}\right)^2}{\frac{\left(k_1 s_1^{*2}\right)^2}{\nu_1} + \frac{\left(k_2 s_2^{*2}\right)^2}{\nu_2}}.$$

Test de Welch

Pour trouver la loi de $S_D^{*2} = \frac{S_X^{*2}}{n} + \frac{S_Y^{*2}}{m}$, étant donné que $(n-1)\frac{S_X^{*2}}{\sigma_X^2} \sim \chi^2_{n-1}$ et $(m-1)\frac{S_Y^{*2}}{\sigma_Y^2} \sim \chi^2_{m-1}$, en utilisant l'approximation de Satterthwaite avec

$$\begin{cases} (S_1^{*2}, S_2^{*2}) &= (S_X^{*2}, S_Y^{*2}) \\ (\nu_1, \nu_2) &= (n-1, m-1) \\ (k_1, k_2) &= \left(\frac{1}{n}, \frac{1}{m}\right). \end{cases}$$

On trouve donc

$$\nu = \frac{\left(\frac{s_X^{*2}}{n} + \frac{s_Y^{*2}}{m}\right)^2}{\frac{\left(\frac{s_X^{*2}}{n}\right)^2}{n-1} + \frac{\left(\frac{s_Y^{*2}}{m}\right)^2}{m-1}},$$

et on obtient

$$\nu \frac{S_D^{*2}}{\sigma_D^2} = \nu \frac{\frac{S_X^{*2}}{n} + \frac{S_Y^{*2}}{m}}{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \overset{\text{approx}}{\sim} \chi^2_\nu.$$

Les hypothèses $H_0 : \mu_X = \mu_Y$ contre $H_1 : \mu_X \neq \mu_Y$, se reformulent donc en $H_0 : \mu_D = 0$ contre $H_1 : \mu_D \neq 0$. On obtient donc en utilisant ce qui précède :

$$\frac{D}{\sqrt{S_D^{*2}}} \overset{\text{approx}}{\underset{H_0}{\sim}} \mathcal{T}_\nu,$$

ce qui implique la région critique :

$$W = \left\{ \frac{|d|}{\sqrt{s_D^{*2}}} > t_{\nu, 1-\frac{\alpha^*}{2}} \right\},$$

avec ν, d et s_D^{*2} tels que précédemment définis.

Remarque 10.1. *Du fait de l'utilisation de l'approximation de Satterthwaite, rien ne contraint ν à être entier. La loi du χ^2 est parfaitement définie avec un nombre de degrés de liberté réel positif.*

Le test de Welch est présent dans GNU R sous la fonction `t.test`.
Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
t.test(x, y, var.equal=FALSE)
```

Le test de Welch est présent dans le module `stats` de `scipy` sous la fonction `ttest_ind` en précisant que les variances sont différentes avec l'argument `equal_var`.
Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
from scipy import stats

stats.ttest_ind(x, y, equal_var=False)
```

10.2.5 Cas de deux échantillons appariés

Dans ce cas, on ne suppose plus que les deux échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ sont indépendants l'un de l'autre, mais qu'ils sont appariés.

Définition 10.1. *Deux échantillons $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ sont appariés si les deux conditions suivantes sont vérifiées :*
- *ils sont de même taille ($n = m$),*
- *la famille des paires est indépendante (pour tout $i \neq j$, (X_i, Y_i) et (X_j, Y_j) sont indépendantes).*

Exemple 10.1. *Un dispositif expérimental mesure pour chaque parcelle le rendement d'une moitié de parcelle traité avec un phytosanitaire et le rendement de l'autre moitié non traité. Les rendements obtenus dans les deux cas constituent des échantillons appariés.*

On introduit la différence

$$D_i = X_i - Y_i,$$

et on suppose que l'échantillon D est iid gaussien. On a alors

$$D_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_D, \sigma_D^2) \text{ avec } \mu_D = \mu_X - \mu_Y,$$

et σ_D inconnu.

Les tests des hypothèses $\{\mu_X = \mu_Y\}$, $\{\mu_X < \mu_Y\}$ et $\{\mu_X > \mu_Y\}$ se traduisent donc sur μ_D en $\{\mu_D = 0\}$, $\{\mu_D < 0\}$ et $\{\mu_D > 0\}$: on appliquera donc dans ces cas un test de conformité sur l'échantillon $D_1, \dots, D_n$.

Le test de comparaison est présent dans GNU R sous la fonction `t.test`.
Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
t.test(x, y, paired=TRUE)
```

Le test de comparaison de Student d'échantillon appariés est présent dans le module `stats` de `scipy` sous la fonction `ttest_rel`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
from scipy import stats

stats.ttest_rel(x, y)
```

10.3 Égalité de deux variances (cas gaussien)

On considère le problème de test :

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

Fonction pivotale du paramètre σ_X^2/σ_Y^2

On a $\frac{(n-1)S_X^{*2}}{\sigma_X^2} \sim \chi^2_{n-1}$ et $\frac{(m-1)S_Y^{*2}}{\sigma_Y^2} \sim \chi^2_{m-1}$. En outre, les deux v.a. étant indépendantes, le rapport suit une loi de Fisher (définition 3.17) :

$$\frac{S_X^{*2}/\sigma_X^2}{S_Y^{*2}/\sigma_Y^2} \sim \mathcal{F}_{n-1,m-1}.$$

Statistique de test

Donc la statistique $F = \frac{S_X^{*2}}{S_Y^{*2}}$ suit sous H_0 : $\{\sigma_X^2 = \sigma_Y^2\}$ la loi $\mathcal{F}_{n-1,m-1}$.

Région critique

$$P_{H_0}(F < k_1 \quad \text{ou} \quad F > k_2) = \alpha^*.$$

$$\begin{cases} \mathbb{P}_{H_0}(F < k_1) = \frac{\alpha^*}{2} \\ \mathbb{P}_{H_0}(F > k_2) = \frac{\alpha^*}{2}. \end{cases}$$

Comme sous H_0, $F \sim \mathcal{F}_{n-1,m-1}$, on a

$$k_1 = f_{n-1,m-1;\frac{\alpha^*}{2}} \quad \text{et} \quad k_2 = f_{n-1,m-1;1-\frac{\alpha^*}{2}}.$$

d'où la région critique :

$$W = \left\{ F < f_{n-1,m-1;\frac{\alpha^*}{2}} \quad \text{ou} \quad F > f_{n-1,m-1;1-\frac{\alpha^*}{2}} \right\}.$$

Ce test est appelé *test de Fisher.*

Le test de Fisher est présent dans `GNU R` sous la fonction `var.test`.
Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
var.test(x, y)
```

Le test de Fisher de comparaison de variances n'est pas implémenté dans le module `stats` de `scipy`. Il est toutefois simple à mettre en œuvre facilement. Pour l'appliquer avec les échantillons dans `x` et `y`, on pourra utiliser :

```
import numpy as np
from scipy import stats

varx = np.var(x, ddof=1)
vary = np.var(y, ddof=1)
fobs = varx/vary
law_under_h0 = stats.f(len(varx), len(vary))
pv1 = law_under_h0.cdf(fobs)
pv2 = law_under_h0.sf(fobs) # note: sf() = 1-cdf()
pval = 2*min(pv1, pv2)
```

Remarque 10.2. *Les tables donnent en général les fractiles de la loi de Fisher pour des ordres supérieurs à* 0.5. *La propriété 3.21 permet d'en déduire les fractiles d'ordres inférieurs à* 0.5.

Remarque 10.3. *Le test de Student nécessitant l'hypothèse d'égalité des variances, on fera souvent le test de Fisher pour choisir entre un test de Student et un test de Welch.*

10.4 Égalité de deux proportions (grands échantillons)

Dans deux échantillons de grandes tailles n et m, on a relevé les proportions $\hat{p}_1 = x/n$ et $\hat{p}_2 = y/m$ d'individus présentant un certain caractère. Soient p_1 et p_2 les probabilités correspondantes. On considère le problème de test suivant,

$$\begin{cases} H_0 : & p_1 = p_2 \\ H_1 : & p_1 \neq p_2. \end{cases}$$

La solution classique à ce problème consiste à raisonner sur la différence des fréquences observées $\hat{p}_1 - \hat{p}_2$, avec $\hat{p}_1 = X/n$ et $\hat{p}_2 = Y/m$. On rejettera H_0 si la différence des fréquences observées est trop grande.

Sous $H_0 : p_1 = p_2 = p$, on a approximativement

$$\hat{p}_1 \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \text{ et } \hat{p}_2 \sim \mathcal{N}\left(p, \frac{p(1-p)}{m}\right),$$

d'où :

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)\right),$$

soit encore :

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim \mathcal{N}(0,1).$$

En remplaçant p par son EMV sous H_0 : $\hat{p} = \frac{X+Y}{n+m}$ et en utilisant le théorème de Slutsky, on peut montrer que l'on a approximativement

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}} \sim \mathcal{N}(0,1),$$

d'où la RC :

$$W = \left\{ \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}} > u_{1-\frac{\alpha^*}{2}} \right\}.$$

10.5 Tests non paramétriques

Dans le cas où les hypothèses de normalité ne sont pas satisfaites, on fait appel à d'autres tests qui ne font plus cette hypothèse.

10.5.1 Test de permutation

Dans cette section, on suppose que $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ sont deux échantillons iid et indépendants. On ne fait pas d'hypothèse sur les variables parentes, mais il est connu que cette méthode fonctionne mieux pour les cas où les lois parentes sont continues. Le test porte sur la comparaison des lois de X et de Y au moyen de l'intermédiaire de la statistique de test.

Le problème de test est donc le suivant :

$$\begin{cases} H_0 : & F_X = F_Y \\ H_1 : & F_X \neq F_Y. \end{cases}$$

L'idée du test est de prendre une statistique mesurant la différence entre X et Y. Très souvent, nous voulons comparer les espérences et nous prenons comme statistique de test celle du test de Student (*cf.* 10.2.3) ici nommée T_{XY} :

$$T_{XY} = \frac{\overline{X} - \overline{Y}}{\sqrt{S^{*2}_{XY}(1/n + 1/m)}}$$

avec

$$S^{*2}_{XY} = \frac{(n-1)S^{*2}_X + (m-1)S^{*2}_Y}{n+m-2}$$

L'observation de la statistique est notée t_{xy}.

Toutefois, les hypothèses de normalité n'étant plus demandée, nous n'avons aucun moyen de connaitre sous H_0 la loi de T_{XY}, notée $\mathscr{L}_0$. L'idée des tests de permutations, c'est de considèrer que sous H_0 toutes les données ont même loi, ainsi si on construit deux échantillons observés x' et y' à partir d'une permutation des valeurs observées des échantillons originaux x et y, alors on obtient une réalisation sous la loi $\mathscr{L}_0$. Il est donc possible d'obtenir par ce moyen des valeurs sous la loi $\mathscr{L}_0$.

Soit $\mathscr{P}_{xy}$ l'ensemble des couples d'échantillon observés à partir des permutations de l'échantillon x et y. On déduit que

$$\mathscr{T}\mathscr{P} = \left\{ t_{x'y'} / (x', y') \in \mathscr{P}_{xy} \right\}$$

est un échantillon sous la loi $\mathscr{L}_0$.

Une illustration est donnée dans un cas où $n = m = 10$, figure 10.2 où à gauche X et Y sont des échantillons gaussiens et où à droite X et Y sont des échantillons suivant des lois exponentielles. Dans le premier cas, on retrouve un résultat que nous connaissions à la section 10.2.3, la loi $\mathscr{L}_0$ est une loi normale (il est inutile d'appliquer un test de permutation dans ce cadre), dans le second cas, les hypothèses de la section 10.2.3 ne sont pas vérifiées, mais il est toujours possible d'obtenir une version approchée de la loi $\mathscr{L}_0$.

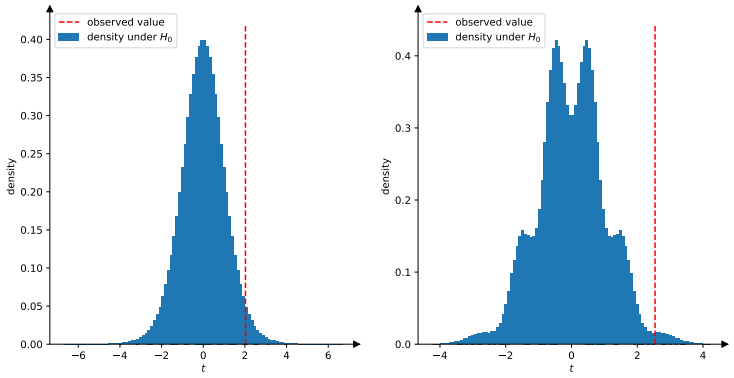


FIGURE 10.2 – Histogramme de la loi de la statistique sous H_0 obtenu à partir de permutation des valeurs de l'échantillon. À gauche, les hypothèses du test de Student étaient vérifiées et on retrouve bien comme attendu une loi de Student, à droite les hypothèses du test de Student ne sont pas vérifiées, et les permutations permettent d'obtenir une version empirique de la loi sous H_0. La valeur observée est représentée avec une barre verticale.

La p-value est définie par :

$$\hat{\alpha} = \mathbb{P}_{\mathscr{L}_0}\left(|T_{XY}| \geq \left|t_{xy}\right|\right)$$

On peut donc approcher la *p-value* par la proportion dans l'échantillon $\mathscr{T}\mathscr{P}$:

$$\tilde{\alpha} = \frac{\sum_{(x',y')\in\mathscr{P}_{xy}} \mathbb{1}_{\left|t_{x'y'}\right|\geq\left|t_{xy}\right|}}{\#\mathscr{P}_{xy}}$$

avec $\#\mathscr{P}_{xy}$ le nombre de permutations.

Cette *p-value* correspond à l'aire sous la courbe de l'histogramme figure 10.2 des données au moins aussi extrêmes que la valeur observée (dans les positifs au delà de la valeur absolue et dans les négatifs en deça de l'opposé de la valeur absolue).

Le nombre de permutations peut vite devenir très élevé, et quand n et m deviennent très grand, il n'est plus possible de parcourir l'ensemble des permutations. Lorsque le nombre de permutations possible est petit, il est possible d'énumérer l'ensemble des permutations, et le test réalisé sera un test dit exact. Lorsque le nombre de permutations est trop élevé, il n'est pas possible d'énumérer l'ensemble des permutations, et un sous ensemble au hasard est choisi, le test est dit approché.

Remarque 10.4. *Naïvement le nombre de permutations est de* $(n+m)!$*, toutefois, il est possible de remarquer que les permutations au sein d'un même échantillon produisent la même statistique puisque l'ordre des éléments de l'échantillon n'est pas considéré. Ainsi le nombre de permutation à explorer est de* $C_{n+m}^{n} = \frac{(n+m)!}{n!m!}$*. En pratique quand* $n+m \leq 24$ *on pourra utiliser un test exact, au delà on utilisera un test approché.*

Le test de Student avec permutation est disponible dans `GNU R` au sein (entre autres) du *package* `MKinfer`. Il s'agit uniquement du test approché, il faut lui préciser en argument `R` le nombre de permutations considérées. Par exemple, pour l'appliquer avec des échantillons dans `x` et `y` :

```
library('MKinfer')

perm.t.test(x, y, R=10**6)
```

Le test de Student avec permutation est disponible dans `scipy` avec la fonction `ttest_ind` du module `stats`. Cette fonction possède un argument `permutations`. Par défaut un test de Student ordinaire est effectuée (celui vu dans les sections précédente), pour utiliser un test de permutation, il faut lui mettre un nombre de permutations maximum autorisé. Si le nombre de permutations à explorer est en deça de ce nombre, un test exact sera effectué, sinon un test approché sera effectué. On peut utiliser $+\infty$ pour forcer un test exact (quel qu'en soit le prix).

Par exemple, pour appliquer un test exact (en le forçant) avec des échantillons dans `x` et `y` :

```
import numpy as np
from scipy import stats

stats.ttest_ind(x, y, permutations=np.inf)
```

ou pour appliquer un test exact si possible (on considère que 10^6 est le nombre maximum possible) et approché sinon :

```
stats.ttest_ind(x, y, permutations=10**6)
```

10.5.2 Test de Wilcoxon-Mann-Whitney

Dans cette section, on suppose que $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ sont deux échantillons iid et indépendants. Les variables parentes X et Y sont supposées continues. On ne fait plus d'hypothèses de normalité : le test porte directement sur la comparaison des lois de X et de Y. On considère les problèmes de test suivants :

Test bilatéral :

$$\begin{cases} H_0 : & F_X = F_Y \\ H_1 : & F_X \neq F_Y. \end{cases}$$

Tests unilatéraux :

$$\begin{cases} H_0 : & F_X = F_Y \\ H_1 : & F_X > F_Y \end{cases}$$

ou

$$\begin{cases} H_0 : & F_X = F_Y \\ H_1 : & F_X < F_Y \end{cases}$$

Lorsque la condition $F_X < F_Y$ est vérifiée, on dit que la v.a. X est *stochastiquement supérieure* à la v.a. Y : pour tout $x \in \mathbb{R}$, on a alors

$$\mathbb{P}(X \leq x) < \mathbb{P}(Y \leq x).$$

Le test de Wilcoxon-Mann-Whitney repose sur un calcul effectué sur les rangs. On introduit les variables aléatoires $R_1, \dots, R_n$ qui sont les rangs des variables aléatoires $X_1, \dots, X_n$ lorsque les échantillons $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$ sont rangés par ordre croissant. On introduit alors la statistique W_X qui est la somme des rangs attribués à l'échantillon $X_1, \dots, X_n$.

$$W_X = \sum_{i=1}^{n} R_i. \tag{10.1}$$

On peut montrer que, sous H_0,

$$\mathbb{E}(W_X) = \frac{n(n+m+1)}{2}$$

et

$$\mathrm{Var}(W_X) = \frac{nm(n+m+1)}{12}.$$

Intuitivement, une valeur de W_X plus faible que la valeur moyenne sous H_0 montre que les rangs des variables échantillon X_i sont faibles, et donc que les X_i sont plutôt plus petits que les Y_i. Plus précisément, on a les régions critiques suivantes

$$W = \left\{w_X \leq w_{n,m,\alpha^*/2}\right\} \cup \left\{w_X \geq n(n+m+1) - w_{n,m,\alpha^*/2}\right\}$$

pour le test bilatéral, $W = \left\{w_X \geq n(n+m+1) - w_{n,m,\alpha^*}\right\}$ pour l'hypothèse alternative $H_1 : F_X < F_Y$ et $W = \left\{w_X \leq w_{n,m,\alpha^*}\right\}$ pour l'hypothèse alternative $H_1 : F_X > F_Y$. Les fractiles w_{n,m,α^*} et $w_{n,m,\alpha^*/2}$ peuvent être lus dans des tables, ou approchés en admettant que W_X suit approximativement normale, l'approximation étant utilisable lorsque n et m sont supérieurs ou égaux à 8.

Remarque 10.5. *Le test de Wilcoxon-Mann-Whitney et les tests de permutations couvrent sensiblement le même cas d'usage. Toutefois, les tests de permutations sont beaucoup plus puissants pour les petits échantillons. Pour les grand échantillons, la puissance des deux méthodes est similaire.*

Le test de Wilcoxon-Mann-Whitney est présent dans GNU R sous la fonction `wilcox.test`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
wilcox.test(x, y)
```

en spécifiant l'argument `alternative` pour les tests unilatéraux.

Le test de Wilcoxon-Mann-Whitney est disponible dans la fonction `mannwhitneyu` du module `stats` de `scipy`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
stats.mannwhitneyu(x, y, method='exact')
```

10.5.3 Cas de deux échantillons appariés

Dans cette section, on suppose que $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ sont deux échantillons appariés. On note D_i les différences $D_i = X_i - Y_i$ et on suppose que les D_i on une médiane commune m. On désire appliquer le test suivant :

$$\begin{cases} H_0 : m = 0 \\ H_1 : m \neq 0, \end{cases} \tag{10.2}$$

ainsi que les variantes unilatérales avec $m > 0$ et $m < 0$. Dans ce qui suit, on notera Z_i les indicatrices des différences positives : $Z_i = 1$ si $D_i > 0$ et $Z_i = 0$ sinon.

Test du signe

Le test du signe repose sur la constatation suivante. Sous H_0, Z_i suit une loi de Bernoulli de paramètre 0.5. En notant $Z = \sum_{i=1}^{n} Z_i$, comme les Z_i sont indépendants par hypothèse, Z suit une loi binomiale de paramètre n et $\frac{1}{2}$. Le test (10.2) est donc équivalent au test sur le paramètre p d'un échantillon Z de longueur 1 suivant une loi binomiale

$$\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p \neq \frac{1}{2}. \end{cases}$$

On remarquera que l'hypothèse $m > 0$ se traduit en l'hypothèse $p > \frac{1}{2}$ et que $m < 0$ devient $p < \frac{1}{2}$ pour les tests unilatéraux.

Z sera donc utilisé en temps que statistique de test, et on a donc $Z \overset{H_0}{\sim} \mathcal{B}\left(n, \frac{1}{2}\right)$. Ainsi donc la région critique s'exprime comme :

$$W = \{z < c_1\} \cup \{z > c_2\}$$

En utilisant le fait que la loi $\mathcal{B}\left(n, \frac{1}{2}\right)$ soit symétrique par rapport à $\frac{n}{2}$, on obtient pour le test bilatéral :

$$W = \left\{\left|z - \frac{n}{2}\right| > c\right\}$$

Pour les tests unilatéraux, on obtient

$$W = \{z < c\} \qquad (H_1 : p < \tfrac{1}{2})$$

$$W = \{z > c\} \qquad (H_1 : p > \tfrac{1}{2})$$

Comme Z est une variable discrete, la détermination de c est fonction de α^* est délicate. Il est plus simple de déterminer la *p-value* $\hat{\alpha}$ pour un z_{obs}.

Dans le cas bilatéral, on a

$$\begin{aligned}
\hat{\alpha} &= \mathbb{P}_{H_0}\left(\left|Z - \frac{n}{2}\right| \geq \left|z_{\text{obs}} - \frac{n}{2}\right|\right) \\
&= \mathbb{P}_{H_0}\left(Z \leq -\left|z_{\text{obs}} - \frac{n}{2}\right| + \frac{n}{2}\right) + \mathbb{P}_{H_0}\left(Z \geq \left|z_{\text{obs}} - \frac{n}{2}\right| + \frac{n}{2}\right) \\
&= 2\mathbb{P}_{H_0}\left(Z \leq -\left|z_{\text{obs}} - \frac{n}{2}\right| + \frac{n}{2}\right) \\
&= 2F_{\mathscr{B}\left(n,\frac{1}{2}\right)}\left(Z \leq -\left|z_{\text{obs}} - \frac{n}{2}\right| + \frac{n}{2}\right)
\end{aligned}$$

Et dans les cas unilatéraux on a :

$$\begin{aligned}
\hat{\alpha} &= \mathbb{P}_{H_0}(Z \leq z_{\text{obs}}) && (H_1 : p < \tfrac{1}{2}) \\
&= F_{\mathscr{B}\left(n,\frac{1}{2}\right)}(z_{\text{obs}})
\end{aligned}$$

$$\begin{aligned}
\hat{\alpha} &= \mathbb{P}_{H_0}(Z \geq z_{\text{obs}}) && (H_1 : p > \tfrac{1}{2}) \\
&= 1 - \mathbb{P}_{H_0}(Z \leq z_{\text{obs}} - 1) \\
&= 1 - F_{\mathscr{B}\left(n,\frac{1}{2}\right)}(z_{\text{obs}} - 1)
\end{aligned}$$

Ce test peut être pratiqué avec la fonction de répartition exacte de la loi $\mathscr{B}\left(n, \frac{1}{2}\right)$ ou avec l'approximation normale.

Le test du signe est réalisable dans `GNU R` au moyen du test sur une proportion sous la fonction `prop.test`. Ce test est approché avec correction de continuité. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
n.di.pos <- sum(x > y)
prop.test(n.di.pos, n, p = 0.5)
```

en spécifiant l'argument `alternative` pour les tests unilatéraux.

Le test du signe est réalisable avec `scipy` au moyen du test sur une proportion avec les fonction `binom_test` du module `stats`. Ce test est exact.

```
from scipy import stats

npos = (x > y).sum()
n = len(x)
stats.binom_test(npos, n, p=0.5)
```

À noter, `binomtest`, qui retourne plusieurs informations dont la *p-value* a été introduite dans la version 1.7.0 de `scipy`. Dans les anciennes versions, on

utilisera `binom_test` qui ne retourne que la *p-value*. `binom_test` est dépréciée dans les nouvelles versions de `scipy`.

Test de Wilcoxon signé

Le test de Wilcoxon signé est une alternative plus puissante au test du signe. En revanche, les conditions d'application sont un peu plus strictes que pour le test de signe. En plus de la médiane commune m, il faut que les différences D_i aient une loi (pas nécessairement la même) symétrique en m.

L'idée est de prendre en compte non seulement le signe de la différence D_i mais aussi leur importance en valeur absolue. On note R_i le rang de la variable aléatoire $|D_i|$ dans l'échantillon $(D_1, \dots, D_n)$ et on pose

$$W^+ = \sum_{i=1}^{n} R_i Z_i. \tag{10.3}$$

La statistique W^+ calcule la somme des rangs attribués aux différences positives. Sous H_0, on montre que

$$\mathbb{E}(W^+) = \frac{n(n+1)}{4}$$

et

$$\mathrm{Var}(W^+) = \frac{n(n+1)(2n+1)}{24}.$$

On a les régions critiques suivantes :

$$W = \left\{w^+ \leq w^+_{\alpha^*/2}\right\} \cup \left\{w^+ \geq \frac{n(n+1)}{2} - w^+_{\alpha^*/2}\right\},$$

pour le test bilatéral et $W = \left\{w^+ \leq w_{\alpha^*}\right\}$ et $W = \left\{w^+ \geq \frac{n(n+1)}{2} - w^+_{\alpha^*}\right\}$ pour les tests unilatéraux. Le seuil $w^+_{\alpha^*}$ peut être lu dans des tables ou approché en admettant que W^+ suit une loi proche de la loi normale, l'approximation étant utilisable dès lors que $n > 20$.

Le test de Wilcoxon signé est présent dans `GNU R` sous le nom `wilcox.test`. Pour l'appliquer avec les échantillons dans `x` et `y`, on utilisera :

```
wilcox.test(x, y, paired = TRUE)
```

en spécifiant l'argument `alternative` pour les tests unilatéraux.

Le test de Wilcoxon signé est présent dans `scipy` sous la fonction `wilcoxon` du module `stats`. Pour l'appliquer avec les échantillons dans les vecteur `x` et `y`, on utilisera :

```
from scipy import stats

stats.wilcoxon(x, y)
```

Il est possible de spécifier l'argument `alternative` pour les tests unilatéraux.

10.6 Preuve de la Proposition 10.1

Calculons la statistique du RV :

$$\Lambda(x_1,\dots,x_n,y_1,\dots,y_m) = \frac{\sup_{(\mu_X,\mu_Y,\sigma^2)\in H_0} L(\mu_X,\mu_Y,\sigma^2;x_1,\dots,x_n,y_1,\dots,y_m)}{\sup_{(\mu_X,\mu_Y,\sigma^2)\in\mathbb{R}^2\times\mathbb{R}_+} L(\mu_X,\mu_Y,\sigma^2;x_1,\dots,x_n,y_1,\dots,y_m)}. \quad (10.4)$$

Sous H_0, $\mu_X = \mu_Y = \mu$ et l'échantillon $X_1,\dots,X_n,Y_1,\dots,Y_m$ est donc iid de v.a. parente $\mathcal{N}(\mu,\sigma^2)$. Les EMV de μ et de σ^2 sont donc :

$$\hat{\mu} = \frac{1}{N}(\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i) = \frac{n\overline{X}+m\overline{Y}}{N}$$

et

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{n}(X_i-\hat{\mu})^2 + \sum_{i=1}^{m}(Y_i-\hat{\mu})^2\right).$$

Le numérateur dans le membre de droite de l'équation (10.4) vaut donc $L(\hat{\mu},\hat{\mu},\hat{\sigma}^2)$. Calculons le dénominateur. Pour cela, déterminons les EMV de μ_X, μ_Y et σ^2 :

$$L = L(\mu_X,\mu_Y,\sigma^2;x_1,\dots,x_n,y_1,\dots,y_m) = \left(\prod_{i=1}^{n} f_X(x_i)\right)\left(\prod_{i=1}^{m} f_Y(y_i)\right)$$
$$= (2\pi\sigma^2)^{-N/2}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_X)^2 + \sum_{i=1}^{m}(y_i-\mu_Y)^2\right)\right],$$

et donc

$$\ell = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_X)^2 + \sum_{i=1}^{m}(y_i-\mu_Y)^2\right).$$

Le système des équations de vraisemblance s'écrit :

$$\begin{cases} \frac{\partial\ell}{\partial\mu_X} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu_X) = 0 \\ \frac{\partial\ell}{\partial\mu_Y} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i-\mu_Y) = 0 \\ \frac{\partial\ell}{\partial\sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\sum_{i=1}^{n}(x_i-\mu_X)^2 + \sum_{i=1}^{m}(y_i-\mu_Y)^2\right) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \mu_X &= \overline{x} \\ \mu_Y &= \overline{y} \\ \sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{i=1}^{m}(y_i-\overline{y})^2\right) \\ &= \frac{1}{N}\left((n-1)s_X^{*2} + (m-1)s_Y^{*2}\right). \end{cases}$$

Les estimateurs du maximum de vraisemblance de μ_X, μ_Y et σ^2 sont donc, respectivement $\overline{X}$, $\overline{Y}$ et $S^2 = \frac{1}{N}\left((n-1)S_X^{*2} + (m-1)S_Y^{*2}\right)$.

Remarquons que

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{N}\left((n-1)\mathbb{E}[S_X^{*2}] + (m-1)\mathbb{E}[S_Y^{*2}]\right)\\ &= \frac{1}{N}\left((n-1)\sigma^2 + (m-1)\sigma^2\right) = \frac{N-2}{N}\sigma^2.\end{aligned}$$

Pour avoir un estimateur sans biais de σ^2, on définit donc :

$$S^{*2} = \frac{N}{N-2}S^2 = \frac{1}{N-2}\left((n-1)S_X^{*2} + (m-1)S_Y^{*2}\right)$$

Revenons à l'expression de Λ (équation 10.4). On a

$$\begin{aligned}\Lambda &= \frac{L(\hat{\mu},\hat{\mu},\hat{\sigma}^2)}{L(\overline{x},\overline{y},s^2)} = \frac{(2\pi\hat{\sigma}^2)^{-N/2}\exp\left[-\frac{1}{2\hat{\sigma}^2}\left(\sum_{i=1}^n (X_i-\hat{\mu})^2 + \sum_{i=1}^m (y_i-\hat{\mu})^2\right)\right]}{(2\pi s^2)^{-N/2}\exp\left[-\frac{1}{2s^2}\left(\sum_{i=1}^n (X_i-\overline{x})^2 + \sum_{i=1}^m (y_i-\overline{y})^2\right)\right]}\\ &= \left(\frac{s^2}{\hat{\sigma}^2}\right)^{N/2}\frac{\exp\left[-\frac{1}{2\hat{\sigma}^2}(N\hat{\sigma}^2)\right]}{\exp\left[-\frac{1}{2s^2}(Ns^2)\right]} = \left(\frac{s^2}{\hat{\sigma}^2}\right)^{N/2} = \left(\frac{\hat{\sigma}^2}{s^2}\right)^{-N/2}.\end{aligned}$$

On a

$$\begin{aligned}\sum_{i=1}^n (x_i-\hat{\mu})^2 &= \sum_{i=1}^n (x_i-\overline{x}+\overline{x}-\hat{\mu})^2 = \sum_{i=1}^n (x_i-\overline{x})^2 + n\left(\overline{x} - \frac{n\overline{x}+m\overline{y}}{N}\right)^2\\ &= \sum_{i=1}^n (x_i-\overline{x})^2 + \frac{nm^2}{N^2}(\overline{x}-\overline{y})^2\end{aligned}$$

Sachant que

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^n (x_i-\hat{\mu})^2 + \sum_{i=1}^m (y_i-\hat{\mu})^2\right)$$

et que

$$s^2 = \frac{1}{N}\left(\sum_{i=1}^n (x_i-\overline{x})^2 + \sum_{i=1}^m (y_i-\overline{y})^2\right),$$

on peut en déduire

$$\hat{\sigma}^2 = s^2 + \frac{nm}{N^2}(\overline{x}-\overline{y})^2.$$

On obtient donc

$$\begin{aligned}\Lambda &= \left(1 + \frac{nm(\overline{x}-\overline{y})^2}{N^2 s^2}\right)^{-N/2}\\ &= \left(1 + \frac{nm(\overline{x}-\overline{y})^2}{N(N-2)s^{*2}}\right)^{-N/2}\\ &= \left(1 + \frac{t^2}{N-2}\right)^{-N/2},\end{aligned}$$

en posant

$$t = \frac{\overline{x}-\overline{y}}{s^*\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

Donc la RC du test du RV pour ce problème est de la forme :

$$W = \{|t| > k\}.$$

Pour déterminer k, déterminons la loi de la statistique T sous l'hypothèse H_0. Remarquons tout d'abord que :

$$\frac{(N-2)S^{*2}}{\sigma^2} = \underbrace{\frac{(n-1)S_X^{*2}}{\sigma^2}}_{\sim\chi^2_{n-1}} + \underbrace{\frac{(m-1)S_Y^{*2}}{\sigma^2}}_{\sim\chi^2_{m-1}},$$

d'où l'on déduit que

$$\frac{(N-2)S^{*2}}{\sigma^2} \sim \chi^2_{N-2}.$$

Par conséquent,

$$T = \frac{\left.\dfrac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right\} \sim \mathcal{N}(0,1)}{\left.\sqrt{\dfrac{S^{*2}}{\sigma^2}}\right\} \sim \sqrt{\dfrac{\chi^2_{N-2}}{N-2}}} \sim \mathcal{T}_{N-2}.$$

On a donc

$$\mathbb{P}_{H_0}(|T| > k) = \alpha^* \Leftrightarrow \mathbb{P}_{H_0}(T > k) = \frac{\alpha^*}{2} \Leftrightarrow k = t_{N-2,1-\frac{\alpha^*}{2}},$$

et finalement

$$W = \left\{ \frac{|\overline{x}-\overline{y}|}{s^*\sqrt{\frac{1}{n}+\frac{1}{m}}} > t_{N-2,1-\frac{\alpha^*}{2}} \right\}.$$

Chapitre 11

Tests d'adéquation

L'objet des méthodes présentées dans ce chapitre est de tester l'accord (l'adéquation) entre un *modèle*, caractérisé par une loi ou une famille de lois de probabilité, et des observations. Plus précisément, soit $X_1, \dots, X_n$ un échantillon iid de variable parente X. Nous considérons des hypothèses nulles de la forme suivante :

$$H_0 \, : \, F_X \in \mathscr{F},$$

où $\mathscr{F}$ est une famille de lois (potentiellement réduite à une seule loi).

Un *flowchart* est présenté figure 11.1 permettant d'appréhender les cas d'usages des différents tests d'adéquation.

Cas particuliers :

- $\mathscr{F} = \{F_0\}$: H_0 est une hypothèse simple (par exemple, $H_0 \, : \, X \sim \mathcal{N}(0,1)$) ;
- $\mathscr{F} = (F_\theta)_{\theta \in \Theta}$: $\mathscr{F}$ est une famille de lois indicées par un paramètre θ (par exemple, $\mathscr{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$; on parle dans ce cas de *test de normalité*.)

11.1 Le test du χ^2

On considère ici l'hypothèse nulle $H_0 \, : \, F_X = F_0$ où F_0 est une loi de probabilité fixée. Connaissant une partition $C_1, \dots, C_K$ en K classes du domaine de $\mathcal{X}$ de X, on définit :

$$N_k = \operatorname{card}\{i \in \{1, \dots, n\} \mid X_i \in C_k\}, \, k = 1, \dots, K.$$

Le vecteur aléatoire $N = (N_1, \dots, N_K)$ suit alors une loi multinomiale (cf. Section 3.5.3) :

$$N \sim \mathscr{M}(n; p_1, \dots, p_K), \quad \text{avec } p_k = \mathbb{P}(X \in C_k).$$

Les paramètres $p_1, \dots, p_K$ dépendent donc de la loi de X. En particulier, sous H_0, on a pour tout k :

$$p_k = p_{k0} = \mathbb{P}_{H_0}(X \in C_k) = \begin{cases} \int_{C_k} f_0(x)dx & \text{si } X \text{ est une v.a. continue} \\ \sum_{x \in C_k} f_0(x) & \text{si } X \text{ est une v.a. discrète,} \end{cases}$$

f_0 désignant la fonction de densité de probabilité, ou la fonction de probabilité de X, selon que X est une v.a. continue ou discrète.

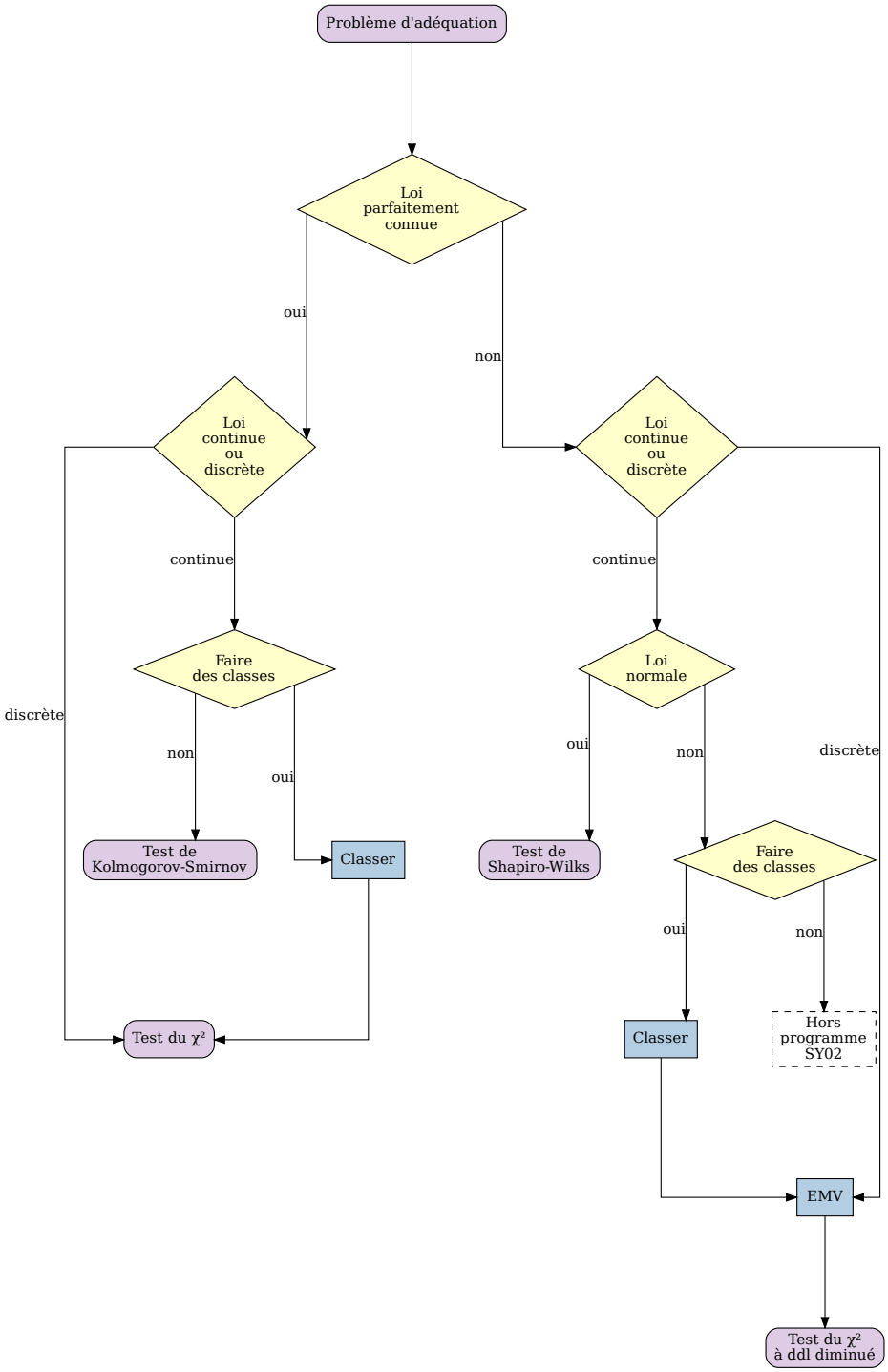


FIGURE 11.1 – *Flowchart* représentant les différents tests d'adéquation et les cas d'usages. Ce diagramme est donné à titre informatif, il ne remplace pas une lecture précise des conditions d'application de chaque test.

On peut donc reformuler ainsi l'hypothèse nulle $H_0 : p_k = p_{k0}, \forall k$.

L'écart entre l'échantillon et l'hypothèse H_0 peut être évalué par la statistique de test

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - np_{k0})^2}{np_{k0}},$$

qui suit asymptotiquement une loi du χ^2 à $K-1$ degrés de liberté si l'hypothèse nulle est vérifiée.

On obtient donc finalement le test du χ^2 :

$$W = \left\{D^2 > \chi^2_{K-1;1-\alpha^*}\right\},$$

dont le niveau est approximativement égal à α^*.

Remarque 11.1. *On considère que l'approximation de la loi de D^2 par la loi du χ^2 est valide dès que $np_{k0} \geq 5, \forall k$. Dans le cas contraire, il faut regrouper des classes.*

Remarque 11.2. *On a*

$$D^2 = \sum_{k=1}^{K} \frac{(n_k - np_{k0})^2}{np_{k0}} = \sum_{k=1}^{K} \frac{n_k^2 - 2n_k np_{k0} + n^2 p_{k0}^2}{np_{k0}}$$
$$= \sum_{k=1}^{K} \left(\frac{n_k^2}{np_{k0}} - 2n_k + np_{k0} \right) = \sum_{k=1}^{K} \frac{n_k^2}{np_{k0}} - 2n + n = \sum_{k=1}^{K} \frac{n_k^2}{np_{k0}} - n.$$

Remarque 11.3. *Dans le raisonnement précédent, nous avons supposé que H_0 était une hypothèse simple. Si $H_0 = (F_\theta)_{\theta\in\Theta}$, avec $\Theta \subset \mathbb{R}^p$, on remplace θ par son EMV $\hat{\theta}$ et on en déduit les probabilités estimées $\hat{p}_{k0}$ $(k = 1, \ldots, K)$ d'appartenance aux classes C_k sous H_0. On montre que, dans ce cas, la statistique*

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - n\hat{p}_{k0})^2}{n\hat{p}_{k0}}$$

suit approximativement une loi du χ^2 à $K - p - 1$ d.d.l.

Le test du χ^2 est présent dans `GNU R` sous le nom `chisq.test`. Pour l'appliquer avec un vecteur de comptage `N`, et `probas` contenant le vecteur des probabilités sous H_0, on utilisera :

```
N <- c(12, 17, 51)
probas <- c(.25, .25, .5)
chisq.test(N, p=probas)
```

Remarquons que via l'argument `simulate.p.value` on peut effectuer le test en simulant la loi exacte de la statistique de test et non la loi asymptotique. Cette méthode pourra être utilisée lorsque les conditions sur les effectifs ne permettent pas de faire l'approximation asymptotique.

Le test du χ^2 est présent dans `scipy` sous la fonction `chisquare` du module `stats`. Il nécessite un vecteur de comptage `N` et un vecteur de valeur obtenue sous H_0 qui peut être déduit du vecteur de probabilité sous H_0. Par exemple, on utilisera :

```
import numpy as np
from scipy import stats

N = np.array([12, 17, 51])
probas = np.array([.25,. .25, .5])
expected_under_H0 = N.sum() * probas

stats.chisquare(N, expected_under_H0)
```

11.2 Diagramme quantile–quantile (Q-Q plot)

Le diagramme quantile–quantile est une méthode permettant de vérifier l'adéquation à une distribution. Bien qu'il ne s'agisse pas d'un test, cette méthode sert de base à certains d'entre eux, et il peut être utile de la connaître pour vérifier visuellement l'adéquation à une distribution.

11.2.1 Diagramme

Soit $(x_1, \dots, x_n)$ un échantillon iid dont on cherche à vérifier la distribution selon une loi de fonction de répartition F. L'idée du diagramme Quantile–Quantile est de comparer les fractiles estimés sur l'échantillon aux fractiles théoriques selon F.

On utilise les fractiles d'ordre $\frac{1}{n+1}, \dots, \frac{n}{n+1}$. Avec la convention utilisée dans ce poly, on peut montrer que les fractiles empiriques d'ordre $\frac{1}{n+1}, \dots, \frac{n}{n+1}$ sont les valeurs de l'échantillon ordonné :

$$x_{(1)}, \dots, x_{(n)}.$$

Les fractiles théoriques d'ordre $\frac{1}{n+1}, \dots, \frac{n}{n+1}$ de la loi de fonction de répartition F sont

$$F^{-1}\left(\frac{1}{n+1}\right), \dots, F^{-1}\left(\frac{n}{n+1}\right).$$

Le diagramme Quantile–Quantile représente les n points ayant pour abscisses les fractiles théoriques

$$F^{-1}\left(\frac{1}{n+1}\right), \dots, F^{-1}\left(\frac{n}{n+1}\right),$$

et pour ordonnées les fractiles empiriques (voir par exemple la figure 11.2)

11.2.2 Interprétation du diagramme

Lorsque la distribution empirique de l'échantillon est proche de celle induite par F, alors on observe que les points du diagramme parcourent la droite d'équation $y = x$.

Lorsqu'après une transformation linéaire (translation et mise à l'échelle) la distribution empirique est proche de celle induite par F, les points du diagramme parcourent une droite dont l'ordonnée à l'origine représente la translation et la pente de la mise à l'échelle.

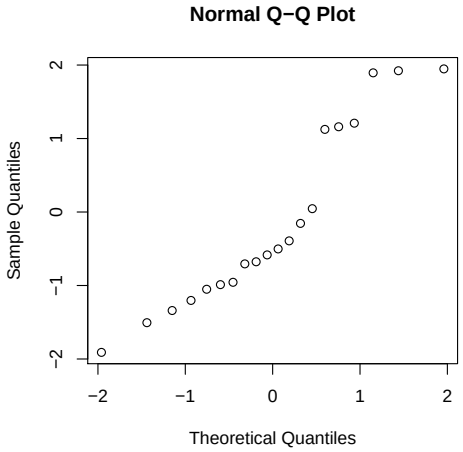


FIGURE 11.2 – Exemple de diagramme quantile–quantile.

Cas particulier d'une loi normale : Lorsqu'on trace un diagramme quantile–quantile avec un échantillon issu d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ contre la loi théorique $\mathcal{N}(0,1)$ les points décrivent une droite d'ordonnée à l'origine μ et de pente σ en application du paragraphe précédent. Cette propriété est souvent utilisée pour vérifier rapidement la normalité d'un échantillon en tracant un diagramme quantile–quantile contre la loi théorique $\mathcal{N}(0,1)$ et en vérifiant l'alignement des points.

Il est aisé de tracer des diagrammes quantile–quantile avec `GNU R` dans le cas d'une comparaison à une loi normale, si `x` contient l'échantillon à analyser :

```
qqnorm(x)
```

Dans le cas d'une autre loi, il faut créer les quantiles auxquels on se compare. Par exemple si on veut analyser l'échantillon contenu dans `x` contre une loi exponentielle (de paramètre d'intensité 1) :

```
quantiles <- qexp(ppoints(length(x)), rate=1)
qqplot(quantiles, x,
       xlab='Theorical Quantiles',
       ylab='Sample Quantiles')
```

Il est possible de tracer des diagramme quantile-quantile à l'aide de `scipy` et de `matplotlib`. Par exemple, si l'on veut vérifier que l'échantillon contenu dans `x` suit une loi normale :

```
from scipy import stats
import matplotlib.pyplot as plt

stats.probplot(x, dist='norm', plot=plt)
plt.show()
```

Dans le cas d'autres loi, on fournira l'objet de la loi tel que connu de `scipy`. Par exemple si on veut analyser un échantillon `x` contre une loi exmponentielle (de paramètre d'intensité 1, donc de paramètre d'echelle 1), et que l'on cherche si cette loi est suivi à un facteur additif et multiplicatif près, on trace la droite la plus plausible :

```
from scipy import stats
import matplotlib.pyplot as plt

distref=stats.expon(scale=1)
stats.probplot(x, dist=distref, plot=plt, fit=True)
plt.show()
```

Dans le même exemple, si l'on cherche si la loi est suivi, sans facteur additif et multiplicatif, on cherche à se comparer à la droite identité :

```
from scipy import stats
import matplotlib.pyplot as plt

distref = stats.expon(scale=1)
stats.probplot(x, dist=distref, plot=plt, fit=False)
spanx = (x.min(), x.max())
plt.plot(spanx, spanx, 'r-')
plt.show()
```

11.3 Test de Shapiro-Wilk

Le test de Shapiro-Wilk permet de tester l'hypothèse qu'un échantillon est issu d'une loi normale. Soit $(X_1, \dots, X_n)$ un échantillon i.i.d. dont on cherche à tester la normalité. On considère l'hypothèse nulle $H_0 = \{\exists(\mu, \sigma^2) \mid X_i \sim \mathcal{N}(\mu, \sigma^2)\}$ et l'hypothèse alternative H_1 complémentaire de H_0.

L'idée essentielle du test de Shapiro-Wilk est de considérer les quantiles empiriques de l'échantillon et les quantiles théoriques d'une loi normale. Comme nous l'avons vu dans la section 11.2 (diagrammes quantile–quantile), si l'échantillon est issu d'une loi normale, on doit observer une relation linéaire entre les quantiles empiriques obtenus avec l'échantillon et les quantiles théoriques de cette distribution.

La statistique de test de Shapiro-Wilk est :

$$\mathscr{W}(X_1, \dots, X_n) = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2},$$

où $a_1, \dots, a_n$ sont des valeurs obtenues à partir des moments théoriques des statistiques d'ordre d'une loi $\mathcal{N}(0,1)$ (aisément calculables au moyen d'un logiciel, elles sont consignées dans des tables). Ces valeurs $a_1, \dots, a_n$ sont antisymétriques ($\forall k$, $a_{1+k} = -a_{n-k}$) : on en déduit donc que $\mathscr{W}$ est invariante par translation ; une mise à l'échelle ayant le même effet au numérateur qu'au dénominateur, $\mathscr{W}$ est donc également invariante par mise à l'échelle.

Proposition 11.1. *La statistique de test $\mathscr{W}$ est invariante par translation et mise à l'echelle, i.e.*

$$\forall a \neq 0 \; \forall b \quad \mathscr{W}(X_1, \dots, X_n) = \mathscr{W}(aX_1 + b, \dots, aX_n + b).$$

La propriété 11.1 suggère que le test permet bien de vérifier l'adéquation à *une* loi normale quelconque, la statistique de test étant invariante aux changements de paramètres μ et σ^2.

Les propriétés de $\mathscr{W}(X)$ indiquent par ailleurs que cette statistique se rapproche de 0 quand X s'éloigne d'une loi normale. On en déduit donc la forme de la région critique :

$$\{\mathscr{W}(X_1, \dots, X_n) < w_{n,\alpha^*}\}$$

Le test de normalité de Shapiro-Wilk est intégré dans `GNU R`. Par exemple pour l'appliquer avec un échantillon stocké dans la variable `x` :

```
shapiro.test(x)
```

Le test de normalité de Shapiro-Wilk est présent dans `scipy` avec la fonction `shapiro` du module `stats`. Par exemple pour l'appliquer avec un échantillon stocké dans la variable `x` :

```
from scipy import stats

stats.shapiro(x)
```

11.4 Test de Kolmogorov-Smirnov

Soit X une v.a. continue de loi F_X. On considère l'hypothèse nulle $H_0 : F_X = F_0$. Soit $X_1, \dots, X_n$ un échantillon iid de X et $\widehat{F}$ la fonction de répartition empirique correspondante :

$$\widehat{F}(x) = \frac{1}{n} \operatorname{card}\{i \in \{1, \dots, n\} / x_i \leq x\}.$$

Sous H_0, $\widehat{F}(x) \xrightarrow{\mathbb{P}} F_X(x)$, $\forall x$. Pour n assez grand, on s'attend donc à avoir $\widehat{F}(x) \approx F_X(x)$, $\forall x$. Le test de Kolmogorov-Smirnov exploite cette idée en se basant sur une mesure de distance entre les fonctions $\widehat{F}$ et F_0. On définit la statistique :

$$D_n = \sup_x \left|\widehat{F}(x) - F_0(x)\right|.$$

On montre que la loi de D_n sous H_0 est la même quelle que soit la distribution F_0 considérée. La RC au niveau α^* est

$$W = \{D_n > d_{n;1-\alpha^*}\},$$

$d_{n;1-\alpha^*}$ étant le fractile d'ordre $1 - \alpha^*$ de la loi de la statistique D_n sous H_0. Des tables donnent les valeurs du seuil critique pour toutes les valeurs de n.

Pour le calcul de D_n, on utilise le fait que le plus grand écart entre $\widehat{F}$ et F_0 se situe nécessairement en un point de discontinuité de $\widehat{F}$, c'est-à-dire pour une valeur x_i, $i = 1, \dots, n$. On a donc

$$D_n = \max_{1 \leq i \leq n} \max \left(\left| \widehat{F}(x_i) - F_0(x_i) \right|, \left| \widehat{F}(x_i^-) - F_0(x_i) \right| \right),$$

avec $\widehat{F}(x_i^-) = \lim_{x \to x_i^-} \widehat{F}(x)$.

Le test de Kolmogorov-Smirnov est intégré dans GNU R. Il faut au préalable créer la fonction de répartition à laquelle on veut se comparer. Par exemple pour se comparer à une loi exponentielle de paramètre d'intensité 12, avec un échantillon stocké dans la variable `x` :

```
FrepH0 <- function(x) pexp(x, rate=12)
ks.test(x, FrepH0)
```

Le test de Kolmogorov-Smirnov est présent dans `scipy` avec la fonction `kstest` du module `stats`. Il faut lui fournir la fonction de répartition à laquelle se comparer. Cette fonction de répartition est généralement obtenue à partir d'un objet `scipy` caractérisant une loi. Par exemple, pour se comparer à une loi exponentielle de paramètre d'intensité 12 (donc de paramètre d'échelle $1/12$), avec un échantillon stocké dans la variable `x` :

```
from scipy import stats

FrepH0 = stats.expon(scale=1/12).cdf
stats.kstest(x, FrepH0)
```

Test de Stephens Le test de Kolmogorov-Smirnov n'est utilisable que si H_0 est une hypothèse simple. Il ne permet pas, par exemple, de tester la normalité d'une v.a. X d'espérance et de variance inconnues. Une modification du test de Kolmogorov, proposée par Stephens, permet de résoudre ce problème.

Le principe du test consiste à calculer la statistique de Kolmogorov en remplaçant dans l'expression de F_0 les paramètres μ et σ^2 par leurs estimateurs $\overline{x}$ et s^{*2}. On définit :

$$D_n^* = \sup_x \left| \widehat{F}(x) - \Phi\left(\frac{x - \overline{x}}{s^*} \right) \right|.$$

La loi de D_n^* sous H_0 n'est plus la même que celle de la statistique de Kolmogorov. Stephens a proposé les RC suivantes :

— Au niveau de 5% :

$$W = \left\{ D_n^* \left(\sqrt{n} + \frac{0.85}{\sqrt{n}} - 0.01 \right) > 0.895 \right\}.$$

— Au niveau de 1% :

$$W = \left\{ D_n^* \left(\sqrt{n} + \frac{0.85}{\sqrt{n}} - 0.01 \right) > 1.035 \right\}.$$

Remarque 11.4. *Le test de Shapiro-Wilk et le test de Stephens recouvrent le même cas d'usage, toutefois, même s'ils contrôlent l'un et l'autre le risque de première espèce, le test de Stephens apparaît dans la pratique comme moins puissant, voir figure 11.3. On préfèrera donc utiliser le test de Shapiro-Wilk pour tester la normalité d'un échantillon dans le cas où sa moyenne et sa variance ne sont pas connues.*

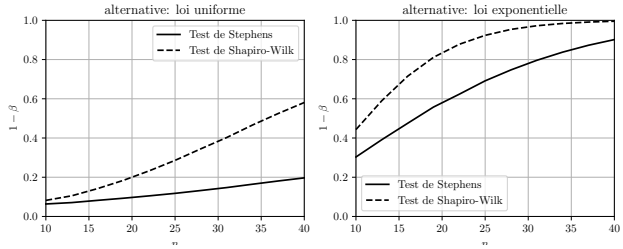


FIGURE 11.3 – Puissances obtenues par simulation des tests de Stephens et Shapiro-Wilk au niveau $\alpha^* = 0.05$ en fonction de n dans le cas de deux hypothèses alternatives.

Chapitre 12

Tests d'indépendance

L'objectif de ce chapitre est d'introduire les tests d'indépendance entre deux variables aléatoires.

Dans ce chapitre, on supposera l'existence de deux échantillons i.i.d. notés $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$, l'hypothèse nulle considérée sera H_0 : « $\forall i\ X_i \perp\!\!\!\perp Y_i$ » où $\perp\!\!\!\perp$ désigne l'indépendance. L'hypothèse alternative sera la complémentaire de l'hypothèse nulle.

Un *flowchart* est présenté figure 12.1 permettant d'appréhender les cas d'usages des différents tests d'indépendance.

12.1 Variables qualitatives : χ^2 de contingence

Le test du χ^2 peut être appliqué pour tester l'indépendance de deux v.a. qualitatives X et Y, respectivement à r et s modalités (X et Y peuvent avoir été obtenues en discrétisant des variables continues).

On dispose d'un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ et on considère l'hypothèse nulle H_0 : X et Y sont indépendantes.

On construit le *tableau de contingence* suivant :

$X \setminus Y$	1	…	j	…	J	
1			⋮			
⋮			⋮			
i	…	…	N_{ij}			$N_{i.}$
⋮						
I						
			$N_{.j}$			

où N_{ij} désigne le nombre de fois où X a pris la modalité i et Y la modalité j et

$$N_{i.} = \sum_{j=1}^{J} N_{ij}, \quad N_{.j} = \sum_{i=1}^{I} N_{ij}.$$

Soit $p_{ij} = \mathbb{P}(X = i; Y = j)$, $p_{i.} = \mathbb{P}(X = i)$ et $p_{.j} = \mathbb{P}(Y = j)$. Si X et Y sont indépendantes, alors $p_{ij} = p_{i.}p_{.j}$ et l'effectif théorique de chaque cas est $np_{i.}p_{.j}$. Par

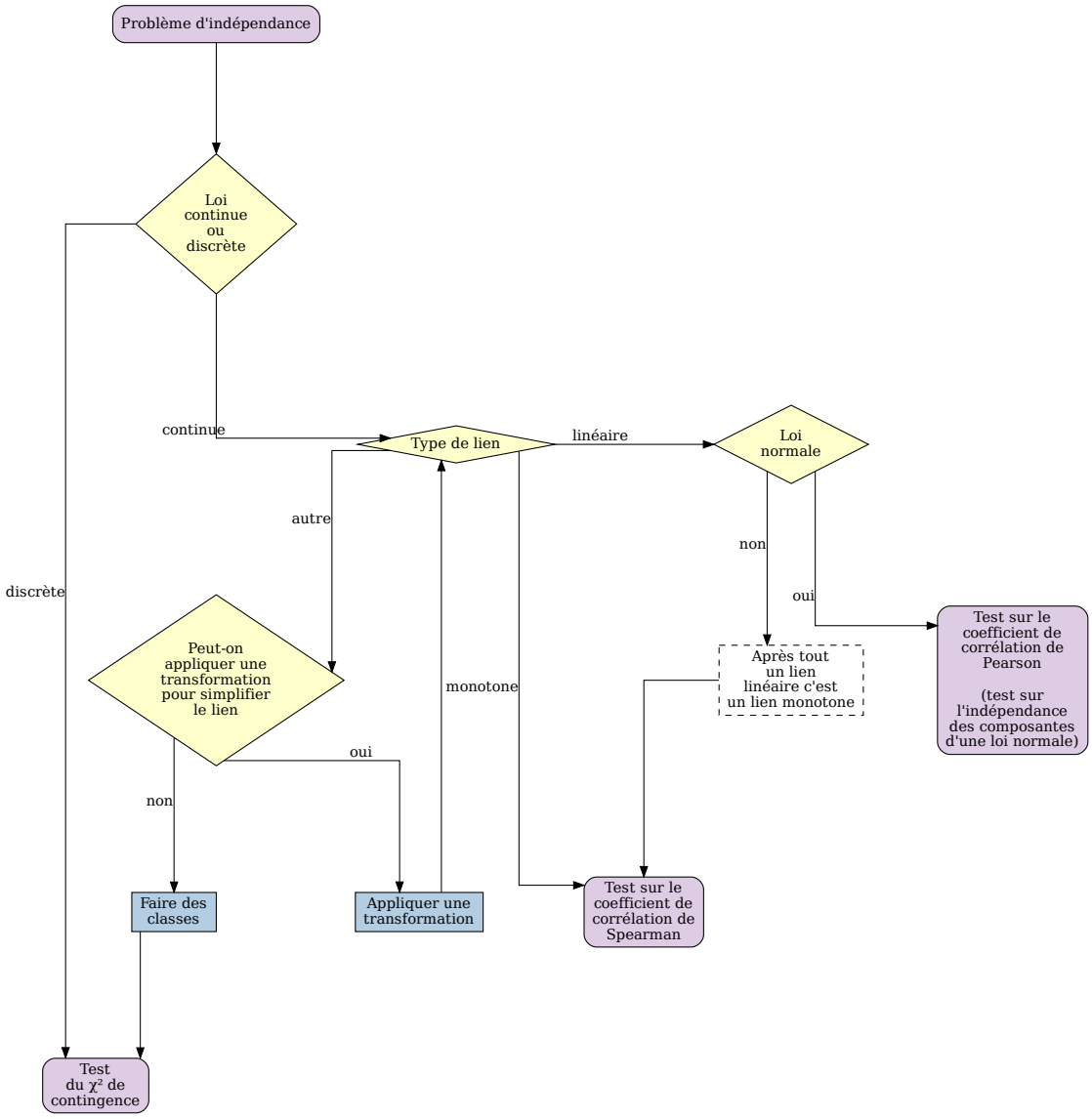


FIGURE 12.1 – *Flowchart* représentant les différents tests d'indépendance et les cas d'usages. Ce diagramme est donné à titre informatif, il ne remplace pas une lecture précise des conditions d'application de chaque test.

conséquent, la statistique

$$D^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(N_{ij} - np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}}$$

suit asymptotiquement sous H_0 une loi χ^2_{IJ-1}.

Toutefois, les valeurs $p_{i.}$ et $p_{.j}$ ne sont en général pas connues et doivent donc être estimées par $N_{i.}/n$ et $N_{.j}/n$. Il faut donc retirer $(I-1)+(J-1) = I+J-2$ d.d.l. Le nombre de d.d.l. devient donc :

$$IJ - 1 - I - J + 2 = IJ - I - J + 1 = (I-1)(J-1);$$

Par conséquent, on a approximativement, sous H_0 :

$$D^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{N_{ij}^2}{\frac{N_{i.}N_{.j}}{n}} - n \sim \chi^2_{(I-1)(J-1)}.$$

La RC du test du χ^2 est donc dans ce cas :

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

Le test du χ^2 de contingence est présent dans `GNU R` via `chisq.test`. Pour l'appliquer sur une matrice de contingence `M` et analyser les résultats :

```
M <- matrix(c(19,53,64,85),nrow=2,ncol=2)
res <- chisq.test(M)
res
res$expected # la prediction selon p[i,j]
res$observed # l'observation
```

Remarquons que via l'argument `simulate.p.value` on peut effectuer le test en utilisant la statistique de test du χ^2 en simulant la loi exacte, sans utiliser la loi asymptotique. Cette méthode pourra être utilisée lorsque les conditions sur les effectifs ne permettent pas d'approcher par l'asymptotique.

Le test du χ^2 de contingence est présent dans `scipy` avec la fonction `chi2_contingency` du module `stats`. Pour l'appliquer sur une matrice de contingence `M` et analyser les résultats :

```
import numpy as np
from scipy import stats

M = np.array([[19,53],[64,85]])
statD2, pval, df, expected = stats.chi2_contingency(M)
```

Pour des tableaux de contingence 2×2, le lecteur est invité à consulter les fonctions `fisher_exact` et `barnard_exact` du module `stats`. *Il est à noter que la fonction `barnard_exact` a été implémentée par un ancien étudiant de SY02 dans le*

cadre de ses études. Cf. scipy-13441[a].

a. https://github.com/scipy/scipy/pull/13441

12.2 Coefficient de corrélation de Pearson

Dans le cas où $(X_1, Y_1), \dots, (X_n, Y_n)$ est un échantillon iid d'une loi normale bidimensionnelle de v.a. parente (X, Y) (cf. section 3.5.3), une méthodologie pour tester la dépendance linéaire (aussi nommée *corrélation*) consiste à estimer le coefficient de corrélation théorique entre X et Y (cf. définition 3.10) par le *coefficient de corrélation de Pearson* défini par

$$R = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

Sous l'hypothèse H_0 d'indépendance, on peut approcher la loi d'une fonction de R par une loi de Student :

$$R\sqrt{\frac{n-2}{1-R^2}} \underset{H_0}{\sim} \mathcal{T}_{n-2}. \tag{12.1}$$

Notons que R s'éloigne de 0 quand le degré de corrélation entre les variables X et Y augmente. On considère donc la région critique suivante :

$$W = \left\{ \left| R\sqrt{\frac{n-2}{1-R^2}} \right| > t_{n-2,1-\frac{\alpha^*}{2}} \right\}.$$

Remarque 12.1. *Le coefficient de corrélation de Pearson ne doit être utilisé que quand le vecteur aléatoire (X, Y) suit au moins approximativement une loi normale bidimensionnelle.*

Remarque 12.2. *Le coefficient de corrélation de Pearson n'indique qu'une dépendance linéaire entre deux variables. Dans le cas de variables issues d'un même vecteur gaussien, l'indépendance linéaire est équivalente à l'indépendance.*

Remarque 12.3. *Le coefficient de détermination R^2 utilisé en régression linéaire est le carré du coefficient de corrélation de Pearson entre les x_i et les y_i (cf. section 7.5.4).*

Le test basé sur le coefficient de corrélation de Pearson est accessible dans `GNU R` en utilisant la fonction `cor.test`. Dans le cas où les deux échantillons sont stockés dans les variables `x` et `y`, on a :

```
cor.test(x, y, method='pearson')
```

Le test basé sur le cefficient de corrélation de Pearson est présent dans `scipy` sous la fonction `pearsonr` du module `stats`. Par exemple dans le cas où les échantillons sont stockés dans les variables `x` et `y`, on a :

```
from scipy import stats

r, pval = stats.pearsonr(x, y)
```

12.3 Coefficient de corrélation de Spearman

Dans le cas où $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ suivent des lois continues quelconques, une méthodologie pour tester la dépendance monotone est d'utiliser le *coefficient de corrélation de Spearman*. L'idée de ce coefficient est de travailler sur les rangs des variables X et Y. On notera $(R_1, \dots, R_n)$ les rangs de $(X_1, \dots, X_n)$, et $(S_1, \dots, S_n)$ les rangs de $(Y_1, \dots, Y_n)$.

Exemple 12.1. *Si* $(x_1, x_2, x_3) = (11.1, 13.2, 7.8)$ *alors* $(r_1, r_2, r_3) = (2, 3, 1)$.

Le coefficient de corrélation de Spearman entre $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ se définit comme le coefficient de corrélation de Pearson entre les rangs $(R_1, \dots, R_n)$ et $(S_1, \dots, S_n)$. Il peut se calculer simplement par la formule suivante :

$$R_s = 1 - \frac{6 \sum_i D_i^2}{n(n^2 - 1)}$$

avec $D_i = R_i - S_i$. Le coefficient R_s mesure le degré de dépendance monotone entre X et Y (croissante ou décroissante si R_s est respectivement positif ou négatif). La loi (12.1) du coefficient de Pearson sous H_0 peut encore être utilisée dès lors que $n \geq 10$. Cependant, une meilleure approximation est obtenue en utilisant la fonction pivotale approchée :

$$\left(\sqrt{\frac{n-3}{1.06}}\right) \frac{1}{2} \ln \frac{1 + R_s}{1 - R_s} \underset{H_0}{\sim} \mathcal{N}(0, 1).$$

Le coefficient de corrélation de Spearman s'éloignant de zéro lorsqu'une dépendance (monotone) est observée, on utilisera la région critique suivante :

$$W = \left\{ \left| \left(\sqrt{\frac{n-3}{1.06}}\right) \frac{1}{2} \ln \frac{1 + R_s}{1 - R_s} \right| > u_{1 - \frac{\alpha^*}{2}} \right\}$$

Remarque 12.4. *Le coefficient de corrélation de Spearman n'indique qu'une dépendance monotone entre deux variables. L'indépendance implique l'indépendance monotone, mais la réciproque est fausse. Deux variables peuvent être liées mais de manière non-monotone. Par exemple, les variables*

$$\begin{cases} X \sim \mathcal{U}([-\pi, \pi]) \\ Y = \cos(X) \end{cases}$$

ne sont pas indépendantes mais il n'existe aucune relation monotone entre ces deux variables.

Le test basé sur le coefficient de corrélation de Spearman est accessible dans `GNU R` en utilisant la fonction `cor.test`. Dans le cas où les deux échantillons sont stockés dans les variables `x` et `y`, on a

```
cor.test(x, y, method='spearman')
```

Le test basé sur le cefficient de corrélation de Spearman est présent dans `scipy` sous la fonction `spearmanr` du module `stats`. Par exemple dans le cas où les échantillons sont stockés dans les variables `x` et `y`, on a :

```
from scipy import stats

r, pval = stats.spearmanr(x, y)
```

Chapitre 13

Analyse de la variance

13.1 Le problème

L'analyse de la variance (en anglais : *ANOVA : Analysis of Variance*) a pour objet l'étude de l'effet de variables qualitatives sur une variable quantitative. Le vocabulaire utilisé est particulier : les variables qualitatives sont appelées *facteurs* et leurs modalités *niveaux*. Lorsqu'il y a plusieurs facteurs, une combinaison de modalités est appelée *traitement*. Nous nous limiterons ici à l'analyse de la variance à un facteur.

Exemple 13.1. *On veut comparer 3 types d'engrais. Pour ceci, on tire au hasard trois ensembles de parcelles sur lesquelles on utilise chacun des engrais. On mesure le rendement obtenu sur chacune des parcelles :*

- *échantillon 1 (engrais 1) : 35, 34, 33, 36, 37*
- *échantillon 2 (engrais 2) : 41, 38, 40, 40, 41*
- *échantillon 3 (engrais 3) : 35, 37, 34, 39, 35.*

Le choix de l'engrais a-t-il une influence sur le rendement ?

Formellement, l'analyse de la variance à un facteur consiste à tester l'égalité des espérances de K v.a. X_k ($k = 1, \dots, K$) supposées gaussiennes et de variance commune σ^2 inconnue. Il s'agit donc d'une généralisation du test de Student de comparaison de deux espérances (cf. paragraphe 10.2.3). On dispose pour chaque v.a. $X_k \sim \mathcal{N}(\mu_k, \sigma^2)$ d'un échantillon i.i.d. $X_k^1, \dots, X_k^{n_k}$ de taille n_k, et on note $N = n_1 + \cdots + n_K$ le nombre total d'observations. On suppose en outre l'indépendance entre les différents échantillons. Le problème de test considéré s'écrit :

$$\begin{cases} H_0 : & \mu_1 = \dots = \mu_K \\ H_1 : & \exists k, \ell \quad \mu_k \neq \mu_\ell. \end{cases}$$

13.2 Test du rapport de vraisemblance

Proposition 13.1. *La statistique du rapport de vraisemblance pour le problème ci-dessus s'écrit*

$$\Lambda(x_1^1, \dots, x_K^{n_K})^{2/N} = \frac{\sum_{k=1}^K \sum_{i=1}^{n_k} (x_k^i - \overline{x}_k)^2}{\sum_{k=1}^K \sum_{i=1}^{n_k} (x_k^i - \overline{x})^2}. \tag{13.1}$$

Preuve. cf. Section 13.6. □

Le numérateur et le dénominateur de l'équation (13.1) sont classiquement notés SSW (*Sum of Squares Within populations*) et SST (*Sum of Squares Total*). On parle en français de sommes des carrés intra-modalités et totale. L'équation (13.1) se simplifie encore en remarquant que

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i-\overline{x})^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i-\overline{x}_k+\overline{x}_k-\overline{x})^2 \tag{13.2}$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i-\overline{x}_k)^2 + \sum_{k=1}^{K} n_k(\overline{x}_k-\overline{x})^2. \tag{13.3}$$

L'équation (13.3) est appelée *équation d'analyse de la variance*, et le second terme du membre de droite est appelé somme des carrés inter-modalités, et noté SSB (*Sum of Squares Between populations*). On a donc

$$SST = SSW + SSB,$$

d'où l'on déduit

$$\Lambda(x_1^1,\dots,x_K^{n_K})^{2/N} = \frac{SSW}{SST} = \frac{SSW}{SSW+SSB} = \left(1+\frac{SSB}{SSW}\right)^{-1}.$$

La région critique du test du RV, de la forme

$$W=\{\Lambda<c\} \quad \text{peut donc écrire} \quad W=\left\{\frac{SSB}{SSW}>c'\right\}$$

pour une constante c' solution de l'équation $\mathbb{P}_{H_0}\left(\frac{SSB}{SSW}>c'\right)=\alpha^*$. Pour résoudre cette équation, il faut connaître la loi de SSB/SSW, ou d'une fonction simple de cette statistique. Remarquons tout d'abord que, d'après le théorème de Fisher :

$$\frac{SSW}{\sigma^2} = \sum_k \frac{(n_k-1)S_k^{*2}}{\sigma^2} \sim \chi^2_{N-K}.$$

D'autre part, sous l'hypothèse H_0,

$$\frac{SST}{\sigma^2} = \frac{N\hat{\sigma}_0^2}{\sigma^2} \sim \chi^2_{N-1}.$$

D'après l'équation d'analyse de la variance (13.3) et en admettant l'indépendance entre SSB et SSW, on en déduit

$$\frac{SSB}{\sigma^2} \overset{H_0}{\sim} \chi^2_{K-1}.$$

Notons $MSB = SSB/(K-1)$ et $MSW = SSW/(N-K)$. On a donc

$$F = \frac{MSB}{MSW} \sim \mathcal{F}_{K-1,N-K}.$$

On obtient finalement la RC du test d'analyse de la variance au niveau α^* :

$$W=\left\{F>f_{K-1,N-K;1-\alpha^*}\right\}.$$

13.3 Mise en œuvre du test

13.3.1 Vérification des hypothèses du modèle

L'analyse de la variance suppose que les K échantillons soient gaussiens et qu'ils aient la même variance. La première condition se vérifie à l'aide d'un test de normalité comme le test de Shapiro–Wilk (cf. section 11.3). Pour effectuer le test de l'égalité des variances

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_K \\ H_1 : \exists k, \ell \quad \sigma_k \neq \sigma_\ell, \end{cases}$$

différentes méthodes ont été proposées, parmi lesquelles on peut citer le *test de Bartlett*, qui repose sur la statistique suivante :

$$B = (N-K)\ln(MSW) - \sum_{k=1}^{K}(n_k - 1)\ln(S_k^{*2}).$$

Cette statistique suit approximativement sous H_0 une loi de χ^2 à $K-1$ degrés de liberté, et elle a tendance à prendre des valeurs plus élevées sous H_1. On rejettera donc H_0, au niveau de signification α^*, lorsque

$$b > \chi^2_{K-1;1-\alpha^*}.$$

13.3.2 Calculs et présentation des résultats

Le calcul de la statistique F repose uniquement sur les moyennes et variances empiriques corrigées des K échantillons. On peut mener les calculs selon la séquence suivante :

1. Calcul de $\overline{x}_k$ et s_k^{*2}, $k = 1, \ldots, K$; calcul de la moyenne empirique $\overline{x}$ des N observations par l'équation (13.5) ;
2. Calcul de $SSW = \sum_{k=1}^{K}(n_k - 1)s_k^{*2}$ et $MSW = SSW/(N-K)$;
3. Calcul de SSB et de $MSB = SSB/(K-1)$. On pourra utiliser la formule suivante :

$$\begin{aligned} SSB &= \sum_{k=1}^{K} n_k(\overline{x}_k - \overline{x})^2 \\ &= \sum_{k=1}^{K} n_k\overline{x}_k^2 - 2\overline{x}\sum_{k=1}^{K} n_k\overline{x}_k + \overline{x}^2\sum_{k=1}^{K} n_k \\ &= \sum_{k=1}^{K} n_k\overline{x}_k^2 - N\overline{x}^2. \end{aligned}$$

4. Calcul de $f = MSB/MSW$ et comparaison au seuil critique $f_{K-1,N-K;1-\alpha^*}$ ou calcul de la *p-value* (degré de signification) associée à f.

On présente souvent les résultats de ces différents calculs sous la forme d'un tableau appelé « tableau d'analyse de la variance », semblable à celui-ci :

Source de variation	Deg. de liberté	Som. des carrés	Moyenne des carrés	Rapport	*p-value*
Inter-moda. Intra-moda.	$K-1$ $N-K$	SSB SSW	$MSB = \frac{SSB}{K-1}$ $MSW = \frac{SSW}{N-K}$	$f = \frac{MSB}{MSW}$	$1 - F_{\mathcal{F}_{K-1,N-K}}(f_{\text{obs}})$
total	$N-1$	SST			

13.3.3 Étude de l'effet du facteur : comparaisons multiples

Quand l'analyse de la variance met en évidence un effet significatif, il peut être intéressant d'aller plus loin pour savoir quels sont les niveaux significativement différents. On ne peut se contenter de comparer les échantillons deux à deux par des tests de Student, car il faudrait effectuer autant de tests que de paires de niveaux du facteur, soit $K(K-1)/2$: bien que le risque de commettre une erreur de première espèce pour une paire de niveau donnée serait de α^*, le risque d'en commettre au moins une sur toute la procédure serait supérieur à α^*.

Pour pallier ce problème, plusieurs méthodes ont été proposées, parmi lesquelles la procédure LSD (*least significant differences*) de Fisher. Cette procédure est la suivante :

- Si le test d'analyse de la variance conclut à l'acceptation de H_0, on en déduit que le facteur n'a pas d'effet et la procédure s'arrête.
- Si le test d'analyse de la variance a mis en évidence un effet significatif du facteur (rejet de H_0), on teste l'égalité des moyennes deux à deux, pour chaque paire de niveaux du facteur. La procédure est alors similaire au test de Student de comparaison de deux populations, mais on utilise ici MSW comme estimateur de la variance. La région critique pour la comparaison des niveaux k et ℓ est :

$$W_{k\ell} = \left\{ \frac{|\overline{X}_k - \overline{X}_\ell|}{\sqrt{MSW\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)}} > t_{N-K,1-\frac{\alpha^*}{2}} \right\}.$$

Une autre solution consiste à utiliser la *correction de Bonferroni*, qui consiste à effectuer les tests de comparaison pour chaque paire au niveau

$$\tilde{\alpha}^* = \frac{\alpha^*}{\frac{K(K-1)}{2}}$$

pour contrôler par α^* le risque de se tromper au moins une fois sous H_0 (c'est-à-dire pour faire en sorte que ce risque soit proche de α^*, en restant inférieur à cette valeur).

13.4 Exemple

On reprend l'exemple 13.1 de l'introduction.

13.4.1 Vérifications des hypothèses du modèle

Normalité des échantillons La statistique $\mathscr{W}$ du test de Shapiro-Wilk vaut respectivement pour les 3 échantillons 0.98676, 0.83274 et 0.90531. La borne à 5% vaut $w_{5;0.05} = 0.775$ et est inférieure à ces trois statistiques ce qui permet d'accepter l'hypothèse de normalité au niveau 5%.

Égalité des variances On a $s_1^{*2} = 2.5$, $s_2^{*2} = 1.5$ et $s_3^{*2} = 4$. On en déduit : $b = 12\ln(2.67) - 4[\ln(2.5) + \ln(1.5) + \ln(4)] = 0.95$. Cette valeur est inférieure au seuil critique au niveau 5%, égal à $\chi^2_{2;0.95} = 5.99$. On accepte donc l'hypothèse d'égalité des variances.

13.4.2 Tableau d'analyse de la variance

Source de variation	degrés de liberté	Somme des carrés	Moyenne des carrés	Rapport
Inter-modalité	2	70	35	13.1
Intra-modalité	12	32	2.67	
total	14	102		

$$\frac{MSB}{MSW} = \frac{35}{2.67} = 13.125 \quad \text{et} \quad f_{2,12;0.95} = 3.89.$$

Le facteur a donc un effet significatif au niveau de 5 %.

13.4.3 Comparaisons multiples

Posons

$$t_{k\ell} = \frac{|\overline{x}_k - \overline{x}_\ell|}{\sqrt{MSW\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)}}.$$

On a $t_{12} = 4.84$, $t_{13} = 0.96$ et $t_{23} = 3.87$. Au niveau de 5%, ces valeurs sont à comparer au seuil $t_{12,0.975} = 2.179$. On constate qu'il y a une différence significative entre les niveaux 1 et 2, et entre les niveaux 2 et 3. En revanche, la différence entre les niveaux 1 et 3 n'est pas significative.

13.5 Le test de Kruskal-Wallis

C'est un équivalent non paramétrique de l'analyse de la variance, qui ne suppose pas la normalité des échantillons. Soient $X_k^1, \dots, X_k^{n_k}$, $(k = 1, \dots, K)$ K échantillons, de v.a. parentes respectives $X_1, \dots, X_K$. Soit F_k la fonction de répartition de X_k. On teste l'hypothèse selon laquelle les K échantillons sont issus de la même loi, c'est-à-dire que l'on considère le problème de test suivant :

$$H_0 : F_1 = \cdots = F_K$$
$$H_1 : \exists k, \ell \text{ t.q. } F_k \neq F_\ell.$$

Le test de Kruskal-Wallis repose sur la statistique $H = SSB/MST$, cette statistique étant calculée sur les rangs.

De manière générale, soit R_k^i le rang de l'observation i de l'échantillon k, et N le nombre total d'observations dans les K échantillons. On remarque que

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} R_k^i = \frac{N(N+1)}{2}$$

et

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} (R_k^i)^2 = \frac{N(N+1)(2N+1)}{6}.$$

Ceci permet de simplifier les expressions de SSB et SST. On obtient en effet :

$$SSB = \sum_{k=1}^{K} \frac{1}{n_k} \left(\sum_{i=1}^{n_k} R_k^i \right)^2 - \frac{N(N+1)^2}{4},$$

$$SST = \frac{N(N+1)(N-1)}{12},$$

et donc

$$MST = \frac{SST}{N-1} = \frac{N(N+1)}{12}.$$

Par conséquent :

$$H = \frac{SSB}{MST} = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{1}{n_k} \left(\sum_{i=1}^{n_k} R_k^i \right)^2 - 3(N+1).$$

On montre que, si N n'est pas trop petit, H suit approximativement sous l'hypothèse H_0 une loi χ^2_{K-1}. La région critique du test est donc :

$$W = \{h > \chi^2_{K-1;1-\alpha^*}\}.$$

13.6 Preuve de la Proposition 13.1

On dispose d'un échantillon indépendant de $N = \sum_{k=1}^{K} n_k$ observations noté $X_1^1, \dots, X_1^{n_1}, \dots, X_K^1, \dots, X_K^{n_K}$, avec

$$X_k^i \sim \mathcal{N}(\mu_k, \sigma^2), \quad k = 1, \dots, K, \ i = 1, \dots, n_k.$$

La fonction de vraisemblance associée à cet échantillon est :

$$L(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_k^i - \mu_k)^2 \right).$$

La statistique du maximum de vraisemblance s'écrit donc :

$$\Lambda(x_1^1, \dots, x_K^{n_K}) = \frac{\sup_{(\mu_1,\dots,\mu_K,\sigma^2)\in H_0} L(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K})}{\sup_{(\mu_1,\dots,\mu_K,\sigma^2)\in \mathbb{R}^K \times \mathbb{R}_+} L(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K})}. \tag{13.4}$$

Sous H_0, l'échantillon complet des N observations est un échantillon gaussien i.i.d de v.a. parente $X \sim \mathcal{N}(\mu, \sigma^2)$, avec $\mu = \mu_1 = \dots = \mu_K$. Les EMV de μ et de σ^2 sont donc respectivement, dans ce cas :

$$\hat{\mu} = \overline{X} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_k^i = \frac{1}{N} \sum_{k=1}^{K} n_k \overline{X}_k, \tag{13.5}$$

avec

$$\overline{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_k^i, \quad \text{et} \quad \hat{\sigma}_0^2 = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_k^i - \overline{X})^2.$$

Le numérateur du RV est donc égal à $L(\overline{x}, \dots, \overline{x}, \hat{\sigma}_0^2; x_1^1, \dots, x_K^{n_K})$. Pour le calcul du dénominateur, il suffit de calculer les EMV non restreints de $\mu_1, \dots, \mu_K$ et σ^2. La log-vraisemblance s'écrit :

$$\ell(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K}) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{n_i}(x_k^i - \mu_k)^2.$$

Le système des $K+1$ équations de vraisemblance s'écrit donc :

$$\left\{\begin{array}{rcl} \dfrac{\partial \ell}{\partial \mu_k}(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K}) & = & 0, \quad \forall k = 1, K \\ \dfrac{\partial \ell}{\partial \sigma^2}(\mu_1, \dots, \mu_K, \sigma^2; x_1^1, \dots, x_K^{n_K}) & = & 0 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl} \dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n_k} 2(x_k^i - \mu_k) & = & 0, \quad \forall k = 1, K \\ -\dfrac{N}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i - \mu_k)^2 & = & 0 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n_k}(x_k^i - \mu_k) & = & 0, \quad \forall k = 1, K \\ \dfrac{N}{2\sigma^4}\left(\dfrac{1}{N}\displaystyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i - \mu_k)^2 - \sigma^2\right) & = & 0 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rcl} \mu_k & = & \overline{x}_k, \quad \forall k = 1, K \\ \sigma^2 & = & \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i - \overline{x}_k)^2. \end{array}\right.$$

On en déduit les estimateurs :

$$\left\{\begin{array}{rcl} \hat{\mu}_k & = & \overline{X}_k, \quad \forall k = 1, K \\ \hat{\sigma}^2 & = & \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_k^i - \overline{X}_k)^2 = \frac{1}{N}\sum_{k=1}^{K}(n_k - 1)S_k^{*2}, \end{array}\right.$$

où S_k^{*2} est la variance empirique corrigée de l'échantillon de la v.a. X_k.

On en déduit l'expression de Λ :

$$\Lambda(x_1^1, \dots, x_K^{n_K}) = \frac{L(\overline{x}, \dots, \overline{x}, \hat{\sigma}_0^2; x_1^1, \dots, x_K^{n_K})}{L(\overline{x}_1, \dots, \overline{x}_K, \hat{\sigma}^2; x_1^1, \dots, x_K^{n_K})},$$

qui se simplifie en :

$$\Lambda(x_1^1, \dots, x_K^{n_K}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{N/2},$$

ce que l'on peut encore écrire :

$$\Lambda(x_1^1, \dots, x_K^{n_K})^{2/N} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i - \overline{x}_k)^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_k^i - \overline{x})^2}.$$

Références bibliographiques

Le lecteur désireux d'approfondir les notions vues en cours ou de les revoir sous un éclairage différent pourra utilement consulter les ouvrages suivants, disponibles à la BUTC :

[1] G. Calot. *Cours de calcul des probabilités*, Dunod, Paris, 1967. *(Un cours classique de probabilités en langue française, ne faisant appel qu'à des connaissances mathématiques de premier cycle).*

[2] V. Girardin et N. Limnios. *Probabilités*, Vuibert, Paris, 2001. *(Ouvrage très complet sur la théorie des probabilités, adoptant une présentation plus moderne et plus rigoureuse que dans l'ouvrage précédent. Contient un chapitre sur la Statistique Mathématique).*

[3] J.-J. Daudin, C. Vuillet et S. Robin. *Statistique inférentielle*, Socité Française de Statistique et Presses Universitaires de Rennes, 1999. *Ouvrage servant de support au cours de statistique de première année à AgroParisTech, avec un programme très similaire à SY02. Cet ouvrage est recommandé pour avoir une approche différente des mêmes notions.*

[4] G. Saporta. *Probabilités, analyse des données et statistiques, 2e édition*, Technip, Paris, 2006. *(Ouvrage très complet traitant de théorie des probabilités, de statistique inférentielle, et d'analyse des données).*

[5] P. Tassi. *Méthodes statistiques*, Economica, Paris, 1992. *(Traité de statistique inférentielle de bon niveau mais néanmoins accessible).*

[6] J.-P. Lecoutre, S. Legait-Maille et P. Tassi. *Statistique. Exercices corrigés avec rappels de cours*, Masson, Paris, 1993. *(Un très bon recueil d'exercices recommandé pour la préparation des examens).*

[7] E. J. Dudewicz et S. N. Mishra. *Modern Mathematical Statistics*, Wiley, New-York, 1988.*(Excellent ouvrage de langue anglaise traitant de probabilités et de statistique ; on y trouvera de nombreux prolongements aux questions abordées en SY02.).*

Index

analyse de la variance, 88, 167
anova, 167
atypique, 13

biais, 63
borne
    de Fréchet, 69
boxplot, 19
boîte
    à moustaches, 19

caractère, 9
coefficient
    de détermination, 98
coefficient de corrélation
    de Pearson, 21, 164
    de Spearman, 21, 165
    théorique, 42
coefficients
    de Fisher, 37
comparaison multiple, 170
conditions
    de Cramer-Rao, 69
convergence
    en loi, 44
    en probabilité, 43
covariance, 41, 42
    empirique, 87
critère
    des moindres carrés, 87

degré de signification, 108
diagramme
    en boîte, 19
    en tige et feuilles, 13
    quantile-quantile, 154
    à bandes, 11
distribution, 10
    de probabilité, 24, 25
    empirique, 10
droite des moindres carrés, 88

écart-type, 37
    empirique, 18
échantillon, 10, 23
    aléatoire, 51
    gaussien, 57
    iid (indépendant, identiquement distribué), 51
échantillonnage, 10
    aléatoire simple, 23
efficacité, 69
équations
    de vraisemblance, 67
erreur
    quadratique
        moyenne, 64
espérance, 40
    mathématique, 36
estimateur, 63
    asymptotiquement sans biais, 63
    convergent, 64
    des moindres carrés, 88
    du maximum de vraisemblance, 66
    efficace, 69
    optimal, 70
    sans biais, 63
    sans biais de variance minimale, 70
estimation, 63
    du maximum de vraisemblance, 66
    par intervalle de confiance, 75
étendue, 18
    inter-quartile, 18
événement, 24
    élémentaire, 23
expérience
    aléatoire, 23

facteur, 167
fonction
    Φ, 32
    de probabilité, 25
    de puissance, 107

de répartition, 10, 25
empirique, 14, 57, 157
théorique, 157
de vraisemblance , 66
Gamma, 45
pivotale, 76
score, 67
fractile, 17, 38
empirique, 58
théorique, 58
fréquence
cumulée, 11
relative, 11

histogramme, 12
hypothèse
complémentaire, 106
simple, 106

individu, 9, 23
indépendance, 42
information
de Fisher, 67
inférence statistique, 10
intervalle
de confiance, 75
de prédiction, 100
invariance fonctionnelle, 67
inégalité
de Bienaymé-Tchebycheff, 37, 44
de Cauchy-Schwarz, 41, 73
de Fréchet-Darmois-Cramer-Rao, 69

loi
binomiale, 26, 55
conditionnelle, 43
continue
uniforme, 30, 37
de Bernoulli, 26
de Fisher, 49, 139
de Gauss, 32
de Poisson, 28
de probabilité, 25
de Student, 47
décentrée, 48
des grands nombres, 53
du χ^2, 45
exponentielle, 34
jointe, 39
marginale, 39
multinomiale, 151
normale, 32, 55
bidimensionnelle, 40
centrée-réduite, 32

matrice
de variance, 41
maximum de vraisemblance, 66
modalité, 9
moment, 37, 40
centré, 37
empirique, 56, 65
théorique, 37, 65
moyenne
empirique, 15, 52
tronquée, 16
multinomiale, 40
médiane, 16
méthode
des moments, 65

niveau, 167
du test, 107

p-value, 108
Permutations, *voir* test de permutations
population, 9, 23
précision, 64
puissance, 107
pvalue, 108

Q-Q plot, 154
quantile, 17, 38
empirique, 58
théorique, 58

rapport de vraisemblance, 114
recensement, 10
règle
de Sturges, 12
risque
de 1re espèce, 107
de 2e espèce, 107
quadratique, 64
région
critique, 106
d'acceptation, 106
régression
linéaire, 85
résidu, 88

statistique, 52
  D_n, 157
  de test, 106
  descriptive, 10
  inférentielle, 10
support, 67

tableau
  d'analyse de la variance, 97, 169
  de contingence, 161
  de fréquences, 12
taille
  de l'échantillon, 51
  minimale, 82
test
  corrélation de Pearson, 164
  corrélation de Spearman, 165
  d'adéquation, 151
  d'ajustement, 151
  d'hypothèses, 105
  d'indépendance, 161
  de Bartlett, 169
  de contingence, 161
  de Fisher, 139
  de Kolmogorov-Smirnov, 157
  de Kruskal-Wallis, 171
  de Neyman-Pearson, 109
  de normalité, 156, 158
  de permutations, 141
  de Shapiro-Wilk, 156
  de Stephens, 158
  de Student, 121, 135
  de Wilcoxon signé, 147
  de Wilcoxon-Mann-Whitney, 143
  du χ^2, 151, 161
  du rapport de vraisemblance, 114
  du signe, 145
  sur l'espérance, 110, 114, 115, 121, 122
  sur la variance, 124
  sur les coefficients de régression, 123
  sur une proportion, 125
  uniformément plus puissant, 108
  UPP, 108
théorème
  central limite, 53
  de Fisher, 57
  de Fréchet-Darmois-Cramer-Rao, 69
  de Moivre-Laplace, 53
  de Neyman-Pearson, 109
  de Slutsky, 44
  de transfert, 36, 41
  de Wilks, 117

variable, 9
  aléatoire, 24
  continue, 9, 26
  discrète, 9, 11, 25
  dépendante, 85
  explicative, 85
  indépendante, 85
  parente, 51
  qualitative, 9
    nominale, 9
    ordinale, 9
  quantitative, 9
  à expliquer, 85
variance, 36
  empirique, 18, 55
    corrigée, 18
  expliquée, 89
  résiduelle, 89
  totale, 89
vecteur
  aléatoire, 38
vraisemblance, 66

Wilcoxon, *voir* test de Wilcoxon-Mann-Whitney, *voir* test de Wilcoxon signé
Wilcoxon-Mann-Whitney, *voir* test de Wilcoxon-Mann-Whitney

échantillonnage
  stratifié, 59