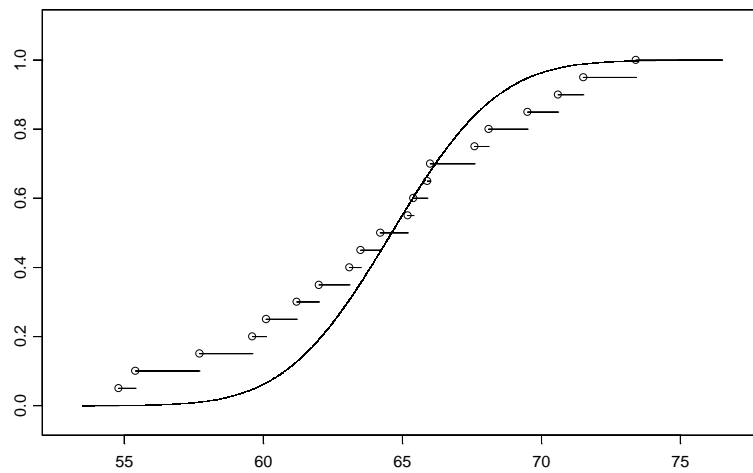


Ensimag - 2^{ème} année



Statistique Inférentielle Avancée

Notes de cours

Olivier Gaudoin

Table des matières

1	Introduction	7
2	Concepts de l'inférence statistique	9
2.1	Le modèle statistique	9
2.2	Modèle paramétrique ou non paramétrique	10
2.3	Fonction de vraisemblance	11
2.4	Statistiques	13
2.5	Exhaustivité	14
2.6	La famille exponentielle	18
3	Estimation paramétrique optimale	23
3.1	Introduction	23
3.2	Réduction de la variance	23
3.3	Complétude	27
3.4	L'estimation sans biais et de variance minimale	28
3.5	Information de Fisher et efficacité	29
3.5.1	Score et matrice d'information	30
3.5.2	Information et exhaustivité	32
3.5.3	Borne de Cramer-Rao et efficacité	33
4	Maximum de vraisemblance et estimation bayésienne	37
4.1	Introduction	37
4.2	Propriétés asymptotiques de l'estimateur de maximum de vraisemblance	37
4.3	Intervalle de confiance asymptotiques	41
4.3.1	Cas d'un paramètre réel	41
4.3.2	Cas d'un paramètre vectoriel	44
4.4	Estimation bayésienne	44
4.4.1	Principe de la méthode	44
4.4.2	Exemple du contrôle de qualité	46
5	Tests d'hypothèses optimaux	49
5.1	Introduction	49
5.2	Définitions	49
5.3	Tests d'hypothèses simples	50
5.4	Tests d'hypothèses composites	55
5.5	Test du rapport des vraisemblances maximales	56

6	Estimation non paramétrique de quantités réelles	59
6.1	Les outils de la statistique non paramétrique	59
6.1.1	Statistiques d'ordre et de rang	59
6.1.2	Loi de probabilité empirique	64
6.2	Estimation de l'espérance d'un échantillon	65
6.2.1	Estimation ponctuelle	65
6.2.2	Intervalle de confiance	65
6.3	Estimation de la variance d'un échantillon	67
6.3.1	Estimation ponctuelle	67
6.3.2	Intervalle de confiance	67
6.3.3	Lien entre moyenne et variance empiriques	68
6.4	Estimation des moments de tous ordres	68
6.5	Estimation des quantiles	69
6.5.1	Propriétés des quantiles empiriques	69
6.5.2	Estimation ponctuelle	70
6.5.3	Intervalle de confiance	70
6.6	Lois asymptotiques des extrêmes	71
7	Estimation fonctionnelle	73
7.1	Estimation de la fonction de répartition	74
7.1.1	Estimation ponctuelle	74
7.1.2	Intervalle de confiance	76
7.2	Estimation de la densité	77
7.2.1	Rappels sur les histogrammes	77
7.2.2	La méthode du noyau	78
8	Tests d'adéquation basés sur la fonction de répartition empirique	83
8.1	Problématique des tests d'adéquation	83
8.2	Rappels sur les graphes de probabilité	84
8.3	Cas d'une loi entièrement spécifiée	85
8.4	Cas d'une famille de lois	87
9	Tests non paramétriques sur un échantillon	91
9.1	Tests d'échantillon	91
9.1.1	Le test de Spearman	92
9.1.2	Le test de Kendall	94
9.2	Tests sur l'espérance et la médiane	95
9.2.1	Tests asymptotiques sur l'espérance	95
9.2.2	Tests sur la médiane	97
10	Tests non paramétriques sur plusieurs échantillons	101
10.1	Test de Kolmogorov-Smirnov	101
10.2	Tests de rang	102
10.2.1	Le test de la médiane	102
10.2.2	Le test de Wilcoxon-Mann-Whitney	104
10.2.3	Le test de Kruskal-Wallis	105

11 Annexe A : Rappels de probabilités pour la statistique	107
11.1 Variables aléatoires réelles	107
11.1.1 Loi de probabilité d'une variable aléatoire	107
11.1.2 Variables aléatoires discrètes et continues	108
11.1.3 Moments et quantiles d'une variable aléatoire réelle	109
11.2 Vecteurs aléatoires réels	110
11.2.1 Loi de probabilité d'un vecteur aléatoire	110
11.2.2 Espérance et matrice de covariance d'un vecteur aléatoire	111
11.3 Convergences et applications	112
11.4 Quelques résultats sur quelques lois de probabilité usuelles	113
11.4.1 Loi binomiale	113
11.4.2 Loi géométrique	114
11.4.3 Loi de Poisson	114
11.4.4 Loi exponentielle	114
11.4.5 Loi gamma et loi du chi-2	114
11.4.6 Loi normale	115
11.4.7 Lois de Student et de Fisher-Snedecor	116
12 Annexe B : Lois de probabilité usuelles	117
12.1 Caractéristiques des lois usuelles	117
12.1.1 Variables aléatoires réelles discrètes	117
12.1.2 Variables aléatoires réelles continues	118
12.1.3 Vecteurs aléatoires dans \mathbb{N}^d et dans \mathbb{R}^d	119
12.2 Tables de lois	120
12.2.1 Table 1 de la loi normale centrée réduite	120
12.2.2 Table 2 de la loi normale centrée réduite	121
12.2.3 Table de la loi du χ^2	122
12.2.4 Table de la loi de Student	123
12.2.5 Tables de la loi de Fisher-Snedecor	124
13 Annexe C : Introduction à R	127
13.1 Les bases de R	127
13.2 Commandes pour les deux premiers TD en R	128
13.3 Quelques commandes utiles de R	129
13.4 Les lois de probabilité usuelles en R	130
13.5 Les principaux tests d'hypothèses en R	132
13.6 Les graphiques dans R	132
13.6.1 Graphique simple	132
13.6.2 Autres fonctions graphiques	133
13.6.3 Paramétrage de la commande plot	134
Bibliographie	135

Chapitre 1

Introduction

Comme son nom l'indique, le cours de premier semestre de Principes et Méthodes Statistiques (PMS) a présenté les principes et les méthodes de base d'une analyse statistique de données. On peut résumer rapidement son contenu de la façon suivante :

- **Statistique descriptive** : le but est de décrire et résumer l'information contenue dans les données à l'aide de représentations graphiques (diagrammes en bâtons, histogrammes, graphes de probabilité) et d'indicateurs statistiques (moyenne, variance, médiane, quantiles, ...). Tous les exemples vus portent sur des données unidimensionnelles. L'extension à des descriptions de données multidimensionnelles sera vue dans le cours d'Analyse Statistique Multidimensionnelle (ASM).
- **Statistique inférentielle** : le but est de faire des prévisions et prendre des décisions au vu des données. Nous avons vu deux grandes catégories de méthodes :
 - L'**estimation**, ponctuelle et par intervalles de confiance, avec la méthode des moments et la méthode du maximum de vraisemblance.
 - Les **tests d'hypothèses**, avec les tests paramétriques sur un ou deux échantillons et les tests du χ^2 .

Le but du cours de Statistique Inférentielle Avancée (SIA) est d'approfondir et d'étendre ces notions, en allant plus loin dans la théorie mathématique sous-jacente.

Nous commencerons par donner des concepts généraux sur l'inférence statistique, en introduisant la notion de **modèle statistique**. Puis nous étudierons des propriétés d'optimalité des notions déjà étudiées : comment trouver un **estimateur optimal** ? Qu'est-ce qu'un **test optimal** et comment le trouver ? Nous étudierons une nouvelle méthode d'estimation, l'**estimation bayésienne**, qui ouvre un champ très important de la statistique moderne.

Nous distinguerons la **statistique paramétrique**, qui suppose l'existence d'un modèle connu avec des paramètres inconnus, et la **statistique non paramétrique**, qui ne fait pas ces hypothèses. Dans ce contexte, nous verrons comment **estimer des fonctions de répartition et des densités de probabilité**.

Enfin, nous étudierons des **tests non paramétriques**, permettant de déterminer si des observations sont indépendantes et de même loi ou présentent une tendance, de tester une moyenne ou de comparer des échantillons sans faire d'hypothèses sur un modèle sous-jacent, ou de tester l'adéquation d'un modèle.

Nous établirons des propriétés sur des **paramètres à plusieurs dimensions** (avec la notion de matrice d'information au lieu de celle de quantité d'information) et étudierons des **résultats asymptotiques** (optimalité asymptotique de l'estimateur de maximum de vraisemblance).

Chapitre 2

Concepts de l'inférence statistique

2.1 Le modèle statistique

Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire.

Une expérience statistique consiste à recueillir une observation x d'un élément aléatoire X , à valeurs dans un espace \mathcal{X} et dont on ne connaît pas exactement la loi de probabilité P . Des considérations de modélisation du phénomène observé amènent à admettre que P appartient à une famille \mathcal{P} de lois de probabilité possibles.

Définition 1 : *Le modèle statistique (ou la structure statistique) associé à cette expérience est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où :*

- \mathcal{X} est l'espace des observations, ensemble de toutes les observations possibles.
- \mathcal{A} est la tribu des événements observables associée.
- \mathcal{P} est une famille de lois de probabilités possibles définie sur \mathcal{A} .

L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

On dit que le modèle est **discret** quand \mathcal{X} est fini ou dénombrable. Dans ce cas, la tribu \mathcal{A} est l'ensemble des parties de \mathcal{X} : $\mathcal{A} = \mathcal{P}(\mathcal{X})$. C'est le cas quand l'élément aléatoire observé X a une loi de probabilité discrète.

On dit que le modèle est **continu** quand $\mathcal{X} \subset \mathbb{R}^p$ et $\forall P \in \mathcal{P}$, P admet une densité (par rapport à la mesure de Lebesgue) dans \mathbb{R}^p . Dans ce cas, \mathcal{A} est la tribu des boréliens de \mathcal{X} (tribu engendrée par les ouverts de \mathcal{X}) : $\mathcal{A} = \mathcal{B}(\mathcal{X})$.

On peut aussi envisager des modèles ni continus ni discrets, par exemple si l'observation a certains éléments continus et d'autres discrets. \mathcal{X} et \mathcal{A} sont alors plus complexes.

Le cas le plus fréquent, celui qui a été principalement vu en PMS, est celui où l'élément aléatoire observé est constitué de variables aléatoires indépendantes et de même loi (i.i.d.) : $X = (X_1, \dots, X_n)$, où les X_i sont i.i.d. On dit que l'on a alors un **modèle d'échantillon**. Dans ce cas, par convention, si on note $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ le modèle correspondant à un échantillon de taille 1, on notera $(\mathcal{X}, \mathcal{A}, \mathcal{P})^n$ le modèle correspondant à un échantillon de taille n .

Exemple 1 : ampoules. L'exemple de référence du cours de PMS a consisté à recueillir les durées de vie, supposées indépendantes et de même loi exponentielle, de n ampoules électriques. L'observation est de la forme $x = (x_1, \dots, x_n)$, où les x_i sont des réalisations de variables aléatoires X_i indépendantes et de même loi exponentielle de paramètre λ inconnu.

Pour tout i , $x_i \in \mathbb{R}^+$, donc l'espace des observations est $\mathcal{X} = \mathbb{R}^{+n}$. Alors la tribu associée est $\mathcal{A} = \mathcal{B}(\mathbb{R}^{+n})$. Le modèle est continu. Comme on admet que la loi est exponentielle mais que son paramètre est inconnu, l'ensemble des lois de probabilités possibles pour chaque X_i est $\{exp(\lambda); \lambda \in \mathbb{R}^+\}$. Comme les X_i sont indépendantes, la loi de probabilité du vecteur (X_1, \dots, X_n) est la loi produit $\mathcal{P} = \{exp(\lambda)^{\otimes n}; \lambda \in \mathbb{R}^+\}$, ensemble des lois de probabilité des vecteurs aléatoires de taille n dont les composantes sont indépendantes et de même loi exponentielle de paramètre inconnu.

Finalement, le modèle statistique associé est :

$$(\mathbb{R}^{+n}, \mathcal{B}(\mathbb{R}^{+n}), \{exp(\lambda)^{\otimes n}; \lambda \in \mathbb{R}^+\})$$

qu'on peut aussi écrire, d'après la convention énoncée :

$$(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{exp(\lambda); \lambda \in \mathbb{R}^+\})^n.$$

Exemple 2 : contrôle de qualité. Une chaîne de production produit un très grand nombre de pièces et on s'intéresse à la proportion inconnue de pièces défectueuses. Pour l'estimer, on prélève indépendamment n pièces dans la production et on les contrôle. L'observation est $x = (x_1, \dots, x_n)$, où :

$$x_i = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ pièce est défectueuse} \\ 0 & \text{sinon} \end{cases}$$

Par conséquent, l'espace des observations est $\mathcal{X} = \{0, 1\}^n$. Il est fini, donc le modèle est discret et $\mathcal{A} = \mathcal{P}(\{0, 1\}^n)$. Les X_i sont indépendants et de même loi de Bernoulli $\mathcal{B}(p)$, où $p = P(X_i = 1)$ est la probabilité qu'une pièce soit défectueuse.

Alors le modèle statistique peut s'écrire :

$$(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), \{\mathcal{B}(p)^{\otimes n}; p \in [0, 1]\})$$

ou

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\mathcal{B}(p); p \in [0, 1]\})^n.$$

Remarque : Quand l'élément aléatoire X est numérique, il admet une fonction de répartition F . La fonction de répartition caractérisant une loi de probabilité, l'ensemble \mathcal{P} des lois de probabilité possibles pour X est en bijection avec l'ensemble \mathcal{F} des fonctions de répartition possibles. Aussi le modèle statistique peut dans ce cas être noté $(\mathcal{X}, \mathcal{A}, \mathcal{F})$ au lieu de $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

2.2 Modèle paramétrique ou non paramétrique

Un **modèle paramétrique** est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu, de dimension d . Alors, la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P} = \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\}$.

C'est évidemment le cas des deux exemples. Le problème principal est alors de faire de l'inférence statistique sur θ : l'estimer, ponctuellement ou par régions de confiance (intervalles si $d = 1$), et effectuer des tests d'hypothèses portant sur θ . On fait alors de la **statistique paramétrique**.

Un **modèle non paramétrique** est un modèle où \mathcal{P} ne peut pas se mettre sous la forme ci-dessus. Par exemple, \mathcal{P} peut être :

- l'ensemble des lois de probabilité continues sur \mathbb{R} ,
- l'ensemble des lois de probabilité dont le support est $[0, 1]$,
- l'ensemble des lois de probabilité sur \mathbb{R} symétriques par rapport à l'origine,
- etc...

Dans ce cadre, il est possible de déterminer des estimations, des intervalles de confiance, d'effectuer des tests d'hypothèses. Mais les objets sur lesquels portent ces procédures statistiques ne sont plus des paramètres de lois de probabilité. On peut vouloir estimer des quantités réelles comme l'espérance et la variance des observations. On a vu en PMS qu'on pouvait utiliser la moyenne et la variance empirique des données. On peut aussi vouloir estimer des fonctions, comme la fonction de répartition et la densité des observations. On a vu en PMS qu'un histogramme est une estimation de densité.

En termes de tests d'hypothèses, on peut effectuer des tests sur la valeur d'une espérance, tester si les observations sont indépendantes, si elles présentent une croissance, si elles proviennent d'une loi normale, tester si plusieurs échantillons proviennent de la même loi, etc... On fait alors de la **statistique non paramétrique**.

De manière générale, la statistique non paramétrique regroupe l'ensemble des méthodes statistiques qui permettent de tirer de l'information pertinente de données sans faire l'hypothèse que la loi de probabilité de ces observations appartient à une famille paramétrée connue.

Un des problèmes de la statistique paramétrique est le risque d'erreur du à un mauvais choix de modèle. Par exemple, on a vu en PMS dans l'exercice sur les niveaux de bruit à Montréal, que l'on obtient des résultats aberrants si on effectue des calculs en supposant que des observations sont de loi exponentielle, alors qu'en fait elles sont de loi normale. L'avantage de la statistique non paramétrique est de ne pas être soumise à cet aléa. En revanche, si les observations sont bien issues d'un modèle précis, les méthodes statistiques paramétriques qui utilisent ce modèle seront plus performantes que celles qui ne l'utilisent pas. Il est donc également important d'établir des méthodes permettant de déterminer si des observations sont issues ou non de tel ou tel modèle paramétrique, les tests d'adéquation.

2.3 Fonction de vraisemblance

Dans un modèle paramétrique, la fonction de vraisemblance joue un rôle fondamental. Nous n'avons vu en PMS que le cas des modèles d'échantillon, en traitant séparément le cas des lois discrètes et des lois continues.

Pour un modèle d'échantillon discret, l'élément aléatoire observé est $X = (X_1, \dots, X_n)$, où les X_i sont indépendantes et de même loi discrète. Alors la fonction de vraisemblance

est :

$$\mathcal{L}(\theta; x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n; \theta) = \prod_{i=1}^n P(X_i = x_i; \theta).$$

Pour un modèle d'échantillon continu, l'élément aléatoire observé est $X = (X_1, \dots, X_n)$, où les X_i sont indépendantes et de même loi continue. Alors la fonction de vraisemblance est :

$$\mathcal{L}(\theta; x_1, \dots, x_n) = f_{(X_1, \dots, X_n)}(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta).$$

Pour définir une fonction de vraisemblance valable dans n'importe quel modèle statistique, pas forcément d'échantillon et pas forcément discret ou continu, il faut utiliser des notions de théorie de la mesure.

Rappels :

- Une mesure μ sur $(\mathcal{X}, \mathcal{A})$ est σ -finie si et seulement si il existe une suite $\{A_n\}_{n \geq 1}$ d'évènements de \mathcal{A} telle que $\bigcup_{n \geq 1} A_n = \mathcal{X}$ et $\forall n \geq 1, \mu(A_n) < +\infty$ (\mathcal{X} est une union dénombrable d'évènements de mesure finie).
- P_θ est absolument continue par rapport à μ si et seulement si :

$$\forall A \in \mathcal{A}, \mu(A) = 0 \Rightarrow P_\theta(A) = 0.$$

On considère un modèle paramétrique quelconque $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$. On supposera qu'il existe une mesure σ -finie μ sur $(\mathcal{X}, \mathcal{A})$ telle que $\forall \theta \in \Theta$, la loi de P_θ est absolument continue par rapport à μ (on dit que μ est la **mesure dominante** du modèle). Alors le théorème de Radon-Nikodym assure que P_θ admet une densité par rapport à μ . Cette densité est appelée fonction de vraisemblance du modèle.

Définition 2 La fonction de vraisemblance du modèle $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$ est la fonction de θ définie par :

$$\forall A \in \mathcal{A}, P_\theta(A) = P(X \in A; \theta) = \int_A \mathcal{L}(\theta; x) d\mu(x).$$

Plus généralement, pour toute fonction φ intégrable, on a :

$$E[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) \mathcal{L}(\theta; x) d\mu(x).$$

En toute rigueur, \mathcal{L} n'est définie qu'à une μ -équivalence près. Mais dans la pratique, il n'y a pas d'ambiguïté, aussi parle-t-on bien de "la" fonction de vraisemblance.

Cas des modèles continus. Si X est un vecteur aléatoire admettant une densité $f_X(x; \theta)$ (par rapport à la mesure de Lebesgue), on sait bien que $P(X \in A; \theta) = \int_A f_X(x; \theta) dx$. Donc la mesure dominante est la mesure de Lebesgue et la fonction de vraisemblance est $\mathcal{L}(\theta; x) = f_X(x; \theta)$.

Cas des modèles discrets. Si X est un vecteur aléatoire de loi discrète, définie par les probabilités élémentaires $P(X = x; \theta)$, alors :

$$P(X \in A; \theta) = \sum_{x \in A} P(X = x; \theta) = \int_A P(X = x; \theta) d\mu_d(x)$$

où μ_d est la mesure de dénombrement sur \mathcal{X} : $\mu_d(A) = \text{card}(A)$ et $\int_A f(x) d\mu_d(x) = \sum_{x \in A} f(x)$. Donc la fonction de vraisemblance est bien $\mathcal{L}(\theta; x) = P(X = x; \theta)$.

L'avantage de cette définition générale est qu'elle permet de traiter des cas plus atypiques que les modèles d'échantillon discrets ou continus.

Exemple. Une expérience aléatoire conduit à observer la réalisation d'un couple de variables aléatoires $X = (Y, N)$, où Y est une variable aléatoire réelle (continue) et N est une variable aléatoire entière (discrète). Y et N ne sont pas forcément indépendantes. Admettons que leur loi conjointe dépende d'un paramètre θ .

Pour calculer la vraisemblance, qui permettra d'estimer θ , il faut être capable de calculer des grandeurs du type $P((Y, N) \in A_1 \times A_2; \theta) = P([Y \in A_1] \cap [N \in A_2]; \theta)$, où A_1 est un intervalle de \mathbb{R} et A_2 est une partie de \mathbb{N} . On a :

$$\begin{aligned} P([Y \in A_1] \cap [N \in A_2]; \theta) &= \sum_{n \in A_2} P([Y \in A_1] \cap [N = n]; \theta) \\ &= \int_{A_2} P([Y \in A_1] \cap [N = n]; \theta) d\mu_d(n) \\ &= \int_{A_2} P(Y \in A_1 | N = n; \theta) P(N = n; \theta) d\mu_d(n) \\ &= \int_{A_2} \int_{A_1} f_{Y|N=n}(y; \theta) dy P(N = n; \theta) d\mu_d(n) \\ &= \int \int_{A_1 \times A_2} f_{Y|N=n}(y; \theta) P(N = n; \theta) dy d\mu_d(n) \\ &= \int \int_{A_1 \times A_2} f_{Y|N=n}(y; \theta) P(N = n; \theta) d\lambda_L \otimes \mu_d(y; n) \end{aligned}$$

ce qui prouve que la fonction de vraisemblance est :

$$\mathcal{L}(\theta; x) = \mathcal{L}(\theta; y, n) = f_{Y|N=n}(y; \theta) P(N = n; \theta).$$

et que la mesure dominante est la mesure produit $\lambda_L \otimes \mu_d$, où λ_L est la mesure de Lebesgue sur \mathbb{R} et μ_d est la mesure de dénombrement sur \mathbb{N} .

2.4 Statistiques

En PMS, on a défini une statistique comme une fonction des observations, $t(x)$. Dans un modèle paramétrique, cette fonction ne doit pas dépendre du paramètre inconnu θ . Autrement dit, elle doit être mesurable. La définition formelle d'une statistique est la suivante.

Définition 3 Dans un modèle statistique $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, une **statistique** est une application mesurable t de $(\mathcal{X}, \mathcal{A})$ dans un espace \mathcal{Y} muni d'une tribu \mathcal{B} .

Rappel : une application t de $(\mathcal{X}, \mathcal{A})$ dans $(\mathcal{Y}, \mathcal{B})$ est mesurable si et seulement si $\forall B \in \mathcal{B}$, l'évènement $t^{-1}(B) = [t(X) \in B]$ est dans \mathcal{A} , c'est-à-dire $\forall A, t(A) = B \Rightarrow A \in \mathcal{A}$. Concrètement, cela signifie que l'on peut calculer la probabilité de tout évènement de la forme $[t(X) \in B]$, donc t ne doit pas dépendre de paramètres inconnus.

Puisque x est une réalisation de l'élément aléatoire X , $t(x)$ est une réalisation de l'élément aléatoire $T = t(X)$.

Définition 4 La loi de probabilité P_T de T est appelée **loi image** par t et le modèle $(\mathcal{Y}, \mathcal{B}, \{P_T; P \in \mathcal{P}\})$ est le **modèle image** par t de $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

Exemple des ampoules. Le modèle est $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{exp(\lambda); \lambda \in \mathbb{R}^+\})^n$. $X = (X_1, \dots, X_n)$, où les X_i sont des variables aléatoires indépendantes et de même loi $exp(\lambda)$. On sait qu'alors $T = \sum_{i=1}^n X_i$ est de loi gamma $G(n, \lambda)$. Donc la loi image par $t(x) = \sum_{i=1}^n x_i$ est la loi $G(n, \lambda)$ et le modèle image est le modèle $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{G(n, \lambda); \lambda \in \mathbb{R}^+\})$.

Remarquons que le modèle image est de dimension 1 alors que le modèle initial était de dimension n . Autrement dit, la statistique $t(x) = \sum_{i=1}^n x_i$ est un résumé des observations $x = (x_1, \dots, x_n)$. On retrouvera cette notion ultérieurement.

Définition 5 Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$ un modèle statistique paramétrique. Si la fonction de vraisemblance admet un maximum unique au point $\hat{\theta}(x)$, alors l'application $x \mapsto \hat{\theta}(x)$ est appelée **statistique de maximum de vraisemblance**. $\hat{\theta}(X)$ est l'**estimateur de maximum de vraisemblance** de θ au vu de X .

2.5 Exhaustivité

On considère un modèle statistique paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})$. On cherche à obtenir le maximum de connaissance possible sur le paramètre θ à partir de l'observation $x \in \mathcal{X}$. Souvent, x est un vecteur (x_1, \dots, x_n) et n est très grand. Il est alors intéressant de réduire les données en les résumant par une statistique $t(x)$ de dimension très inférieure à n . Il est logique de s'attendre à ce que le résumé $t(x)$ des observations contienne moins d'information sur θ que l'ensemble des données initiales. Or il existe des statistiques qui résument les observations tout en conservant l'intégralité de l'information sur θ , les statistiques exhaustives.

Définition 6 Une statistique t est **exhaustive** pour θ si et seulement si la loi de probabilité conditionnelle de X sachant $[T = t]$ ne dépend pas de θ .

Justification. Si la loi de X sachant $[T = t]$ ne dépend pas de θ , cela signifie que, quand on connaît le résumé de l'observation $t(x)$, la connaissance de la totalité de l'observation x n'apporte aucun renseignement supplémentaire sur θ . Donc la totalité de l'information sur θ est contenue dans $t(x)$. Par conséquent, il faut s'attendre à ne se servir que de $t(x)$ (au lieu de x tout entier) pour estimer θ .

Exemple du contrôle de qualité. Le modèle est $(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\mathcal{B}(p); p \in [0, 1]\})^n$. $x = (x_1, \dots, x_n)$, où

$$x_i = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ pièce est défectueuse} \\ 0 & \text{sinon} \end{cases}$$

Les X_i sont des variables aléatoires indépendantes et de même loi $\mathcal{B}(p)$, où p est la probabilité qu'une pièce soit défectueuse.

Il semble évident que, pour avoir toute l'information sur p , il est inutile de savoir, pour chaque pièce contrôlée, si elle est défectueuse ou pas. Il suffit de connaître le pourcentage (ou le nombre total) de pièces défectueuses. D'ailleurs on a vu en PMS que l'estimateur optimal (ESBVM) de p était bien la proportion de pièces défectueuses $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

On doit donc s'attendre à ce que $\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^n x_i$ soit une statistique exhaustive. Pour des raisons de simplicité d'écriture, on va plutôt montrer que le nombre total de pièces défectueuses $t(x) = \sum_{i=1}^n x_i$ est une statistique exhaustive.

On sait que $T = \sum_{i=1}^n X_i$ est de loi binomiale $\mathcal{B}(n, p)$. Alors :

$$\begin{aligned} P(X = x | T = t) &= P(X_1 = x_1, \dots, X_n = x_n | \sum_{i=1}^n X_i = t) \\ &= \frac{P\left(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = t\right)}{P\left(\sum_{i=1}^n X_i = t\right)} \\ &= \begin{cases} 0 & \text{si } \sum_{i=1}^n x_i \neq t \\ \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P\left(\sum_{i=1}^n X_i = t\right)} & \text{si } \sum_{i=1}^n x_i = t \end{cases} \end{aligned}$$

$$P(X_i = x_i) = \begin{cases} p & \text{si } x_i = 1 \\ 1 - p & \text{si } x_i = 0 \end{cases} = p^{x_i} (1 - p)^{1 - x_i}$$

et comme les X_i sont indépendants, on a :

$$\frac{P(X_1 = x_1, \dots, X_n = x_n)}{P\left(\sum_{i=1}^n X_i = t\right)} = \frac{\prod_{i=1}^n P(X_i = x_i)}{P(T = t)} = \frac{\prod_{i=1}^n p^{x_i} (1 - p)^{1 - x_i}}{C_n^t p^t (1 - p)^{n - t}}$$

$$= \frac{p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}}{C_n^t p^t (1-p)^{n-t}} = \frac{1}{C_n^t} \text{ si } \sum_{i=1}^n x_i = t$$

$$\text{Donc } P(X = x|T = t) = \begin{cases} 0 & \text{si } \sum_{i=1}^n x_i \neq t \\ \frac{1}{C_n^t} & \text{si } \sum_{i=1}^n x_i = t \end{cases}.$$

On reconnaît la loi uniforme sur $\left\{ (x_1, \dots, x_n) \in \{0, 1\}^n; \sum_{i=1}^n x_i = t \right\}$.

La loi conditionnelle de X sachant $[T = t]$ ne dépend pas de p , donc $t(x) = \sum_{i=1}^n x_i$ est une statistique exhaustive pour p .

La vérification de la propriété définissant les statistiques exhaustives n'étant pas forcément facile, il est plus pratique d'utiliser le théorème de Fisher-Neyman, qui caractérise très simplement l'exhaustivité.

Théorème 1 . Théorème de factorisation de Fisher-Neyman. *Pour qu'une statistique t soit exhaustive pour θ , il faut et il suffit qu'il existe deux fonctions mesurables g et h telles que :*

$$\forall x \in \mathcal{X}, \forall \theta \in \Theta, \mathcal{L}(\theta; x) = g(t(x); \theta) h(x).$$

Démonstration. Effectuons la démonstration dans le cas d'un modèle discret. On a donc $\mathcal{L}(\theta; x) = P(X = x; \theta)$.

(\Rightarrow) Si t est exhaustive, $P(X = x|T = t)$ ne dépend pas de θ . Par conséquent :

$$\begin{aligned} \mathcal{L}(\theta; x) &= P(X = x; \theta) = P(X = x \cap t(X) = t(x); \theta) \\ &= P(X = x \cap T = t(x); \theta) = P(X = x|T = t(x)) P(T = t(x); \theta) \\ &= h(x) P(T = t(x); \theta) \end{aligned}$$

qui est bien de la forme $g(t(x); \theta) h(x)$.

(\Leftarrow) On suppose que $\mathcal{L}(\theta; x) = P(X = x; \theta) = g(t(x); \theta) h(x)$. Il faut montrer qu'alors $P(X = x|T = t)$ ne dépend pas de θ . On a :

$$\begin{aligned} P(X = x|T = t_0; \theta) &= \frac{P(X = x \cap T = t_0; \theta)}{P(T = t_0; \theta)} = \frac{P(X = x \cap t(X) = t_0; \theta)}{P(T = t_0; \theta)} \\ &= \begin{cases} 0 & \text{si } t(x) \neq t_0 \\ \frac{P(X = x; \theta)}{P(T = t_0; \theta)} & \text{si } t(x) = t_0 \end{cases} \end{aligned}$$

$$\text{Or } P(T = t_0; \theta) = P(t(X) = t_0; \theta) = \sum_{y: t(y)=t_0} P(X = y; \theta).$$

Donc, pour $t(x) = t_0$, on a :

$$P(X = x|T = t_0; \theta) = \frac{P(X = x; \theta)}{\sum_{y: t(y)=t_0} P(X = y; \theta)} = \frac{g(t(x); \theta) h(x)}{\sum_{y: t(y)=t_0} g(t(y); \theta) h(y)}$$

$$= \frac{g(t_0; \theta) h(x)}{\sum_{y:t(y)=t_0} g(t_0; \theta) h(y)} = \frac{h(x)}{\sum_{y:t(y)=t_0} h(y)}$$

qui ne dépend pas de θ . Donc t est exhaustive, d'où le théorème. ■

Exemple 1 : contrôle de qualité. On a vu que :

$$\mathcal{L}(p; x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}.$$

C'est de la forme $g(\sum_{i=1}^n x_i; p)$, donc on retrouve immédiatement que $\sum_{i=1}^n x_i$ est une statistique exhaustive.

Exemple 2 : échantillon de loi normale $\mathcal{N}(m; \sigma^2)$. On suppose que $X = (X_1, \dots, X_n)$, où les X_i sont indépendantes et de même loi $\mathcal{N}(m; \sigma^2)$. La vraisemblance est :

$$\begin{aligned} \mathcal{L}(m, \sigma^2; x_1, \dots, x_n) &= \prod_{i=1}^n f_{X_i}(x_i; m, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - m)^2}{2\sigma^2}} \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2} \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n x_i^2 - 2m \sum_{i=1}^n x_i + nm^2 \right]} \end{aligned}$$

qui est de la forme $g\left(\left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right); m, \sigma^2\right)$. Donc le couple $\left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$ est une statistique exhaustive pour le paramètre $\theta = (m, \sigma^2)$ d'un échantillon de loi normale.

Propriété 1 Si t est exhaustive et si $t = \varphi \circ s$, alors s est exhaustive.

Démonstration. t est exhaustive donc

$$\mathcal{L}(\theta; x) = g(t(x); \theta) h(x) = g(\varphi[s(x)]; \theta) h(x) = G(s(x); \theta) h(x)$$

donc s est exhaustive. ■

Exemple : échantillon de loi normale. $\left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right) = \varphi(\bar{x}_n, s_n^2)$, donc (\bar{x}_n, s_n^2) est une statistique exhaustive pour (m, σ^2) (c'est la statistique de maximum de vraisemblance).

Remarque : Si t est exhaustive, $\varphi \circ t$ ne l'est pas forcément ! Par exemple, $\varphi(\bar{x}_n, s_n^2) = \bar{x}_n$ n'est pas exhaustive pour (m, σ^2) .

Propriété 2 Si t est une statistique exhaustive et si $\hat{\theta}$ est la statistique de maximum de vraisemblance, alors il existe une fonction φ telle que $\hat{\theta} = \varphi \circ t$.

Démonstration. t est exhaustive donc $\mathcal{L}(\theta; x) = g(t(x); \theta) h(x)$. h n'intervient pas dans la maximisation de cette fonction par rapport à θ , donc la statistique de maximum de vraisemblance ne dépend de x qu'à travers $t(x)$. Par conséquent, il existe une fonction φ telle que $\hat{\theta} = \varphi \circ t$. ■

C'est bien le cas de la loi normale avec $t(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right)$ et $\hat{\theta}(x) = (\bar{x}_n, s_n^2)$.

La statistique de maximum de vraisemblance est fonction d'une statistique exhaustive, mais elle n'est pas forcément exhaustive elle-même.

En fait, on peut caractériser facilement les lois de probabilité pour lesquelles les modèles d'échantillon admettent une statistique exhaustive : celles qui appartiennent à la famille exponentielle.

2.6 La famille exponentielle

Définition 7 Soit X une variable aléatoire réelle, dont la loi de probabilité dépend d'un paramètre $\theta \in \mathbb{R}^d$. On dit que la loi de X **appartient à la famille exponentielle** si et seulement si $P(X = x; \theta)$ (cas discret) ou $f_X(x; \theta)$ (cas continu) est de la forme :

$$e^{-\sum_{j=1}^d a_j(x) \alpha_j(\theta) - b(x) - \beta(\theta)}$$

La plupart des lois usuelles appartiennent à la famille exponentielle :

- Loi de Bernoulli $\mathcal{B}(p)$:

$$\begin{aligned} P(X = x; p) &= \begin{cases} p & \text{si } x = 1 \\ 1 - p & \text{si } x = 0 \end{cases} = p^x (1 - p)^{1-x} = e^{x \ln p + (1-x) \ln(1-p)} \\ &= e^{x[\ln p - \ln(1-p)] + \ln(1-p)} = e^{x \ln \frac{p}{1-p} + \ln(1-p)} \end{aligned}$$

qui est de la forme souhaitée avec $d = 1$, $a(x) = x$, $\alpha(p) = \ln \frac{p}{1-p}$, $b(x) = 0$ et $\beta(p) = \ln(1-p)$.

- Loi exponentielle $\exp(\lambda)$:

$$f_X(x; \lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln \lambda}$$

qui est de la forme souhaitée avec $d = 1$, $a(x) = x$, $\alpha(\lambda) = -\lambda$, $b(x) = 0$ et $\beta(\lambda) = \ln \lambda$.

- Loi normale $\mathcal{N}(m, \sigma^2)$:

$$f_X(x; m, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} = e^{-\frac{x^2}{2\sigma^2} + \frac{mx}{\sigma^2} - \frac{m^2}{2\sigma^2} - \ln \sigma \sqrt{2\pi}}$$

qui est de la forme souhaitée avec $d = 2$, $a_1(x) = x^2$, $\alpha_1(m, \sigma^2) = -\frac{1}{2\sigma^2}$, $a_2(x) = x$, $\alpha_2(m, \sigma^2) = \frac{m}{\sigma^2}$, $b(x) = 0$ et $\beta(m, \sigma^2) = -\frac{m}{2\sigma^2} - \ln \sigma \sqrt{2\pi}$.

Mais par exemple, la loi de Weibull $\mathcal{W}(\eta, \beta)$ n'appartient pas à la famille exponentielle :

$$f_X(x; \eta, \beta) = \beta \frac{x^{\beta-1}}{\eta^\beta} e^{-\left(\frac{x}{\eta}\right)^\beta} = e^{-\frac{x^\beta}{\eta^\beta} + (\beta-1) \ln x - \beta \ln \eta + \ln \beta}$$

Le terme x^β fait que $\frac{x^\beta}{\eta^\beta}$ ne peut pas être mis sous la forme $a(x)\alpha(\eta, \beta)$, donc la loi de Weibull n'appartient pas à la famille exponentielle.

Le lien entre famille exponentielle et exhaustivité est donné par le théorème de Dar-mois :

Théorème 2 . Théorème de Dar-mois. *Dans un modèle d'échantillon $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})^n$, où le support de la loi des observations ne dépend pas de θ , il existe une statistique exhaustive si et seulement si cette loi appartient à la famille exponentielle. Alors $t(x) = \left(\sum_{i=1}^n a_1(x_i), \dots, \sum_{i=1}^n a_d(x_i)\right)$ est une statistique exhaustive.*

Démonstration. On effectue la démonstration pour des lois continues.

(\Leftarrow) Si la loi des observations appartient à la famille exponentielle, la fonction de vraisemblance est :

$$\begin{aligned} \mathcal{L}(\theta; x_1, \dots, x_n) &= \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n e^{\sum_{j=1}^d a_j(x_i)\alpha_j(\theta) + b(x_i) + \beta(\theta)} \\ &= e^{\sum_{i=1}^n \sum_{j=1}^d a_j(x_i)\alpha_j(\theta) + \sum_{i=1}^n b(x_i) + n\beta(\theta)} \\ &= e^{\sum_{j=1}^d \alpha_j(\theta) \sum_{i=1}^n a_j(x_i) + \sum_{i=1}^n b(x_i) + n\beta(\theta)} \end{aligned}$$

Le théorème de Fisher-Neyman permet alors d'en déduire que $t(x) = \left(\sum_{i=1}^n a_1(x_i), \dots, \sum_{i=1}^n a_d(x_i)\right)$ est une statistique exhaustive pour θ .

(\Rightarrow) Montrons la réciproque pour $d = 1$, c'est-à-dire $\theta \in \mathbb{R}$. On suppose qu'il existe une statistique exhaustive t . Alors :

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta) = g(t(x_1, \dots, x_n); \theta) h(x_1, \dots, x_n)$$

Il faut montrer qu'alors forcément $f(x; \theta)$ est de la forme $e^{a(x)\alpha(\theta) + b(x) + \beta(\theta)}$. On a :

$$\ln \mathcal{L}(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \ln f(x_i; \theta) = \ln g(t(x_1, \dots, x_n); \theta) + \ln h(x_1, \dots, x_n)$$

Et comme h ne dépend pas de θ :

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(x_i; \theta) = \frac{\partial}{\partial \theta} \ln g(t(x_1, \dots, x_n); \theta)$$

Pour un i quelconque fixé dans $\{1, \dots, n\}$, on a :

$$\begin{aligned} \frac{\partial^2}{\partial \theta \partial x_i} \ln \mathcal{L}(\theta; x_1, \dots, x_n) &= \frac{\partial^2}{\partial \theta \partial x_i} \ln f(x_i; \theta) = \frac{\partial^2}{\partial \theta \partial x_i} \ln g(t(x_1, \dots, x_n); \theta) \\ &= \frac{\partial}{\partial x_i} t(x_1, \dots, x_n) \frac{\partial^2}{\partial \theta \partial y} \ln g(y; \theta)|_{y=t(x_1, \dots, x_n)} \end{aligned}$$

Pour i et j distincts, on obtient donc :

$$\frac{\frac{\partial^2}{\partial \theta \partial x_i} \ln f(x_i; \theta)}{\frac{\partial^2}{\partial \theta \partial x_j} \ln f(x_j; \theta)} = \frac{\frac{\partial}{\partial x_i} t(x_1, \dots, x_n) \frac{\partial^2}{\partial \theta \partial y} \ln g(y; \theta)|_{y=t(x_1, \dots, x_n)}}{\frac{\partial}{\partial x_j} t(x_1, \dots, x_n) \frac{\partial^2}{\partial \theta \partial y} \ln g(y; \theta)|_{y=t(x_1, \dots, x_n)}} = \frac{\frac{\partial}{\partial x_i} t(x_1, \dots, x_n)}{\frac{\partial}{\partial x_j} t(x_1, \dots, x_n)}$$

qui ne dépend pas de θ . On est donc dans la situation d'une fonction φ telle que $\frac{\varphi(x; \theta)}{\varphi(y; \theta)}$ ne dépend pas de θ . Alors forcément $\varphi(x; \theta)$ est de la forme $\varphi(x; \theta) = u(x)v(\theta)$. Par conséquent, on a $\frac{\partial^2}{\partial \theta \partial x} \ln f(x; \theta) = u(x)v(\theta)$.

D'où $\frac{\partial}{\partial \theta} \ln f(x; \theta) = a(x)v(\theta) + w(\theta)$ et $\ln f(x; \theta) = a(x)\alpha(\theta) + \beta(\theta) + b(x)$.

Finalement, la densité est bien de la forme $f(x; \theta) = e^{a(x)\alpha(\theta) + b(x) + \beta(\theta)}$. ■

Pour finir ce chapitre, appliquons le théorème de Darrois aux lois usuelles.

- *Loi de Bernoulli* $\mathcal{B}(p)$: $a(x) = x$, donc on retrouve le fait que $\sum_{i=1}^n x_i$ est une statistique exhaustive. L'ESBVM de p est une fonction de cette statistique : $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

- *Loi exponentielle* $\exp(\lambda)$: $a(x) = x$, donc $\sum_{i=1}^n x_i$ est une statistique exhaustive.

L'ESBVM de λ est une fonction de cette statistique : $\hat{\lambda}_n = \frac{n-1}{\sum_{i=1}^n X_i}$.

- *Loi normale* $\mathcal{N}(m, \sigma^2)$: $a_1(x) = x^2$ et $a_2(x) = x$, donc on retrouve le fait que $\left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right)$ ou (\bar{x}_n, s_n^2) est une statistique exhaustive.
- *Loi de Weibull* $\mathcal{W}(\eta, \beta)$. Elle n'appartient pas à la famille exponentielle, donc il n'y a pas de statistique exhaustive. Cela peut se voir autrement en écrivant la vraisemblance :

$$\mathcal{L}(\eta, \beta; x_1, \dots, x_n) = \prod_{i=1}^n \beta \frac{x_i^{\beta-1}}{\eta^\beta} e^{-\left(\frac{x_i}{\eta}\right)^\beta} = \frac{\beta^n}{\eta^{n\beta}} \left[\prod_{i=1}^n x_i^{\beta-1} \right] e^{-\frac{1}{\eta^\beta} \sum_{i=1}^n x_i^\beta}$$

Elle ne peut pas être factorisée sous la forme du théorème de Fisher-Neyman $g(t(x_1, \dots, x_n); \eta, \beta) h(x_1, \dots, x_n)$, sauf si on prend $t(x_1, \dots, x_n) = (x_1, \dots, x_n)$. Autrement dit, on ne peut pas résumer l'ensemble des données en conservant la totalité de l'information sur les paramètres.

Remarque : on a relié la notion d'exhaustivité à celle d'information sans définir précisément l'information. Il y a en fait un lien entre l'exhaustivité et l'information de Fisher, comme on le verra plus tard.

Chapitre 3

Estimation paramétrique optimale

3.1 Introduction

On se place dans un modèle statistique paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})$. On cherche à estimer au mieux le paramètre θ à partir de l'observation x à l'aide d'une statistique $t(x)$. L'estimateur $T = t(X)$ doit vérifier certaines propriétés pour être de bonne qualité. Il est sans biais si $E(T) = \theta$. Quand $\theta \in \mathbb{R}$ ($d = 1$), on a vu qu'il fallait que l'erreur quadratique moyenne $EQM(T) = E[(T - \theta)^2]$ soit la plus petite possible. Quand T est sans biais, $EQM(T) = Var(T)$. Donc pour $\theta \in \mathbb{R}$, un estimateur optimal sera un **estimateur sans biais et de variance minimale (ESBVM)**.

En PMS, nous avons vu qu'un estimateur sans biais et efficace (sa variance est égale à la borne de Cramer-Rao) était forcément un ESBVM, mais nous n'avons pas donné de procédure générale permettant de trouver un ESBVM. C'est le but essentiel de ce chapitre. Cela nécessite d'utiliser la notion d'exhaustivité, vue au chapitre précédent, et de complétude, que nous allons introduire.

Les résultats seront d'abord introduits dans le cas simple où θ est de dimension 1 (sections 3.2. à 3.4.), puis nous regarderons le cas où θ est de dimension d quelconque en abordant la notion d'information de Fisher.

3.2 Réduction de la variance

Le théorème suivant permet, à partir d'un estimateur sans biais, de construire un autre estimateur sans biais de variance inférieure, pour peu qu'il existe une statistique exhaustive.

Théorème 3 . Théorème de Rao-Blackwell. *S'il existe une statistique exhaustive T et un estimateur sans biais $\hat{\theta}$ de θ , alors $Z = E[\hat{\theta} | T]$ est un estimateur sans biais de θ , de variance inférieure à celle de $\hat{\theta}$.*

Rappels.

- $E[Y | X]$ est une variable aléatoire fonction de X . $E[Y | X = x]$ en est une réalisation.
- Théorème de l'espérance totale : $E[E[Y | X]] = E(Y)$.
- Pour toute fonction φ mesurable, $E[\varphi(X) | X] = \varphi(X)$.

- Pour toute fonction φ mesurable, $E[\varphi(X)Y | X] = \varphi(X)E[Y | X]$.

Démonstration. Comme T est exhaustive, la loi de X sachant T ne dépend pas de θ , donc celle de $\hat{\theta}$ sachant T non plus. Par conséquent, $E[\hat{\theta} | T = t]$ ne dépend pas de θ , donc $z(x) = E[\hat{\theta} | T = t(x)]$ est bien une statistique. Ce résultat est indispensable puisque, si Z dépendait de θ , on ne pourrait pas l'utiliser pour estimer θ .

D'après le théorème de l'espérance totale, $E(Z) = E[E[\hat{\theta} | T]] = E(\hat{\theta})$. Donc si $\hat{\theta}$ est un estimateur sans biais de θ , Z est aussi un estimateur sans biais de θ . La variance de $\hat{\theta}$ est :

$$\begin{aligned} \text{Var}(\hat{\theta}) &= E[(\hat{\theta} - E(\hat{\theta}))^2] = E[(\hat{\theta} - \theta)^2] \\ &= E[(\hat{\theta} - Z + Z - \theta)^2] \\ &= E[(\hat{\theta} - Z)^2] + E[(Z - \theta)^2] + 2E[(\hat{\theta} - Z)(Z - \theta)]. \end{aligned}$$

Les 3 termes de cette somme vérifient :

1. $E[(\hat{\theta} - Z)^2] \geq 0$.
2. $E[(Z - \theta)^2] = E[(Z - E(Z))^2] = \text{Var}(Z)$.
3. $E[(\hat{\theta} - Z)(Z - \theta)] = E[(\hat{\theta} - Z)Z] - \theta E(\hat{\theta} - Z) = E[(\hat{\theta} - Z)Z]$
car $E(\hat{\theta} - Z) = E(\hat{\theta}) - E(Z) = \theta - \theta = 0$.

Enfin :

$$\begin{aligned} E[(\hat{\theta} - Z)Z] &= E[E[(\hat{\theta} - Z)Z | T]] \text{ d'après le théorème de l'espérance totale} \\ &= E[E[(\hat{\theta} - E[\hat{\theta} | T]) E[\hat{\theta} | T] | T]] \\ &= E[E[\hat{\theta} | T] E[\hat{\theta} - E[\hat{\theta} | T] | T]] \\ &= E[E[\hat{\theta} | T] [E[\hat{\theta} | T] - E[\hat{\theta} | T]]] \\ &= 0. \end{aligned}$$

D'où $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - Z)^2] + \text{Var}(Z)$, ce qui prouve que $\text{Var}(Z) \leq \text{Var}(\hat{\theta})$, d'où le théorème. ■

Exemple des ampoules. Modèle d'échantillon de loi exponentielle. On souhaite estimer la fiabilité d'une ampoule à l'instant x , c'est-à-dire la probabilité qu'elle fonctionne toujours au bout d'une durée x :

$$R(x) = P(X_i > x) = e^{-\lambda x}.$$

On sait que l'estimateur de maximum de vraisemblance de λ est $\hat{\lambda}_n = 1/\bar{X}_n = n/\sum_{i=1}^n X_i$, donc l'estimateur de maximum de vraisemblance de $R(x)$ est :

$$\hat{R}_n(x) = e^{-\hat{\lambda}_n x} = e^{-nx/\sum_{i=1}^n X_i}.$$

On a dit en PMS (mais sans le prouver) que l'ESBVM de λ est $\hat{\lambda}'_n = (n-1)/\sum_{i=1}^n X_i$, donc on peut aussi proposer d'estimer $R(x)$ par $\hat{R}'_n(x) = e^{-(n-1)x/\sum_{i=1}^n X_i}$.

Mais le biais de ces estimateurs est difficile à calculer. En effet, étant donné que $\sum_{i=1}^n X_i$ est de loi $G(n, \lambda)$, on a par exemple :

$$E[\hat{R}_n(x)] = \int_0^{+\infty} e^{-nx/y} \frac{\lambda^n}{(n-1)!} e^{-\lambda y} y^{n-1} dy$$

qu'on ne sait pas calculer.

Une autre solution consiste à estimer la probabilité qu'une ampoule fonctionne toujours à l'instant x par le pourcentage d'ampoules observées qui fonctionnent toujours à l'instant x . C'est ce qu'on appelle la fiabilité empirique :

$$\mathbb{R}_n(x) = 1 - \mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i > x\}}.$$

Les propriétés de cet estimateur sont faciles à établir. En effet, les $Y_i = \mathbb{1}_{\{X_i > x\}}$ sont des variables aléatoires indépendantes et de même loi de Bernoulli $\mathcal{B}(P(Y_i = 1)) = \mathcal{B}(P(X_i > x)) = \mathcal{B}(R(x))$.

La fiabilité empirique n'est autre que la moyenne empirique des Y_i : $\mathbb{R}_n(x) = \bar{Y}_n$. Donc on sait que $\mathbb{R}_n(x)$ est un estimateur sans biais et convergent de $E(Y_i) = R(x)$:

$$E[\mathbb{R}_n(x)] = R(x) \quad \text{et} \quad \text{Var}[\mathbb{R}_n(x)] = \frac{\text{Var}(Y_i)}{n} = \frac{R(x)[1-R(x)]}{n}.$$

On a vu que $t(x) = \sum_{i=1}^n x_i$ était une statistique exhaustive pour λ . Par conséquent, le théorème de Rao-Blackwell permet d'affirmer que $Z = E\left[\mathbb{R}_n(x) \mid \sum_{i=1}^n X_i\right]$ est un estimateur sans biais de $R(x)$, de variance inférieure à celle de $\mathbb{R}_n(x)$.

$$\begin{aligned} \text{Soit } z(x, t) &= E\left[\mathbb{R}_n(x) \mid \sum_{i=1}^n X_i = t\right] \\ &= E\left[\frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{X_j > x\}} \mid \sum_{i=1}^n X_i = t\right] \\ &= \frac{1}{n} \sum_{j=1}^n E\left[\mathbb{1}_{\{X_j > x\}} \mid \sum_{i=1}^n X_i = t\right] \end{aligned}$$

$$= E \left[\mathbb{1}_{\{X_1 > x\}} \mid \sum_{i=1}^n X_i = t \right]$$

car les X_i sont interchangeables, donc toutes les espérances sont égales

$$= P(X_1 > x \mid \sum_{i=1}^n X_i = t).$$

Comme les X_i sont positives, il est impossible que l'on ait à la fois $X_1 > x$ et $\sum_{i=1}^n X_i = t$ quand $t \leq x$. On fera donc le calcul sous l'hypothèse $t > x$ et on rajoutera à la fin l'indicatrice $\mathbb{1}_{\{t > x\}}$. On a :

$$P(X_1 > x \mid \sum_{i=1}^n X_i = t) = \int_x^{+\infty} f_{X_1 \mid \sum_{i=1}^n X_i = t}(u) du$$

avec :

$$f_{X_1 \mid \sum_{i=1}^n X_i = t}(u) = \frac{f_{(X_1, \sum_{i=1}^n X_i)}(u, t)}{f_{\sum_{i=1}^n X_i}(t)} = \frac{f_{(X_1, \sum_{i=2}^n X_i)}(u, t-u)}{f_{\sum_{i=1}^n X_i}(t)}$$

Pour les mêmes raisons que précédemment, le numérateur est nul quand $t \leq u$. Donc dans l'intégrale, la borne sup est en fait t au lieu de $+\infty$.

Pour $u < t$, on a :

$$f_{X_1 \mid \sum_{i=1}^n X_i = t}(u) = \frac{f_{X_1}(u) f_{\sum_{i=2}^n X_i}(t-u)}{f_{\sum_{i=1}^n X_i}(t)}$$

car X_1 et $\sum_{i=2}^n X_i$ sont indépendantes. Comme $\sum_{i=2}^n X_i$ est de loi $G(n-1, \lambda)$, on a :

$$f_{X_1 \mid \sum_{i=1}^n X_i = t}(u) = \frac{\lambda e^{-\lambda u} \frac{\lambda^{n-1}}{(n-2)!} e^{-\lambda(t-u)} (t-u)^{n-2}}{\frac{\lambda^n}{(n-1)!} e^{-\lambda t} t^{n-1}} = (n-1) \frac{(t-u)^{n-2}}{t^{n-1}}$$

D'où :

$$\begin{aligned} P(X_1 > x \mid \sum_{i=1}^n X_i = t) &= \int_x^t (n-1) \frac{(t-u)^{n-2}}{t^{n-1}} du = \frac{1}{t^{n-1}} [-(t-u)^{n-1}]_x^t \\ &= \frac{(t-x)^{n-1}}{t^{n-1}} = \left(1 - \frac{x}{t}\right)^{n-1}, \text{ avec } x < t. \end{aligned}$$

Donc finalement $z(x, t) = \left(1 - \frac{x}{t}\right)^{n-1} \mathbb{1}_{\{t > x\}}$ et l'estimateur recherché est :

$$Z = \left(1 - \frac{x}{\sum_{i=1}^n X_i}\right)^{n-1} \mathbb{1}_{\{\sum_{i=1}^n X_i > x\}}.$$

Autant les estimateurs $\hat{R}_n(x)$, $\hat{R}'_n(x)$ et $\mathbb{R}_n(x)$ semblent naturels, autant celui-ci n'est pas intuitif. Pourtant, c'est le meilleur des 4.

On a vu qu'on pouvait diminuer la variance d'un estimateur sans biais, mais peut-on atteindre la variance minimale? Pour le déterminer, on doit introduire la notion de statistique complète.

3.3 Complétude

Définition 8 Une statistique t est **complète** ou **totale** si et seulement si pour toute fonction mesurable φ , on a :

$E[\varphi(T)] = 0, \forall \theta \in \Theta \Rightarrow \varphi = 0$ presque partout sur le support de la loi de T , c'est-à-dire partout sauf sur un ensemble de mesure nulle.

Exemple 1 : contrôle de qualité. $X = (X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi de Bernoulli $\mathcal{B}(p)$. On sait que $t(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ est une statistique exhaustive pour p . Est-elle complète?

On sait que $T = \sum_{i=1}^n X_i$ est de loi binomiale $\mathcal{B}(n, p)$, donc :

$$E[\varphi(T)] = \sum_{k=0}^n \varphi(k) P(T = k) = \sum_{k=0}^n \varphi(k) C_n^k p^k (1-p)^{n-k}.$$

Il faut montrer que

$$\sum_{k=0}^n \varphi(k) C_n^k p^k (1-p)^{n-k} = 0, \forall p \in [0, 1] \Rightarrow \forall k \in \{0, \dots, n\}, \varphi(k) = 0.$$

En effet, comme le support de T est fini, φ doit être nulle partout sur le support.

$$\text{Or } \sum_{k=0}^n \varphi(k) C_n^k p^k (1-p)^{n-k} = (1-p)^n \sum_{k=0}^n \varphi(k) C_n^k \left(\frac{p}{1-p}\right)^k.$$

Soit $\theta = \frac{p}{1-p}$. On a :

$$\sum_{k=0}^n \varphi(k) C_n^k p^k (1-p)^{n-k} = 0, \forall p \in [0, 1] \Rightarrow \sum_{k=0}^n \varphi(k) C_n^k \theta^k = 0, \forall \theta \in \mathbb{R}^+.$$

C'est un polynôme de degré n en θ qui est identiquement nul, donc tous ses coefficients sont nuls. Par conséquent, $\forall k \in \{0, \dots, n\}, \varphi(k) C_n^k = 0$ et donc $\forall k \in \{0, \dots, n\}, \varphi(k) = 0$, ce qui prouve que $t(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ est une statistique complète.

Exemple 2 : ampoules. $X = (X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi exponentielle $\exp(\lambda)$. On sait que $t(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ est une statistique exhaustive pour λ . Est-elle complète?

On sait que $T = \sum_{i=1}^n X_i$ est de loi gamma $G(n, \lambda)$, donc :

$$E[\varphi(T)] = \int_0^{+\infty} \varphi(y) \frac{\lambda^n}{(n-1)!} e^{-\lambda y} y^{n-1} dy.$$

$$E[\varphi(T)] = 0, \forall \lambda \in \mathbb{R}^+ \Rightarrow \int_0^{+\infty} \varphi(y) y^{n-1} e^{-\lambda y} dy = 0, \forall \lambda \in \mathbb{R}^+.$$

Or cette intégrale est la transformée de Laplace de la fonction $\varphi(y) y^{n-1}$ au point λ . Comme la transformée de Laplace est injective, la seule fonction dont la transformée soit 0 est la fonction nulle.

Donc on a $\forall y \in \mathbb{R}^+, \varphi(y) y^{n-1} = 0$, d'où $\forall y \in \mathbb{R}^{+*}, \varphi(y) = 0$. φ n'est peut-être pas nulle en 0, mais elle est nulle presque partout sur \mathbb{R}^+ , support de la loi $G(n, \lambda)$. Par conséquent, $t(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ est une statistique complète.

3.4 L'estimation sans biais et de variance minimale

Les notions d'exhaustivité et de complétude permettent de trouver un ESBVM de θ à partir d'un estimateur sans biais.

Théorème 4 . Théorème de Lehmann-Scheffé. *Si $\hat{\theta}$ est un estimateur sans biais de θ et t est une statistique exhaustive et complète, alors $Z = E[\hat{\theta} | T]$ est l'unique estimateur sans biais de θ , de variance minimale parmi tous les estimateurs sans biais de θ .*

Démonstration. D'après le théorème de Rao-Blackwell, si un estimateur sans biais $\hat{\theta}$ n'est pas fonction de la statistique exhaustive T , on peut toujours trouver un autre estimateur sans biais de θ , de variance inférieure, qui soit fonction de $T : Z = E[\hat{\theta} | T]$. Donc un ESBVM est forcément fonction de T .

Supposons qu'il existe 2 estimateurs sans biais fonction de T , $\hat{\theta}_1(T)$ et $\hat{\theta}_2(T)$.

$$E[\hat{\theta}_1(T)] = E[\hat{\theta}_2(T)] = \theta \text{ donc } \forall \theta \in \Theta, E[\hat{\theta}_1(T) - \hat{\theta}_2(T)] = E[(\hat{\theta}_1 - \hat{\theta}_2)(T)] = 0.$$

Comme t est complète, on en déduit que $\hat{\theta}_1 - \hat{\theta}_2 = 0$ presque partout, d'où $\hat{\theta}_1 = \hat{\theta}_2$ presque partout. Il n'existe donc qu'un seul estimateur sans biais fonction de T et cet estimateur est de variance minimale. ■

Corollaire 1 . *Pour trouver un estimateur optimal, il suffit de trouver un estimateur sans biais fonction d'une statistique exhaustive et complète.*

Exemple 1 : contrôle de qualité. $\hat{p}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de p , fonction de la statistique exhaustive et complète $\sum_{i=1}^n X_i$, donc c'est l'ESBVM de p .

Cela conforte l'intuition : la meilleure façon d'estimer la probabilité qu'une pièce soit défectueuse, c'est de prendre le pourcentage de pièces défectueuses dans le lot contrôlé.

Exemple 2 : ampoules. L'estimateur de maximum de vraisemblance de λ est $\hat{\lambda}_n = n / \sum_{i=1}^n X_i$.

On a vu qu'il était biaisé et que $\hat{\lambda}'_n = (n-1) / \sum_{i=1}^n X_i$ était sans biais. On a affirmé en PMS que $\hat{\lambda}'_n$ était l'ESBVM de λ , sans pouvoir le justifier. On sait maintenant que c'est parce que $\hat{\lambda}'_n$ est un estimateur sans biais fonction de la statistique exhaustive et complète $\sum_{i=1}^n X_i$.

Propriété 3 *Le théorème de Lehmann-Scheffé reste valable si on remplace θ par $\varphi(\theta)$, où φ est une fonction mesurable quelconque. Autrement dit, l'ESBVM de $\varphi(\theta)$ est un estimateur sans biais de $\varphi(\theta)$ fonction d'une statistique exhaustive et complète.*

Dans l'exemple des ampoules, on a vu que $Z = \left(1 - \frac{x}{\sum_{i=1}^n X_i}\right)^{n-1} \mathbb{1}_{\{\sum_{i=1}^n X_i > x\}}$ est un estimateur sans biais de $R(x) = e^{-\lambda x}$. Comme il est fonction de la statistique exhaustive et complète $\sum_{i=1}^n X_i$, cela signifie que Z est l'ESBVM de $R(x)$. $\mathbb{R}_n(x)$ est aussi un estimateur sans biais de $R(x)$, mais comme il n'est pas fonction de $\sum_{i=1}^n X_i$, ce n'est pas l'ESBVM.

Théorème 5 *Dans un modèle d'échantillon où la loi des observations appartient à la famille exponentielle, si $\alpha(\theta)$ est bijective, alors la statistique exhaustive $\sum_{i=1}^n a(x_i)$ est complète.*

Ce théorème permet de retrouver directement que $\sum_{i=1}^n x_i$ est complète dans les exemples du contrôle de qualité et des ampoules.

3.5 Information de Fisher et efficacité

On a dit qu'une statistique exhaustive contenait autant d'information sur θ que l'observation x toute entière, mais on n'a pas défini ce qu'était l'information sur un paramètre. Il y a en fait plusieurs façons de la définir. On ne parlera ici que de l'information de Fisher, mais on pourrait aussi parler de l'information de Kullback ou de Shannon. Intuitivement, l'information mesure la capacité de l'observation à estimer avec précision le paramètre θ .

En PMS, on a défini la quantité d'information de Fisher dans le cas de modèles paramétriques d'échantillon, pour un paramètre θ de dimension 1 :

$$\begin{aligned} \mathcal{I}_n(\theta) &= \text{Var} \left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, \dots, X_n) \right] \\ &= E \left[\left(\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, \dots, X_n) \right)^2 \right] = -E \left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X_1, \dots, X_n) \right] \end{aligned}$$

L'intérêt principal de la quantité d'information est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur sans biais de θ , grâce à l'inégalité FDCR : pour n'importe quelle statistique T ,

$$\text{Var}(T) \geq \frac{\left[\frac{\partial}{\partial \theta} E(T) \right]^2}{\mathcal{I}_n(\theta)}$$

En particulier, si T est un estimateur sans biais de θ , alors $\text{Var}(T) \geq \frac{1}{\mathcal{I}_n(\theta)}$.

Un estimateur efficace est un estimateur pour lequel l'inégalité FDCR est une égalité. Si un estimateur sans biais est efficace, alors il est forcément de variance minimale et sa variance est égale à la borne de Cramer-Rao $1/\mathcal{I}_n(\theta)$.

Dans cette section, nous allons approfondir cette notion d'information de Fisher, en commençant par la définir pour un paramètre θ de dimension d quelconque.

3.5.1 Score et matrice d'information

On se place dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})$. Le paramètre θ s'écrit donc $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$.

Quand on estime un paramètre de dimension d , les notions usuelles liées à l'estimation s'écrivent sous forme vectorielle. Par exemple :

- Le vecteur aléatoire $T = \begin{pmatrix} T_1 \\ \vdots \\ T_d \end{pmatrix}$ est un estimateur sans biais de θ si $E(T) = \theta$, ce qui s'écrit vectoriellement $\begin{pmatrix} E(T_1) \\ \vdots \\ E(T_d) \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$ où $\forall j \in \{1, \dots, d\}, E(T_j) = \theta_j$.
- L'erreur quadratique moyenne de l'estimateur T est

$$E[||T - \theta||^2] = \sum_{j=1}^d E[(T_j - \theta_j)^2]$$

- Les théorèmes de Rao-Blackwell et Lehmann-Scheffé se généralisent en remplaçant la notion de variance par celle d'erreur quadratique moyenne : on réduit l'EQM en prenant l'espérance conditionnelle à une statistique exhaustive et on a l'EQM minimale si cette statistique est complète.

Pour pouvoir traiter à la fois les modèles discrets et continus, nous allons revenir à la définition générale de la fonction de vraisemblance. Soit μ la mesure de référence. On

rappelle que la vraisemblance $\mathcal{L}(\theta, x)$ vérifie :

$$\forall A \in \mathcal{A}, \forall \theta \in \Theta, P(X \in A; \theta) = \int_A \mathcal{L}(\theta; x) d\mu(x)$$

et pour toute fonction φ intégrable :

$$E[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) \mathcal{L}(\theta; x) d\mu(x).$$

Pour définir les notions qui vont suivre, on a besoin de faire les hypothèses suivantes :

- Le support de P_θ ne dépend pas de θ (ce qui, par exemple, exclut la loi uniforme sur $[0, \theta]$).
- $\forall \theta, \forall x, \mathcal{L}(\theta; x) > 0$.
- $\ln \mathcal{L}(\theta; x)$ est dérivable 2 fois par rapport à chaque composante θ_j de θ .
- On peut dériver 2 fois sous le signe somme par rapport à chaque composante de θ : pour toute fonction mesurable g et tous j et k dans $\{1, \dots, d\}$,

$$\frac{\partial}{\partial \theta_j} \int_A g(x) \mathcal{L}(\theta; x) d\mu(x) = \int_A g(x) \frac{\partial}{\partial \theta_j} \mathcal{L}(\theta; x) d\mu(x)$$

et

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \int_A g(x) \mathcal{L}(\theta; x) d\mu(x) = \int_A g(x) \frac{\partial^2}{\partial \theta_j \partial \theta_k} \mathcal{L}(\theta; x) d\mu(x).$$

Sous ces hypothèses, on peut définir les quantités suivantes.

Définition 9 *Le score est le gradient de la log-vraisemblance :*

$$Z(\theta; X) = \nabla \ln \mathcal{L}(\theta; X) = \begin{pmatrix} Z_1(\theta; X) \\ \vdots \\ Z_d(\theta; X) \end{pmatrix}$$

où $\forall j \in \{1, \dots, d\}, Z_j(\theta; X) = \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; X)$.

Le score est un vecteur aléatoire de dimension d . Quand $\theta \in \mathbb{R}$, c'est simplement la variable aléatoire $Z(\theta; X) = \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X)$. L'estimateur de maximum de vraisemblance $\hat{\theta}$ de θ est la valeur de θ qui annule le score : $Z(\hat{\theta}; X) = 0$.

Définition 10 *La matrice d'information de Fisher $\mathcal{I}(\theta)$ est la matrice de covariance du score, de terme général*

$$\mathcal{I}_{jk}(\theta) = \text{Cov}[Z_j(\theta; X); Z_k(\theta; X)].$$

Quand $\theta \in \mathbb{R}$, on retrouve bien $\mathcal{I}(\theta) = \text{Var}[Z(\theta; X)] = \text{Var}\left[\frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta; X)\right]$.

Propriété 4 *Le score est centré : $E[Z(\theta; X)] = 0$.*

Démonstration. $\forall j \in \{1, \dots, d\}$,

$$\begin{aligned} E[Z_j(\theta; X)] &= E\left[\frac{\partial}{\partial\theta_j} \ln \mathcal{L}(\theta; X)\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_j} \ln \mathcal{L}(\theta; x) \mathcal{L}(\theta; x) d\mu(x) \\ &= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta_j} \mathcal{L}(\theta; x)}{\mathcal{L}(\theta; x)} \mathcal{L}(\theta; x) d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_j} \mathcal{L}(\theta; x) d\mu(x) \\ &= \frac{\partial}{\partial\theta_j} \int_{\mathcal{X}} \mathcal{L}(\theta; x) d\mu(x) \quad \text{d'après les hypothèses effectuées} \\ &= \frac{\partial}{\partial\theta_j} P(X \in \mathcal{X}) = \frac{\partial}{\partial\theta_j} 1 = 0 \end{aligned}$$

■

On en déduit que :

$$\begin{aligned} \mathcal{I}_{jk}(\theta) &= \text{Cov}[Z_j(\theta; X); Z_k(\theta; X)] = E[Z_j(\theta; X)Z_k(\theta; X)] - E[Z_j(\theta; X)]E[Z_k(\theta; X)] \\ &= E[Z_j(\theta; X)Z_k(\theta; X)] = E\left[\frac{\partial}{\partial\theta_j} \ln \mathcal{L}(\theta; X) \frac{\partial}{\partial\theta_k} \ln \mathcal{L}(\theta; X)\right] \end{aligned}$$

Pour $\theta \in \mathbb{R}$, on retrouve que $\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta; X)\right)^2\right]$.

De la même manière, on montre que $\mathcal{I}_{jk}(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ln \mathcal{L}(\theta; X)\right]$.

Propriété 5 *Pour les modèles d'échantillon de taille n , la matrice d'information est notée $\mathcal{I}_n(\theta)$ et vérifie $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$.*

Cette propriété traduit l'idée naturelle que, dans un échantillon, chaque observation porte la même quantité d'information sur θ , et que la quantité d'information est additive. La démonstration de ce résultat est similaire à celle effectuée en PMS p. 43.

3.5.2 Information et exhaustivité

Définition 11 *La quantité d'information d'une statistique t , $\mathcal{I}_t(\theta)$, est la quantité d'information du modèle image par t .*

Si on résume les données x par une statistique $t(x)$, on a dit qu'on s'attendait à perdre de l'information, sauf si la statistique est exhaustive. C'est exactement ce qui se passe et qui se traduit de la façon suivante. On présente le résultat pour $\theta \in \mathbb{R}$ pour simplifier.

Propriété 6 .

Dégradation de l'information : pour toute statistique t , $\mathcal{I}_t(\theta) \leq \mathcal{I}(\theta)$.

Information et exhaustivité : $\mathcal{I}_t(\theta) = \mathcal{I}(\theta) \Leftrightarrow t$ est exhaustive.

3.5.3 Borne de Cramer-Rao et efficacité

L'inégalité FDCR vue plus haut pour $\theta \in \mathbb{R}$ s'exprime en fait pour θ de dimension quelconque.

Théorème 6 . Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR). *On considère un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})$ vérifiant les hypothèses de cette section et tel que la matrice d'information $\mathcal{I}(\theta)$ soit inversible.*

Soit t une statistique à valeurs dans \mathbb{R}^q , Λ_T la matrice de covariance de T et Δ la matrice de terme général $\Delta_{ij} = \frac{\partial}{\partial \theta_j} E(T_i)$, $1 \leq i \leq q$, $1 \leq j \leq d$.

Alors $\forall \theta \in \mathbb{R}^d$, la matrice $\Lambda_T - \Delta \mathcal{I}^{-1}(\theta) {}^t \Delta$ est semi-définie positive.

Rappel : La matrice M est semi-définie positive si et seulement si $\forall x \neq 0, {}^t x M x \geq 0$.

Quand $d = q = 1$, $\Lambda_T = \text{Var}(T)$ et $\Delta = \frac{\partial}{\partial \theta} E(T)$. Alors on obtient :

$$\text{Var}(T) - \frac{\left[\frac{\partial}{\partial \theta} E(T) \right]^2}{\mathcal{I}(\theta)} \geq 0.$$

C'est bien le résultat attendu.

Démonstration. Démontrons le théorème pour $d = q = 1$. On a :

$$\begin{aligned} \text{Cov}[T; Z(\theta; X)] &= E[TZ(\theta; X)] - E[T] E[Z(\theta; X)] \\ &= E[TZ(\theta; X)] \text{ car le score est centré} \\ &= E \left[T \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X) \right] = \int_{\mathcal{X}} t(x) \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x) \mathcal{L}(\theta; x) d\mu(x) \\ &= \int_{\mathcal{X}} t(x) \frac{\partial}{\partial \theta} \mathcal{L}(\theta; x) d\mu(x) = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} t(x) \mathcal{L}(\theta; x) d\mu(x) \\ &= \frac{\partial}{\partial \theta} E(T). \end{aligned}$$

L'inégalité de Cauchy-Schwarz permet d'écrire :

$$\text{Cov}[T; Z(\theta; X)]^2 \leq \text{Var}(T) \text{Var}[Z(\theta; X)]$$

D'où :

$$\text{Var}(T) \geq \frac{\text{Cov}[T; Z(\theta; X)]^2}{\text{Var}[Z(\theta; X)]} = \frac{\left[\frac{\partial}{\partial \theta} E(T) \right]^2}{\mathcal{I}(\theta)}.$$

■

Quand $\theta \in \mathbb{R}^d$, l'inégalité FDCR appliquée aux termes diagonaux de Λ_T permet d'obtenir une borne inférieure pour la variance de chaque composante de T :

Propriété 7 $\forall i \in \{1, \dots, q\}$, on a :

$$\text{Var}(T_i) \geq \sum_{j=1}^d \sum_{k=1}^d \mathcal{I}_{jk}^{-1}(\theta) \frac{\partial E(T_i)}{\partial \theta_j} \frac{\partial E(T_i)}{\partial \theta_k}.$$

En particulier, si T est un estimateur sans biais de θ , on a pour tout i , $E(T_i) = \theta_i$. Donc $\frac{\partial E(T_i)}{\partial \theta_j} = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$, d'où $\text{Var}(T_i) \geq \mathcal{I}_{ii}^{-1}(\theta)$, qui est la borne de Cramer-Rao.

L'estimateur T est efficace si l'inégalité FDCR est une égalité.

Définition 12 Un estimateur sans biais T est **efficace** si et seulement si $\Lambda_T = \mathcal{I}^{-1}(\theta)$. Alors, pour tout i , $\text{Var}(T_i) = \mathcal{I}_{ii}^{-1}(\theta)$.

Le dernier théorème de ce chapitre donne une condition d'existence d'un estimateur efficace dans les modèles d'échantillon, liée à la famille exponentielle.

Théorème 7 Dans un modèle d'échantillon $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})^n$, la borne de Cramer-Rao ne peut être atteinte que si P_θ appartient à la famille exponentielle. La vraisemblance s'écrit :

$$\mathcal{L}(\theta; x_1, \dots, x_n) = e^{\sum_{i=1}^n \sum_{j=1}^d a_j(x_i) \alpha_j(\theta) + \sum_{i=1}^n b(x_i) + n\beta(\theta)}$$

Alors, à une transformation linéaire près, la seule fonction de θ qui peut être estimée efficacement est $h(\theta) = -A^{-1}(\theta) \nabla \beta(\theta)$, où $A(\theta)$ est la matrice de terme générique $A_{ij}(\theta) = \frac{\partial \alpha_i(\theta)}{\partial \theta_j}$.

Quand $\theta \in \mathbb{R}$, on a simplement $h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$. On montre alors en plus que l'estimateur efficace de $h(\theta)$ est $T = \frac{1}{n} \sum_{i=1}^n a(X_i)$ et la variance minimale est $\text{Var}(T) = \frac{h'(\theta)}{n\alpha'(\theta)}$.

Exemple des ampoules : échantillon de la loi $\exp(\lambda)$.

$$f_X(x; \lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln \lambda}.$$

La loi exponentielle appartient à la famille exponentielle avec $d = 1$, $a(x) = x$, $\alpha(\lambda) = -\lambda$, $b(x) = 0$ et $\beta(\lambda) = \ln \lambda$.

Alors $h(\lambda) = -\frac{\beta'(\lambda)}{\alpha'(\lambda)} = -\frac{1/\lambda}{-1} = \frac{1}{\lambda}$. Donc on peut estimer efficacement $1/\lambda$ mais pas

λ . C'est bien ce qu'on avait vu : $\hat{\lambda}'_n = (n-1)/\sum_{i=1}^n X_i$ est l'ESBVM de λ , mais il n'est pas efficace.

L'estimateur efficace de $h(\lambda) = \frac{1}{\lambda}$ est $\frac{1}{n} \sum_{i=1}^n a(X_i) = \bar{X}_n$ et la variance minimale est

$$\text{Var}(\bar{X}_n) = \frac{h'(\lambda)}{n\alpha'(\lambda)} = \frac{-1/\lambda^2}{n(-1)} = \frac{1}{n\lambda^2}.$$

C'est logique car $\frac{1}{\lambda} = E(X)$, $\frac{1}{\lambda^2} = \text{Var}(X)$, $E(\bar{X}_n) = E(X)$ et $\text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n}$.

Chapitre 4

Maximum de vraisemblance et estimation bayésienne

4.1 Introduction

On se place dans ce chapitre dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})$. Le chapitre précédent s'est intéressé à la qualité des estimateurs de θ dans ces modèles : variance minimale et efficacité. Mais au préalable, il faut disposer de méthodes permettant d'obtenir de tels estimateurs. On a vu en PMS la méthode des moments et la méthode du maximum de vraisemblance. Il existe de très nombreuses autres méthodes d'estimation. Nous verrons dans ce chapitre une troisième méthode, de plus en plus populaire, l'estimation bayésienne. Mais d'abord nous allons approfondir les propriétés des estimateurs de maximum de vraisemblance, en nous intéressant à leurs propriétés asymptotiques. Les résultats établis permettront en particulier de construire des intervalles de confiance asymptotiques pour les paramètres du modèle sous-jacent.

4.2 Propriétés asymptotiques de l'estimateur de maximum de vraisemblance

Rappelons que si la fonction de vraisemblance $\mathcal{L}(\theta; x)$ admet un maximum unique au point $\hat{\theta}(x)$, alors l'application $x \mapsto \hat{\theta}(x)$ est appelée statistique de maximum de vraisemblance et $\hat{\theta}(X)$ est l'estimateur de maximum de vraisemblance (EMV) de θ . Dans la suite, on notera plus simplement $\hat{\theta}$ cet estimateur. On a donc :

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta; X).$$

Comme d'habitude, on préférera maximiser le logarithme de la vraisemblance :

$$\hat{\theta} = \arg \max_{\theta} \ln \mathcal{L}(\theta; X).$$

Dans la plupart des cas, on maximisera la log-vraisemblance en annulant sa dérivée par rapport à chaque composante de θ . Mais on a vu (voir le cas de la loi uniforme) que cette méthode ne fonctionnait pas toujours. Nous allons nous placer dans ce chapitre dans le cas où cette méthode va fonctionner. Il faut pour cela faire les mêmes hypothèses

(dérivabilité, intégration,...) que celles qui ont été introduites dans la section 3.5.1 pour définir la matrice d'information. Dans ces conditions, l'EMV $\hat{\theta}$ est solution du système des équations de vraisemblance :

$$\forall j \in \{1, \dots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; X) = 0.$$

Mais comme le score est défini par $Z(\theta; X) = \nabla \ln \mathcal{L}(\theta; X)$, $\hat{\theta}$ est finalement la valeur de θ qui annule le score :

$$Z(\hat{\theta}; X) = 0.$$

Nous allons maintenant énoncer les propriétés asymptotiques de l'EMV, vues en PMS pour $\theta \in \mathbb{R}$, pour un paramètre de dimension d quelconque. Nous nous intéressons ici uniquement aux modèles d'échantillon, mais il existe des résultats analogues pour de nombreux autres modèles. Pour un échantillon de taille n , l'EMV sera noté $\hat{\theta}_n$, le score $Z_n(\theta; X)$ et la matrice d'information $\mathcal{I}_n(\theta)$.

Théorème 8 Dans un modèle paramétrique d'échantillon $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})^n$ vérifiant les hypothèses annoncées, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \mathcal{I}_1^{-1}(\theta))$$

où $\mathcal{I}_1(\theta)$ est la matrice d'information de Fisher pour un échantillon de taille 1 et \mathcal{N}_d est la loi normale dans \mathbb{R}^d .

Interprétation : Comme $E[\hat{\theta}_n]$ tend vers θ , l'EMV est asymptotiquement sans biais. Comme la matrice de covariance de $\hat{\theta}_n$ est asymptotiquement équivalente à la borne de Cramer-Rao $[n\mathcal{I}_1]^{-1}(\theta) = \mathcal{I}_n^{-1}(\theta)$, l'EMV est asymptotiquement efficace. Enfin, l'EMV est asymptotiquement gaussien. De plus, la vitesse de convergence de $\hat{\theta}_n$ vers θ est $1/\sqrt{n}$, ce qui signifie que la variance de chaque composante de $\hat{\theta}_n$ tend vers 0 comme $1/n$. Il s'avère que la plupart des autres estimateurs convergent moins vite. Par ailleurs, $\hat{\theta}_n$ converge également presque sûrement vers θ .

Démonstration : Nous allons montrer le résultat pour un paramètre réel ($d = 1$). Alors la quantité d'information est simplement un réel $\mathcal{I}_n(\theta)$, et comme on est dans un modèle d'échantillon, $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$.

Par commodité d'écriture, on suppose que la loi sous-jacente est continue, de densité f . Alors la vraisemblance s'écrit $\mathcal{L}(\theta; x) = \mathcal{L}(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$ et le score est :

$$Z_n(\theta; X) = \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i; \theta).$$

On a déjà vu que $E[Z_n(\theta; X)] = 0$ et :

$$\mathcal{I}_n(\theta) = \text{Var}[Z_n(\theta; X)] = -E \left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X) \right] = -E \left[\frac{\partial}{\partial \theta} Z_n(\theta; X) \right].$$

En particulier, $\mathcal{I}_1(\theta) = \text{Var} \left[\frac{\partial}{\partial \theta} \ln f(X_1; \theta) \right] = -E \left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta) \right]$.

4.2 Propriétés asymptotiques de l'estimateur de maximum de vraisemblance 39

Les variables aléatoires $\frac{\partial}{\partial \theta} \ln f(X_i; \theta)$ sont indépendantes, de même loi, centrées et de variance $\mathcal{I}_1(\theta)$.

Pour éviter des confusions d'écriture, on va noter dans la suite θ_0 la vraie valeur du paramètre θ .

Le théorème des accroissements finis permet d'écrire qu'il existe un θ'_n dans $\left] \min(\hat{\theta}_n, \theta_0), \max(\hat{\theta}_n, \theta_0) \right[$ tel que :

$$Z_n(\hat{\theta}_n; X) = Z_n(\theta_0; X) + (\hat{\theta}_n - \theta_0) \frac{\partial}{\partial \theta} Z_n(\theta; X) \Big|_{\theta'_n}.$$

Or $Z_n(\hat{\theta}_n; X) = 0$. Multiplions par $1/\sqrt{n}$.

$$\frac{1}{\sqrt{n}} Z_n(\theta_0; X) + \frac{1}{\sqrt{n}} (\hat{\theta}_n - \theta_0) \frac{\partial}{\partial \theta} Z_n(\theta; X) \Big|_{\theta'_n} = 0$$

$$\text{ou } \frac{1}{\sqrt{n}} Z_n(\theta_0; X) + \sqrt{n} (\hat{\theta}_n - \theta_0) \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta'_n} = 0.$$

Or :

$$\frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta'_n} = \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta'_n} - \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta_0} + \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta_0} + \mathcal{I}_1(\theta_0) - \mathcal{I}_1(\theta_0).$$

On pose :

$$\begin{aligned} A_n &= \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta_0} + \mathcal{I}_1(\theta_0) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \Big|_{\theta_0} - E \left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta) \right] \Big|_{\theta_0}. \end{aligned}$$

Comme les X_i sont indépendantes et de même loi, la loi des grands nombres permet d'affirmer que :

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \Big|_{\theta_0} \xrightarrow{PS} E \left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta) \right] \Big|_{\theta_0}$$

donc $A_n \xrightarrow{PS} 0$. On pose :

$$B_n = \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta'_n} - \frac{\partial}{\partial \theta} \frac{1}{n} Z_n(\theta; X) \Big|_{\theta_0}.$$

Puisque $\hat{\theta}_n \xrightarrow{PS} \theta_0$ et $\theta'_n \in \left] \min(\hat{\theta}_n, \theta_0), \max(\hat{\theta}_n, \theta_0) \right[$, on a forcément $\theta'_n \xrightarrow{PS} \theta_0$, donc $B_n \xrightarrow{PS} 0$.

D'où $\frac{1}{\sqrt{n}} Z_n(\theta_0; X) + \sqrt{n} (\hat{\theta}_n - \theta_0) [B_n + A_n - \mathcal{I}_1(\theta_0)] = 0$, avec $A_n \xrightarrow{PS} 0$ et $B_n \xrightarrow{PS} 0$.

De plus, le théorème central-limite appliqué aux $\frac{\partial}{\partial \theta} \ln f(X_i; \theta)$ s'écrit :

$$\frac{\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i; \theta) - 0}{\sqrt{n\mathcal{I}_1(\theta)}} = \frac{Z_n(\theta; X)}{\sqrt{n\mathcal{I}_1(\theta)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Finalement, $\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} Z_n(\theta_0; X)}{\mathcal{I}_1(\theta_0) - B_n - A_n}$ a même limite en loi que $\frac{Z_n(\theta_0; X)}{\sqrt{n\mathcal{I}_1(\theta_0)}}$
 $= \frac{Z_n(\theta_0; X)}{\sqrt{\mathcal{I}_1(\theta_0)}\sqrt{n\mathcal{I}_1(\theta_0)}}$, c'est-à-dire la loi $\mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$, d'où le résultat. ■

Si au lieu d'estimer directement θ , on veut estimer une fonction de θ , on sait que $\varphi(\hat{\theta}_n)$ est l'estimateur de maximum de vraisemblance de $\varphi(\theta)$. Les propriétés de cet estimateur sont données par le théorème suivant. Il porte le nom de méthode delta car ce résultat fournit une méthode pour construire des intervalles de confiance asymptotiques.

Théorème 9 . Méthode delta. Si φ est une fonction de \mathbb{R}^d dans \mathbb{R}^q dérivable par rapport à chaque composante de θ , on a :

$$\sqrt{n} \left[\varphi(\hat{\theta}_n) - \varphi(\theta) \right] \xrightarrow{\mathcal{L}} \mathcal{N}_q \left(0, \Delta(\theta) \mathcal{I}_1^{-1}(\theta)^t \Delta(\theta) \right)$$

où $\Delta(\theta)$ est la matrice de terme général $\Delta_{ij}(\theta) = \frac{\partial}{\partial \theta_j} \varphi_i(\theta)$, $1 \leq i \leq q$, $1 \leq j \leq d$.

Démonstration pour $d = q = 1$. Dans ce cas, $\Delta(\theta) = \varphi'(\theta)$, donc le résultat s'écrit :

$$\sqrt{n} \left[\varphi(\hat{\theta}_n) - \varphi(\theta) \right] \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{\varphi'(\theta)^2}{\mathcal{I}_1(\theta)} \right)$$

On le montre facilement à l'aide du théorème des accroissements finis. Il existe θ'_n dans $\left] \min(\hat{\theta}_n, \theta), \max(\hat{\theta}_n, \theta) \right[$ tel que :

$$\varphi(\hat{\theta}_n) = \varphi(\theta) + (\hat{\theta}_n - \theta) \varphi'(\theta'_n).$$

Donc $\sqrt{n} \left[\varphi(\hat{\theta}_n) - \varphi(\theta) \right] = \sqrt{n}(\hat{\theta}_n - \theta) \varphi'(\theta'_n)$. Comme $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{1}{\mathcal{I}_1(\theta)} \right)$ et $\varphi'(\theta'_n) \rightarrow \varphi'(\theta)$, on a bien le résultat ci-dessus. ■

Exemple des ampoules. X_1, \dots, X_n sont indépendantes et de même loi $\exp(\lambda)$. L'information de Fisher est :

$$\begin{aligned} \mathcal{I}_1(\lambda) &= \text{Var} \left[\frac{\partial}{\partial \lambda} \ln f(X; \lambda) \right] = \text{Var} \left[\frac{\partial}{\partial \lambda} \ln \lambda e^{-\lambda X} \right] \\ &= \text{Var} \left[\frac{\partial}{\partial \lambda} (\ln \lambda - \lambda X) \right] = \text{Var} \left[\frac{1}{\lambda} - X \right] = \text{Var}(X) = \frac{1}{\lambda^2} \end{aligned}$$

L'EMV de λ est $\hat{\lambda}_n = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{i=1}^n X_i}$. Le résultat asymptotique sur l'EMV s'écrit :

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}_1^{-1}(\lambda)) = \mathcal{N}(0, \lambda^2).$$

Donc $Var(\sqrt{n}\hat{\lambda}_n) = nVar(\hat{\lambda}_n)$ tend vers λ^2 , d'où $Var(\hat{\lambda}_n) \approx \lambda^2/n$ quand n tend vers l'infini. Or en PMS, on a vu que $Var(\hat{\lambda}_n) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}$, qui est bien équivalent à λ^2/n .

L'EMV de $R(x) = \varphi(\lambda) = e^{-\lambda x}$ est $\hat{R}_n(x) = e^{-\hat{\lambda}_n x}$. On a vu qu'on ne pouvait pas calculer son biais et sa variance pour n fini. Mais la méthode delta montre que $\hat{R}_n(x)$ est asymptotiquement sans biais et que sa variance asymptotique est :

$$Var_{as}(\hat{R}_n(x)) = \frac{\varphi'(\lambda)^2}{n\mathcal{I}_1(\lambda)} = \frac{x^2 e^{-2\lambda x}}{n/\lambda^2} = \frac{\lambda^2 x^2}{n} e^{-2\lambda x}.$$

4.3 Intervalles de confiance asymptotiques

On a vu en PMS que la meilleure façon de déterminer un intervalle de confiance pour un paramètre réel d'un modèle paramétrique, est de trouver une fonction pivotale, fonction des observations et du paramètre, dont la loi de probabilité ne dépend pas du paramètre. Mais il n'est pas forcément facile de trouver une telle fonction. Nous allons voir dans cette section que les propriétés asymptotiques de l'estimateur de maximum de vraisemblance permettent de déterminer assez facilement des intervalles de confiance asymptotiques pour des fonctions presque quelconques des paramètres.

Si $\theta \in \mathbb{R}$, un intervalle de confiance (exact) de seuil α pour θ est un intervalle aléatoire $[Y, Z]$ qui a une probabilité $1 - \alpha$ de contenir θ . Comme on se place dans le cadre de modèles d'échantillon de taille n , on notera $[Y_n, Z_n]$ l'intervalle de confiance. On a donc $P(\theta \in [Y_n, Z_n]) = 1 - \alpha$.

Définition 13 $[Y_n, Z_n]$ est un intervalle de confiance asymptotique de seuil α pour θ si et seulement si :

$$\lim_{n \rightarrow +\infty} P(\theta \in [Y_n, Z_n]) = 1 - \alpha.$$

Dans la pratique, si on sait calculer un intervalle de confiance exact, on n'a pas besoin de calculer un intervalle de confiance asymptotique. Mais quand on ne sait pas calculer un intervalle de confiance exact, on utilise un intervalle de confiance asymptotique : si n est suffisamment grand, $P(\theta \in [Y_n, Z_n])$ ne devrait pas être trop éloigné de $1 - \alpha$.

4.3.1 Cas d'un paramètre réel

Si $\theta \in \mathbb{R}$, $\mathcal{I}_1(\theta)$ est un réel et le résultat asymptotique sur l'EMV s'écrit : $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$ ou $\sqrt{n\mathcal{I}_1(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Le terme $\sqrt{n\mathcal{I}_1(\theta)}(\hat{\theta}_n - \theta)$ est une fonction pivotale asymptotique : fonction de θ et des observations (par l'intermédiaire de $\hat{\theta}_n$), dont la loi asymptotique ne dépend pas de θ .

D'après les propriétés usuelles de la loi $\mathcal{N}(0, 1)$, on a donc :

$$\begin{aligned} \lim_{n \rightarrow +\infty} P\left(-u_\alpha \leq \sqrt{n\mathcal{I}_1(\theta)}(\hat{\theta}_n - \theta) \leq +u_\alpha\right) &= 1 - \alpha \\ &= \lim_{n \rightarrow +\infty} P\left(\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}} \leq \theta \leq \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}\right). \end{aligned}$$

Donc $\left[\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}, \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}\right]$ est un intervalle de confiance asymptotique de seuil α pour θ . Mais cet intervalle est inutilisable à cause du terme $\mathcal{I}_1(\theta)$ qui est inconnu. L'idée naturelle est de le remplacer par $\mathcal{I}_1(\hat{\theta}_n)$. Pour savoir quel est l'impact de cette transformation, il faut utiliser le résultat suivant.

Théorème 10 .Théorème de Slutsky. *Soit $\{U_n\}_{n \geq 1}$ une suite de variables aléatoires convergeant en loi et $\{V_n\}_{n \geq 1}$ une suite de variables aléatoires convergeant en probabilité vers une constante c . Alors pour toute fonction continue g , la suite $\{g(U_n, V_n)\}_{n \geq 1}$ a même limite en loi que la suite $\{g(U_n, c)\}_{n \geq 1}$.*

Ici, on pose $U_n = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$.

On sait que $\hat{\theta}_n \xrightarrow{PS} \theta$, donc $\sqrt{\mathcal{I}_1(\hat{\theta}_n)} \xrightarrow{PS} \sqrt{\mathcal{I}_1(\theta)}$. Comme la convergence presque sûre entraîne la convergence en probabilité, on a également $\sqrt{\mathcal{I}_1(\hat{\theta}_n)} \xrightarrow{P} \sqrt{\mathcal{I}_1(\theta)}$.

Soit $g(u, v) = uv$, $V_n = \sqrt{\mathcal{I}_1(\hat{\theta}_n)}$ et $c = \sqrt{\mathcal{I}_1(\theta)}$. Le théorème de Slutsky permet d'écrire que $g(U_n, V_n) = \sqrt{n\mathcal{I}_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta)$ a même limite en loi que $g(U_n, c) = \sqrt{n\mathcal{I}_1(\theta)}(\hat{\theta}_n - \theta)$, donc $\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Alors, en appliquant la même démarche que précédemment, on obtient la propriété suivante.

Propriété 8 *Un intervalle de confiance asymptotique de seuil α pour θ est :*

$$\left[\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}\right].$$

Exemple 1 : contrôle de qualité. X_1, \dots, X_n sont indépendantes et de même loi $\mathcal{B}(p)$. On a vu en PMS que $\mathcal{I}_n(p) = n\mathcal{I}_1(p) = \frac{n}{p(1-p)}$. Donc un intervalle de confiance asymptotique de seuil α pour p est :

$$\left[\hat{p}_n - u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right].$$

Ce résultat avait été obtenu en PMS (propriété 9) par une méthode bien différente.

Exemple 2 : ampoules. X_1, \dots, X_n sont indépendantes et de même loi $\exp(\lambda)$. $\mathcal{I}_n(\lambda) = n\mathcal{I}_1(\lambda) = \frac{n}{\lambda^2}$. Donc un intervalle de confiance asymptotique de seuil α pour λ est :

$$\left[\hat{\lambda}_n - u_\alpha \frac{\hat{\lambda}_n}{\sqrt{n}}, \hat{\lambda}_n + u_\alpha \frac{\hat{\lambda}_n}{\sqrt{n}} \right] = \left[\hat{\lambda}_n \left(1 - \frac{u_\alpha}{\sqrt{n}} \right), \hat{\lambda}_n \left(1 + \frac{u_\alpha}{\sqrt{n}} \right) \right].$$

Rappelons que l'intervalle de confiance exact est :

$$\left[\hat{\lambda}_n \frac{z_{2n, 1-\alpha/2}}{2n}, \hat{\lambda}_n \frac{z_{2n, \alpha/2}}{2n} \right].$$

Pour n grand, les deux intervalles de confiance sont équivalents.

Intéressons-nous maintenant à des intervalles de confiance asymptotiques pour une fonction $\varphi(\theta)$ du paramètre θ , où $\theta \in \mathbb{R}$ et φ est continue et dérivable. Le résultat de la méthode delta s'écrit :

$$\sqrt{n} \left[\varphi(\hat{\theta}_n) - \varphi(\theta) \right] \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{\varphi'(\theta)^2}{\mathcal{I}_1(\theta)} \right)$$

ou :

$$\frac{\sqrt{n\mathcal{I}_1(\theta)}}{|\varphi'(\theta)|} \left[\varphi(\hat{\theta}_n) - \varphi(\theta) \right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

On peut encore appliquer le théorème de Slutsky et on obtient le résultat suivant.

Propriété 9 *Un intervalle de confiance asymptotique de seuil α pour $\varphi(\theta)$ est :*

$$\left[\varphi(\hat{\theta}_n) - u_\alpha \frac{|\varphi'(\hat{\theta}_n)|}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}, \varphi(\hat{\theta}_n) + u_\alpha \frac{|\varphi'(\hat{\theta}_n)|}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}} \right].$$

Exemple des ampoules. X_1, \dots, X_n sont indépendantes et de même loi $\exp(\lambda)$. L'estimateur de maximum de vraisemblance de $R(x) = \varphi(\lambda) = e^{-\lambda x}$ est $e^{-\hat{\lambda}_n x}$. On a vu que

$$\frac{\varphi'(\lambda)^2}{n\mathcal{I}_1(\lambda)} = \frac{\lambda^2 x^2}{n} e^{-2\lambda x}.$$

Donc un intervalle de confiance asymptotique de seuil α pour $R(x)$ est :

$$\left[e^{-\hat{\lambda}_n x} - u_\alpha \frac{\hat{\lambda}_n x}{\sqrt{n}} e^{-\hat{\lambda}_n x}, e^{-\hat{\lambda}_n x} + u_\alpha \frac{\hat{\lambda}_n x}{\sqrt{n}} e^{-\hat{\lambda}_n x} \right].$$

4.3.2 Cas d'un paramètre vectoriel

Si $\theta \in \mathbb{R}^d$, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \mathcal{I}_1^{-1}(\theta)).$$

$\mathcal{I}_1(\theta)$ est une matrice symétrique définie positive, donc on peut en prendre la racine carrée et écrire :

$$\sqrt{n}\mathcal{I}_1^{1/2}(\theta)(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, Id).$$

où Id est la matrice identité.

Sous des conditions de régularité (continuité des composantes de $\mathcal{I}_1(\theta)$ par rapport à chaque composante de θ), on peut appliquer une version vectorielle du théorème de Slutsky et on obtient :

$$\sqrt{n}\mathcal{I}_1^{1/2}(\hat{\theta}_n)(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, Id).$$

De même, le résultat de la méthode delta s'écrit :

$$\sqrt{n}[\varphi(\hat{\theta}_n) - \varphi(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}_q(0, \Delta(\theta)\mathcal{I}_1^{-1}(\theta)^t\Delta(\theta))$$

ou :

$$\sqrt{n}[\Delta(\theta)\mathcal{I}_1^{-1}(\theta)^t\Delta(\theta)]^{-1/2}[\varphi(\hat{\theta}_n) - \varphi(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}_q(0, Id).$$

Sous des conditions de régularité, on a alors :

$$\sqrt{n}[\Delta(\hat{\theta}_n)\mathcal{I}_1^{-1}(\hat{\theta}_n)^t\Delta(\hat{\theta}_n)]^{-1/2}[\varphi(\hat{\theta}_n) - \varphi(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}_q(0, Id)$$

ce qui permet de donner des intervalles de confiance asymptotiques pour chaque composante de $\varphi(\theta)$.

4.4 Estimation bayésienne

4.4.1 Principe de la méthode

La philosophie de la méthode d'estimation bayésienne est très différente de celles des méthodes vues jusque là. Dans les méthodes du maximum de vraisemblance ou des moments, le paramètre θ est inconnu mais constant, déterministe. L'estimation est menée en considérant qu'on ignore tout de θ , mis à part son ensemble de définition.

Or parfois, on dispose d'une connaissance partielle sur θ . Cette information, dite a priori, peut provenir d'expériences similaires effectuées auparavant ou d'avis d'experts du phénomène étudié qui peuvent anticiper le résultat de l'expérience. Le principe de l'estimation bayésienne est de considérer que le paramètre θ est en fait la réalisation d'une variable aléatoire, et d'intégrer dans sa loi de probabilité toutes les informations a priori dont on dispose sur lui.

Soit T la variable aléatoire dont θ est une réalisation. La loi de probabilité de T est appelée **loi a priori**. En général, cette loi est supposée continue et admettre une densité $f_T(\theta)$ (qu'on note aussi usuellement $\pi(\theta)$).

Les données observées x vont maintenant être considérées comme étant issues de la loi conditionnelle de X sachant $[T = \theta]$. Cela signifie que la fonction de vraisemblance s'écrit :

$$\mathcal{L}(\theta; x) = \begin{cases} P(X = x|T = \theta) & \text{si le modèle est discret} \\ f_{X|T=\theta}(x) & \text{si le modèle est continu} \end{cases}$$

La loi de X , appelée **loi marginale**, est alors obtenue de la façon suivante :

- Modèle discret : $P(X = x) = \int P(X = x|T = \theta) f_T(\theta) d\theta$
- Modèle continu : $f_X(x) = \int f_{X|T=\theta}(x) f_T(\theta) d\theta$

On peut résumer les deux cas en un seul en disant que la **vraisemblance marginale** ou **vraisemblance prédictive** est :

$$\mathcal{L}(x) = \int \mathcal{L}(\theta; x) f_T(\theta) d\theta.$$

Estimer θ dans ce contexte va consister à enrichir l'a priori sur θ (exprimé par $f_T(\theta)$) à l'aide de l'information apportée par l'observation x . On est alors amenés à s'intéresser à la loi conditionnelle de T sachant $[X = x]$, appelée **loi a posteriori**. Les caractéristiques de cette loi sont déterminées grâce à la formule de Bayes :

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

d'où le nom d'estimation bayésienne.

La loi a posteriori est déterminée par sa densité :

- Modèle discret : $f_{T|X=x}(\theta) = \frac{P(X = x|T = \theta)f_T(\theta)}{P(X = x)} = \frac{P(X = x|T = \theta)f_T(\theta)}{\int P(X = x|T = u) f_T(u) du}$
- Modèle continu : $f_{T|X=x}(\theta) = \frac{f_{X|T=\theta}(x) f_T(\theta)}{f_X(x)} = \frac{f_{X|T=\theta}(x) f_T(\theta)}{\int f_{X|T=u}(x) f_T(u) du}$

On résume les deux cas en un seul en disant que la **densité a posteriori** ou **vraisemblance a posteriori** est :

$$f_{T|X=x}(\theta) = \frac{\mathcal{L}(\theta; x) f_T(\theta)}{\int \mathcal{L}(u; x) f_T(u) du} = \frac{\mathcal{L}(\theta; x) f_T(\theta)}{\mathcal{L}(x)}.$$

La loi a posteriori prend en compte à la fois l'information a priori et celle fournie par les données. On l'utilise donc pour estimer θ . On peut prendre comme estimateur la médiane ou le mode de la loi a posteriori, mais la façon la plus courante de procéder est de prendre l'espérance de la loi a posteriori :

$$\hat{\theta}^B = E[T|X].$$

Pour $\theta \in \mathbb{R}$, l'estimation bayésienne correspondante est donc :

$$E[T|X = x] = \int \theta f_{T|X=x}(\theta) d\theta = \frac{\int \theta \mathcal{L}(\theta; x) f_T(\theta) d\theta}{\int \mathcal{L}(\theta; x) f_T(\theta) d\theta}.$$

Elle nécessite donc le calcul de deux intégrales. C'est évidemment un avantage si on peut les calculer explicitement. Ce sera le cas si les lois a priori et a posteriori appartiennent à la même famille. On dit alors que ce sont des **lois conjuguées**. Mais une loi a priori qui reflète de façon réaliste l'information a priori sur θ n'aura pas forcément ces propriétés. Dans ce cas, les intégrales n'ont pas d'expression explicite. Pour les calculer, on utilise alors des méthodes de simulation de Monte-Carlo (méthodes MCMC). L'estimation bayésienne est alors coûteuse en temps de calcul. Les algorithmes de calcul les plus connus sont les échantillonneurs de Gibbs et d'Hastings-Metropolis.

Les paramètres de la loi a priori sont appelés les **hyperparamètres**. Ils sont déterminés par la connaissance a priori que l'on a sur θ , donc ce sont normalement des quantités connues. Mais on peut pousser la logique bayésienne un cran plus loin en considérant que ces hyperparamètres sont inconnus. On peut alors les estimer en maximisant la vraisemblance marginale. L'estimateur bayésien obtenu avec la loi a priori estimée est alors appelé **estimateur bayésien empirique**.

Puisque l'on connaît la loi a posteriori de T sachant $[X = x]$, on est capables de calculer des intervalles $[y, z]$ tels que $P(y \leq T \leq z | X = x) = 1 - \alpha$. $[y, z]$ fournit un encadrement du paramètre θ avec un certain niveau de confiance. Le concept est proche de celui d'intervalle de confiance, mais il est différent. De tels intervalles sont appelés **intervalles de crédibilité**. On peut les utiliser pour mesurer la précision de l'estimation bayésienne de θ .

Dans la densité a posteriori

$$f_{T|X=x}(\theta) = \frac{\mathcal{L}(\theta; x) f_T(\theta)}{\int \mathcal{L}(u; x) f_T(u) du},$$

on constate que l'on peut multiplier $f_T(\theta)$ par une constante sans changer le résultat. Aussi on peut s'autoriser à prendre pour $f_T(\theta)$ une fonction qui n'est pas forcément une densité de probabilité. On a alors ce qu'on appelle des lois a priori impropres. Bien que surprenante, cette démarche permet d'aboutir à des estimateurs bayésiens simples et cohérents.

4.4.2 Exemple du contrôle de qualité

Les données sont des variables aléatoires X_1, \dots, X_n indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$. X_i vaut 1 si la $i^{\text{ème}}$ pièce est défectueuse et 0 sinon.

On cherche à estimer la proportion p de pièces défectueuses. Il est naturel de s'attendre à ce que cette proportion soit faible si la machine est de bonne qualité. Il est également possible que des experts soient capables de donner un ordre de grandeur de cette proportion. Pour tenir compte de cette information, il faut choisir une loi a priori pour p dont le support est $[0, 1]$, et qui soit concentrée sur les petites valeurs. C'est le cas par exemple de certaines lois beta.

On va donc supposer que la loi a priori pour p est la loi beta de première espèce $\beta_1(a, b)$ dont la densité est :

$$f_P(p) = \frac{1}{\beta(a, b)} p^{a-1} (1-p)^{b-1} \mathbf{1}_{[0,1]}(p)$$

$$\text{où } \beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

L'espérance et la variance de cette loi sont :

$$E(P) = \frac{a}{a+b} \quad \text{et} \quad \text{Var}(P) = \frac{ab}{(a+b)^2(a+b+1)}.$$

La connaissance a priori sur p peut se traduire par une valeur moyenne et une variabilité, qui permettent de donner des valeurs aux hyperparamètres a et b .

La vraisemblance habituelle est maintenant considérée comme la densité (par rapport à la mesure de dénombrement) de X sachant $[P = p]$. Autrement dit :

$$\mathcal{L}(p; x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n | P = p) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

La vraisemblance marginale est :

$$\begin{aligned} \mathcal{L}(x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= \int P(X_1 = x_1, \dots, X_n = x_n | P = p) f_P(p) dp \\ &= \int_0^1 p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \frac{1}{\beta(a, b)} p^{a-1} (1-p)^{b-1} dp \\ &= \frac{1}{\beta(a, b)} \int_0^1 p^{\sum_{i=1}^n x_i + a - 1} (1-p)^{n - \sum_{i=1}^n x_i + b - 1} dp \\ &= \frac{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)}{\beta(a, b)} \int_0^1 \frac{p^{\sum_{i=1}^n x_i + a - 1} (1-p)^{n - \sum_{i=1}^n x_i + b - 1}}{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)} dp \\ &= \frac{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)}{\beta(a, b)} \int_0^1 f_{\beta_1\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)}(p) dp \\ &= \frac{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)}{\beta(a, b)} \end{aligned}$$

La loi a posteriori est déterminée par sa densité :

$$\begin{aligned} f_{P|X_1=x_1, \dots, X_n=x_n}(p) &= \frac{P(X_1 = x_1, \dots, X_n = x_n | P = p) f_P(p)}{P(X_1 = x_1, \dots, X_n = x_n)} \\ &= \frac{\beta(a, b)}{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)} \frac{1}{\beta(a, b)} p^{\sum_{i=1}^n x_i + a - 1} (1-p)^{n - \sum_{i=1}^n x_i + b - 1} \\ &= \frac{1}{\beta\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)} p^{\sum_{i=1}^n x_i + a - 1} (1-p)^{n - \sum_{i=1}^n x_i + b - 1} \quad \text{pour } p \in [0, 1]. \end{aligned}$$

On reconnaît la densité de la loi $\beta_1\left(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b\right)$. L'estimateur bayésien est

l'espérance de cette loi, d'où finalement :

$$\hat{p}^B = \frac{\sum_{i=1}^n X_i + a}{\sum_{i=1}^n X_i + a + n - \sum_{i=1}^n X_i + b} = \frac{\sum_{i=1}^n X_i + a}{n + a + b}.$$

Rappelons que l'estimateur de maximum de vraisemblance usuel est :

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

On constate que les 2 estimateurs sont équivalents quand on a beaucoup de données. Quand on a peu de données, la différence peut être importante et dépend du choix de l'a priori. C'est logique : c'est précisément quand on a peu de données qu'il est intéressant de compenser ce manque par de l'information a priori. A la limite, si on n'a pas du tout de données ($n = 0$), on peut quand même estimer p . En effet, dans ce cas l'estimateur bayésien est $\hat{p}^B = \frac{a}{a+b}$. C'est l'espérance de la loi a priori. C'est logique puisqu'en l'absence de données, la seule base pour faire une estimation est l'information a priori. Plus on a d'observations, plus le poids de l'information a priori diminue. La subtilité de l'estimation bayésienne est dans le dosage entre les poids respectifs des observations et de l'information a priori.

L'ignorance complète sur p consiste à prendre comme loi a priori la loi uniforme sur $[0, 1]$, qui n'est autre que la loi $\beta_1(1, 1)$. Alors l'estimateur bayésien est ;

$$\hat{p}^B = \frac{\sum_{i=1}^n X_i + 1}{n + 2},$$

qui est très proche de l'EMV.

On constate que les lois a priori et a posteriori sont toutes les deux des lois beta. C'est ce qu'on a appelé des lois conjuguées. C'est cette propriété qui permet d'avoir des estimateurs bayésiens explicites.

Chapitre 5

Tests d'hypothèses optimaux

5.1 Introduction

Les principes généraux des tests d'hypothèses ont été introduits dans le cours de PMS. Rappelons les rapidement.

- Un test d'hypothèses a pour but de trancher, au vu d'observations, entre une hypothèse nulle H_0 , et une hypothèse alternative H_1 .
- Le seuil du test α est la probabilité maximale de l'erreur de première espèce, erreur qui consiste à rejeter H_0 à tort (conclure H_1 alors que H_0 est vraie). La valeur de α est fixée par l'utilisateur en fonction de la gravité des conséquences de l'erreur de première espèce.
- La puissance β est liée à la probabilité de rejeter H_0 à raison. Sa définition dépend de la nature du test (test d'hypothèses simples ou d'hypothèses composites).
- La région critique W est l'ensemble des valeurs des observations pour lesquelles on rejettera H_0 .

En PMS, on a déterminé les régions critiques essentiellement à l'aide du bon sens ou de l'intuition, ou en utilisant la dualité entre tests d'hypothèses et intervalles de confiance. Nous allons donner dans ce chapitre un procédé systématique de construction de tests d'hypothèses paramétriques.

Comme on ne peut pas minimiser les deux risques d'erreur en même temps, on a choisi de privilégier l'erreur de première espèce, c'est-à-dire de construire des tests en fixant le seuil α . À α fixé, le meilleur des tests possibles est celui qui minimisera la probabilité de l'erreur de deuxième espèce, ou maximisera la puissance. Nous donnerons dans ce chapitre les moyens de déterminer des tests optimaux.

Mais pour commencer, nous allons proposer une définition plus formelle des tests d'hypothèses, qui va permettre d'élargir le cadre vu en PMS.

5.2 Définitions

On se place dans un modèle statistique $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. L'observation x est la réalisation d'un élément aléatoire X de loi $P \in \mathcal{P}$. Les hypothèses que l'on peut effectuer sur cette

observation portent sur la nature de la loi P .

Donc on peut écrire que les hypothèses d'un test sont de la forme $H_0 : "P \in \mathcal{P}_0"$ et $H_1 : "P \in \mathcal{P}_1"$, où \mathcal{P}_0 et \mathcal{P}_1 sont des parties de \mathcal{P} . Au vu de x , on doit décider s'il est plus raisonnable de considérer que $P \in \mathcal{P}_0$ ou que $P \in \mathcal{P}_1$.

Définition 14 *Un test d'hypothèses de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P \in \mathcal{P}_1"$ est une statistique*

$$\begin{aligned} \psi : \mathcal{X} &\rightarrow [0, 1] \\ x &\mapsto \psi(x) = \text{probabilité de rejeter } H_0 \text{ au profit de } H_1 \text{ quand l'observation est } x. \end{aligned}$$

Définition 15 *Un test d'hypothèses est **déterministe** si et seulement si ψ est une indicatrice : $\psi(x) = \mathbf{1}_W(x)$. Autrement dit, on rejettera H_0 si $x \in W$ et on ne rejettera pas H_0 si $x \notin W$.*

On voit que l'on retrouve ici la notion de **région critique**. Tous les tests vus en PMS sont déterministes. Mais la définition proposée ici est plus large : un test n'est pas forcément une indicatrice, donc on peut imaginer des tests pour lesquels la valeur de l'observation x ne permet pas immédiatement de trancher entre H_0 et H_1 . On va voir qu'il est indispensable de définir un test de cette façon si on veut être capables de traiter l'optimalité des tests.

Une hypothèse est **simple** si elle est réduite à un singleton : " $P = P_0$ ". Une hypothèse est **composite** ou **multiple** quand elle n'est pas simple : " $P \in \mathcal{P}_0$ " où \mathcal{P}_0 n'est pas réduit à un singleton.

5.3 Tests d'hypothèses simples

Un **test d'hypothèses simples** est un test dans lequel H_0 et H_1 sont simples. C'est donc un test de $H_0 : "P = P_0"$ contre $H_1 : "P = P_1"$.

Définition 16 *Le seuil du test est $\alpha = E_{P_0} [\psi(X)]$ et la **puissance** du test est $\beta = E_{P_1} [\psi(X)]$.*

Explication : Le seuil du test est la probabilité de rejeter à tort H_0 , c'est-à-dire la probabilité de décider que la loi de X est P_1 alors qu'en fait la loi de X est P_0 . Or on a défini le test de sorte que $\psi(x)$ soit la probabilité de rejeter H_0 quand l'observation est x . Pour obtenir α , il faut donc considérer $\psi(x)$ pour toutes les valeurs possibles de x quand la loi de X est P_0 . Autrement dit, il faut prendre l'espérance de $\psi(X)$ sous la loi P_0 .

La loi de X étant caractérisée par sa fonction de vraisemblance, on note $\mathcal{L}(P; x)$ la fonction de vraisemblance quand la loi de X est P . Alors on peut réécrire α sous la forme :

$$\alpha = E_{P_0} [\psi(X)] = \int \psi(x) dP_0(x) = \int \psi(x) \mathcal{L}(P_0; x) d\mu(x).$$

La puissance du test est la probabilité de rejeter à raison H_0 , c'est-à-dire la probabilité de décider à juste titre que la loi de X est P_1 . On a donc :

$$\beta = E_{P_1} [\psi(X)] = \int \psi(x) dP_1(x) = \int \psi(x) \mathcal{L}(P_1; x) d\mu(x).$$

Quand le test est déterministe, $\psi(x) = \mathbf{1}_W(x)$, donc :

$$\alpha = E_{P_0} [\psi(X)] = \int \mathbf{1}_W(x) \mathcal{L}(P_0; x) d\mu(x) = \int_W \mathcal{L}(P_0; x) d\mu(x) = P_0(X \in W).$$

De même, $\beta = E_{P_1} [\psi(X)] = P_1(X \in W)$.

On retrouve bien le fait que, pour un test d'hypothèses simples déterministe, le seuil est la probabilité sous H_0 que les observations soient dans la région critique et la puissance est la probabilité sous H_1 que les observations soient dans la région critique.

La probabilité d'erreur de deuxième espèce est $1 - \beta$. Un test ψ_1 est meilleur qu'un test ψ_2 si les deux probabilités d'erreur sont inférieures pour ψ_1 à ce qu'elles sont pour ψ_2 . Donc ψ_1 a un seuil inférieur et une puissance supérieure à ψ_2 :

$$\alpha_{\psi_1} \leq \alpha_{\psi_2} \text{ et } \beta_{\psi_1} \geq \beta_{\psi_2}.$$

D'où la définition suivante :

Définition 17 *Un test ψ de $H_0 : "P = P_0"$ contre $H_1 : "P = P_1"$ est dit le meilleur à son niveau de signification si et seulement si tout test de seuil inférieur est moins puissant. Autrement dit :*

$$\forall \psi', \alpha_{\psi'} \leq \alpha_{\psi} \Rightarrow \beta_{\psi'} \leq \beta_{\psi}.$$

Cela signifie en particulier que, quand la probabilité d'erreur de première espèce est fixée, le meilleur test est celui qui minimise la probabilité d'erreur de deuxième espèce.

Le résultat le plus important de ce chapitre est le lemme de Neyman-Pearson qui permet, d'une part de construire des tests d'hypothèses simples de façon systématique, et d'autre part de déterminer les meilleurs tests d'hypothèses simples.

Théorème 11 . Lemme de Neyman-Pearson. $\forall \alpha \in [0, 1]$, il existe $k_\alpha \in \mathbb{R}^+$ et $\gamma_\alpha \in [0, 1]$ tels que le meilleur test de seuil α de $H_0 : "P = P_0"$ contre $H_1 : "P = P_1"$ est :

$$\psi(x) = \begin{cases} 1 & \text{si } \mathcal{L}(P_1; x) > k_\alpha \mathcal{L}(P_0; x) \\ \gamma_\alpha & \text{si } \mathcal{L}(P_1; x) = k_\alpha \mathcal{L}(P_0; x) \\ 0 & \text{si } \mathcal{L}(P_1; x) < k_\alpha \mathcal{L}(P_0; x) \end{cases}$$

Remarque. Quand $\mathcal{L}(P_0; x) \neq 0$, on voit que le test consiste à comparer $\mathcal{L}(P_1; x)/\mathcal{L}(P_0; x)$ à k_α . Aussi le test ψ est-il appelé **test du rapport de vraisemblances**. Intuitivement,

si ce rapport est grand, alors P_1 est “plus vraisemblable” que P_0 et donc on rejettera H_0 au profit de H_1 . Et inversement si le rapport est petit.

Démonstration. Soit ψ' un test tel que $\alpha_{\psi'} \leq \alpha$. Il faut montrer que ψ' est forcément moins puissant que ψ , c'est-à-dire que $\beta_{\psi'} \leq \beta$.

Posons $A(x) = \psi(x) - \psi'(x)$, $B(x) = \mathcal{L}(P_1; x) - k_\alpha \mathcal{L}(P_0; x)$ et $g(x) = A(x)B(x)$. On a :

- Si $B(x) > 0$, $\psi(x) = 1$, donc $A(x) = 1 - \psi'(x) \geq 0$ d'où $g(x) \geq 0$.
- Si $B(x) = 0$, $g(x) = 0$.
- Si $B(x) < 0$, $\psi(x) = 0$, donc $A(x) = -\psi'(x) \leq 0$ d'où $g(x) \geq 0$.

Par conséquent, $\forall x \in \mathcal{X}$, $g(x) \geq 0$, donc $\int g(x) d\mu(x) \geq 0$. Or :

$$\begin{aligned} \int g(x) d\mu(x) &= \int \psi(x) \mathcal{L}(P_1; x) d\mu(x) - \int \psi'(x) \mathcal{L}(P_1; x) d\mu(x) \\ &\quad - k_\alpha \left[\int \psi(x) \mathcal{L}(P_0; x) d\mu(x) - \int \psi'(x) \mathcal{L}(P_0; x) d\mu(x) \right] \\ &= \beta_\psi - \beta_{\psi'} - k_\alpha [\alpha_\psi - \alpha_{\psi'}] \\ &= \beta - \beta_{\psi'} - k_\alpha [\alpha - \alpha_{\psi'}] \geq 0 \end{aligned}$$

Finalement $\beta - \beta_{\psi'} \geq k_\alpha [\alpha - \alpha_{\psi'}] \geq 0$, donc $\beta \geq \beta_{\psi'}$, ce qui prouve que ψ est bien le meilleur test à son niveau de signification α . ■

Dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$, à chaque loi P_θ correspond un paramètre θ . Donc l'hypothèse “ $P = P_{\theta_0}$ ” peut s'écrire “ $\theta = \theta_0$ ” et la vraisemblance peut s'écrire $\mathcal{L}(P_\theta; x) = \mathcal{L}(\theta; x)$. Les tests d'hypothèses correspondant sont appelés **tests paramétriques**. Dans le cas contraire, on parle de **tests non paramétriques**.

Exemple du contrôle de qualité. Dans le modèle $(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\mathcal{B}(p); p \in [0, 1]\})^n$, on veut tester $H_0 : “P = \mathcal{B}(p_0)^{\otimes n}”$ contre $H_1 : “P = \mathcal{B}(p_1)^{\otimes n}”$. Plus simplement, il s'agit de tester $H_0 : “p = p_0”$ contre $H_1 : “p = p_1”$ dans un modèle d'échantillon de loi de Bernoulli. On reconnaît le problème de test d'hypothèses simples sur une proportion vu en PMS.

On sait que la fonction de vraisemblance est $\mathcal{L}(p; x_1, \dots, x_n) = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$. Par conséquent :

$$\begin{aligned} \mathcal{L}(p_1; x) > k_\alpha \mathcal{L}(p_0; x) &\iff p_1^{\sum_{i=1}^n x_i} (1-p_1)^{n-\sum_{i=1}^n x_i} > k_\alpha p_0^{\sum_{i=1}^n x_i} (1-p_0)^{n-\sum_{i=1}^n x_i} \\ &\iff \left[\frac{p_1(1-p_0)}{p_0(1-p_1)} \right]^{\sum_{i=1}^n x_i} > k_\alpha \left[\frac{1-p_0}{1-p_1} \right]^n \\ &\iff \left[\sum_{i=1}^n x_i \right] \ln \frac{p_1(1-p_0)}{p_0(1-p_1)} > \ln k_\alpha + n \ln \frac{1-p_0}{1-p_1} \end{aligned}$$

On va maintenant isoler la statistique de test, c'est-à-dire ce qui ne dépend que des x_i . Il faut alors prendre en compte le signe de $p_1 - p_0$. On a :

$$p_0 < p_1 \Rightarrow 1 - p_1 < 1 - p_0 \Rightarrow p_0(1 - p_1) < p_1(1 - p_0) \Rightarrow \ln \frac{p_1(1 - p_0)}{p_0(1 - p_1)} > 0.$$

Donc, pour $p_0 < p_1$,

$$\mathcal{L}(p_1; x) > k_\alpha \mathcal{L}(p_0; x) \iff \sum_{i=1}^n x_i > \frac{\ln k_\alpha + n \ln \frac{1 - p_0}{1 - p_1}}{\ln \frac{p_1(1 - p_0)}{p_0(1 - p_1)}} = l_\alpha,$$

ce qui signifie que le meilleur test est de la forme :

$$\psi(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n x_i > l_\alpha \\ \gamma_\alpha & \text{si } \sum_{i=1}^n x_i = l_\alpha \\ 0 & \text{si } \sum_{i=1}^n x_i < l_\alpha \end{cases}$$

Réciproquement, pour $p_0 > p_1$, le meilleur test est de la forme :

$$\psi(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n x_i < l_\alpha \\ \gamma_\alpha & \text{si } \sum_{i=1}^n x_i = l_\alpha \\ 0 & \text{si } \sum_{i=1}^n x_i > l_\alpha \end{cases}$$

Il reste à déterminer les valeurs de l_α et de γ_α , ce qui se fait en explicitant le seuil du test. On pourrait ensuite déterminer k_α en fonction de l_α, p_0, p_1 et n , mais ça n'a aucun intérêt : seuls l_α et γ_α sont importants.

Faisons le calcul dans le cas où $p_0 < p_1$:

$$\begin{aligned} \alpha = E_{P_0} [\psi(X)] &= 1 \times P_0(\psi(X) = 1) + \gamma_\alpha \times P_0(\psi(X) = \gamma_\alpha) + 0 \times P_0(\psi(X) = 0) \\ &= P_0\left(\sum_{i=1}^n X_i > l_\alpha\right) + \gamma_\alpha P_0\left(\sum_{i=1}^n X_i = l_\alpha\right) \end{aligned}$$

Sous H_0 , $\sum_{i=1}^n X_i$ est de loi binomiale $\mathcal{B}(n, p_0)$, donc, pour l_α entier, $P_0\left(\sum_{i=1}^n X_i = l_\alpha\right) = C_n^{l_\alpha} p_0^{l_\alpha} (1 - p_0)^{n - l_\alpha}$ et $P_0\left(\sum_{i=1}^n X_i > l_\alpha\right) = \sum_{k=l_\alpha+1}^n C_n^k p_0^k (1 - p_0)^{n-k}$.

S'il existe l_0 tel que $P_0\left(\sum_{i=1}^n X_i > l_0\right) = \alpha$, on prend $l_\alpha = l_0$ et $\gamma_\alpha = 0$. Sinon, il existe forcément l_0 tel que $P_0\left(\sum_{i=1}^n X_i > l_0\right) < \alpha < P_0\left(\sum_{i=1}^n X_i > l_0 - 1\right)$. Alors on prend $l_\alpha = l_0$

$$\text{et } \gamma_\alpha = \frac{\alpha - P_0\left(\sum_{i=1}^n X_i > l_0\right)}{P_0\left(\sum_{i=1}^n X_i = l_0\right)}.$$

Ayant obtenu l_α et γ_α , on peut calculer la puissance du test :

$$\beta = E_{P_1} [\psi(X)] = P_1 \left(\sum_{i=1}^n X_i > l_\alpha \right) + \gamma_\alpha P_1 \left(\sum_{i=1}^n X_i = l_\alpha \right),$$

où, sous P_1 , $\sum_{i=1}^n X_i$ est de loi binomiale $\mathcal{B}(n, p_1)$.

Si n est assez grand, on peut utiliser le théorème central-limite et l'approximation de la loi binomiale par la loi normale :

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

Comme la loi normale est continue, $\lim_{n \rightarrow +\infty} P_0 \left(\sum_{i=1}^n X_i = l_\alpha \right) = 0$. Donc il suffit de prendre un test déterministe ($\gamma_\alpha = 0$) et

$$\alpha = P_0 \left(\sum_{i=1}^n X_i > l_\alpha \right) = P_0 \left(\frac{\sum_{i=1}^n X_i - np_0}{\sqrt{np_0(1-p_0)}} > \frac{l_\alpha - np_0}{\sqrt{np_0(1-p_0)}} \right),$$

qui tend, quand n tend vers l'infini, vers $1 - \phi \left(\frac{l_\alpha - np_0}{\sqrt{np_0(1-p_0)}} \right)$.

On va donc prendre $l_\alpha = np_0 + \sqrt{np_0(1-p_0)} \phi^{-1}(1-\alpha) = np_0 + \sqrt{np_0(1-p_0)} u_{2\alpha}$, et on obtient que le meilleur test asymptotique de seuil α de $H_0 : "p = p_0"$ contre $H_1 : "p = p_1"$, avec $p_0 < p_1$, est le test déterministe défini par la région critique

$$W = \left\{ \sum_{i=1}^n x_i > np_0 + \sqrt{np_0(1-p_0)} u_{2\alpha} \right\} = \left\{ \frac{\sum_{i=1}^n x_i - np_0}{\sqrt{np_0(1-p_0)}} > u_{2\alpha} \right\}.$$

On retrouve le test vu en PMS pour les hypothèses " $p \leq p_0$ " contre " $p > p_0$ ". Sa puissance est :

$$\beta = P_1 \left(\sum_{i=1}^n X_i > l_\alpha \right) = P_1 \left(\frac{\sum_{i=1}^n X_i - np_1}{\sqrt{np_1(1-p_1)}} > \frac{l_\alpha - np_1}{\sqrt{np_1(1-p_1)}} \right),$$

qui tend, quand n tend vers l'infini, vers $1 - \phi \left(\frac{n(p_0 - p_1) + \sqrt{np_0(1-p_0)} u_{2\alpha}}{\sqrt{np_1(1-p_1)}} \right)$.

On constate que le meilleur test de seuil α pour n fini n'est pas un test déterministe. Donc la définition des tests avec des régions critiques ne suffisait pas pour déterminer des tests optimaux.

5.4 Tests d'hypothèses composites

Un test d'hypothèses est composite quand au moins une des deux hypothèses est composite. C'est donc un test de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P \in \mathcal{P}_1"$ où \mathcal{P}_0 et \mathcal{P}_1 ne sont pas toutes les deux réduites à un singleton.

Les tests paramétriques d'hypothèses composites les plus usuels sont :

- *test bilatéral* : test de $H_0 : "\theta = \theta_0"$ contre $H_1 : "\theta \neq \theta_0"$.
- *tests unilatéraux* : test de $H_0 : "\theta \leq \theta_0"$ contre $H_1 : "\theta > \theta_0"$ et test de $H_0 : "\theta \geq \theta_0"$ contre $H_1 : "\theta < \theta_0"$.

Dans ces deux exemples, H_0 et H_1 sont complémentaires : des 2 hypothèses, l'une est forcément vraie. C'est ce cas qui est important en pratique.

Définition 18 *La fonction puissance d'un test d'hypothèses composites est la fonction*

$$\begin{aligned} \beta : \mathcal{P} &\rightarrow [0, 1] \\ P &\mapsto \beta(P) = \text{probabilité de rejeter } H_0 \text{ quand la vraie loi de } X \text{ est } P \\ &= E_P [\psi(X)] = \int \psi(x) \mathcal{L}(P; x) d\mu(x). \end{aligned}$$

Le **seuil** du test est la probabilité maximale de rejeter H_0 à tort :

$$\alpha = \sup_{P \in \mathcal{P}_0} \beta(P).$$

Pour les tests paramétriques, la puissance peut être considérée comme une fonction du paramètre :

$$\beta(\theta) = \int \psi(x) \mathcal{L}(\theta; x) d\mu(x).$$

Pour le test bilatéral, on a simplement $\alpha = \beta(\theta_0)$.

Un test ψ est meilleur qu'un test ψ' si $\forall P \in \mathcal{P}$, la probabilité de rejeter à tort H_0 est plus forte pour ψ' que pour ψ et la probabilité de rejeter à raison H_0 est plus forte pour ψ que pour ψ' :

$$\forall P \in \mathcal{P}_0, \beta_\psi(P) \leq \beta_{\psi'}(P) \quad \text{et} \quad \forall P \in \mathcal{P}_1, \beta_\psi(P) \geq \beta_{\psi'}(P).$$

Définition 19 *Un test ψ de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P \in \mathcal{P}_1"$ est dit **uniformément le plus puissant (UPP)** si et seulement si tout test de seuil inférieur est moins puissant. Autrement dit :*

$$\forall \psi', \alpha_{\psi'} \leq \alpha_\psi \implies \forall P \in \mathcal{P}_1, \beta_{\psi'}(P) \leq \beta_\psi(P).$$

Dans le cas particulier des tests d'hypothèses simples ($\mathcal{P}_0 = \{P_0\}$ et $\mathcal{P}_1 = \{P_1\}$), le test du rapport de vraisemblances donné par le lemme de Neyman-Pearson est UPP.

Il n'existe pas de théorème analogue au lemme de Neyman-Pearson pour les tests composites. Pour rechercher des tests UPP, on utilise alors les résultats suivants :

Théorème 12 .

1. Un test ψ de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P \in \mathcal{P}_1"$ est UPP si et seulement si il est UPP de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P = P_1"$, $\forall P_1 \in \mathcal{P}_1$.
2. Soit $\mathcal{P}'_0 \subset \mathcal{P}_0$. Soit ψ un test de seuil α de $H_0 : "P \in \mathcal{P}_0"$ contre $H_1 : "P \in \mathcal{P}_1"$. Si ψ considéré comme un test de $"P \in \mathcal{P}'_0"$ contre $"P \in \mathcal{P}_1"$ est UPP et de seuil α , alors ψ est UPP.

Démonstration. 1. est immédiat. Pour 2., soit ψ' un test de $"P \in \mathcal{P}_0"$ contre $"P \in \mathcal{P}_1"$ de seuil $\alpha_{\psi'} \leq \alpha$. Il faut montrer que $\forall P \in \mathcal{P}_1$, $\beta_{\psi'}(P) \leq \beta_{\psi}(P)$.

$$\text{Or } \sup_{P \in \mathcal{P}'_0} \beta_{\psi'}(P) \leq \sup_{P \in \mathcal{P}_0} \beta_{\psi'}(P) = \alpha_{\psi'} \leq \alpha.$$

Donc ψ' , considéré comme un test de $"P \in \mathcal{P}'_0"$ contre $"P \in \mathcal{P}_1"$ est de seuil inférieur à α . Si ψ est UPP pour cette situation, on en déduit que $\forall P \in \mathcal{P}_1$, $\beta_{\psi'}(P) \leq \beta_{\psi}(P)$, ce qui prouve que ψ est aussi UPP pour le problème de test initial. ■

La partie 1 du théorème permet de réduire l'hypothèse alternative à une hypothèse simple. La partie 2 permet de réduire l'hypothèse nulle à une hypothèse simple en prenant $\mathcal{P}'_0 = \{P_0\}$. Pour traiter un problème de test d'hypothèses composites, il faut donc commencer par traiter le problème de test d'hypothèses simples sous-jacent.

5.5 Test du rapport des vraisemblances maximales

On se place dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$ et on souhaite tester $H_0 : "\theta \in \Theta_0"$ contre $H_1 : "\theta \notin \Theta_0"$, où Θ_0 est une partie de Θ .

Définition 20 La statistique du rapport des vraisemblances maximales est :

$$v(x) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta; x)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta; x)}.$$

Il est clair que $v(x) \in [0, 1]$. S'il existe une statistique de maximum de vraisemblance $\hat{\theta}(x)$, le dénominateur est $\sup_{\theta \in \Theta} \mathcal{L}(\theta; x) = \mathcal{L}(\hat{\theta}(x); x)$. Ce dénominateur est la vraisemblance maximale globale alors que le numérateur peut être considéré comme la vraisemblance maximale sous H_0 .

Si $\hat{\theta}(x) \in \Theta_0$, $v(x) = 1$. Comme $\hat{\theta}(x)$ est une bonne estimation de θ , si H_0 est vraie, $v(x)$ ne doit pas être trop loin de 1. Inversement, si $v(x)$ est trop loin de 1, on peut douter du fait que $\theta \in \Theta_0$. D'où l'idée de construire un test qui va rejeter H_0 si $v(x)$ est trop petit.

Définition 21 Le test du rapport des vraisemblances maximales est le test déterministe de la forme :

$$\psi(x) = \mathbf{1}_{\{v(x) < l_\alpha\}}.$$

Autrement dit, sa région critique est de la forme $W = \{v(x) < l_\alpha\}$.

Pour un test d'hypothèses simples de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, on se retrouve dans ce cadre si on admet qu'il n'y a que deux valeurs possibles pour $\theta : \Theta = \{\theta_0, \theta_1\}$. Alors :

$$v(x) = \frac{\mathcal{L}(\theta_0; x)}{\sup(\mathcal{L}(\theta_0; x), \mathcal{L}(\theta_1; x))} = \begin{cases} 1 & \text{si } \mathcal{L}(\theta_0; x) \geq \mathcal{L}(\theta_1; x) \\ \frac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\theta_1; x)} & \text{si } \mathcal{L}(\theta_0; x) < \mathcal{L}(\theta_1; x) \end{cases}$$

On ne rejettera H_0 que dans le second cas, ce qui signifie que :

$$\psi(x) = \mathbb{1} \left\{ \frac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\theta_1; x)} < l_\alpha \right\} = \mathbb{1} \left\{ \mathcal{L}(\theta_1; x) > \frac{1}{l_\alpha} \mathcal{L}(\theta_0; x) \right\} = \mathbb{1} \{ \mathcal{L}(\theta_1; x) > k_\alpha \mathcal{L}(\theta_0; x) \}$$

et on retrouve bien le test du rapport de vraisemblances dans le cas où il est déterministe.

Pour déterminer l_α , il faut connaître la loi de $v(X)$ sous H_0 . Donnons le résultat dans un cas particulier.

Propriété 10 On considère un modèle d'échantillon $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})^n$ et le test bilatéral de $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. On a :

$$v(x) = \frac{\mathcal{L}(\theta_0; x)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta; x)} = \frac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\hat{\theta}_n; x)}.$$

Alors, sous H_0 , on a :

$$-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi_d^2.$$

Donc le test déterministe dont la région critique est

$$W = \{-2 \ln v(x) > z_{d, \alpha}\}$$

est asymptotiquement de seuil α pour tester H_0 contre H_1 .

Démonstration. On considère le cas où $d = 1$ ($\theta \in \mathbb{R}$) et la loi des observations est continue, de densité f . On utilise le développement limité déjà vu pour démontrer les propriétés asymptotiques de l'estimateur de maximum de vraisemblance, mais on le prend cette fois à l'ordre 2 :

$$\ln \mathcal{L}(\theta_0; x) = \ln \mathcal{L}(\hat{\theta}_n; x) + (\theta_0 - \hat{\theta}_n) \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x) \Big|_{\hat{\theta}_n} + \frac{1}{2} (\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; x) \Big|_{\theta'_n},$$

où θ'_n est compris entre θ_0 et $\hat{\theta}_n$.

Par définition de l'EMV, $\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x) \Big|_{\hat{\theta}_n} = 0$. Donc on a :

$$\begin{aligned} -2 \ln v(X) &= -2 \left[\ln \mathcal{L}(\theta_0; X) - \ln \mathcal{L}(\hat{\theta}_n; X) \right] \\ &= -(\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X) \Big|_{\theta'_n} \end{aligned}$$

$$\begin{aligned}
&= -(\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \ln f(X_i; \theta) \Big|_{\theta'_n} \\
&= - \left[\sqrt{n} (\theta_0 - \hat{\theta}_n) \right]^2 \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \Big|_{\theta'_n}
\end{aligned}$$

$\hat{\theta}_n \xrightarrow{PS} \theta_0$ donc $\theta'_n \xrightarrow{PS} \theta_0$. Par la loi des grands nombres :

$$-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \Big|_{\theta'_n} \xrightarrow{PS} -E \left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta) \right] \Big|_{\theta_0} = \mathcal{I}_1(\theta_0).$$

Par ailleurs, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$,

donc $\sqrt{\mathcal{I}_1(\theta_0)}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$

et $\mathcal{I}_1(\theta_0) \left[\sqrt{n}(\hat{\theta}_n - \theta_0) \right]^2 \xrightarrow{\mathcal{L}} \chi_1^2$,

ce qui prouve que $-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi_1^2$.

Revenons à d quelconque : $-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi_d^2$. Le test du rapport des vraisemblances maximales est de la forme :

$$\psi(x) = \mathbb{1}\{v(x) < l_\alpha\} = \mathbb{1}\{-2 \ln v(x) > -2 \ln l_\alpha\}.$$

Le seuil du test est $\alpha = P_{H_0}(-2 \ln v(X) > -2 \ln l_\alpha)$. Or :

$$\lim_{n \rightarrow +\infty} P_{H_0}(-2 \ln v(X) > -2 \ln l_\alpha) = 1 - F_{\chi_d^2}(-2 \ln l_\alpha).$$

Donc on peut prendre $-2 \ln l_\alpha = F_{\chi_d^2}^{-1}(1 - \alpha) = z_{d, \alpha}$ et le test s'écrit :

$$\psi(x) = \mathbb{1}\{-2 \ln v(x) > z_{d, \alpha}\}.$$

■

Ce résultat est aussi valable pour d'autres modèles que les modèles d'échantillon (par exemple pour des cas où les X_i sont indépendantes mais pas de même loi), mais malheureusement pas dans tous les cas.

Chapitre 6

Estimation non paramétrique de quantités réelles

Comme on l'a dit dans l'introduction, la statistique non paramétrique regroupe l'ensemble des méthodes statistiques qui permettent de tirer de l'information pertinente de données sans faire l'hypothèse que la loi de probabilité de ces observations appartient à une famille paramétrée connue.

On se place dans le cadre d'un modèle d'échantillon : l'observation x est un vecteur (x_1, \dots, x_n) , constitué de réalisations de variables aléatoires réelles X_1, \dots, X_n indépendantes et de même loi, de fonction de répartition F . On notera f leur densité, si elle existe.

En statistique paramétrique, la loi des X_i dépend d'un paramètre θ . Les problèmes statistiques que l'on traite consistent essentiellement à estimer θ (par exemple par la méthode du maximum de vraisemblance) et à effectuer des tests d'hypothèses sur ce paramètre. L'estimation du paramètre permet alors d'estimer toutes les caractéristiques intéressantes de la loi de probabilité sous-jacente. En particulier, on peut estimer l'espérance $E(X)$ et la variance $Var(X)$ de cette loi.

Mais il n'est pas nécessaire d'avoir un cadre paramétrique pour estimer ces quantités. Le but de ce chapitre est d'étudier des méthodes d'estimation non paramétrique de quantités réelles, comme les moments et les quantiles de l'échantillon. Pour cela, il faut d'abord introduire les outils de base de la statistique non paramétrique : statistiques d'ordre et de rang, loi de probabilité empirique.

Remarque. En toute rigueur, on devrait parler des moments *de la loi de probabilité* d'un échantillon. Pour simplifier, on parle de moments d'un échantillon.

6.1 Les outils de la statistique non paramétrique

6.1.1 Statistiques d'ordre et de rang

Rappelons que si x_1, \dots, x_n sont n réels, on note $x_1^* \leq x_2^* \leq \dots \leq x_n^*$ ces n réels rangés dans l'ordre croissant.

Définition 22 . La statistique d'ordre associée à l'échantillon X_1, \dots, X_n est le vecteur $X^* = (X_1^*, \dots, X_n^*)$. X_i^* est appelée la $i^{\text{ème}}$ statistique d'ordre.

Remarques :

- On note parfois $X_{(i)}$ ou $X_{(i:n)}$ au lieu de X_i^* .
- X^* est à valeurs dans $\tilde{\mathbb{R}}^n = \{(y_1, \dots, y_n) \in \mathbb{R}^n; y_1 \leq y_2 \leq \dots \leq y_n\}$
- $X_1^* = \text{Min}(X_1, \dots, X_n)$, $X_n^* = \text{Max}(X_1, \dots, X_n)$.

La statistique d'ordre contient toute l'information de l'échantillon de départ, sauf l'ordre dans lequel les observations ont été obtenues. Cet ordre est indiqué par les rangs r_i des observations.

exemple 1 (sans ex-aequos) : $n = 5$

x_i	2.3	-3.5	1.7	0.5	-1.4
x_i^*	-3.5	-1.4	0.5	1.7	2.3
r_i	5	1	4	3	2

exemple 2 (avec ex-aequos) : $n = 5$

x_i	0.5	-3.5	1.7	0.5	-1.4
x_i^*	-3.5	-1.4	0.5	0.5	1.7
r_i	3	1	5	3	2

Définition 23 . La **statistique de rang** associée à l'échantillon (X_1, \dots, X_n) est le vecteur $R = (R_1, \dots, R_n)$ où $\forall i \in \{1, \dots, n\}$,

$$\begin{aligned} R_i &= 1 + \sum_{j=1}^n \mathbf{1}_{\{X_j < X_i\}} \\ &= 1 + \text{nombre d'observations strictement inférieures à } X_i \\ &= \text{rang de } X_i \text{ dans l'échantillon ordonné} \end{aligned}$$

Le rang R_i de la $i^{\text{ème}}$ observation X_i est aussi appelé la $i^{\text{ème}}$ **statistique de rang**.

Remarque : on ne définit pas R_i comme le nombre d'observations inférieures ou égales à X_i , pour pouvoir traiter le cas des ex-aequos.

Propriété 11 . Si on connaît les statistiques d'ordre et de rang, on peut reconstruire l'échantillon initial car $X_i = X_{R_i}^*$.

On constate que s'il n'y a pas d'ex-aequos dans l'échantillon, les rangs seront les entiers de 1 à n dans un ordre quelconque. On est sûrs de ne pas avoir d'ex-aequos si et seulement si $\forall (i, j) \in \{1, \dots, n\}^2, i \neq j \Rightarrow P(X_i = X_j) = 0$. En théorie, c'est bien ce qui se passe si la loi des X_i est continue. Mais en pratique, même si cette loi est continue, il est possible qu'il y ait des ex-aequos, du fait de la limitation de la précision des mesures et des erreurs d'arrondis. Il faudra donc être très attentifs à la présence d'ex-aequos dans les données. Sur le plan théorique, nous éviterons cette difficulté en nous limitant aux lois continues.

Théorème 13 . Soit X_1, \dots, X_n un échantillon d'une loi continue. Alors :

1. La loi de R est la loi uniforme sur l'ensemble Σ_n des permutations des entiers de 1 à n .
2. Les statistiques d'ordre et de rang sont indépendantes.

Démonstration.

1. La loi est continue donc il n'y a pas d'ex-aequo. Les R_i prennent toutes les valeurs entières de 1 à n , donc R est bien à valeurs dans Σ_n . Puisque les X_i sont indépendantes et de même loi, elles sont interchangeables et les permutations sont équiprobables, d'où le résultat.

$$\forall r = (r_1, \dots, r_n) \in \Sigma_n, P(R = r) = P(R_1 = r_1, \dots, R_n = r_n) = \frac{1}{\text{card } \Sigma_n} = \frac{1}{n!}.$$

Par exemple, pour $n = 3$, on a :

$$\begin{aligned} P(X_1 < X_2 < X_3) &= P(X_1 < X_3 < X_2) = P(X_2 < X_1 < X_3) = P(X_2 < X_3 < X_1) \\ &= P(X_3 < X_1 < X_2) = P(X_3 < X_2 < X_1) = \frac{1}{6}. \end{aligned}$$

2. Il faut montrer que pour tout borélien B de $\widetilde{\mathbb{R}}_n$ et toute permutation r de Σ_n , on a :

$$P(X^* \in B \cap R = r) = P(X^* \in B)P(R = r).$$

Commençons par un exemple simple :

$$P((X_1^*, X_2^*) \in [2, 4] \times [7, 8] \cap R = (2, 1)) = P(X_2 \in [2, 4] \cap X_1 \in [7, 8]).$$

Or l'interchangeabilité des X_i fait que :

$$\begin{aligned} P(X_2 \in [2, 4] \cap X_1 \in [7, 8]) &= P(X_1 \in [2, 4] \cap X_2 \in [7, 8]) \\ &= P((X_1, X_2) \in [2, 4] \times [7, 8]). \end{aligned}$$

Plus généralement, pour tous B et r , on obtient :

$$P(X^* \in B \cap R = r) = P(X \in B).$$

D'autre part, le théorème des probabilités totales permet d'écrire :

$$P(X^* \in B) = \sum_{r \in \Sigma_n} P(X^* \in B \cap R = r) = \sum_{r \in \Sigma_n} P(X \in B) = n! P(X \in B).$$

D'où $\forall B \in \mathcal{B}(\widetilde{\mathbb{R}}_n), \forall r \in \Sigma_n,$

$$P(X \in B) = \frac{1}{n!} P(X^* \in B) = P(R = r)P(X^* \in B) = P(X^* \in B \cap R = r),$$

ce qui prouve que X^* et R sont indépendantes. ■

La principale conséquence de ce théorème est que la loi de R ne dépend pas de la loi des X_i . On en déduit que **toute variable aléatoire qui ne s'exprime qu'à l'aide des**

rangs des observations a une loi de probabilité indépendante de la loi de ces observations. C'est bien ce qu'on cherche à obtenir en statistique non paramétrique, où la loi des observations n'appartient pas à une famille paramétrée connue. On pourra donc faire de l'estimation et des tests non paramétriques à partir des rangs des observations.

Remarques.

- Il n'y a pas d'équivalent de ce théorème pour les lois non continues, ce qui limite beaucoup l'intérêt de la statistique non paramétrique basée sur les rangs dans ce cas.
- Toute fonction symétrique des observations initiales est une fonction des statistiques d'ordre. Par exemple, $\sum_{i=1}^n X_i = \sum_{i=1}^n X_i^*$.

Propriété 12 . Si la loi des X_i est continue, X^* admet pour densité :

$$f_{(X_1^*, \dots, X_n^*)}(x_1^*, \dots, x_n^*) = n! \prod_{i=1}^n f(x_i^*) \mathbb{1}_{\widetilde{\mathbb{R}}^n}(x_1^*, \dots, x_n^*)$$

Démonstration. Etant donné que pour tout borélien B de $\widetilde{\mathbb{R}}_n$, on a $P(X^* \in B) = n! P(X \in B)$, on obtient pour tout B :

$$\begin{aligned} \int_B f_{(X_1^*, \dots, X_n^*)}(x_1^*, \dots, x_n^*) dx_1^*, \dots, dx_n^* &= n! \int_B f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) dx_1, \dots, dx_n \\ &= \int_B n! \prod_{i=1}^n f_{X_i}(x_i) dx_1, \dots, dx_n \\ &= \int_B n! \prod_{i=1}^n f(x_i^*) dx_1^*, \dots, dx_n^* \end{aligned}$$

d'où le résultat. ■

Propriété 13 . $\forall i \in \{1, \dots, n\}$, la fonction de répartition de la $i^{\text{ème}}$ statistique d'ordre X_i^* est :

$$\forall x \in \mathbb{R}, F_{X_i^*}(x) = \sum_{k=i}^n C_n^k [F(x)]^k [1 - F(x)]^{n-k}$$

Démonstration :

$$\begin{aligned} F_{X_i^*}(x) &= P(X_i^* \leq x) = P(i \text{ au moins des } X_j \text{ sont inférieurs à } x) \\ &= \sum_{k=i}^n P(k \text{ exactement des } X_j \text{ sont inférieurs à } x) \\ &= \sum_{k=i}^n C_n^k P(X_1 \leq x, \dots, X_k \leq x, X_{k+1} > x, \dots, X_n > x) \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=i}^n C_n^k [P(X_i \leq x)]^k [P(X_i > x)]^{n-k} \\
&= \sum_{k=i}^n C_n^k [F(x)]^k [1 - F(x)]^{n-k}
\end{aligned}$$

■

Corollaire 2 . Si la loi des X_i est continue, alors $\forall i \in \{1 \dots n\}$, X_i^* admet pour densité :

$$\forall x \in \mathbb{R}, f_{X_i^*}(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x).$$

Démonstration. Une première solution est de dériver directement l'expression de $F_{X_i^*}(x)$ donnée par la propriété 13.

Une autre façon de faire, qui permet de mieux comprendre le sens des statistiques d'ordre, est la suivante :

$$\begin{aligned}
f_{X_i^*}(x) &= F'_{X_i^*}(x) = \lim_{dx \rightarrow 0} \frac{1}{dx} (F_{X_i^*}(x + dx) - F_{X_i^*}(x)) = \lim_{dx \rightarrow 0} \frac{1}{dx} P(x < X_i^* \leq x + dx) \\
&= \lim_{dx \rightarrow 0} \frac{1}{dx} P((i-1) \text{ des } X_j \text{ sont } \leq x, \text{ un des } X_j \text{ est compris entre } x \text{ et } x + dx, \\
&\quad (n-i) \text{ des } X_j \text{ sont } > x + dx) \\
&= \lim_{dx \rightarrow 0} \frac{1}{dx} C_n^{i-1} [P(X_j \leq x)]^{i-1} C_{n-i+1}^1 P(x < X_j \leq x + dx) [P(X_j > x + dx)]^{n-i} \\
&= \frac{n!}{(i-1)!(n-i+1)!} (n-i+1) [F(x)]^{i-1} [1 - F(x)]^{n-i} \lim_{dx \rightarrow 0} \frac{1}{dx} P(x < X_j \leq x + dx) \\
&= \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x)
\end{aligned}$$

■

Les lois de probabilité du minimum et du maximum d'un échantillon peuvent s'obtenir comme cas particuliers des résultats précédents :

- $X_1^* = \text{Min}(X_1, \dots, X_n)$: $F_{X_1^*}(x) = 1 - [1 - F(x)]^n$
 $f_{X_1^*}(x) = n f(x) [1 - F(x)]^{n-1}$
- $X_n^* = \text{Max}(X_1, \dots, X_n)$: $F_{X_n^*}(x) = [F(x)]^n$
 $f_{X_n^*}(x) = n f(x) [F(x)]^{n-1}$

Plus généralement, on peut déterminer la loi de probabilité de n'importe quel sous-ensemble de la statistique d'ordre. Dans le cas où la loi des X_i est continue, on obtient :

Propriété 14 . Pour tous r_1, \dots, r_k entiers tels que $1 \leq r_1 < r_2 < \dots < r_k \leq n$, on a :

$$f_{(X_{r_1}^*, \dots, X_{r_k}^*)}(x_1, \dots, x_k) = \frac{n!}{(r_1 - 1)! \prod_{i=2}^k (r_i - r_{i-1} - 1)! (n - r_k)!} [F(x_1)]^{r_1 - 1} \prod_{i=1}^k f(x_i) \\ \left[\prod_{i=2}^k [F(x_i) - F(x_{i-1})]^{r_i - r_{i-1} - 1} \right] [1 - F(x_k)]^{n - r_k} \mathbb{1}_{\mathbb{R}^k}(x_1, \dots, x_k)$$

6.1.2 Loi de probabilité empirique

La loi de probabilité empirique est une loi de probabilité créée directement à partir de l'échantillon observé x_1, \dots, x_n .

Définition 24 . La loi de probabilité empirique \mathbb{P}_n associée à l'échantillon x_1, \dots, x_n est la loi uniforme (discrète) sur $\{x_1, \dots, x_n\}$. Si X_e est une variable aléatoire de loi \mathbb{P}_n , alors :

- X_e est à valeurs dans $\{x_1, \dots, x_n\}$.
- $\forall i \in \{1, \dots, n\}, P(X_e = x_i) = \mathbb{P}_n(x_i) = \frac{1}{n}$.

On peut aussi écrire $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$.

Les caractéristiques essentielles de la loi de probabilité empirique sont en fait des quantités bien connues :

- La fonction de répartition de la loi de probabilité empirique est la fonction de répartition empirique F_n :

$$P(X_e \leq x) = \sum_{x_i \leq x} P(X_e = x_i) = \frac{1}{n} \times \text{nombre de } x_i \leq x = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x\}} = F_n(x).$$

- L'espérance de la loi de probabilité empirique est la moyenne empirique \bar{x}_n :

$$E(X_e) = \sum_{i=1}^n x_i P(X_e = x_i) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n.$$

- La variance de la loi de probabilité empirique est la variance empirique s_n^2 :

$$\text{Var}(X_e) = E[(X_e - E[X_e])^2] = \sum_{i=1}^n (x_i - \bar{x}_n)^2 P(X_e = x_i) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = s_n^2.$$

- Le moment empirique d'ordre k est :

$$m_k^e = E[X_e^k] = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

- Le moment empirique centré d'ordre k est :

$$\mu_k^e = E[(X_e - E[X_e])^k] = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^k.$$

- Les quantiles de la loi de probabilité empirique sont les quantiles empiriques :

$$\forall p \in]0, 1[, \tilde{q}_{n,p} = \begin{cases} \frac{1}{2}(x_{np}^* + x_{np+1}^*) & \text{si } np \text{ est entier,} \\ x_{[np]+1}^* & \text{sinon.} \end{cases}$$

Remarque. Puisqu'on considère les observations x_1, \dots, x_n comme des réalisations de variables aléatoires X_1, \dots, X_n , toutes les quantités définies dans cette section sont elles-mêmes des réalisations de variables aléatoires :

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

$$\tilde{Q}_{n,p} = \begin{cases} \frac{1}{2}(X_{np}^* + X_{np+1}^*) & \text{si } np \text{ est entier} \\ X_{[np]+1}^* & \text{sinon} \end{cases}$$

6.2 Estimation de l'espérance d'un échantillon

6.2.1 Estimation ponctuelle

On a déjà vu que la moyenne empirique \bar{X}_n est un estimateur sans biais et convergent (presque sûrement et en moyenne quadratique) de $E(X)$:

$$E(\bar{X}_n) = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} nE(X) = E(X)$$

$$Var(\bar{X}_n) = Var\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{Var(X)}{n}$$

qui tend vers 0 quand n tend vers l'infini. La convergence presque sûre est une conséquence directe de la loi forte des grands nombres.

6.2.2 Intervalle de confiance

Donner un intervalle de confiance de seuil α pour $E(X)$, c'est donner un intervalle aléatoire I tel que $P(E(X) \in I) = 1 - \alpha$.

Etant donné que \bar{X}_n est un bon estimateur de $E(X)$, il est logique de chercher un intervalle de confiance de la forme $I = [\bar{X}_n - a_\alpha, \bar{X}_n + a_\alpha]$. a_α est déterminé en écrivant :

$$P(\bar{X}_n - a_\alpha \leq E(X) \leq \bar{X}_n + a_\alpha) = P(|\bar{X}_n - E(X)| \leq a_\alpha) = 1 - \alpha.$$

Il est donc nécessaire de connaître la loi de probabilité de $|\bar{X}_n - E(X)|$ pour déterminer a_α . Dans un cadre paramétrique, c'est parfois possible, mais ça ne l'est pas si on ne fait pas d'hypothèses particulières sur la loi des X_i . Aussi est-on obligés de recourir à un résultat asymptotique. Le théorème central-limite dit que :

$$\sqrt{n} \frac{\bar{X}_n - E(X)}{\sigma(X)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

On dit aussi qu'asymptotiquement, \bar{X}_n est de loi $\mathcal{N}(E(X), Var(X)/n)$. Par conséquent, quand n est suffisamment grand, on a :

$$P(|\bar{X}_n - E(X)| \leq a_\alpha) = P\left(\sqrt{n} \frac{|\bar{X}_n - E(X)|}{\sigma(X)} \leq \sqrt{n} \frac{a_\alpha}{\sigma(X)}\right) = P(|U| \leq \sqrt{n} \frac{a_\alpha}{\sigma(X)})$$

où U est une variable aléatoire de loi $\mathcal{N}(0, 1)$.

Alors, avec les notations habituelles, on a asymptotiquement :

$$P(|\bar{X}_n - E(X)| \leq a_\alpha) = 1 - \alpha \implies \sqrt{n} \frac{a_\alpha}{\sigma(X)} = u_\alpha \implies a_\alpha = u_\alpha \frac{\sigma(X)}{\sqrt{n}}.$$

Et un intervalle de confiance asymptotique pour $E(X)$ est donc :

$$\left[\bar{X}_n - u_\alpha \frac{\sigma(X)}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{\sigma(X)}{\sqrt{n}}\right].$$

Comme d'habitude, cet intervalle de confiance est inexploitable car il est fonction de $\sigma(X)$, qui est inconnu. Une solution naturelle est alors de remplacer $\sigma(X)$ par l'écart-type empirique S_n dans l'expression de l'intervalle de confiance.

Il reste alors à déterminer quelles conséquences a ce remplacement de l'écart-type théorique par l'écart-type empirique. Pour cela, il faut utiliser le théorème de Slutsky, vu au chapitre 4.

Ce théorème dit que, si $\{U_n\}_{n \geq 1}$ est une suite de variables aléatoires convergeant en loi et $\{V_n\}_{n \geq 1}$ une suite de variables aléatoires convergeant en probabilité vers une constante c , alors pour toute fonction continue g , la suite $\{g(U_n, V_n)\}_{n \geq 1}$ a même limite en loi que la suite $\{g(U_n, c)\}_{n \geq 1}$.

Ici, soit $U_n = \sqrt{n}(\bar{X}_n - E(X))$. $\{U_n\}_{n \geq 1}$ converge en loi vers la loi $\mathcal{N}(0, Var(X))$.

La loi des grands nombres appliquée aux X_i^2 permet d'écrire que $\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{PS} E(X^2)$.

Comme par ailleurs, $\bar{X}_n \xrightarrow{PS} E(X)$, on obtient que :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \xrightarrow{PS} E(X^2) - E(X)^2 = Var(X).$$

Comme la convergence presque sûre entraîne la convergence en probabilité, on obtient que $S_n^2 \xrightarrow{P} \text{Var}(X)$, d'où $V_n = S_n \xrightarrow{P} \sigma(X)$.

Alors, puisque la fonction $g(u, v) = \frac{u}{v}$ est continue sur $\mathbb{R} \times \mathbb{R}^*$, le théorème de Slutsky prouve que :

$$\sqrt{n} \frac{\bar{X}_n - E(X)}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Il suffit alors d'appliquer la même démarche que précédemment, et on obtient :

Propriété 15 . *Un intervalle de confiance asymptotique de seuil α pour $E(X)$ est :*

$$\left[\bar{X}_n - u_\alpha \frac{S_n}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{S_n}{\sqrt{n}} \right].$$

Remarque. Rappelons que dans un contexte paramétrique, un intervalle de confiance de seuil α pour la moyenne m de la loi normale $\mathcal{N}(m, \sigma^2)$ au vu d'un échantillon est :

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

Pour n grand, la loi de Student se rapproche de la loi normale et l'intervalle de confiance proposé est équivalent à celui de la propriété 15.

6.3 Estimation de la variance d'un échantillon

6.3.1 Estimation ponctuelle

On sait déjà que la variance empirique S_n^2 est un estimateur biaisé de la variance de l'échantillon et que la variance estimée $S_n'^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ est un estimateur sans biais et convergent en moyenne quadratique de $\text{Var}(X)$.

Dans la section précédente, on a montré que S_n^2 converge presque sûrement vers $\text{Var}(X)$. C'est évidemment aussi le cas de $S_n'^2$.

Enfin, on montre que, si $E[X^4] < \infty$, alors la variance de la variance estimée est :

$$\text{Var}(S_n'^2) = \frac{1}{n(n-1)} [(n-1)\mu_4 - (n-3)\mu_2^2]$$

avec $\mu_4 = E[(X - E[X])^4]$ et $\mu_2 = \text{Var}(X)$.

6.3.2 Intervalle de confiance

On peut montrer que, si $E[X^4] < \infty$, alors le comportement asymptotique de la variance estimée est déterminé par :

$$\sqrt{n} \frac{S_n'^2 - \mu_2}{\sqrt{\mu_4 - \mu_2^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

En utilisant le théorème de Slutsky, on montre que :

Propriété 16 . Un intervalle de confiance asymptotique de seuil α pour $Var(X) = \mu_2$ est :

$$\left[S_n'^2 - \frac{u_\alpha}{\sqrt{n}} \sqrt{\mu_4^e - S_n'^4}, S_n'^2 + \frac{u_\alpha}{\sqrt{n}} \sqrt{\mu_4^e - S_n'^4} \right]$$

où $\mu_4^e = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^4$.

6.3.3 Lien entre moyenne et variance empiriques

Dans la mesure où la moyenne et la variance empiriques sont deux quantités calculées à l'aide des mêmes observations, ce ne sont a priori pas des variables aléatoires indépendantes.

Propriété 17 . Si $E(X^3) < \infty$, alors $Cov(\bar{X}_n, S_n'^2) = \frac{\mu_3}{n}$.

On en déduit que \bar{X}_n et $S_n'^2$ sont corrélées mais asymptotiquement non corrélées.

On peut montrer que si la loi des X_i est symétrique, alors $\mu_3 = 0$. Donc dans ce cas, \bar{X}_n et $S_n'^2$ sont non corrélées pour tout n .

On sait que l'indépendance entraîne la non-corrélation mais que la réciproque est fautive. En fait, on montre que \bar{X}_n et $S_n'^2$ sont indépendantes si et seulement si les X_i sont de loi normale.

6.4 Estimation des moments de tous ordres

Comme pour l'espérance et la variance, on peut estimer les moments d'ordre k , $m_k = E[X^k]$, et les moments centrés d'ordre k , $\mu_k = E[(X - E[X])^k]$, de la loi de l'échantillon par les moments empiriques correspondants $m_k^e = \frac{1}{n} \sum_{i=1}^n X_i^k$ et $\mu_k^e = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k$.

Les propriétés de ces estimateurs sont données par la propriété suivante :

Propriété 18 .

$$\begin{aligned} m_k^e &\xrightarrow{PS} m_k \\ \mu_k^e &\xrightarrow{PS} \mu_k \\ \sqrt{n} \frac{m_k^e - m_k}{\sqrt{m_{2k} - m_k^2}} &\xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \\ \sqrt{n} \frac{\mu_k^e - \mu_k}{\sqrt{k^2 \mu_{k-1} \mu_{k+1} + 2k \mu_{k-1} \mu_{k+1} + \mu_{2k} - \mu_k^2}} &\xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \end{aligned}$$

Les résultats de convergence en loi et le théorème de Slutsky permettent d'obtenir des intervalles de confiance asymptotiques pour tous les moments.

On n'a pas de résultat non asymptotique, par exemple sur le biais de ces estimateurs.

Enfin, ces résultats interviennent dans l'établissement des propriétés de la méthode d'estimation paramétrique des moments.

6.5 Estimation des quantiles

On s'intéresse maintenant à l'estimation des quantiles de la loi de l'échantillon. Pour simplifier, on se bornera ici au cas où la loi des observations est continue et F est strictement croissante. Alors le quantile d'ordre p est $q_p = F^{-1}(p)$. On se propose de l'estimer par le quantile empirique d'ordre p ,

$$\tilde{Q}_{n,p} = \begin{cases} \frac{1}{2}(X_{np}^* + X_{np+1}^*) & \text{si } np \text{ est entier,} \\ X_{\lfloor np \rfloor + 1}^* & \text{sinon.} \end{cases}$$

6.5.1 Propriétés des quantiles empiriques

Connaissant la loi d'une statistique d'ordre et la loi conjointe d'un couple de statistiques d'ordre, il est facile de déterminer la loi d'un quantile empirique, donnée par sa densité :

Théorème 14 . Si np est entier,

$$f_{\tilde{Q}_{n,p}}(x) = \frac{2n!}{(np-1)!(n-np-1)!} \int_x^{+\infty} F(2x-y)^{np-1} (1-F(y))^{n-np-1} f(2x-y) f(y) dy.$$

Si np n'est pas entier,

$$f_{\tilde{Q}_{n,p}}(x) = \frac{n!}{\lfloor np \rfloor! (n - \lfloor np \rfloor - 1)!} F(x)^{\lfloor np \rfloor} (1-F(x))^{n-\lfloor np \rfloor - 1} f(x).$$

Démonstration. Le cas où np n'est pas entier est immédiat car on a directement la densité de $X_{\lfloor np \rfloor + 1}^*$.

Quand np est entier, on utilise la loi conjointe de (X_{np}^*, X_{np+1}^*) en écrivant :

$$F_{\tilde{Q}_{n,p}}(x) = P\left(\frac{1}{2}(X_{np}^* + X_{np+1}^*) \leq x\right) = \int \int_{\frac{z+y}{2} \leq x} f_{(X_{np}^*, X_{np+1}^*)}(z, y) dz dy$$

et on obtient le résultat annoncé par dérivation. ■

On obtient donc entre autres ainsi la loi de probabilité de la médiane d'un échantillon, mais cette loi dépend de f et F , qui sont inconnues.

On a également un résultat sur la loi asymptotique d'un quantile empirique :

Théorème 15 . **Théorème de Mosteller** :

$$\forall p \in]0, 1[, \sqrt{n} \frac{\tilde{Q}_{n,p} - q_p}{\sqrt{p(1-p)}} f(q_p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

6.5.2 Estimation ponctuelle

Le théorème 14 montre que le calcul de $E(\tilde{Q}_{n,p})$ n'est pas simple. En fait, on n'a pas de résultat non asymptotique sur la qualité d'estimation de q_p par $\tilde{Q}_{n,p}$. En revanche, le théorème de Mosteller permet d'établir un résultat asymptotique.

Propriété 19 . $\tilde{Q}_{n,p}$ est un estimateur de q_p asymptotiquement sans biais et convergent presque sûrement.

Démonstration. Le théorème de Mosteller dit que la loi asymptotique de $\tilde{Q}_{n,p}$ est la loi $\mathcal{N}(q_p, \frac{p(1-p)}{nf^2(q_p)})$, ce qui prouve directement que $\tilde{Q}_{n,p}$ est asymptotiquement sans biais et est convergent en moyenne quadratique. Pour la convergence presque sûre, il faut utiliser un résultat sur la fonction de répartition empirique, le théorème de Glivenko-Cantelli, qui sera énoncé dans le chapitre suivant. ■

En conclusion, il est justifié, au moins si on a beaucoup d'observations, d'estimer un quantile théorique par un quantile empirique. Dans certains cas, certains moments et quantiles théoriques sont confondus. C'est le cas par exemple pour les lois symétriques pour lesquelles l'espérance et la médiane sont confondus. Il est alors important de déterminer lequel des deux estimateurs empiriques correspondants est le meilleur.

6.5.3 Intervalle de confiance

Contrairement à ce qu'on avait pour les moments, le théorème de Mosteller ne permet pas de construire un intervalle de confiance asymptotique pour q_p en utilisant le théorème de Slutsky, car on ne sait pas estimer simplement $f(q_p)$. En fait, on a ici un résultat non asymptotique.

Théorème 16 . $\forall(i, j), 1 \leq i < j \leq n$, on a :

$$P(X_i^* \leq q_p \leq X_j^*) = \sum_{k=i}^{j-1} C_n^k p^k (1-p)^{n-k}.$$

Démonstration. Il suffit d'écrire :

$$\begin{aligned} P(X_i^* \leq q_p \leq X_j^*) &= P(X_i^* \leq q_p) - P(X_j^* < q_p) = F_{X_i^*}(q_p) - F_{X_j^*}(q_p) \\ &= \sum_{k=i}^n C_n^k [F(q_p)]^k [1 - F(q_p)]^{n-k} - \sum_{k=j}^n C_n^k [F(q_p)]^k [1 - F(q_p)]^{n-k} \\ &= \sum_{k=i}^{j-1} C_n^k [F(q_p)]^k [1 - F(q_p)]^{n-k}. \end{aligned}$$

Or $F(q_p) = p$, donc on obtient $P(X_i^* \leq q_p \leq X_j^*) = \sum_{k=i}^{j-1} C_n^k p^k (1-p)^{n-k}$. ■

Corollaire 3 . S'il existe i et j tels que $\sum_{k=i}^{j-1} C_n^k p^k (1-p)^{n-k} = 1 - \alpha$, alors $[X_i^*, X_j^*]$ est un intervalle de confiance de seuil α pour q_p .

Dans la pratique, on cherche le couple (i, j) tel que cette somme soit la plus proche possible de $1 - \alpha$. $[X_i^*, X_j^*]$ sera alors un intervalle de confiance de seuil proche de α (et connu) pour q_p .

Par exemple, si on s'intéresse à la médiane, on a $p = \frac{1}{2}$. On cherche donc i et j tels que $\sum_{k=i}^{j-1} C_n^k p^k (1-p)^{n-k} = \frac{1}{2^n} \sum_{k=i}^{j-1} C_n^k$ soit proche de $1 - \alpha$.

Pour $n = 10$, on a $\frac{1}{2^{10}} \sum_{k=3}^7 C_{10}^k \simeq 89\%$. On en déduit que $[X_3^*, X_8^*]$ est un intervalle de confiance de seuil approximativement égal à 11% pour la médiane de l'échantillon (rappelons que la médiane empirique dans ce cas est $\frac{1}{2}(X_5^* + X_6^*)$).

L'intérêt principal de ce résultat est qu'il n'est pas asymptotique, ce qui est assez rare en statistique non paramétrique. Cependant, ces intervalles sont en général très larges, ce qui les rend assez peu utiles en pratique si on a peu d'observations.

6.6 Lois asymptotiques des extrêmes

Le théorème de Mosteller dit que, pour p fixé, la loi asymptotique de $X_{[np]+1}^*$ est la loi $\mathcal{N}(q_p, \frac{p(1-p)}{nf^2(q_p)})$.

On peut ainsi obtenir la loi asymptotique des statistiques d'ordre "centrales", par exemple de la médiane empirique. En revanche, le théorème de Mosteller ne permet pas d'obtenir la loi asymptotique des statistiques d'ordre "extrêmes", c'est à dire X_1^* et X_n^* .

En effet, $X_{[np]+1}^* = X_1^* \Leftrightarrow [np] = 0 \Leftrightarrow np < 1 \Leftrightarrow p < \frac{1}{n}$.

Or, pour p fixé, en faisant tendre n vers l'infini, on finira forcément par avoir $\frac{1}{n} < p$.

De même, $X_{[np]+1}^* = X_n^* \Leftrightarrow [np] = n - 1 \Leftrightarrow np \geq n - 1 \Leftrightarrow p \geq 1 - \frac{1}{n}$.

Et pour p fixé, en faisant tendre n vers l'infini, on finira forcément par avoir $1 - \frac{1}{n} > p$.

Par conséquent, les lois asymptotiques de X_1^* et X_n^* ne peuvent pas être obtenues à l'aide du résultat sur la loi asymptotique de $X_{[np]+1}^*$.

En fait, X_1^* et X_n^* convergent en loi vers les bornes inférieure et supérieure du support de la loi de l'échantillon.

En effet, $\lim_{n \rightarrow +\infty} F_{X_1^*}(x) = \lim_{n \rightarrow +\infty} [1 - (1 - F(x))^n] = \begin{cases} 0 & \text{si } F(x) = 0 \\ 1 & \text{si } F(x) > 0 \end{cases}$.

Par exemple, si la loi des X_i est la loi uniforme sur $[a, b]$, X_1^* converge en loi vers a et X_n^* converge en loi vers b . Si c'est la loi exponentielle, X_1^* converge en loi vers 0.

En fait, au lieu de s'intéresser à la loi asymptotique de X_1^* , on va s'intéresser à celle de $\frac{X_1^* - b_n}{a_n}$ où $\{a_n\}_{n \geq 1}$ et $\{b_n\}_{n \geq 1}$ sont des suites de réels bien choisies.

Le théorème de Gnedenko dit que, dans ce cas, il n'y a que 3 familles de lois limites possibles.

Théorème 17 . Théorème de Gnedenko : Soit X_1, \dots, X_n un échantillon d'une loi continue. S'il existe des suites de réels strictement positifs $\{a_n\}_{n \geq 1}$ et de réels $\{b_n\}_{n \geq 1}$ telles que $\frac{X_1^* - b_n}{a_n}$ converge en loi vers une loi limite, alors les seules lois limites possibles, définies par leur fonction de répartition G , sont :

- $G(x) = 1 - e^{-e^x}, x \in \mathbb{R}$ (première loi de Gumbel).
- $G(x) = 1 - e^{-x^\beta}, x \geq 0, \beta > 0$ (loi de Weibull $\mathcal{W}(1, \beta)$).
- $G(x) = 1 - e^{-(-x)^{-\beta}}, x \leq 0, \beta > 0$ (loi de $-\frac{1}{X}$ quand X est de loi $\mathcal{W}(1, \beta)$).

De même, les seules lois limites possibles pour les suites $\frac{X_n^* - b_n}{a_n}$ sont :

- $G(x) = e^{-e^{-x}}, x \in \mathbb{R}$ (deuxième loi de Gumbel).
- $G(x) = e^{-x^{-\beta}}, x \geq 0, \beta > 0$ (loi de $\frac{1}{X}$ quand X est de loi $\mathcal{W}(1, \beta)$).
- $G(x) = e^{-(-x)^\beta}, x \leq 0, \beta > 0$ (loi de $-X$ quand X est de loi $\mathcal{W}(1, \beta)$).

Pour une loi donnée, même s'il existe plusieurs suites $\{a_n\}_{n \geq 1}$ et $\{b_n\}_{n \geq 1}$ possibles, la famille de lois limite est toujours la même.

Par exemple, si $\frac{X_1^* - b_n}{a_n}$ converge en loi vers la loi de Weibull, on dit que la loi des X_i appartient au **domaine d'attraction du minimum** de la loi de Weibull.

La constante 0 (loi de Dirac en 0) est une loi limite particulière qui correspond à β infini.

Ce qui est remarquable dans ce résultat, c'est que, pour une fois, les lois asymptotiques ne sont pas des lois normales. Il existe donc une différence de comportement notable entre les statistiques d'ordre "centrales" et les statistiques d'ordre "extrêmes".

D'un point de vue pratique, dès qu'un phénomène peut s'interpréter comme un maximum ou un minimum (par exemple une durée de vie ou bien un pic d'ozone), les lois de probabilité du théorème de Gnedenko peuvent être utilisées comme modèles. C'est essentiellement pour cela que les lois de Weibull et de Gumbel sont utilisées.

Chapitre 7

Estimation fonctionnelle

Les hypothèses de ce chapitre sont les mêmes que celles du chapitre précédent : on suppose que les observations x_1, \dots, x_n sont des réalisations de variables aléatoires réelles X_1, \dots, X_n indépendantes et de même loi, de fonction de répartition F , et de densité f , si elle existe.

Dans le chapitre précédent, on s'est intéressé à l'estimation de quantités réelles caractéristiques de la loi de probabilité de l'échantillon, les moments et les quantiles. Aussi riches d'enseignement que soient ces quantités, elles ne suffisent pas à déterminer entièrement la loi de probabilité de l'échantillon.

C'est pourquoi nous allons maintenant nous intéresser à l'estimation de la fonction de répartition et, si elle existe, de la densité de l'échantillon. Par rapport au chapitre précédent, il s'agit maintenant d'estimer des fonctions, d'où le nom d'**estimation fonctionnelle**. De plus l'une comme l'autre de ces fonctions caractérisent entièrement la loi de probabilité de l'échantillon.

La fonction de répartition empirique est un estimateur simple et performant de la fonction de répartition de l'échantillon. Il est beaucoup plus difficile d'estimer une densité. On connaît déjà l'estimateur de base de la densité d'un échantillon, l'histogramme. Bien que très connu et très utilisé, il est de médiocre qualité. Aussi allons-nous proposer une méthode d'estimation de densité bien plus performante, la méthode du noyau.

Remarquons que l'estimation des quantiles peut être considérée comme de l'estimation fonctionnelle dans la mesure où estimer $q_p = F^{-1}(p)$ quel que soit p revient à estimer la fonction F^{-1} .

Estimer une fonction g , c'est d'abord estimer $g(x)$ pour tout x donné. Il faut ensuite juger de la qualité de l'estimation de $g(x)$ pour chaque x , puis de l'estimation de g dans son ensemble.

Si $\hat{g}(x)$ est un estimateur de $g(x)$, la qualité de l'estimation pour un x donné est usuellement mesurée par le biais, la variance et l'Erreur Quadratique Moyenne (ou risque quadratique), qu'on notera $EQM_x(\hat{g})$:

$$EQM_x(\hat{g}) = E[(\hat{g}(x) - g(x))^2] = [E(\hat{g}(x)) - g(x)]^2 + Var(\hat{g}(x)).$$

On voit que l'erreur quadratique moyenne se décompose en un terme de biais et un terme de variance. Si $\hat{g}(x)$ est un estimateur sans biais de $g(x)$, l'erreur quadratique moyenne se réduit à la variance. On verra que, si on peut trouver facilement un estimateur sans biais pour la fonction de répartition en un point x , il n'en est pas de même pour la

densité. Aussi utilisera-t-on l'erreur quadratique moyenne plutôt que la variance dans ce cas.

Pour juger de la qualité de l'estimation de g dans son ensemble, il faut utiliser des mesures de l'écart entre g et \hat{g} . Suivant les cas, on utilisera :

- l'**Erreur Quadratique Moyenne Intégrée** (EQMI) :

$$EQMI(\hat{g}) = \int_{-\infty}^{+\infty} EQM_x(\hat{g}) dx = \int_{-\infty}^{+\infty} [E(\hat{g}(x)) - g(x)]^2 dx + \int_{-\infty}^{+\infty} Var(\hat{g}(x)) dx.$$

- l'écart maximum entre les deux fonctions :

$$\sup\{|\hat{g}(x) - g(x)|; x \in \mathbb{R}\}.$$

7.1 Estimation de la fonction de répartition

7.1.1 Estimation ponctuelle

Rappelons que la fonction de répartition empirique \mathbb{F}_n de l'échantillon est définie par :

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}} = \text{pourcentage d'observations inférieures à } x$$

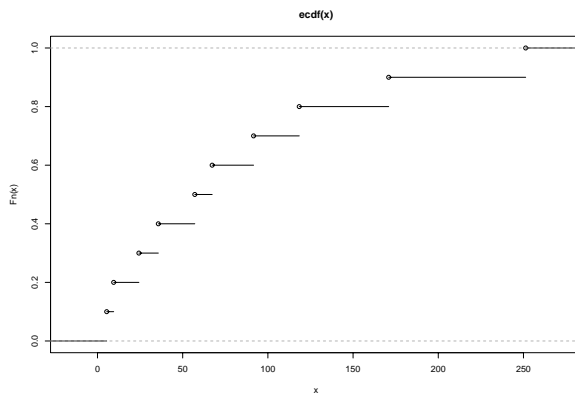


FIGURE 7.1 – Fonction de répartition empirique

Il s'avère que \mathbb{F}_n est un excellent estimateur de F , ce que l'on peut montrer en plusieurs étapes.

Propriété 20 . $\forall x \in \mathbb{R}$, $n\mathbb{F}_n(x)$ est de loi binomiale $\mathcal{B}(n, F(x))$.

Démonstration. $n\mathbb{F}_n(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}$ est une somme de n variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $P(X_i \leq x) = F(x)$, donc c'est une variable aléatoire de loi $\mathcal{B}(n, F(x))$.

On peut dire aussi que $n\mathbb{F}_n(x)$ est le nombre de X_i inférieurs à x , qui peut s'interpréter comme le nombre de fois où, en n expériences identiques et indépendantes, un évènement de probabilité $P(X_i \leq x) = F(x)$ s'est produit. Donc c'est une variable aléatoire de loi $\mathcal{B}(n, F(x))$. ■

On en déduit facilement les qualités de l'estimation de $F(x)$ par $\mathbb{F}_n(x)$.

Propriété 21 . $\forall x \in \mathbb{R}$, $\mathbb{F}_n(x)$ est un estimateur sans biais et convergent en moyenne quadratique de $F(x)$.

Démonstration. $E(\mathbb{F}_n(x)) = \frac{1}{n} E(n\mathbb{F}_n(x)) = \frac{1}{n} nF(x) = F(x)$.

$$\begin{aligned} \text{Var}(\mathbb{F}_n(x)) &= \frac{1}{n^2} \text{Var}(n\mathbb{F}_n(x)) = \frac{1}{n^2} nF(x)(1 - F(x)) \\ &= \frac{F(x)(1 - F(x))}{n}, \end{aligned}$$

qui tend vers 0 quand n tend vers l'infini. ■

En fait, la convergence est presque sûre :

Propriété 22 : $\forall x \in \mathbb{R}$, $\mathbb{F}_n(x) \xrightarrow{PS} F(x)$.

Démonstration. Il suffit d'appliquer la loi des grands nombres aux variables aléatoires de loi de Bernoulli $\mathbb{1}_{\{X_i \leq x\}}$:

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \xrightarrow{PS} E(\mathbb{1}_{\{X_i \leq x\}}) = F(x).$$

Vue sous cet angle, la loi des grands nombres dit que la probabilité d'un évènement est la limite de la fréquence d'occurrence de cet évènement dans une suite d'expériences identiques et indépendantes. On en déduit que l'on peut estimer la probabilité que X soit inférieure à x , $F(x)$, par le pourcentage d'observations inférieures à x , $\mathbb{F}_n(x)$. Cette estimation est d'excellente qualité.

Pour juger de la qualité globale de l'estimation de F par \mathbb{F}_n , on utilise le théorème de Glivenko-Cantelli, qui dit que \mathbb{F}_n est un estimateur convergent uniformément et presque sûrement de F :

Théorème 18 . **Théorème de Glivenko-Cantelli.**

$$D_n = \sup\{|\mathbb{F}_n(x) - F(x)|; x \in \mathbb{R}\} \xrightarrow{PS} 0.$$

Par ailleurs, l'erreur quadratique moyenne intégrée est :

$$EQMI(\mathbb{F}_n) = \int_{-\infty}^{+\infty} \text{Var}(\mathbb{F}_n(x)) dx = \frac{1}{n} \int_{-\infty}^{+\infty} F(x)(1 - F(x)) dx.$$

On ne peut pas calculer explicitement cette erreur, mais on sait qu'elle tend vers 0 quand n tend vers l'infini à la vitesse $1/n$.

7.1.2 Intervalle de confiance

Soit x fixé. Un intervalle de confiance de seuil α pour $F(x)$ est un intervalle aléatoire I tel que $P(F(x) \in I) = 1 - \alpha$.

Si on reprend la démarche vue en 6.2.2. pour l'espérance de l'échantillon, on va chercher un intervalle de confiance de la forme $I = [\mathbb{F}_n(x) - a_\alpha, \mathbb{F}_n(x) + a_\alpha]$, où a_α est déterminé en écrivant :

$$\begin{aligned}
 P(F(x) \in I) &= P(\mathbb{F}_n(x) - a_\alpha \leq F(x) \leq \mathbb{F}_n(x) + a_\alpha) \\
 &= P(F(x) - a_\alpha \leq \mathbb{F}_n(x) \leq F(x) + a_\alpha) \\
 &= P(n(F(x) - a_\alpha) \leq n\mathbb{F}_n(x) \leq n(F(x) + a_\alpha)) \\
 &= \sum_{k=\lfloor n(F(x)-a_\alpha) \rfloor + 1}^{\lfloor n(F(x)+a_\alpha) \rfloor} C_n^k [F(x)]^k [1 - F(x)]^{n-k} \\
 &= 1 - \alpha
 \end{aligned}$$

On ne peut pas déduire la valeur de a_α de cette expression car elle implique $F(x)$, qui est inconnue. En revanche, on peut obtenir un résultat asymptotique par un raisonnement similaire à celui que l'on a utilisé pour l'espérance.

En effet, l'application du théorème central-limite sur les $\mathbb{1}_{\{X_i \leq x\}}$, variables aléatoires indépendantes de loi de Bernoulli, d'espérance $F(x)$ et de variance $F(x)(1 - F(x))$ permet d'écrire :

$$\frac{\sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} - nE(\mathbb{1}_{\{X_i \leq x\}})}{\sqrt{n \text{Var}(\mathbb{1}_{\{X_i \leq x\}})}} = \frac{n\mathbb{F}_n(x) - nF(x)}{\sqrt{nF(x)(1 - F(x))}} = \sqrt{n} \frac{\mathbb{F}_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Grâce au théorème de Slutsky et à la convergence presque sûre de $\mathbb{F}_n(x)$ vers $F(x)$, on a également :

$$\sqrt{n} \frac{\mathbb{F}_n(x) - F(x)}{\sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Alors on obtient que, pour n suffisamment grand :

$$\begin{aligned}
 P(F(x) \in I) &= P(-a_\alpha \leq \mathbb{F}_n(x) - F(x) \leq a_\alpha) = P(|\mathbb{F}_n(x) - F(x)| \leq a_\alpha) \\
 &= P\left(\sqrt{n} \frac{|\mathbb{F}_n(x) - F(x)|}{\sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}} \leq \sqrt{n} \frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}}\right) \\
 &= P(|U| \leq \sqrt{n} \frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}}) \\
 &= 1 - \alpha
 \end{aligned}$$

où U est de loi $\mathcal{N}(0, 1)$.

$$\text{D'où } \sqrt{n} \frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}} = u_\alpha \text{ et } a_\alpha = \frac{u_\alpha}{\sqrt{n}} \sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))}.$$

Et on obtient finalement :

Propriété 23 . $\forall x \in \mathbb{R}$, un intervalle de confiance asymptotique de seuil α pour $F(x)$ est :

$$\left[\mathbb{F}_n(x) - \frac{u_\alpha}{\sqrt{n}} \sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))} \quad , \quad \mathbb{F}_n(x) + \frac{u_\alpha}{\sqrt{n}} \sqrt{\mathbb{F}_n(x)(1 - \mathbb{F}_n(x))} \right].$$

En fait, on a des résultats sur les intervalles de confiance pour le paramètre de la loi binomiale qui nous donnent directement le résultat suivant :

Théorème 19 . $\forall x \in \mathbb{R}$, un intervalle de confiance exact de seuil α pour $F(x)$ est :

$$\left[\frac{1}{1 + \frac{n - n\mathbb{F}_n(x) + 1}{n\mathbb{F}_n(x)} f_{2(n - n\mathbb{F}_n(x) + 1), 2n\mathbb{F}_n(x), \alpha/2}} \quad , \quad \frac{1}{1 + \frac{n - n\mathbb{F}_n(x)}{n\mathbb{F}_n(x) + 1} f_{2(n - n\mathbb{F}_n(x)), 2(n\mathbb{F}_n(x) + 1), 1 - \alpha/2}} \right]$$

où $f_{\nu_1, \nu_2, \alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi de Fisher-Snedecor à (ν_1, ν_2) degrés de liberté.

7.2 Estimation de la densité

Dans cette section, on suppose que la loi de l'échantillon est continue et on cherche à estimer sa densité f . f est la dérivée de F , mais la fonction de répartition empirique \mathbb{F}_n n'est pas dérivable, puisque c'est une fonction en escalier. On ne peut donc pas utiliser directement les résultats sur la fonction de répartition empirique pour estimer la densité.

On peut se demander quelle est l'utilité d'estimer la densité alors que l'on a déjà un très bon estimateur de la fonction de répartition. La principale raison est que la forme d'une densité est beaucoup plus facile à interpréter que celle d'une fonction de répartition. Par exemple, on pourra facilement avoir, grâce à une estimation de densité, des informations sur la symétrie ou la multimodalité de la loi de l'échantillon, alors que ce n'est pas du tout facile au seul vu de la fonction de répartition empirique. De même, une estimation de densité est une aide importante au choix d'un modèle approprié pour la loi de l'échantillon. Par exemple, une densité estimée en forme de cloche symétrique peut conduire à l'adoption d'un modèle de loi normale.

Nous allons commencer par donner des rappels sur la méthode d'estimation de densité la plus élémentaire, celle de l'histogramme. Puis nous présenterons la méthode plus sophistiquée du noyau.

7.2.1 Rappels sur les histogrammes

On se fixe une borne inférieure de l'échantillon $a_0 < x_1^*$ et une borne supérieure $a_k > x_n^*$. On partitionne l'intervalle $]a_0, a_k]$, contenant toutes les observations, en k classes $]a_{j-1}, a_j]$. La largeur de la classe j est $h_j = a_j - a_{j-1}$.

L'effectif de la classe j est le nombre d'observations appartenant à cette classe : $n_j = \sum_{i=1}^n \mathbb{1}_{]a_{j-1}, a_j]}(x_i)$. La fréquence de la classe j est $\frac{n_j}{n}$.

L'histogramme est constitué de rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes. Donc l'histogramme est la fonction en

escalier constante sur les classes et qui vaut $\frac{n_j}{nh_j}$ sur la classe $]a_{j-1}, a_j]$. Cette fonction peut s'écrire :

$$\hat{f}(x) = \sum_{j=1}^k \frac{n_j}{nh_j} \mathbb{1}_{]a_{j-1}, a_j]}(x) = \frac{1}{n} \sum_{j=1}^k \frac{1}{h_j} \mathbb{1}_{]a_{j-1}, a_j]}(x) \sum_{i=1}^n \mathbb{1}_{]a_{j-1}, a_j]}(x_i).$$

Dans l'histogramme à pas fixe, les classes sont de même largeur $h = \frac{a_k - a_0}{k}$. Dans ce cas, la hauteur d'un rectangle est proportionnelle à l'effectif de sa classe.

On a vu en PMS qu'il était plus pertinent de choisir un histogramme à classes de même effectif. Admettons pour simplifier que n soit divisible par k . Alors chaque classe doit contenir n/k observations. Les limites des classes seront alors les j/k quantiles empiriques :

$$a_j = \tilde{q}_{n,j/k} = \frac{1}{2}(x_{\frac{n_j}{k}}^* + x_{\frac{n_j}{k}+1}^*), \quad j = 1, \dots, k-1;$$

Les bornes des classes sont donc cette fois aléatoires, puisqu'elles sont fonction des observations.

Enfin, le polygone des fréquences est la ligne brisée reliant les milieux des sommets des rectangles, et prolongée de part et d'autre de l'histogramme de façon à ce que l'aire totale délimitée par le polygone soit égale à 1, comme pour une densité.

Prenons l'exemple vu en PMS du bruit à Montréal. Les histogrammes à classe de même largeur et de même effectif, avec leurs polygones des fréquences, sont donnés par la figure 7.2.

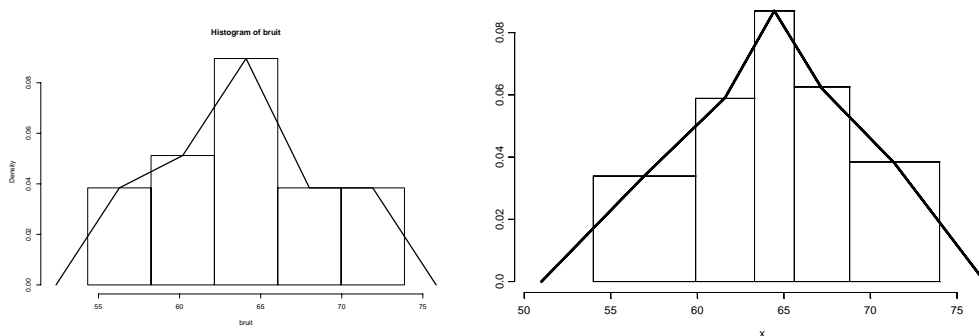


FIGURE 7.2 – Histogramme à classes de même largeur et à classes de même effectif pour les niveaux de bruit à Montréal

La forme de ces histogrammes est assez proche d'une cloche symétrique, ce qui nous amène à envisager l'hypothèse que les données proviennent d'une loi normale.

7.2.2 La méthode du noyau

Les histogrammes et les polygones des fréquences ne sont pas des estimations très satisfaisantes de la densité de l'échantillon car ce sont des fonctions en escalier et des

lignes brisées alors que la densité à estimer est en général plus lisse, avec au moins sa dérivée continue.

D'autre part, l'aléa du au choix du nombre de classes et des bornes des classes est un élément très perturbant de l'analyse, puisque des choix différents peuvent aboutir à des histogrammes d'allures assez nettement différentes.

L'estimation par noyau a pour but de répondre à ces deux écueils et de proposer des estimations de densité ayant de bonnes propriétés.

Pour cela, on commence par remarquer que la densité est la dérivée de la fonction de répartition, ce qui permet d'écrire pour tout x :

$$f(x) = F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Donc pour un $h > 0$ fixé "petit", on peut penser à estimer $f(x)$ par :

$$\hat{f}(x) = \frac{1}{2h} (\mathbb{F}_n(x+h) - \mathbb{F}_n(x-h)) = \frac{1}{2nh} \sum_{i=1}^n \mathbb{1}_{]x-h, x+h]}(X_i).$$

On a alors :

$$E[\hat{f}(x)] = \frac{1}{2h} (E[\mathbb{F}_n(x+h)] - E[\mathbb{F}_n(x-h)]) = \frac{1}{2h} (F(x+h) - F(x-h))$$

qui tend vers $f(x)$ quand h tend vers 0. Il faut donc faire dépendre h de la taille de l'échantillon, et le faire tendre vers 0 quand n tend vers l'infini, de sorte que $\hat{f}(x)$ soit un estimateur asymptotiquement sans biais de $f(x)$. h sera donc dorénavant noté h_n .

Cette démarche est proche de celle de l'histogramme au sens où cela revient à mettre x au centre d'une classe de largeur $2h$ et à calculer l'estimateur histogramme correspondant. La fonction \hat{f} obtenue a des sauts aux points $X_i \pm h$ et est constante autrement.

La grande différence par rapport à l'histogramme est qu'il n'y a pas de classe fixée à l'avance : on crée une classe en chaque point où on veut estimer la densité.

L'estimateur \hat{f} reste une fonction en escalier. Pour obtenir quelque chose de plus lisse, on peut remarquer que :

$$\begin{aligned} \hat{f}(x) &= \frac{1}{2nh_n} \sum_{i=1}^n \mathbb{1}_{]x-h_n, x+h_n]}(X_i) = \frac{1}{nh_n} \sum_{i=1}^n \frac{1}{2} \mathbb{1}_{\{x-h_n < X_i \leq x+h_n\}} \\ &= \frac{1}{nh_n} \sum_{i=1}^n \frac{1}{2} \mathbb{1}_{[-1, +1[}\left(\frac{x - X_i}{h_n}\right) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \end{aligned}$$

où $K(u) = \frac{1}{2} \mathbb{1}_{[-1, +1[}(u)$.

La **méthode du noyau** consiste à généraliser cette approche à d'autres fonctions K .

Définition 25 . Un **estimateur à noyau** de la densité f est une fonction \hat{f} définie par :

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)$$

où $\{h_n\}_{n \geq 1}$ est une suite de réels positifs appelés **paramètres de lissage** ou **largeurs de la fenêtre**, qui tend vers 0 quand n tend vers l'infini, et K est une densité de probabilité appelée **noyau**.

Les noyaux les plus communs sont :

- le **noyau rectangulaire** : $K(u) = \frac{1}{2} \mathbb{1}_{[-1, +1[}(u)$. C'est celui qui donne l'estimateur de type histogramme.
- le **noyau triangulaire** : $K(u) = (1 - |u|) \mathbb{1}_{[-1, +1[}(u)$.
- le **noyau gaussien** : $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$.
- le **noyau d'Epanechnikov** : $K(u) = \frac{3}{4\sqrt{5}} \left(1 - \frac{u^2}{5}\right) \mathbb{1}_{[-\sqrt{5}, +\sqrt{5}[}(u)$.

Dans l'estimation de $f(x)$ par le noyau rectangulaire, le même poids est accordé à toutes les observations comprises entre $x - h$ et $x + h$. Dans les 3 derniers noyaux, le poids d'une observation est d'autant plus fort qu'elle est proche de x .

\hat{f} a les mêmes propriétés de continuité et de différentiabilité que K . Par exemple, si K est le noyau gaussien, \hat{f} admet des dérivées de tous ordres.

Propriété 24 . Un estimateur à noyau est une densité.

Démonstration.

$$\begin{aligned} \int_{-\infty}^{+\infty} \hat{f}(x) dx &= \frac{1}{nh_n} \sum_{i=1}^n \int_{-\infty}^{+\infty} K\left(\frac{x - X_i}{h_n}\right) dx \\ &= \frac{1}{nh_n} \sum_{i=1}^n \int_{-\infty}^{+\infty} K(u) h_n du \quad (\text{changement de variable } u = \frac{x - X_i}{h_n}) \\ &= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{+\infty} K(u) du = \frac{1}{n} n = 1. \end{aligned}$$

■

Pour choisir quel noyau prendre et surtout choisir le paramètre de lissage h_n , il faut étudier la qualité de l'estimation de f par \hat{f} .

Comme les expressions du biais et de la variance de l'estimateur à noyau ne sont pas simples à traiter, on en donne des équivalents pour pouvoir étudier leur comportement asymptotique :

Propriété 25 . Si K est la densité d'une loi de probabilité symétrique par rapport à l'origine et de variance μ_2 , si f admet des dérivées continues de tous ordres, alors, quand n tend vers l'infini, on a :

- $E[\hat{f}(x)] - f(x) \sim \frac{h_n^2 \mu_2}{2} f''(x).$
- $Var[\hat{f}(x)] \sim \frac{f(x)}{nh_n} \int_{-\infty}^{+\infty} K(u)^2 du.$
- $EQMI(\hat{f}) \sim \frac{h_n^4 \mu_2^2}{4} \int_{-\infty}^{+\infty} f''(x)^2 dx + \frac{1}{nh_n} \int_{-\infty}^{+\infty} K(u)^2 du.$

On voit que, dans l'erreur quadratique moyenne intégrée, le terme de biais est une fonction croissante de h_n , alors que le terme de variance est une fonction décroissante de h_n . Si h_n est grand, la variance sera faible, mais le biais sera fort. Si h_n est petit, c'est l'inverse. La valeur de h_n optimale, qui minimise l'EQMI, réalise donc un compromis entre biais et variance.

Cette valeur optimale est une fonction de f , qui est inconnue. On ne peut donc en donner qu'une valeur approchée. En pratique, on choisit souvent :

$$h_n = \left(\frac{4}{3}\right)^{1/5} n^{-1/5} \min\left(s'_n, \frac{1}{1.34}(\tilde{q}_{n,3/4} - \tilde{q}_{n,1/4})\right).$$

En fait, la valeur optimale de h_n dépend aussi de K . On montre que l'erreur quadratique moyenne intégrée minimale est obtenue en choisissant le noyau d'Epanechnikov. Mais l'écart de performance entre les différents noyaux usuels est assez faible, aussi on a plutôt tendance en pratique à choisir le noyau le plus facile à utiliser, qui est le noyau gaussien.

Le biais étant un $O(h_n^2)$, on voit que le biais optimal est un $O(n^{-2/5})$. Par conséquent, $\hat{f}(x)$ est un estimateur asymptotiquement sans biais de $f(x)$, mais la convergence est lente car $n^{-2/5}$ tend lentement vers 0.

De la même façon, la variance optimale est un $O(n^{-4/5})$. Donc $\hat{f}(x)$ est un estimateur convergent de $f(x)$, mais la convergence est plus lente que celle de $\mathbb{F}_n(x)$ vers $F(x)$ car $n^{-4/5}$ tend plus lentement que n^{-1} vers 0.

Ces deux résultats font que, pour pouvoir estimer efficacement une densité, il faut avoir beaucoup de données.

Dans l'exemple des niveaux de bruit, l'estimation de densité par la méthode du noyau gaussien avec le paramètre de lissage ci-dessus est donnée par la commande :

```
> lines(density(bruit,n=200))
```

On obtient la figure 7.3, la densité estimée semble bien proche de celle d'une loi normale.

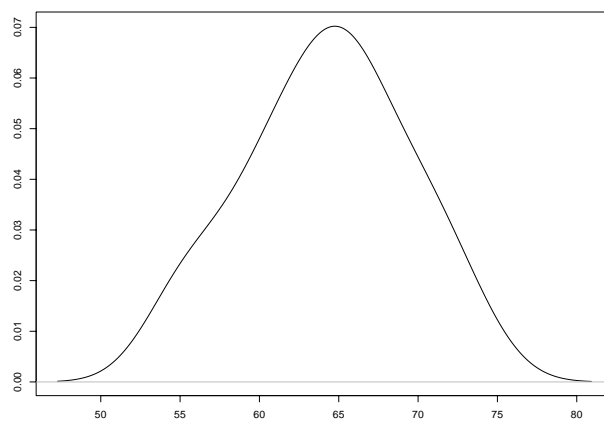


FIGURE 7.3 – Estimation de densité par la méthode du noyau

Chapitre 8

Tests d'adéquation basés sur la fonction de répartition empirique

Grâce aux méthodes de statistique non paramétrique, il est tout à fait possible d'extraire des informations pertinentes d'un échantillon sans connaître la loi de probabilité dont il est issu. Cependant, si c'est possible, il est quand même préférable d'adopter un modèle probabiliste. En effet, les estimations seront toujours plus précises dans un cadre paramétrique que dans un cadre non paramétrique. Par ailleurs, un grand nombre de procédures statistiques standard ne sont utilisables que si on fait des hypothèses particulières sur la loi de probabilité des observations (par exemple, les tests dans les modèles linéaires gaussiens).

Par conséquent, il est fondamental de disposer de méthodes permettant de déterminer s'il est vraisemblable de considérer que des observations proviennent d'un modèle probabiliste donné. Ces méthodes sont appelées les **tests d'adéquation**. On a vu en PMS deux types de méthodes : les graphes de probabilité, qui sont des tests d'adéquation graphiques, et les tests du χ^2 . Nous allons dans ce chapitre étudier des tests plus puissants, qui sont basés sur la fonction de répartition empirique.

8.1 Problématique des tests d'adéquation

Tester l'adéquation d'un échantillon (x_1, \dots, x_n) à une loi de probabilité donnée, c'est déterminer s'il est vraisemblable que x_1, \dots, x_n soient les réalisations de variables aléatoires X_1, \dots, X_n indépendantes et de cette loi.

On note F la fonction de répartition inconnue de l'échantillon, que l'on supposera pour simplifier continue. Dans le cas de lois discrètes, les procédures présentées ici nécessiteront des aménagements, pas toujours simples.

On distinguera deux cas, suivant que l'on veut tester l'adéquation de l'échantillon à une loi de probabilité entièrement spécifiée ou à une famille de lois de probabilité.

- *Cas 1* : Test d'adéquation à une loi entièrement spécifiée.

Test de $H_0 : "F = F_0"$ contre $H_1 : "F \neq F_0"$.

Par exemple, on se demande si les observations sont issues d'une loi normale de moyenne 10 et de variance 4.

- *Cas 2* : Test d'adéquation à une famille de lois de probabilité.

Test de H_0 : " $F \in \mathcal{F}$ " contre H_1 : " $F \notin \mathcal{F}$ ".

Le plus souvent, la famille \mathcal{F} est une famille paramétrée : $\mathcal{F} = \{F(\cdot; \theta); \theta \in \Theta\}$. C'est le cas quand on se demande simplement si les observations sont issues d'une loi normale, sans donner de valeur a priori aux paramètres. Si le modèle de loi normale est adopté, on pourra toujours estimer les paramètres ultérieurement.

En théorie, on devrait toujours appliquer un test d'adéquation avant d'utiliser n'importe quel modèle probabiliste sur des données. En pratique, on ne le fait pas toujours, ce qui entraîne parfois l'utilisation de modèles complètement erronés.

8.2 Rappels sur les graphes de probabilité

On a vu que la fonction de répartition empirique \mathbb{F}_n était un excellent estimateur de la fonction de répartition inconnue F . Si on teste l'hypothèse " $F = F_0$ ", il est naturel de tracer les graphes de \mathbb{F}_n et de F_0 , et de juger visuellement si les deux courbes sont proches (voir figure 8.1). Cependant, il est difficile de juger si les deux courbes sont "significativement proches", surtout si on dispose de peu de données. De plus, toutes les fonctions de répartition ont des formes voisines.

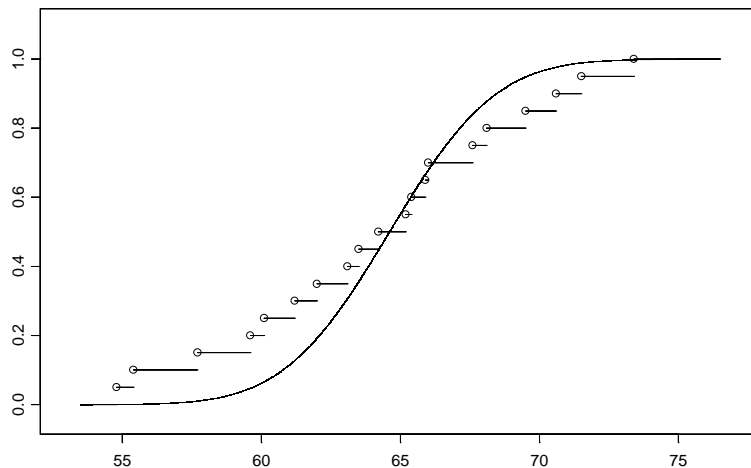


FIGURE 8.1 – Fonctions de répartition empirique et testée

De la même façon, on peut comparer visuellement une estimation de la densité (par histogramme ou par noyau) et la densité testée f_0 . Cela peut permettre d'écarter certaines hypothèses manifestement fausses. Par exemple, si l'estimation de densité n'est pas du tout en forme de cloche symétrique, il est peu probable que les observations proviennent d'une loi normale. Dans ce cas, il n'est pas forcément nécessaire d'effectuer un test d'adéquation pour confirmer cette hypothèse. Inversement, même si la forme de l'estimation de densité n'est pas très éloignée d'une cloche, rien ne prouve que la loi des

observations est normale. De toutes façons, il est toujours difficile d'évaluer visuellement la proximité de deux courbes.

L'idée des graphes de probabilité est de chercher, à partir de la fonction de répartition F , une relation linéaire caractéristique de la loi à tester. On trace alors un nuage de points qui, si la vraie fonction de répartition est F , devraient être approximativement alignés.

Le problème essentiel de cette procédure graphique est de déterminer à partir de quand on peut considérer que des points sont "suffisamment alignés". Une idée naturelle est de déterminer la droite des moindres carrés pour le nuage de points, et de considérer que l'adéquation est bonne si le coefficient de corrélation linéaire empirique correspondant dépasse une certaine valeur. Malheureusement, la loi de probabilité de ce coefficient de corrélation sous H_0 est trop complexe pour que l'on puisse construire un test d'adéquation statistique simple par ce moyen.

Les graphes de probabilité sont une première étape indispensable dans une étude statistique, car ils sont faciles à mettre en oeuvre et permettent de rejeter facilement de trop mauvais modèles. Il est cependant nécessaire de les compléter par des tests statistiques si l'on veut obtenir des résultats plus précis.

8.3 Cas d'une loi entièrement spécifiée

Quand on doit tester si " $F = F_0$ ", il est logique de ne pas rejeter cette hypothèse si \mathbb{F}_n et F_0 sont significativement proches, d'autant plus que l'on sait, d'après le théorème de Glivenko-Cantelli, que $D_n = \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F_0(x)|$ converge presque sûrement vers 0 sous H_0 .

Il s'agit donc de définir une distance, ou plutôt un écart, entre \mathbb{F}_n et F_0 , et de rejeter H_0 : " $F = F_0$ " si cet écart est "trop grand". Les mesures d'écart les plus usuelles sont :

- La **statistique de Kolmogorov-Smirnov (KS)** - Commande R : `ks.test` :

$$K_n = \sqrt{n}D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F_0(x)|.$$

- La **statistique de Cramer-von Mises (CM)** - Commande R : `cvm.test` :

$$W_n^2 = n \int_{-\infty}^{+\infty} [\mathbb{F}_n(x) - F_0(x)]^2 dF_0(x).$$

- La **statistique d'Anderson-Darling (AD)** - Commande R : `ad.test` :

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{[\mathbb{F}_n(x) - F_0(x)]^2}{F_0(x)(1 - F_0(x))} dF_0(x).$$

Un test de seuil α de H_0 : " $F = F_0$ " contre H_1 : " $F \neq F_0$ " aura donc une région critique de la forme $W = \{K_n > k_\alpha\}$, avec $\alpha = P_{H_0}(K_n > k_\alpha)$. Il faut donc connaître la loi des variables aléatoires K_n , W_n^2 et A_n^2 sous H_0 . Ces lois ne sont pas facilement accessibles pour n fini. En revanche, on a un résultat asymptotique.

Théorème 20 . Sous H_0 , K_n converge en loi vers la loi de Kolmogorov-Smirnov, de fonction de répartition : $\forall z \in \mathbb{R}^+$, $F_{KS}(z) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2 z^2}$.

Ce qui est remarquable dans ce théorème, c'est que la loi limite de K_n est la même, quelle que soit la loi de l'échantillon. C'est en cela que la procédure est non paramétrique et c'est pour cela que l'on peut construire un test.

Ainsi, dans la région critique définie plus haut, k_α est le quantile d'ordre $1 - \alpha$ de la loi de Kolmogorov-Smirnov.

L'inconvénient de ce résultat est qu'il n'est qu'asymptotique. En pratique, on ne peut utiliser ce test tel quel que pour $n > 80$. Pour $n \leq 80$, on peut utiliser les lois exactes de K_n , qui ont été tabulées pour tout n , mais c'est fastidieux. On préfère utiliser le résultat suivant.

Propriété 26 . Pour tout $n \geq 5$, la variable aléatoire $D_n \left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right)$ est approximativement de loi de Kolmogorov-Smirnov sous H_0 .

Par conséquent, on appelle **test de Kolmogorov-Smirnov**, le test, valable quelle que soit la taille de l'échantillon, ayant pour région critique :

$$W = \left\{ D_n \left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right) > F_{KS}^{-1}(1 - \alpha) \right\}$$

On montre que, sous H_0 , W_n^2 et A_n^2 convergent aussi en loi vers des lois qui ne dépendent pas de F . Mais cette fois, les fonctions de répartition des lois limites n'ont pas d'expressions simples, et on est obligés de se référer à des tables. Comme pour K_n , on dispose de résultats permettant d'appliquer les tests quelle que soit la loi de l'échantillon :

Propriété 27 . Pour tout $n \geq 5$, on a, sous H_0 :

- $\left(W_n^2 - \frac{0.4}{n} + \frac{0.6}{n^2} \right) \left(1 + \frac{1}{n} \right)$ est approximativement de loi de Cramer-von Mises.
- A_n^2 est approximativement de loi d'Anderson-Darling.

La table 8.1 donne quelques quantiles usuels des lois limites de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling.

Enfin, pour calculer facilement les statistiques de test, il est pratique d'utiliser le résultat suivant.

Propriété 28 . Pour $i \in \{1, \dots, n\}$, on pose $U_i = F_0(X_i)$. On a :

- $K_n = \sqrt{n} \max \left[\max \left\{ \frac{i}{n} - U_i^*, i = 1..n \right\}, \max \left\{ U_i^* - \frac{i-1}{n}, i = 1..n \right\} \right]$.

	15%	10%	5%	2.5%	1%
KS	1.138	1.224	1.358	1.480	1.628
CM	0.284	0.347	0.461	0.581	0.743
AD	1.610	1.933	2.492	3.070	3.857

TABLE 8.1 – Valeurs usuelles des quantiles des lois de KS, CM et AD dans le cas 1

$$\bullet W_n^2 = \sum_{i=1}^n \left(U_i^* - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}.$$

$$\bullet A_n^2 = -n + \frac{1}{n} \sum_{i=1}^n [(2i-1-2n) \ln(1-U_i^*) - (2i-1) \ln U_i^*].$$

Il est impossible de calculer la puissance de tels tests puisque l'hypothèse alternative $H_1 : "F \neq F_0"$ est beaucoup trop vaste. Des études intensives prenant en compte un grand nombre d'alternatives possibles ont montré que, de manière générale, le test d'Anderson-Darling était le plus puissant des trois et le test de Kolmogorov-Smirnov le moins puissant.

8.4 Cas d'une famille de lois

On teste $H_0 : "F \in \mathcal{F} = \{F(\cdot; \theta); \theta \in \Theta\}"$ contre $H_1 : "F \notin \mathcal{F}"$.

Puisque θ est un paramètre inconnu, une démarche naturelle est d'en déterminer un estimateur $\hat{\theta}(X_1, \dots, X_n)$ et de calculer les statistiques K_n , W_n^2 et A_n^2 en remplaçant $F_0(x)$ par $F(x; \hat{\theta}(X_1, \dots, X_n))$. On notera \hat{K}_n , \hat{W}_n^2 et \hat{A}_n^2 les statistiques correspondantes.

Les expressions de \hat{K}_n , \hat{W}_n^2 et \hat{A}_n^2 en fonction des U_i restent valables à condition de remplacer $U_i = F_0(X_i)$ par $\hat{U}_i = F(X_i; \hat{\theta}(X_1, \dots, X_n))$.

Malheureusement, le fait d'estimer θ entraîne que les lois limites sous H_0 de \hat{K}_n , \hat{W}_n^2 et \hat{A}_n^2 ne sont pas les mêmes que celles de K_n , W_n^2 et A_n^2 . Cela tient au fait que, sous H_0 , les U_i étaient indépendantes et de même loi uniforme sur $[0, 1]$, alors que les \hat{U}_i ne sont plus ni indépendantes, ni de loi uniforme.

Dans le cas général, les lois limites des statistiques de test dépendent de la loi testée F , de la procédure d'estimation utilisée (maximum de vraisemblance, moments, moindres carrés, ...), et de la vraie valeur de θ . Contrairement au cas d'une loi entièrement spécifiée, on ne peut donc pas obtenir de test d'adéquation applicable dans tous les cas de figure.

Pour faire un test d'adéquation, il faut au minimum que la loi limite des statistiques de test soit indépendante de θ , puisque cette valeur est inconnue.

Propriété 29 . Si θ est un paramètre de position, d'échelle ou de position-échelle, alors les lois de probabilité sous H_0 de \hat{K}_n , \hat{W}_n^2 et \hat{A}_n^2 ne dépendent pas de θ .

Rappelons que :

- m est un **paramètre de position** (ou de localisation) si et seulement si la loi de $X - m$ est indépendante de m ou bien si et seulement si la densité de X est de la forme $f(x; m) = g(x - m)$.
- σ est un **paramètre d'échelle** si et seulement si la loi de $\frac{X}{\sigma}$ est indépendante de σ ou bien si et seulement si la densité de X est de la forme $f(x; \sigma) = \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right)$.
- $\theta = (m, \sigma)$ est un **paramètre de position-échelle** si et seulement si la loi de $\frac{X - m}{\sigma}$ est indépendante de m et de σ ou bien si et seulement si la densité de X est de la forme $f(x; m, \sigma) = \frac{1}{\sigma} g\left(\frac{x - m}{\sigma}\right)$.

Exemples :

- loi normale : $f(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x - m)^2}{2\sigma^2}}$. (m, σ) est un paramètre de position-échelle.
- loi exponentielle : $f(x; \lambda) = \lambda e^{-\lambda x}$. $\frac{1}{\lambda}$ est un paramètre d'échelle.
- loi gamma : $f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}$. (α, λ) n'est pas un paramètre de position-échelle.

Par conséquent, les méthodes KS, CM et AD permettent de tester l'adéquation d'un échantillon à la loi normale et à la loi exponentielle, mais pas à la loi gamma.

Pour les lois dont le paramètre est de position-échelle, la loi limite des statistiques de test ne dépend pas de θ , mais elle dépend du type de loi testée et de la procédure d'estimation. Aucune des lois limites n'a d'expression explicite, donc il faut recourir à des tables. D'autre part, il existe encore des modifications des statistiques de test à effectuer pour pouvoir utiliser les tests même pour de petits échantillons.

Exemple 1 : la loi normale.

Les estimateurs de maximum de vraisemblance de la moyenne et la variance pour un échantillon de loi normale sont $\hat{m}_n = \bar{X}_n$ et $\hat{\sigma}_n^2 = S_n^2$. Donc $\hat{U}_i = F(X_i; \hat{m}_n, \hat{\sigma}_n^2) = \Phi\left(\frac{X_i - \bar{X}_n}{S_n}\right)$. Les modifications des statistiques sont :

- Statistique de Kolmogorov-Smirnov modifiée : $\hat{D}_n \left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right)$.
- Statistique de Cramer-von Mises modifiée : $\hat{W}_n^2 \left(1 + \frac{0.5}{n} \right)$.
- Statistique d'Anderson-Darling modifiée : $\hat{A}_n^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right)$.

Et les valeurs usuelles des quantiles sont données par la table 8.2.

	15%	10%	5%	2.5%	1%
KS	0.775	0.819	0.895	0.995	1.035
CM	0.091	0.104	0.126	0.148	0.179
AD	0.561	0.631	0.752	0.873	1.035

TABLE 8.2 – Valeurs usuelles des quantiles des lois de KS, CM et AD dans le cas 2 pour la loi normale avec estimation par maximum de vraisemblance

Exemple 2 : la loi exponentielle.

L'estimateur de maximum de vraisemblance du paramètre λ pour un échantillon exponentiel est $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$. Donc $\hat{U}_i = F(X_i; \hat{\lambda}_n) = 1 - e^{-X_i/\bar{X}_n}$.

Les modifications des statistiques sont :

- Statistique de Kolmogorov-Smirnov modifiée : $\left(\hat{D}_n - \frac{0.2}{n}\right) \left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$.
- Statistique de Cramer-von Mises modifiée : $\hat{W}_n^2 \left(1 + \frac{0.16}{n}\right)$.
- Statistique d'Anderson-Darling modifiée : $\hat{A}_n^2 \left(1 + \frac{0.6}{n}\right)$.

Et les valeurs usuelles des quantiles sont données par la table 8.3.

	15%	10%	5%	2.5%	1%
KS	0.926	0.995	1.094	1.184	1.298
CM	0.148	0.175	0.222	0.271	0.338
AD	0.916	1.062	1.321	1.591	1.959

TABLE 8.3 – Valeurs usuelles des quantiles des lois de KS, CM et AD dans le cas 2 pour la loi exponentielle avec estimation par maximum de vraisemblance

L'estimateur de maximum de vraisemblance est biaisé. Il faut néanmoins le conserver car la table 8.3 a été obtenue avec cet estimateur biaisé.

Il s'avère que les puissances des différents tests sont plus proches quand on estime les paramètres que pour une loi entièrement spécifiée. Cependant, Anderson-Darling est toujours le meilleur et Kolmogorov-Smirnov le moins bon.

Ces tests sont plus puissants que les tests du χ^2 car le regroupement en classes fait perdre de l'information sur les données.

Chapitre 9

Tests non paramétriques sur un échantillon

Comme précédemment, on suppose dans ce chapitre que les observations x_1, \dots, x_n sont des réalisations de variables aléatoires réelles X_1, \dots, X_n . Dans les chapitres précédents, on a supposé que les X_i étaient indépendantes et de même loi. Tout ce qui a été fait jusqu'ici n'a de sens que si cette hypothèse est vérifiée. Il est donc fondamental de déterminer si cette hypothèse est valide ou pas. Les tests qui permettent d'y parvenir sont appelés **tests d'échantillon**.

Si on a admis que les observations forment un échantillon, on peut utiliser les procédures d'estimation des moments, quantiles, fonction de répartition et densité de l'échantillon, vues précédemment. L'étape statistique suivante est d'effectuer des tests d'hypothèses sur ces quantités. Par exemple, on peut vouloir faire un test de " $E(X) \leq m$ " contre " $E(X) > m$ ". Dans ce chapitre, on se contentera d'étudier des tests portant sur la moyenne et la médiane de la loi de l'échantillon.

Dans les deux cas, on supposera que la loi est continue et on utilisera les statistiques de rang pour effectuer les tests. En effet, on a vu que, si la loi de l'échantillon est continue, alors la loi des statistiques de rang ne dépend pas de la loi de l'échantillon.

9.1 Tests d'échantillon

Le problème est de déterminer si les observations forment un échantillon. L'hypothèse nulle d'un test d'échantillon sera donc :

$$H_0 : "X_1, \dots, X_n \text{ sont indépendantes et de même loi (i.i.d.)"}.$$

Le choix d'un test d'échantillon dépend fortement des hypothèses alternatives que l'on choisit. L'hypothèse $H_1 = \bar{H}_0$: " X_1, \dots, X_n ne sont pas i.i.d." est trop vaste. Les alternatives les plus fréquemment retenues sont celles qui portent sur l'existence d'une *tendance* :

- H_1 : "Les X_i sont de plus en plus grandes"
- H_2 : "Les X_i sont de plus en plus petites"

Par exemple, si X_i est la cotation d'un titre au jour i , il est intéressant de déterminer si le titre évolue à la hausse (H_1), à la baisse (H_2), ou ni l'un ni l'autre (H_0).

Ou bien, si les X_i sont les durées de bon fonctionnement successives entre les pannes d'un système réparable, l'usure va faire en sorte que les pannes se produiront de plus en plus souvent, donc les X_i seront de plus en plus petits (H_2).

Il est nécessaire de définir ce que l'on entend par "des variables aléatoires de plus en plus grandes". Cela peut vouloir dire par exemple que la suite des $E(X_i)$ est croissante. On peut en fait définir plusieurs ordres de ce type, appelés **ordres stochastiques**. L'ordre le plus fréquemment retenu est le suivant :

Définition 26 . On dira que la suite de variables aléatoires $\{X_i\}_{i \geq 1}$ est **stochastiquement croissante** (resp. **décroissante**) si et seulement si les fonctions de répartition des X_i diminuent (resp. augmentent) au sens où :

$$\forall x \in \mathbb{R}, \quad i < j \Rightarrow F_{X_i}(x) \geq F_{X_j}(x) \text{ (resp. } \leq \text{)}$$

En effet, pour n'importe quel x , si X_i est "plus petit" que X_j , X_i a une plus forte chance que X_j d'être inférieure à x .

On se contentera ici d'étudier les hypothèses :

- H_1 : "Les X_i sont stochastiquement croissantes"
- H_2 : "Les X_i sont stochastiquement décroissantes"

sachant que d'autres alternatives sont possibles comme par exemple :

- "Les X_i sont stochastiquement périodiques"
- "Les X_i sont de même moyenne mais de variances croissantes"

Sous H_0 , les X_i sont i.i.d. donc leur ordre n'a aucune importance. Ce n'est évidemment pas le cas sous H_1 et H_2 . Il semble donc logique d'utiliser les statistiques d'ordre et de rang pour construire les tests.

Remarque. N'oublions pas que le résultat d'un test n'est probant que si on rejette H_0 . Donc on pourra éventuellement conclure qu'il n'est pas improbable que les X_i forment un échantillon, mais on ne pourra jamais accepter cette hypothèse.

9.1.1 Le test de Spearman

La première idée consiste à étudier le lien entre les rangs R_i des observations et leurs indices i . En effet, si les X_i sont strictement croissants, alors les observations sont directement ordonnées dans l'ordre croissant, donc $\forall i, R_i = i$. Inversement, si les X_i sont strictement décroissants, alors $\forall i, R_i = n - i + 1$.

D'où l'idée d'utiliser le coefficient de corrélation linéaire empirique entre les rangs et les indices, $R_{RI,n}$. Sous H_1 , $R_{RI,n}$ doit être proche de 1, sous H_2 il doit être proche de -1, et sous H_0 , il doit être proche de 0.

$$R_{RI,n} = \frac{c_{RI}}{s_{RSI}} = \frac{\frac{1}{n} \sum_{i=1}^n R_i i - \bar{R}_n \bar{i}_n}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n R_i^2 - \bar{R}_n^2\right) \left(\frac{1}{n} \sum_{i=1}^n i^2 - \bar{i}_n^2\right)}}$$

Par exemple, un test de H_0 contre H_1 de seuil α aura une région critique de la forme $W = \{R_{RI,n} > k_\alpha\}$. k_α est déterminé en écrivant que $P_{H_0}(R_{RI,n} > k_\alpha) = \alpha$. Il faut donc connaître la loi de $R_{RI,n}$ sous H_0 pour effectuer le test : k_α sera le quantile d'ordre $1 - \alpha$ de cette loi.

On a $\bar{i}_n = \frac{1}{n} \sum_{i=1}^n i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$ et $\bar{R}_n = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2}$, car, s'il n'y a pas d'ex-aequos (loi continue), alors pour toute fonction φ , $\sum_{i=1}^n \varphi(R_i) = \sum_{i=1}^n \varphi(i)$.

De même,

$$\begin{aligned} s_R^2 = s_I^2 &= \frac{1}{n} \sum_{i=1}^n i^2 - \bar{i}_n^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n+1}{2} \left[\frac{2n+1}{3} - \frac{n+1}{2} \right] = \frac{n+1}{12} [4n+2-3n-3] = \frac{(n+1)(n-1)}{12} \\ &= \frac{n^2-1}{12} \end{aligned}$$

$$\text{D'où } R_{RI,n} = \frac{\frac{1}{n} \sum_{i=1}^n R_i i - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}} = \frac{12}{n(n^2-1)} \sum_{i=1}^n R_i i - 3 \frac{n+1}{n-1}.$$

Sachant que la loi du vecteur des rangs $R = (R_1, \dots, R_n)$ sous H_0 est la loi uniforme sur l'ensemble des permutations des entiers de 1 à n , il est possible d'en déduire la loi de $R_{RI,n}$ sous H_0 . Cette loi est appelée **loi de Spearman**. D'où le test d'échantillon suivant :

Définition 27 . *Le test de Spearman est le test d'échantillon basé sur la statistique $R_{RI,n} = \frac{12}{n(n^2-1)} \sum_{i=1}^n R_i i - 3 \frac{n+1}{n-1}$. Plus précisément, on a :*

- Test de H_0 contre H_1 (test de croissance) : $W = \{R_{RI,n} > s_{n,\alpha}\}$,
- Test de H_0 contre H_2 (test de décroissance) : $W = \{R_{RI,n} < s_{n,1-\alpha}\}$,

où $s_{n,\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi de Spearman de paramètre n .

Il existe une table des quantiles de la loi de Spearman. Mais quand la taille de l'échantillon est suffisamment grande, on utilise les résultats suivants.

Propriété 30 .

- Sous H_0 , pour $n > 10$, $\sqrt{n-2} \frac{R_{RI,n}}{\sqrt{1-R_{RI,n}^2}}$ est approximativement de loi de Student $St(n-2)$.
- Sous H_0 , $\sqrt{n-1} R_{RI,n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

En pratique, pour $n < 10$, on utilise la table de la loi de Spearman. Pour $11 \leq n \leq 30$, on utilise l'approximation de Student, et pour $n > 30$, on utilise l'approximation normale.

9.1.2 Le test de Kendall

Si les X_i sont strictement croissants, alors $\forall(i, j), i < j \Rightarrow X_i < X_j$. Inversement, si les X_i sont strictement décroissants, alors $\forall(i, j), i < j \Rightarrow X_i > X_j$.

D'où l'idée de compter le nombre Q_n de couples (i, j) tels que $i < j$ et $X_i < X_j$:
 $Q_n = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbf{1}_{\{X_i < X_j\}}$. Le nombre total de couples (i, j) tels que $i < j$ est le nombre de façons de choisir 2 entiers distincts parmi n , c'est-à-dire $C_n^2 = \frac{n(n-1)}{2}$.

Donc, sous H_1 , Q_n doit être proche de $\frac{n(n-1)}{2}$, et sous H_2 , Q_n doit être proche de 0. Sous H_0 , $\forall(i, j), P(X_i < X_j) = \frac{1}{2}$. Donc Q_n doit être proche de $\frac{n(n-1)}{4}$.

Pour rendre la statistique de test plus facile à interpréter, on pose $\tau_n = \frac{4Q_n}{n(n-1)} - 1$. τ_n est appelée le **tau de Kendall**. Sous H_1 , τ_n doit être proche de 1, sous H_2 , τ_n doit être proche de -1 et sous H_0 , τ_n doit être proche de 0. Ainsi l'interprétation de τ_n est similaire à celle du coefficient de corrélation de Spearman. On peut déterminer la loi de τ_n sous H_0 , appelée **loi de Kendall**.

Définition 28 . *Le test de Kendall est le test d'échantillon basé sur la statistique*

$$\tau_n = \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbf{1}_{\{X_i < X_j\}} - 1. \text{ Plus précisément, on a :}$$

- Test de H_0 contre H_1 (test de croissance) : $W = \{\tau_n > k_{n,\alpha}\}$
- Test de H_0 contre H_2 (test de décroissance) : $W = \{\tau_n < k_{n,1-\alpha}\}$

où $k_{n,\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi de Kendall de paramètre n .

Propriété 31 . Sous H_0 , $\sqrt{\frac{9n(n-1)}{2(2n+5)}} \tau_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

En pratique, pour $n \leq 10$, on utilise une table de quantiles de la loi de Kendall, et pour $n > 10$, on utilise l'approximation normale.

Suivant les cas, le test de Kendall sera plus ou moins puissant que le test de Spearman. On peut montrer que $\rho(R_{RI,n}, \tau_n) = \frac{2(n+1)}{\sqrt{2n(2n+5)}}$, qui tend vers 1 quand n tend vers l'infini, ce qui signifie que, quand on a beaucoup de données, les deux tests sont équivalents.

9.2 Tests sur l'espérance et la médiane

Dans cette section, on suppose que les observations forment un échantillon, ce qui a pu être confirmé par les tests de la section précédente. On peut alors vouloir effectuer des tests d'hypothèses sur les diverses caractéristiques de la loi de l'échantillon.

Les tests les plus utilisés portent sur la valeur de l'espérance de l'échantillon. On a vu dans le chapitre 2 que le moyenne empirique \bar{X}_n est un excellent estimateur de $E(X)$. Il est donc logique de construire des tests sur l'espérance à partir de la moyenne empirique. Mais comme on ne connaît que la loi asymptotique de \bar{X}_n , seuls des tests asymptotiques seront possibles.

Au lieu de faire porter les tests sur l'espérance de la loi, il est aussi intéressant de les faire porter sur la médiane de cette loi. Il s'avère qu'il est plus facile de construire des tests sur la médiane que des tests sur l'espérance à partir des statistiques de rang. Par ailleurs, espérance et médiane sont égales dans le cas des lois symétriques.

9.2.1 Tests asymptotiques sur l'espérance

Les hypothèses des tests portant sur l'espérance de l'échantillon sont les suivantes :

$$H_0 : "E(X) = m" \qquad H_1 : "E(X) \neq m"$$

$$H_2 : "E(X) \geq m" \qquad H_3 : "E(X) \leq m"$$

Au chapitre 6, on a vu qu'un intervalle de confiance asymptotique de seuil α pour $E(X)$ est :

$$\left[\bar{X}_n - u_\alpha \frac{S_n}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{S_n}{\sqrt{n}} \right]$$

Par conséquent, pour tester $H_0 : "E(X) = m"$ contre $H_1 : "E(X) \neq m"$ au seuil α , il suffit de rejeter H_0 si et seulement si m n'est pas dans l'intervalle de confiance ci-dessus. On obtient comme région critique :

$$\begin{aligned} W &= \{m \notin [\bar{X}_n - u_\alpha \frac{S_n}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{S_n}{\sqrt{n}}]\} = \{m < \bar{X}_n - u_\alpha \frac{S_n}{\sqrt{n}} \text{ ou } m > \bar{X}_n + u_\alpha \frac{S_n}{\sqrt{n}}\} \\ &= \{\bar{X}_n - m < -u_\alpha \frac{S_n}{\sqrt{n}} \text{ ou } \bar{X}_n - m > +u_\alpha \frac{S_n}{\sqrt{n}}\} = \left\{ \left| \frac{\bar{X}_n - m}{S_n} \sqrt{n} \right| > u_\alpha \right\}. \end{aligned}$$

On peut vérifier que cette région critique convient. On a vu que :

$$\sqrt{n} \frac{\bar{X}_n - E(X)}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Donc, sous H_0 , $\sqrt{n} \frac{\bar{X}_n - m}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Alors $P_{H_0}((X_1, \dots, X_n) \in W) = P_{H_0}(|\frac{\bar{X}_n - m}{S_n} \sqrt{n}| > u_\alpha) \xrightarrow{n \rightarrow \infty} \alpha$. La probabilité de rejeter à tort H_0 est bien asymptotiquement égale à α .

Intuitivement, on rejette l'hypothèse " $E(X) = m$ " si \bar{X}_n est significativement éloigné de m , c'est-à-dire si $|\bar{X}_n - m|$ est "trop grand".

Supposons maintenant que l'on veuille tester $H_3 : "E(X) \leq m"$ contre $H_2 : "E(X) > m"$. L'idée naturelle est de rejeter " $E(X) \leq m$ " si \bar{X}_n est significativement grand, d'où une région critique de la forme $W = \{\bar{X}_n > k_\alpha\}$.

k_α est déterminé en écrivant que le seuil du test est :

$$\alpha = \sup_{H_3} P(\bar{X}_n > k_\alpha) = \sup_{E(X) \leq m} P\left(\sqrt{n} \frac{\bar{X}_n - E(X)}{S_n} > \sqrt{n} \frac{k_\alpha - E(X)}{S_n}\right)$$

Donc asymptotiquement, $\alpha = \sup_{E(X) \leq m} [1 - \Phi(\sqrt{n} \frac{k_\alpha - E(X)}{S_n})]$, où Φ est la fonction de répartition de la loi normale centrée-réduite.

Φ est une fonction croissante, donc $1 - \Phi(\sqrt{n} \frac{k_\alpha - E(X)}{S_n})$ est une fonction croissante de $E(X)$. Par conséquent, son maximum quand $E(X) \leq m$ est atteint pour $E(X) = m$.

On en déduit que $\alpha = 1 - \Phi(\sqrt{n} \frac{k_\alpha - m}{S_n})$, d'où $\sqrt{n} \frac{k_\alpha - m}{S_n} = \Phi^{-1}(1 - \alpha) = u_{2\alpha}$ et finalement $k_\alpha = m + \frac{u_{2\alpha} S_n}{\sqrt{n}}$.

Le test de H_3 contre H_2 aura donc comme région critique $W = \{\bar{X}_n > m + \frac{u_{2\alpha} S_n}{\sqrt{n}}\}$, ce qu'on peut aussi écrire sous la forme plus pratique $W = \{\frac{\bar{X}_n - m}{S_n} \sqrt{n} > u_{2\alpha}\}$.

Le test symétrique de H_2 contre H_3 s'établit de la même manière et on obtient au bout du compte la propriété suivante.

Propriété 32 . *Tests asymptotiques de seuil α sur l'espérance de l'échantillon, parfois appelés tests de Student :*

- Test de $H_3 : "E(X) \leq m"$ contre $H_2 : "E(X) > m"$: $W = \{\frac{\bar{X}_n - m}{S_n} \sqrt{n} > u_{2\alpha}\}$.
- Test de $H_2 : "E(X) \geq m"$ contre $H_3 : "E(X) < m"$: $W = \{\frac{\bar{X}_n - m}{S_n} \sqrt{n} < -u_{2\alpha}\}$.
- Test de $H_0 : "E(X) = m"$ contre $H_1 : "E(X) \neq m"$: $W = \{|\frac{\bar{X}_n - m}{S_n} \sqrt{n}| > u_\alpha\}$.

L'inconvénient de ces tests est qu'ils sont asymptotiques, donc ils ne sont pas valables pour des échantillons de petite taille.

En pratique, on peut disposer de tests valables même pour de petits échantillons, à condition de supposer en plus que la loi de l'échantillon est symétrique. Le principe est d'effectuer des tests portant sur la médiane, puisque, quand la loi est symétrique, la médiane $q_{1/2}$ est égale à l'espérance $E(X)$.

9.2.2 Tests sur la médiane

Dans cette section, on va s'intéresser à des tests non paramétriques portant sur la médiane $q_{1/2}$. Quand la loi est symétrique, ces tests pourront être considérés comme des tests sur l'espérance de l'échantillon.

Les hypothèses des tests portant sur la médiane de l'échantillon sont les suivantes :

$$H_0 : "q_{1/2} = m", \quad H_1 : "q_{1/2} \neq m",$$

$$H_2 : "q_{1/2} \geq m", \quad H_3 : "q_{1/2} \leq m".$$

Sous H_0 , il y a une chance sur deux qu'une observation soit inférieure à m et une chance sur deux qu'elle soit supérieure à m .

9.2.2.1. Le test du signe

Le principe de ce test est de considérer le nombre d'observations supérieures à m , appelé **statistique du signe** : $S_n^+ = \sum_{i=1}^n \mathbf{1}_{\{X_i > m\}}$.

Sous H_0 , puisque la probabilité qu'une observation soit supérieure à m est $\frac{1}{2}$, S_n^+ doit être proche de $\frac{n}{2}$. Sous H_2 , S_n^+ doit être "grand" et sous H_3 , S_n^+ doit être "petit". Sous H_1 , S_n^+ doit être "éloigné de $\frac{n}{2}$ ".

Propriété 33 . S_n^+ est de loi binomiale $\mathcal{B}(n, 1 - F(m))$. Sous H_0 , S_n^+ est de loi $\mathcal{B}(n, \frac{1}{2})$.

Démonstration. S_n^+ est une somme de n variables aléatoires indépendantes et de même loi de Bernoulli $\mathcal{B}(P(X_i > m))$. Donc S_n^+ est de loi binomiale $\mathcal{B}(n, P(X_i > m)) = \mathcal{B}(n, 1 - F(m))$. Sous H_0 , $F(m) = F(q_{1/2}) = \frac{1}{2}$, donc S_n^+ est de loi $\mathcal{B}(n, \frac{1}{2})$. ■

Pour tester H_3 contre H_2 , on prendra logiquement une région critique de la forme $W = \{S_n^+ > k_\alpha\}$. k_α est déterminé en écrivant :

$$\begin{aligned} \alpha &= \sup_{H_3} P(S_n^+ > k_\alpha) = \sup_{q_{1/2} \leq m} \left[1 - F_{\mathcal{B}(n, 1 - F(m))}(k_\alpha) \right] = 1 - F_{\mathcal{B}(n, \frac{1}{2})}(k_\alpha) \\ &= \sum_{i=[k_\alpha]+1}^n C_n^i \left[\frac{1}{2}\right]^i \left[1 - \frac{1}{2}\right]^{n-i} = \frac{1}{2^n} \sum_{i=[k_\alpha]+1}^n C_n^i. \end{aligned}$$

Le problème est que la fonction de répartition de la loi binomiale n'est pas inversible. Donc il n'est pas forcément possible, pour un α donné, de trouver k_α vérifiant l'équation ci-dessus. Cela signifie qu'on ne peut effectuer le test que pour quelques valeurs de α bien déterminées.

Ce problème se résoud quand n est grand en utilisant l'approximation de la loi binomiale $\mathcal{B}(n, \frac{1}{2})$ par la loi normale $\mathcal{N}(\frac{n}{2}, \frac{n}{4})$. On a en fait, sous H_0 :

$$\frac{S_n^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2S_n^+ - n}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Alors, en reprenant le calcul précédent, on a, asymptotiquement :

$$\alpha = P_{H_0}(S_n^+ > k_\alpha) = P_{H_0}\left(\frac{2S_n^+ - n}{\sqrt{n}} > \frac{2k_\alpha - n}{\sqrt{n}}\right) = 1 - \Phi\left(\frac{2k_\alpha - n}{\sqrt{n}}\right)$$

$$\text{d'où } \frac{2k_\alpha - n}{\sqrt{n}} = \Phi^{-1}(1 - \alpha) = u_{2\alpha} \text{ et } k_\alpha = \frac{n + \sqrt{n}u_{2\alpha}}{2}.$$

Le test de H_3 contre H_2 aura donc comme région critique $W = \{S_n^+ > \frac{n + \sqrt{n}u_{2\alpha}}{2}\}$, ce qu'on peut aussi écrire sous la forme plus pratique $W = \{\frac{2S_n^+ - n}{\sqrt{n}} > u_{2\alpha}\}$.

Finalement, on obtient :

Définition 29 . *Le test du signe est le test sur la médiane basé sur la statistique $S_n^+ = \sum_{i=1}^n \mathbb{1}_{\{X_i > m\}}$. Plus précisément, on a, asymptotiquement :*

- Test de H_3 : “ $q_{1/2} \leq m$ ” contre H_2 : “ $q_{1/2} > m$ ” : $W = \left\{ \frac{2S_n^+ - n}{\sqrt{n}} > u_{2\alpha} \right\}$.
- Test de H_2 : “ $q_{1/2} \geq m$ ” contre H_3 : “ $q_{1/2} < m$ ” : $W = \left\{ \frac{2S_n^+ - n}{\sqrt{n}} < -u_{2\alpha} \right\}$.
- Test de H_0 : “ $q_{1/2} = m$ ” contre H_1 : “ $q_{1/2} \neq m$ ” : $W = \left\{ \left| \frac{2S_n^+ - n}{\sqrt{n}} \right| > u_\alpha \right\}$.

En pratique, on admet que l'approximation normale est valide dès que $n > 10$.

9.2.2.2. Le test des rangs signés de Wilcoxon

Dans cette section, on suppose que $m = 0$. Il est possible de généraliser à m quelconque.

Définition 30 *Le vecteur des rangs signés associé à l'échantillon (X_1, \dots, X_n) est le vecteur $R^+ = (R_1^+, \dots, R_n^+)$ défini par :*

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \quad R_i^+ &= 1 + \sum_{j=1}^n \mathbb{1}_{\{|X_j| < |X_i|\}} \\ &= \text{rang de } |X_i| \text{ dans la suite } |X_1|^*, \dots, |X_n|^* \end{aligned}$$

Les rangs signés R_i^+ sont aux $|X_i|$ ce que les rangs R_i sont aux X_i .

Exemple : $n = 5$.

x_i	2.3	-3.5	1.7	0.5	-1.4
x_i^*	-3.5	-1.4	0.5	1.7	2.3
r_i	5	1	4	3	2
$ x_i $	2.3	3.5	1.7	0.5	1.4
$ x_i ^*$	0.5	1.4	1.7	2.3	3.5
r_i^+	4	5	3	1	2

Définition 31 . *Le test des rangs signés de Wilcoxon est le test de nullité de la médiane basé sur la somme des rangs signés des observations strictement positives, appelé statistique des rangs signés de Wilcoxon : $W_n^+ = \sum_{i=1}^n R_i^+ \mathbf{1}_{\{X_i > 0\}}$.*

L'idée est que, sous H_2 : " $q_{1/2} > 0$ ", il y aura plus de X_i positifs que de X_i négatifs, et que les valeurs absolues des X_i positifs seront dans l'ensemble plus grandes que les valeurs absolues des X_i négatifs. Donc, sous H_2 , W_n^+ sera "grand". Réciproquement, sous H_3 , W_n^+ sera "petit".

Propriété 34 .

- W_n^+ est à valeurs dans $\{0, \dots, \frac{n(n+1)}{2}\}$.
- $W_n^+ = \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{X_i + X_j > 0\}}$.
- $W_n^+ = \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{X_i + X_j > 0\}} + S_n^+$.
- Sous H_0 , $E(W_n^+) = \frac{n(n+1)}{4}$ et $Var(W_n^+) = \frac{n(n+1)(2n+1)}{24}$.
- Sous H_0 , $\frac{W_n^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

En pratique, pour $n \leq 15$, on utilise une table de la loi de W_n^+ sous H_0 . Pour $n > 15$, on utilise l'approximation gaussienne.

On montre que le test des rangs signés est plus puissant que le test du signe. De plus, il est utilisable sans problèmes même pour les très petits échantillons. Donc il est conseillé d'utiliser le test des rangs signés plutôt que le test du signe.

Chapitre 10

Tests non paramétriques sur plusieurs échantillons

Dans ce chapitre, on suppose que l'on dispose de plusieurs échantillons, que l'on souhaite comparer. Par exemple, il peut s'agir des résultats de l'application de plusieurs traitements d'une même maladie à plusieurs groupes de malades. Il est important de déterminer si les traitements ont des efficacités comparables ou si l'un s'avère plus efficace que les autres. Mathématiquement, cela revient à comparer les lois de probabilité de chaque échantillon. Dans un contexte paramétrique, on dispose pour cela de méthodes bien connues comme l'**analyse de variance**. On s'intéressera dans ce chapitre à un point de vue non paramétrique sur ce problème.

La situation de base est la comparaison de deux échantillons indépendants, notés X_1, \dots, X_{n_1} et Y_1, \dots, Y_{n_2} . Les X_i sont supposés indépendants et de même loi, de fonction de répartition F inconnue, et les Y_j sont supposés indépendants et de même loi, de fonction de répartition G inconnue. Tester l'hypothèse que les deux échantillons sont issus de la même loi de probabilité, c'est tester :

$$H_0 : "F = G" \text{ contre } H_1 : "F \neq G".$$

Mais on peut aussi s'intéresser aux hypothèses :

- $H_2 : "F > G"$, qui signifie que les X_i sont stochastiquement inférieurs aux Y_j .
- $H_3 : "F < G"$, qui signifie que les X_i sont stochastiquement supérieurs aux Y_j .

C'est ce genre d'hypothèses que l'on utilisera si on cherche à déterminer si un traitement est plus efficace qu'un autre.

Pour pouvoir utiliser les propriétés des statistiques de rang, on se contentera d'étudier le cas où les lois des échantillons sont continues.

10.1 Test de Kolmogorov-Smirnov

Si les deux échantillons proviennent de la même loi, ils ont la même fonction de répartition, donc leurs fonctions de répartition empiriques \mathbb{F}_{n_1} et \mathbb{G}_{n_2} doivent être très proches. Le test de Kolmogorov-Smirnov consiste à rejeter $H_0 : "F = G"$ si et seulement si $D_{n_1, n_2} = \sup_{x \in \mathbb{R}} |\mathbb{F}_{n_1}(x) - \mathbb{G}_{n_2}(x)|$ est "trop grand".

On montre alors que, sous H_0 , la variable aléatoire $K_{n_1, n_2} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1, n_2}$ a une loi de probabilité qui ne dépend pas de F et converge en loi vers la loi de Kolmogorov-Smirnov. Donc le test de comparaison d'échantillon résultant est similaire au test d'adéquation de Kolmogorov-Smirnov.

Si $n_1 = n_2 = m$, la loi de $D_{m, m}$ sous H_0 est très simple et a une expression explicite même pour m fini :

$$\forall k \in \mathbb{N}, P\left(D_{m, m} \geq \frac{k}{m}\right) = 2 \sum_{i=1}^{\lfloor m/k \rfloor} (-1)^{j+1} \frac{(m!)^2}{(m - jk)!(m + jk)!}$$

10.2 Tests de rang

Pour un seul échantillon, on a utilisé le fait que le vecteur des rangs a une loi de probabilité indépendante de la loi de l'échantillon (loi uniforme sur l'ensemble Σ_n des permutations des entiers de 1 à n). Dans le cas de deux échantillons, on a une propriété équivalente.

Théorème 21 . Soient S et R les vecteurs des rangs respectifs de (X_1, \dots, X_{n_1}) et (Y_1, \dots, Y_{n_2}) lorsque ces $n = n_1 + n_2$ variables aléatoires sont ordonnées toutes ensemble. Alors, sous H_0 : " $F = G$ ", on a :

- (S, R) est de loi uniforme sur Σ_n .
- $\forall s = (s_1, \dots, s_{n_1}), \{s_1, \dots, s_{n_1}\} \subset \{1, \dots, n\}, P(S = s) = \frac{n_2!}{n!}$.
- $\forall r = (r_1, \dots, r_{n_2}), \{r_1, \dots, r_{n_2}\} \subset \{1, \dots, n\}, P(R = r) = \frac{n_1!}{n!}$.

Démonstration. Si $F = G$, $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ est un échantillon de taille n de la loi de fonction de répartition F et (S, R) est son vecteur des rangs. (S, R) est donc de loi uniforme sur Σ_n , ce qui signifie que $\forall (s, r) \in \Sigma_n, P(S = s \cap R = r) = \frac{1}{n!}$.

Alors, $P(S = s) = \sum_r P(S = s \cap R = r) = \frac{1}{n!} \times$ nombre de vecteurs r possibles. Puisque les rangs des X_i sont déterminés par s , il reste à choisir les rangs des $n_2 Y_j$. Il y a $n_2!$ possibilités pour cela. On obtient donc $P(S = s) = \frac{n_2!}{n!}$ et symétriquement $P(R = r) = \frac{n_1!}{n!}$. ■

Puisque la loi de (S, R) ne dépend pas de F sous H_0 , on pourra construire des tests de H_0 contre H_1 à partir de statistiques ne dépendant que de (S, R) . De tels tests et statistiques s'appellent **tests de rang** et **statistiques de rang**.

10.2.1 Le test de la médiane

L'idée de ce test est que, si les X_i sont stochastiquement inférieurs aux Y_j , alors les rangs des X_i dans l'échantillon complet (les S_i) seront dans l'ensemble inférieurs aux rangs

des Y_j (les R_j). En particulier, les Y_j seront dans l'ensemble supérieurs à la médiane de l'échantillon complet, ou bien les rangs des Y_j seront dans l'ensemble supérieurs au rang médian de l'échantillon complet, qui vaut $\frac{n+1}{2}$. D'où :

Définition 32 . La statistique de la médiane M_{n_1, n_2} est le nombre d'observations du deuxième échantillon strictement supérieures à la médiane de l'échantillon complet :

$$M_{n_1, n_2} = \sum_{j=1}^{n_2} \mathbf{1}_{\{R_j > \frac{n+1}{2}\}}$$

Sous H_2 , M_{n_1, n_2} doit être "grand", sous H_3 , M_{n_1, n_2} doit être "petit", et sous H_0 , M_{n_1, n_2} doit être "ni grand, ni petit".

Propriété 35 . Sous H_0 , M_{n_1, n_2} est de loi hypergéométrique :

- $\mathcal{H}\left(n, n_2, \frac{n}{2}\right)$ si n est pair.
- $\mathcal{H}\left(n, n_2, \frac{n-1}{2}\right)$ si n est impair.

Démonstration. Rappelons qu'une variable aléatoire K est de loi hypergéométrique $\mathcal{H}(N, m, n)$ si et seulement si on est dans la situation suivante : on a N objets dont m ont une certaine caractéristique ; on tire n objets sans remise parmi ces N ; K représente alors le nombre d'objets possédant la caractéristique en question parmi les n tirés.

Ici, on a n observations parmi lesquelles n_2 sont des Y_j et M_{n_1, n_2} représente le nombre de Y_j parmi les observations strictement supérieures à la médiane. Celles-ci sont au nombre de $\frac{n}{2}$ si n est pair et $\frac{n-1}{2}$ si n est impair. ■

Connaissant l'espérance et la variance de la loi hypergéométrique, on peut en déduire celles de la statistique de la médiane sous H_0 . Un argument de type théorème central-limite permet d'en déduire la loi asymptotique de M_{n_1, n_2} sous H_0 .

Propriété 36 . Sous H_0 , $\frac{2M_{n_1, n_2} - n_2}{\sqrt{n_1 n_2}} \sqrt{n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Finalement, on a :

Définition 33 : Le test de la médiane est le test de comparaison de deux échantillons basé sur la statistique de la médiane $M_{n_1, n_2} = \sum_{j=1}^{n_2} \mathbf{1}_{\{R_j > \frac{n+1}{2}\}}$.

Les régions critiques des différents tests possibles sont établis à l'aide des quantiles des lois hypergéométrique ou normale. En pratique, on considère que l'approximation normale est valide si $n_1 \geq 8$ et $n_2 \geq 8$.

10.2.2 Le test de Wilcoxon-Mann-Whitney

Le principe de ce test est similaire à celui du test de la médiane : si les Y_j sont dans l'ensemble supérieurs aux X_i , alors les rangs R_j des Y_j seront dans l'ensemble supérieurs aux rangs S_i des X_i dans l'échantillon complet.

Définition 34 : La statistique de Wilcoxon W_{n_1, n_2} est la somme des rangs des observations du deuxième échantillon dans l'échantillon complet :

$$W_{n_1, n_2} = \sum_{j=1}^{n_2} R_j.$$

Dans le cas extrême où les Y_j sont tous inférieurs aux X_i , $W_{n_1, n_2} = \sum_{j=1}^{n_2} j = \frac{n_2(n_2 + 1)}{2}$.

Inversement, si les Y_j sont tous supérieurs aux X_i , $W_{n_1, n_2} = \sum_{j=n_1+1}^{n_1+n_2} j = \frac{n_2(n_2 + 1)}{2} + n_1 n_2$.

Sous H_0 , le mélange des deux échantillons est homogène, donc W_{n_1, n_2} devrait être de l'ordre de $\frac{n_2(n_2 + 1)}{2} + \frac{n_1 n_2}{2} = \frac{n_2(n + 1)}{2}$.

Par conséquent, sous H_2 , W_{n_1, n_2} doit être "grand", sous H_3 , W_{n_1, n_2} doit être "petit", et sous H_0 , W_{n_1, n_2} doit être "proche" de $\frac{n_2(n + 1)}{2}$.

Définition 35 . Le test de Wilcoxon est le test de comparaison de deux échantillons basé sur la statistique de Wilcoxon.

Propriété 37 . Sous H_0 , $\frac{2W_{n_1, n_2} - (n + 1)n_2}{\sqrt{(n + 1)n_1 n_2}} \sqrt{3} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Quand n est petit, on utilise des tables de la loi de la statistique de Wilcoxon sous H_0 . En pratique, on considère que l'approximation normale est valide si $n_1 \geq 8$ et $n_2 \geq 8$.

On peut aborder le problème différemment, en remarquant que, sous H_0 , comme les X_i et les Y_j sont indépendants et de même loi, on a $\forall (i, j), P(X_i \leq Y_j) = \frac{1}{2}$.

Définition 36 . La statistique de Mann-Whitney est le nombre de couples (i, j) tels que $X_i \leq Y_j$:

$$U_{n_1, n_2} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{1}_{\{X_i \leq Y_j\}}.$$

Sous H_0 , U_{n_1, n_2} doit être de l'ordre de la moitié des couples (X_i, Y_j) possibles, à savoir $\frac{n_1 n_2}{2}$. Sous H_2 , U_{n_1, n_2} doit être "grand", et sous H_3 , U_{n_1, n_2} doit être "petit".

Définition 37 . *Le test de Mann-Whitney est le test de comparaison de deux échantillons basé sur la statistique de Mann-Whitney.*

Propriété 38 . Sous H_0 , $\frac{2U_{n_1, n_2} - n_1 n_2}{\sqrt{(n+1)n_1 n_2}} \sqrt{3} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

La condition de validité de l'approximation normale est la même que pour les tests précédents : $n_1 \geq 8$ et $n_2 \geq 8$.

Propriété 39 . $U_{n_1, n_2} = W_{n_1, n_2} - \frac{n_2(n_2 + 1)}{2}$.

Cette propriété a pour conséquence que les tests de Mann-Whitney et Wilcoxon sont en fait équivalents, au sens où ils donneront exactement la même réponse. C'est pourquoi on peut utiliser indifféremment l'un ou l'autre, en leur donnant le nom de **test de Wilcoxon-Mann-Whitney**.

On montre que ce test est globalement plus puissant que le test de Kolmogorov-Smirnov et le test de la médiane.

10.2.3 Le test de Kruskal-Wallis

Après avoir comparé deux échantillons, on souhaite maintenant comparer k échantillons, avec $k > 2$. Pour i allant de 1 à k , le $i^{\text{ème}}$ échantillon est noté $X_1^i, \dots, X_{n_i}^i$. Le nombre total d'observations est $n = \sum_{i=1}^k n_i$.

Des hypothèses comparables à " $F > G$ " ne sont plus possibles quand on a plus de deux échantillons. Aussi on se contentera de tester :

$$H_0 : \text{"Les } k \text{ échantillons sont de même loi"} \text{ contre } H_1 = \overline{H_0}.$$

Pour cela, on ordonne l'ensemble des n observations et on note :

- R_j^i = rang de X_j^i dans l'échantillon global.
- $R^i = \sum_{j=1}^{n_i} R_j^i$ = somme des rangs des observations du $i^{\text{ème}}$ échantillon dans l'échantillon global.

Définition 38 . *Le test de Kruskal-Wallis est le test de comparaison de k échantillons basé sur la statistique de Kruskal-Wallis :*

$$K_n = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R^i{}^2}{n_i} - 3(n+1)$$

Propriété 40 . Sous H_0 , $K_n \xrightarrow{\mathcal{L}} \chi_{k-1}^2$.

En pratique, l'approximation par la loi du χ^2 est valide dès qu'il y a au moins 5 observations par échantillon.

Le test de Kruskal-Wallis consiste à rejeter l'hypothèse d'égalité des k lois si K_n est "trop grand". Si l'approximation du χ^2 est valide, la région critique du test sera $W = \{K_n > z_{k-1,\alpha}\}$, où $z_{k-1,\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi χ_{k-1}^2 .

Chapitre 11

Annexe A : Rappels de probabilités pour la statistique

Cette annexe rappelle quelques résultats de base du calcul des probabilités utiles pour la statistique. Les notions sont présentées sans aucune démonstration. Les détails sont à aller chercher dans le cours de Probabilités Appliquées de première année.

11.1 Variables aléatoires réelles

11.1.1 Loi de probabilité d'une variable aléatoire

Mathématiquement, une variable aléatoire est définie comme une application mesurable. On se contentera ici de la conception intuitive suivante.

Une **variable aléatoire** est une grandeur dépendant du résultat d'une expérience aléatoire, c'est-à-dire non prévisible à l'avance avec certitude. Par exemple, on peut dire que la durée de vie d'une ampoule électrique ou le résultat du lancer d'un dé sont des variables aléatoires. Pour une expérience donnée, ces grandeurs prendront une valeur donnée, appelée réalisation de la variable aléatoire. Si on recommence l'expérience, on obtiendra une réalisation différente de la même variable aléatoire.

On ne s'intéresse ici qu'aux **variables aléatoires réelles**, c'est-à-dire à valeurs dans \mathbb{R} ou un sous-ensemble de \mathbb{R} . On note traditionnellement une variable aléatoire par une lettre majuscule (X) et sa réalisation par une lettre minuscule (x).

Le calcul des probabilités va permettre de calculer des grandeurs comme la durée de vie moyenne d'une ampoule ou la probabilité d'obtenir un 6 en lançant le dé. Ces grandeurs sont déterminées par la **loi de probabilité** de ces variables aléatoires.

Il y a plusieurs moyens de caractériser la loi de probabilité d'une variable aléatoire. Le plus simple est la fonction de répartition.

On appelle **fonction de répartition** de la variable aléatoire X la fonction

$$\begin{aligned} F_X : \mathbb{R} &\rightarrow [0, 1] \\ x &\mapsto F_X(x) = P(X \leq x) \end{aligned}$$

F_X est croissante, continue à droite, telle que $\lim_{x \rightarrow -\infty} F_X(x) = 0$ et $\lim_{x \rightarrow +\infty} F_X(x) = 1$. Elle permet de calculer la probabilité que X appartienne à n'importe quel intervalle de \mathbb{R} :

$$\forall (a, b) \in \mathbb{R}^2, a < b, P(a < X \leq b) = F_X(b) - F_X(a)$$

Les variables aléatoires peuvent être classées selon le type d'ensemble dans lequel elles prennent leurs valeurs. Dans la pratique, on ne s'intéressera qu'à deux catégories : les variables aléatoires discrètes et les variables aléatoires continues (ou à densité).

11.1.2 Variables aléatoires discrètes et continues

Une **variable aléatoire** X est dite **discrète (v.a.d.)** si et seulement si elle est à valeurs dans un ensemble E fini ou dénombrable. On peut noter $E = \{x_1, x_2, \dots\}$.

Exemples :

- Face obtenue lors du lancer d'un dé : $E = \{1, 2, 3, 4, 5, 6\}$.
- Nombre de bugs dans un programme : $E = \mathbb{N}$.

La loi de probabilité d'une v.a.d. X est entièrement déterminée par les probabilités élémentaires $P(X = x_i), \forall x_i \in E$.

La fonction de répartition de X est alors $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$.

Une **variable aléatoire** X est dite **continue (v.a.c.)** si et seulement si sa fonction de répartition F_X est continue et presque partout dérivable. Sa dérivée f_X est alors appelée densité de probabilité de X , ou plus simplement **densité** de X . Une v.a.c. est forcément à valeurs dans un ensemble non dénombrable.

Exemples :

- Appel de la fonction Random d'une calculatrice : $E = [0, 1]$.
- Durée de bon fonctionnement d'un système : $E = \mathbb{R}^+$.

On a alors $\forall (a, b) \in \mathbb{R}^2, a < b, P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$.

Plus généralement, $\forall B \subset \mathbb{R}, P(X \in B) = \int_B f_X(x) dx$. Donc la densité détermine entièrement la loi de probabilité de X .

f_X est une fonction positive telle que $\int_{-\infty}^{+\infty} f_X(x) dx = P(X \in \mathbb{R}) = 1$.

Connaissant la loi de X , on est souvent amenés à déterminer celle de $Y = \varphi(X)$. Quand X est discrète, il suffit d'écrire $P(Y = y) = P(\varphi(X) = y)$. Si φ est inversible, on obtient $P(Y = y) = P(X = \varphi^{-1}(y))$. Quand X est continue, on commence par déterminer la fonction de répartition de Y en écrivant $F_Y(y) = P(Y \leq y) = P(\varphi(X) \leq y)$, puis on en déduit sa densité par dérivation. Quand φ est inversible, on obtient la **formule du changement de variable** :

$$f_Y(y) = \frac{1}{|\varphi'(\varphi^{-1}(y))|} f_X(\varphi^{-1}(y))$$

Remarque : Il existe des lois de probabilité de variables aléatoires réelles qui ne sont ni discrètes ni continues. Par exemple, si X est la durée de bon fonctionnement d'un système qui a une probabilité non nulle p d'être en panne à l'instant initial, on a $\lim_{x \rightarrow 0^-} F_X(x) = 0$

(une durée ne peut pas être négative) et $F_X(0) = P(X \leq 0) = P(X = 0) = p$. Par conséquent F_X n'est pas continue en 0. La loi de X ne peut donc pas être continue, et elle n'est pas non plus discrète puisqu'elle est à valeurs dans \mathbb{R}^+ . Ce type de variable aléatoire ne sera pas étudié dans ce cours.

11.1.3 Moments et quantiles d'une variable aléatoire réelle

Si X est une variable aléatoire discrète, son **espérance mathématique** est définie par :

$$E(X) = \sum_{x_i \in E} x_i P(X = x_i)$$

Si X est une variable aléatoire continue, son espérance mathématique est définie par :

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

Concrètement, $E(X)$ est ce qu'on s'attend à trouver comme moyenne des résultats obtenus si on répète l'expérience un grand nombre de fois. Par exemple, si on lance une pièce de monnaie 10 fois, on s'attend à trouver en moyenne 5 piles.

Plus généralement, on peut s'intéresser à l'espérance mathématique d'une fonction de X :

- Si X est une v.a.d., $E[\varphi(X)] = \sum_{x_i \in E} \varphi(x_i) P(X = x_i)$.
- Si X est une v.a.c., $E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f_X(x) dx$.

Ce résultat permet de calculer l'espérance de $\varphi(X)$ sans avoir à déterminer entièrement sa loi.

Deux espérances de ce type sont particulièrement utiles :

- Si X est une v.a.d., sa **fonction génératrice** est définie par $G_X(z) = E[z^X] = \sum_{x_i \in E} z^{x_i} P(X = x_i)$.
- Si X est une v.a.c., sa **fonction caractéristique** est définie par $\phi_X(t) = E[e^{itX}] = \int_{-\infty}^{+\infty} e^{itx} f_X(x) dx$.

Au même titre que la fonction de répartition et la densité, les fonctions génératrices et caractéristiques définissent entièrement les lois de probabilité concernées.

Soit k un entier naturel quelconque. Le **moment d'ordre k** de X est $E[X^k]$ et le **moment centré d'ordre k** est $E[(X - E(X))^k]$.

De tous les moments, le plus important est le moment centré d'ordre 2, appelé aussi **variance**. La variance de X est $Var(X) = E[(X - E(X))^2]$, qui se calcule plus facilement sous la forme $Var(X) = E(X^2) - [E(X)]^2$.

L'**écart-type** de X est $\sigma(X) = \sqrt{Var(X)}$.

La variance et l'écart-type sont des indicateurs de la dispersion de X : plus la variance de X est petite, plus les réalisations de X seront concentrées autour de son espérance.

Le **coefficient de variation** de X est $CV(X) = \frac{\sigma(X)}{E(X)}$. C'est également un indicateur de dispersion, dont l'avantage est d'être sans dimension. Il permet de comparer les dispersions de variables aléatoires d'ordres de grandeur différents ou exprimées dans des unités différentes. En pratique, on considère que, quand $CV(X)$ est inférieur à 15%, l'espérance peut être considérée comme un bon résumé de la loi.

Soit $p \in]0, 1[$. Le **quantile d'ordre p** (ou **p -quantile**) de la loi de X est tout réel q_p vérifiant $P(X < q_p) \leq p \leq P(X \leq q_p)$.

- Si F est continue et strictement croissante (donc inversible), on a simplement $P(X < q_p) = P(X \leq q_p) = F_X(q_p) = p$, d'où $q_p = F_X^{-1}(p)$.
- Si F_X est constante égale à p sur un intervalle $[a, b]$, n'importe quel réel de $[a, b]$ est un quantile d'ordre p . En général, on choisit de prendre le milieu de l'intervalle : $q_p = \frac{a+b}{2}$.
- Si F_X est discontinue en q et telle que $\lim_{x \rightarrow q^-} F_X(x) < p \leq F_X(q)$, alors $q_p = q$.

Les tables fournies donnent les quantiles les plus usuels des lois normale, du chi-deux, de Student et de Fisher-Snedecor.

11.2 Vecteurs aléatoires réels

On ne s'intéressera ici qu'aux vecteurs aléatoires (X_1, \dots, X_n) constitués de n variables aléatoires réelles toutes discrètes ou toutes continues.

11.2.1 Loi de probabilité d'un vecteur aléatoire

La loi d'un vecteur aléatoire (X_1, \dots, X_n) est déterminée par sa fonction de répartition :

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

Si les X_i sont discrètes, cette loi est aussi déterminée par les probabilités élémentaires $P(X_1 = x_1, \dots, X_n = x_n)$.

Si les X_i sont continues, la densité de (X_1, \dots, X_n) est définie, si elle existe, par :

$$f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_{(X_1, \dots, X_n)}(x_1, \dots, x_n)$$

On a alors $\forall B \subset \mathbb{R}^n, P((X_1, \dots, X_n) \in B) = \int \dots \int_B f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) dx_1 \dots dx_n$.

Les variables aléatoires X_1, \dots, X_n sont (mutuellement) **indépendantes** si et seulement si :

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \prod_{i=1}^n P(X_i \leq x_i)$$

Pour des variables discrètes cela donne $P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$.

Et pour des variables continues, $f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$.

Concrètement, l'indépendance signifie que la valeur prise par l'une des variables n'a aucune influence sur la valeur prise par les autres.

11.2.2 Espérance et matrice de covariance d'un vecteur aléatoire

L'**espérance mathématique** d'un vecteur aléatoire est le vecteur des espérances mathématiques de ses composantes : $E[(X_1, \dots, X_n)] = (E[X_1], \dots, E[X_n])$.

L'équivalent de la variance en dimension n est la **matrice de covariance** du vecteur (X_1, \dots, X_n) , notée $K_{(X_1, \dots, X_n)}$ ou K , dont les coefficients sont donnés par

$$k_{ij} = Cov(X_i, X_j), \forall (i, j) \in \{1, \dots, n\}^2$$

$Cov(X_i, X_j)$ est la covariance des variables aléatoires X_i et X_j et est définie par :

$$Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] = E(X_i X_j) - E(X_i)E(X_j)$$

Pour $i = j$, $Cov(X_i, X_i) = E(X_i^2) - [E(X_i)]^2 = Var(X_i)$.

Pour $i \neq j$, la covariance de X_i et X_j traduit le degré de corrélation entre ces deux variables. En particulier, si X_i et X_j sont indépendantes, $Cov(X_i, X_j) = 0$ (mais la réciproque est fautive). Par conséquent, si X_1, \dots, X_n sont indépendantes, leur matrice de covariance K est diagonale.

Le **coefficient de corrélation linéaire** entre X_i et X_j est $\rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}$.

On montre que :

- $\rho(X_i, X_j) \in [-1, +1]$.
- $\rho(X_i, X_j) = +1 \Leftrightarrow X_i = aX_j + b$, avec $a > 0$ et $b \in \mathbb{R}$.
- $\rho(X_i, X_j) = -1 \Leftrightarrow X_i = -aX_j + b$, avec $a > 0$ et $b \in \mathbb{R}$.
- si $\rho(X_i, X_j) > 0$, X_i et X_j sont corrélées positivement, ce qui signifie qu'elles varient dans le même sens. Par exemple, X_i et X_j peuvent être la taille et le poids d'individus pris au hasard.
- si $\rho(X_i, X_j) < 0$, X_i et X_j sont corrélées négativement, ce qui signifie qu'elles varient en sens contraire. Par exemple, X_i et X_j peuvent être l'âge et la résistance d'un matériau.
- si $\rho(X_i, X_j) = 0$, il n'y a pas de corrélation linéaire entre X_i et X_j . Cela ne signifie pas que X_i et X_j sont indépendantes. Il peut éventuellement y avoir une corrélation non linéaire.

L'espérance mathématique est linéaire : si X et Y sont des variables aléatoires et a, b et c des réels, alors $E(aX + bY + c) = aE(X) + bE(Y) + c$.

En revanche, la variance n'est pas linéaire : si X et Y sont des variables aléatoires et a , b et c des réels, alors $Var(aX + bY + c) = a^2Var(X) + 2abCov(X, Y) + b^2Var(Y)$.

Si X et Y sont indépendantes, $Cov(X_i, X_j) = 0$, donc $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y)$. En particulier, la variance de la somme de variables aléatoires indépendantes est égale à la somme des variances de ces variables. Mais ce résultat est faux si les variables ne sont pas indépendantes.

11.3 Convergences et applications

Deux des résultats les plus importants des probabilités sont le théorème central-limite et la loi des grands nombres. Ces résultats nécessitent d'utiliser la notion de convergence d'une suite de variables aléatoires.

Une suite de variables aléatoires $\{X_n\}_{n \geq 1}$ **converge en loi** vers la loi de probabilité de fonction de répartition F si et seulement si $\lim_{n \rightarrow \infty} F_{X_n}(x) = F(x)$ en tout point x où F est continue. Cela signifie que, quand n est grand, la loi de probabilité de X_n est approximativement la loi de fonction de répartition F .

Théorème Central-Limite : Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X)$ et d'écart-type $\sigma(X) = \sqrt{Var(X)}$ finis. Pour tout $n \geq 1$, on pose :

$$Z_n = \frac{\sum_{i=1}^n X_i - nE(X)}{\sqrt{nVar(X)}} = \sqrt{n} \frac{\bar{X}_n - E(X)}{\sigma(X)}$$

Alors la suite $\{Z_n\}_{n \geq 1}$ converge en loi vers la loi normale centrée-réduite, ce qui s'écrit :

$$\sqrt{n} \frac{\bar{X}_n - E(X)}{\sigma(X)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Concrètement, cela signifie que la loi de toute variable aléatoire égale à la somme d'un nombre "suffisamment grand" de variables aléatoires indépendantes et de même loi est approximativement une loi normale. Plus précisément, pour n grand, $\sum_{i=1}^n X_i$ est approximativement de loi $\mathcal{N}(nE(X), nVar(X))$. Ce qui est remarquable, c'est que ce résultat est vrai quelle que soit la loi des X_i .

De très nombreux phénomènes naturels sont la résultante d'un grand nombre de phénomènes élémentaires identiques, indépendants et additifs ce qui justifie l'importance (et le nom) de la loi normale.

La plus forte des convergences de suites de variables aléatoires est la convergence presque sûre. Ce concept nécessite d'avoir défini une variable aléatoire comme une application mesurable d'un espace probabilisé dans un autre. Une suite de variables aléatoires $\{X_n\}_{n \geq 1}$ **converge presque sûrement** vers la variable aléatoire X si et seulement si $P\left(\left\{\omega; \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1$.

Une suite de variables aléatoires $\{X_n\}_{n \geq 1}$ **converge en probabilité** vers la variable aléatoire X si et seulement si $\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$.

On montre que la convergence presque sûre entraîne la convergence en probabilité, qui elle-même entraîne la convergence en loi.

Loi forte des grands nombres : Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X)$. Soit $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Alors la suite $\{\bar{X}_n\}_{n \geq 1}$ converge presque sûrement vers $E(X)$, ce qui s'écrit :

$$\bar{X}_n \xrightarrow{ps} E(X)$$

Concrètement, cela signifie que quand on fait un très grand nombre d'expériences identiques et indépendantes, la moyenne des réalisations de la variable aléatoire à laquelle on s'intéresse tend vers l'espérance de sa loi. Ce résultat permet de justifier l'idée naturelle d'estimer une espérance par une moyenne et une probabilité par une proportion.

En fait, la convergence la plus utile en statistique est la convergence en moyenne quadratique ou dans L^2 . L^2 est l'ensemble des variables aléatoires réelles X telles que $E(X^2) < \infty$. Une suite de variables aléatoires $\{X_n\}_{n \geq 1}$ de L^2 **converge en moyenne quadratique** vers la variable aléatoire X si et seulement si $\lim_{n \rightarrow \infty} E(|X_n - X|^2) = 0$.

On montre que la convergence en moyenne quadratique entraîne la convergence en probabilité, qui elle-même entraîne la convergence en loi. Mais il n'y a pas de lien entre la convergence en moyenne quadratique et la convergence presque sûre.

11.4 Quelques résultats sur quelques lois de probabilité usuelles

Les tables de lois de probabilité fournies donnent notamment, pour les lois les plus usuelles, les probabilités élémentaires ou la densité, l'espérance, la variance, et la fonction génératrice ou la fonction caractéristique. On présente dans cette section quelques propriétés supplémentaires de quelques unes de ces lois.

11.4.1 Loi binomiale

Une variable aléatoire K est de loi binomiale $\mathcal{B}(n, p)$ si et seulement si elle est à valeurs dans $\{0, 1, \dots, n\}$ et $P(K = k) = C_n^k p^k (1 - p)^{n-k}$.

Le nombre de fois où, en n expériences identiques et indépendantes, un évènement de probabilité p s'est produit, est une variable aléatoire de loi $\mathcal{B}(n, p)$.

La loi de Bernoulli $\mathcal{B}(p)$ est la loi $\mathcal{B}(1, p)$.

Si X_1, \dots, X_n sont indépendantes et de même loi $\mathcal{B}(m, p)$, alors $\sum_{i=1}^n X_i$ est de loi $\mathcal{B}(nm, p)$. En particulier, la somme de n v.a. indépendantes et de même loi $\mathcal{B}(p)$ est de loi $\mathcal{B}(n, p)$.

11.4.2 Loi géométrique

Une variable aléatoire K est de loi géométrique $\mathcal{G}(p)$ si et seulement si elle est à valeurs dans \mathbb{N}^* et $P(K = k) = p(1 - p)^{k-1}$.

Dans une suite d'expériences identiques et indépendantes, le nombre d'expériences nécessaires pour que se produise pour la première fois un évènement de probabilité p , est une variable aléatoire de loi $\mathcal{G}(p)$.

Si X_1, \dots, X_n sont indépendantes et de même loi $\mathcal{G}(p)$, alors $\sum_{i=1}^n X_i$ est de loi binomiale négative $\mathcal{BN}(n, p)$.

11.4.3 Loi de Poisson

Une variable aléatoire K est de loi de Poisson $\mathcal{P}(\lambda)$ si et seulement si elle est à valeurs dans \mathbb{N} et $P(K = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.

Pour $n \geq 50$ et $p \leq 0.1$, la loi binomiale $\mathcal{B}(n, p)$ peut être approchée par la loi de Poisson $\mathcal{P}(np)$. On dit que la loi de Poisson est la loi des évènements rares : loi du nombre de fois où un évènement de probabilité très faible se produit au cours d'un très grand nombre d'expériences identiques et indépendantes.

Si X_1, \dots, X_n sont indépendantes et de même loi $\mathcal{P}(\lambda)$, alors $\sum_{i=1}^n X_i$ est de loi $\mathcal{P}(n\lambda)$.

11.4.4 Loi exponentielle

Une variable aléatoire X est de loi exponentielle $exp(\lambda)$ si et seulement si elle est à valeurs dans \mathbb{R}^+ et $f_X(x) = \lambda e^{-\lambda x}$.

La loi exponentielle est dite sans mémoire : $\forall (t, x) \in \mathbb{R}^{+2}, P(X > t + x | X > t) = P(X > x)$.

Si X_1, \dots, X_n sont indépendantes et de même loi $exp(\lambda)$, alors $\sum_{i=1}^n X_i$ est de loi gamma $G(n, \lambda)$.

Si X_1, \dots, X_n sont indépendantes et de même loi $exp(\lambda)$, et représentent les durées entre occurrences successives d'un même évènement, alors le nombre d'évènements survenus sur une période de longueur t est une variable aléatoire de loi de Poisson $\mathcal{P}(\lambda t)$.

11.4.5 Loi gamma et loi du chi-2

Une variable aléatoire X est de loi gamma $G(a, \lambda)$ si et seulement si elle est à valeurs dans \mathbb{R}^+ et $f_X(x) = \frac{\lambda^a}{\Gamma(a)} e^{-\lambda x} x^{a-1}$. Les propriétés de la fonction gamma sont rappelées sur les tables.

La loi $G(1, \lambda)$ est la loi $exp(\lambda)$.

La loi $G\left(\frac{n}{2}, \frac{1}{2}\right)$ est appelée loi du chi-2 à n degrés de liberté, notée χ_n^2 .

Si X est de loi $G(a, \lambda)$ et α est un réel strictement positif, alors αX est de loi $G\left(a, \frac{\lambda}{\alpha}\right)$.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $G(\alpha, \lambda)$ et $G(\beta, \lambda)$, alors $X + Y$ est de loi $G(\alpha + \beta, \lambda)$. En particulier, si X et Y sont indépendantes et de lois respectives χ_n^2 et χ_m^2 , alors $X + Y$ est de loi χ_{n+m}^2 .

11.4.6 Loi normale

Une variable aléatoire X est de loi normale $\mathcal{N}(m, \sigma^2)$ si et seulement si elle est à valeurs dans \mathbb{R} et $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$.

Si X est de loi $\mathcal{N}(m, \sigma^2)$, alors $aX + b$ est de loi $\mathcal{N}(am + b, a^2\sigma^2)$. En particulier, $\frac{X - m}{\sigma}$ est de loi $\mathcal{N}(0, 1)$.

$$P(X \in [m - \sigma, m + \sigma]) = 68.3\% \quad P(X \in [m - 2\sigma, m + 2\sigma]) = 95.4\%.$$

$$P(X \in [m - 3\sigma, m + 3\sigma]) = 99.7\%.$$

Si X est de loi $\mathcal{N}(0, 1)$, alors X^2 est de loi χ_1^2 .

Si (X_1, X_2) est un vecteur gaussien tel que X_1 est de loi $\mathcal{N}(m_1, \sigma_1^2)$ et X_2 est de loi $\mathcal{N}(m_2, \sigma_2^2)$, alors $aX_1 + bX_2$ est de loi $\mathcal{N}(am_1 + bm_2, a^2\sigma_1^2 + 2ab\text{Cov}(X_1, X_2) + b^2\sigma_2^2)$.

Théorème de Fisher. Si X_1, \dots, X_n sont indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors, en posant $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, on a :

- $\sum_{i=1}^n X_i$ est de loi $\mathcal{N}(nm, n\sigma^2)$.
- \bar{X}_n est de loi $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.
- $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - m)^2$ est de loi χ_n^2 .
- $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{nS_n^2}{\sigma^2}$ est de loi χ_{n-1}^2 .
- \bar{X}_n et S_n^2 sont indépendantes.
- $\sqrt{n-1} \frac{\bar{X}_n - m}{S_n}$ est de loi de Student $St(n-1)$.

11.4.7 Lois de Student et de Fisher-Snedecor

Soit U une variable aléatoire de loi $\mathcal{N}(0, 1)$ et X une variable aléatoire de loi χ_n^2 . Si U et X sont indépendantes, alors $\sqrt{n} \frac{U}{\sqrt{X}}$ est de loi de Student à n degrés de liberté $St(n)$.

Soit X une variable aléatoire de loi χ_n^2 et Y une variable aléatoire de loi χ_m^2 . Si X et Y sont indépendantes, alors $\frac{mX}{nY}$ est de loi de Fisher-Snedecor $F(n, m)$.

Ces deux définitions entraînent que si T est de loi $St(n)$, alors T^2 est de loi $F(1, n)$.

Les lois de Student et de Fisher-Snedecor sont toujours utilisées par l'intermédiaire de tables ou à l'aide d'un logiciel de statistique. Il n'est donc pas nécessaire de donner l'expression de leur densité.

Chapitre 12

Annexe B : Lois de probabilité usuelles

12.1 Caractéristiques des lois usuelles

12.1.1 Variables aléatoires réelles discrètes

Dans le tableau ci-dessous, on suppose $n \in \mathbb{N}^*$, $p \in]0, 1[$ et $\lambda \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Probabilités	$E(X)$	$Var(X)$	Fonction caractéristique $\varphi_X(t) = E(e^{itX})$
Bernouilli $\mathcal{B}(p)$	$P(X = 0) = 1 - p$ $P(X = 1) = p$	p	$p(1 - p)$	$1 - p + pe^{it}$
Binomiale $\mathcal{B}(n, p)$	$P(X = k) = C_n^k p^k (1 - p)^{n-k} \mathbb{1}_{\{0, \dots, n\}}(k)$	np	$np(1 - p)$	$(1 - p + pe^{it})^n$
Binomiale négative $\mathcal{BN}(n, p)$	$P(X = k) = C_{k-1}^{n-1} p^n (1 - p)^{k-n} \mathbb{1}_{\{n, \dots\}}(k)$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	$\left(\frac{pe^{it}}{1-(1-p)e^{it}}\right)^n$
Poisson $\mathcal{P}(\lambda)$	$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \mathbb{1}_{\mathbb{N}}(k)$	λ	λ	$e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$P(X = k) = p(1 - p)^{k-1} \mathbb{1}_{\mathbb{N}^*}(k)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$
Hypergéométrique $\mathcal{H}(N, m, n)$ $(m, n) \in \{1, \dots, N\}^2$	$P(X = k) = \frac{C_m^k C_{N-m}^{n-k}}{C_N^n} \mathbb{1}_{\{0, \dots, \min(m, n)\}}(k)$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	

12.1.2 Variables aléatoires réelles continues

La fonction Gamma est définie pour $a > 0$ par $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$.

$$\text{On a : } \forall n \in \mathbb{N}^*, \quad \Gamma(n) = (n-1)!, \quad \Gamma(1) = 1, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},$$

$$\forall a \in]1, +\infty[, \quad \Gamma(a) = (a-1)\Gamma(a-1).$$

Dans le tableau ci dessous, $[a, b] \subset \mathbb{R}$, $m \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $\lambda \in \mathbb{R}_+^*$, $\alpha \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$

Loi et Symbole $X \rightsquigarrow$	Densité	Espérance	Var (X)	Fonction caractéristique $\varphi_X(t) = E(e^{itX})$
Loi Uniforme $\mathcal{U}[a, b]$	$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
Loi Normale $\mathcal{N}(m, \sigma^2)$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \mathbb{1}_{\mathbb{R}}(x)$	m	σ^2	$e^{itm - \frac{\sigma^2 t^2}{2}}$
Loi Exponentielle $\exp(\lambda) = G(1, \lambda)$	$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$(1 - \frac{it}{\lambda})^{-1}$
Loi Gamma $G(\alpha, \lambda)$	$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \mathbb{1}_{\mathbb{R}_+}(x)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$(1 - \frac{it}{\lambda})^{-\alpha}$
Loi du Chi-deux $\chi_n^2 = G(\frac{n}{2}, \frac{1}{2})$	$f_X(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \mathbb{1}_{\mathbb{R}_+}(x)$	n	$2n$	$(1 - 2it)^{-\frac{n}{2}}$
Première loi de Laplace	$f_X(x) = \frac{1}{2} e^{- x } \mathbb{1}_{\mathbb{R}}(x)$	0	2	$\frac{1}{1+t^2}$

La fonction Beta est définie pour $a > 0$ et $b > 0$ par

$$\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

Dans le tableau suivant, on suppose $a \in \mathbb{R}_+^*$, $b \in \mathbb{R}_+^*$ et $\eta \in \mathbb{R}_+^*$, $\beta \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Densité	$E(X)$	$Var(X)$
Loi Beta de 1 ^{ère} espèce $\beta_1(a, b)$	$f_X(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Loi Beta de 2 ^{ème} espèce $\beta_2(a, b)$	$f_X(x) = \frac{1}{\beta(a,b)} \frac{x^{a-1}}{(1+x)^{a+b}} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{a}{b-1}$ si $b > 1$	$\frac{a(a+b-1)}{(b-1)^2(b-2)}$ si $b > 2$
Loi de Weibull $\mathcal{W}(\eta, \beta)$	$f_X(x) = \frac{\beta}{\eta^\beta} x^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\eta\Gamma\left(1 + \frac{1}{\beta}\right)$	$\eta^2 \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma\left(1 + \frac{1}{\beta}\right)^2 \right]$

12.1.3 Vecteurs aléatoires dans \mathbb{N}^d et dans \mathbb{R}^d

Dans le tableau suivant, on a :

$$n \in \mathbb{N}^*, p = (p_1, p_2, \dots, p_d) \in]0, 1[^d, \sum_{i=1}^d p_i = 1 \text{ et } k = (k_1, k_2, \dots, k_d) \in \mathbb{N}^d, \sum_{i=1}^d k_i = n.$$

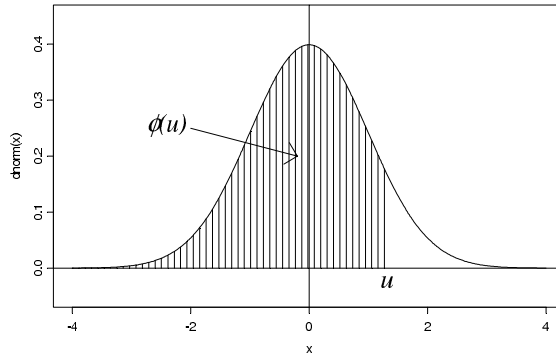
$$m \in \mathbb{R}^d \text{ et } \Sigma \in M_{d,d}.$$

Loi et Symbole $X \rightsquigarrow$	Probabilités ou Densité	$E(X)$	Matrice de covariance	Fonction Caractéristique
Loi Multinomiale $\mathcal{M}_d(n, p)$	$P(X = k) = \frac{n!}{k_1! \dots k_d!} p_1^{k_1} p_2^{k_2} \dots p_d^{k_d} \mathbb{1}_{\mathbb{N}^d}(k)$	np	$c_{i,i} = np_i(1 - p_i)$ $c_{i,j} = -np_i p_j, i \neq j$	$\left[\sum_{i=1}^d p_i z_i \right]^n$
Loi normale $\mathcal{N}_d(m, \Sigma)$	$f_X(x) = \frac{1}{\sqrt{\det \Sigma} (\sqrt{2\pi})^d} e^{-\frac{1}{2}(x-m)\Sigma^{-1}(x-m)}$	m	Σ	$e^{i^t m t - \frac{1}{2} t^t \Sigma t}$

12.2 Tables de lois

12.2.1 Table 1 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0, 1)$, la table donne la valeur de $\phi(u) = P(U \leq u)$. En R, la commande correspondante est `pnorm(u)`.



u	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

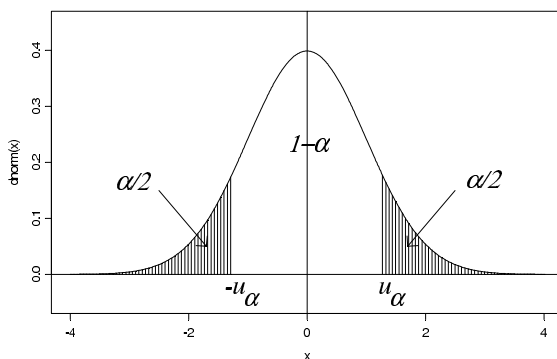
Lecture de la table : $\phi(1.25) = \phi(1.2 + 0.05) = 0.8944$.

Grandes valeurs de u

u	3.0	3.5	4.0	4.5
$\phi(u)$	0.9987	0.99977	0.999968	0.999997

12.2.2 Table 2 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0, 1)$ et α un réel de $[0, 1]$, la table donne la valeur de $u_\alpha = \phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ telle que $P(|U| > u_\alpha) = \alpha$. En R, la commande correspondante est `qnorm(1-alpha/2)`.



α	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	$+\infty$	2.5758	2.3263	2.1701	2.0537	1.96	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

Lecture de la table : $u_{0.25} = u_{0.2+0.05} = 1.1503$.

Petites valeurs de α

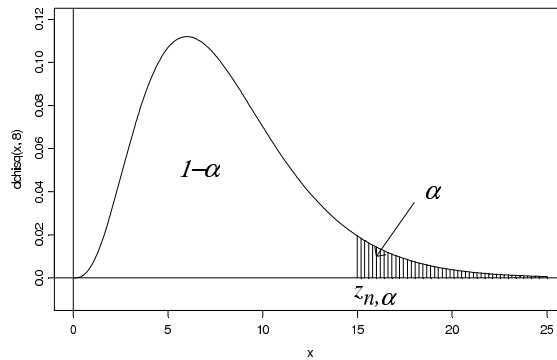
α	0.002	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_α	3.0902	3.2905	3.8906	4.4171	4.8916	5.3267	5.7307	6.1094

Pour $p < \frac{1}{2}$, $\phi^{-1}(p) = -u_{2p}$.

Pour $p \geq \frac{1}{2}$, $\phi^{-1}(p) = u_{2(1-p)}$.

12.2.3 Table de la loi du χ^2

X étant une variable aléatoire de loi du χ^2 à n degrés de libertés et α un réel de $[0, 1]$, la table donne la valeur de $z_{n,\alpha} = F_{\chi_n^2}^{-1}(1 - \alpha)$ telle que $P(X > z_{n,\alpha}) = \alpha$. En R, la commande correspondante est `qchisq(1-alpha, n)`.



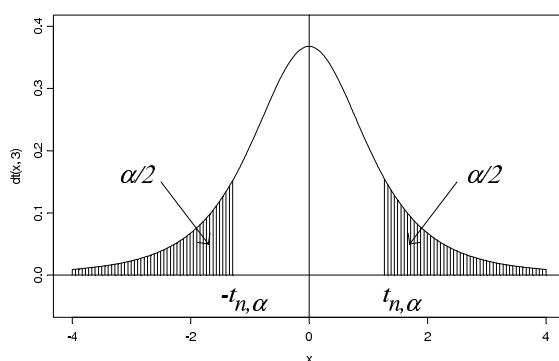
$n \backslash \alpha$	0.995	0.990	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00004	0.0002	0.001	0.004	0.02	0.06	0.15	0.45	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.6	13.82
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70

Pour $n > 30$, on admet que $z_{n,\alpha} \approx \frac{1}{2} (u_{2\alpha} + \sqrt{2n-1})^2$ si $\alpha < \frac{1}{2}$

et $z_{n,\alpha} \approx \frac{1}{2} (\sqrt{2n-1} - u_{2(1-\alpha)})^2$ si $\alpha \geq \frac{1}{2}$.

12.2.4 Table de la loi de Student

X étant une variable aléatoire de loi $St(n)$ et α un réel de $[0, 1]$, la table donne la valeur de $t_{n,\alpha} = F_{St(n)}^{-1}\left(1 - \frac{\alpha}{2}\right)$ telle que $P(|X| > t_{n,\alpha}) = \alpha$. En R, la commande correspondante est `qt(1-alpha/2,n)`. Pour $n = +\infty$, $t_{+\infty,\alpha} = u_\alpha$.



$n \ \alpha$	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$+\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

12.2.5 Tables de la loi de Fisher-Snedecor

X étant une variable aléatoire de loi $F(\nu_1, \nu_2)$, les tables donnent les valeurs de $f_{\nu_1, \nu_2, \alpha} = F_{F(\nu_1, \nu_2)}^{-1}(1 - \alpha)$ telles que $P(X > f_{\nu_1, \nu_2, \alpha}) = \alpha$ pour $\alpha = 5\%$ et $\alpha = 1\%$.

En R, la commande correspondante est `qf(1-alpha, nu1, nu2)`. $f_{\nu_2, \nu_1, \alpha} = \frac{1}{f_{\nu_1, \nu_2, 1-\alpha}}$.

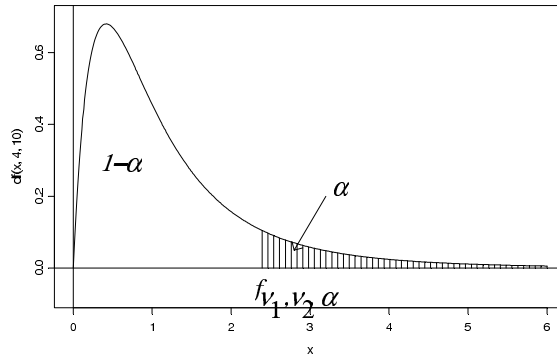


Table 1 : $\alpha = 5\%$.

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	241.9	243.9	246.5	248	249	251.1	252.2	253	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.40	19.41	19.43	19.45	19.45	19.47	19.48	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.69	8.66	8.64	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.84	5.80	5.77	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.60	4.56	4.53	4.46	4.43	4.41	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.92	3.87	3.84	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.49	3.44	3.41	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.20	3.15	3.12	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.99	2.94	2.90	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.83	2.77	2.74	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.70	2.65	2.61	2.53	2.49	2.46	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.60	2.54	2.51	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.51	2.46	2.42	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.44	2.39	2.35	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.38	2.33	2.29	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.33	2.28	2.24	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.29	2.23	2.19	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.25	2.19	2.15	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.21	2.16	2.11	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.18	2.12	2.08	1.99	1.95	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.16	2.10	2.05	1.96	1.92	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.13	2.07	2.03	1.94	1.89	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.11	2.05	2.01	1.91	1.86	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.09	2.03	1.98	1.89	1.84	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.07	2.01	1.96	1.87	1.82	1.78	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.99	1.93	1.89	1.79	1.74	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.90	1.84	1.79	1.69	1.64	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.03	1.95	1.85	1.78	1.74	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.82	1.75	1.70	1.59	1.53	1.48	1.39
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.77	1.70	1.65	1.54	1.48	1.43	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.75	1.68	1.63	1.52	1.45	1.39	1.28
$+\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.64	1.57	1.52	1.39	1.32	1.24	1.00

Table 2 : $\alpha = 1\%$.

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	4052	4999	5403	5625	5764	5859	5928	5981	6056	6106	6170	6209	6235	6287	6313	6334	6366
2	98.5	99.0	99.17	99.25	99.3	99.33	99.36	99.37	99.4	99.42	99.44	99.45	99.46	99.47	99.48	99.49	99.5
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.23	27.05	26.83	26.69	26.60	26.41	26.32	26.24	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.55	14.37	14.15	14.02	13.93	13.75	13.65	13.58	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.05	9.89	9.68	9.55	9.47	9.29	9.20	9.13	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.52	7.40	7.31	7.14	7.06	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.28	6.16	6.07	5.91	5.82	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.48	5.36	5.28	5.12	5.03	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.92	4.81	4.73	4.57	4.48	4.41	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.52	4.41	4.33	4.17	4.08	4.01	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.21	4.10	4.02	3.86	3.78	3.71	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	3.97	3.86	3.78	3.62	3.54	3.47	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.78	3.66	3.59	3.43	3.34	3.27	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.94	3.80	3.62	3.51	3.43	3.27	3.18	3.11	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.80	3.67	3.49	3.37	3.29	3.13	3.05	2.98	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.69	3.55	3.37	3.26	3.18	3.02	2.93	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.59	3.46	3.27	3.16	3.08	2.92	2.83	2.76	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.51	3.37	3.19	3.08	3.00	2.84	2.75	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.43	3.30	3.12	3.00	2.92	2.76	2.67	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.37	3.23	3.05	2.94	2.86	2.69	2.61	2.54	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.31	3.17	2.99	2.88	2.80	2.64	2.55	2.48	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.26	3.12	2.94	2.83	2.75	2.58	2.50	2.42	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.21	3.07	2.89	2.78	2.70	2.54	2.45	2.37	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.17	3.03	2.85	2.74	2.66	2.49	2.40	2.33	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.13	2.99	2.81	2.70	2.62	2.45	2.36	2.29	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.98	2.84	2.66	2.55	2.47	2.30	2.21	2.13	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.80	2.66	2.48	2.37	2.29	2.11	2.02	1.94	1.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.70	2.56	2.38	2.27	2.18	2.01	1.91	1.82	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.63	2.50	2.31	2.20	2.12	1.94	1.84	1.75	1.60
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.55	2.42	2.23	2.12	2.03	1.85	1.75	1.65	1.49
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.50	2.37	2.19	2.07	1.98	1.80	1.69	1.60	1.43
$+\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.32	2.18	2.00	1.88	1.79	1.59	1.47	1.36	1.00

Chapitre 13

Annexe C : Introduction à R

Ce chapitre fournit une introduction élémentaire à R. Pour plus de détails, voir les liens présentés sur le Kiosk.

13.1 Les bases de R

R est un logiciel de statistique dédié à l'analyse des données et à leur visualisation. Il contient une collection d'outils pour la statistique, un environnement graphique et un langage de programmation orienté objet. La plupart des entités créées en R sont permanentes. Ces entités sont les objets "données, résultats, fonctions", et sont stockées dans le répertoire `.RData` créé par défaut. Le résultat d'une procédure statistique peut être ainsi réutilisé lors de différentes sessions. Il est donc important de créer un répertoire pour chaque projet statistique effectué en R.

On ouvre une session de R par la commande :

```
$ R
```

Pour clôturer une session, utiliser :

```
> q()
```

L'historique d'une session est conservé dans le fichier `.Rhistory`.

R possède une documentation en ligne accessible par :

```
> help.start()
```

Techniquement, R est un langage fonctionnel. Les commandes élémentaires sont constituées d'expressions et d'affectations. Par exemple :

```
> 2 + 5
```

```
[1] 7
```

```
> a <- c(9,3,7,5)
```

```
> a
```

```
[1] 9 3 7 5
```

```
> a + 3
```

```
[1] 12 6 10 8
```

```
> a[2:4]
```

```
[1] 3 7 5
```

```
> a[a>6]
[1] 9 7
```

R peut être complété en écrivant de nouvelles fonctions. Voici un exemple où l'on souhaite calculer la statistique `stat.log(x) = -\frac{1}{n} \sum_{i=1}^n \ln x_i` où $\forall i, x_i > 0$. On pourra définir une fonction de la façon suivante (même si l'on peut faire bien plus rapide en pratique) :

```
> stat.log <- fonction(x)
+ {
+ n <- length(x)
+ s <- 0
+ for(i in (1:n)) { s <- s + log(x[i]) }
+ -s/n
+ }
```

La fonction `stat.log` pourra être désormais utilisée comme une fonction standard de R. D'un point de vue pratique, on peut éditer ses fonctions dans un éditeur externe (`nedit`, `emacs`, ...) puis faire du copier/coller vers R ou bien utiliser la commande `source`.

13.2 Commandes pour les deux premiers TD en R

Pour enregistrer une figure dans un fichier au format postscript, commencer par rediriger la sortie graphique vers le fichier de sauvegarde, ici "nomfichier.eps" :

```
postscript("nomfichier.ps", horizontal=FALSE)
```

Puis tracer la figure voulue, par exemple un histogramme :

```
hist(x)
```

Et enfin rediriger la sortie graphique vers la fenêtre initiale :

```
dev.off()
```

Même chose en pdf avec `pdf("nomfichier.pdf")`.

Pour tracer un histogramme des données `x` dont l'aire est égale à 1, les bornes des classes sont données par le vecteur `bornes`, et les plages de valeurs des abscisses par le vecteur `xlim` :

```
histx <- hist(x, prob=T, breaks=bornes, xlim=xlim, ...)
```

Pour un histogramme à classes de même effectif, les bornes des classes peuvent être calculées comme des quantiles empiriques, à l'aide d'une commande du type :

```
breaks <- c(a0, quantile(x,seq(1,k-1)/k),ak)
```

La droite de régression linéaire sur le nuage des points d'abscisses `abs` et d'ordonnées `ord` est obtenue à l'aide de :

```
reg <- lm(ord~abs)
```

La pente de la droite des moindres carrés est donnée par `reg$coefficient[2]` et l'ordonnée à l'origine par `reg$coefficient[1]`.

Pour tracer la droite obtenue, l'une des commandes suivantes pourra être utilisée :

```
lines(abs, fitted.values(reg)) ou abline(reg).
```

13.3 Quelques commandes utiles de R

<code>help(mean)</code>	aide sur la commande <code>mean</code>
<code>x <- c(3,14,15,9)</code>	créé un vecteur ligne $x = (3, 14, 15, 9)$
<code>n <- length(x)</code>	taille du vecteur x
<code>sum(x^2)</code>	$\sum_i x_i^2$
<code>mean(x)</code>	moyenne empirique de l'échantillon x
<code>round(x)</code>	valeurs de x arrondies à l'entier le plus proche
<code>seq(from=1,to=10,by=2)</code>	séquence $(1 + 2k; k \text{ entier}, 1 + 2k \leq 10)$
<code>rep(x,3)</code>	concaténation de 3 répliques du vecteur x
<code>solve(a,b)</code>	solution du système linéaire $ax = b$
<code>diag(x)</code>	matrice diagonale de diagonale x
<code>var(x)</code>	variance estimée $s_n'^2$
<code>sqrt(x)</code>	racine carrée de x , élément par élément.
<code>summary(x)</code>	moyenne, médiane, quartiles et valeurs extrêmes
<code>hist(x)</code>	histogramme de x
<code>sort(x)</code>	tri de x par valeurs croissantes
<code>qqnorm(x)</code>	graphe de probabilités pour la loi normale
<code>plot(x,y)</code>	trace le nuage de points $\{(x_i, y_i)\}_i$
<code>abline(b,a)</code>	superpose au graphique précédent la droite d'équation $y = ax + b$
<code>points(x,z)</code>	superpose au graphique précédent le nuage de points $\{(x_i, z_i)\}_i$
<code>lines(x,z)</code>	superpose au graphique précédent la ligne polygonale reliant les points $\{(x_i, z_i)\}_i$
<code>lm(y~x)</code>	régression linéaire de y sur x
<code>lm(y~x)\$coefficients[2]</code>	pente de la droite de régression
<code>lm(y~x)\$coefficients[1]</code>	ordonnée à l'origine de la droite de régression
<code>lines(x,fitted.values(lm(y~x)))</code>	superpose au graphique précédent la droite de régression
<code>postscript("nom.eps")</code>	redirection de la sortie graphique vers le fichier <code>nom.eps</code>
<code>dev.off()</code>	termine la redirection graphique vers un fichier

<code>par()</code>	contrôle des paramètres graphiques
<code>par(mfcol=c(2,1))</code>	graphique à 2 lignes et 1 colonnes
<code>cat("bonjour", "\n")</code>	imprime à l'écran le mot <code>bonjour</code> et retourne à la ligne
<code>source("nom.R")</code>	charge les commandes R contenues dans le fichier <code>nom.R</code> dans R
<code>if, else</code>	structure de contrôle ou d'itération
<code>for, while, repeat</code>	...

13.4 Les lois de probabilité usuelles en R

Toutes les lois de probabilité usuelles ont été implémentées en R. Chaque loi est identifiée par une abréviation :

- loi binomiale : `binom`
- loi de Poisson : `pois`
- loi géométrique : `geom`. Attention, la commande `geom` concerne en fait la loi de $X-1$, où X est de loi géométrique.
- loi exponentielle : `exp`
- loi gamma : `gamma`
- loi du chi 2 : `chisq`
- loi normale : `norm`
- loi de Student : `t`
- loi de Fisher-Snedecor : `f`
- Loi uniforme : `unif`
- Loi beta de première espèce : `beta`
- Loi de Cauchy : `cauchy`
- Loi hypergéométrique : `hyper`
- Loi log-normale : `lnorm`
- Loi logistique : `logis`
- Loi négative binomiale : `nbinom`
- Loi de Weibull : `weibull`
- Loi de Wilcoxon : `wilcox`

Pour chaque loi, 4 fonctions sont disponibles, identifiées par un préfixe :

- Probabilités élémentaires pour les v.a.d. ou densité pour les v.a.c. : `d`
- Fonction de répartition : `p`
- Quantiles : `q`
- Simulation : `r`

Une commande R pour une loi de probabilité est constituée d'un préfixe suivi de l'abréviation de la loi. Les paramètres dépendent de la loi choisie.

Exemples :

- `pnorm(u)` donne la fonction de répartition de la loi normale centrée-réduite $\mathcal{N}(0, 1)$ au point u , $\phi(u)$. On retrouve la table 1 de la loi normale.


```
> pnorm(0.61)
[1] 0.7290691
```
- `dnorm(x, m, σ)` donne la densité de la loi normale $\mathcal{N}(m, \sigma^2)$ au point x .


```
> dnorm(1.2, 2, 5)
[1] 0.07877367
```
- `qnorm(p)` donne le quantile d'ordre p de la loi $\mathcal{N}(0, 1)$, $\phi^{-1}(p)$. On retrouve la table 2 de la loi normale en prenant $p = 1 - \alpha/2$.


```
> qnorm(1-0.05/2)
[1] 1.959964
```
- `rnorm(n, m, σ)` simule un échantillon de taille n de la loi $\mathcal{N}(m, \sigma^2)$.


```
> rnorm(10, 20, 1)
[1] 21.63128 20.16724 17.21667 18.76593 20.48102 20.46236 20.41822
[8] 19.91344 21.19312 19.89164
```
- `dbinom(k, n, p)` donne $P(K = k)$ quand K est de loi binomiale $\mathcal{B}(n, p)$.


```
> dbinom(3, 5, 0.2)
[1] 0.0512
```
- `rpois(n, λ)` simule un échantillon de taille n de la loi de Poisson $\mathcal{P}(\lambda)$.


```
> rpois(15, 4)
[1] 8 3 2 1 6 6 7 5 3 3 4 4 6 1 1
```
- `qchisq(p, n)` donne le quantile d'ordre p de la loi du chi 2 χ_n^2 . On retrouve la table de la loi du chi 2 en prenant $p = 1 - \alpha$.


```
> qchisq(1-0.05, 20)
[1] 31.41043
```
- `qt(p, n)` donne le quantile d'ordre p de la loi de Student $St(n)$. On retrouve la table de la loi de Student en prenant $p = 1 - \alpha/2$.


```
> qt(1-0.3/2, 12)
[1] 1.083211
```
- `qf(p, ν_1, ν_2)` donne le quantile d'ordre p de la loi de Fisher-Snedecor $F(\nu_1, \nu_2)$. On retrouve la table de la loi de Fisher-Snedecor en prenant $p = 1 - \alpha$.


```
> qf(1-0.05, 8, 22)
[1] 2.396503
```

13.5 Les principaux tests d'hypothèses en R

<code>t.test(x, ...)</code>	test de Student sur l'espérance d'une loi normale
<code>binom.test()</code>	test sur une proportion
<code>var.test(x, y, ...)</code>	test de Fisher sur la variance de 2 échantillons gaussiens indépendants
<code>t.test(x, y, ...)</code>	test de Student sur l'espérance de 2 échantillons gaussiens indépendants
<code>prop.test()</code>	test de comparaison de proportions
<code>chisq.test(x, ...)</code>	test du χ^2 sur les probabilités d'évènements et tables de contingence
<code>ks.test(x, ...)</code>	test de Kolmogorov-Smirnov sur un ou deux échantillons
<code>wilcox.test(x, ...)</code>	test de Wilcoxon-Mann-Whitney sur un ou deux échantillons

13.6 Les graphiques dans R

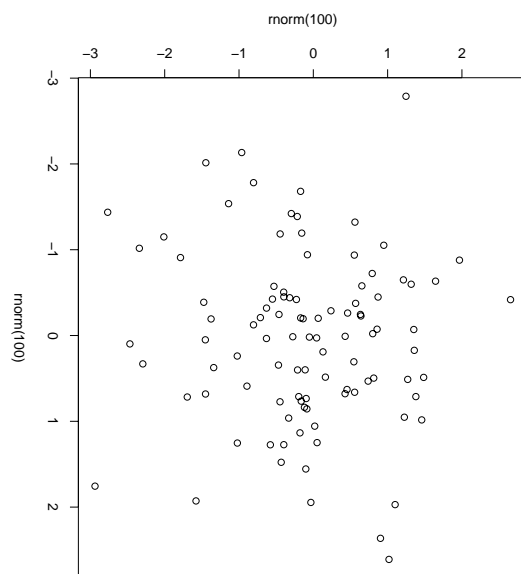
13.6.1 Graphique simple

Le script suivant en R permet de tracer un nuage de 100 points dont les coordonnées sont des variables aléatoires indépendantes et de même loi normale centrée-réduite $\mathcal{N}(0, 1)$, et de le sauvegarder au format postscript dans le fichier "rnorm.ps".

```
postscript("rnorm.ps")
plot(rnorm(100), rnorm(100))
dev.off()
```

Les instructions suivantes permettent d'insérer cette figure dans un document Latex et de pouvoir la référencer sous le nom de figure 13.1.

```
\begin{figure}[htbp]
\begin{center}
% Requires \usepackage{graphicx}
\includegraphics[width=8 cm, angle=270]{rnorm.ps}
\caption{{\it Utilisation de rnorm}}\label{rnorm}
\end{center}
\end{figure}
```

FIGURE 13.1 – Utilisation de *rnorm*

13.6.2 Autres fonctions graphiques

<code>abline(h=u)</code>	ajoute une droite d'équation $y=u$.
<code>abline(v=u)</code>	ajoute une droite d'équation $x=u$.
<code>legend(x,y,legend,...)</code>	ajoute une légende d'utilisation très flexible
<code>text(x,y,labels,...)</code>	ajoute du texte dans un graphe
<code>axis(side,at, labels..)</code>	ajoute un axe au graphique
<code>arrows(x0,y0,x1,y1,...)</code>	dessine des flèches
<code>symbols(x,y,....)</code>	dessine des cercles, des carrés, ...
<code>box(...)</code>	ajoute une boîte
<code>polygon(x,y)</code>	ajoute un polygone
voir aussi <code>image()</code> , <code>pairs()</code> , <code>persp()</code> ,...	

13.6.3 Paramétrage de la commande plot

Le script suivant :

```
postscript("graphesR.ps")
x<- seq(-2*pi,2*pi,0.05)
y <- sin(x)
par(mfrow=c(2,2))
plot(x,y,xlab="x",ylab="Sinus de x")
plot(x,y,type="l", main="trait continu")
plot(x[seq(5,1000,by=5)],y[seq(5,1000,by=5)], type="b",axes=F)
plot(x,y,type="n", ylim=c(-2,1))
text(0,0.05,"Divers paramétrages de la fonction plot")
dev.off()
```

permet d'obtenir la figure 13.2.

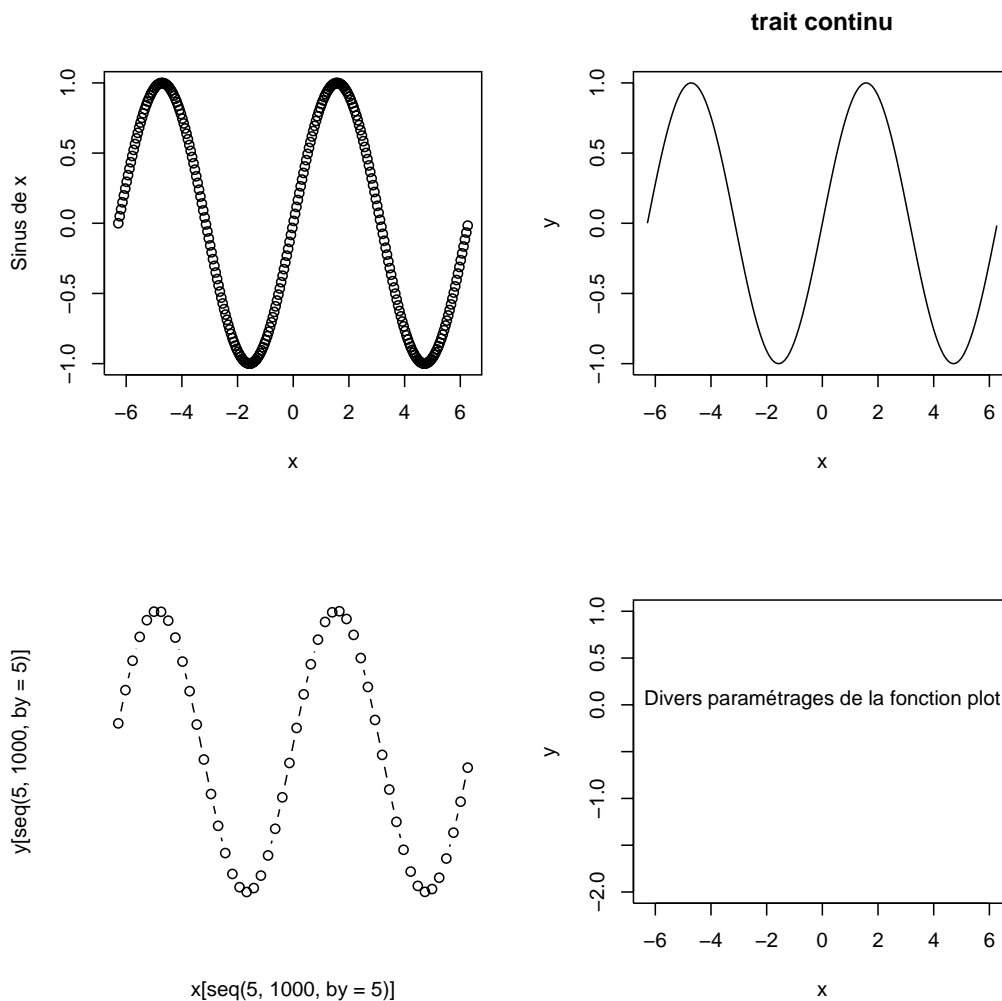


FIGURE 13.2 – R permet de créer plusieurs types de graphiques

Bibliographie

- [1] FOURDRINIER D., *Statistique inférentielle*, Dunod, 2002.
- [2] LEJEUNE M., *Statistique, la théorie et ses applications*, Springer, 2004.
- [3] MONFORT A., *Cours de statistique mathématique*, Economica, 1997.
- [4] RICE J.A., *Mathematical Statistics and Data Analysis*, Duxbury Press, 1995.
- [5] SAPORTA G., *Probabilités, analyse des données et statistique*, Technip, 2006.
- [6] SHAO J., *Mathematical statistics*, Springer, 1998.
- [7] TASSI P., *Méthodes statistiques*, Economica, 1989.