École des Ponts
ParisTech

Statistique Numérique et Analyse des Données

Arnak DALALYAN

Septembre 2011

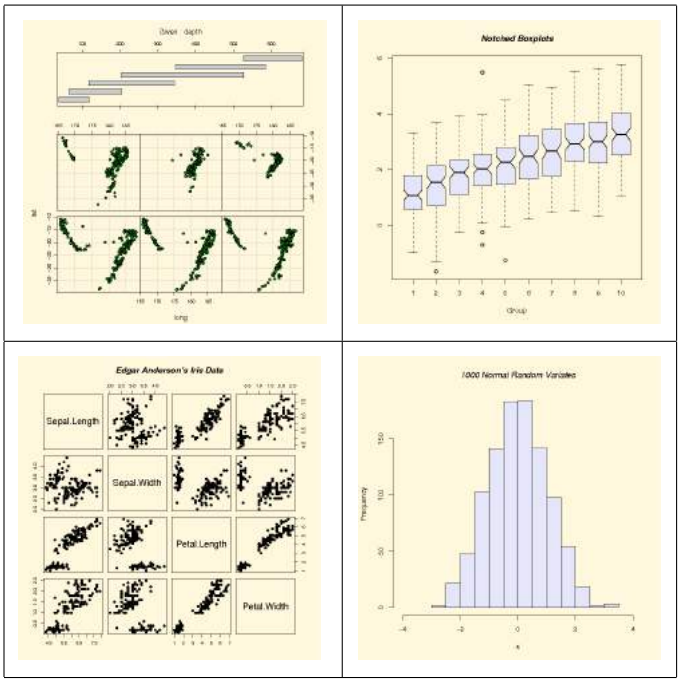


Table des matières

Table des figures

Liste des tableaux

1

Éléments de statistique descriptive

Le but de ce chapitre est de présenter les outils graphiques les plus répandus de la statistique descriptive. On considérera les cas d'une série numérique unidimensionnelle et bidimensionnelle.

Avant de rentrer dans le vif du sujet, apportons une petite précision à une idée très largement répandue, selon laquelle le but de la discipline statistique est d'analyser des données issues d'une expérience à caractère aléatoire. Cela sous-entend qu'il n'est pas possible ou qu'il n'est pas utile d'appliquer la méthodologie statistique aux données recueillies par un procédé déterministe (non aléatoire). Cette une déduction erronée. La bonne définition de l'objectif de la Statistique en tant que discipline scientifique, à notre avis, serait d'explorer les «propriétés fréquentielles» d'un jeu de données. Par «propriétés fréquentielles», on comprend les propriétés qui restent invariantes par toute transformation des données (comme, par exemple, la permutation) qui ne modifie pas la fréquence des résultats.

Le but de ce chapitre est d'introduire les statistiques principales et de donner un aperçu des outils graphiques les plus utilisés.

1.1 Répartition d'une série numérique unidimensionnelle

Supposons que les données qu'on a à notre disposition représentent n valeurs réelles – notées $x_1, \ldots, x_n$ – constituant les résultats d'une certaine expérience répétée n fois. Des exemples de source de telles données sont : les sondages, les expériences scientifiques (physiques, chimiques, médicales,...), les enregistrements historiques (météorologiques, socio-économiques,...). Dans certains cas, ces données sont volumineuses et difficiles à interpréter. On a alors besoin de les résumer et de trouver des outils pertinents pour les visualiser.

Afin que l'analyse statistique d'une série numérique ait un sens, il faut que les différents éléments de cette série représentent la même quantité mesurée sur des entités différentes. Par exemple, $x_1, \ldots, x_n$ peuvent être les hauteurs de n immeubles choisis au hasard à Paris, ou les températures journalières moyennes à Paris enregistrées au cours de l'année 2009, etc. On dit alors que $x_1, \ldots, x_n$ sont les valeurs d'une variable (statistique) observées sur n individus.

On va différencier deux types de séries numériques : celles qui représentent une variable discrète et celles qui représentent une variable continue [1]. On dit qu'une variable est discrète, si le nombre de valeurs différentes parmi $x_1, \ldots, x_n$ est petit devant n. Cette définition est loin d'être rigoureuse, mais cela n'est en général pas très gênant. Dans les deux exemples donnés au paragraphe précédent, les variables «hauteur d'immeuble» et «température journalière moyenne» sont continues. Si au lieu de mesurer la hauteur d'un immeuble, on comptait le nombre d'étages, ce serait une variable discrète.

1.1.1 Histogramme

Pour les séries numériques représentantes une variable discrète, on définit l'histogramme comme la fonction $h : \mathbb{R} \to \mathbb{N}$ qui à chaque $x \in \mathbb{R}$ associe le nombre d'éléments dans la série $x_1, \ldots, x_n$ égaux à x. Par exemple, l'histogramme de la série numérique

$$\begin{array}{cccccccccccc} 10 & 8 & 9 & 6 & 5 & 9 & 8 & 7 & 6 & 5 & 6 & 9 & 10 \\ 8 & 7 & 8 & 7 & 8 & 7 & 6 & 9 & 10 & 9 & 8 & 5 & 9 \end{array} \tag{1.1}$$

est tracé dans la Figure 1.1 (à gauche). Une approche alternative consiste à définir $h(x)$ comme la proportion des éléments dans la série égaux à x. On utilise alors la forme analytique

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

Dans le cas où la série numérique qu'on cherche à analyser est continue, on commence par choisir une partition de $\mathbb{R}$ en un nombre fini d'intervalles : $I_0, I_1, \ldots, I_k$. Ayant fixé la partition, on définit l'histogramme de la série $x_1, \ldots, x_n$ comme la fonction $h : \mathbb{R} \to \mathbb{R}_+$ donnée par la formule

$$\boxed{h(x) = \frac{n_j}{n|I_j|}, \qquad \text{si } x \in I_j,}$$

où n_j est le nombre d'éléments de la série qui se trouvent dans le $j$ème intervalle I_j de la partition et $|I_j|$ est la longueur de l'intervalle I_j. Le choix de la partition est une question délicate que l'on n'approfondira pas ici. Dans la plupart des cas, on choisit une partition uniforme (c'est-à-dire, tous les I_j sont de même longueur) d'un intervalle contenant toutes les valeurs de la série numérique. De plus, on essaye de faire en sorte qu'il y ait au moins 5 observations dans chaque intervalle non-vide.

Par exemple, l'histogramme de la série numérique

$$\begin{array}{cccccccccc} 0.11 & 0.81 & 0.94 & 0.62 & 0.50 & 0.29 & 0.48 & 0.17 & 0.26 & 0.55 \\ 0.68 & 0.17 & 0.28 & 0.57 & 0.98 & 0.77 & 0.56 & 0.49 & 0.31 & 0.89 \\ 0.76 & 0.39 & 0.64 & 0.05 & 0.91 & 0.78 & 0.59 & 0.79 & 0.07 & 0.86 \end{array} \tag{1.2}$$

est tracé dans la Figure 1.1 (à droite).

1.1.2 Fonction de répartition empirique

Une représentation alternative des fréquences des valeurs contenues dans une série numérique est la fonction de répartition, appelée également histogramme cumulé. Pour un $x \in \mathbb{R}$,

1. Le terme *variable continue* n'est pas très bien choisi, mais cela ne pose pas de problème majeur.

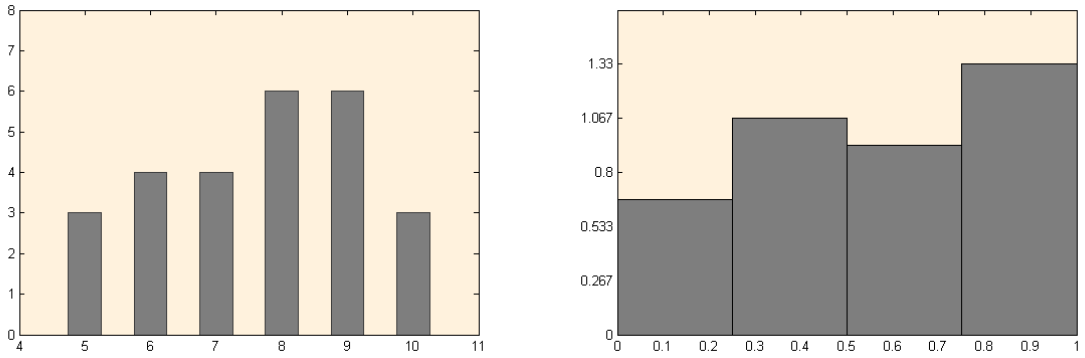


FIGURE 1.1 – Exemples d'histogrammes. A gauche : l'histogramme de la série discrète (1.1). A droite : l'histogramme de la série (1.2).

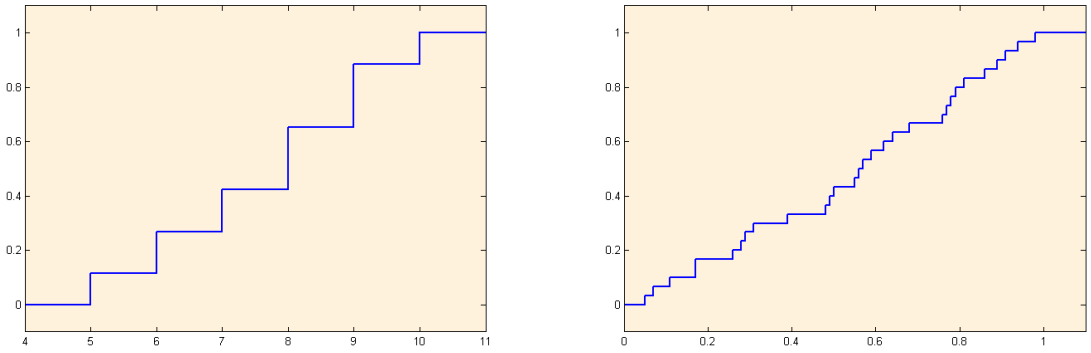


FIGURE 1.2 – Fonction de répartition empirique (FDRE). A gauche : la FDRE de la série discrète (1.1). A droite : la FDRE de la série (1.2). On voit bien que c'est une fonction en escalier croissante, qui vaut 0 sur l'intervalle $]-\infty, \min_i x_i[$ et qui vaut 1 sur l'intervalle $]\max_i x_i, +\infty[$.

la valeur en x de la fonction de répartition d'une série numérique $x_1, \ldots, x_n$ est la proportion des éléments de la série inférieurs ou égaux à x, c'est-à-dire :

$$\boxed{\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x).}$$

L'avantage de la fonction de répartition, comparé à l'histogramme, est que sa définition est identique dans le cas d'une variable discrète et dans le cas d'une variable continue.

1.2 Statistiques d'une série numérique unidimensionnelle

On appelle une statistique toute fonction qui associe aux données $x_1, \ldots, x_n$ un vecteur $S(x_1, \ldots, x_n) \in \mathbb{R}^p$. On utilise les statistiques pour résumer les données.

1.2.1 Statistiques de tendance centrale et de dispersion

Les trois statistiques de tendance centrale les plus utilisées sont la moyenne, la médiane et le mode. On les appelle également les statistiques de position.

La moyenne, notée $\bar{x}$, est définie par :

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

La médiane, notée Med_x, est un nombre réel tel qu'au moins la moitié des données sont $\leq Med_x$ et au moins la moitié des données sont $\geq Med_x$.

Le mode, noté $Mode_x$, est la valeur la plus fréquente à l'intérieur de l'ensemble des données.

Contrairement à la moyenne, la médiane et le mode ne sont pas toujours uniques.

Les trois statistiques de dispersion les plus utilisées sont la variance, l'écart-type et l'écart interquartile.

La variance, notée v_x, est la valeur moyenne des carrés des écarts entre les données et la moyenne :

$$\boxed{v_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.}$$

L'écart-type, notée s_x, est la racine carré de la variance : $s_x = \sqrt{v_x}$.

L'écart interquartile est la différence entre le troisième et le premier quartile : $Q_3 - Q_1$, où le premier quartile Q_1 (respectivement, le troisième quartile Q_3) est la médiane des données $< Med_x$ (resp. $> Med_x$).

1.2.2 Statistiques d'ordre et quantiles

Etant donné une série de données unidimensionnelles $x_1, \ldots, x_n$, on s'intéresse souvent à la plus petite valeur $\min_i x_i$ ou à la plus grande valeur $\max_i x_i$ prise par les x_i. En statistique, on utilise les notations

$$\boxed{x_{(1)} = \min_{1\leq i\leq n} x_i,} \qquad \boxed{x_{(n)} = \max_{1\leq i\leq n} x_i,}$$

et on les appelle première et dernière statistiques d'ordre. Plus généralement, on définit la statistique d'ordre de rang k, notée $x_{(k)}$, comme la $k^{\text{ème}}$ plus petite valeur parmi $x_1, \ldots, x_n$. Plus précisément, soit $(i_1, \ldots, i_n)$ une permutation (il peut y en avoir plusieurs) des indices $(1, \ldots, n)$ qui classe les données dans l'ordre croissant :

$$x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_n}.$$

On appelle alors statistique d'ordre k la valeur $x_{(k)} = x_{i_k}$.

Pour toute valeur $\alpha \in [0, 1]$, on appelle quantile d'ordre α, noté q^x_α, de la série $x_1, \ldots, x_n$, la statistique d'ordre $x_{(m)}$ avec $m = [\alpha n]$. En utilisant la notion de quantile, on peut redéfinir les quartiles et la médiane comme suit :

$$Q_1 = q^x_{0.25}, \qquad Med_x = q^x_{0.5}, \qquad Q_3 = q^x_{0.75}.$$

En pratique, ces définitions de quartiles et médiane conduisent vers des résultats qui diffèrent légèrement de ceux obtenus par la première définition, mais généralement la différence n'est pas importante et décroît lorsque la taille n de la série augmente.

1.2.3 Statistiques de forme*

Les deux statistiques de forme les plus utilisées sont le coefficient d'asymétrie et le coefficient d'aplatissement. Le coefficient d'asymétrie (skewness), notée α_x, et le coefficient d'aplatissement (kurtosis), notée β_x, sont définis par :

$$\alpha_x = \frac{1}{ns_x^3}\sum_{i=1}^{n}(x_i - \bar{x})^3, \qquad \beta_x = -3 + \frac{1}{ns_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

On peut facilement vérifier que le coefficient d'asymétrie de toute série numérique symétrique est nul. (On dit qu'une série numérique est symétrique par rapport à un nombre réel mu, si pour tout $a > 0$ la fréquence de la valeur $\mu + a$ dans la série est égale à celle de $\mu - a$.

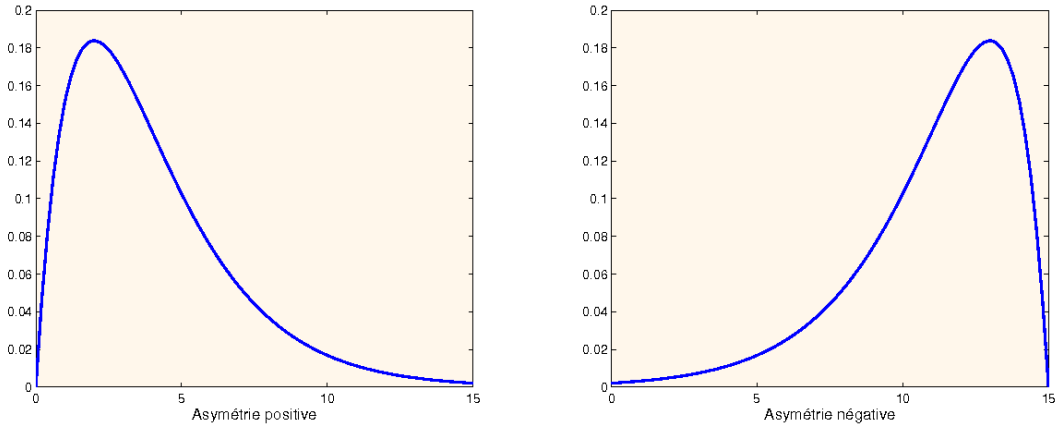


FIGURE 1.3 – Exemples de répartitions asymétriques : le coefficient d'asymétrie est positive pour la distribution à gauche et négative pour celle de droite.

On peut également vérifier que le coefficient d'aplatissement tend vers zéro lorsque $n \to \infty$ si la série numérique représente des réalisations indépendantes de la loi gaussienne $\mathcal{N}(0,1)$.

1.2.4 Box plots (Boîtes à moustaches)

Un résumé simple et pratique de la répartition d'une série $x_1, \ldots, x_n$ est donné par le quintuplé (A, Q_1, Med_x, Q_3, B), où

- A et B représentent les limites inférieure et supérieure de l'intervalle en dehors duquel les données sont considérées comme aberrantes (on les appelle aussi atypiques ou des outliers).
- Q_1 et Q_3 sont respectivement le premier et le troisième quartile.
- Med_x est la médiane de l'échantillon.

Ce quintuplé est utilisé pour construire le diagramme en boîte ou à moustaches que nous appellerons désormais boxplot. La forme générale d'un boxplot est montrée dans la Figure 1.4. Les valeurs A et B sont déterminées par les formules

$$A = \min\Big\{x_i : x_i \geq Q_1 - 1.5(Q_3 - Q_1)\Big\},$$
$$B = \max\Big\{x_i : x_i \leq Q_3 + 1.5(Q_3 - Q_1)\Big\}.$$

Si la série numérique a une répartition normale (Gaussienne), la probabilité qu'une valeur de la série se trouve en dehors de l'intervalle $[A, B]$ est de 0.7%.

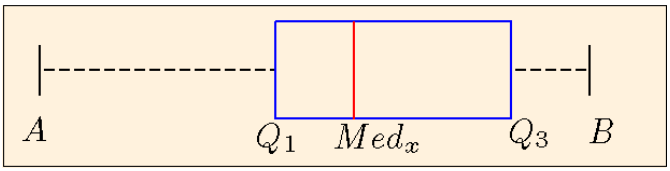


FIGURE 1.4 – La forme typique d'une boîte à moustaches (ou boxplot), le rectangle bleu étant la boîte et les segments $[A, Q_1]$ et $[Q_3, B]$ étant les moustaches.

Pour compléter le boxplot, on fait apparaître les valeurs aberrantes. Toutes les valeurs qui se trouvent en dehors de l'intervalle $[A, B]$ sont désignées par un symbole (souvent par une étoile). Dans l'exemple de la Fig. 1.4, il n'y a pas de valeur aberrante.

Pour interpréter un boxplot, il faut noter que
- la moitié des valeurs de la série se trouvent entre Q_1 et Q_3, c'est-à-dire dans la boîte du boxplot,
- la moitié des valeurs de la série se trouvent à gauche de la médiane,
- s'il n'y a pas de valeurs aberrantes, toutes les valeurs de la série se trouvent entre A et B.

Les boxplots sont pratiques pour comparer deux séries statistiques.

1.3 Statistiques et représentations graphiques de deux séries numériques

Considérons maintenant le cas de deux séries numériques $x_1, \dots, x_n$ et $y_1, \dots, y_n$ correspondant aux valeurs de deux variables prélevées sur le même individu. Par exemple, x_i et y_i peuvent constituer la taille et le poids d'une personne, la température moyenne et le niveau de pollution à Paris un jour donné,...

1.3.1 Covariance et corrélation

La statistique la plus utilisée dans le contexte de deux séries numériques est la corrélation. Pour la définir, la notion de covariance doit être introduite. On appelle covariance des séries numériques $x_1, \dots, x_n$ et $y_1, \dots, y_n$ la valeur

$$\boxed{s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),}$$

où $\bar{x}$ et $\bar{y}$ sont respectivement la moyenne des x_i et celle des y_i.

On appelle coefficient corrélation ou coefficient corrélation linéaire des séries numériques $x_1, \dots, x_n$ et $y_1, \dots, y_n$ la valeur

$$\boxed{\rho_{xy} = \frac{s_{xy}}{s_x s_y},}$$

où s_x et s_y sont respectivement l'écart-type des x_i et celui des y_i. Par convention, on pose $\rho_{xy} = 0$ si au moins l'un des deux écart-types s_x, s_y est nul.

Proposition 1.1. *Le coefficient de corrélation est toujours entre -1 et $+1$:*

$$\boxed{-1 \leq \rho_{xy} \leq 1.}$$

De plus, $|\rho_{xy}| = 1$ si et seulement si les séries $x_1, \dots, x_n$ et $y_1, \dots, y_n$ sont liées par une relation affine, c'est-à-dire $x_i = ay_i + b$ pour tout $i = 1, \dots, n$.

Démonstration. En utilisant l'inégalité de Cauchy-Schwarz, on vérifie que

$$|s_{xy}| \leq \frac{1}{n}\sum_{i=1}^{n} |(x_i - \bar{x})(y_i - \bar{y})| \leq \frac{1}{n}\Big(\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\Big)^{\frac{1}{2}} = s_x \cdot s_y.$$

Cela implique que le coefficient de corrélation $\rho_{xy} = s_{xy}/(s_x s_y)$ est toujours entre -1 et $+1$. De plus, l'inégalité de Cauchy-Schwarz est une égalité si et seulement si $x_i - \bar{x} = a(y_i - \bar{y})$, ce qui entraîne la seconde assertion de la proposition. □

1.3.2 Nuage de points et droite de régression

Supposons que l'on dispose de deux séries numériques $x_1, \dots, x_n$ et $y_1, \dots, y_n$ représentant les valeurs de deux variables prélevées sur n individus. Il est naturel et pratique de représenter ces données sous forme d'un nuage de points. Il s'agit de représenter par un symbole (losange, dans l'exemple de la Fig. 4.1) les n points de coordonnées (x_i, y_i).

A titre d'exemple, considérons les données présentées dans la Table 1.1. Ces données représentent deux variables dont les valeurs sont enregistrées pour $n = 38$ individus. Les individus sont des pays, alors que les deux variables X et Y sont respectivement le PIB (produit intérieur brut) par habitant et la consommation d'énergie par habitant. Le nuage de point de ces données est affiché dans la partie haute de la Figure 4.1. Dans ce contexte, l'identité des individus représente un intérêt (cela n'est pas toujours le cas). Il est alors pratique de marquer à côté de chaque point du nuage une chaîne de caractère permettant l'identification de l'individu représenté par le point. C'est ce qui est fait dans la partie basse de la Fig. 4.1.

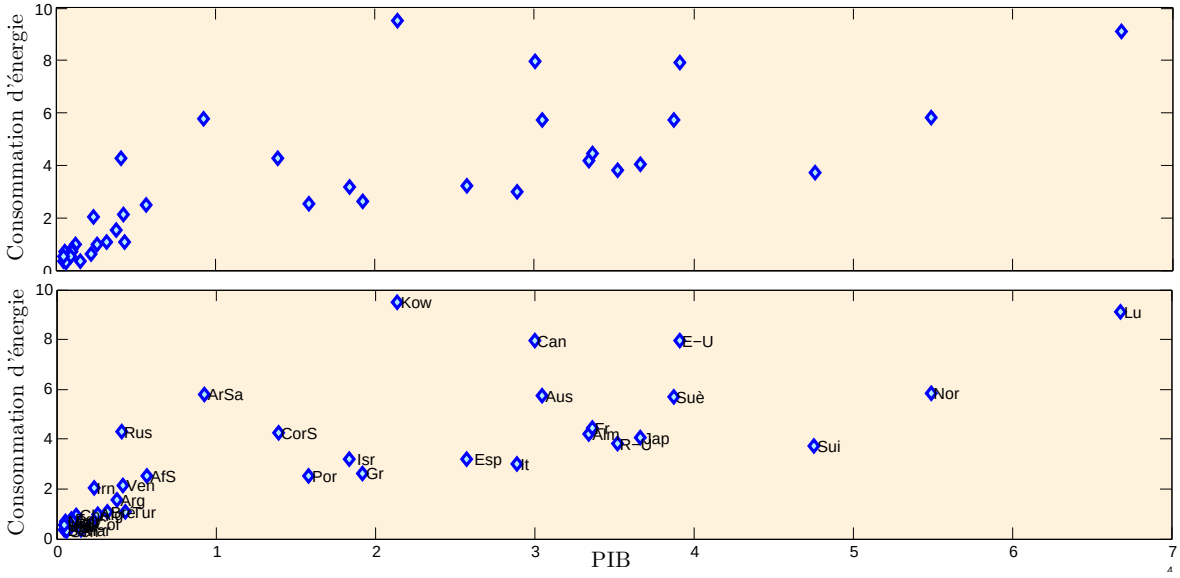


FIGURE 1.5 – Le nuage de points représentant les données de la Table 1.1. En haut : le nuage simple. En bas : le nuage annoté

Pour rendre le nuage de point plus lisible, on a souvent recours à une transformation d'une ou des deux variables. Dans l'exemple de la Table 1.1, on obtient un nuage de point plus interprétable (voir la Fig. 4.3) en prenant le logarithme des deux variables.

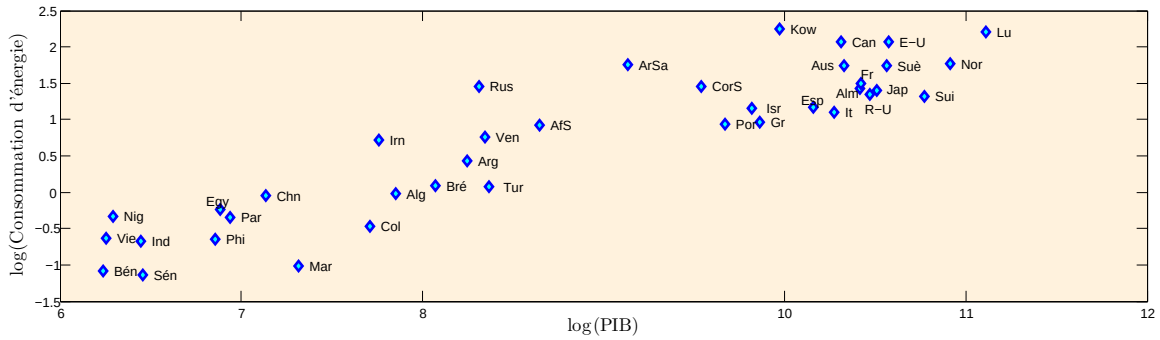


FIGURE 1.6 – Le nuage de points représentant les logarithmes des données de la Table 1.1.

Afin d'obtenir une droite approximant le nuage de points, on calcule la droite de régression de Y sur X, donnée par l'équation $y = ax + b$ où

$$\boxed{a = \frac{s_{xy}}{s_x^2},} \qquad \boxed{b = \bar{y} - a\bar{x}.} \tag{1.3}$$

Pour les données de la Table 1.1, la droite de régression ainsi que son équation sont données dans la Fig. 1.7. On voit dans la formule (1.3) que la droite de régression de Y sur X ne coïncide pas, en général, avec la droite de régression de X sur Y. Si l'on note M_i le point qui a pour coordonnées (x_i, y_i) et par d_i la distance entre M_i et le point $M_i' = (x_i, ax_i + b)$, alors la droite de régression est la droite pour laquelle la somme des d_i au carré est minimale. C'est la raison pour laquelle on dit que la droite de régression est obtenue par la méthode des moindres carrés. On reparlera de cette propriété dans un cadre plus général plus loin dans ce document.

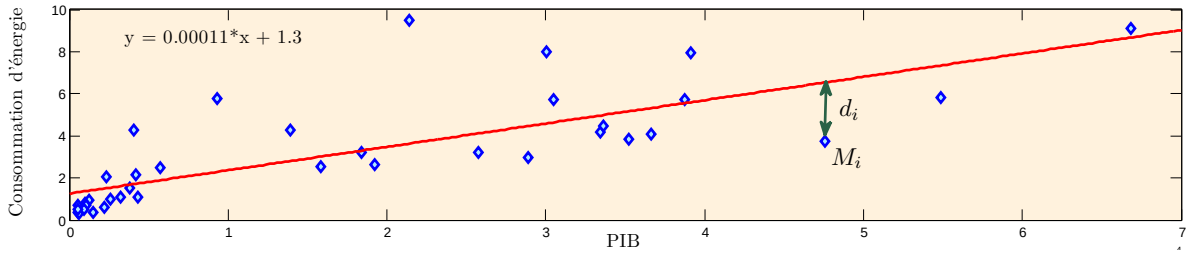


FIGURE 1.7 – Le nuage de points représentant les données de la Table 1.1 superposé de la droite de régression.

1.3.3 QQ-plot* (graphiques quantile-quantile)

Un QQ-plot permet de voir rapidement l'adéquation d'une série numérique à une distribution, ou comparer les répartitions de deux séries numériques.

1er cas : Lorsque l'on s'intéresse à l'adéquation à une distribution, l'axe des ordonnées porte les quantiles q_j de la distribution observée, tandis que l'axe des abscisses porte les quantiles q_j^* correspondants de la loi théorique.

2ème cas : Lorsque l'on s'intéresse à la comparaison de deux distributions, l'axe des ordonnées porte les quantiles q_j^x de la série $x_1, \ldots, x_n$, tandis que l'axe des abscisses porte les quantiles q_j^y de la série $y_1, \ldots, y_n$.

Le nuage des points (q_j^*, q_j) (respectivement (q_j^y, q_j^x)) s'aligne sur la première bissectrice lorsque la distribution théorique proposée est une bonne représentation des observations (resp., lorsque les répartitions des x_i et des y_i sont égales).

Si le nuage des points (q_j^*, q_j) s'aligne sur une droite, alors il existe une transformation affine des observations telle que la distribution théorique proposée est une bonne représentation des observations transformées.

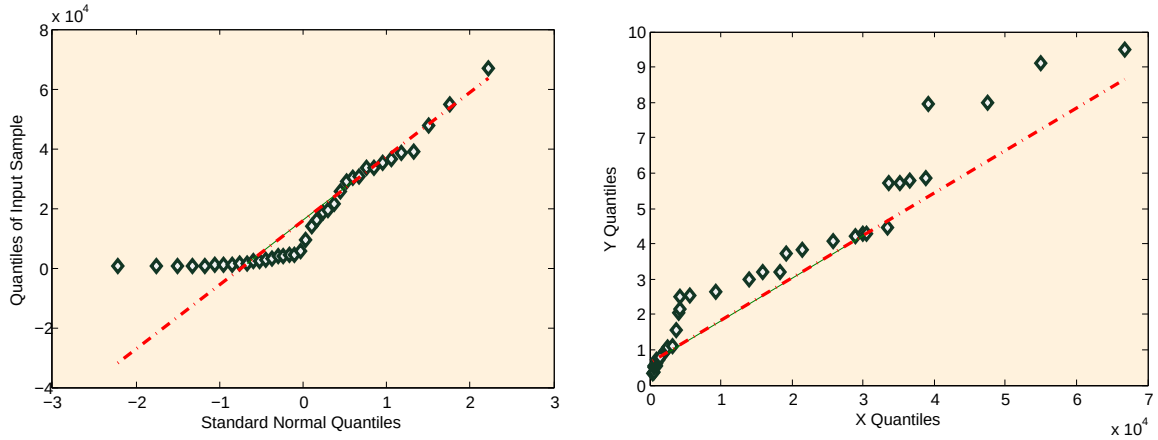
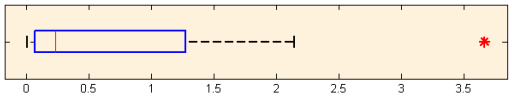


FIGURE 1.8 – QQ-plots pour les données de la Table 1.1. Le graphe de gauche indique que la répartition du PIB est significativement différente d'une loi normale. Le graphe de droite montre que les répartitions du PIB et de la consommation d'énergie ne sont pas liées par une transformation affine.

Exercice 1.1. *Le tableau suivant présente les données du PIB par habitant pour 15 pays dont la majeure partie se trouve en Asie. Ces données ont été obtenues sur le site* `http://www.statistiques-mondiales.com/`. *Le boxplot de ces données a la forme suivante :*



1. *Selon ce diagramme, quelle est la valeur médiane du PIB/habitant en Asie ?*
2. *Y a-t-il des données atypiques ?*
3. *La répartition du PIB/habitant est-elle symétrique ? Comment s'interprète cette asymétrie ?*
4. *Répondre à la question 3 en utilisant l'information que la moyenne des 15 observations qu'on dispose est de 7670.*

Pays	PIB / habitant (en $ US, 2004)
Afghanistan	174
Arabie Saoudite	9285
Arménie	1034
Chine	1258
Corée du Sud	13929
Inde	631
Iran	2350
Israël	18404
Japon	36647
Koweït	21420
Pakistan	81
Philippines	948
Russie	4071
Turquie	4296
Vietnam	520

1.4 Résumé du Chapitre 1

Série numérique :

Variable discrète :

Variable continue :

Histogramme :
- variable discrète :
- variable continue :

Fonction de répartition empirique :

Statistiques de tendance centrale :
- moyenne :
- médiane :
- mode :

Statistiques de dispersion :
- variance :
- écart-type :
- écart interquartile :

Statistiques d'ordre :

Quantiles :

Boxplots :

Covariance :

Corrélation :

Nuage de points :

Droite de régression :

QQ-plot :

	Pays	PIB par habitant (en $ US, en 2004)	Consommation d'énergie par habitant (en Tonnes d'équivalent pétrole, en 2002)
	Afrique du sud	5700	2.502
	Algérie	2575	0.985
	Bénin	512	0.340
	Egypte	976	0.789
	Maroc	1505	0.363
	Nigeria	541	0.718
	Sénégal	637	0.319
	Allemagne	33422	4.198
	Espagne	25777	3.215
	France	33614	4.470
	Grèce	19226	2.637
	Italie	28909	2.994
	Luxembourg	66808	9.112
	Norvège	54894	5.843
	Portugal	15835	2.546
	Royaume-Uni	35193	3.824
	Suède	38746	5.718
	Suisse	47577	3.723
	Arabie Saoudite	9285	5.775
	Chine	1258	0.960
	Corée du Sud	13929	4.272
	Inde	631	0.513
	Iran	2350	2.044
	Israël	18404	3.191
	Japon	36647	4.058
	Koweït	21420	9.503
	Philippines	948	0.525
	Russie	4071	4.288
	Turquie	4296	1.083
	Vietnam	520	0.530
	Argentine	3808	1.543
	Brésil	3210	1.093
	Canada	30014	7.973
	Colombie	2234	0.625
	Etats-Unis	39114	7.943
	Paraguay	1032	0.709
	Venezuela	4203	2.141
	Australie	30498	5.732

TABLE 1.1 – Ces données sont obtenues du site `http://www.statistiques-mondiales.com/`

2

Analyse des données multivariées

2.1 Introduction

2.1.1 Objectif

Dans toute étude appliquée, la démarche première du statisticien est de décrire et d'explorer les données dont il dispose, avant d'en tirer de quelconques lois ou modèles prédictifs. Or la statistique traite généralement du grand nombre et, les outils informatiques aidant, les bases de données deviennent de plus en plus volumineuses, tant en largeur (quantité d'informations recueillies) qu'en hauteur (nombre d'unités sur lesquelles ces informations sont recueillies).

Cette phase d'exploration descriptive des données n'est en conséquence pas aisée. Si le statisticien est déjà outillé pour analyser la distribution d'une variable ou la relation entre deux variables, ces outils basiques ne permettent pas d'appréhender ce vaste ensemble informatif dans sa globalité. Il ne s'agit naturellement pas d'en donner alors une vision exhaustive, mais bien de répondre à l'une des principales missions du statisticien : extraire d'une masse de données ce qu'il faut en retenir, en la synthétisant ou en simplifiant les structures.

Les techniques d'analyse de données répondent à ce besoin. On présentera ici l'Analyse en Composantes Principales (ACP) qui s'appuie sur la réduction de rang découlant des travaux de décomposition matricielle d'Eckart et Young (1936). Le but principal de l'ACP est de déterminer les principales relations linéaires dans un ensemble de variables numériques. Il s'agit bien de réduire un ensemble complexe et de grande dimension à ses principaux éléments, de façon à en mieux comprendre les structures sous-jacentes.

2.1.2 Notations

On dispose de p variables $X^1, \ldots, X^j, \ldots, X^p$, que l'on observe sur n unités statistiques - ou individus : on note x_i^j la valeur de la variable X^j observée sur le i-ème individu. Cet ensemble de données peut donc être mis sous la forme d'un tableau X à n lignes et p colonnes, et de terme courant x_i^j.

Dans la suite - et c'est très généralement le cas en analyse des données, contrairement aux autres domaines de la statistique - on confondra la notion de variable avec le vecteur de dimension n qui la définit sur notre échantillon, c'est-à-dire $X^j = (x_1^j, \ldots, x_n^j)$. De même, chaque individu sera assimilé au vecteur de dimension p qui compile ses valeurs sur les variables : $X_i = (x_i^1, \ldots, x_i^p)$.

$$\boldsymbol{X} = \underbrace{\begin{pmatrix} x_1^1 & \ldots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \ldots & x_n^p \end{pmatrix}}_{p \text{ variables}} \begin{matrix} \rightarrow \text{ individu \# 1, noté } X_1 \\ \vdots \\ \rightarrow \text{ individu \# } n\text{, noté } X_n \end{matrix}$$

2.2 Exemple : billets suisses

Nous choisirons ici un exemple décrivant 6 mesures, notée $X^1, \ldots, X^6$, relevées sur 200 billets de 1000 Francs Suisses. La Figure 2.1 présente la nature des mesures effectuées alors que l'ensemble des données recueillies est donné dans les Tables 2.1 et 2.2. Sur les 200 billets examinés, il y a eu 100 billets authentiques et 100 billets contrefaits. Cet exemple comporte volontairement un nombre réduit de variables, pour en faciliter la compréhension.

Pour comprendre ce qu'apportent les méthodes d'analyse de données, menons au préalable une brève analyse descriptive de ces tableaux du point de vue des variables.

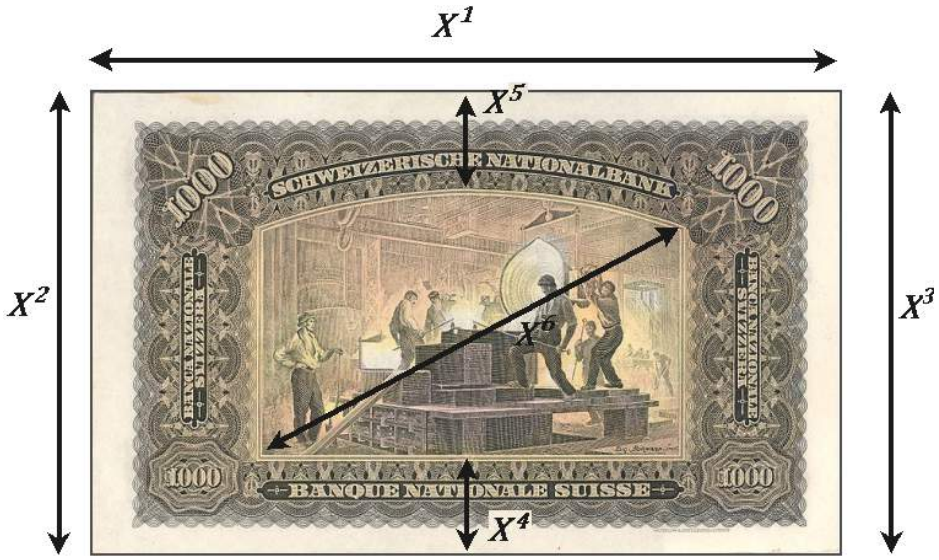


FIGURE 2.1 – Cette figure montre une coupure de 1000 Francs Suisses (anciens) avec les 6 mesures effectuées.

Etude descriptive des variables

Classiquement, on peut se livrer à une analyse de la distribution de chaque variable. Cela peut se faire, par exemple, en visualisant les boxplots de chacune des 6 variables X^i. La Fig. 2.2 montre ces boxplots, qui nous renseignent sur les caractéristiques individuelles des variables X^i. On y voit, entre autre, qu'il y a 2 billets dont la longueur est anormalement

grande et un billet dont la longueur est anormalement petite. On remarque également, en comparant les boxplots de X^2 et X^3, que la largeur à gauche est typiquement légèrement plus grande que la largeur à droite.

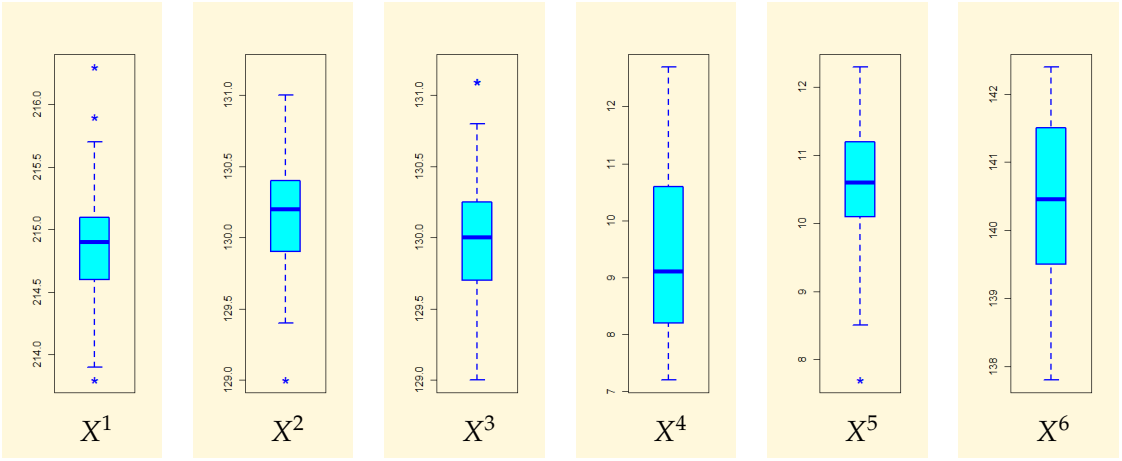


FIGURE 2.2 – Les boxplots des données de billets suisses

Cette figure ne dit cependant rien sur la relation entre les variables. Pour appréhender les distributions bivariées, des outils d'analyse vus à la fin du chapitre précédent peuvent être appliqués à tous les paires de variables. Exemples de tels outils sont la matrice de «scatter plots» (voir Fig. 2.3), ou la matrice des coefficients de corrélation linéaire. Ce dernier représente un intérêt surtout lorsque les nuages sont aplatis ou les répartitions bidimensionnelles sont approximativement gaussiennes.

Voici le tableau des corrélations :

	X^1	X^2	X^3	X^4	X^5	X^6
X^1	1.00	0.23	0.15	-0.19	-0.06	0.19
X^2	0.23	1.00	0.74	0.41	0.36	-0.50
X^3	0.15	0.74	1.00	0.49	0.40	-0.52
X^4	-0.19	0.41	0.49	1.00	0.14	-0.62
X^5	-0.06	0.36	0.40	0.14	1.00	-0.59
X^6	0.19	-0.50	-0.52	-0.62	-0.59	1.00

Ce tableau montre que les variables X^2 et X^3 sont les plus corrélées, ce qui est tout à fait logique et cela se voyait déjà sur le scatter plot de la Fig. 2.3.

On voit donc qu'on dispose des outils qui nous permettent d'analyser les variables individuellement ou deux par deux. Il nous manque cependant des outils de synthèse, qui permettraient de dégager la structure globale de ces données. Nous allons en développer un, parmi les plus utilisés.

2.3 La théorie de l'Analyse en Composantes Principales

2.3.1 Problématique

On se place ici dans la situation où les p variables d'intérêt $X^1, \ldots, X^j, \ldots, X^p$, sont numériques. Pour appréhender l'information contenue dans le tableau numérique $\boldsymbol{X}$, on peut tenter de visualiser le nuage de points représentant les n individus dans $\mathbb{R}^p$. Mais très souvent, le nombre de variables p peut atteindre quelques dizaines. Quoiqu'il en soit, même

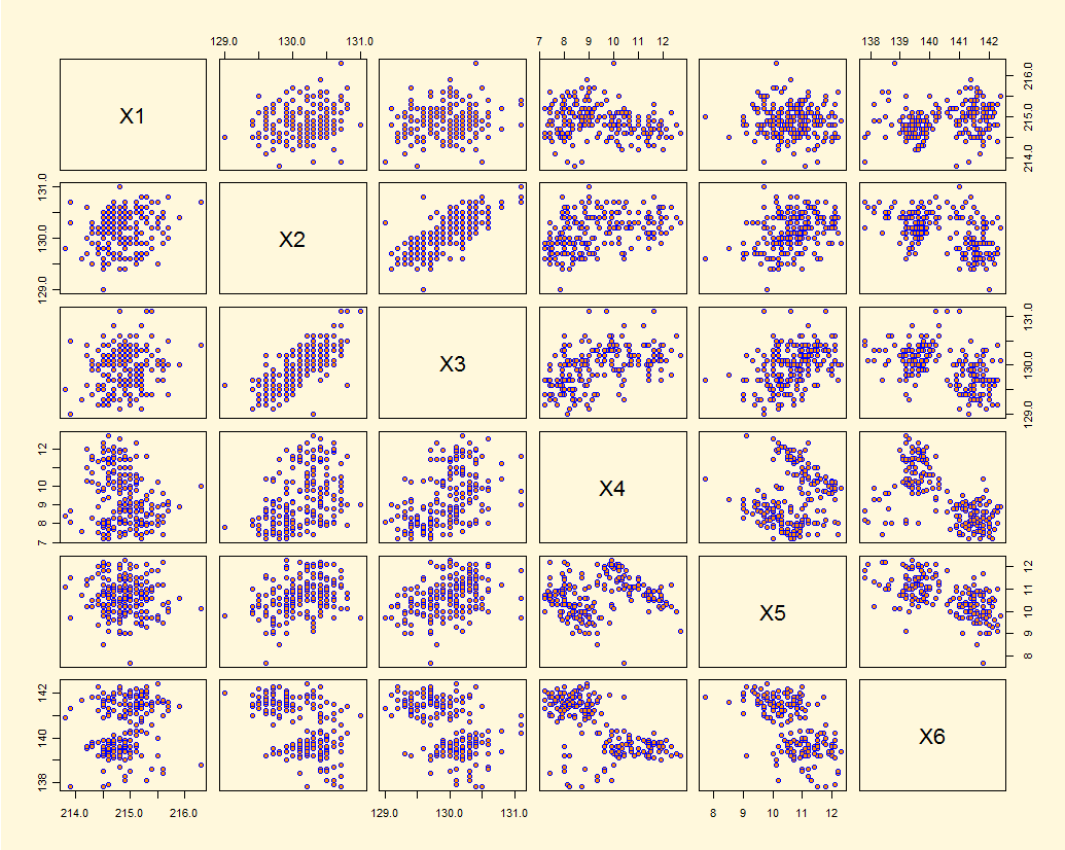


FIGURE 2.3 – Scatter plots des différentes variables

avec des outils de visualisation performants, X ne peut être appréhendé de façon simple dans sa globalité, ni les relations entre les variables.

La problématique est alors double :
– Comment visualiser la forme du nuage des individus ?
– Comment synthétiser les relations entre variables ?
L'ACP permet justement de répondre à ce type de besoin.

2.3.2 Choix de la métrique

La méthode d'Analyse en Composantes Principales requiert un espace vectoriel muni d'un produit scalaire. Dans ce chapitre, nous considérerons l'espace euclidien $\mathbb{R}^p$ muni de son produit scalaire canonique. La métrique associée est donnée par

$$\|X_i - X_{i'}\|^2 = \sum_{j=1}^{p} \left(x_i^j - x_{i'}^j\right)^2.$$

Définition 2.1. *Soient* $\bar{x}^j = \frac{1}{n}\sum_{i=1}^n x_i^j$ *et* $s_j^2 = \frac{1}{n}\sum_{i=1}^n \left(x_i^j - \bar{x}^j\right)^2$ *la moyenne et la variance de la variable d'intérêt* X^j. *La représentation centrée de l'individu* i *est donnée par* $\tilde{x}_1^j, \ldots, \tilde{x}_p^j$, *où pour tout* $1 \leq j \leq p$,

$$\tilde{x}_i^j = x_i^j - \bar{x}^j.$$

La représentation centrée-réduite de l'individu i est donnée par $\tilde{x}_1^j, \dots, \tilde{x}_p^j$, où pour tout $1 \leq j \leq p$,

$$\tilde{x}_i^j = \frac{x_i^j - \bar{x}^j}{s_j}.$$

Une ACP normée est une ACP menée sur la représentation centrée-réduite.

L'ACP opère toujours sur les représentations centrées. Pour simplifier la présentation, on considérera dans la suite que les variables ont été déjà centrées, dans le sens où $\sum_{i=1}^n X_i = 0$. Les différentes variables X^j pouvant être hétérogènes, et correspondre à des échelles de mesure disparates, la représentation centrée-réduite est utilisée pour éviter que le choix de ces unités ait une influence dans le calcul des distances. Cette représentation rend les variables centrées et de variance 1.

2.3.3 Moindre déformation du nuage

Pour visualiser le nuage des individus (et donc en connaître la forme, pour savoir comment sont liées nos p variables), il est nécessaire de réduire la dimension de l'espace qui le porte. L'ACP réduit cette dimension par projection orthogonale sur des sous-espaces affines.

Définition 2.2. *Soit $X_1, \dots, X_n$ un nuage de points dont le barycentre coïncide avec l'origine (c'est le cas pour des variables réduites). L'inertie du nuage $X_1, \dots, X_n$ est donnée par*

$$I = \frac{1}{n} \sum_{i=1}^n \|X_i\|^2.$$

L'inertie J_H du nuage autour du sous-espace linéaire H est donnée par

$$J_H = \frac{1}{n} \sum_{i=1}^n \|X_i - P_H X_i\|^2,$$

où $P_H X_i$ le projeté orthogonal de X_i sur H.

L'inertie J_H autour de H mesure la déformation du nuage lorsque celui-ci est projeté orthogonalement sur H. Pour que la représentation des données par leur projection sur un sous-espace affine ait un sens, il faut qu'elle modifie peu la forme du nuage de points, donc qu'elle minimise l'inertie J_H.
Remarquons que d'après le théorème de Pythagore, on a

$$I = \frac{1}{n} \sum_{i=1}^n \left(\|X_i - P_H X_i\|^2 + \|P_H X_i\|^2 \right) \overset{\text{déf}}{=} J_H + I_H.$$

Par conséquent, la moindre déformation d'un nuage de points par projection orthogonale sur un sous-espace linéaire est obtenue, de manière équivalente, par minimisation de l'inertie par rapport au sous-espace linéaire ou par maximisation de l'inertie du nuage projeté.
Dans le but de pouvoir visualiser le nuage de points des individus, on aimerait trouver dans $\mathbb{R}^p$ un sous-espace linéaire de dimension 2 (c'est-à-dire, un plan) qui approche bien les données. On est donc tout naturellement intéressé par la résolution du problème

$$H_2 = \underbrace{\arg \min_{H:\dim(H)=2} J_H}_{\text{minimisation de la déformation du nuage}} = \underbrace{\arg \max_{H:\dim(H)=2} I_H}_{\text{maximisation de l'inértie du nuage projeté}}$$

D'une façon plus générale, on s'intéresse aux sous-espaces linéaires H_k, pour $k \in \{1, \dots, p-1\}$, définis par

$$H_k = \arg \min_{H:\dim(H)=k} J_H = \arg \max_{H:\dim(H)=k} I_H. \tag{2.1}$$

Par exemple, si le nuage des individus dans $\mathbb{R}^p$ n'est pas bien approximable par un plan, il pourrait être plus intéressant de considérer une visualisation 3 dimensionnelle en projetant les données sur H_3. Dans certains cas, cela peut considérablement augmenter l'inertie du nuage projeté.

Montrons maintenant que la recherche d'un sous-espace affine de dimension fixée maximisant l'inertie du nuage projeté peut être menée de manière séquentielle et que l'inertie se décompose en la somme des inerties moyennes du nuage projeté sur des droites orthogonales, dites directions principales de l'ACP.

Soit Γ la matrice de variance-covariance associée au nuage de points (dans la représentation centrée, les moyennes $\bar{X}^j$ sont nulles) :

$$\Gamma = \frac{1}{n}(\boldsymbol{X})^t \boldsymbol{X},$$

autrement dit $\Gamma_{j,j'} = \sum_{i=1}^n x_i^j x_i^{j'}$ est la covariance entre les variables d'intérêt X^j et $X^{j'}$. Notons au passage que lorsqu'on considère des variables réduites, la matrice Γ est également la matrice des corrélations des variables X^j.

Théorème 2.1. *Les assertions suivantes caractérisent la résolution séquentielle du problème de réduction de dimension par moindre déformation.*

- *Soit u_k un vecteur propre unitaire de Γ associée à la k-ième plus grande valeur propre. Alors $H_k = Vect(u_1, \dots, u_k)$ est l'espace vectoriel engendré par les k premiers vecteurs propres de Γ.*
- *La k-ième plus grande valeur propre λ_k de Γ vaut l'inertie du nuage projeté sur le k-ième axe propre u_k :*

$$I_{u_k} = \lambda_k.$$

- *l'inertie sur H_k est la somme des inerties moyennes sur les k axes propres principaux :*

$$I_{H_k} = \sum_{l=1}^{k} \lambda_l.$$

Démonstration. Cherchons d'abord le vecteur unitaire, i.e. de norme 1, u maximisant l'inertie du nuage projeté sur u. Considérons la projection du nuage sur la direction donnée par le vecteur unitaire u. Le projeté X_i^* de l'individu i s'écrit

$$X_i^* = \langle u, X_i \rangle u$$

et l'inertie du nuage projeté (nous nous plaçons toujours dans le cadre de la représentation réduite) est

$$I_u = \frac{1}{n}\sum_{i=1}^n \|\langle u, X_i\rangle u\|^2 = \frac{1}{n}\sum_{i=1}^n \langle u, X_i\rangle^2 = \frac{1}{n}\sum_{i=1}^n u^t X_i (X_i)^t u = u^t \Gamma u.$$

La matrice Γ est symétrique, semi-définie positive ; elle est diagonalisable, a toutes ses valeurs propres réelles, et il existe une base orthonormale de vecteurs propres de $\mathbb{R}^p$. Notons $\lambda_1 \geq ... \geq \lambda_p$ les valeurs propres triées par ordre décroissant, et $u_1, ..., u_p$ les vecteurs propres unitaires associés. Alors

$$I_u = \sum_{j=1}^p \lambda_j \langle u, u_j \rangle^2 \leq \lambda_1 \sum_{j=1}^p \langle u, u_j \rangle^2 = \lambda_1 ||u||^2 = \lambda_1.$$

Il suffit alors de choisir $u = u_1$ pour maximiser I_u.

Par conséquent, la meilleure droite de projection du nuage est celle de vecteur directeur u_1, associé à la plus grande valeur propre λ_1 de la matrice Γ.

On admet sans démonstration que pour tout entier $k < p$, l'espace H_{k+1} est obtenu à partir de l'espace H_k par $H_{k+1} = \text{Vect}(H_k, v_{k+1})$ où v_{k+1} est un vecteur orthogonal à H_k.

Pour les H_k suivants, on procède par récurrence. Ansi, pour H_2 on cherche le vecteur directeur u_2 orthogonal à u_1 portant l'inertie maximale. Pour tout vecteur u orthogonal à u_1, on a

$$I_u = \sum_{j=2}^{p} \lambda_j \langle u, u_j \rangle^2 \leq \lambda_2.$$

Donc le maximum est atteint pour $u = u_2$, et ainsi de suite.

Au passage, on a également prouvé la deuxième assertion du théorème : $I_{u_k} = \lambda_k$. La troisième assertion découle alors du théorème de Pythagore. □

L'inertie I du nuage de points est donc égale à la trace de matrice de variance-covariance, ce qui implique $I = p$, en ACP normée. (En ACP non normée, elle vaut la somme des variances : $I = \sum_{j=1}^{p} s_j^2 = \sum_{l=1}^{p} \lambda_l$.) On définit la part d'inertie expliquée sur le l-ième axe propre : $\tau_l = \lambda_l / I$. L'inertie portée par un sous-espace de dimension k est donc au mieux $\sum_{l=1}^{k} \tau_l$ pour cent de l'inertie totale I.

2.4 Représentations graphiques et interprétation

Sur notre exemple concernant les billets suisses, on peut chercher à visualiser les proximités (en termes de distance normée sur les 6 caractéristiques) entre billets sur le premier plan factoriel (u_1 horizontalement, u_2 verticalement) (voir Fig.2.4 à gauche). Dans cet exemple,

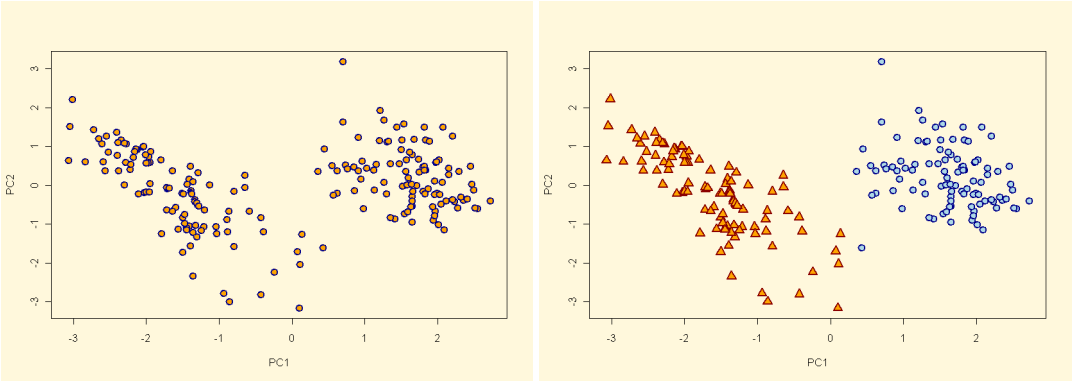


FIGURE 2.4 – A gauche : projection des individus sur le premier plan factoriel. A droite : la même projection avec des symboles différents pour les billets authentiques et les billets contrefaits. Les triangles correspondent aux billets contrefaits, alors que les cercles représentent les billets authentiques.

l'inertie $I = 4.494$ se décompose sur les premiers axes ainsi : $I_1 = 3$ (donc $\tau_1 = 66.7\%$), $I_2 = 0,93$ (donc $\tau_2 = 20.8\%$). On visualise donc de façon simplifiée, mais optimale ($\tau_{1\oplus 2} = I_{u_1 \oplus u_2} / I$ =87.5% de l'inertie représentée sur ce plan), les proximités entre les billets.

Les vecteurs directeurs de ces deux premiers axes s'expriment ainsi, dans l'ancienne base :

Vecteur propre	X_1	X_2	X_3	X_4	X_5	X_6
u_1	-0.04	0.11	0.14	0.77	0.20	-0.58
u_2	0.01	0.07	0.07	-0.56	0.66	-0.49

Reste à interpréter véritablement ces axes, et à comprendre quels sont les principales relations linéaires entre les caractéristiques techniques...

2.4.1 Principales relations entre variables

Les composantes principales

La diagonalisation vue précédemment permet de définir p nouvelles variables[1] appelées composantes principales :

$$C^\alpha = \sum_{j=1}^{p} u_\alpha^j X^j = \boldsymbol{X} u_\alpha \in \mathbb{R}^n,$$

ou encore $C_i^\alpha = \langle X_i, u_\alpha \rangle$. Elles sont donc combinaisons linéaires des variables d'intérêt X^j initiales. Elles sont centrées puisque les X^j le sont, et on a :

$$\text{Cov}\left(C^\alpha, C^\beta\right) = \sum_{j=1}^{p}\sum_{j'=1}^{p} u_\alpha^j u_\beta^{j'} \text{Cov}\left(X^j, X^{j'}\right) = u_\alpha^t \Gamma u_\beta = \lambda_\beta u_\alpha^t u_\beta.$$

Donc $\text{Cov}\left(C^\alpha, C^\beta\right) = \begin{cases} 0 & \text{si } \alpha \neq \beta, \\ \lambda_\alpha & \text{si } \alpha = \beta, \end{cases}$ ce qui veut dire que les différentes composantes principales sont non-corrélées.

On peut calculer la covariance entre les composantes principales et les variables initiales :

$$\text{Cov}\left(C^\alpha, X^j\right) = \sum_{j'=1}^{p} u_\alpha^{j'} \text{Cov}\left(X^{j'}, X^j\right) = \sum_{j'=1}^{p} u_\alpha^{j'} \Gamma_{j',j} = \lambda_\alpha u_\alpha^j.$$

Il s'ensuit que

$$\text{Corr}\left(C^\alpha, X^j\right) = \frac{\text{Cov}\left(C^\alpha, X^j\right)}{\sqrt{\mathbf{Var}(C_\alpha)\mathbf{Var}(X_j)}} = \sqrt{\lambda_\alpha} u_\alpha^j / s_j.$$

Donc $\sum_{j=1}^{p} s_j^2 \text{Corr}^2\left(C^\alpha, X^j\right) = \lambda_\alpha$.

Pour visualiser les corrélations entre les composantes principales et les X^j, on établit des représentations planes où, en prenant par exemple $\left(C^1, C^2\right)$ comme base orthogonale de ce plan, chaque X^j est figuré par un vecteur de coordonnées $\left(\text{Corr}\left(C^1, X^j\right), \text{Corr}\left(C^2, X^j\right)\right)$, à l'intérieur du cercle unité[2], dit des corrélations.

1. De même que précédemment, on confondra sous le vocable variable la forme linéaire, et sa réalisation sur nos n individus, soit encore le vecteur de $\mathbb{R}^n$ associé.

2. Ce vecteur est dans le cercle unité car, dans $\mathbb{R}^n$ muni du produit scalaire $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$, c'est le vecteur projeté orthogonal du vecteur unitaire X^j / s_j sur le plan engendré par les vecteurs orthonormés $C^1 / \sqrt{\mathbf{Var}(C^1)}$ et $C^2 / \sqrt{\mathbf{Var}(C^2)}$.

Retour à l'exemple

On voit, dans cet exemple (voir la partie droite de la Fig. 2.5), que les variables X^1, X^2 et X^3 sont mal expliquées par les deux premiers axes principaux, car les points représentant ces variables sont éloignés du cercle. En revanche, les 3 autres points sont quasiment sur le cercle, ce qui veut dire que les variables X^4, X^5, X^6 sont très bien expliquées par C^1 et C^2.

De plus, comme l'angle formé par les vecteurs $\overrightarrow{OX}_4$ et $\overrightarrow{OX}_5$ est proche de 90°, les variables X^4 et X^5 sont très faiblement corrélées.

2.4.2 Nombre d'axes (ou de composantes) à analyser

Combien d'axes analyser ? Il existe plusieurs critères de décision.

Le premier (Kaiser) veut qu'on ne s'intéresse en général qu'aux axes dont les valeurs propres sont supérieures à la moyenne (qui vaut 1 en ACP normée).

Un second (dit du coude, ou de Cattell) utilise le résultat suivant : lorsque des variables sont peu corrélées, les valeurs propres de la matrice d'inertie décroissent régulièrement - et l'ACP présente alors peu d'intérêt. A l'inverse, lorsqu'il existe une structure sur les données, on observe des ruptures dans la décroissance des valeurs propres (cf. Fig.2.5). On cherchera donc à ne retenir que les axes correspondant aux valeurs qui précèdent la décroissance régulière. Analytiquement, cela revient à chercher un point d'inflexion dans la décroissance des valeurs propres, et de ne pas aller au-delà dans l'analyse.

Ainsi, dans notre exemple, on ne s'intéressera qu'aux 2 premiers axes.

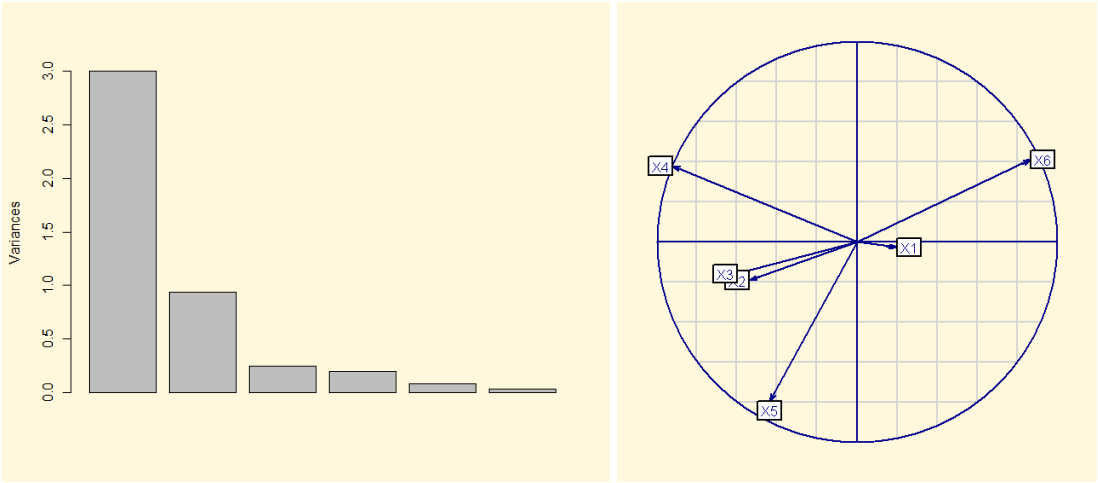


FIGURE 2.5 – Représentation des valeurs propres et cercle des corrélations pour le premier plan factoriel

2.4.3 Aides à l'interprétation*

Si, pour les variables numériques, la visualisation des vecteurs à l'intérieur du cercle des corrélations donne toute l'information nécessaire à l'analyse, il peut être utile de définir, pour chaque individu, les aides suivantes :

– La contribution à l'inertie du nuage, qui croît avec l'excentricité de l'individu :

$$CTR\,(X_i) = \frac{\|X_i\|^2}{I}$$

– La contribution à l'inertie portée par un axe (O, u_α) :

$$CTR_\alpha(X_i) = \frac{\left(C_i^\alpha\right)^2}{\lambda_\alpha}$$

Par construction : $\sum_{i=1}^n CTR(X_i) = 1$, et $\sum_{i=1}^n CTR_\alpha(X_i) = 1$. La valeur de ces contributions dépend donc fortement du nombre d'individus : une contribution de 5% sera considérée comme forte si l'on manipule les données de milliers d'individus, nettement moins si l'on n'en a qu'une vingtaine (de façon générale, on considèrera que l'individu i a une contribution importante si elle dépasse son poids $1/n$).

– La qualité de projection sur l'axe (O, u_α) est donnée par le carré du cosinus de l'angle :

$$CO2_\alpha(X_i) = \frac{\left(C_i^\alpha\right)^2}{\|X_i\|^2}.$$

Par orthogonalité des u_α, la qualité de projection d'un individu sur un sous-espace principal est additive : $CO2_{\alpha+\beta}(X_i) = CO2_\alpha(X_i) + CO2_\beta(X_i)$. D'autre part, on remarque que $\sum_{\alpha=1}^p CO2_\alpha(X_i) = 1$; de même que précédemment, cette qualité dépend fortement du nombre initial de variables : on pourra être exigeant si l'on n'en manipule qu'une poignée, on le sera moins s'il y en a davantage.

Pour un axe donné, l'examen parallèle des CTR et des $CO2$ des individus qui s'y projettent peut donner lieu à quatre cas de figure, dont un pose problème ($CO2$ faible-CTR forte), qui apparaît lorsqu'un individu a un poids trop fort par rapport aux autres :

	CTR faible	CTR forte
$CO2$ faible	Elément peu contributif quasi indépendant de l'axe	Elément très contributif mais peu illustratif de l'axe
$CO2$ forte	Elément peu contributif mais bien illustratif de l'axe	Elément particulièrement caractéristique de l'axe

2.5 Résumé du Chapitre 2

Tableau de données multivariées :

– variables :

– individus :

Matrice des corrélations :

Matrice de scatter-plots :

Représentation centrée :

Représentation centrée-réduite :

Analyse en Composantes Principales (ACP) :

- ACP normée :
- Inertie du nuage :
- Inertie autour d'un sous-espace :
- Composantes principales :

Représentation graphiques dérivées de l'ACP :

- Projection des individus :
- Scree-graph :
- Projection des variables :

X_1	X_2	X_3	X_4	X_5	X_6
214.8	131	131.1	9	9.7	141
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142
215	129.6	129.7	10.4	7.7	141.8
215.7	130.8	130.5	9	10.1	141.4
215.5	129.5	129.7	7.9	9.6	141.6
214.5	129.6	129.2	7.2	10.7	141.7
214.9	129.4	129.7	8.2	11	141.9
215.2	130.4	130.3	9.2	10	140.7
215.3	130.4	130.3	7.9	11.7	141.8
215.1	129.5	129.6	7.7	10.5	142.2
215.2	130.8	129.6	7.9	10.8	141.4
214.7	129.7	129.7	7.7	10.9	141.7
215.1	129.9	129.7	7.7	10.8	141.8
214.5	129.8	129.8	9.3	8.5	141.6
214.6	129.9	130.1	8.2	9.8	141.7
215	129.9	129.7	9	9	141.9
215.2	129.6	129.6	7.4	11.5	141.5
214.7	130.2	129.9	8.6	10	141.9
215	129.9	129.3	8.4	10	141.4
215.6	130.5	130	8.1	10.3	141.6
215.3	130.6	130	8.4	10.8	141.5
215.7	130.2	130	8.7	10	141.6
215.1	129.7	129.9	7.4	10.8	141.1
215.3	130.4	130.4	8	11	142.3
215.5	130.2	130.1	8.9	9.8	142.4
215.1	130.3	130.3	9.8	9.5	141.9
215.1	130	130	7.4	10.5	141.8
214.8	129.7	129.3	8.3	9	142
215.2	130.1	129.8	7.9	10.7	141.8
214.8	129.7	129.7	8.6	9.1	142.3
215	130	129.6	7.7	10.5	140.7
215.6	130.4	130.1	8.4	10.3	141
215.9	130.4	130	8.9	10.6	141.4
214.6	130.2	130.2	9.4	9.7	141.8
215.5	130.3	130	8.4	9.7	141.8
215.3	129.9	129.4	7.9	10	142
215.3	130.3	130.1	8.5	9.3	142.1
213.9	130.3	129	8.1	9.7	141.3
214.4	129.8	129.2	8.9	9.4	142.3
214.8	130.1	129.6	8.8	9.9	140.9
214.9	129.6	129.4	9.3	9	141.7
214.9	130.4	129.7	9	9.8	140.9
214.8	129.4	129.1	8.2	10.2	141
214.3	129.5	129.4	8.3	10.2	141.8
214.8	129.9	129.7	8.3	10.2	141.5
214.8	129.9	129.7	7.3	10.9	142
214.6	129.7	129.8	7.9	10.3	141.1
214.5	129	129.6	7.8	9.8	142

X_1	X_2	X_3	X_4	X_5	X_6
214.6	129.8	129.4	7.2	10	141.3
215.3	130.6	130	9.5	9.7	141.1
214.5	130.1	130	7.8	10.9	140.9
215.4	130.2	130.2	7.6	10.9	141.6
214.5	129.4	129.5	7.9	10	141.4
215.2	129.7	129.4	9.2	9.4	142
215.7	130	129.4	9.2	10.4	141.2
215	129.6	129.4	8.8	9	141.1
215.1	130.1	129.9	7.9	11	141.3
215.1	130	129.8	8.2	10.3	141.4
215.1	129.6	129.3	8.3	9.9	141.6
215.3	129.7	129.4	7.5	10.5	141.5
215.4	129.8	129.4	8	10.6	141.5
214.5	130	129.5	8	10.8	141.4
215	130	129.8	8.6	10.6	141.5
215.2	130.6	130	8.8	10.6	140.8
214.6	129.5	129.2	7.7	10.3	141.3
214.8	129.7	129.3	9.1	9.5	141.5
215.1	129.6	129.8	8.6	9.8	141.8
214.9	130.2	130.2	8	11.2	139.6
213.8	129.8	129.5	8.4	11.1	140.9
215.2	129.9	129.5	8.2	10.3	141.4
215	129.6	130.2	8.7	10	141.2
214.4	129.9	129.6	7.5	10.5	141.8
215.2	129.9	129.7	7.2	10.6	142.1
214.1	129.6	129.3	7.6	10.7	141.7
214.9	129.9	130.1	8.8	10	141.2
214.6	129.8	129.4	7.4	10.6	141
215.2	130.5	129.8	7.9	10.9	140.9
214.6	129.9	129.4	7.9	10	141.8
215.1	129.7	129.7	8.6	10.3	140.6
214.9	129.8	129.6	7.5	10.3	141
215.2	129.7	129.1	9	9.7	141.9
215.2	130.1	129.9	7.9	10.8	141.3
215.4	130.7	130.2	9	11.1	141.2
215.1	129.9	129.6	8.9	10.2	141.5
215.2	129.9	129.7	8.7	9.5	141.6
215	129.6	129.2	8.4	10.2	142.1
214.9	130.3	129.9	7.4	11.2	141.5
215	129.9	129.7	8	10.5	142
214.7	129.7	129.3	8.6	9.6	141.6
215.4	130	129.9	8.5	9.7	141.4
214.9	129.4	129.5	8.2	9.9	141.5
214.5	129.5	129.3	7.4	10.7	141.5
214.7	129.6	129.5	8.3	10	142
215.6	129.9	129.9	9	9.5	141.7
215	130.4	130.3	9.1	10.2	141.1
214.4	129.7	129.5	8	10.3	141.2
215.1	130	129.8	9.1	10.2	141.5
214.7	130	129.4	7.8	10	141.2

TABLE 2.1 – Les données de billets suisses **authentiques**. Le tableau comprend 100 lignes (individus) et 6 colonnes (variables). Ces variables sont décrites dans la Fig. 2.1. Toutes les valeurs sont en mm.

X_1	X_2	X_3	X_4	X_5	X_6
214.4	130.1	130.3	9.7	11.7	139.8
214.9	130.5	130.2	11	11.5	139.5
214.9	130.3	130.1	8.7	11.7	140.2
215	130.4	130.6	9.9	10.9	140.3
214.7	130.2	130.3	11.8	10.9	139.7
215	130.2	130.2	10.6	10.7	139.9
215.3	130.3	130.1	9.3	12.1	140.2
214.8	130.1	130.4	9.8	11.5	139.9
215	130.2	129.9	10	11.9	139.4
215.2	130.6	130.8	10.4	11.2	140.3
215.2	130.4	130.3	8	11.5	139.2
215.1	130.5	130.3	10.6	11.5	140.1
215.4	130.7	131.1	9.7	11.8	140.6
214.9	130.4	129.9	11.4	11	139.9
215.1	130.3	130	10.6	10.8	139.7
215.5	130.4	130	8.2	11.2	139.2
214.7	130.6	130.1	11.8	10.5	139.8
214.7	130.4	130.1	12.1	10.4	139.9
214.8	130.5	130.2	11	11	140
214.4	130.2	129.9	10.1	12	139.2
214.8	130.3	130.4	10.1	12.1	139.6
215.1	130.6	130.3	12.3	10.2	139.6
215.3	130.8	131.1	11.6	10.6	140.2
215.1	130.7	130.4	10.5	11.2	139.7
214.7	130.5	130.5	9.9	10.3	140.1
214.9	130	130.3	10.2	11.4	139.6
215	130.4	130.4	9.4	11.6	140.2
215.5	130.7	130.3	10.2	11.8	140
215.1	130.2	130.2	10.1	11.3	140.3
214.5	130.2	130.6	9.8	12.1	139.9
214.3	130.2	130	10.7	10.5	139.8
214.5	130.2	129.8	12.3	11.2	139.2
214.9	130.5	130.2	10.6	11.5	139.9
214.6	130.2	130.4	10.5	11.8	139.7
214.2	130	130.2	11	11.2	139.5
214.8	130.1	130.1	11.9	11.1	139.5
214.6	129.8	130.2	10.7	11.1	139.4
214.9	130.7	130.3	9.3	11.2	138.3
214.6	130.4	130.4	11.3	10.8	139.8
214.5	130.5	130.2	11.8	10.2	139.6
214.8	130.2	130.3	10	11.9	139.3
214.7	130	129.4	10.2	11	139.2
214.6	130.2	130.4	11.2	10.7	139.9
215	130.5	130.4	10.6	11.1	139.9
214.5	129.8	129.8	11.4	10	139.3
214.9	130.6	130.4	11.9	10.5	139.8
215	130.5	130.4	11.4	10.7	139.9
215.3	130.6	130.3	9.3	11.3	138.1
214.7	130.2	130.1	10.7	11	139.4
214.9	129.9	130	9.9	12.3	139.4

X_1	X_2	X_3	X_4	X_5	X_6
214.9	130.3	129.9	11.9	10.6	139.8
214.6	129.9	129.7	11.9	10.1	139
214.6	129.7	129.3	10.4	11	139.3
214.5	130.1	130.1	12.1	10.3	139.4
214.5	130.3	130	11	11.5	139.5
215.1	130	130.3	11.6	10.5	139.7
214.2	129.7	129.6	10.3	11.4	139.5
214.4	130.1	130	11.3	10.7	139.2
214.8	130.4	130.6	12.5	10	139.3
214.6	130.6	130.1	8.1	12.1	137.9
215.6	130.1	129.7	7.4	12.2	138.4
214.9	130.5	130.1	9.9	10.2	138.1
214.6	130.1	130	11.5	10.6	139.5
214.7	130.1	130.2	11.6	10.9	139.1
214.3	130.3	130	11.4	10.5	139.8
215.1	130.3	130.6	10.3	12	139.7
216.3	130.7	130.4	10	10.1	138.8
215.6	130.4	130.1	9.6	11.2	138.6
214.8	129.9	129.8	9.6	12	139.6
214.9	130	129.9	11.4	10.9	139.7
213.9	130.7	130.5	8.7	11.5	137.8
214.2	130.6	130.4	12	10.2	139.6
214.8	130.5	130.3	11.8	10.5	139.4
214.8	129.6	130	10.4	11.6	139.2
214.8	130.1	130	11.4	10.5	139.6
214.9	130.4	130.2	11.9	10.7	139
214.3	130.1	130.1	11.6	10.5	139.7
214.5	130.4	130	9.9	12	139.6
214.8	130.5	130.3	10.2	12.1	139.1
214.5	130.2	130.4	8.2	11.8	137.8
215	130.4	130.1	11.4	10.7	139.1
214.8	130.6	130.6	8	11.4	138.7
215	130.5	130.1	11	11.4	139.3
214.6	130.5	130.4	10.1	11.4	139.3
214.7	130.2	130.1	10.7	11.1	139.5
214.7	130.4	130	11.5	10.7	139.4
214.5	130.4	130	8	12.2	138.5
214.8	130	129.7	11.4	10.6	139.2
214.8	129.9	130.2	9.6	11.9	139.4
214.6	130.3	130.2	12.7	9.1	139.2
215.1	130.2	129.8	10.2	12	139.4
215.4	130.5	130.6	8.8	11	138.6
214.7	130.3	130.2	10.8	11.1	139.2
215	130.5	130.3	9.6	11	138.5
214.9	130.3	130.5	11.6	10.6	139.8
215	130.4	130.3	9.9	12.1	139.6
215.1	130.3	129.9	10.3	11.5	139.7
214.8	130.3	130.4	10.6	11.1	140
214.7	130.7	130.8	11.2	11.2	139.4
214.3	129.9	129.9	10.2	11.5	139.6

TABLE 2.2 – Les données de billets suisses **contrefaits**. Le tableau comprend 100 lignes (individus) et 6 colonnes (variables). Ces variables sont décrites dans la Fig. 2.1. Toutes les valeurs sont en mm.

3

Rappel des bases de la statistique paramétrique

Dans ce chapitre, nous survolons rapidement les bases du calcul des probabilités et de la statistique. Toutes les notions et tous les résultats présentés ci-dessous constituent les prérequis pour ce cours de « Statistique numérique et analyse des données ». Pour une présentation plus détaillée des sujets traités dans ce chapitre voir le polycopié du cours de 1ère année [1]

3.1 Introduction

Les problèmes statistiques que nous allons étudier dans le cadre de ce module peuvent se résumer de la façon suivante : nous disposons d'un jeu de données qui sont supposées être générées par un phénomène aléatoire. (Rappelons que tout phénomène aléatoire est entièrement caractérisé par sa loi de probabilité.) De plus, nous considérons qu'un travail de modélisation a été effectué à l'issue duquel la loi de probabilité régissant les données a été déterminée à un paramètre inconnu près. Dans ce contexte, les trois types de problèmes que nous allons étudier sont :

estimation : trouver une valeur approchée du paramètre inconnu,

région de confiance : déterminer une région (aussi petite que possible) qui contient le paramètre inconnu avec une probabilité prescrite (généralement 95%),

test d'hypothèse : pour un ensemble Θ_0 de valeurs possibles du paramètre inconnu, décider au vu des données si oui ou non le paramètre inconnu appartient à Θ_0.

Afin de faciliter la compréhension, les différentes notions introduites dans ce chapitre seront illustrées dans les deux exemples suivants.

Exemple 1. (Qualité de l'air) On cherche à évaluer la fréquence des jours où l'indice ATMO (mesurant la qualité de l'air) à Paris dépasse [2] le niveau 8. Pour avoir une estimation simple, on choisit au hasard n jours dans le passé et on regarde si oui ou non le niveau 8 a été dépassé ces jours-là. On obtient ainsi un échantillon $x_1, \ldots, x_n$ où chaque x_i prend

1. B. Jourdain, Probabilités et statistique, `http://cermics.enpc.fr/~jourdain/probastat/poly.pdf`.

2. L'indice ATMO varie sur une échelle allant de 1 (très bonne) à 10 (exécrable). Lorsque la valeur de cet indice dépasse le niveau 8, la qualité de l'air est considérée comme mauvaise.

deux valeurs : 0 ou 1. Par convention, la valeur 0 correspond à un jour où le niveau 8 n'a pas été dépassé. Nous modélisons le dépassement du niveau 8 par l'indice ATMO par une variable aléatoire X de loi de Bernoulli ;

$$\textbf{Proba}(X = 1) = \vartheta^*, \qquad \textbf{Proba}(X = 0) = 1 - \vartheta^*$$

pour une valeur $\vartheta^* \in]0, 1[$ qui nous est inconnue. Cette valeur représente la fréquence moyenne des jours où la qualité de l'air a été mauvaise à Paris.

Exemple 2. (Vitesse du vent) Afin d'étudier la possibilité de l'installation d'une centrale éolienne sur un site donné, on cherche à estimer la probabilité de l'événement « la vitesse du vent sur le site en question est inférieure à 10km/h ». L'intérêt à l'égard de cet événement vient du fait que, lorsque la vitesse du vent est inférieure à 10km/h, une centrale éolienne s'arrête en raison des forces de frottement sec qui s'opposent à la rotation de l'hélice. L'approche la plus simple consiste à modéliser la vitesse du vent sur le site en question à un instant donné par une variable aléatoire de loi exponentielle. En d'autres termes, si X représente la vitesse du vent, on suppose que

$$\textbf{Proba}(X \in [a, b]) = \int_a^b p(\vartheta^*; x)\, dx, \qquad p(\vartheta^*; x) = \frac{1}{\vartheta^*}\, e^{-x/\vartheta^*} \mathbb{1}_{[0,\infty[}(x),$$

où $\vartheta^* > 0$ est un paramètre inconnu. Si l'on admet que cette modélisation est correcte, on peut calculer la probabilité de l'événement $A =$ « la vitesse du vent est inférieure à 10km/h » par la formule

$$\textbf{Proba}(A) = \int_0^{10} \frac{1}{\vartheta^*} e^{-x/\vartheta^*}\, dx = 1 - e^{-10/\vartheta^*}.$$

Par conséquent, une valeur approchée de ϑ^* nous permettrait de calculer une valeur approchée de la probabilité de l'événement A. Pour pouvoir estimer ϑ^*, nous mesurons la vitesse du vent à n instants suffisamment espacés dans le temps, ce qui nous fournit les observations $x_1, \ldots, x_n$. Le but d'un statisticien est, entre autre, d'utiliser ces observations pour estimer le paramètre ϑ^*.

3.2 Modèle statistique

Nous commençons par donner la définition générale d'un modèle statistique, que nous illustrons par la suite dans les deux exemples présentés ci-dessus.

Définition 3.1. *On appelle modèle statistique la donnée d'un espace mesurable $(\mathcal{X}_n, \mathcal{F}_n)$ et d'une famille de mesures de probabilité $\mathcal{P}_n = \{\boldsymbol{P}_{n,\vartheta}, \vartheta \in \Theta\}$ définies sur $(\mathcal{X}_n, \mathcal{F}_n)$. L'espace $\mathcal{X}_n$, appelé espace d'états, est constitué de toutes les valeurs qu'aurait pu prendre le jeu de données étudié. La famille $\mathcal{P}_n$ décrit l'ensemble des lois de probabilité pouvant avoir généré le jeu de données étudié.*

Pour un modèle statistique donné, la problématique générale de la théorie statistique s'énonce de la façon suivante : au vu d'une réalisation $x^{(n)} \in \mathcal{X}_n$ tiré au hasard selon une loi $\boldsymbol{P}_{n,\vartheta^*} \in \mathcal{P}$, étudier certaines propriétés de $\boldsymbol{P}_{n,\vartheta^*}$. Le plus souvent $x^{(n)}$ est un vecteur. On cherche donc à caractériser la loi d'un vecteur aléatoire $X^{(n)}$ à partir d'une réalisation $x^{(n)}$. Bien-entendu, si l'on autorise la famille $\mathcal{P}_n$ à être une collection quelconque de lois sur $(\mathcal{X}_n, \mathcal{F}_n)$, la tâche de l'extraction de l'information fiable sur la loi du vecteur aléatoire $X^{(n)}$ à partir d'une seule réalisation est irréalisable. Afin de pouvoir élaborer une théorie raisonnable et utile pour les

applications, on se restreint au cas où la famille $\mathscr{P}_n$ a une certaine «structure». Exemples de telles structures sont le modèle à observations i.i.d., le modèle de régression linéaire, etc.

Ce chapitre est entièrement dédié à l'étude du modèle à observations i.i.d. (indépendantes et identiquement distribuées). Il s'agit du cas où $x^{(n)} = (x_1, \dots, x_n)$ est un vecteur dans $\mathbb{R}^n$ dont les coordonnées représentent n copies indépendantes d'une même variable aléatoire X. Cela revient à postuler que $x^{(n)}$ est une réalisation du vecteur aléatoire $X^{(n)} = (X_1, \dots, X_n)$ composé de n variables aléatoires indépendantes distribuées selon la même loi que X. Dans ce cas, la loi de $X^{(n)}$ est entièrement caractérisée par celle de X, car

$$\begin{aligned}\mathbf{Proba}(X_1 \in A_1, \dots, X_n \in A_n) &= \mathbf{Proba}(X_1 \in A_1) \cdot \ldots \cdot \mathbf{Proba}(X_n \in A_n) \\ &= \mathbf{Proba}(X \in A_1) \cdot \ldots \cdot \mathbf{Proba}(X \in A_n)\end{aligned}$$

quels que soient les intervalles $A_1, \dots, A_n \subset \mathbb{R}$. Si $\boldsymbol{P}$ désigne la loi de X, on dit alors que $X^{(n)}$ est un échantillon i.i.d. de loi $\boldsymbol{P}$. Par conséquent, pour définir un modèle à observations i.i.d., il suffit de décrire la famille $\mathscr{P} = \{\boldsymbol{P}_\vartheta\}$ qui est sensée contenir la loi $\boldsymbol{P}$ de X. Les deux exemples présentés dans l'introduction correspondent à des modèles à observations i.i.d. : dans le premier exemple $\mathscr{P} = \{\mathcal{B}(\vartheta) : \vartheta \in]0,1[\}$ où $\mathcal{B}(\vartheta)$ désigne la loi de Bernoulli de paramètre ϑ, tandis que dans le deuxième exemple $\mathscr{P} = \{\mathcal{E}(\vartheta^{-1}) : \vartheta > 0\}$, où $\mathcal{E}(\lambda)$ désigne la loi exponentielle de paramètre $\lambda > 0$. En conséquence, dans le premier exemple $\Theta = [0,1]$ alors que dans le deuxième exemple $\Theta =]0, \infty[$.

Tout au long de ce chapitre, on appellera statistique toute fonction de l'échantillon $X^{(n)}$.

3.3 Estimation

Supposons maintenant qu'on dispose d'un échantillon i.i.d. $X_1, \dots, X_n$ de loi $\boldsymbol{P} \in \mathscr{P} = \{\boldsymbol{P}_\vartheta : \vartheta \in \Theta\}$. Cela veut dire que pour un $\vartheta^* \in \Theta$ inconnu, on a

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \boldsymbol{P}_{\vartheta^*}.$$

Par la suite, on appellera ϑ^* la vraie valeur du paramètre. La première question qu'on se pose est celle du calcul d'une valeur approchée de ϑ^* en utilisant uniquement l'échantillon observé.

Définition 3.2. *Soit $X_1, \dots, X_n$ un échantillon i.i.d. de loi $\boldsymbol{P} \in \mathscr{P} = \{\boldsymbol{P}_\vartheta : \vartheta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^p$ pour un $p \in \mathbb{N}$. On appelle estimateur de ϑ^* toute application mesurable*

$$\bar{\vartheta} : \mathbb{R}^n \to \mathbb{R}^p.$$

Dans la statistique théorique, on identifie l'application $\bar{\vartheta}$ au vecteur aléatoire $\bar{\vartheta}(X_1, \dots, X_n)$.

Un estimateur a pour objectif d'approcher la vraie valeur ϑ^*. Cependant, la définition ci-dessus ne reflète absolument pas cet objectif. En effet, même si $\bar{\vartheta}(X_1, \dots, X_n)$ est très éloigné de ϑ^*, $\bar{\vartheta}$ sera appelé un estimateur si peu qu'il soit mesurable. Afin de restreindre la classe de tous les estimateurs à ceux qui représentent un intérêt pratique, on spécifie des propriétés qu'on aimerait voir satisfaites par un estimateur. Par la suite, pour souligner le fait que l'estimateur $\bar{\vartheta}$ dépend de n (la taille de l'échantillon), on utilisera la notation $\bar{\vartheta}_n$.

Définition 3.3. *On dit que l'estimateur $\bar{\vartheta}_n$ est sans biais, si*

$$\mathbf{E}_{\vartheta^*}[\bar{\vartheta}_n] = \vartheta^*, \qquad \forall\, \vartheta^* \in \Theta,$$

où l'expression $\mathbf{E}_{\vartheta^*}[\bar{\vartheta}_n]$ *doit être lue comme « espérance du vecteur aléatoire* $\bar{\vartheta}_n(X_1,\dots,X_n)$ *sachant que* $X_1,\dots,X_n \overset{iid}{\sim} \boldsymbol{P}_{\vartheta^*}$ *». On dit que l'estimateur* $\bar{\vartheta}_n$ *est convergent (ou consistant), s'il converge en probabilité vers la vraie valeur, c'est-à-dire*

$$\lim_{n\to\infty} \boldsymbol{P}_{\vartheta^*}(|\bar{\vartheta}_n - \vartheta^*| > \varepsilon) = 0, \qquad \forall \varepsilon > 0, \quad \forall \vartheta^* \in \Theta.$$

La propriété de convergence est centrale en statistique, car elle indique que la valeur estimée de ϑ^* calculée à l'aide de l'estimateur $\bar{\vartheta}$ est proche de ϑ^* si la taille n de l'échantillon est suffisamment grande. Dans beaucoup de situations, il existe de nombreux estimateurs convergents. On s'intéresse alors aux propriétés plus raffinées des estimateurs : la vitesse à laquelle $\bar{\vartheta}_n$ tend vers ϑ^* et la loi asymptotique de la différence $\bar{\vartheta}_n - \vartheta^*$ proprement normalisée.

Définition 3.4. *On dit que l'estimateur convergent* $\bar{\vartheta}_n$ *est asymptotiquement de loi* $\boldsymbol{P}^{\infty}_{\vartheta}$ *avec la vitesse* $n^{-\gamma}$*, où* $\gamma > 0$*, si*

$$n^{\gamma}(\bar{\vartheta}_n - \vartheta^*) \overset{\mathscr{L}}{\to} \boldsymbol{P}^{\infty}_{\vartheta^*}, \qquad \forall \vartheta^* \in \Theta,$$

où $\overset{\mathscr{L}}{\to}$ *désigne la convergence en loi. Si* $\boldsymbol{P}^{\infty}_{\vartheta^*}$ *est la loi gaussienne* $\mathcal{N}(0, \sigma^2_{\vartheta^*})$*, on dit alors que* $\bar{\vartheta}_n$ *est asymptotiquement normal avec la vitesse* $n^{-\gamma}$ *et la variance limite* $\sigma^2_{\vartheta^*}$*.*

Pour démontrer la convergence et la normalité asymptotique des estimateurs, on utilise le plus souvent les résultats probabilistes présentés dans le paragraphe suivant.

3.3.1 Quelques résultats sur la convergence des variables aléatoires

Soit $\xi_1, \xi_2, \dots, \xi_n, \dots$ et ξ_∞ des variables aléatoires et soit $F_{\xi_n}(x) = \boldsymbol{P}(\xi_n \le x)$ la fonction de répartition de ξ_n, $n \in \mathbb{N} \cup \{\infty\}$. On distingue les quatre types de convergence (de $\{\xi_n\}$ vers ξ_∞) suivants :

1. convergence en probabilité : pour tout $\varepsilon > 0$, on a $\lim_{n\to\infty} \boldsymbol{P}(|\xi_n - \xi_\infty| > \varepsilon) = 0$,
2. convergence presque sûr : $\boldsymbol{P}(\limsup_{n\to\infty} |\xi_n - \xi_\infty| = 0) = 1$,
3. convergence en moyenne quadratique : $\lim_{n\to\infty} \mathbf{E}[(\xi_n - \xi_\infty)^2] = 0$,
4. convergence en loi : $\lim_{n\to\infty} F_{\xi_n}(x) = F_{\xi_\infty}(x)$ pour tout $x \in \mathbb{R}$ tel que F_{ξ_∞} est continue en x.

Rappelons que les convergences presque sûr et en moyenne quadratique entraînent la convergence en probabilité et cette dernière entraîne à son tour la convergence en loi. Notons aussi que la définition de la convergence en loi, contrairement aux autres types de convergences précitées, ne sous-entend pas que les variables ξ_n soient définies sur le même espace probabilisé.

Théorème 3.1 (Loi forte des grands nombres)**.** *Soit* $X_1,\dots,X_n$ *des variables aléatoires i.i.d. intégrables :* $\mathbf{E}[|X_1|] < \infty$*. Alors,*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{p.s.}{\longrightarrow} \mathbf{E}[X_1], \qquad \textit{lorsque} \quad n \to \infty,$$

où $\overset{p.s.}{\to}$ *désigne la convergence presque-sûr.*

Théorème 3.2 (Théorème de la limite centrale)**.** *Soit* $X_1,\dots,X_n$ *des variables aléatoires i.i.d. de carré intégrables :* $\mathbf{E}[X_1^2] < \infty$*. Alors,*

$$\sqrt{n}(\overline{X}_n - \mathbf{E}[X_1]) \overset{\mathscr{L}}{\longrightarrow} \mathcal{N}(0, \mathbf{Var}[X_1]), \qquad \textit{lorsque} \quad n \to \infty.$$

Théorème 3.3 (Méthode delta). *Soit $X_1, \dots, X_n$ des variables aléatoires i.i.d. de carré intégrables et soit G une fonction continûment différentiable sur un ensemble ouvert A tel que $\boldsymbol{P}(X_1 \in A) = 1$. Alors,*

$$\sqrt{n}\big(G(\overline{X}_n) - G(\mathbf{E}[X_1])\big) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \sigma^2), \qquad \textit{lorsque} \quad n \to \infty,$$

avec la variance limite $\sigma^2 = G'(\mathbf{E}[X_1])^2 \mathbf{Var}[X_1]$.

Ces résultats se généralisent à une suite de vecteurs aléatoires, auquel cas la variance est remplacée par la matrice de covariance $\mathbf{Var}[X_1] = \mathbf{E}[X_1 X_1^\top] - \mathbf{E}[X_1]\mathbf{E}[X_1^\top]$ et la variance limite dans la méthode delta est donnée par $\sigma^2 = \nabla G(\mathbf{E}[X_1])^\top \mathbf{Var}[X_1] \nabla G(\mathbf{E}[X_1])$.

Théorème 3.4 (Théorème de Slutsky). *Soit $\{\xi_n\}_{n\in\mathbb{N}}$ $\{\eta_n\}_{n\in\mathbb{N}}$ deux suites de variables aléatoires définies sur le même espace probabilisé. Si pour une constante $a \in \mathbb{R}$ et pour une variable aléatoire ξ_∞ on a*

$$\xi_n \xrightarrow[n\to\infty]{\mathscr{L}} \xi_\infty, \quad \textit{et} \quad \eta_n \xrightarrow[n\to\infty]{P} a$$

alors

$$\xi_n + \eta_n \xrightarrow[n\to\infty]{\mathscr{L}} \xi_\infty + a, \quad \textit{et} \quad \xi_n \eta_n \xrightarrow[n\to\infty]{\mathscr{L}} a\xi_\infty.$$

3.3.2 Estimateur du maximum de vraisemblance

Après avoir vu ce que c'est un estimateur et quelles sont les propriétés souhaitées d'un estimateur, on s'intéresse naturellement à la mise en place d'une procédure générique permettant la construction d'un estimateur pour une large classe de modèles. On se focalise ici sur la méthode d'estimation la plus utilisée : le maximum de vraisemblance. De plus, pour éviter le rappel de notions abstraites (absolue continuité, théorème de Radon-Nykodim) de la théorie de la mesure, on ne donnera pas la définition de l'estimateur du maximum de vraisemblance (EMV) dans le cas le plus général des modèles dominés, mais seulement dans le cadre des modèles i.i.d. discrets et à densité.

Définition 3.5. *On dira que le modèle à observations i.i.d. $\{\boldsymbol{P}_\vartheta : \vartheta \in \Theta\}$ est discret, s'il existe un ensemble $A = \{a_1, a_2, \dots\}$ au plus dénombrable tel que $\boldsymbol{P}_\vartheta(A) = 1$ pour tout $\vartheta \in \Theta$. En d'autres termes, l'ensemble A contient toutes les valeurs possibles prises par les variables de l'échantillon.*

L'exemple 1 considéré au début de ce chapitre porte sur un modèle discret, car les variables aléatoires constituant l'échantillon sont des variables de Bernoulli et, par conséquent, prennent leurs valeurs dans l'ensemble fini $\{0, 1\}$.
On caractérise un modèle discret par les probabilités discrètes

$$p(\vartheta; a_k) = \mathbf{Proba}(X_i = a_k),\ \forall a_k \in A \qquad \text{où } X_1, \dots, X_n \overset{\text{iid}}{\sim} \boldsymbol{P}_\vartheta. \tag{3.1}$$

Définition 3.6. *On dira que le modèle à observations i.i.d. $\{\boldsymbol{P}_\vartheta : \vartheta \in \Theta\}$ est à densité, si pour tout $\vartheta \in \Theta$ il existe une fonction (appelée densité) $p(\vartheta; \cdot) : \mathbb{R} \to \mathbb{R}$ telle que*

$$\boldsymbol{P}_\vartheta([a, b]) = \mathbf{Proba}(X_i \in [a, b]) = \int_a^b p(\vartheta; x)\, dx, \qquad \textit{où } X_i \sim \boldsymbol{P}_\vartheta, \tag{3.2}$$

pour tout $a, b \in \mathbb{R}$.

Définition 3.7. *Soit $\mathscr{P} = \{P_\vartheta : \vartheta \in \Theta\}$ un modèle i.i.d. discret ou à densité et soit $p(\vartheta, x)$ la fonction définie par (3.1) dans le cas discret et par (3.2) dans le cas à densité. On appelle fonction de vraisemblance l'application*

$$p_n : \Theta \times \mathbb{R}^n \to \mathbb{R}_+, \qquad p_n(\vartheta; x_1, \dots, x_n) = \prod_{i=1}^n p(\vartheta; x_i). \tag{3.3}$$

On appelle estimateur du maximum de vraisemblance (EMV), noté $\hat{\vartheta}_n^{MV}$, le point du maximum global (s'il existe) de l'application $\vartheta \mapsto p_n(\vartheta, X_1, \dots, X_n)$. On écrit alors

$$\hat{\vartheta}_n^{MV} = \arg\max_{\vartheta \in \Theta} p_n(\vartheta; X_1, \dots, X_n).$$

3.3.3 L'EMV dans l'exemple 1

Dans l'exemple 1 portant sur la qualité de l'air, on dispose d'un échantillon i.i.d. $X_1, \dots, X_n$ de loi de Bernoulli $\mathcal{B}(\vartheta^*)$ avec $\vartheta^* \in \Theta =]0,1[$. Il s'agit d'un modèle discret avec $A = \{0,1\}$ et

$$p(\vartheta; x) = \begin{cases} \vartheta, & \text{si } x = 1, \\ 1 - \vartheta, & \text{si } x = 0. \end{cases}$$

On vérifie facilement que cela équivaut à

$$p(\vartheta; x) = \vartheta^x (1-\vartheta)^{1-x}, \qquad x \in \{0,1\}.$$

Par conséquent, la fonction de vraisemblance s'écrit comme

$$p_n(\vartheta; x_1, \dots, x_n) = \prod_{i=1}^n \vartheta^{x_i}(1-\vartheta)^{1-x_i} = \vartheta^{\sum_i x_i}(1-\vartheta)^{n-\sum_i x_i}.$$

On remarque d'abord que la fonction de vraisemblance est strictement positive sur $]0,1[$. Il en résulte qu'on peut remplacer le problème de maximisation de p_n par celui de maximisation de $l_n = \log p_n$:

$$\hat{\vartheta}_n^{\mathrm{MV}} = \arg\max_{\vartheta \in]0,1[} \log p_n(\vartheta; X_1, \dots, X_n) = \arg\max_{\vartheta \in]0,1[} \Big\{ n\overline{X} \log \vartheta + n(1-\overline{X}) \log(1-\vartheta) \Big\},$$

où $\overline{X} = \frac{1}{n}\sum_i X_i$. On vérifie aisément que la fonction

$$l_n(\vartheta) = n\overline{X} \log \vartheta + n(1-\overline{X}) \log(1-\vartheta),$$

appelée fonction de log-vraisemblance est strictement concave sur $]0,1[$ et que $\overline{X}$ est le seul point où la dérivée de l_n s'annule. Or, si la dérivée d'une fonction concave s'annule en un point alors c'est le point de maximum global. Il en découle que dans le modèle de Bernoulli

$$\hat{\vartheta}_n^{\mathrm{MV}} = \overline{X}.$$

Par la linéarité de l'espérance, on montre que cet estimateur est sans biais :

$$\mathbf{E}_\vartheta[\overline{X}] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}_\vartheta[X_i] = \frac{1}{n}\sum_{i=1}^n \vartheta = \vartheta, \quad \forall \vartheta \in [0,1].$$

De plus, c'est un estimateur consistant et asymptotiquement normal de vitesse $1/\sqrt{n}$ et de variance limite $\vartheta(1-\vartheta)$.

La courbe de la fonction de log-vraisemblance l_n pour trois échantillons i.i.d. de loi $\mathcal{B}(1/2)$ est représentée dans la Figure 3.1. On y voit clairement la nature aléatoire de l'estimateur du maximum de vraisemblance, qui est dû au fait que l'échantillon a été obtenu par un tirage aléatoire.

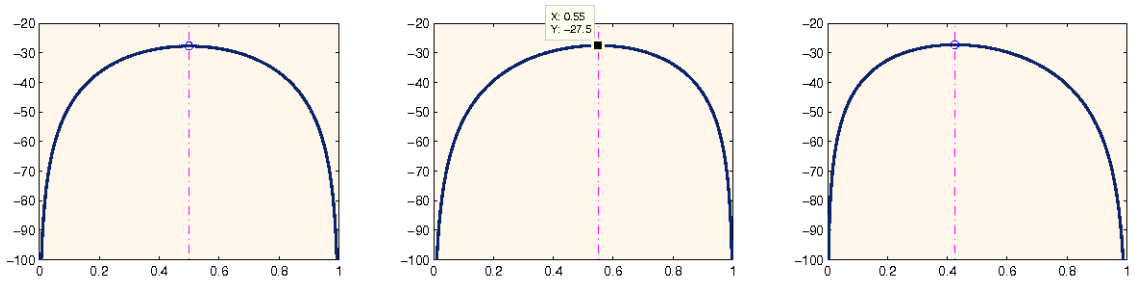


FIGURE 3.1 – Modèle de Bernoulli : la fonction de log-vraisemblance et son maximum global. Les trois courbes représentent la log-vraisemblance pour trois échantillons différents de taille 40. La vraie valeur du paramètre dans les trois cas est $\vartheta^* = 1/2$. Les valeurs estimées qu'on obtient pour ces échantillons sont $\hat{\vartheta}_n^{\mathrm{MV}} = 0.5\,;\, 0.55\,;\, 0.425$.

3.3.4 L'EMV dans l'exemple 2

Dans l'exemple 2 portant sur la vitesse du vent, on dispose d'un échantillon i.i.d. $X_1, \ldots, X_n$ de loi Exponentielle $\mathcal{E}(1/\vartheta^*)$ avec $\vartheta^* \in \Theta =]0, +\infty[$. Il s'agit d'un modèle à densité avec :

$$p(\vartheta; x) = \vartheta^{-1} e^{-x/\vartheta} \mathbb{1}_{[0,\infty[}(x).$$

On en déduit la fonction de vraisemblance

$$p_n(\vartheta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \vartheta^{-1} e^{-x_i/\vartheta} = \vartheta^{-n} \exp\Big\{ -\frac{1}{\vartheta} \sum_{i=1}^{n} x_i \Big\}$$

pour tout $x_1, \ldots, x_n \geq 0$. Comme on sait que l'échantillon $X_1, \ldots, X_n$ est généré par une loi exponentielle, $\boldsymbol{P}(X_i \geq 0; \forall i = 1, \ldots, n) = 1$. On a donc la fonction de log-vraisemblance

$$l_n(\vartheta) = -n(\log \vartheta + \vartheta^{-1} \overline{X}), \qquad \forall \vartheta > 0.$$

Cette fonction n'est pas concave sur $\mathbb{R}_+$, mais on vérifie aisément qu'elle est croissante sur $]0, \overline{X}]$ et décroissante sur $[\overline{X}, +\infty[$. Il en découle que $\overline{X}$ est le point de maximum global de l_n, ce qui entraîne que

$$\hat{\vartheta}_n^{\mathrm{MV}} = \overline{X}_n.$$

Comme dans l'exemple précédent, ici aussi l'estimateur $\overline{X}$ est sans biais. De plus, en vertu de la loi forte des grands nombres et du théorème de la limite centrale, $\overline{X}$ est consistant et asymptotiquement normal de vitesse $n^{-1/2}$ et de variance limite ϑ^{*2}, c'est-à-dire

$$\sqrt{n}(\overline{X} - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \vartheta^{*2}).$$

Remarque 3.1. *Dans les deux exemples précédents la méthode du maximum de vraisemblance nous a conduit à des estimateurs sans biais, consistants et asymptotiquement normaux de vitesse $n^{-1/2}$. On peut naturellement se demander si ces propriétés sont caractéristiques aux deux modèles considérés ou si elles restent valables dans un cadre plus général. Nous ne donnerons pas ici une réponse exhaustive à cette question, mais seulement quelques éléments de réponse :*

- *l'EMV n'est en général pas sans biais (on dit qu'il est biaisé), mais son biais tend vers zéro lorsque $n \to \infty$ sous certaines conditions de régularité ;*
- *il existe des conditions de régularité assez faibles sur l'application $(\vartheta, x) \mapsto p(\vartheta; x)$ garantissant la consistance de l'EMV ainsi que sa normalité asymptotique avec la vitesse $n^{-1/2}$.*

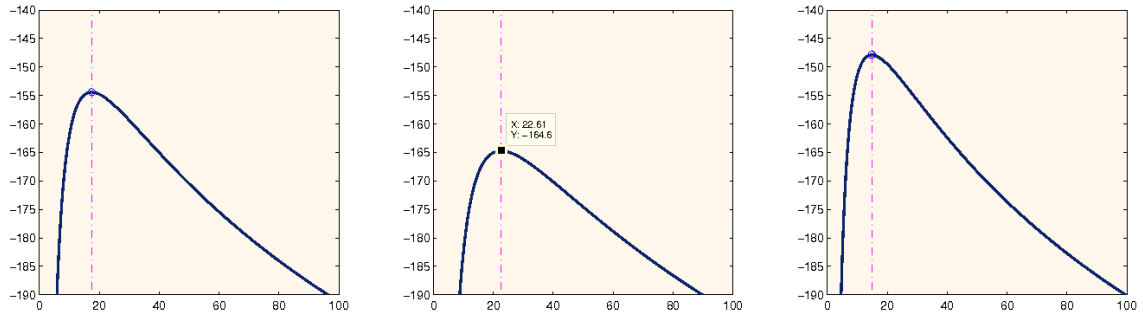


FIGURE 3.2 – Modèle exponentiel : la fonction de log-vraisemblance et son maximum global. Les trois courbes représentent la log-vraisemblance pour trois échantillons différents de taille 40. La vraie valeur du paramètre dans les trois cas est $\vartheta^* = 20$. Les valeurs estimées qu'on obtient pour ces échantillons sont $\hat{\vartheta}_n^{\mathrm{MV}} = 17.48\,;\, 22.61\,;\, 14.84$.

3.3.5 Un exemple de modèle irrégulier : modèle uniforme

Pour se convaincre que l'EMV n'est pas toujours sans biais et qu'il peut converger à une vitesse différente de $n^{-1/2}$, considérons le modèle suivant. On dispose d'un échantillon i.i.d. $X_1, \ldots, X_n$ de loi uniforme sur l'intervalle $[0, \vartheta^*]$, notée $\mathcal{U}([0, \vartheta^*])$. Le paramètre inconnu ϑ^* est supposé appartenir à l'ensemble $\mathbb{R}_+$. C'est un modèle à densité avec

$$p(\vartheta; x) = \frac{1}{\vartheta} \mathbb{1}_{[0,\vartheta]}(x).$$

Par conséquent, la fonction de vraisemblance a la forme

$$\begin{aligned} p_n(\vartheta; x_1, \ldots, x_n) &= \frac{1}{\vartheta^n} \begin{cases} 1, & \text{si } x_i \in [0, \vartheta] \ \forall i, \\ 0, & \text{sinon} \end{cases} \\ &= \vartheta^{-n} \mathbb{1}_{[x_{(n)}, +\infty[}(\vartheta), \end{aligned}$$

où $x_{(n)} = \max_{i=1,\ldots,n} x_i$. L'EMV est donc défini par

$$\hat{\vartheta}_n^{\mathrm{MV}} = \arg\max_{\vartheta > 0} \quad \vartheta^{-n} \mathbb{1}_{[X_{(n)}, +\infty[}(\vartheta) = X_{(n)} \quad (= \max_{1 \le i \le n} X_i).$$

Vérifions d'abord que $X_{(n)}$ est biaisé. Pour cela, on introduit l'événement

$$A = \{X_1 \le \vartheta^*/2; \ldots X_n \le \vartheta^*/2\}$$

qui vérifie $\boldsymbol{P}_{\vartheta^*}(A) = (1/2)^n > 0$. Comme sur cet événement $X_{(n)} \le \vartheta^*/2$, on a

$$\begin{aligned} \mathbf{E}_{\vartheta^*}[X_{(n)}] &= \mathbf{E}_{\vartheta^*}[X_{(n)} \mathbb{1}_A] + \mathbf{E}_{\vartheta^*}[X_{(n)} \mathbb{1}_{A^c}] \\ &\le \frac{1}{2} \vartheta^* \boldsymbol{P}_{\vartheta^*}(A) + \vartheta^* \boldsymbol{P}_{\vartheta^*}(A^c) \\ &= \vartheta^* - \frac{1}{2} \vartheta^* \boldsymbol{P}_{\vartheta^*}(A) < \vartheta^*. \end{aligned}$$

Il en résulte que $\hat{\vartheta}_n^{\mathrm{MV}} = X_{(n)}$ est un estimateur biaisé.

Exercice 3.1. *Soit* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}([0, \vartheta^*])$ *avec* $\vartheta^* \in]0, +\infty[$ *et soit* $\hat{\vartheta}_n^{MV} = X_{(n)}$.

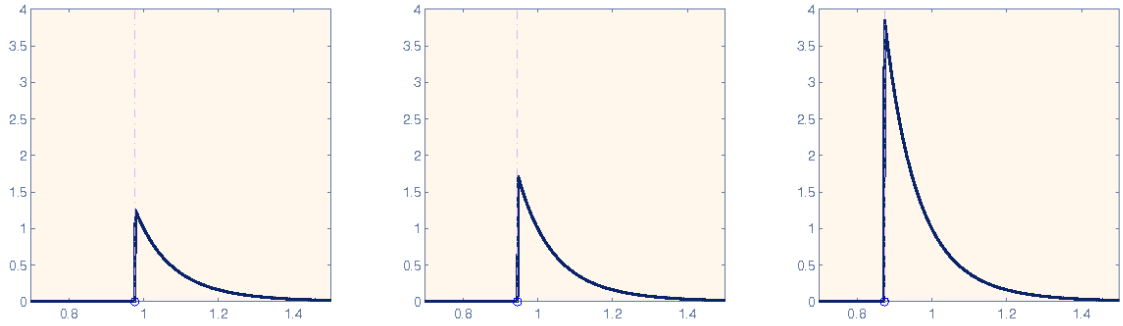


FIGURE 3.3 – Modèle uniforme : la fonction de vraisemblance et son maximum global. Les trois courbes représentent la vraisemblance pour trois échantillons différents de taille 10. La vraie valeur du paramètre dans les trois cas est $\vartheta^* = 1$. Les valeurs estimées qu'on obtient pour ces échantillons sont $\hat{\vartheta}_n^{\mathrm{MV}} = 0.98\,; 0.95\,; 0.87$.

1. *Vérifier que la fonction de répartition F_n de $X_{(n)}$ est donnée par*

$$F_n(\vartheta^*, x) = \begin{cases} 0, & si\ x \in]-\infty, 0], \\ (x/\vartheta^*)^n, & si\ x \in]0, \vartheta^*], \\ 1, & si\ x \in]\vartheta^*, +\infty]. \end{cases}$$

   En déduire la densité de $\hat{\vartheta}_n^{MV}$.
2. *Vérifier que la quantité*

$$B_n(\vartheta^*) = |\mathbf{E}_{\vartheta^*}[\hat{\vartheta}_n^{MV}] - \vartheta^*|,$$

   appelée le biais de $\hat{\vartheta}_n^{MV}$, est égale à $\vartheta^*/(n+1)$.
3. *En utilisant la définition de la convergence en loi, prouver que $\hat{\vartheta}_n^{MV}$ est asymptotiquement de loi exponentielle $-\mathcal{E}(1/\vartheta^*)$ avec la vitesse $1/n$, c'est-à-dire*

$$n(\vartheta^* - \hat{\vartheta}_n^{MV}) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{E}(1/\vartheta^*).$$

3.4 Intervalle de confiance

La méthode du maximum de vraisemblance nous permet de calculer une estimation de la vraie valeur du paramètre. Cependant, ayant calculé cette estimation, on peut naturellement s'interroger sur sa qualité. Une façon largement répandue pour décrire la qualité de l'estimation consiste à fournir un intervalle de confiance ou, plus généralement, une région de confiance.

Définition 3.8. *Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de loi $\boldsymbol{P}_{\vartheta^*}$ avec $\vartheta^* \in \Theta \subset \mathbb{R}^p$. On appelle région de confiance de niveau prescrit $1-\alpha$, avec $\alpha \in [0,1]$, tout sous-ensemble $I_n = I(X_1, \ldots, X_n)$ de $\mathbb{R}^p$ tel que*

$$\boldsymbol{P}_{\vartheta^*}\big(I_n \text{ contient } \vartheta^*\big) \geq 1-\alpha, \qquad \forall \vartheta^* \in \Theta. \tag{3.4}$$

Si $p = 1$ et I_n est un intervalle, on l'appelle intervalle de confiance. Si au lieu d'avoir (3.4) pour n fixé, on l'a de façon asymptotique, c'est-à-dire

$$\lim_{n\to\infty} \boldsymbol{P}_{\vartheta^*}\big(I_n \text{ contient } \vartheta^*\big) \geq 1-\alpha, \quad \forall \vartheta^* \in \Theta, \tag{3.5}$$

alors on dit que I_n est une région de confiance de niveau asymptotique $1-\alpha$.

La démarche générale pour construire un intervalle de confiance peut se résumer de la manière suivante.

1. On détermine un estimateur consistant $\bar{\vartheta}_n$; dans la plupart des cas, la loi de $\bar{\vartheta}_n$ est concentrée autour de la vraie valeur ϑ^*.
2. On cherche un $\delta_n = \delta(X_1, \ldots, X_n) > 0$ tel que

$$\boldsymbol{P}_{\vartheta^*}(|\bar{\vartheta}_n - \vartheta^*| > \delta_n) \leq \alpha, \qquad \forall \vartheta^* \in \Theta,$$

et l'on définit $I_n = [\bar{\vartheta}_n - \delta_n, \bar{\vartheta}_n + \delta_n]$.

Remarque 3.2. *Si la loi de $\bar{\vartheta}_n - \vartheta^*$ est fortement asymétrique, on remplace la seconde étape par la recherche de deux variables aléatoires $\delta_n = \delta(X_1, \ldots, X_n) > 0$ et $\delta'_n = \delta'(X_1, \ldots, X_n) > 0$ telles que*

$$\boldsymbol{P}_{\vartheta^*}\big(\bar{\vartheta}_n - \vartheta^* < -\delta_n\big) \leq \frac{\alpha}{2}, \quad \textit{et} \quad \boldsymbol{P}_{\vartheta^*}\big(\bar{\vartheta}_n - \vartheta^* > \delta'_n\big) \leq \frac{\alpha}{2},$$

pour tout $\vartheta^* \in \Theta$, et l'on définit $I_n = [\bar{\vartheta}_n - \delta'_n, \bar{\vartheta}_n + \delta_n]$.

Afin de clarifier le schéma présenté ci-dessus, considérons deux exemples.

3.4.1 Modèle de Bernoulli : intervalle de confiance par excès

Rappelons que dans l'exemple 1 portant sur la qualité de l'air, on dispose de n variables i.i.d. de loi $\mathcal{B}(\vartheta^*)$ avec $\vartheta^* \in]0,1[$. Nous avons déjà vu que l'EMV $\hat{\vartheta}_n^{\mathrm{MV}} = \overline{X}$ est consistant dans ce modèle. On cherche donc un δ_n tel que

$$\boldsymbol{P}_{\vartheta^*}\big(|\overline{X}_n - \vartheta^*| > \delta_n\big) \leq \alpha, \qquad \forall \vartheta^* \in]0,1[. \tag{3.6}$$

D'après l'inégalité de Tchebychev, on a

$$\boldsymbol{P}_{\vartheta^*}\big(|\overline{X}_n - \vartheta^*| > \delta_n\big) \leq \frac{\mathbf{E}_{\vartheta^*}[(\overline{X}_n - \vartheta^*)^2]}{\delta_n^2}.$$

Or, comme $\overline{X}_n$ est sans biais, il vient

$$\mathbf{E}_{\vartheta^*}[(\overline{X}_n - \vartheta^*)^2] = \mathbf{Var}_{\vartheta^*}(\overline{X}_n) = \frac{\mathbf{Var}_{\vartheta^*}\left[\sum_{i=1}^n X_i\right]}{n^2} = \frac{\vartheta^*(1-\vartheta^*)}{n}.$$

En combinant les deux inégalités précédentes avec l'inégalité élémentaire $ab \leq (a+b)^2/4$, on obtient

$$\boldsymbol{P}_{\vartheta^*}\big(|\overline{X}_n - \vartheta^*| > \delta_n\big) \leq \frac{\vartheta^*(1-\vartheta^*)}{n\delta_n^2} \leq \frac{1}{4n\delta_n^2}.$$

Il en résulte qu'en choisissant $\delta_n^2 = 1/(4n\alpha)$, l'inégalité (3.6) sera satisfaite. Par conséquent,

$$I_n = \Big[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}; \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\Big]$$

est un intervalle de confiance (IC) de niveau $1-\alpha$ pour ϑ^*. On remarque que le δ_n qu'on a trouvé n'est pas aléatoire. En d'autres termes, la longueur de l'IC ne dépend pas de l'échantillon qu'au travers de sa taille n.

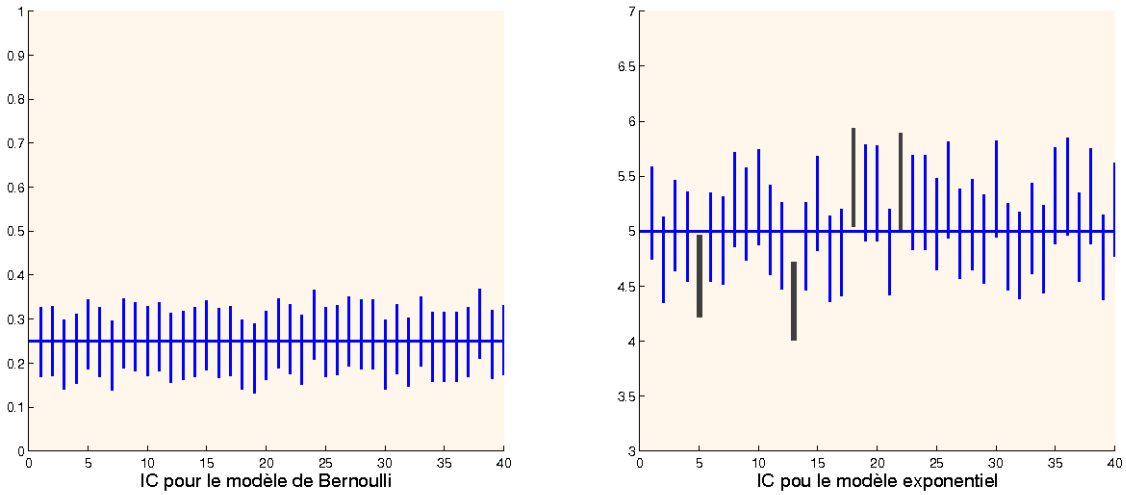


FIGURE 3.4 – A gauche : les intervalles de confiance de niveau 90% pour $\vartheta^* = 0.25$ dans le modèle de Bernoulli. On a tiré au hasard 40 échantillons de taille 400. En particulier, on remarque sur le graphe ci-dessus que tous les 40 intervalles contiennent la valeur 0.25 et sont tous de même taille. A droite : les intervalles de confiance de niveau 90% pour $\vartheta^* = 5$ dans le modèle exponentiel. On remarque que sur 40 échantillons de taille 400 tirés au hasard, 4 fois l'intervalle de confiance calculé ne contient pas la vraie valeur.

3.4.2 Modèle exponentiel : intervalle de confiance asymptotique

Considérons maintenant l'exemple de modèle exponentielle :

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(1/\vartheta^*), \qquad \vartheta^* \in]0, \infty[.$$

Nous avons vu que dans cet exemple l'EMV de ϑ^* est la moyenne empirique $\overline{X}_n$. De plus, en vertu de la loi des grands nombres $\overline{X}_n$ est un estimateur consistant. On cherche donc un intervalle de confiance sous la forme $[\overline{X}_n - \delta_n, \overline{X}_n + \delta_n]$. Dans ce cas, il est impossible d'appliquer la stratégie utilisée dans l'exemple précédent, car la variance de $\overline{X}_n$ égale à ϑ^{*2}/n n'est pas bornée sur $\Theta =]0, +\infty[$.

Supposons que la taille n de l'échantillon est suffisamment grande. On peut alors utiliser une approximation de la loi de $\overline{X}_n$ par une loi normale, car en vertu du théorème de la limite centrale (TLC),

$$\sqrt{n}(\overline{X}_n - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \vartheta^{*2}).$$

(L'utilisation du TLC est justifiée puisque $\mathbf{E}^*_\vartheta[X_1^2] = \mathbf{Var}_{\vartheta^*}[X_1] + (\mathbf{E}_{\vartheta^*}[X_1])^2 = 2\vartheta^{*2} < \infty$.) Cela implique que

$$\sqrt{n}\left(\frac{\overline{X}_n}{\vartheta^*} - 1\right) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

et, par conséquent,

$$\lim_{n\to\infty} \boldsymbol{P}_{\vartheta^*}\left(\sqrt{n}\left(\frac{\overline{X}_n}{\vartheta^*} - 1\right) \in A\right) = \boldsymbol{P}(\xi \in A), \qquad \forall A \in \mathscr{B}_{\mathbb{R}},$$

où $\xi \sim \mathcal{N}(0,1)$. On peut démontrer que le plus petit ensemble A tel que $\boldsymbol{P}(\xi \in A) = 1 - \alpha$ pour $\xi \sim \mathcal{N}(0,1)$ est $A = [-q^N_{1-\alpha/2}, q^N_{1-\alpha/2}]$ où $q^N_{1-\alpha/2}$ désigne le quantile d'ordre $1 - \alpha/2$ de

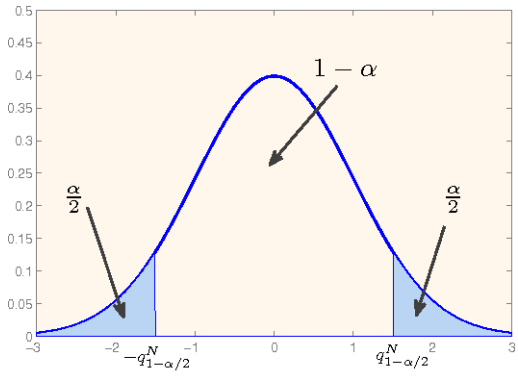


FIGURE 3.5 – La courbe de la densité de la loi normale centrée réduite et les quantiles d'ordre $1-\alpha/2$.

la loi normale centrée réduite (voir la Figure 3.5). En choisissant A de cette façon, on obtient

$$\lim_{n\to\infty} \boldsymbol{P}_{\vartheta^*}\left(\sqrt{n}\left(\frac{\overline{X}_n}{\vartheta^*}-1\right) \in [-q^N_{1-\alpha/2}, q^N_{1-\alpha/2}]\right) = 1-\alpha.$$

Pour conclure, il suffit de remarquer que

$$\begin{aligned}\sqrt{n}\left(\frac{\overline{X}_n}{\vartheta^*}-1\right) \in [-q^N_{1-\alpha/2}, q^N_{1-\alpha/2}] \quad &\Longleftrightarrow \quad \frac{\overline{X}_n}{\vartheta^*} \in \left[1-\frac{q^N_{1-\alpha/2}}{\sqrt{n}}, 1+\frac{q^N_{1-\alpha/2}}{\sqrt{n}}\right] \\ &\Longleftrightarrow \quad \vartheta^* \in \left[\frac{\overline{X}_n}{1+(q^N_{1-\alpha/2}/\sqrt{n})}, \frac{\overline{X}_n}{1-(q^N_{1-\alpha/2}/\sqrt{n})}\right].\end{aligned}$$

On en déduit que

$$I_n = \left[\frac{\overline{X}_n}{1+(q^N_{1-\alpha/2}/\sqrt{n})}, \frac{\overline{X}_n}{1-(q^N_{1-\alpha/2}/\sqrt{n})}\right]$$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour ϑ^*.

Exercice 3.2. *Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de loi $\mathcal{E}(1/\vartheta^*)$ avec $\vartheta^* \in]0, \infty[$.*

1. *Prouver que*
$$\frac{\sqrt{n}}{\overline{X}_n}(\overline{X}_n - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,1).$$
2. *En déduire que*
$$\tilde{I}_n = \left[\overline{X}_n\left(1-\frac{q^N_{1-\alpha/2}}{\sqrt{n}}\right), \overline{X}_n\left(1+\frac{q^N_{1-\alpha/2}}{\sqrt{n}}\right)\right]$$
est in IC de niveau asymptotique α pour ϑ^*.
3. *Démontrer que, pour les grandes valeurs de n, les intervalles I_n et $\tilde{I}_n$ sont très proches. Plus précisément, montrer que si $\frac{q^N_{1-\alpha/2}}{\sqrt{n}} \leq 1/2$ alors*
$$\frac{|I_n \setminus \tilde{I}_n| + |\tilde{I}_n \setminus I_n|}{|\tilde{I}_n|} \leq \frac{2q^N_{1-\alpha/2}}{\sqrt{n}}.$$

Exercice 3.3. *Vérifier que, dans le modèle de Bernoulli $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{B}(\vartheta^*)$,*

$$\tilde{I}_n = \left[\overline{X}_n - \frac{q^N_{1-\alpha/2}}{\sqrt{n}}; \overline{X}_n + \frac{q^N_{1-\alpha/2}}{\sqrt{n}}\right]$$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour le paramètre ϑ^*

3.5 Test d'hypothèses

On termine ce chapitre par un rappel très succinct des tests d'hypothèses. On se place toujours dans le contexte des modèles à observations i.i.d., où un échantillon $X_1, \dots, X_n$ de loi $\boldsymbol{P}_{\vartheta^*}$ sur $\mathbb{R}$ est à notre disposition, mais le paramètre $\vartheta^* \in \Theta$ est inconnu. Le but des tests statistiques est de désigner des procédures automatiques qui, pour un sous-ensemble (propre) $\Theta_0 \subset \Theta$ donné, permettent de décider avec une probabilité d'erreur contrôlée si oui ou non l'hypothèse «Θ_0 contient ϑ^*» est satisfaite.

3.5.1 Définitions principales

On est donc intéressé par tester l'hypothèse

$$H_0 : \vartheta^* \in \Theta_0 \qquad \text{contre} \qquad H_1 : \vartheta^* \in \Theta_0^c = \Theta \setminus \Theta_0. \tag{3.7}$$

On dit que H_0 est l'hypothèse nulle et H_1 est l'hypothèse alternative. La décision quant au rejet (ou pas) de l'hypothèse nulle doit bien-entendu être prise au vu de l'échantillon observé. Par conséquent, une procédure de test peut être considérée comme une partition de l'ensemble $\mathbb{R}^n$ (c'est l'ensemble des valeurs prises par l'échantillon) en deux classes. Si l'échantillon observé appartient à la première classe de la partition, on rejette l'hypothèse nulle, sinon on l'accepte. Ce raisonnement nous conduit à la définition suivante.

Définition 3.9. *On appelle région critique ou région de rejet, notée R_n, toute partie mesurable de $\mathbb{R}^n$. La procédure de test associée à la région critique R_n consiste à*
- *rejeter H_0 si $(x_1, \dots, x_n) \in R_n$,*
- *ne pas rejeter H_0 si $(x_1, \dots, x_n) \notin R_n$.*

Lorsqu'on effectue un test en utilisant une procédure basée sur la région critique R_n, deux types d'erreurs sont possibles. L'erreur de première espèce consiste à rejeter à tort l'hypothèse H_0. Par opposition, l'erreur de deuxième espèce consiste à accepter à tort l'hypothèse H_0. Comme la décision est prise au vu d'un échantillon aléatoire, chacune de ces deux erreurs a une certaine probabilité (généralement non nulle) d'être commise.

Définition 3.10. *Le risque de première espèce d'une procédure de test R_n, noté $\alpha(R_n)$ est la plus grande valeur atteinte par la probabilité de commettre l'erreur de première espèce :*

$$\alpha(R_n) = \sup_{\vartheta^* \in \Theta_0} \boldsymbol{P}_{\vartheta^*}\big((X_1, \dots, X_n) \in R_n\big).$$

De la même façon, le risque de deuxième espèce d'une procédure de test R_n, noté $\beta(R_n)$ est la plus grande valeur atteinte par la probabilité de commettre l'erreur de deuxième espèce :

$$\beta(R_n) = \sup_{\vartheta^* \notin \Theta_0} \boldsymbol{P}_{\vartheta^*}\big((X_1, \dots, X_n) \notin R_n\big).$$

On appelle puissance d'une procédure de test R_n l'application qui à chaque valeur $\vartheta \notin \Theta_0$ associe la probabilité de rejeter H_0 : $\pi_{R_n}(\vartheta) = \boldsymbol{P}_{\vartheta}\big((X_1, \dots, X_n) \in R_n\big)$.

En utilisant ce vocabulaire, une procédure de test R_n serait idéale si les risques de première et de deuxième espèce étaient tous les deux égaux à zéro : $\alpha(R_n) = \beta(R_n) = 0$. Malheureusement, sauf dans des cas très spécifiques, il n'existe pas de procédure idéale et on doit se contenter par des procédures dont les risques sont contrôlés.

Définition 3.11. *Soit $\alpha \in]0,1[$ une valeur donnée. Une procédure de test R_n est dite de niveau α si son risque de première espèce ne dépasse pas le niveau α :*

$$\alpha(R_n) \leq \alpha.$$

On dit que R_n est asymptotiquement de niveau α si $\lim_{n\to\infty} \alpha(R_n) \leq \alpha$.

Il existe en général un grand nombre de procédures de test de niveau α. L'une des approches les plus répandues pour départager deux procédures de niveau α est de donner la préférence à celle dont la puissance est plus grande partout sur Θ_0^c. Dans la même logique, un test de niveau asymptotique α est dit convergent (et considéré comme un bon test) si pour tout $\vartheta \notin \Theta_0$ fixé, la puissance $\pi_{R_n}(\vartheta)$ tend vers 1. Même si l'évaluation de la puissance est une étape importante dans l'étude d'une procédure de test, nous avons fait le choix de ne pas approfondir cette question dans ce cours.

3.5.2 Stratégie générale

Nous présentons ici un schéma générique qui comprend la plupart des stratégies usuelles de construction des procédures de test pour le problème (3.7). Il s'agit d'effectuer les étapes suivantes :

1. Déterminer un estimateur consistant, noté $\hat{\vartheta}_n$, du paramètre inconnu ϑ^*.
2. Déterminer une fonction $T : \mathbb{R} \times \Theta \to \mathbb{R}$ telle que
   (a) pour tout $\vartheta^* \in \Theta$, la fonction $u \mapsto T(\vartheta^* + u, \vartheta^*)$ est continue et ne s'annule qu'en 0, c'est-à-dire $T(\vartheta^* + u, \vartheta^*) = 0$ si et seulement si $u = 0$.
   (b) La loi de la variable aléatoire $T(\hat{\vartheta}_n, \vartheta^*)$ ne dépend pas de ϑ^*.
3. Définir, pour deux valeurs réelles a, b telles que $a \leq 0 \leq b$,

$$R_n = \{(x_1, \ldots, x_n) : T(\hat{\vartheta}_n, \vartheta) \notin [a,b] \quad \forall\, \vartheta \in \Theta_0\}.$$

4. Choisir a et b de telle sorte que R_n soit de niveau α.

La justification de cette stratégie est simple. La fonction T joue le rôle d'une distance (signée) entre l'estimateur et les valeurs possibles du paramètre inconnu ϑ^*. Comme $\hat{\vartheta}_n$ est consistant et T est continue par rapport à la première variable, on a $T(\hat{\vartheta}_n, \vartheta^*) \approx T(\vartheta^*, \vartheta^*) = 0$. Par conséquent, si l'hypothèse nulle $H_0 : \vartheta^* \in \Theta_0$ est vraie, il existe un élément ϑ de Θ_0 tel que $T(\hat{\vartheta}_n, \vartheta)$ se trouve dans un voisinage de 0. Cela nous conduit à accepter H_0 si $T(\hat{\vartheta}_n, \vartheta) \in [a,b]$ pour un élément $\vartheta \in \Theta_0$ et de la rejeter dans le cas contraire. D'où la définition de la région critique ci-dessus.

Remarque 3.3 (Loi symétrique)**.** *Dans la plupart des exemples que nous allons considérer par la suite, la loi de la variable aléatoire $T(\hat{\vartheta}_n, \vartheta^*)$ sera symétrique par rapport à zéro. On prendra alors $a = -b$ et on pourra réécrire R_n sous la forme*

$$R_n = \{(x_1, \ldots, x_n) : \min_{\vartheta \in \Theta_0} |T(\hat{\vartheta}_n, \vartheta)| > b\}.$$

Remarque 3.4 (Test asymptotique)**.** *Si on cherche un test de niveau asymptotique α, la condition 2(b) peut être remplacée par la suivante : pour tout $\vartheta^* \in \Theta_0$, la variable aléatoire $T(\hat{\vartheta}_n, \vartheta^*)$ converge en loi vers une variable aléatoire dont la loi ne dépend pas de ϑ^*.*

3.5.3 P-value d'un test

Lorsqu'on effectue un test statistique, on a souvent envie de quantifier l'évidence ou la pertinence de la decision dictée par le test. La notion qui nous permet d'atteindre cet objectif est la p-value d'un test. Afin de motiver la définition rigoureuse de la p-value donnée ci-dessous, remarquons que la majorité des tests peut être écrite comme

$$R_{n,\alpha} = \big\{(x_1, \dots, x_n) : S_n(x_1, \dots, x_n) \geq C_\alpha\big\}$$

où S_n est une statistique de test et C_α est un nombre réel appelé seuil critique du test. Ici, on a ajouté un indice α à la région critique R_n pour souligner le fait que le test est de niveau α. Considérons le cas où

$$\sup_{\vartheta^* \in \Theta_0} \boldsymbol{P}_{\vartheta^*}\big((X_1, \dots, X_n) \in R_{n,\alpha}\big) = \alpha.$$

Intuitivement, il est claire que la région $R_{n,\alpha}$ grossit lorsque α augmente. Il existe donc une valeur α^* pour laquelle R_{n,α^*} contient la réalisation observée $x_1, \dots, x_n$, alors que pour tous les $\alpha < \alpha^*$ R_{n,α^*} ne contient pas la réalisation observée. Cette valeur α^* est appelée p-value du test $R_{n,\alpha}$.

Définition 3.12. *On appelle p-value d'un test $R_n = R_{n,\alpha}$, notée $\alpha^*(R_n)$, la plus petite valeur de α pour laquelle le test R_n rejette l'hypothèse H_0.*

En pratique, si la p-value d'un test est inférieure à 5%, alors l'hypothèse H_0 sera rejeté au seuil de 5%. De plus, une p-value très petite traduit l'évidence de la décision concernant le rejet de H_0.

3.5.4 Exemple 1 : test bilatéral dans le modèle de Bernoulli

On observe $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{B}(\vartheta^*)$ et on cherche à tester l'hypothèse

$$H_0 : \vartheta^* = \vartheta_0$$

contre l'alternative bilatérale

$$H_1 : \vartheta^* \neq \vartheta_0$$

où $\vartheta_0 = 10\%$. En suivant le schéma générique, on utilise comme estimateur de ϑ^* la proportion empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. D'après le théorème de la limite centrale, on a

$$\sqrt{n}(\overline{X}_n - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \vartheta^*(1-\vartheta^*))$$

ou encore

$$\frac{\sqrt{n}(\overline{X}_n - \vartheta^*)}{\sqrt{\vartheta^*(1-\vartheta^*)}} \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,1).$$

Par conséquent, on pose

$$T(\overline{X}_n, \vartheta) = \frac{\sqrt{n}(\overline{X}_n - \vartheta)}{\sqrt{\vartheta(1-\vartheta)}}$$

et

$$R_n = \{(x_1, \dots, x_n) : |T(\overline{X}_n, \vartheta_0)| > b\}.$$

Pour que R_n soit de niveau α, il faut que $\lim_{n\to\infty} \boldsymbol{P}_{\vartheta_0}(|T(\overline{X}_n, \vartheta_0)| > b) \leq \alpha$. Or, la convergence en loi établie ci-dessus implique que

$$\lim_{n\to\infty} \boldsymbol{P}_{\vartheta_0}(|T(\overline{X}_n, \vartheta_0)| > b) = \boldsymbol{P}(|\xi| > b), \qquad \xi \sim \mathcal{N}(0,1).$$

Par conséquent, on choisit b de telle sorte que la probabilité de l'événement $|\xi| > b$ soit égale à α. Cela nous conduit vers $b = q^N_{1-\alpha/2}$ (voir la Figure 3.5).
Nous avons donc construit la procédure de test suivant :
- on rejette $H_0 : \vartheta^* = \vartheta_0$, avec $\vartheta_0 = 10\%$, si

$$\left|\frac{\sqrt{n}(\overline{X}_n - \vartheta_0)}{\sqrt{\vartheta_0(1-\vartheta_0)}}\right| > q^N_{1-\alpha/2} \Longleftrightarrow |\overline{X}_n - 0.1| > \frac{0.3 \cdot q^N_{1-\alpha/2}}{\sqrt{n}}.$$

- on ne rejette pas H_0 si l'inégalité ci-dessus n'est pas satisfaite.

3.5.5 Exemple 2 : test unilatéral dans le modèle exponentiel

Plaçons-nous maintenant dans la situation où on observe $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(1/\vartheta^*)$ et on cherche à tester l'hypothèse

$$H_0 : \vartheta^* \geq \vartheta_0$$

contre l'alternative unilatérale

$$H_1 : \vartheta^* < \vartheta_0$$

avec, par exemple, $\vartheta_0 = 2$. Comme dans l'exemple précédent, on utilise l'EMV de ϑ^* qui n'est autre que la moyenne empirique $\hat{\vartheta}^{\text{MV}}_n = \overline{X}_n$ et qui vérifie

$$\sqrt{n}(\overline{X}_n - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0, \vartheta^{*2})$$

en vertu du théorème de la limite centrale. Posons

$$T(\overline{X}_n, \vartheta) = \frac{\sqrt{n}(\overline{X}_n - \vartheta)}{\vartheta} \quad \text{et} \quad R_n = \big\{(x_1, \ldots, x_n) : \min_{\vartheta \geq \vartheta_0} |T(\overline{X}_n, \vartheta)| > b\big\}.$$

On vérifie aisément que

$$\begin{aligned} \min_{\vartheta \geq \vartheta_0} |T(\overline{X}_n, \vartheta)| > b \quad &\Longleftrightarrow \quad |\vartheta^{-1}\overline{X}_n - 1| > bn^{-1/2},\ \forall \vartheta \geq \vartheta_0 \\ &\Longleftrightarrow \quad \overline{X}_n < \vartheta_0(1 - bn^{-1/2}). \end{aligned}$$

On veut donc déterminer b de telle sorte que

$$\lim_{n\to\infty} \sup_{\vartheta^* \geq \vartheta_0} \boldsymbol{P}_{\vartheta^*}\Big(\overline{X}_n < \vartheta_0(1 - bn^{-1/2})\Big) = \alpha.$$

On peut vérifier que la loi de $\bar{X}_n/\vartheta^*$ est absolument continue et ne dépend pas de ϑ^*, ce qui implique que

$$\begin{aligned} \sup_{\vartheta^* \geq \vartheta_0} \boldsymbol{P}_{\vartheta^*}\Big(\overline{X}_n < \vartheta_0(1 - bn^{-1/2})\Big) &= \sup_{\vartheta^* \geq \vartheta_0} F_{\overline{X}_n/\vartheta^*}\left(\frac{\vartheta_0(1 - bn^{-1/2})}{\vartheta^*}\right) \\ &= F_{\overline{X}_n/\vartheta^*}\big(1 - bn^{-1/2}\big) \\ &= \boldsymbol{P}_{\vartheta^*}\Big(\overline{X}_n/\vartheta^* < 1 - bn^{-1/2}\Big) \\ &= \boldsymbol{P}_{\vartheta^*}\Big(\frac{\sqrt{n}(\overline{X}_n - \vartheta^*)}{\vartheta^*} < -b\Big) \\ &\xrightarrow[n\to\infty]{} \boldsymbol{P}(\xi < -b) = \boldsymbol{P}(\xi > b) = 1 - \boldsymbol{P}(\xi \leq b) \end{aligned}$$

où $\xi \sim \mathcal{N}(0,1)$. Pour que le test soit de niveau asymptotique α, on choisit $b = q^N_{1-\alpha}$.
En conclusion, nous rejetons l'hypothèse $H_0 : \vartheta^* \geq \vartheta_0$ si et seulement si $\overline{X}_n < \vartheta_0\Big(1 - \frac{q^N_{1-\alpha}}{\sqrt{n}}\Big)$.

3.6 Exercices

Exercice 1. On observe un échantillon $X_1, \dots, X_n$ de loi double exponentielle translatée. C'est-à-dire, $X_1, \dots, X_n \overset{\text{iid}}{\sim} \boldsymbol{P}_{\vartheta^*}$ où $\vartheta^* \in \mathbb{R}$ et $\boldsymbol{P}_{\vartheta^*}$ a pour densité la fonction

$$p(\vartheta^*, x) = \frac{1}{2}\, e^{-|x-\vartheta^*|}, \qquad \forall x \in \mathbb{R}.$$

1. Vérifier que $p(\vartheta^*, \cdot)$ est bien une densité de probabilité et prouver que la médiane empirique de l'échantillon $X_1, \dots, X_n$ est l'EMV de ϑ^*.
2. Montrer que la moyenne empirique de l'échantillon est un estimateur consistant et asymptotiquement normal de ϑ^*.
3. On suppose que n est grand et on admet le résultat suivant : si $X_1, \dots, X_n$ sont i.i.d. de densité p dont la médiane est ϑ^*, alors la médiane empirique $\widehat{Me}_n$ de l'échantillon $X_1, \dots, X_n$ vérifie $2p(\vartheta^*)\sqrt{n}(\widehat{Me}_n - \vartheta^*) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,1)$.

   Au vu de ce résultat et de celui de la question 2, lequel des deux estimateurs $\widehat{Me}_n$ et $\overline{X}_n$ préféreriez-vous.

Exercice 2. Soient $X_1, \dots, X_n$ des variables i.i.d. de loi exponentielle $\mathcal{E}(1/\vartheta^*)$ avec $\vartheta^* > 0$.

1. Montrer que la fonction $T(x, \vartheta) = (x - \vartheta)/\vartheta$ vérifie les conditions 2(a) et 2(b) (voir paragraphe 3.5.2) avec $\hat{\vartheta}_n = \overline{X}_n$.
2. En déduire un test d'hypothèse $H_0 : \vartheta^* = 1$ contre $H_1 : \vartheta^* \neq 1$.

3.7 Résumé du Chapitre 3

3.7.1 Modèle statistique

1. **Définition :** on appelle modèle statistique le triplet $(\mathcal{X}_n, \mathscr{T}_n, \{\boldsymbol{P}_{n,\vartheta}, \vartheta \in \Theta\})$, où $\mathcal{X}_n$ est l'espace d'états et Θ est l'espace des paramètres. La problématique statistique est alors la suivante : ayant observé un élément $x^{(n)}$ de $\mathcal{X}_n$ tiré au hasard selon la loi $\boldsymbol{P}_{n,\vartheta^*}$ (avec un ϑ^* que l'on ignore), caractériser la loi $\boldsymbol{P}_{n,\vartheta^*}$.
2. **Modèle à observations i.i.d. :** $x^{(n)}$ est une réalisation d'un vecteur aléatoire $X^{(n)} = (X_1, \dots, X_n)$ dont les coordonnées sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.).
3. **Modèle discret :** un modèle à observations i.i.d. tel que X_1 prend ces valeurs dans un ensemble fini ou dénombrable, noté $A = \{a_1, a_2, \dots\}$. Un modèle discret est caractérisé par les valeurs $p(\vartheta; a_k) = \boldsymbol{P}_\vartheta(X_1 = a_k)$.
4. **Modèle à densité :** un modèle à observations i.i.d. tel que X_1 admet une densité par rapport à la mesure de Lebesgue, noté $p(\vartheta; x)$. Cela équivaut à $\boldsymbol{P}_\vartheta(X_1 \in I) = \int_I p(\vartheta; x)\, dx$ pour tout intervalle I et tout $\vartheta \in \Theta$.
5. **Echantillon :** le vecteur aléatoire dont on a observé une réalisation. Dans le modèle à observations i.i.d., c'est simplement une suite $X_1, \dots, X_n$ de variables aléatoires i.i.d. de loi $\boldsymbol{P}_{\vartheta^*}$.
6. **Statistique :** toute variable aléatoire de forme $Y = g(X_1, \dots, X_n)$ où g est une fonction mesurable.

7. **Vraisemblance :** pour un modèle à observations i.i.d., qu'il soit discret ou à densité, la fonction de vraisemblance est donnée par la formule :

$$p_n(\vartheta; x_1, \ldots, x_n) = \prod_{i=1}^{n} p(\vartheta; x_i).$$

   Pour un modèle discret, $p(\vartheta; x_i)$ est la probabilité de la valeur x_i si la vraie valeur du paramètre est ϑ. Pour un modèle à densité, $p(\vartheta; x_i)$ est la valeur de la densité, lorsque la vraie valeur du paramètre est ϑ, évaluée au point x_i.

8. **Log-vraisemblance :** étant donné les observations $X_1, \ldots, X_n$, la log-vraisemblance est :

$$l_n(\vartheta) = \log p_n(\vartheta; X_1, \ldots, X_n) = \sum_{i=1}^{n} \log p(\vartheta; X_i).$$

   Cette fonction peut prendre la valeur $-\infty$ si l'argument du log s'annule.

3.7.2 Estimation

Pour un échantillon $X_1, \ldots, X_n$ donné, on appelle estimateur toute statistique des $X_1, \ldots, X_n$: $\hat{\vartheta}_n = g_n(X_1, \ldots, X_n)$.

1. **Estimateur sans biais :** $\mathbf{E}_\vartheta[\hat{\vartheta}_n] = \vartheta$ pour tout ϑ.
2. **Estimateur convergent (consistant) :** $\hat{\vartheta}_n \xrightarrow[n\to\infty]{P} \vartheta^*$ si $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \boldsymbol{P}_{\vartheta^*}$.
3. **Estimateur asymptotiquement normal (de vitesse $1/\sqrt{n}$ et de variance limite $\sigma^2_{\vartheta^*}$) :** $\sqrt{n}(\hat{\vartheta}_n - \vartheta^*) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, \sigma^2_{\vartheta^*})$.
4. **Estimateur du maximum de vraisemblance :** la valeur du paramètre ϑ qui maximise la vraisemblance $p_n(\vartheta; X_1, \ldots, X_n)$ ou, de façon équivalente, la log-vraisemblance $l_n(\vartheta)$, est appelée estimateur du maximum de vraisemblance et est notée $\hat{\vartheta}_n^{MV}$.

3.7.3 Intervalle de confiance

1. **Intervalle de confiance de niveau $1-\alpha$:** on dit que l'intervalle I qui dépend de l'échantillon $X_1, \ldots, X_n$ est un intervalle de confiance de niveau $1-\alpha$ pour le paramètre ϑ, si $\boldsymbol{P}_{\vartheta^*}(\vartheta^* \in I) \geq 1-\alpha$. Si cette inégalité est stricte, on parle d'un intervalle de confiance par excès.
2. **Intervalle de confiance de niveau asymptotique $1-\alpha$:** on dit que l'intervalle I_n qui dépend de l'échantillon $X_1, \ldots, X_n$ est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour le paramètre ϑ, si[3] $\lim_{n\to\infty} \boldsymbol{P}_{\vartheta^*}(\vartheta^* \in I_n) \geq 1-\alpha$.
3. **Exemple générique :** si $\hat{\vartheta}_n$ est un estimateur consistant de ϑ tel que $\sqrt{n}(\hat{\vartheta}_n - \vartheta^*) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, \sigma^2_{\vartheta^*})$ et l'application $\vartheta \mapsto \sigma^2_\vartheta$ est continue, alors

$$I_n = \left[\hat{\vartheta}_n - \frac{\sigma_{\hat{\vartheta}_n}}{\sqrt{n}}\, q^N_{1-\alpha/2}; \hat{\vartheta}_n + \frac{\sigma_{\hat{\vartheta}_n}}{\sqrt{n}}\, q^N_{1-\alpha/2}\right]$$

   est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour ϑ^*. Ici, $q^N_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite : $\mathcal{N}(0,1)$.

3. La limite ici est en réalité une limité inférieure

3.7.4 Test d'hypothèses

On cherche à tester l'hypothèse nulle $H_0 : \vartheta^* \in \Theta_0$ contre l'alternative $H_1 : \vartheta^* \in \Theta_1$.

1. On dit que l'hypothèse H_0 est **simple**, si Θ_0 ne contient qu'un seul élément :$\Theta_0 = \{\vartheta_0\}$. Une hypothèse qui n'est pas simple est dite **composite**.
2. **Région critique ou région de rejet :** on appelle région critique d'un test l'ensemble R des valeurs possibles de l'échantillon pour lesquelles l'hypothèse nulle est rejetée.
3. **Erreur de première espèce :** le fait de rejeter à tort l'hypothèse nulle.
4. **Erreur de deuxième espèce :** le fait de ne pas rejeter l'hypothèse H_0, alors qu'il fallait le faire.
5. **Risque de première espèce :** la probabilité de l'erreur de première espèce :
$$\sup_{\vartheta^* \in \Theta_0} \boldsymbol{P}_{\vartheta^*}\big((X_1, \ldots, X_n) \in R\big).$$
6. **Risque de deuxième espèce :** la probabilité de l'erreur de deuxième espèce :
$$\sup_{\vartheta^* \in \Theta_1} \boldsymbol{P}_{\vartheta^*}\big((X_1, \ldots, X_n) \notin R\big).$$
7. **Test de niveau α :** le risque de première espèce ne dépasse pas le niveau α.
8. **Test de niveau asymptotique α :** la limite (inférieure) lorsque $n \to \infty$ du risque de première espèce ne dépasse pas α.
9. **Puissance d'un test :** la fonction
$$\vartheta^* \mapsto \boldsymbol{P}_{\vartheta^*}\big((X_1, \ldots, X_n) \in R\big), \qquad \forall \vartheta^* \in \Theta_1.$$
Entre deux tests de niveau α, celui dont la puissance est plus grande est préférable.
10. **P-value d'un test :** soit R_α la région critique d'un test (de niveau α). Etant donné les observations $x_1, \ldots, x_n$, la p-value du test R_α est la plus grande valeur de α pour laquelle l'hypothèse H_0 n'est pas rejetée : $\max\{\alpha : (x_1, \ldots, x_n) \notin R_\alpha\}$.
11. **Interprétation :** une p-value élevée suggère que l'hypothèse nulle ne doit pas être rejetée. Typiquement, si la p-value est $> 5\%$ on ne rejette pas l'hypothèse nulle.

4
Régression linéaire multiple

4.1 Généralités

4.1.1 Plans d'expériences

Le statisticien planifie une expérience statistique en fonction d'un objectif qui est souvent l'étude de l'effet de certains **facteurs** de variabilité d'un phénomène. Ces facteurs sont présents sous plusieurs **modalités**.

La technique de bon sens lorsque plusieurs facteurs sont à étudier est de ne modifier qu'un facteur à la fois. Par exemple, si on dispose de 3 facteurs présents chacun sous p modalités, cette technique conduirait à fixer 2 facteurs puis étudier dans chacun des cas l'effet du troisième facteur, soit $3p^2$ expériences. Dans beaucoup de cas le coût, l'efficacité, ou les possibilités effectives d'expérimentation, recommandent de minimiser le nombre d'expériences tout en conservant un cadre expérimental rigoureux. En répondant à ces critères, la méthode des plans d'expérience initiée au début du $XX^{\text{ème}}$ siècle par Ronald A.Fisher s'est imposée dans le cadre industriel pour tester des médicaments, des variétés de plantes, des procédés de fabrication, etc...

L'objectif de la construction de plans d'expérience est de mettre en place un dispositif expérimental permettant d'aboutir à une interprétation statistique des résultats notamment à l'aide de tests d'hypothèses. Pour cela il faut construire un modèle statistique qui distinguera parmi les facteurs de variabilité les **facteurs contrôlés** et les **facteurs aléatoires**.

4.1.2 Le modèle général

Ce type d'expérience statistique peut être décrit avec le modèle général suivant :

$$Y = f(\vartheta) + \varepsilon,$$

où
- $Y = (Y_i)_{i=1,\dots,n}$ désigne les observations effectuées.
- $\vartheta = (\vartheta_1, \dots, \vartheta_p)$ est un vecteur de paramètres inconnu caractérisant les facteurs contrôlés que l'on souhaite étudier à l'aide de ces observations.

– $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$ sont des variables aléatoires indépendantes et centrées, représentant l'erreur expérimentale. Le modèle est gaussien si ε est un vecteur gaussien centré.
– $f(\cdot)$ est une application connue qui fixe le modèle. Ce modèle est linéaire si $f(\vartheta)$ est une application $\vartheta \mapsto \boldsymbol{X}\vartheta$ où $\boldsymbol{X}$ est une matrice. Le modèle s'écrit alors matriciellement :

$$Y = \boldsymbol{X}\vartheta + \varepsilon.$$

Dans la suite nous considérerons des modèles linéaires gaussiens. Ces deux hypothèses (linéarité et caractère gaussien de l'erreur) doivent être validées. Pour les vérifier on peut, soit utiliser la connaissance *a priori* que l'on a du modèle, soit construire des tests.

Dans certains cas, lorsqu'il y a plusieurs observations, le caractère gaussien peut être une conséquence du théorème de la limite centrale. Enfin, dans de nombreux cas, on peut rendre le modèle gaussien et linéaire en effectuant des transformations sur les observations.

4.1.3 Exemples

Dans ce paragraphe nous proposons des exemples illustrant la problématique précédente. Dans les sections suivantes, nous donnerons les éléments permettant de résoudre ce type de problèmes.

Exemple 4.1. *Le tableau ci-dessous représente des mesures de hauteurs d'arbres en mètres effectuées dans 3 forêts distinctes. On rassemble dans un même tableau les mesures effectuées dans les* 3 *forêts dans le but de les comparer.*

Foret 1 $n_1 = 13$ arbres	Foret 2 $n_2 = 14$	Foret 3 $n_3 = 10$
23.4	22.5	18.9
24.4	22.9	21.1
24.6	23.7	21.2
24.9	24.0	22.1
25.0	24.4	22.5
26.2	24.5	23.5
26.3	25.3	24.5
26.8	26.0	24.6
26.8	26.2	26.2
26.9	26.4	26.7
27.0	26.7	
27.6	26.9	
27.7	27.4	
	28.5	

TABLE 4.1 – Hauteurs d'arbres dans 3 forêts

Le facteur étudié est ici l'influence de la forêt sur la hauteur de ces arbres. La variabilité de la hauteur due ici au tirage d'un échantillon aléatoire dans chaque forêt se décompose donc naturellement en une partie controlée, le facteur (forêt), et une partie aléatoire, la variabilité intrinsèque à la pousse des arbres due au terrain, à la lumière, à la présence ou non d'un autre arbre à proximité...

On peut supposer que les hauteurs des différents arbres sont indépendantes (ce qui exige que l'on ne mesure pas des arbres trop rapprochés les uns des autres), et que, pour la forêt numéro k, la mesure d'un arbre suit une loi gaussienne de moyenne m_k et de variance σ_k^2 ; on peut alors comparer les 3

échantillons 2 à 2. Mais si la variabilité des hauteurs des arbres peut être considérée comme identique d'une forêt à l'autre ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$) on observe trois échantillons gaussiens de même variance σ^2 et de moyennes différentes qui représentent l'effet de chaque forêt (les modalités du facteur "forêt") sur la pousse des arbres. L'hypothèse d'égalité des variances est appelée **homoscédasticité**. *Avec ces hypothèses on peut alors écrire :*

$$Y_{i,j} = m_i + \varepsilon_{i,j} \quad \text{pour la j-ième mesure de la forêt } i, \quad j = 1, \ldots, n_i, \quad i = 1, 2, 3,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Ceci s'écrit avec une notation matricielle :

$$Y = \boldsymbol{X}\vartheta + \varepsilon,$$

où ε est un vecteur aléatoire gaussien, et

$$Y = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2}, Y_{3,1}, \ldots, Y_{3,n_3})^t,$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ - & - & - \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ - & - & - \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}, \quad \vartheta = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$$

Ce problème est un problème **d'analyse de la variance à un facteur**. *Pour répondre à la question "existe-t-il un effet forêt", on construira un test statistique dont l'hypothèse nulle est :*

$$H_0 : m_1 = m_2 = m_3.$$

Exemple 4.2. *Le tableau suivant donne le nombre de jours de pluie et la hauteur de pluie en mm, observés pendant toute l'année à Paris de 1956 à 1995.*

Une représentation sur un graphique (fig. 4.1) des données avec en abscisse le nombre de jours de pluie et en ordonnée la hauteur de pluie permet de constater que l'ensemble des points forme un nuage allongé et que la quantité de pluie augmente lorsque le nombre de jours de pluie augmente.

Le facteur hauteur de pluie *est alors un facteur à expliquer par le facteur explicatif contrôlé* nombre de jours de pluie.

La question que l'on se pose est de savoir si ces deux quantités sont liées par une relation affine, de calculer les paramètres de cette relation et d'avoir une indication sur le caractère prédictif de ce modèle (autrement dit, peut-on déduire de façon satisfaisante la hauteur de pluie à partir du nombre de jours de pluie ?).

Le modèle statistique que l'on propose est le suivant :

$$Y_i = \beta + \alpha\ X_i + \varepsilon_i$$

où :

Années	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Jours	154	161	193	131	198	152	159	159	146	196
Hauteur	545	536	783	453	739	541	528	559	521	880
Années	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Jours	192	161	176	173	199	141	170	156	198	164
Hauteur	834	592	634	618	631	508	740	576	668	658
Années	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Jours	135	179	171	172	170	197	173	177	177	163
Hauteur	417	717	743	729	690	746	700	623	745	501
Années	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Jours	176	180	167	140	149	140	154	155	192	162
Hauteur	611	707	734	573	501	472	645	663	699	670

TABLE 4.2 – Jour et quantité de pluie par années

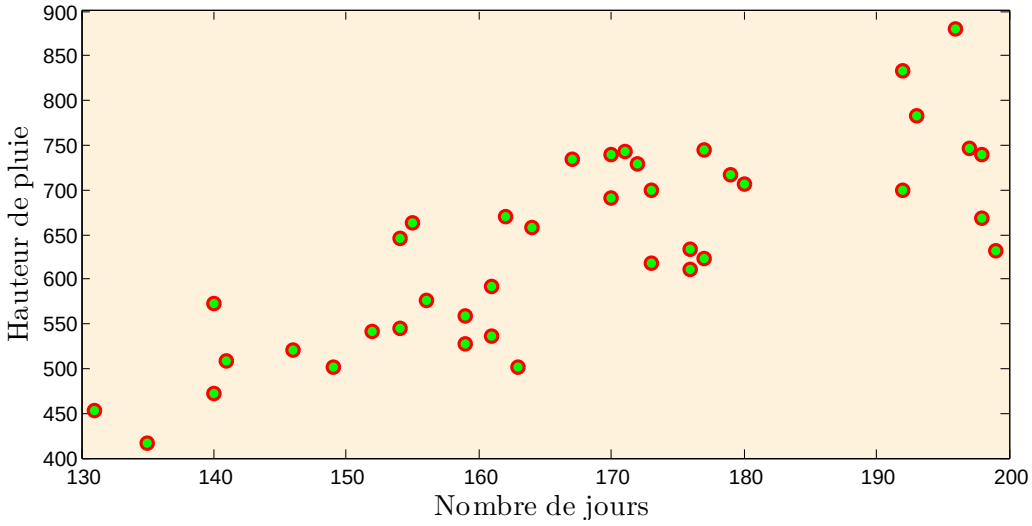


FIGURE 4.1 – Représentation des données

– $Y = (Y_i)_{i=1,\dots,n}$ *désigne la hauteur de pluie.*
– $(X_i)_{i=1,\dots,n}$ *désigne le nombre de jours de pluie*
– *la droite d'équation*

$$y = \alpha\, x + \beta$$

est appelée droite de régression ; α et β sont à estimer à partir des observations.
– $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$ *représente les écarts aléatoires entre les observations et la droite. On supposera que c'est une suite de variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$.*

Le modèle peut alors s'écrire :

$$Y = \boldsymbol{X}\vartheta + \varepsilon$$

en notant :

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad et \quad \vartheta = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$

C'est un modèle de **régression linéaire simple** *qui sera étudié en 4.4.*

4.2 Lois associées aux échantillons gaussiens

Rappelons pour commencer les définitions des lois associées aux échantillons gaussiens qui nous serons utiles dans la suite.

Définition 4.1.

- *Si $(X_1, \ldots, X_n)$ est un échantillon de loi normale $\mathcal{N}(0,1)$, alors la loi de la v.a. $\sum_{i=1}^n X_i^2$ est la loi du chi-deux à n degrés de liberté, notée $\chi^2(n)$.*
- *Si $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$ et que X et Y sont indépendantes, alors $\frac{X}{\sqrt{Y/n}} \sim t(n)$, loi de Student à n degrés de liberté.*
- *Si $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ et que X et Y sont indépendantes, alors $\frac{X/n}{Y/m} \sim \mathcal{F}(n,m)$, loi de Fisher (ou de Fisher-Snedecor) à n et m degrés de liberté.*

Enfin, on utilise souvent la convention pratique suivante : si une v.a. X a pour loi F, on note aF la loi de aX. Ainsi, on notera $\sigma^2\chi^2(n)$ la loi de $\sum_{i=1}^n X_i^2$ dans le cas où $(X_1, \ldots, X_n)$ forment un n-échantillon de la loi $\mathcal{N}(0, \sigma^2)$.

4.2.1 Théorème de Cochran

C'est l'outil fondamental pour l'étude des échantillons gaussiens et du modèle linéaire gaussien (la notation $||\cdot||$ désigne la norme euclidienne dans $\mathbb{R}^n$).

Théorème 4.1. *Soit $Y = (Y_1, \ldots, Y_n)$ un n-échantillon de $\mathcal{N}(0,1)$, et $E_1, \ldots, E_p$ une suite de p sous-espaces deux-à-deux orthogonaux de $\mathbb{R}^n$, avec $dim(E_j) = d_j$, $j = 1, \ldots, p$. Alors on a :*

(i) Les composantes de Y dans toute base orthonormale de $\mathbb{R}^n$ forment encore un n-échantillon de $\mathcal{N}(0,1)$.

(ii) Les vecteurs aléatoires $Y_{E_1}, \ldots, Y_{E_p}$, qui sont les projections de Y sur $E_1, \ldots, E_p$, sont indépendants.

(iii) Les variables aléatoires $||Y_{E_1}||, \ldots, ||Y_{E_p}||$ sont indépendantes, et $||Y_{E_j}||^2 \sim \chi^2(d_j)$, $j = 1, \ldots, p$.

Une formulation équivalente consiste à dire (par exemple avec $p = 2$), que si P_1 et P_2 sont deux projecteurs orthogonaux de $\mathbb{R}^n$ sur deux sous-espaces orthogonaux E_1 et E_2 de dimensions d_1 et d_2, alors $P_1Y = Y_{E_1}$ et $P_2Y = Y_{E_2}$ sont indépendants, et $||P_1Y||^2$ et $||P_2Y||^2$ sont indépendants et ont pour lois respectivement $\chi^2(d_1)$ et $\chi^2(d_2)$.

4.2.2 Statistiques fondamentales

Plaçons-nous donc dans le cas où $(Y_1, \ldots, Y_n)$ est un n-échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$. Les statistiques utiles pour les problèmes de test ou d'intervalle de confiance sur les paramètres

μ et σ^2 sont fonction de la moyenne empirique, que nous notons

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

et de la variance empirique, dont nous choisissons ici la "version sans biais" (voir 4.3.1) :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\bar{Y})^2\right].$$

Utilisons le théorème 4.1 dans le cas où $p = 2$ et où on projette Y sur le sous-espace E de dimension 1 engendré par le vecteur (normé) de $\mathbb{R}^n$, $e_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ (où on note $\mathbf{1}_n$ le vecteur de dimension n ayant toute ses coordonnées égales à 1). On obtient $Y_E = \sqrt{n}\bar{Y}\frac{1}{\sqrt{n}}\mathbf{1}_n$. La norme de la projection de Y sur l'orthogonal de E (de dimension $n-1$) est

$$||Y - Y_E||^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

qui suit la loi $\sigma^2\chi^2(n-1)$ (c'est le point (iii) du théorème de Cochran à ceci près qu'il faut tenir compte de la variance σ^2). On en déduit les résultats suivants, utiles pour le statisticien :

Proposition 4.1. *Soit $Y = (Y_1, \ldots, Y_n)$ un n-échantillon de $\mathcal{N}(\mu, \sigma^2)$. Alors on a :*

(i) Les v.a. $\bar{Y}$ et S^2 sont indépendantes.

(ii) $(n-1)S^2 \sim \sigma^2\chi^2(n-1)$.

(iii) $\dfrac{\sqrt{n}(\bar{Y} - \mu)}{S} \sim t(n-1)$.

Remarquons que la v.a. $\sum_{i=1}^{n}(Y_i - \mu)^2$ suit elle-même la loi $\sigma^2\chi^2(n)$ mais, si μ est inconnu, son calcul n'est pas accessible. Le point (ii) exprime intuitivement le fait que l'on perd un degré de liberté en raison du remplacement de μ, inconnu, par son estimateur $\bar{Y}$. De même la v.a. $\sqrt{n}(\bar{Y} - \mu)/\sigma \sim \mathcal{N}(0,1)$, autrement dit le point (iii) signifie que la loi de Student remplace la loi normale comme loi de la moyenne empirique normalisée dans le cas où σ est inconnu et doit être remplacé par son estimateur S.

4.3 Le modèle gaussien

Nous illustrons dans un premier temps les concepts du modèle paramètrique sur le modèle gaussien. Ce modèle est très (trop ?) couramment utilisé pour analyser des données continues. Cet usage fréquent est dû à la simplicité des calculs et à la généralité du TCL (sous des hypothèses très faibles, la somme de nombreux petits bruits suit asymptotiquement une loi gaussienne).

4.3.1 Un exemple de données réelles à loi gaussienne

On a enregistré le taux d'alcool dans le sang (en dg/l) de n sujets : voici le tableau des observations, avec $n = 30$ (extrait de l'ouvrage de D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion).

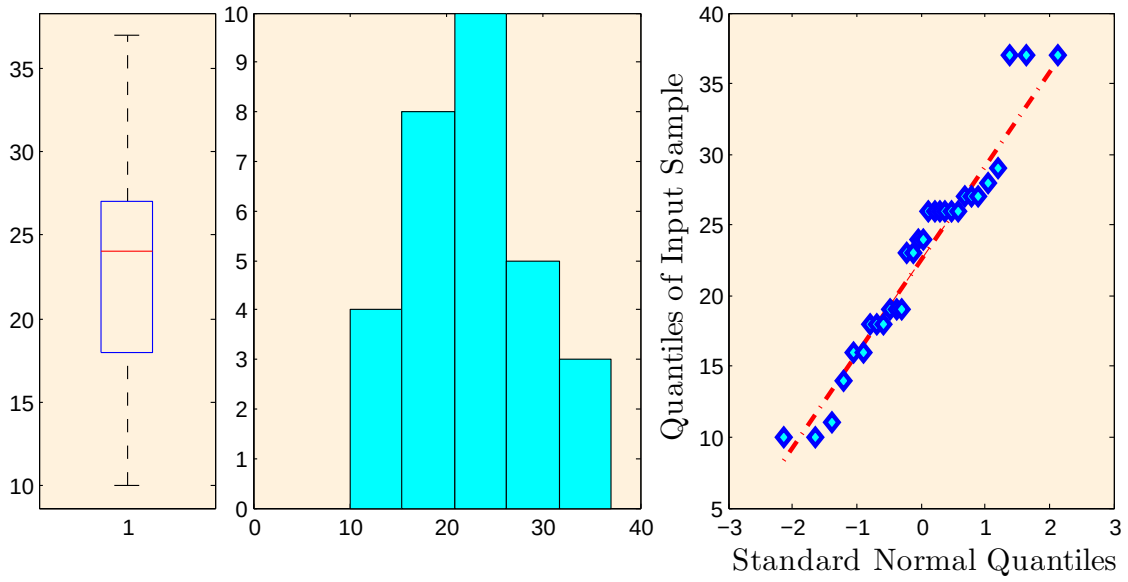


FIGURE 4.2 – Le boxplot, l'histogramme et le QQ-plot des données du taux d'alcool.

27 , 26 , 26 , 29 , 10 , 28 , 26 , 23 , 14 , 37 , 16 , 18 , 26 , 27 , 24
19 , 11 , 19 , 16 , 18 , 27 , 10 , 37 , 24 , 18 , 26 , 23 , 26 , 19 , 37

On notera $(x_1, \ldots, x_{30})$ cette suite de résultats observée. Les valeurs s'échelonnant entre 10 et 37, la précision étant l'unité, il serait maladroit de modéliser ceci comme les réalisations de v.a. discrètes : le nombre de valeurs distinctes envisageables devrait être grand, de l'ordre de la quarantaine, car rien n'interdit de penser qu'auraient pu être observées des valeurs en dehors de l'intervalle ici présent. Il est plus raisonnable de considérer qu'il y a, sous-jacent à ces observations, un phénomène à valeurs réelles, dont les observations recueillies sont une discrétisation, l'arrondi se faisant à la précision du décigramme par litre.

Les modèles les plus simples que l'on puisse envisager ici sont des modèles d'échantillonnage : on admet que l'on a observé les réalisations de n v.a. Y_i indépendantes et identiquement distribuées.

Pour voir si un tel modèle est approprié, il faut d'abord se demander comment a été constitué cet échantillon.

Le problème essentiel est, comme dans le premier paragraphe, celui de la source de variabilité (cause de l'aléatoire). Celle-ci a en fait ici plusieurs origines simultanées : variation d'individu à individu et, pour chaque individu, imprécision de l'appareil de mesure et effet de l'erreur d'arrondi.

Il est assez évident que, quelles que soient les conditions de recueil, elles ont dû assurer l'indépendance des n v.a. Y_i dont les observations résultent. Le problème de l'identité de leurs lois et du choix de la famille à laquelle serait supposée appartenir cette loi commune est plus délicat.

Nous l'avons dit, les praticiens utilisent souvent dans un tel contexte une modélisation avec pour loi commune une loi normale, de moyenne μ et variance σ^2 (non nulle) inconnues, $\mathcal{N}(\mu, \sigma^2)$. Le paramètre est donc bi-dimensionnel $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$. La probabilité $\mathcal{N}(\mu, \sigma^2)$ a pour support $\mathbb{R}$ tout entier, alors qu'ici (comme presque toujours dans la pratique) les données sont fondamentalement bornées ; cet usage suppose donc que, pour la zone de valeurs de μ et σ envisageables, la probabilité du complémentaire de l'intervalle des

valeurs effectivement atteignables par les taux d'alcool soit négligeable.

4.3.2 Étude du modèle

On considère donc un échantillon $(Y_1, \dots, Y_n)$ de v.a. indépendantes et de même loi gaussienne : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \vartheta = (\mu, \sigma^2) \in \mathbb{R} \times]0, \infty[\}$. La densité de la loi $\mathcal{N}(\mu, \sigma^2)$ est

$$p(y_1; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y_1-\mu)^2/2\sigma^2}.$$

La vraisemblance du modèle est pour $y = (y_1, \dots, y_n) \in \mathbb{R}^n$,

$$\begin{aligned} p_n(y; \mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \mathrm{e}^{-\sum_{i=1}^n (y_i-\mu)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} \mathrm{e}^{-n\frac{(\bar{y}_n-\mu)^2+v_n}{2\sigma^2}}, \end{aligned}$$

où $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ et $v_n = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y}_n)^2$. Traditionnellement, on considère

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \frac{1}{n-1}\sum_{i=1}^n Y_i^2 - \frac{n}{n-1}(\bar{Y}_n)^2,$$

au lieu de $V_n = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ (car S_n^2 est un estimateur sans biais de σ^2), cf la proposition 4.1. La loi de la statistique $(\bar{Y}_n, S_n^2)$ est donnée dans la proposition 4.1.

4.3.3 Estimation

Pour calculer l'estimateur du maximum de vraisemblance de (μ, σ^2), on considère la log-vraisemblance

$$\ell_n(y; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - n\frac{(\bar{y}_n - \mu)^2 + v_n}{2\sigma^2}.$$

En calculant les dérivées partielles, il vient

$$\frac{\partial}{\partial\mu}\ell_n(y; \mu, \sigma^2) = n\frac{\bar{y}_n - \mu}{\sigma^2},$$

et

$$\frac{\partial}{\partial\sigma^2}\ell_n(y; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + n\frac{(\bar{y}_n - \mu)^2 + v_n}{2\sigma^4}.$$

En particulier, les dérivées de la log-vraisemblance s'annulent pour $\mu = \bar{y}_n$ et $\sigma^2 = v_n$. Ensuite, on vérifie sans difficulté que la log-vraisemblance atteint son maximum pour $(\mu, \sigma^2) = (\bar{y}_n, v_n)$. On en déduit donc que l'EMV de $\vartheta = (\mu, \sigma^2)$ est $(\bar{Y}_n, V_n)$. On déduit de la proposition 4.1 que $\mathbf{E}_\vartheta[\bar{Y}_n] = \mu$ et que $\mathbf{E}_\vartheta[S_n^2] = \sigma^2$. (En revanche V_n est un estimateur biaisé de σ^2, d'où le choix traditionnel de S_n^2). Ainsi l'estimateur $\hat{\vartheta}_n = (\bar{Y}_n, S_n^2)$ est un estimateur sans biais de ϑ.

Par la loi forte des grands nombre $\bar{Y}_n$ et S_n^2 sont des estimateurs convergents. Ainsi $\hat{\vartheta}_n$ est un estimateur convergent de ϑ. (On peut également vérifier qu'il est asymptotiquement normal, mais cela ne nous sera pas utile par la suite).

4.3.4 Intervalle de confiance et tests pour la moyenne

On déduit de la proposition 4.1, que la loi de $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ est la loi $t(n-1)$. La loi de Student est symétrique, ainsi si $q_{1-\alpha/2}(t_{n-1})$ est le quantile d'ordre $1-\alpha/2$ de la loi $t(n-1)$, alors $-q_{1-\alpha/2}(t_{n-1})$ est le quantile d'ordre $\alpha/2$. En particulier, une v.a. de loi $t(n-1)$ appartient à $[-q_{1-\alpha/2}(t_{n-1}), q_{1-\alpha/2}(t_{n-1})]$ avec probabilité $1-\alpha$. Comme

$$\frac{\sqrt{n}(\bar{Y}_n - \mu)}{S_n} \in [-q_{1-\alpha/2}(t_{n-1}), q_{1-\alpha/2}(t_{n-1})] \Longleftrightarrow \mu \in \Big[\bar{Y}_n \pm q_{1-\alpha/2}(t_{n-1})\frac{S_n}{\sqrt{n}}\Big],$$

on en déduit que $[\bar{Y}_n \pm q_{1-\alpha/2}(t_{n-1})\frac{S_n}{\sqrt{n}}]$ est un intervalle de confiance de niveau $1-\alpha$ pour μ.

On remarque que la longueur de l'intervalle de confiance $\big[\bar{y}_n \pm q_{1-\alpha/2}(t_{n-1})\frac{s_n}{\sqrt{n}}\big]$, où $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$ tend bien vers 0 quand la taille de l'échantillon tend vers l'infini (à $\bar{y}_n$ et s_n fixé). Il est aussi d'autant plus long que s_n est plus élevé (ceci est naturel : la fluctuation des données contrarie la confiance que l'on a en elles, confiance qui se traduirait par un intervalle de confiance assez court).

Exercice 4.1. *Si la variance est connue et égale à σ_0^2, c'est-à-dire si l'on considère le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$, vérifier que l'intervalle $[\bar{Y}_n \pm q^N_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}]$ (où $q^N_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$) est alors un intervalle de confiance de niveau $1-\alpha$ pour μ.*

On considère les hypothèses $H_0 : \mu = \mu_0$ et $H_1 : \mu \neq \mu_0$, où μ_0 est donné. (On parle d'hypothèse bilatérale, par opposition à l'exercice 4.2, où parle d'hypothèse unilatérale). Il est naturel de comparer la moyenne empirique avec moyenne proposée, μ_0. Toutefois, sous H_0, la loi de $\bar{Y}_n - \mu_0$ est la loi $\mathcal{N}(0, \sigma^2/n)$, qui dépend du paramètre inconnu σ^2. On considère donc la statistique de test

$$\zeta_n = \sqrt{n}\frac{\bar{Y}_n - \mu_0}{S_n}.$$

La loi de la statistique de test sous H_0 est la loi de Student de paramètre $n-1$. La loi de ζ_n sous H_1 est la loi de Student décentrée, mais nous ne l'expliciterons pas ici. On remarque que sous H_1, $\bar{Y}_n - \mu_0$ converge p.s. vers $\mu - \mu_0 \neq 0$ quand $n \to \infty$. On a toujours que S_n converge p.s. vers σ^2. On en déduit donc que sous H_1, p.s.

$$\lim_{n\to\infty} |\zeta_n| = +\infty.$$

Il est donc naturel de considérer la région critique

$$W_n = \{(y_1, \ldots, y_n);\ |\zeta_n^{\text{obs}}| \geq a_n\}, \tag{4.1}$$

où $\zeta_n^{\text{obs}} = \sqrt{n}\frac{\bar{y}_n - \mu_0}{s_n}$, avec $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ et $s_n = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$. D'après le comportement de la statistique de test sous H_1, on en déduit que le test W_n est convergent.

Comme sous H_0, la loi de ζ_n est la loi de Student de paramètre $n-1$, on en déduit que le niveau du test W_n est

$$\sup_{\vartheta \in H_0} \boldsymbol{P}_\vartheta(W_n) = \boldsymbol{P}(|Z| \geq a_n),$$

où Z est de loi $t(n-1)$. Pour obtenir un test de niveau α, on choisit $a_n = q_{1-\alpha/2}(t_{n-1})$, le quantile d'ordre $1-\alpha/2$ de loi de Student de paramètre $n-1$.

La p-valeur du test est donnée par

$$p\text{-valeur} = \boldsymbol{P}(|Z| \geq |\zeta_n^{\text{obs}}|), \tag{4.2}$$

où ζ_n^{obs} est la statistique de test évaluée en les observations.

Remarque 4.1. *On peut étudier la réponse du test en fonction de n, $\bar{y}_n$ et s_n.*
- *à n et s_n fixés, si $\bar{y}_n$ s'éloigne de μ_0, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test.*
- *à n et $\bar{y}_n$ fixés, si s_n diminue, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test. Cela traduit le fait que si s_n est petit alors la variabilité des données est petite et $\bar{y}_n$ donne une estimation précise du vrai paramètre μ. Des petits écarts entre $\bar{y}_n$ et μ_0 deviennent significatifs.*
- *à $\bar{y}_n$ et s_n fixés, si n augmente, alors $|\zeta_n|$ augmente et on a tendance à rejeter le test. En effet, plus la taille de l'échantillon est grande est plus $\bar{y}_n$ donne une estimation précise du vrai paramètre μ.*

Exercice 4.2. *Écrire le test pour les hypothèses unilatérales $H_0 : \mu \leq \mu_0$ et $H_1 : \mu > \mu_0$.*

Exercice 4.3. *Tester les hypothèses $H_0 : \mu = \mu_0$ et $H_1 : \mu \neq \mu_0$, où μ_0 est donné dans le modèle gaussien à variance connue : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$.*

4.3.5 Intervalles de confiance et tests pour la variance

Le raisonnement est identique dans le cas de la variance : la construction d'intervalles de confiance ou de tests se fait à partir de la connaissance de la loi, sous l'hypothèse nulle ou à la frontière de celle-ci, de l'estimateur du paramètre d'intérêt.

Intervalles de confiance pour la variance

L'estimateur (sans biais) de σ^2 est la variance empirique sans biais S^2, et le point (ii) de la proposition 4.1 permet d'écrire par exemple que, si $\chi^2_{n,1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(n)$,

$$\boldsymbol{P}\left(q_{\alpha/2}(\chi^2_{n-1}) < \frac{(n-1)S^2}{\sigma^2} < q_{1-\alpha/2}(\chi^2_{n-1})\right) = 1-\alpha,$$

d'où l'on déduit un intervalle de confiance pour la variance (bilatéral dans cet exemple) de niveau de confiance $(1-\alpha)$:

$$\left[\frac{(n-1)S^2}{q_{1-\alpha/2}(\chi^2_{n-1})}; \frac{(n-1)S^2}{q_{\alpha/2}(\chi^2_{n-1})}\right].$$

Tests pour la variance

On peut aussi, en suivant la démarche introduite au chapitre 3, construire des tests pour des hypothèses relatives au paramètre σ^2. Considérons par exemple le test de

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{contre} \quad H_1 : \sigma^2 > \sigma_0^2.$$

à la frontière de H_0, i.e. lorsque la valeur du paramètre est σ_0^2, la statistique

$$Z = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Cette statistique aura "tendance à croître" avec σ sous l'hypothèse alternative (et de plus $S^2 \to \sigma^2$ p.s. en vertu de la loi forte des grands nombres), d'où le choix d'une région de rejet de la forme $]c, +\infty[$, où c est calibré (le plus petit possible) de sorte que $\boldsymbol{P}_{\sigma_0}(Z > c) = \alpha$. Ceci amène donc à choisir pour c le quantile d'ordre $(1-\alpha)$ de la loi $\chi^2(n-1)$, autrement dit à conclure

$$\text{Rejet de } H_0 \text{ si} \quad \frac{(n-1)S^2}{\sigma_0^2} > q_{1-\alpha}(\chi^2_{n-1}).$$

Le lecteur pourra construire les tests relatifs aux situations suivantes :

$$H_0 : \{\sigma^2 \geq \sigma_0^2\} \text{ contre } H_1 : \{\sigma^2 < \sigma_0^2\},$$
$$H_0 : \{\sigma^2 = \sigma_0^2\} \text{ contre } H_1 : \{\sigma^2 \neq \sigma_0^2\}.$$

4.3.6 Analyse des données réelles

On choisit le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. On obtient l'estimation de (μ, σ^2) à l'aide de $(\bar{y}_n, s_n^2)$:

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i = 22.9 \quad \text{et} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = 53.0.$$

L'intervalle de confiance de niveau 95% de μ est donné par

$$[\bar{y}_n \pm q_{1-\alpha/2}(t_{n-1})\frac{s_n}{\sqrt{n}}] = [20.2, 25.6].$$

La p-valeur associée au test de région critique (4.1), définie par (4.2), est pour $\mu_0 = 20$,

$$p\text{-valeur} = \boldsymbol{P}(|Z| \geq \zeta_n^{\text{obs}}) = 0.037, \quad \text{où} \quad \zeta_n = \sqrt{n}\frac{\bar{y}_n - \mu_0}{s_n} = 2.18.$$

En particulier on rejette $H_0 : \{\mu = \mu_0\}$ au niveau de 5%.

4.4 Régression linéaire multiple

Rappel de la problématique

La problématique a été introduite sur un exemple en 4.1. Reprenons-la avec une autre situation. Il s'agit ici de modéliser un phénomène aléatoire observé par une combinaison linéaire ou affine de *variables explicatives*, dont les valeurs sont déterministes et connues pour chaque expérience (ou observation) réalisée. Par exemple, si l'on souhaite "expliquer" la durée d'une certaine maladie (en jours) après l'admission de patients à l'hôpital, on peut penser que cette durée est liée à certaines variables quantitatives (i.e., à valeur numériques). On relèvera par exemple les nombres de bactéries de certains types présentes dans l'organisme du patient à son arrivée, ainsi que des indicateurs de son état général (poids, température,...).

Si l'on dispose de n observations de ces variables explicatives ainsi que de la variable à expliquer (l'observation de la variable à expliquer est donc faite a posteriori dans cet exemple, lorsque les n patients ont quitté l'hopital) on peut étudier la pertinence de cette modélisation linéaire. Il est possible de tester la significativité du modèle, et celle de certaines variables explicatives. Il est possible aussi d'estimer les liens entre variables explicatives et variable à expliquer et éventuellement de faire ensuite de la *prédiction*, c'est à dire ici d'estimer la durée d'hospitalisation d'un nouveau patient à partir de la connaissance des valeurs des variables explicatives dans son cas.

4.4.1 Cadre général du modèle linéaire gaussien

L'introduction générale et l'exemple précédent permettent de dégager le cadre formel ci-dessous. On effectue n observations $Y = (Y_1, \dots, Y_n)$, et chaque observation est l'addition d'un "effet moyen" et d'un "bruit". Si on considère le vecteur des observations $Y \in \mathbb{R}^n$, le modèle s'écrit

$$Y = \mu + \varepsilon,$$

et on fait les hypothèses (de modèle) suivantes :

- $M1$ l'effet moyen μ est inconnu et non observable, mais $\mu \in E$, sous espace vectoriel de $\mathbb{R}^n$, fixé et de dimension k ;
- $M2$ le vecteur aléatoire ε (non observable) a pour loi $\mathcal{N}(0, \sigma^2 I_n)$ et le paramètre $\sigma^2 > 0$ est inconnu.

Estimation

Ayant observé Y, le point de E le plus proche de Y est sa projection sur E, $Y_E = \mu + \varepsilon_E$, qui est l'estimateur intuitif de μ. La projection sur l'orthogonal de E, $Y - Y_E = \varepsilon - \varepsilon_E$ ne contient pas d'information sur μ (elle est centrée) : c'est un indicateur de la dispersion des observations, qu'il est naturel d'utiliser pour estimer σ^2. On précise ceci dans le résultat suivant, conséquence directe du théorème 4.1.

Proposition 4.2. *On observe $Y = \mu + \varepsilon$ avec les hypothèses $M1$ et $M2$. Alors on a :*

(i) Y_E *est un estimateur sans biais de* μ.

(ii) $||Y - Y_E||^2/(n-k)$ *est un estimateur sans biais de* σ^2.

(iii) Y_E *et* $Y - Y_E$ *sont indépendants.*

(iv) $||Y_E - \mu||^2 \sim \sigma^2 \chi^2(k)$ *et* $||Y - Y_E||^2 \sim \sigma^2 \chi^2(n-k)$.

On peut montrer également que, pour tout vecteur $u \in \mathbb{R}^n$, le produit scalaire $\langle u, Y_E \rangle$ est l'estimateur de $\langle u, \mu \rangle$ sans biais de variance minimum.

4.4.2 Définition du modèle

On observe un phénomène aléatoire Y et l'on suppose ce phénomène influencé par p variables explicatives ou *régresseurs*, $R^1, \dots, R^p$. Parfois, Y est aussi appelée la *variable dépendante*, et $R^1, \dots, R^p$ les *variables indépendantes*.

On réalise n observations, autrement dit $Y = (Y_1, \dots, Y_n)$, et on note $R_i^1, \dots, R_i^p$ les conditions expérimentales pour la i-ème observation Y_i, c'est à dire les valeurs (déterministes) des p régresseurs lors de l'expérience i. On fait comme on l'a dit l'hypothèse d'une relation linéaire ou affine entre les régresseurs et la variable à expliquer Y et, comme en analyse de la variance, on suppose observer la somme de l'effet de ces régresseurs et d'un ensemble de perturbations non observables, que l'on résume par un "bruit" gaussien centré. Ce modèle s'écrit ainsi

$$Y_i = \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i, \quad \text{ou bien} \quad Y_i = \beta + \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i, \quad i = 1, \dots, n,$$

où $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ est un n-échantillon de la loi $\mathcal{N}(0, \sigma^2)$ (l'hypothèse d'homoscédasticité est présente ici aussi, puisque σ^2 ne dépend pas de i). Les paramètres inconnus à estimer sont $(\beta, \alpha_1, \dots, \alpha_p, \sigma^2)$ dans le cas affine (on retire β dans le cas linéaire sans constante).

Notation vectorielle

Considérons par exemple le cas affine, et posons

$$\boldsymbol{X} = \begin{bmatrix} 1 & R_1^1 & \cdots & R_1^p \\ \vdots & \vdots & \ldots & \vdots \\ 1 & R_n^1 & \cdots & R_n^p \end{bmatrix} = \left[\mathbf{1}_n\, R^1 \cdots R^p\right],$$

la matrice $n \times (p+1)$ des régresseurs (la colonne de 1, $\mathbf{1}_n$, étant considérée comme un régresseur particulier lorsqu'elle figure dans le modèle). Posons aussi $\vartheta \in \mathbb{R}^{p+1}$ le paramètre du modèle, où $\vartheta = (\beta, \alpha_1, \ldots, \alpha_p)^t$. Le modèle s'écrit vectoriellement :

$$Y = \boldsymbol{X}\vartheta + \varepsilon, \quad \text{avec } \boldsymbol{X}\vartheta \in E \text{ et } \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

où $E = \{\boldsymbol{X}u, u \in \mathbb{R}^{p+1}\}$ est le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\boldsymbol{X}$. Ce modèle s'inscrit ainsi dans le cadre général du modèle linéaire gaussien décrit en 4.4.1, avec adoption des hypothèses $M1$ et $M2$ qui y ont été faites.

On suppose que la dimension de E est $p+1$, c'est à dire que les p régresseurs et $\mathbf{1}_n$ sont linéairement indépendants, ou ce qui revient au même que $\text{rang}(\boldsymbol{X}) = p+1$, ou encore que la matrice symétrique $\boldsymbol{X}^T\boldsymbol{X}$ est elle-même de rang $p+1$. Cette hypothèse n'est pas une réelle perte de généralité puisque, si elle n'est pas vérifiée, cela signifie que l'un des régresseurs est combinaison linéaire des autres ; il n'apporte alors pas d'explication supplémentaire et il suffit de le retirer.

Exemple 4.3. ***La régression simple.*** *C'est la situation où l'on dispose d'un seul régresseur ($p = 1$) que nous notons simplement R. Le modèle s'écrit*

$$Y_i = \beta + \alpha R_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

ce qui revient à dire que $Y_i \sim \mathcal{N}(\beta + \alpha R_i, \sigma^2)$. On visualise ce modèle dans l'espace des variables (R, Y) par le fait que les observations "tombent" dans un "tunnel gaussien" d'amplitude σ le long de la droite d'équation $x = \beta + \alpha r$. L'exemple 4.2 des données de pluie est de ce type.

4.4.3 Estimation

On applique dans ce cadre les résultats de la proposition 4.2. La projection de Y sur E est l'estimateur sans biais de $\boldsymbol{X}\vartheta$. Il s'écrit $Y_E = \boldsymbol{X}\hat{\vartheta}$, où $\hat{\vartheta} \in \mathbb{R}^{p+1}$ est l'estimateur sans biais de ϑ. Il est tel que $Y - \boldsymbol{X}\hat{\vartheta}$ est orthogonal à tout vecteur de E, autrement dit pour tout vecteur $u \in \mathbb{R}^{p+1}$, $\langle \boldsymbol{X}u, Y - \boldsymbol{X}\hat{\vartheta}\rangle = 0$, ce qui donne

$$\boxed{\hat{\vartheta} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t Y.}$$

Remarquons que, si l'on note P le projecteur sur E (donc tel que $Y_E = PY$), celui-ci s'écrit $P = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$. La résiduelle est $||Y - Y_E||^2 = \langle Y, Y - Y_E\rangle = Y^t(I - P)Y$, soit

$$||Y - Y_E||^2 = Y^t\left[I - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\right]Y.$$

D'après le point (iv) de la proposition 4.2, $||Y - Y_E||^2 \sim \sigma^2\chi^2\big(n - (p+1)\big)$, et l'on estime (sans biais) la variance par

$$\boxed{\hat{\sigma}^2 = \frac{||Y - Y_E||^2}{n - (p+1)}.}$$

Remarque : dans le cas de la régression sans constante, il suffit de retirer la colonne $\mathbf{1}_n$ de $\boldsymbol{X}$ et de remplacer $p+1$ par p.

Variances des estimateurs

On déduit immédiatement de l'expression de $\hat{\vartheta}$ que sa matrice de variances-covariances est

$$\mathbf{Var}(\hat{\vartheta}) = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}.$$

Exemple 4.4. ***La régression simple*** *(suite de l'exemple 4.3).*
Il est facile de mener les calculs "à la main" dans le cas de la régression simple. La matrice des régresseurs est $\boldsymbol{X} = [\mathbf{1}_n\ R]$*, d'où*

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^n (R_i - \bar{R})^2}\begin{bmatrix} \sum_{i=1}^n R_i^2 & -\sum_{i=1}^n R_i \\ -\sum_{i=1}^n R_i & n \end{bmatrix},$$

et le calcul de $\hat{\vartheta}$ *donne*

$$\hat{\alpha} = \frac{\mathrm{Cov}(R,Y)}{\mathrm{Var}(R)}, \qquad \hat{\beta} = \bar{Y} - \hat{\alpha}\bar{R},$$

où $\bar{R} = \sum_{i=1}^n R_i/n$ *est la moyenne empirique de* R*, et*

$$\mathrm{Var}(R) = \frac{1}{n}\sum_{i=1}^n (R_i - \bar{R})^2, \quad \mathrm{Cov}(R,Y) = \frac{1}{n}\sum_{i=1}^n (R_i - \bar{R})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^n R_iY_i - \bar{R}\bar{Y},$$

sont les variances et covariances empiriques (qui ont le sens de mesures descriptives ici puisque R *n'est pas aléatoire). On peut remarquer que ces estimateurs coïncident avec les estimateurs des moindres carrés de la droite de régression de* Y *sur* R*, c'est à dire la pente et la constante de la droite d'équation* $Y = b + aR$ *qui minimisent les carrés des écarts* $\sum_{i=1}^n (Y_i - b - aR_i)^2$*.*
On déduit immédiatement de l'expression de $\mathbf{Var}(\hat{\vartheta})$ *l'expression des variances de* $\hat{\alpha}$ *et* $\hat{\beta}$*, ainsi que la covariance entre les deux estimateurs (ils ne sont pas indépendants). Comme ils sont des estimateurs sans biais des paramètres qu'ils estiment, et suivent des lois gaussiennes (car combinaisons linéaires de* Y*), on a finalement :*

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha\,,\ \frac{\sigma^2}{\sum_{i=1}^n (R_i - \bar{R})^2}\right), \qquad \hat{\beta} \sim \mathcal{N}\left(\beta\,,\ \frac{\sigma^2\sum_{i=1}^n R_i^2}{n\sum_{i=1}^n (R_i - \bar{R})^2}\right).$$

Le projeté est $Y_E = \hat{\beta}\mathbf{1}_n + \hat{\alpha}R$*, et on peut écrire directement la résiduelle*

$$SSE = ||Y - Y_E||^2 = \sum_{i=1}^n (Y_i - \hat{\beta} - \hat{\alpha}R_i)^2,$$

écarts entre les valeurs observées et les valeurs ajustées par le modèle. Elle suit la loi $\sigma^2\chi^2(n-2)$*, et l'estimateur sans biais de la variance est* $||Y - Y_E||^2/(n-2)$ *qui est indépendant de* $\hat{\vartheta}$*. La connaissance des lois des estimateurs de* (β, α)*, qui dépendent de* σ^2*, ainsi que de la loi de l'estimateur de* σ^2 *et cette propriété d'indépendance permet de construire des intervalles de confiance ou des tests pour* β *et* α *analogues aux intervalles de confiance et tests de Student construits en 4.3.5.*

4.4.4 Test de l'utilité des régresseurs

Dans le modèle $Y = \boldsymbol{X}\vartheta + \varepsilon$ avec p régresseurs et la constante (cas affine), on souhaite souvent tester l'utilité d'une partie des régresseurs, autrement dit une hypothèse nulle de la forme

$$H_0 : \{\text{“}R^{q+1}, \ldots, R^p \text{ sont inutiles”}\} \quad \text{contre} \quad H_1 : \{\text{“c'est faux”}\},$$

où $1 \leq q < p$, et où on a éventuellement effectué une permutation de l'ordre des régresseurs. La contre-hypothèse se comprend comme H_1 : "l'un des $R^j, q+1 \leq j \leq p$ au moins est utile". L'hypothèse nulle, si elle n'est pas rejetée, permet alors de simplifier le modèle de régression, en ne conservant qu'une partie des variables qui étaient a priori explicatives. Les hypothèses peuvent se reformuler comme

$$H_0 : \{\alpha_j = 0, j = q+1, \ldots, p\} \quad \text{contre} \quad H_1 : \{\text{il existe au moins un } \alpha_j \neq 0\},$$

autrement dit comme l'appartenance, sous H_0, de l'effet moyen à un sous-espace vectoriel de E de dimension plus petite, ce qui nous ramène à la méthode employée pour le test d'homogénéité en analyse de la variance (voir le passage "Généralisation" en **??**, p. **??**). En effet, le sous-modèle associé à H_0 s'écrit $Y_i = \beta + \sum_{j=1}^{q} \alpha_j R_i^j + \varepsilon_i$, $i = 1, \ldots, n$, ou vectoriellement (en indiçant par 0 les quantités qui diffèrent sous l'hypothèse nulle)

$$Y = \boldsymbol{X}_0 \vartheta_0 + \varepsilon, \quad \boldsymbol{X}_0 = [\mathbf{1}_n \, R^1 \cdots R^q], \quad \vartheta_0 = (\beta, \alpha_1, \ldots, \alpha_q)^t,$$

et donc $\boldsymbol{X}_0 \vartheta_0 \in H = \{\boldsymbol{X}_0 w : w \in \mathbb{R}^{q+1}\}$, où H est de dimension $q+1$. On teste ainsi

$$H_0 : \{\boldsymbol{X}\vartheta \in H\} \quad \text{contre} \quad H_1 : \{\boldsymbol{X}\vartheta \in E \setminus H\}.$$

Sous H_0, on estime l'effet moyen par la projection de Y sur H c'est à dire $Y_H = \boldsymbol{X}_0 \hat{\vartheta}_0$ avec $\hat{\vartheta}_0 = (\boldsymbol{X}_0^t \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^t Y$. On procède ensuite comme pour le test d'homogénéité : sous H_0, $||Y_E - Y_H||^2 \sim \sigma^2 \chi^2(p-q)$ mais σ est inconnu. On prend le rapport avec la résiduelle normalisée qui, elle, suit toujours la loi $\chi^2(n-p-1)$, pour construire la statistique de test

$$F = \frac{||Y_E - Y_H||^2/(p-q)}{||Y - Y_E||^2/(n-p-1)} \sim \mathcal{F}(p-q, n-p-1) \quad \text{sous } H_0.$$

La loi du χ^2 du numérateur (normalisé convenablement) se décentre sous l'hypothèse alternative, d'où le test au niveau α qui conduit à

$$\text{rejeter } H_0 \text{ dès que } F > q_{1-\alpha}(\mathcal{F}(p-q, n-p-1)).$$

Table d'analyse de la variance pour le modèle de régression

Lorsqu'ils traitent un modèle de régression, la plupart des logiciels de statistique calculent les estimateurs des paramètres et effectuent des tests individuels de nullité de ces paramètres (p tests de Student de $H_0 : \alpha_j = 0$, $j = 1, \ldots, p$, fondés sur les lois que nous avons donné plus haut). Ils fournissent également une *table d'analyse de variance associée au modèle de régression*. Il s'agit du résultat du test de Fisher pour l'hypothèse nulle "pas de modèle de régression", autrement dit "aucun régresseur n'est significatif". C'est la réalisation du test ci-dessus pour $H = \{\lambda \mathbf{1}_n, \lambda \in \mathbb{R}\}$.

Coefficient de détermination

Lorsque il y a une constante dans la régression, on appelle coefficient de détermination, ou R^2, le nombre

$$R^2 = \frac{||Y_E - \bar{Y}\mathbf{1}_n||^2}{||Y - \bar{Y}\mathbf{1}_n||^2} \in [0, 1].$$

C'est un indicateur de la "qualité" de la régression : plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données (on parle aussi de pourcentage de la variabilité expliquée par le modèle de régression).

Remarquons que pour le test de Fisher associé à l'hypothèse nulle "aucun régresseur n'est significatif", le sous-espace vectoriel H est celui engendré par $\mathbf{1}_n$ ce qui entraîne que $Y_H = \bar{Y}\mathbf{1}_n$. Dans ce cas il existe un lien simple entre le R^2 et la statistique du test de Fisher :

$$F = \frac{(n-(p+1))}{p}\frac{R^2}{1-R^2}.$$

Exemple 4.5. ***La régression simple****, (suite et fin de l'exemple 4.4).*
Nous terminons l'étude détaillée de la régression simple avec le test de non effet du seul régresseur présent dans le modèle :

$$H_0 : \{\alpha = 0\} \quad \textit{contre} \quad H_1 : \{\alpha \neq 0\}.$$

Remarquons que, ici, il est possible de construire ce test de deux manières : à partir de la loi de $\hat{\alpha}$ en utilisant la loi de Student (qui provient, rappelons-le, de l'obligation d'estimer σ^2 par la résiduelle), ou bien à partir du test de Fisher. On vérifie que les statistiques de ces deux tests sont liées par la relation $F = T^2$, et ils donnent la même p-valeur. Nous allons utiliser ici la seconde méthode.
Sous H_0, le modèle est simplement $Y = \beta\mathbf{1}_n + \varepsilon$ (il s'agit donc d'un n-échantillon de $\mathcal{N}(\beta, \sigma^2)$), et $Y_H = \bar{Y}\mathbf{1}_n$. Nous avons déjà précisé l'expression de la résiduelle dans ce cas. La "somme des carrés du modèle" est

$$SSM = ||Y_E - Y_H||^2 = \sum_{i=1}^{n}(\hat{\beta} + \hat{\alpha}R_i - \bar{Y})^2,$$

et la statistique de test

$$F = \frac{||Y_E - Y_H||^2}{||Y - Y_E||^2/(n-2)} \sim \mathcal{F}(1, n-2) \quad \textit{sous } H_0.$$

On rejette donc H_0 au niveau γ si $F > q_{1-\gamma}(\mathcal{F}(1, n-2))$. Enfin, si on a observé la valeur f de la statistique F, la p-valeur de ce test est $\boldsymbol{P}(F > f)$, où $F \sim \mathcal{F}(1, n-2)$.
Dans le cas de la régression simple, le coefficient de détermination $R^2 = SSM/SST$ est aussi le carré du coefficient de corrélation entre Y et R.

Exemple 4.6. *Si on reprend l'exemple 4.2, on obtient les résultats suivants :*
– *Les estimations des paramètres valent : $\hat{\alpha} = 4.55$ et $\hat{\beta} = -128.07$. Sur le graphique (Fig. 4.3) on a représenté la droite de régression.*

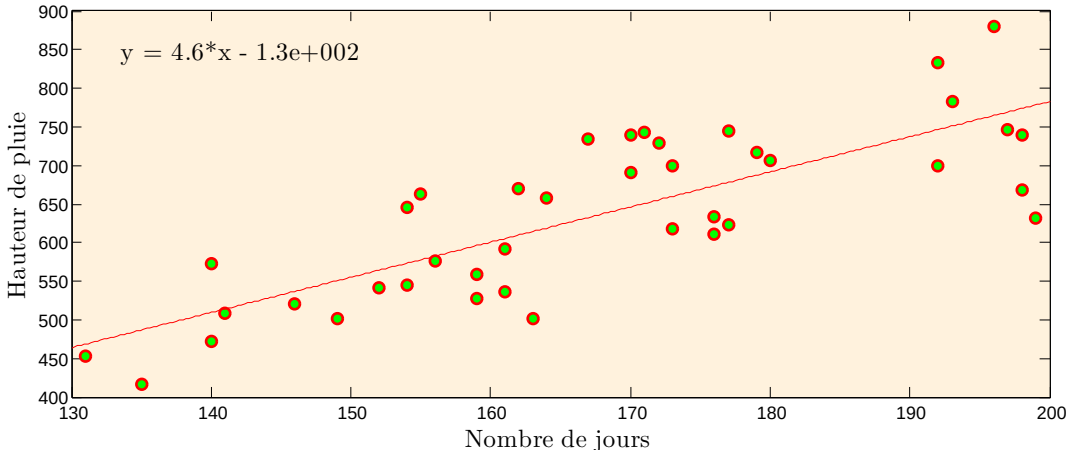


FIGURE 4.3 – Droite de régression sur le nuage de points

– *Les intervalles de confiance de Student sont :* $I_{0.05}(\alpha) = [3.40; 5.70]$ *et* $I_{0.05}(\beta) = [-322; 66]$
– *Le calcul du* R^2 *et du test de* $H_0 : \{\alpha = 0\}$ *donnent :*

R^2	`Fisher`	p-`valeur`
`0.6294`	`64.52`	$< 10^{-4}$

donc on rejette clairement H_0.

4.5 Exercices

4.6 Résumé du Chapitre 4

4.6.1 Le modèle gaussien à variance connue

1. Modèle : $(Y_k, 1 \leq k \leq n)$ suite de v.a. i.i.d. de loi gaussienne à variance, σ_0^2, connue : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0^2), \mu \in \mathbb{R}\}$.
2. $H_0 : \{\mu = \mu_0\}$, $H_1 : \{\mu \neq \mu_0\}$, avec $\mu_0 \in \mathbb{R}$.
3. Statistique de test : $\zeta_n = \sqrt{n}\dfrac{\bar{Y}_n - \mu_0}{\sigma_0}$.
4. Loi sous H_0 : $\mathcal{N}(0,1)$.
5. Loi sous H_1 : gaussienne réduite décentrée.
6. Région critique : $W_n = \{|\zeta_n| \geq a\}$.
7. Niveau exact α : $a = q^N_{1-\alpha/2}$, où $q^N_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0,1)$.
8. Test convergent.
9. p-valeur : $\boldsymbol{P}(|G| \geq |\zeta_n^{\text{obs}}|)$ où G de loi $\mathcal{N}(0,1)$.
10. Variante : $H_0 : \{\mu \leq \mu_0\}$, $H_1 : \{\mu > \mu_0\}$. Même statistique de test. Région critique : $W_n = \{\zeta_n \geq a\}$. Niveau exact α : $a = q^N_{1-\alpha}$. Test convergent. p-valeur : $\boldsymbol{P}(G \geq \zeta_n^{\text{obs}})$.

4.6.2 Le modèle gaussien à variance inconnue

1. Modèle : $(Y_k, 1 \leq k \leq n)$ suite de v.a. i.i.d. de loi gaussienne : $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$.
2. $H_0 = \{\mu = \mu_0\}$, $H_1 : \{\mu \neq \mu_0\}$, avec $\mu_0 \in \mathbb{R}$.
3. Statistique de test : $\zeta_n = \sqrt{n}\dfrac{\bar{Y}_n - \mu_0}{S_n}$.
4. Loi sous H_0 : Student de paramètre $n - 1$.
5. Comportement asymptotique sous H_1 : ζ_n converge p.s. vers $-\infty$ ou $+\infty$.
6. Région critique : $W_n = \{|\zeta_n| \geq a\}$.
7. Niveau exact α : $a = q_{1-\alpha/2}(t_{n-1})$, où $q_{1-\alpha/2}(t_{n-1})$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student de paramètre $n - 1$.
8. Test convergent.
9. p-valeur : $\boldsymbol{P}(|T| \geq \zeta_n^{\text{obs}})$ où T de loi de Student de paramètre $n - 1$.
10. Variante : $H_0 : \{\mu \leq \mu_0\}$, $H_1 : \{\mu > \mu_0\}$. Région critique : $W_n = \{\zeta_n \geq a\}$. Niveau exact α : $a = q_{1-\alpha}(t_{n-1})$. Test convergent. p-valeur : $\boldsymbol{P}(T \geq \zeta_n^{\text{obs}})$.

4.6.3 Régression multiple

1. Modèle : pour $i = 1 \dots n$
$$Y_i = \beta + \sum_{j=1}^{p} \alpha_j R_i^j + \varepsilon_i.$$
Les v.a. $\varepsilon_i, i = 1 \dots n$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Les coefficients de la régression $\beta, \alpha_1, \dots, \alpha_p$ et la variance σ^2 sont inconnues.
2. $H_0 : \{\alpha_{q+1} = \dots = \alpha_p = 0\}$ (les $p - q$ régresseurs $R^{q+1}, \dots, R^p$ sont inutiles),
$H_1 : \{\exists j \in \{q+1, \dots, p\}, \alpha_j \neq 0\}$ (un au moins des $p - q$ régresseurs $R^{q+1}, \dots, R^p$ est utile).
3. Statistique de test :
$$F = \frac{||Y_E - Y_H||^2/(p-q)}{||Y - Y_E||^2/(n-p-1)},$$
où Y_E est la projection orthogonale de Y sur l'espace vectoriel, E, engendré par $\mathbf{1}, R^1, \dots, R^p$, et Y_H est la projection orthogonale de Y sur l'espace vectoriel, H, engendré par $\mathbf{1}, R^1, \dots, R^q$.
4. Comportement sous H_0 : F suit une loi de Fischer : $\mathcal{F}(p-q, n-p-1)$.
5. Comportement sous H_1 : $F \to \infty$ quand $n \to \infty$.
6. Région critique : $W_n = \{F > a\}$.
7. Niveau α : $a = q_{1-\alpha}(\mathcal{F}(p-q, n-p-1))$, où $q_{1-\alpha}(\mathcal{F}(p-q, n-p-1))$ est le quantile d'ordre $1 - \alpha$ de la loi de Fisher $\mathcal{F}(p-q, n-p-1)$.
8. Le test est convergent.
9. p-valeur : $\boldsymbol{P}(F \geq f^{\text{obs}})$.

5
Tests d'adéquation

5.1 Introduction

En pratique, dans la plupart des situations, il est impossible de savoir quelle est la loi de probabilité des données que nous souhaitons analyser. Au chapitre 1, nous avons vu quelques outils graphiques—histogramme, fonction de répartition empirique, QQ-plot—offrant une évaluation visuelle de la pertinence de modélisation des données observées par telle ou telle loi. Le but de ce chapitre est de fournir des méthodes statistiques permettant une évaluation quantitative de la modélisation des données par une loi donnée ou par une famille des lois.

5.2 Tests du chi-deux

5.2.1 Test d'adéquation à une loi discrète

Le problème

On observe n v.a. $(X_i)_{1\leq i\leq n}$, indépendantes et de même loi, à valeurs dans un espace fini $A = \{a_1, \ldots, a_k\}$. Cette loi, inconnue, est caractérisée par la suite $\underline{p} = (p_1, \ldots p_k)$ (avec $\sum_{j=1}^k p_j = 1$), où pour tout $j = 1, \ldots, k$, la quantité p_j désigne la probabilité d'observer a_j (indépendante de i en raison de l'identique distribution des X_i) ; soit $p_j = \boldsymbol{P}(X_i = a_j)$. La loi jointe du n-uplet $\underline{X} = (X_i)_{1\leq i\leq n}$ est : pour tout $(x_1, \cdots, x_n) \in A^n$,

$$\boldsymbol{P}_{\underline{p}}(X_i = x_i,\ \ 1 \leq i \leq n) = \prod_{i=1}^{n} \boldsymbol{P}_{\underline{p}}(X_i = x_i) = \prod_{j=1}^{k} p_j^{\mathrm{card}(\{i\,;\, x_i = a_j\})}.$$

Remarque 5.1. *Il en est ainsi, par exemple, si on procède à un sondage dans une population divisée en k catégories, les tirages des n individus pouvant être considérés comme indépendants, et, à chaque fois, la probabilité d'être dans une catégorie donnée étant égale à la proportion (inconnue) d'individus de cette catégorie dans la population totale. C'est bien le cas si on effectue des tirages "avec remises" et "brassage" de la population, mais un tel "modèle d'urne", quoique traditionnel, n'est pas très réaliste. Cependant, on peut considérer qu'on est approximativement dans le modèle proposé si on fait porter*

le tirage sur des individus distincts (tirage "sans remise") mais dans un contexte où la taille totale de la population est très grande par rapport à celle de l'échantillon.

On avance l'hypothèse que le paramètre est $\underline{p}^0 = (p_1^0, \ldots, p_k^0)$, où $p_j^0 > 0$, pour tout $j = 1, \ldots, k$. Le but est de tester, à un niveau donné α, cette hypothèse nulle simple, $H_0 : \underline{p} = \underline{p}^0$, contre l'hypothèse alternative $H_1 : \underline{p} \neq \underline{p}^0$.

Intuitions

Pour tout $j = 1, \ldots, k$ on note $N_j = \text{card}(\{i : X_i = a_j\}) = \sum_{i=1}^n \mathbf{1}_{\{X_i = a_j\}}$ la variable aléatoire de comptage du nombres de fois où l'état a_j est visité par les v.a. X_i, $i = 1, \ldots, n$. La v.a. N_j suit une loi binomiale de paramètres (n, p_j). On rappelle que $\mathbf{E}[N_j] = np_j$, que la v.a. $\hat{P}_j = \frac{N_j}{n}$ est un **estimateur convergent sans biais** de p_j.

Il y a donc lieu de penser que, s'il est vrai que $\underline{p} = \underline{p}^0$, la suite des effectifs observés $n_j = \text{card}(\{i : x_i = a_j\})$ sera telle que la suite des fréquences observées, $\underline{\hat{p}} = (\hat{p}_1, \ldots, \hat{p}_k) = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$, sera "proche" (en raison de la loi forte des grands nombres citée précèdemment) de la suite mise en test $\underline{p}^0 = (p_1^0, \ldots, p_k^0)$.

Avec cette notation, il vient que $\boldsymbol{P}_{\underline{p}}(X_i = x_i,\ 1 \leq i \leq n) = \prod_{j=1}^k p_j^{n_j}$. On peut en déduire que $\underline{\hat{p}}$ est l'estimation par maximum de vraisemblance de $\underline{p}$, ce qui justifie que nous fassions porter notre test sur cette suite des fréquences observées $\underline{\hat{P}} = (\hat{P}_j)_{1 \leq j \leq k}$.

On souhaite donc pouvoir caractériser une "distance" entre la suite des fréquences observées $\underline{\hat{p}}$ et la suite des fréquences théoriques $\underline{p}^0$, de manière à rejeter l'hypothèse nulle si cette distance est supérieure à une certaine valeur frontière. Pour réaliser ce programme, il faut que :

- la loi, sous l'hypothèse nulle, de cette distance soit (au moins approximativement) connue de sorte que la frontière sera le quantile d'ordre $1 - \alpha$ de cette loi (le rejet à tort de l'hypothèse nulle sera bien alors de probabilité approximativement égale à α),
- **si l'hypothèse nulle n'est pas satisfaite**, cette distance ait tendance à prendre des valeurs d'autant plus grandes que la vraie valeur du paramètre $\underline{p}$ est plus "éloignée" de $\underline{p}^0$ (ce qui, là aussi, conduit à souhaiter disposer d'une distance entre $\underline{p}$ et $\underline{p}^0$, gouvernant la loi de la distance entre la v.a. $\underline{\hat{P}}$ et $\underline{p}^0$).

Outils

On définit la **distance du χ^2** (ou **distance du chi-deux**) , entre deux probabilités sur un ensemble fini à k éléments, $\underline{p} = (p_j)_{1 \leq j \leq k}$ et $\underline{q} = (q_j)_{1 \leq j \leq k}$, par :

$$D(\underline{p}, \underline{q}) = \sum_{j=1}^k \frac{(p_j - q_j)^2}{q_j} \, .$$

Remarquons que, faute de symétrie entre $\underline{p}$ et $\underline{q}$, cet objet n'est pas une "distance" au sens mathématique traditionnel du terme (on parle parfois de "pseudo-distance" du χ^2).

On démontre (*nous l'admettrons*) que, **si l'hypothèse nulle est satisfaite**, la loi de la v.a. $n.D(\underline{\hat{P}}, \underline{p}^0)$ tend, quand n tend vers l'infini, vers la loi du chi-deux à $k - 1$ degrés de liberté. Ceci conduit, pour n "assez grand" (notion qui sera précisée empiriquement dans la suite),

à fonder sur $n.D(\underline{\hat{P}}, \underline{p}^0)$ le test, au niveau α, de l'hypothèse $H_0 = \{\underline{p} = \underline{p}^0\}$, le rejet ayant lieu si

$$n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j^0)^2}{p_j^0} \geq \chi^2_{k-1,1-\alpha},$$

où $q_{1-\alpha}(\chi^2_{k-1})$ désigne le quantile d'ordre $1-\alpha$ de la loi du chi-deux à $k-1$ degrés de liberté, disponible dans des tables ou via les ordinateurs. C'est ce que l'on appelle le **test du χ^2**.

Critère pratique. *On considère souvent que l'approximation fournie par la loi du χ^2 à $k-1$ degrés de liberté pour la loi de $n.D(\underline{\hat{P}}, \underline{p}^0)$ est valide si tous les produits $np_j^0(1-p_j^0)$ sont supérieurs ou égaux à 5.*

Remarque 5.2 (Pour les lecteurs de niveau avancé)**.** *Intéressons nous maintenant à la puissance de ce test, c'est-à-dire considérons les situations où $\underline{p} \neq \underline{p}^0$. On démontre (nous l'admettrons) que, si la loi commune des v.a. X_i est caractérisée par la valeur $\underline{p}$ du paramètre, alors la loi de $n.D(\underline{\hat{P}}, \underline{p}^0)$ est bien approchée, quand n tend vers l'infini, par la loi dite du χ^2 décentré à $k-1$ degrés de liberté, $\chi^2_{k-1,\delta}$, avec pour coefficient d'excentricité $\delta = n.D(p, p_0)$.*

Il se produit alors une circonstance heureuse concernant la famille des lois $\chi^2_{k-1,\delta}$: elle est, à nombre de degrés de liberté fixé (ici $k-1$) stochastiquement croissante avec le coefficient d'excentricité δ, c'est-à-dire que, pour tout $t > 0$, la probabilité qu'une v.a. suivant la loi $\chi^2_{k-1,\delta}$ dépasse t est fonction croissante de δ. Afin d'illustrer davantage le phénomène d'excentricité engendré par δ nous pouvons rappeler que $\mathbf{E}[\chi^2_{k,\delta}] = k + \delta$ et $\mathrm{Var}(\chi^2_{k,\delta}) = 2(k+2\delta)$.

5.2.2 Test d'adéquation à une famille de lois discrètes

Présentation générale

Le modèle est ici le même qu'en 5.2.1 : on observe n v.a. X_i, indépendantes et de même loi, à valeurs dans un espace fini, soit $A = \{a_1, \ldots, a_k\}$. Cette loi, inconnue, est caractérisée par la suite $\underline{p} = (p_1, \ldots p_k)$, où, pour tout j (avec $1 \leq j \leq k$), p_j désigne la probabilité d'observer a_j.

Ici l'hypothèse à tester n'est plus réduite à une valeur bien déterminée $\underline{p}^0$, mais elle exprime que le paramètre appartient à une famille $(\underline{p}_\vartheta,\ \vartheta \in \Theta)$, où l'on note $\underline{p}_\vartheta = (p_{1,\vartheta}, \ldots, p_{k,\vartheta})$ un vecteur de poids de probabilité indexé par un paramètre ϑ. Attention : Θ n'est pas ici l'ensemble des paramètres du modèle tout entier mais paramétrise seulement l'hypothèse nulle.

Une idée naturelle est de reprendre la méthode du test d'adéquation vue en 5.2.1 en y remplaçant $\underline{p}^0$ par $\underline{p}_{\hat{\vartheta}}$, où $\hat{\vartheta}$ est une estimation de ϑ. C'est ce que l'on appelle un **test du χ^2 adaptatif**. On démontre alors que **si l'ensemble Θ des valeurs possibles pour ϑ est une partie ouverte d'intérieur non vide de $\mathbb{R}^h$** (avec $h < k-1$) la loi de $nD(\underline{\hat{P}}, \underline{p}_{\hat{\vartheta}})$ tend, sous l'hypothèse nulle, vers la loi du χ^2 à $k-h-1$ degrés de liberté, sous des conditions de régularité que nous ne préciserons pas ici, mais qui sont satisfaites si $\hat{\vartheta}$ est une estimation par maximum de vraisemblance. Donc **on procède comme dans le test du χ^2 d'adéquation, en remplaçant seulement le nombre de degrés de liberté $k-1$ par $k-h-1$.**

Exemple : test du χ^2 d'indépendance

Les v.a. i.i.d. X_i sont ici de la forme (Y_i, Z_i), où les "premières composantes" Y_i sont à valeurs dans $A = \{a_1, \ldots, a_k\}$, et les "secondes composantes" Z_i sont à valeurs dans $B = \{b_1, \ldots, b_m\}$.

On note, pour tout $j = 1, \dots, k$, et tout $\ell = 1, \dots, m$, $p_{j,\ell} = \boldsymbol{P}((Y_i, Z_i) = (a_j, b_\ell))$. Le paramètre est donc $\underline{p} = (p_{j,\ell})_{1 \le j \le k, 1 \le \ell \le m}$.
On veut tester l'hypothèse que les 2 composantes sont indépendantes, autrement dit que la loi commune des couples (Y_i, Z_i) est une loi produit, c'est-à-dire encore que tous les $p_{j,\ell}$ sont de la forme :

$$\forall (j,\ell) \in A \times B, \quad p_{j,\ell} = \boldsymbol{P}(Y_i = a_j, Z_i = b_\ell) = \boldsymbol{P}(Y_i = a_j)\boldsymbol{P}(Z_i = b_\ell) = q_j r_\ell,$$

où nécessairement, pour tout j, $q_j = \sum_{\ell=1}^{m} p_{j,\ell}$ et, pour tout ℓ, $r_\ell = \sum_{j=1}^{k} p_{j,\ell}$. Les q_j caractérisent la loi commune des v.a. Y_i et les r_ℓ caractérisent la loi commune des v.a. Z_i ; ces lois sont appelées aussi première et seconde lois marginales des X_i.

Ainsi, **sous l'hypothèse nulle**, le paramètre, caractérisé d'une part par les k valeurs q_j (de somme égale à 1) et d'autre part par les m valeurs r_ℓ (aussi de somme égale à 1), appartient à un espace de dimension $h = k + m - 2$. On supposera que les q_j et les r_ℓ sont tous non nuls, ce qui assure que, sous l'hypothèse nulle, l'ensemble de paramétrage est une partie ouverte de $\mathbb{R}^{k+m-2}$

Étant observé un échantillon de taille n, soit $(y_i, z_i)_{1 \le i \le n}$, notons, pour tout couple (j,ℓ), $n_{j,\ell}$ l'effectif des observations égales à (a_j, b_ℓ) et $\hat{p}_{j,\ell}$ leur fréquence ($\hat{p}_{j,\ell} = \frac{n_{j,\ell}}{n}$). On estime alors chaque q_j de la première marge par la fréquence marginale correspondante $\hat{q}_j = \frac{1}{n}\sum_{\ell=1}^{m} n_{j,\ell}$ et de même, pour la seconde marge, chaque r_ℓ par la fréquence marginale correspondante $\hat{r}_\ell = \frac{1}{n}\sum_{j=1}^{k} n_{j,\ell}$.

Alors, **si l'hypothèse nulle est satisfaite**, on estime, pour tout couple (j,ℓ), $p_{j,\ell}$, par le produit des fréquences marginales $\hat{q}_j \hat{r}_\ell$ (pour mimer la formule d'indépendance citée plus haut).

Nous admettons que les conditions de validité de la méthode sont satisfaites, $\hat{q}_j$ et $\hat{r}_\ell$ étant respectivement des estimateurs par maximum de vraisemblance de q_j et r_ℓ. Le test, au seuil α, consiste donc à rejeter l'hypothèse d'indépendance si :

$$n \sum_{j=1}^{k} \sum_{\ell=1}^{m} \frac{(\hat{p}_{j,\ell} - \hat{q}_j \hat{r}_\ell)^2}{\hat{q}_j \hat{r}_\ell} \ge q_{1-\alpha}(\chi^2_{(k-1)(m-1)}),$$

autrement dit

$$n \sum_{j=1}^{k} \sum_{\ell=1}^{m} \frac{(\frac{n_{j,\ell}}{n} - \frac{n'_j . n''_\ell}{n^2})^2}{\frac{n'_j . n''_\ell}{n^2}} \ge q_{1-\alpha}(\chi^2_{(k-1)(m-1)}),$$

où :
- $n_{j,\ell}$ est le nombre d'observations égales à (a_j, b_ℓ),
- $n'_j = \sum_{\ell=1}^{m} n_{j,\ell}$ est le nombre d'observations dont la première composante est égale à a_j,
- $n''_\ell = \sum_{j=1}^{k} n_{j,\ell}$ est le nombre d'observations dont la seconde composante est égale à b_ℓ,
- $q_{1-\alpha}(\chi^2_{(k-1)(m-1)})$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $(k-1)(m-1)$ degrés de liberté (en effet $km - (k+m-2) - 1 = (k-1)(m-1)$).

5.3 Test de Kolmogorov

C'est un test d'ajustement à une loi, comme le test du χ^2, mais qui s'applique à une variable continue. On veut tester l'hypothèse selon laquelle les données observées sont tirées d'une loi dont la fonction de répartition est F_0. Dans toute cette section, on considère que la vraie fonction de répartion inconnue F^* et F_0 sont continues.

Le test est basé sur la différence entre la fonction de répartition F_0 de cette loi théorique et la fonction de répartition empirique $\hat{F}_n$ dont on rappelle la définition :

Définition 5.1. *On définit la fonction de répartition empirique du n-échantillon $(X_1, ..., X_n)$, par la fonction en escalier suivante :*

$$\hat{F}_n(t) = \frac{Card(\{1 \leq i \leq n : X_i \leq t\})}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}.$$

Remarque 5.3. *Notons que $\hat{F}_n$ est continue à droite.*

Le test de Kolmogorov [1] permet de tester l'hypothèse H_0 : "Les observations sont un échantillon de la loi F_0" contre sa négation. La statistique D_n de ce test est alors basée sur la distance maximale entre F_0 et $\hat{F}_n$, c'est à dire :

$$D_n = \sup_{t \in \mathbb{R}} \left| F_0(t) - \hat{F}_n(t) \right|.$$

Il s'agit d'un choix de distance raisonnable, car d'après le théorème de Glivenko-Cantelli, sous H_0, D_n converge presque sûrement vers 0 lorsque n tend vers l'infini. La zone de rejet est alors de la forme : $\{D_n > a\}$. Notons que comme $\hat{F}_n$ est constante et égale à i/n sur l'intervalle $[X_{(i)}, X_{(i+1)}[$ tandis que F_0 est croissante sur cet intervalle,

$$\sup_{t \in [X_{(i)}, X_{(i+1)}[} \left| F_0(t) - \hat{F}_n(t) \right| = \max\left(\left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i+1)}) - \frac{i}{n} \right| \right).$$

On en déduit l'expression suivante très utile en pratique

$$\boxed{D_n = \max_{1 \leq i \leq n} \max\left(\left| F_0(X_{(i)}) - \frac{i-1}{n} \right|, \left| F_0(X_{(i)}) - \frac{i}{n} \right| \right).}$$

La légitimité du choix de D_n comme statistique de test repose sur la proposition suivante :

Proposition 5.1. *Sous H_0, la loi de D_n ne dépend pas de F^*. On dit alors que D_n est une statistique libre.*

Démonstration. On vérifie facilement que

$$D_n = \sup_{t \in \mathbb{R}} \left| F_0(t) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq F_0(t)\}} \right|$$

où les variables $U_i = F_0(X_i)$ sont i.i.d. suivant la loi uniforme sur $[0, 1]$. Il suffit ensuite de faire le changement de variable $u = F_0(t)$ pour conclure. □

La loi de D_n sous H_0 a été tabulée, ce qui donne des valeurs seuils a_α à ne pas dépasser pour que H_0 soit acceptable au niveau α. Les moyens actuels de calcul informatique permettent également d'approcher la loi de D_n à l'aide de simulations. Pour n grand, il existe une approximation décrite par la proposition suivante :

1. Ce test est également appelé test de Kolmogorov-Smirnov à un échantillon

Proposition 5.2. *Sous H_0, en posant $\zeta_n = \sqrt{n}D_n$, on dispose du résultat asymptotique suivant : la suite $(\zeta_n, n \geq 1)$ converge en loi et pour tout $y > 0$, on a*

$$\mathbb{P}(\zeta_n \leq y) \xrightarrow[n\to\infty]{} \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2y^2\right).$$

Démonstration. Comme pour $t \in \mathbb{R}$, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}}$ où les variables $\mathbf{1}_{\{X_i \leq t\}}$ sont i.i.d. suivant la loi de Bernoulli $\mathcal{B}(F_0(t))$, le TCL entraîne que $\sqrt{n}(F_0(t) - \hat{F}_n(t))$ converge en loi vers Y_t de loi normale centrée $\mathcal{N}(0, F_0(t)(1 - F_0(t)))$. Plus généralement, le théorème de la limite centrale multidimensionnel assure que $\sqrt{n}(F_0(t_1) - \hat{F}_n(t_1), \ldots, F_0(t_k) - \hat{F}_n(t_k))$ converge en loi vers un vecteur gaussien centré $(Y_{t_1}, \ldots, Y_{t_k})$ de covariance donnée par $\mathrm{Cov}(Y_{t_i}, Y_{t_j}) = F_0(\min(t_i, t_j)) - F_0(t_i)F_0(t_j)$. En fait on montre que le processus $\sqrt{n}(F_0(t) - \hat{F}_n(t))_{t\in\mathbb{R}}$ converge en loi vers "un processus gaussien centré" tel que $\mathrm{Cov}(Y_s, Y_t) = F_0(\min(s,t)) - F_0(s)F_0(t)$ et on montre que pour tout $y > 0$,

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |Y_t| \leq y\right) = \sum_{-\infty}^{+\infty} (-1)^k \exp\left(-2k^2y^2\right).$$

□

Proposition 5.3. *Sous H_1, $\zeta_n = \sqrt{n}D_n$ tend p.s. vers $+\infty$ avec n.*

Le test est donc nécessairement unilatéral à droite (rejet des valeurs trop grandes).

Démonstration. Sous H_1 la fonction de répartition commune des X_i, notée F est différente de F_0. Soit $t_1 \in \mathbb{R}$ tel que $F_0(t_1) \neq F(t_1)$. D'après la loi forte des grands nombres $\hat{F}_n(t_1) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t_1\}}$ converge p.s. vers $\mathbb{E}\left[\mathbf{1}_{\{X_i \leq t_1\}}\right] = F(t_1)$. Donc $\sqrt{n}|F_0(t_1) - \hat{F}_n(t_1)|$ tend p.s. vers $+\infty$ de même pour $\sqrt{n}D_n$. □

Remarque 5.4. *Si F_0 est **non continue** (par exemple lorsqu'il s'agit d'une loi discrète), le test de Kolmogorov sous sa forme classique n'est pas valide (la proposition 5.2 n'est valable que si F_0 est continue) : on peut montrer que D_n est alors plus «concentrée» à proximité de zéro que quand F est continue.*

Remarque 5.5. *On peut aussi envisager des contre-hypothèses plus fines, du type unilatéral : «la loi des données a une fontion de répartition F telle que $F \prec F_0$ au sens où $\forall t \in \mathbb{R}$, $F(t) \leq F_0(t)$ et $\exists t_0 \in \mathbb{R}$, $F(t_0) < F_0(t_0)$». Dans ce cas, la statistique de test s'écrit sans la valeur absolue (et sa loi est différente).*

5.3.1 Un exemple

On dispose des 10 données suivantes :

$$\underline{x} = (2.2\ ,\ 3.3\ ,\ 5.3\ ,\ 1.8\ ,\ 4.3\ ,\ 6.3\ ,\ 4.5\ ,\ 3.8\ ,\ 6.4\ ,\ 5.5)$$

La question naïve « ces observations proviennent-elles d'une loi normale de moyenne 4 et de variance 4 ? » va être formalisée sous l'énoncé : « tester, au niveau de signification 0.05, l'hypothèse nulle selon laquelle ces observations, supposées indépendantes et identiquement distribuées, ont pour loi commune la loi normale de moyenne 4 et variance 4 ». On calcule la fonction empirique dessinée sur la figure 5.1. Elle montre que $D_{\underline{x}} = 0.163$, écart maximal

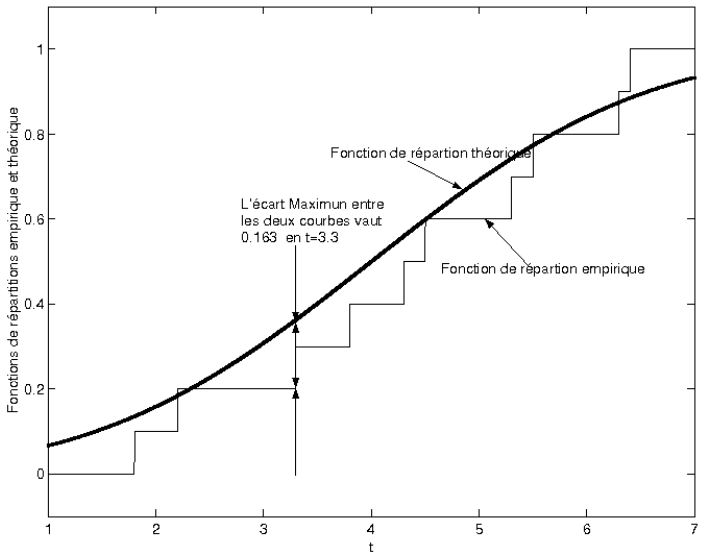


FIGURE 5.1 – Le test de Kolmogorov s'appuie sur la distance entre fonction de répartition empirique et théorique.

obtenu en $t = 3.3$. Cette valeur est-elle plausible, au niveau 0.05, sous l'hypothèse H_0 ? Les praticiens ont l'habitude de faire la transformation de l'axe des abscisses $u = F(t)$. Cette transformation permet de travailler dans le carré $[0,1] \times [0,1]$ (cf figure 5.2) où D_n mesure alors l'écart de la fonction de répartition empirique par rapport à la première bissectrice.

En utilisant une table ou bien en approchant les quantiles de la loi de D_n sous H_0 par simulation d'un grand nombre de réalisations suivant cette loi, on remarque que la valeur observée $D_{\underline{x}} = 0.163$ est inférieure au quantile d'ordre 0.95 de la loi de D_n : 0.410. (La p-valeur est de 0.963.)

L'hypothèse de référence H_0 est acceptée.

5.3.2 Test de normalité

Revenons à l'exemple des mesures de taux d'alcoolémie. On peut de la même manière tester H_0 : "Les données suivent une loi gaussienne de moyenne 23 et de variance 49" contre l'alternative : "c'est faux". On trouve $D_{\underline{x}} = 0.132$ donc on ne rejette pas H_0 pour les niveaux habituellement utilisés (quantile asymptotique d'ordre 0.95 égal à 0.242, et p-valeur asymptotique égale à 0.637). Dans ce problème on pourrait tester H_0 : "Les données suivent une loi gaussienne" contre l'alternative : "c'est faux", à l'aide du test de normalité de Lilliefors : ce test utilise la statistique de Kolmogorov déterminée par la distance entre la loi empirique et la loi gaussienne dont l'espérance est la moyenne empirique et la variance, la variance empirique. Les quantiles sont différents des quantiles du test de Kolmogorov et peuvent être calculés par simulation. Il existe de nombreux tests de normalité (test de Pearson construit avec une approche de discrétisation et un test du χ^2, test de Shapiro-Wilk, ...).

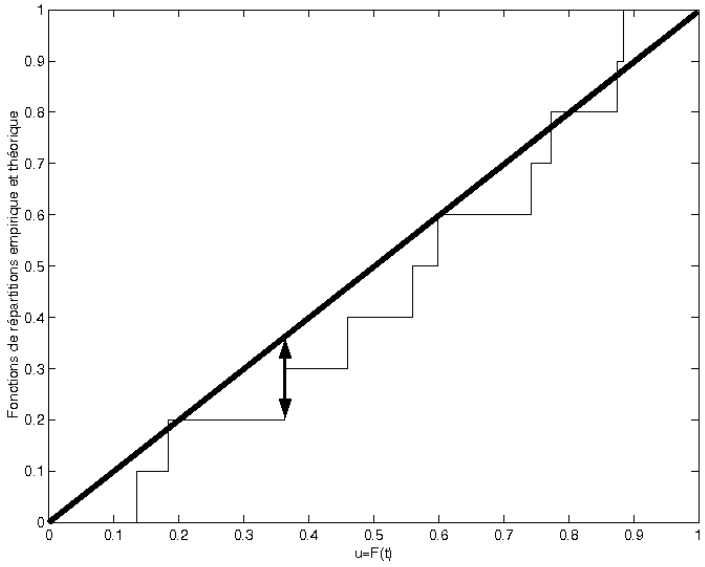


FIGURE 5.2 – Présentation usuelle de la distance de Kolmogorov.

5.4 Résumé du Chapitre 5

5.4.1 Test d'adéquation à une loi discrète : le test du χ^2

L'objectif est de déterminer si les données discrètes observées proviennent d'une loi donnée ou non.

1. Description du modèle : $(X_j, 1 \leq j \leq n)$ est une suite de v.a. i.i.d. à valeurs dans $A = \{a_1, \ldots, a_k\}$. Une loi P_p sur A est décrite par le paramètre $p = (p_1, \ldots, p_k)$, où $p_i = \boldsymbol{P}_p(X_1 = a_i)$.
2. Les hypothèses : $H_0 : p = p^0$ et $H_1 : p \neq p^0$, où p_0 est donné.
3. La statistique de test :
$$\zeta_n = n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0},$$
où $\hat{p}_i$ est le nombre d'occurrence de a_i divisé par n.
4. Sous H_0, $(\zeta_n, n \geq 1)$ converge en loi vers $\chi^2(k-1)$.
5. Sous H_1, $(\zeta_n, n \geq 1)$ diverge vers $+\infty$.
6. Région de critique du test asymptotique : $[a, +\infty[$.
7. Niveau asymptotique du test égal à α : a est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k-1)$.
8. Le test est convergent.
9. La p-valeur asymptotique est donnée par
$$p\text{-valeur} = \boldsymbol{P}(Z \geq \zeta_n^{\mathrm{obs}}),$$
où Z est de loi $\chi^2(k-1)$, et ζ_n^{obs} est la statistique de test calculée avec les observations.

Le test asymptotique est considéré valide si $np_i^0(1-p_i^0) \geq 5$ pour tout i.

5.4.2 Test d'indépendance entre deux variables qualitatives

L'objectif est de vérifier si deux variables catégorielles sont indépendantes ou non.

1. Description du modèle : $((Y_i, Z_i), 1 \leq i \leq n)$ est une suite de v.a. i.i.d. respectivement à valeurs dans $A = \{a_1, \ldots, a_k\}$ et $B = \{b_1, \ldots, b_m\}$. Une loi commune P_p des couples (Y_i, Z_i) sur (A, B) est décrite par le paramètre $p = (p_{j,h})_{1\leq j\leq k, 1\leq l\leq m}$ où $p_{j,l} = \boldsymbol{P}_p((Y_i, Z_i) = (a_j, b_l))$.
2. Les hypothèses : $H_0 = \{p_{j,l} = q_j r_l\}_{1\leq j\leq k, 1\leq l\leq m}$ et $H_1 = \{\exists j, l; p_{j,l} \neq q_j r_l\}$, où $q_j = \sum_{l=1}^{m} p_{j,l}$ et $r_l = \sum_{j=1}^{k} p_{j,l}$.
3. La statistique de test :
$$\zeta_n = n \sum_{j=1}^{k} \sum_{l=1}^{m} \frac{(\hat{p}_{j,l} - \hat{q}_j \hat{r}_l)^2}{\hat{q}_j \hat{r}_l},$$
où $\hat{p}_{j,l}$, $\hat{q}_j$ et $\hat{r}_l$ sont respectivement les nombres d'occurrence de (a_j, b_l), de a_j et de b_l divisé par n.
4. Sous H_0, $(\zeta_n, n \geq 1)$ converge en loi vers $\chi^2((k-1)(m-1))$.
5. Sous H_1, $(\zeta_n, n \geq 1)$ diverge vers $+\infty$.
6. Région de critique du test asymptotique : $[a, +\infty[$.
7. Niveau asymptotique du test égal à α : a est le quantile d'ordre $1-\alpha$ de la loi $\chi^2((k-1)(m-1))$.
8. Le test est convergent.
9. La p-valeur asymptotique est donnée par
$$p\text{-valeur} = \boldsymbol{P}(Z \geq \zeta_n^{\text{obs}}),$$
où Z est de loi $\chi^2((k-1)(m-1))$, et ζ_n^{obs} est la statistique de test calculée avec les observations.

Le test asymptotique est considéré valide si $n\hat{q}_j\hat{r}_l(1-\hat{q}_j\hat{r}_l) \geq 5$ pour tout (j, l).

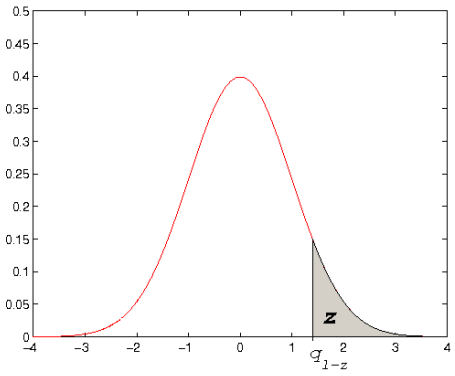
5.4.3 Test de Kolmogorov

1. Modèle non paramétrique : $(X_i, 1 \leq i \leq n)$ i.i.d. de fonction de répartition F continue.
2. Hypothèses : $H_0 : F^* = F_0$ et $H_1 : F^* \neq F_0$
3. Statistique de Kolmogorov
$$D_n = \max_{1\leq i\leq n} \max\left(|F_0(X_{(i)}) - \frac{i-1}{n}|, |F_0(X_{(i)}) - \frac{i}{n}|\right)$$
où $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ est le réordonnement croissant des X_i.
Statistique de test : $\zeta_n = \sqrt{n} D_n$.
4. Sous H_0, lorsque n tend vers l'infini, ζ_n converge en loi vers la loi de fonction de répartition $\mathbf{1}_{\{y>0\}} \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$.
5. Sous H_1, ζ_n tend p.s. vers $+\infty$.
6. Région critique : $[a, +\infty[$, avec $a > 0$.
7. Test convergent pour $n \to +\infty$.
8. Pour un niveau asymptotique α, a est donné par $\sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2a^2) = 1-\alpha$.

6
Tables numériques

6.1 Quantiles de la loi normale centrée réduite

La table suivante donne les valeurs numériques des quantiles $q_{1-z}(\mathcal{N})$ de la loi normale centrée réduite $\mathcal{N}(0,1)$. Rappelons que, par définition, $q_{1-z}(\mathcal{N})$ est l'unique solution de l'équation $\Phi(q) = 1 - z$, où $\Phi(\cdot)$ désigne la fonction de répartition de la loi normale centrée réduite.



Utilisation de la table : si, par exemple, on souhaite déterminer $q_{0.975}(\mathcal{N})$,
– on calcule $z = 1 - 0.975 = 0.025$ et on écrit $0.025 = 0.02 + 0.005$,
– on cherche la ligne correspondant à 0.02 et la colonne correspondant à 0.005,
– à l'intersection de la ligne et de la colonne trouvée on lit 1.96. Donc $q_{0.975}(\mathcal{N}) = 1.96$.

z	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.00	$+\infty$	3.090	2.878	2.748	2.652	2.576	2.512	2.457	2.409	2.366
0.01	2.326	2.290	2.257	2.226	2.197	2.170	2.144	2.120	2.097	2.075
0.02	2.054	2.034	2.014	1.995	1.977	1.960	1.943	1.927	1.911	1.896
0.03	1.881	1.866	1.852	1.838	1.825	1.812	1.799	1.787	1.774	1.762
0.04	1.751	1.739	1.728	1.717	1.706	1.695	1.685	1.675	1.665	1.655
0.05	1.645	1.635	1.626	1.616	1.607	1.598	1.589	1.580	1.572	1.563
0.06	1.555	1.546	1.538	1.530	1.522	1.514	1.506	1.499	1.491	1.483
0.07	1.476	1.468	1.461	1.454	1.447	1.440	1.433	1.426	1.419	1.412
0.08	1.405	1.398	1.392	1.385	1.379	1.372	1.366	1.359	1.353	1.347
0.09	1.341	1.335	1.329	1.323	1.317	1.311	1.305	1.299	1.293	1.287
0.10	1.282	1.276	1.270	1.265	1.259	1.254	1.248	1.243	1.237	1.232
0.11	1.227	1.221	1.216	1.211	1.206	1.200	1.195	1.190	1.185	1.180
0.12	1.175	1.170	1.165	1.160	1.155	1.150	1.146	1.141	1.136	1.131
0.13	1.126	1.122	1.117	1.112	1.108	1.103	1.098	1.094	1.089	1.085
0.14	1.080	1.076	1.071	1.067	1.063	1.058	1.054	1.049	1.045	1.041
0.15	1.036	1.032	1.028	1.024	1.019	1.015	1.011	1.007	1.003	0.999
0.16	0.994	0.990	0.986	0.982	0.978	0.974	0.970	0.966	0.962	0.958
0.17	0.954	0.950	0.946	0.942	0.938	0.935	0.931	0.927	0.923	0.919
0.18	0.915	0.912	0.908	0.904	0.900	0.896	0.893	0.889	0.885	0.881
0.19	0.878	0.874	0.870	0.867	0.863	0.860	0.856	0.852	0.849	0.845
0.20	0.842	0.838	0.834	0.831	0.827	0.824	0.820	0.817	0.813	0.810
0.21	0.806	0.803	0.799	0.796	0.792	0.789	0.786	0.782	0.779	0.776
0.22	0.772	0.769	0.765	0.762	0.759	0.755	0.752	0.749	0.745	0.742
0.23	0.739	0.736	0.732	0.729	0.726	0.722	0.719	0.716	0.713	0.709
0.24	0.706	0.703	0.700	0.697	0.693	0.690	0.687	0.684	0.681	0.678
0.25	0.674	0.671	0.668	0.665	0.662	0.659	0.656	0.653	0.649	0.646
0.26	0.643	0.640	0.637	0.634	0.631	0.628	0.625	0.622	0.619	0.616
0.27	0.613	0.610	0.607	0.604	0.601	0.598	0.595	0.592	0.589	0.586
0.28	0.583	0.580	0.577	0.574	0.571	0.568	0.565	0.562	0.559	0.556
0.29	0.553	0.550	0.548	0.545	0.542	0.539	0.536	0.533	0.530	0.527
0.30	0.524	0.521	0.519	0.516	0.513	0.510	0.507	0.504	0.501	0.499
0.31	0.496	0.493	0.490	0.487	0.484	0.482	0.479	0.476	0.473	0.470
0.32	0.468	0.465	0.462	0.459	0.456	0.454	0.451	0.448	0.445	0.443
0.33	0.440	0.437	0.434	0.432	0.429	0.426	0.423	0.421	0.418	0.415
0.34	0.413	0.410	0.407	0.404	0.402	0.399	0.396	0.393	0.391	0.388
0.35	0.385	0.383	0.380	0.377	0.374	0.372	0.369	0.366	0.364	0.361
0.36	0.358	0.356	0.353	0.350	0.348	0.345	0.342	0.340	0.337	0.334
0.37	0.332	0.329	0.327	0.324	0.321	0.319	0.316	0.313	0.311	0.308
0.38	0.305	0.303	0.300	0.298	0.295	0.292	0.290	0.287	0.284	0.282
0.39	0.279	0.277	0.274	0.271	0.269	0.266	0.264	0.261	0.258	0.256
0.40	0.253	0.251	0.248	0.246	0.243	0.240	0.238	0.235	0.233	0.230
0.41	0.227	0.225	0.222	0.220	0.217	0.215	0.212	0.210	0.207	0.204
0.42	0.202	0.199	0.197	0.194	0.192	0.189	0.187	0.184	0.181	0.179
0.43	0.176	0.174	0.171	0.169	0.166	0.164	0.161	0.159	0.156	0.153
0.44	0.151	0.148	0.146	0.143	0.141	0.138	0.136	0.133	0.131	0.128
0.45	0.126	0.123	0.121	0.118	0.116	0.113	0.110	0.108	0.105	0.103
0.46	0.100	0.098	0.095	0.093	0.090	0.088	0.085	0.083	0.080	0.078
0.47	0.075	0.073	0.070	0.068	0.065	0.063	0.060	0.058	0.055	0.053
0.48	0.050	0.048	0.045	0.043	0.040	0.038	0.035	0.033	0.030	0.028
0.49	0.025	0.023	0.020	0.017	0.015	0.012	0.010	0.007	0.005	0.002

6.2 Table de la loi du khi-deux

La table suivante donne la valeur du quantile d'ordre $(1-z)$ de la loi du khi-deux à k degrés de liberté. Par exemple, le quantile $q_{0.05}(t_{19})$ se trouve à l'intersection de la ligne 19 et de la colonne $1-0.05 = 0.95$, donc $q_{0.05}(t_{19}) = 10.12$. En d'autres termes, si $X \sim t_{19}$, alors $\boldsymbol{P}(X \leq 10.12) = 5\%$.

$k \backslash z$	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.65
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
33	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16	64.18
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.70	32.35	34.76	37.68	63.16	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.18	46.45	74.39	79.08	83.29	88.37	91.95
70	43.27	45.44	48.75	51.73	55.32	85.52	90.53	95.02	100.42	104.21
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17

6.3 Table de la loi de Student

La table suivante donne la valeur du quantile d'ordre $(1-z)$ de la loi de Student à k degrés de liberté. Par exemple, le quantile $q_{0.6}(t_{19})$ se trouve à l'intersection de la ligne 19 et de la colonne $1-0.6=0.4$, donc $q_{0.6}(t_{19})=0.2569$.

$k\backslash z$	0.450	0.400	0.300	0.200	0.100	0.050	0.025	0.010	0.005
8	0.1297	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.1293	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.1289	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.1286	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.1283	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.1281	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.1280	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.1278	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.1277	0.2576	0.5350	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.1276	0.2573	0.5344	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.1274	0.2571	0.5338	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.1274	0.2569	0.5333	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.1273	0.2567	0.5329	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.1272	0.2566	0.5325	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.1271	0.2564	0.5321	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.1271	0.2563	0.5317	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.1270	0.2562	0.5314	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.1269	0.2561	0.5312	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.1269	0.2560	0.5309	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.1268	0.2559	0.5306	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.1268	0.2558	0.5304	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.1268	0.2557	0.5302	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.1267	0.2556	0.5300	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.1267	0.2555	0.5298	0.8534	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.1267	0.2555	0.5297	0.8530	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.1266	0.2554	0.5295	0.8526	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.1266	0.2553	0.5294	0.8523	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.1266	0.2553	0.5292	0.8520	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.1266	0.2552	0.5291	0.8517	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.1265	0.2552	0.5289	0.8514	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.1265	0.2551	0.5288	0.8512	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.1265	0.2551	0.5287	0.8509	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.1265	0.2550	0.5286	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045
50	0.1263	0.2547	0.5278	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778
60	0.1262	0.2545	0.5272	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603
70	0.1261	0.2543	0.5268	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479
80	0.1261	0.2542	0.5265	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387
90	0.1260	0.2541	0.5263	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316
100	0.1260	0.2540	0.5261	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259

6.4 Quantiles pour le test de Kolmogorov

La Table ci-dessous contient les quantiles d'ordre $1-\alpha$ de la loi de la statistique $D_n = \sqrt{n}\sup_{x\in\mathbb{R}}|\hat{F}_n(x) - F^*(x)|$, utilisée dans le test d'adéquation de Kolmogorov.

n	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$	n	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 1\%$
1	0.9500	0.9750	0.9950	41	0.1869	0.2076	0.2490
2	0.7764	0.8419	0.9293	42	0.1847	0.2052	0.2461
3	0.6360	0.7076	0.8290	43	0.1826	0.2028	0.2433
4	0.5652	0.6239	0.7342	44	0.1806	0.2006	0.2406
5	0.5094	0.5633	0.6685	45	0.1786	0.1984	0.2380
6	0.4680	0.5193	0.6166	46	0.1767	0.1963	0.2354
7	0.4361	0.4834	0.5758	47	0.1748	0.1942	0.2330
8	0.4096	0.4543	0.5418	48	0.1730	0.1922	0.2306
9	0.3875	0.4300	0.5133	49	0.1713	0.1903	0.2283
10	0.3687	0.4093	0.4889	50	0.1696	0.1884	0.2260
11	0.3524	0.3912	0.4677	51	0.1680	0.1866	0.2239
12	0.3382	0.3754	0.4491	52	0.1664	0.1848	0.2217
13	0.3255	0.3614	0.4325	53	0.1648	0.1831	0.2197
14	0.3142	0.3489	0.4176	54	0.1633	0.1814	0.2177
15	0.3040	0.3376	0.4042	55	0.1619	0.1798	0.2157
16	0.2947	0.3273	0.3920	56	0.1605	0.1782	0.2138
17	0.2863	0.3180	0.3809	57	0.1591	0.1767	0.2120
18	0.2785	0.3094	0.3706	58	0.1577	0.1752	0.2102
19	0.2714	0.3014	0.3612	59	0.1564	0.1737	0.2084
20	0.2647	0.2941	0.3524	60	0.1551	0.1723	0.2067
21	0.2587	0.2873	0.3443	61	0.1539	0.1709	0.2051
22	0.2529	0.2809	0.3367	62	0.1526	0.1696	0.2034
23	0.2475	0.2749	0.3296	63	0.1515	0.1682	0.2018
24	0.2425	0.2693	0.3229	64	0.1503	0.1669	0.2003
25	0.2377	0.2641	0.3166	65	0.1492	0.1657	0.1988
26	0.2333	0.2591	0.3106	66	0.1480	0.1644	0.1973
27	0.2290	0.2544	0.3050	67	0.1469	0.1632	0.1958
28	0.2250	0.2500	0.2997	68	0.1459	0.1620	0.1944
29	0.2212	0.2457	0.2947	69	0.1448	0.1609	0.1930
30	0.2176	0.2417	0.2899	70	0.1438	0.1598	0.1917
31	0.2142	0.2379	0.2853	71	0.1428	0.1586	0.1903
32	0.2109	0.2343	0.2809	72	0.1418	0.1576	0.1890
33	0.2078	0.2308	0.2768	73	0.1409	0.1565	0.1878
34	0.2048	0.2275	0.2728	74	0.1399	0.1554	0.1865
35	0.2019	0.2243	0.2690	75	0.1390	0.1544	0.1853
36	0.1991	0.2212	0.2653	76	0.1381	0.1534	0.1841
37	0.1965	0.2183	0.2618	77	0.1372	0.1524	0.1829
38	0.1940	0.2155	0.2584	78	0.1364	0.1515	0.1817
39	0.1915	0.2127	0.2552	79	0.1355	0.1505	0.1806
40	0.1892	0.2101	0.2521	80	0.1347	0.1496	0.1795

TABLE 6.1 – Quantiles de la statistique de Kolmogorov