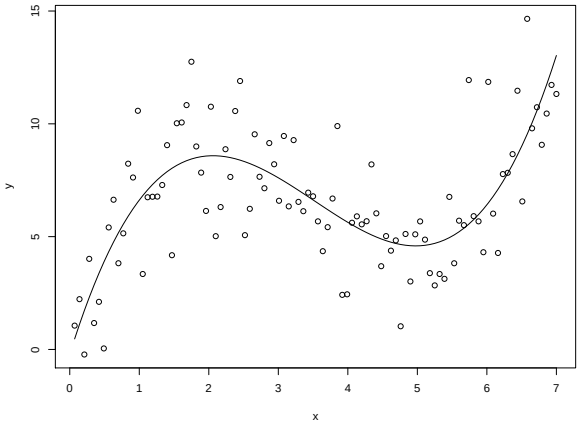
Ensimag - 1ère année



Principes et Méthodes Statistiques

Notes de cours

Olivier Gaudoin

Table des matières

Chapitre 1

Introduction

1.1 Définition et domaines d'application de la statistique

La **statistique** est la science dont l'objet est de recueillir, de traiter et d'analyser des **données** issues de l'observation de phénomènes **aléatoires**, c'est-à-dire dans lesquels le hasard intervient.

L'analyse des données est utilisée pour **décrire** les phénomènes étudiés, **faire des prévisions** et **prendre des décisions** à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie, et bien sûr les sciences de l'ingénieur.

Donnons quelques exemples d'utilisation de la statistique dans divers domaines.

- *Médecine, biologie* : diagnostic médical, imagerie médicale, essais thérapeutiques, épidémiologie, dynamique des populations, analyse du génôme, détection des maladies génétiques, impact des OGM ou des perturbateurs endocriniens, ...
- *Sciences de la terre, environnement* : prévisions météorologiques, analyse du réchauffement climatique, prévision de l'intensité et de la trajectoire des cyclones tropicaux, prévision des pics de pollution, exploration pétrolière, ...
- *Economie, assurance, finance* : prévisions économétriques, analyse de la consommation des ménages, fixation des primes d'assurance et franchises, études quantitatives de marchés, gestion de portefeuille, gestion des risques financiers, ...
- *Sciences humaines* : enquêtes d'opinion, sondages, démographie, études de populations, ...
- *Sciences de l'ingénieur* : voiture autonome, maîtrise des risques industriels, sûreté de fonctionnement (fiabilité, disponibilité, sécurité, maintenance,...), contrôle de qualité, maîtrise statistique des procédés (méthode "six-sigma"), évaluation des performances des systèmes complexes, ...
- *Sciences de l'information et de la communication* : traitement des images et des signaux, reconnaissance faciale, traitement automatique du langage naturel, analyse des grandes masses de données (big data), publicité ciblée sur le web, sys-

tèmes de recommandation, analyse des réseaux de communication,...
- *Physique* : mécanique statistique, théorie cinétique des gaz, astrophysique,...
- etc...

Le point fondamental est que les données sont entâchées d'**incertitudes** et présentent des **variations** pour plusieurs raisons :

- Le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude. Par exemple on ne sait pas prévoir avec certitude les cours de la bourse, la météo du week-end prochain, les pannes des voitures, etc.
- Toute mesure est entâchée d'erreur.
- Seuls quelques individus sont observés et on doit extrapoler les conclusions de l'étude à toute une population (contexte des sondages).
- etc...

Il y a donc intervention du **hasard** et des **probabilités**. L'objectif essentiel de la statistique est de maîtriser au mieux cette incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations.

Nous ne nous intéresserons pas à la collecte des données, qui est une tâche importante et difficile, mais qui ne relève pas des mathématiques. Si on omet la collecte des données, les méthodes statistiques se répartissent en deux classes :

- La **statistique descriptive**, **statistique exploratoire** ou **analyse des données**, a pour but de **résumer l'information** contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des **représentations de données** sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

- La **statistique inférentielle** va au delà de la simple description des données. Elle a pour but de **faire des prévisions** et de **prendre des décisions** au vu des observations. En général, il faut pour cela proposer des **modèles probabilistes** du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental.

Pour le grand public, les statistiques désignent les résumés de données fournis par la statistique descriptive. Par exemple, on parle des "statistiques du chômage" ou des "statistiques de l'économie américaine". Mais on oublie en général les aspects les plus importants liés aux prévisions et à l'aide à la décision apportés par la statistique inférentielle.

Les différents aspects de la statistique peuvent être illustrés sur l'exemple de la pandémie de covid 19.

- **Description** : le suivi de l'épidémie se fait à l'aide des courbes de contaminations, admissions à l'hôpital et en soins critiques, etc.
- **Prévision** : Prévoir l'évolution de l'épidémie.
  — Prévoir le nombre de contaminations, d'hospitalisations, d'admissions en ré-

animation, de décès à court et moyen terme.
— Prévoir quand on passera au-dessus ou au-dessous d'un certain seuil de contaminations par jour.
— Prévoir quand commencer ou finir un confinement.
— On utilise pour cela des modèles épidémiologiques (SIR,...).

- **Aide à la décision** :
  — Déterminer les facteurs influents sur le développement des formes graves (âge, sexe, vaccination, comorbidité, groupe sanguin,...).
  — Evaluer l'efficacité des traitements et des vaccins.

L'informatique et la statistique sont deux éléments du **traitement de l'information** : l'informatique acquiert et traite l'information tandis que la statistique l'analyse. Les deux disciplines sont donc étroitement liées. En particulier, l'augmentation considérable de la puissance des ordinateurs et la facilité de transmission des données par internet ont rendu possible l'analyse de très grandes masses de données (**big data**). La **science des données** ou **data science** désigne l'ensemble des méthodes permettant d'extraire des informations utiles de ces grandes masses de données et de les traiter. Cela nécessite des compétences en informatique (bases de données, calcul parallèle, systèmes distribués, visualisation,...), en optimisation et en statistique (fouille de données/**data mining**, apprentissage). L'apprentissage statistique (**machine learning**) et l'apprentissage profond (**deep learning**) sont des techniques de science des données combinant statistique et informatique, particulièrement appropriées en **intelligence artificielle**. Enfin, l'**informatique décisionnelle** ou **business intelligence** regroupe les outils d'**aide à la décision** devenus essentiels dans la gestion des entreprises. Ces outils nécessitent un recours important aux méthodes statistiques.

Plus généralement, tout ingénieur est amené à prendre des décisions au vu de certaines informations, dans des contextes où de nombreuses incertitudes demeurent. Il importe donc qu'un ingénieur soit formé aux techniques de gestion du risque et de traitement de données expérimentales.

1.2 La démarche statistique

La statistique et les probabilités sont les deux aspects complémentaires de l'étude des phénomènes aléatoires. Ils sont cependant de natures bien différentes.

Les **probabilités** peuvent être envisagées comme une branche des mathématiques pures, basée sur la théorie de la mesure, abstraite et complètement déconnectée de la réalité.

Les **probabilités appliquées** proposent des **modèles probabilistes** du déroulement de phénomènes aléatoires concrets. On peut alors, **préalablement à toute expérience**, faire des prévisions sur ce qui va se produire.

Par exemple, il est usuel de modéliser la durée de bon fonctionnement ou durée de vie d'un système, mettons une ampoule électrique, par une variable aléatoire X de loi exponentielle de paramètre λ. Ayant adopté ce modèle probabiliste, on peut effectuer tous les calculs que l'on veut. Par exemple :

- La probabilité que l'ampoule ne soit pas encore tombée en panne à la date t est $P(X > t) = e^{-\lambda t}$.

- La durée de vie moyenne est $E[X] = 1/\lambda$.
- Si n ampoules identiques sont mises en fonctionnement en même temps, et qu'elles fonctionnent indépendamment les unes des autres, le nombre N_t d'ampoules qui tomberont en panne avant un instant t est une variable aléatoire de loi binomiale $\mathcal{B}(n, P(X \leq t)) = \mathcal{B}\left(n, 1 - e^{-\lambda t}\right)$. Donc on s'attend à ce que, en moyenne, $E[N_t] = n\left(1 - e^{-\lambda t}\right)$ ampoules tombent en panne entre 0 et t.

Dans la pratique, l'utilisateur de ces ampoules est très intéressé par ces résultats. Il souhaite évidemment avoir une évaluation de leur durée de vie, de la probabilité qu'elles fonctionnent correctement pendant plus d'un mois, un an, etc... Mais si l'on veut utiliser les résultats théoriques énoncés plus haut, il faut d'une part pouvoir s'assurer qu'on a choisi un bon modèle, c'est-à-dire que la durée de vie de ces ampoules est bien une variable aléatoire de loi exponentielle, et, d'autre part, pouvoir calculer d'une manière ou d'une autre la valeur du paramètre λ. C'est la statistique qui va permettre de résoudre ces problèmes. Pour cela, il faut faire une expérimentation, recueillir des données et les analyser.

On met donc en place ce qu'on appelle un **essai** ou une **expérience**. On fait fonctionner en parallèle et indépendamment les unes des autres $n = 10$ ampoules identiques, dans les mêmes conditions expérimentales, et on relève leurs durées de vie. Admettons que l'on obtienne les durées de vie suivantes, exprimées en heures :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

Notons $x_1, \ldots, x_n$ ces observations. Il est bien évident que la durée de vie des ampoules n'est pas prévisible avec certitude à l'avance. On va donc considérer que $x_1, \ldots, x_n$ sont les **réalisations** de variables aléatoires $X_1, \ldots, X_n$. Cela signifie qu'avant l'expérience, la durée de vie de la $i^{\text{ème}}$ ampoule est inconnue et que l'on traduit cette incertitude en modélisant cette durée par une variable aléatoire X_i. Mais après l'expérience, la durée de vie a été observée. Il n'y a donc plus d'incertitude, cette durée est égale au réel x_i. On dit que x_i est la réalisation de X_i sur l'essai effectué.

Puisque les ampoules sont identiques, il est naturel de supposer que les X_i sont de même loi. Cela signifie qu'on observe plusieurs fois le même phénomène aléatoire. Mais le hasard fait que les réalisations de ces variables aléatoires de même loi sont différentes, d'où la variabilité dans les données. Puisque les ampoules ont fonctionné indépendamment les unes des autres, on pourra également supposer que les X_i sont des variables aléatoires indépendantes. On peut alors se poser les questions suivantes :

1. Au vu de ces observations, est-il raisonnable de supposer que la durée de vie d'une ampoule est une variable aléatoire de loi exponentielle ? Si non, quelle autre loi serait plus appropriée ? C'est un problème de **choix de modèle** ou de **test d'adéquation**.

2. Si le modèle de loi exponentielle a été retenu, comment proposer une valeur (ou un ensemble de valeurs) vraisemblable pour le paramètre λ ? C'est un problème d'**estimation paramétrique**.

3. Dans ce cas, peut-on garantir que λ est inférieur à une valeur fixée λ_0 ? Cela garantira alors que $E[X] = 1/\lambda \geq 1/\lambda_0$, autrement dit que les ampoules seront suffisamment fiables. C'est un problème de **test d'hypothèses paramétriques**.

4. Sur un parc de 100 ampoules, à combien de pannes peut-on s'attendre en moins de 50 h ? C'est un problème de **prévision**.

Le premier problème central est celui de l'**estimation** : comment proposer, au vu des observations, une approximation des grandeurs inconnues du problème qui soit la plus proche possible de la réalité ? La première question peut se traiter en estimant la fonction de répartition ou la densité de la loi de probabilité sous-jacente, la seconde revient à estimer un paramètre de cette loi, la quatrième à estimer un nombre moyen de pannes sur une période donnée.

Le second problème central est celui des **tests d'hypothèses** : il s'agit de se prononcer sur la validité d'une hypothèse liée au problème : la loi est-elle exponentielle ? λ est-il inférieur à λ_0 ? un objectif de fiabilité est-il atteint ? En répondant oui ou non à ces questions, il est possible que l'on se trompe. Donc, à toute réponse statistique, il faudra associer le **degré de confiance** que l'on peut accorder à cette réponse. C'est une caractéristique importante de la statistique par rapport aux mathématiques classiques, pour lesquelles un résultat est soit juste, soit faux.

Pour résumer, la démarche probabiliste suppose que la nature du hasard est connue. Cela signifie que l'on adopte un modèle probabiliste particulier (ici la loi exponentielle), qui permettra d'effectuer des prévisions sur les observations futures. Dans la pratique, la nature du hasard est inconnue. La statistique va, au vu des observations, formuler des hypothèses sur la nature du phénomène aléatoire étudié. Maîtriser au mieux cette incertitude permettra de traiter les données disponibles. Probabilités et statistiques agissent donc en aller-retour dans le traitement mathématique des phénomènes aléatoires.

L'exemple des ampoules est une illustration du cas le plus fréquent où les données se présentent sous la forme d'une suite de nombres. C'est ce cas que nous traiterons dans ce cours, mais il faut savoir que les données peuvent être beaucoup plus complexes : des fonctions, des images, etc... Les principes et méthodes généraux que nous traiterons dans ce cours seront adaptables à tous les types de données.

1.3 Objectifs et plan du cours

Ce cours a pour but de présenter les principes de base d'une analyse statistique de données (description, estimation, tests), ainsi que les méthodes statistiques les plus usuelles. Ces méthodes seront toujours illustrées par des problèmes concrets, issus de divers domaines d'application. Il ne s'agit pas de donner un catalogue de recettes. Les méthodes statistiques seront la plupart du temps justifiées mathématiquement, ce qui permettra de comprendre d'où elles viennent et d'éviter un certain nombre d'erreurs d'interprétation des résultats, fréquentes dans la pratique. Néanmoins, le cours privilégie l'application à la théorie.

Mis à part dans le chapitre consacré à la régression linéaire, les données étudiées dans le cours seront à une dimension et indépendantes. Dans la pratique, les données

à étudier seront en général à plusieurs dimensions, avec des dépendances, et évolutives dans le temps et/ou l'espace. Les principes généraux de statistique descriptive, estimation et tests vus ici dans le cas le plus simple se généralisent dans ces cas plus complexes.

Les approfondissements théoriques et la généralisation à des données plus complexes seront étudiés dans les cours de $2^{\text{ème}}$ année de Statistique Inférentielle Avancée, Modélisation Statistique et Analyse des Données, et Statistical Learning and Applications.

Les méthodes présentées seront mises en œuvre à l'aide du logiciel R. La plupart du temps, on associera à chaque méthode la syntaxe et les sorties (tableaux, graphiques) correspondantes de R. Le cours donnera donc lieu à une introduction succincte à R. L'objectif n'est pas de présenter la programmation en R et la puissance de l'outil, mais simplement d'illustrer les notions de statistique du cours.

Le chapitre 2 présente les techniques de base en statistique descriptive, représentations graphiques et indicateurs statistiques. Le chapitre 3 est consacré aux problèmes d'estimation paramétrique ponctuelle, le chapitre 4 aux intervalles de confiance et le chapitre 5 aux tests d'hypothèses. Le dernier chapitre est consacré à une des méthodes statistiques les plus utilisées, la régression linéaire. Enfin, des annexes donnent quelques rappels de probabilités utiles en statistique, des tables des lois de probabilité usuelles et une courte introduction à R.

Chapitre 2

Statistique descriptive

La **statistique descriptive** a pour but de **résumer l'information** contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible. Les deux principaux outils de la statistique descriptive sont les **représentations graphiques** et les **indicateurs statistiques**.

2.1 Terminologie

Les données dont nous disposons sont des mesures faites sur des **individus** (ou unités statistiques) issus d'une **population**. On s'intéresse à une ou plusieurs particularités des individus appelées **variables** ou **caractères**. L'ensemble des individus constitue l'**échantillon** étudié.

Exemple : si l'échantillon est un groupe de TD à l'Ensimag,

- un individu est un étudiant,
- la population peut être l'ensemble des étudiants de l'Ensimag, des élèves ingénieur de France, des habitants de Grenoble, etc...
- les variables étudiées peuvent être la taille, la filière choisie, la moyenne d'année, la couleur des yeux, la catégorie socio-professionnelle des parents,...

Si l'échantillon est constitué de tous les individus de la population, on dit que l'on fait un **recensement**. Il est extrêmement rare que l'on se trouve dans cette situation, essentiellement pour des raisons de coût. Quand l'échantillon n'est qu'une partie de la population, on parle de **sondage**. Le principe des sondages est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon. Pour que cela ait un sens, il faut que l'échantillon soit représentatif de la population. Il existe des méthodes pour y parvenir, dont nous ne parlerons pas ici.

Remarque : le mot "variable" désigne à la fois la grandeur que l'on veut étudier (variable statistique) et l'objet mathématique qui la représente (variable aléatoire).

Une variable statistique peut être **discrète** ou **continue**, **qualitative** ou **quantitative**. Les méthodes de représentation des données diffèrent suivant la nature des variables étudiées.

Dans ce chapitre, on ne s'intéresse qu'au cas où on ne mesure qu'une seule variable sur les individus, comme dans l'exemple des ampoules. On dit alors que l'on fait de la **statistique unidimensionnelle**. Dans ce cas, les données sont sous la forme de la série des valeurs prises par la variable pour les n individus, notées $x_1, ..., x_n$. On supposera que ces données sont les réalisations de n variables aléatoires $X_1, ..., X_n$ indépendantes et de même loi. On notera X une variable aléatoire de cette loi. Le terme d'**échantillon** désignera à la fois les séries $x_1, ..., x_n$ et $X_1, ..., X_n$.

Quand on mesure plusieurs variables sur les mêmes individus, on dit que l'on fait de la **statistique multidimensionnelle**. Des données de ce type seront traitées dans le chapitre consacré à la régression linéaire. Les cours de $2^{\text{ème}}$ année de Modélisation Statistique et Analyse des Données, et Statistical Learning and Applications leur sont largement consacrés.

L'objectif premier de la statistique descriptive est un objectif de représentation des données, et pas d'estimation. On peut cependant utiliser les outils de statistique descriptive dans un but d'estimation. Notamment, on s'intéressera au choix d'un modèle probabiliste pertinent, ce qui reviendra à estimer la fonction de répartition F ou la densité f de la variable aléatoire X sous-jacente, quand celle-ci est quantitative.

2.2 Représentations graphiques

2.2.1 Variables discrètes

Une variable discrète est une variable à valeurs dans un ensemble fini ou dénombrable. Mais l'ensemble des valeurs prises par cette variable dans un échantillon de taille n est forcément fini. Les variables qui s'expriment par des nombres réels sont appelées **variables quantitatives** ou **numériques** (ex : longueur, durée, coût,...). Les variables qui s'expriment par l'appartenance à une catégorie sont appelées **variables qualitatives** ou **catégorielles** (ex : couleur, catégorie socio-professionnelle, ...).

2.2.1.1. Variables qualitatives

Si la variable est qualitative, on appelle **modalités** les valeurs possibles de cette variable. L'ensemble des modalités est noté $E = \{e_1, \ldots, e_k\}$.

Par exemple, si la variable est la couleur des yeux d'un individu, l'ensemble des modalités est E = {bleu, vert, brun, pers, noir}. Si on interroge n = 200 personnes, les données brutes se présenteront sous la forme d'une suite du type : brun, vert, vert, bleu, ..., noir, vert. Cette suite n'est pas lisible. La meilleure manière de représenter ces données est d'utiliser les fréquences absolues et relatives.

Définition 1 *On appelle **fréquence absolue** de la modalité e_j le nombre total n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j : $n_j = \sum_{i=1}^{n} \mathbb{1}_{\{e_j\}}(x_i)$.*

*On appelle **fréquence relative** de la modalité e_j le pourcentage n_j/n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j.*

Dans l'exemple, on obtient un tableau du type du tableau 2.1.

couleur des yeux	bleu	vert	brun	pers	noir
fréquences absolues	66	34	80	15	5
fréquences relatives	33%	17%	40%	7.5%	2.5%

TABLE 2.1 – couleur des yeux d'un échantillon de 200 personnes

De même, dans le cas des résultats d'élection en France, les individus sont les n = 49 millions d'électeurs et la variable est la personne ou la liste pour laquelle l'individu a voté. La suite des 49 millions de votes n'a aucun intérêt. Le résultat est exprimé directement sous forme du tableau des fréquences relatives. Par exemple, le tableau 2.2 donne le résultat des élections européennes de 2024.

Listes	LFI	PC	EELV	PSPP	ECOC	ENS	AN	DVD	LR	FREX	RN	REC	Autres
% Voix	9.9	2.4	5.5	13.8	1.3	14.6	2.0	2.4	7.2	1.0	31.4	5.5	3.0

TABLE 2.2 – résultats des élections européennes de 2024

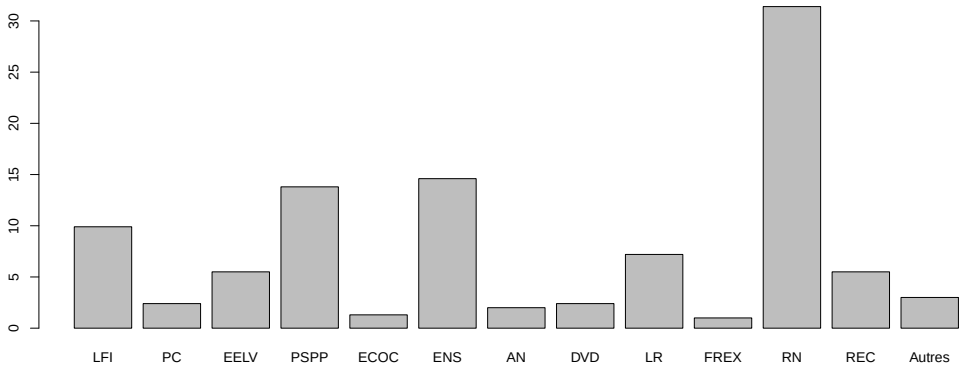


FIGURE 2.1 – élections européennes, diagramme en colonnes

Les représentations graphiques correspondantes sont de deux types :

- **diagrammes en colonnes** ou **en bâtons** : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de cette modalité
- **diagrammes sectoriels** ou **camemberts** : à chaque modalité correspond un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative de cette modalité

Les commandes `R` pour les diagrammes en colonnes et sectoriels sont `barplot(x)` et `pie(x)`. Dans l'exemple des élections, les figures 2.1 et 2.2 sont obtenues à l'aide des commandes :

```
>  x<-c(2.5, 6.3, 3.3, 6.2, 13.5, 22.4, 2.5, 8.5, 3.5, 23.3,
  2.2, 5.8)
>  partis<-c("PC","LFI","Gen","EEES","EE","LREM","UDI","UDC",
  "DLF","RN","PA","Autres")
>  barplot(x,names=partis)
>  pie(x,labels=partis)
```

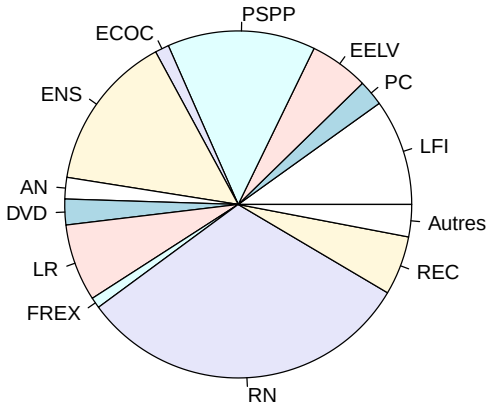


FIGURE 2.2 – élections européennes, diagramme sectoriel

2.2.1.2. Variables quantitatives

Quand la variable est quantitative, on utilise les mêmes représentations à l'aide des fréquences absolues et relatives. La différence fondamentale entre les représentations pour des variables qualitatives et quantitatives tient au fait qu'il existe un ordre naturel sur les modalités (qui sont des nombres réels) pour les variables quantitatives, alors qu'aucun ordre n'est prédéfini pour les variables qualitatives. C'est pourquoi les diagrammes en bâtons sont toujours utilisés, mais pas les diagrammes sectoriels.

Par exemple, on a effectué une enquête auprès de 1000 couples en leur demandant notamment leur nombre d'enfants. Le tableau 2.3 donne les fréquences et la figure 2.3 donne le diagramme en bâtons, obtenu à l'aide de la commande :

```
>  barplot(c(235,183,285,139,88,67,3),names=c(0,1,2,3,4,5,6))
```

2.2.2 Variables continues

Une variable continue est à valeurs dans un ensemble non dénombrable comme $\mathbb{R}$ ou $[a, b]$. Dans ce cas, les représentations du type diagramme en bâtons sont sans

Nombre d'enfants	0	1	2	3	4	5	6	> 6
fréquence absolue	235	183	285	139	88	67	3	0
fréquence relative	23.5%	18.3%	28.5%	13.9%	8.8%	6.7%	0.3%	0

TABLE 2.3 – nombre d'enfants de 1000 couples

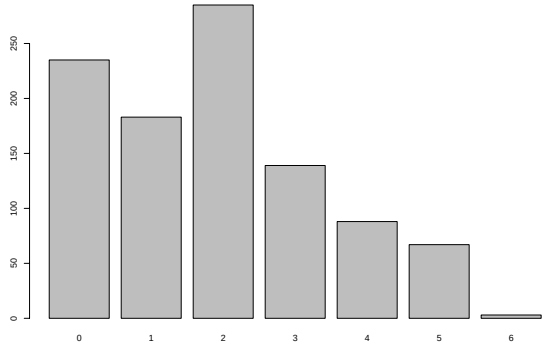


FIGURE 2.3 – nombre d'enfants de 1000 couples, diagramme en bâtons

intérêt, car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1.

On considèrera ici deux types de représentations graphiques :

- l'**histogramme**.
- la **fonction de répartition empirique**, qui permet notamment de construire des **graphes de probabilités**.

Ces deux types de représentations nécessitent d'**ordonner** les données. Si l'échantillon initial est noté $x_1, ..., x_n$, l'échantillon ordonné sera noté $x_1^*, ..., x_n^*$.

Dans l'exemple des ampoules, l'échantillon initial est :

$$91.6 \quad 35.7 \quad 251.3 \quad 24.3 \quad 5.4 \quad 67.3 \quad 170.9 \quad 9.5 \quad 118.4 \quad 57.1$$

et l'échantillon ordonné est :

$$5.4 \quad 9.5 \quad 24.3 \quad 35.7 \quad 57.1 \quad 67.3 \quad 91.6 \quad 118.4 \quad 170.9 \quad 251.3$$

On a donc, par exemple :

$x_1 = 91.6 =$ durée de vie de la première ampoule.
$x_1^* = \min(x_1, \ldots, x_n) = 5.4 =$ plus petite des durées de vie des 10 ampoules.

En R, l'échantillon `x` est créé par la commande :

```
x<-c(91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4,
57.1)
```

La $i^{\text{ème}}$ observation est donnée par `x[i]`.

L'échantillon ordonné est obtenu par la commande `sort(x)`.

```
>  x[1]
[1] 91.6

>  y<-sort(x)

>  y[1]
[1] 5.4
```

2.2.2.1. Histogramme

Le principe de cette représentation est de regrouper les observations "proches" en classes. Pour cela, on se fixe une borne inférieure de l'échantillon $a_0 < x_1^*$ et une borne supérieure $a_k > x_n^*$. On partitionne l'intervalle $]a_0, a_k]$, contenant toutes les observations, en k intervalles $]a_{j-1}, a_j]$ appelés **classes**. La largeur de la classe j est $h_j = a_j - a_{j-1}$.

Si toutes les classes sont de même largeur $h = (a_k - a_0)/k$, on dit que l'on fait un **histogramme à pas fixe**. Si les h_j ne sont pas tous égaux, on dit que l'on fait un **histogramme à pas variable**.

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

La **fréquence** (ou fréquence relative) de la classe j est n_j/n.

Définition 2 *: L'***histogramme** *est la figure constituée des rectangles dont les* **bases** *sont les classes et dont les* **aires** *sont égales aux fréquences de ces classes. Autrement dit, la hauteur du $j^{ème}$ rectangle est n_j/nh_j.*

Notons $\hat{f}$ la fonction en escalier constante sur les classes et qui vaut n_j/nh_j sur la classe $]a_{j-1}, a_j]$. L'aire du $j^{\text{ème}}$ rectangle est la fréquence de la classe j : $n_j/n = \int_{a_{j-1}}^{a_j} \hat{f}(x)dx$. Or cette fréquence est le pourcentage d'observations appartenant à la classe j, donc c'est une estimation naturelle de la probabilité qu'une observation appartienne à cette classe. Cette probabilité est $P(a_{j-1} < X \leq a_j) = F(a_j) - F(a_{j-1}) = \int_{a_{j-1}}^{a_j} f(x)dx$, c'est-à-dire l'aire délimitée par l'axe des abcisses et la densité f sur la classe j.

On en déduit que **l'histogramme fournit une estimation de la densité des observations**. L'estimation de la densité en un point x, $f(x)$, est égale à la hauteur $\hat{f}(x)$ du rectangle correspondant à la classe à laquelle x appartient.

L'allure de l'histogramme permettra donc de proposer des modèles probabilistes vraisemblables pour la loi de X en comparant la forme de $\hat{f}$ à celle de densités de lois de probabilité usuelles.

On voit que l'estimation proposée par l'histogramme dépend de plusieurs paramètres : les bornes inférieure et supérieure a_0 et a_k, le nombre et la largeur des classes. Cela fait que plusieurs histogrammes peuvent être dessinés à partir des mêmes données et avoir des allures assez différentes, pouvant donner lieu à des interprétations trompeuses. En pratique, il est conseillé de suivre les règles suivantes :

- Il est recommandé d'avoir entre 5 et 20 classes. La *règle de Sturges* préconise de choisir un nombre de classes égal à $k \approx 1 + \log_2 n = 1 + \ln n / \ln 2$. Cela donne par exemple $k = 5$ pour $n \leq 22$, $k = 6$ pour $23 \leq n \leq 45$, etc...
- Le choix des bornes a_0 et a_k doit être fait de façon à respecter une certaine homogénéité des largeurs de classes. Un choix fréquent est $a_0 = x_1^* - 0.025(x_n^* - x_1^*)$ et $a_k = x_n^* + 0.025(x_n^* - x_1^*)$.

Le choix le plus fréquent est celui de l'histogramme à pas fixe, où les classes sont de même largeur $h = (a_k - a_0)/k$. Dans ce cas, la hauteur d'un rectangle est proportionnelle à l'effectif de sa classe.

Prenons l'exemple des ampoules. On a $n = 10$ données, donc la règle de Sturges dit de choisir $k = 5$ classes. Comme $x_1^* = 5.4$ et $x_n^* = 251.3$, la règle énoncée donne $a_0 = -0.747$ et $a_5 = 257.4$, qu'on peut arrondir à $a_0 = 0$ et $a_5 = 260$. Si on veut un histogramme à 5 classes de même largeur, cette largeur sera donc $h = 260/5 = 52$. On obtient alors le tableau 2.4 et l'histogramme correspondant est donné par la figure 2.4. La commande `R` permettant de construire cette figure est :

```
> hist(x, prob=T, breaks=seq(0,260,52))
```

classes $]a_{j-1}, a_j]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_j	4	3	1	1	1
fréquences n_j/n	40%	30%	10%	10%	10%
hauteurs n_j/nh	0.0077	0.0058	0.0019	0.0019	0.0019

TABLE 2.4 – Ampoules, répartition en classes de même largeur

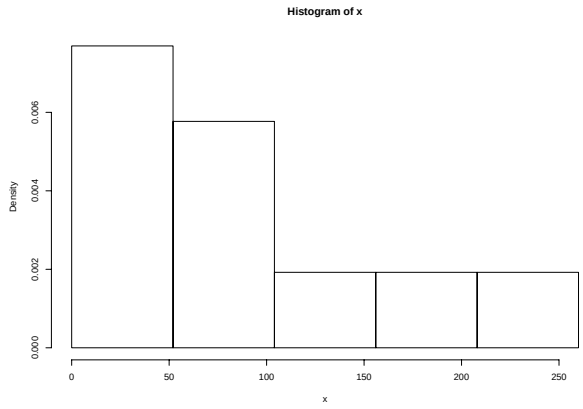


FIGURE 2.4 – Ampoules, histogramme à classes de même largeur

Dans cette commande, `prob=T` signifie que l'on représente bien en ordonnées les hauteurs (avec `prob=F`, on aurait représenté les effectifs) et `breaks=seq(0,260,52)` signifie que les bornes des classes sont obtenues en partageant l'intervalle]0,260] en intervalles de même largeur 52.

Le *mode* de l'histogramme est le milieu de la classe correspondant au rectangle le

plus haut, ici 26. C'est une estimation du point où la densité est maximum (que l'on appelle également le mode de la densité).

L'histogramme fournit bien une visualisation de la répartition des données. Ici, le phénomène marquant est la concentration des observations sur les petites valeurs et le fait que, plus la durée de vie grandit, moins il y a d'observations. Autrement dit, la densité de la variable aléatoire représentant la durée de vie d'une ampoule est une fonction décroissante.

L'inconvénient d'un histogramme à pas fixe est que certaines classes peuvent être très chargées et d'autres pratiquement vides. Par exemple ici, la classe 1 contient plus d'observations à elle seule que les classes 3, 4 et 5 réunies. Pour connaître la répartition des observations dans les classes chargées, on a envie de scinder celles-ci. De même, on peut regrouper des classes trop peu chargées.

Une façon d'y parvenir est de faire en sorte que toutes les classes aient le même effectif. Dans ce cas, elles ne peuvent pas être de même largeur. Les bornes des classes sont cette fois aléatoires, puisqu'elles sont fonction des observations.

Dans l'exemple des ampoules, on peut faire en sorte d'avoir 2 observations par classe. On détermine par exemple les limites des classes en prenant le milieu de deux observations ordonnées successives. On obtient alors le tableau 2.5 et l'histogramme 2.5.

classes $]a_{j-1}, a_j]$	$]0, 17]$	$]17, 46]$	$]46, 79]$	$]79, 145]$	$]145, 260]$
largeurs h_j	17	29	33	66	115
effectifs n_j	2	2	2	2	2
fréquences n_j/n	20%	20%	20%	20%	20%
hauteurs n_j/nh_j	0.0118	0.0069	0.0061	0.0030	0.0017

TABLE 2.5 – Ampoules, répartition en classes de même effectif

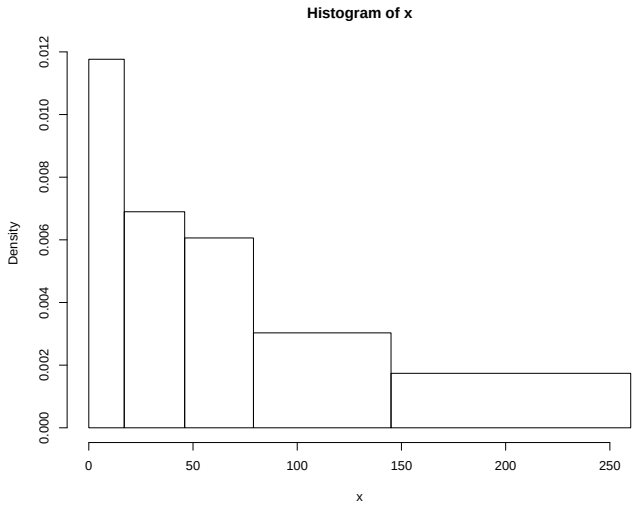


FIGURE 2.5 – Ampoules, histogramme à classes de même effectif

On constate que cet histogramme décrit plus finement la distribution que le pré-

cédent. C'est toujours le cas des histogrammes à classes de même effectif. Mais leur usage est moins répandu que celui des histogrammes à classes de même largeur, car ils sont moins faciles à tracer.

On voit que des histogrammes distincts sur les mêmes données peuvent être sensiblement différents. Donc il faudra se méfier des histogrammes si on veut estimer la densité des observations. On se contentera de dire que l'histogramme donne une allure générale de cette densité.

Par exemple ici, il est clair que la forme des deux histogrammes n'est pas très éloignée de la densité d'une loi exponentielle ($f(x) = \lambda e^{-\lambda x}$). En revanche, ils ne ressemblent pas du tout à la densité d'une loi normale (en forme de cloche). On en conclura qu'il est très peu probable que la durée de vie d'une ampoule soit de loi normale, et qu'il est possible, voire vraisemblable, qu'elle soit de loi exponentielle. Ce jugement est pour l'instant purement visuel. Il faudra l'affiner par des techniques quantitatives plus précises.

Remarque : Si au lieu des effectifs n_j, on considère les effectifs cumulés $m_j = \sum_{l=1}^{j} n_l$, on construit un **histogramme cumulé**, qui fournit une estimation de la fonction de répartition de la variable étudiée.

2.2.2.2. Fonction de répartition empirique

Définition 3 *: La* **fonction de répartition empirique** *(FdRE) F_n associée à un échantillon $x_1, \ldots, x_n$ est la fonction définie par :*

$\forall x \in \mathbb{R}$, $F_n(x) =$ *pourcentage d'observations inférieures à* x

$$\forall x \in \mathbb{R}\ ,\ F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}} = \begin{cases} 0 & \textit{si } x < x_1^* \\ \dfrac{i}{n} & \textit{si } x_i^* \leq x < x_{i+1}^* \\ 1 & \textit{si } x \geq x_n^* \end{cases}$$

La fonction de répartition de X, $F(x) = P(X \leq x)$, donne la probabilité qu'une observation soit inférieure à x, tandis que $F_n(x)$ est le pourcentage d'observations inférieures à x. Par conséquent, $F_n(x)$ est une estimation de $F(x)$. On peut montrer que cette estimation est d'excellente qualité, en un sens que l'on verra plus tard.

$F_n(x)$ est une fonction en escalier qui fait des sauts de hauteur $1/n$ en chaque point de l'échantillon. Par exemple, la figure 2.6 représente la fonction de répartition empirique de l'échantillon des durées de vie d'ampoules. La commande `R` permettant de tracer cette fonction sur cet exemple est `plot(ecdf(x))`.

2.2.2.3. Les graphes de probabilités

La fonction de répartition empirique est très utile en statistique. Intéressons-nous ici uniquement à son utilisation pour déterminer un modèle probabiliste acceptable pour les observations.

A priori, la première idée est de tracer le graphe de la fonction de répartition empirique et de déterminer si ce graphe ressemble à celui de la fonction de répartition d'une

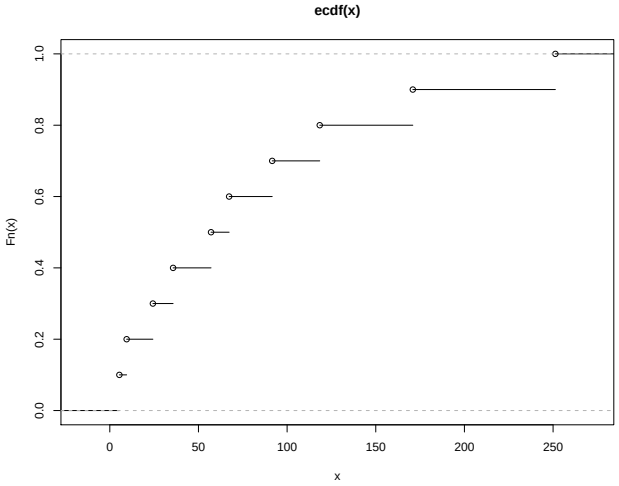


FIGURE 2.6 – Ampoules, fonction de répartition empirique

loi connue. En fait, il est très difficile de procéder ainsi car les fonctions de répartition de toutes les lois de probabilité se ressemblent : à vue d'oeil, il n'y a pas de grande différence entre les fonctions de répartition des lois normale et exponentielle, alors que leurs densités ne se ressemblent pas du tout.

Une seconde idée est alors d'appliquer une transformation à la fonction de répartition empirique qui permette de reconnaître visuellement une caractéristique d'une loi de probabilité. Un **graphe de probabilités** (en anglais **probability plot** ou **Q-Q plot**) est un nuage de points tracé à partir de la fonction de répartition empirique, tel que les points doivent être approximativement alignés si les observations proviennent d'une loi de probabilité bien précise.

Si on souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F, le principe est de chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Ainsi, si la vraie fonction de répartition des observations est F, $h[F_n(x)]$ devrait être "proche" de $\alpha(\theta)g(x) + \beta(\theta)$, pour tout x. Pour $x = x_i^*$, $h[F_n(x_i^*)] = h(i/n)$. Donc, si la vraie fonction de répartition est F, les points $\Big(g(x_i^*), h(i/n)\Big)$ seront approximativement alignés. La pente et l'ordonnée à l'origine de cette droite fourniront des estimations de $\alpha(\theta)$ et $\beta(\theta)$, donc la plupart du temps de θ.

Définition 4 *: Soit F la fonction de répartition d'une loi de probabilité, dépendant d'un paramètre inconnu θ. S'il existe des fonctions h, g, α et β telles que,*

$$\forall x \in \mathbb{R},\ h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$$

alors le nuage des points

$$\Big(g(x_i^*), h(i/n)\Big),\ i \in \{1, \ldots, n\}$$

est appelé **graphe de probabilités** *pour la loi de fonction de répartition F. Si les points du nuage sont approximativement alignés, on admettra que F est une fonction de répartition plausible pour les observations.*

Exemple 1 : Graphe de probabilités pour la loi exponentielle

Si X est de loi $exp(\lambda)$, $F(x) = 1 - e^{-\lambda x}$, d'où $\ln(1 - F(x)) = -\lambda x$. C'est de la forme voulue avec $\theta = \lambda$, $h(u) = \ln(1-u)$, $\alpha(\lambda) = -\lambda$, $g(x) = x$ et $\beta(\lambda) = 0$.

Par conséquent, le graphe de probabilités pour la loi exponentielle est le nuage des points $\Big(x_i^*, \ln(1 - i/n)\Big),\ i \in \{1, \ldots, n-1\}$ (le point correspondant à $i = n$ doit être enlevé car $\ln(1 - n/n) = -\infty$).

Si ces points sont approximativement alignés sur une droite de pente négative et passant par l'origine, on pourra considérer que la loi exponentielle est un modèle probabiliste vraisemblable pour ces observations. La pente de la droite fournit alors une estimation graphique de λ. Inversement, si ce n'est pas le cas, il est probable que les observations ne soient pas issues d'une loi exponentielle.

En `R`, le vecteur des entiers de 1 à 9 est obtenu par la commande :

```
>  seq(1:9)
[1] 1 2 3 4 5 6 7 8 9
```

Le vecteur des $\ln(1 - i/n)$ est obtenu par :

```
>  log(1-seq(1:9)/10)
[1] -0.1053605 -0.2231436 -0.3566749 -0.5108256 -0.6931472
[6] -0.9162907 -1.2039728 -1.6094379 -2.3035851
```

Sur l'exemple des ampoules, le graphe de probabilités pour la loi exponentielle, donné par la figure 2.7, est obtenu par :

```
>  plot(sort(x)[1:9],log(1-seq(1:9)/10),ylim=c(-2.5,0.1))
>  abline(v=0)
>  abline(h=0)
```

`sort(x)[1:9]` signifie que l'on ne prend que les 9 premières composantes du vecteur des observations triées x_i^*. Les commandes `abline` ont rajouté sur la figure les axes des abscisses et des ordonnées, ce qui permet de juger si les points peuvent être considérés comme approximativement alignés sur une droite de pente négative et passant par l'origine. Apparemment, c'est bien le cas, donc on peut considérer qu'il est vraisemblable que la durée de vie d'une ampoule soit une variable aléatoire de loi exponentielle. Cette conclusion est cohérente avec celle des histogrammes.

La droite en question a pour équation $y = -\lambda x$. Sa pente fournit donc une estimation du paramètre λ. Pour déterminer cette pente, la méthode la plus usuelle est la méthode des moindres carrés, qui sera étudiée dans le chapitre consacré à la régression linéaire. On obtient ici une estimation de l'ordre de 0.013.

Exemple 2 : Graphe de probabilités pour la loi normale

Si X est de loi $\mathcal{N}(m, \sigma^2)$, $U = \dfrac{X - m}{\sigma}$ est de loi $\mathcal{N}(0, 1)$. Alors $F(x) = P(X \leq x) = P\Big(U \leq \dfrac{x-m}{\sigma}\Big) = \phi\Big(\dfrac{x-m}{\sigma}\Big)$, où ϕ est la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Etant

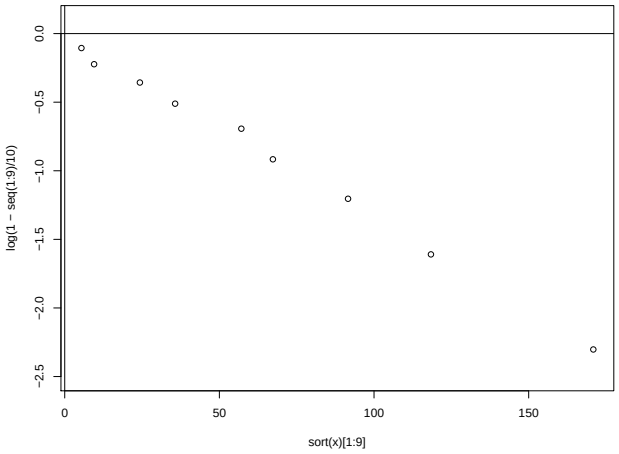


FIGURE 2.7 – Ampoules, graphe de probabilités pour la loi exponentielle

donné que ϕ est strictement croissante, elle est inversible, et on a $\phi^{-1}(F(x)) = \dfrac{x-m}{\sigma} = \dfrac{1}{\sigma}x - \dfrac{m}{\sigma}$. C'est de la forme voulue avec $\theta = (m, \sigma^2)$, $h(u) = \phi^{-1}(u)$, $\alpha(m, \sigma^2) = 1/\sigma$, $g(x) = x$ et $\beta(m, \sigma^2) = -m/\sigma$.

Par conséquent, le graphe de probabilités pour la loi normale est le nuage des points $\left(x_i^*, \phi^{-1}(i/n)\right),\ i \in \{1, \ldots, n-1\}$ (le point correspondant à $i = n$ doit être enlevé car $\phi^{-1}(1) = +\infty$).

Si ces points sont approximativement alignés, on admettra que la loi normale est un modèle plausible pour les observations. Si c'est le cas, la droite en question est appelée **droite de Henry** (du nom d'un ingénieur de l'armée française qui s'est intéressé à la dispersion des tirs d'obus au dix-neuvième siècle). Sa pente et son ordonnée à l'origine fournissent des estimations graphiques de m et σ.

En R, $\phi^{-1}(p)$ est donné par `qnorm(p)`. Donc le vecteur des $\phi^{-1}(i/n)$ est obtenu par la commande :

```
>  qnorm(seq(1:9)/10)
 [1] -1.2815516 -0.8416212 -0.5244005 -0.2533471  0.0000000
 [6]  0.2533471  0.5244005  0.8416212  1.2815516
```

Par ailleurs, la table 2 de la loi normale donne pour $\alpha \in [0, 1]$ la valeur de $u_\alpha = \phi^{-1}(1 - \alpha/2)$. On a donc :

- pour $p < 1/2, \phi^{-1}(p) = -u_{2p}$.
- $\phi^{-1}(1/2) = 0$.
- pour $p > 1/2, \phi^{-1}(p) = u_{2(1-p)}$.

Sur l'exemple des ampoules, le graphe de probabilités pour la loi normale, donné par la figure 2.8, est obtenu par :

```
>  plot(sort(x)[1:9], qnorm(seq(1:9)/10))
```

```
> abline(h=0)
```

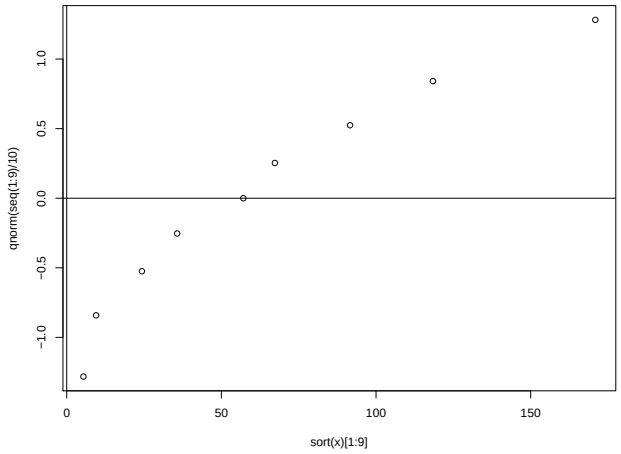


FIGURE 2.8 – Ampoules, graphe de probabilités pour la loi normale

Le graphe de probabilités semble plus proche d'un logarithme que d'une droite. On en conclura donc que la loi normale n'est pas un modèle approprié pour ces données, ce qui est cohérent avec le résultat précédent.

On constate ici le principal défaut de la méthode : comment juger visuellement si des points sont "suffisamment alignés" ? La réponse est soumise à la subjectivité de l'utilisateur. Il est donc nécessaire de compléter cette approche graphique par des techniques objectives : les tests d'adéquation. Néanmoins, les graphes de probabilités sont une première étape indispensable dans une étude statistique, car ils sont faciles à mettre en oeuvre et permettent de détecter facilement des modèles clairement pas adaptés aux données.

Remarque : En `R`, la commande `qqnorm(x)` trace le nuage des points $\left(\phi^{-1}\left(\frac{i-1/2}{n}\right), x_i^*\right)$, qui est quasiment la même chose que le graphe de probabilités pour la loi normale.

2.3 Indicateurs statistiques

Les représentations graphiques présentées dans la section précédente ne permettent qu'une analyse visuelle de la répartition des données. Pour des variables quantitatives, il est intéressant de donner des indicateurs numériques permettant de caractériser au mieux ces données. On donne en général deux indicateurs : un indicateur de localisation et un indicateur de dispersion.

2.3.1 Indicateurs de localisation ou de tendance centrale

Le but est de donner un ordre de grandeur général des observations, un nombre unique qui résume au mieux les données. On pense immédiatement à la moyenne des observations.

2.3.1.1. La moyenne empirique

La **moyenne empirique** de l'échantillon est la moyenne arithmétique des observations, notée $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Son interprétation est évidente. La commande `R` correspondante est `mean(x)`.

Pour l'exemple des ampoules, $\bar{x}_{10}$= 83.15, donc on dira que la durée de vie moyenne d'une ampoule est de 83.15 h. Les représentations graphiques nous ont amenés à admettre que la durée de vie d'une ampoule était une variable aléatoire de loi exponentielle. On rappelle que l'espérance de la loi $\exp(\lambda)$ est $1/\lambda$. D'après la loi des grands nombres, la moyenne empirique converge vers l'espérance de la loi. Il est donc logique de considérer qu'une estimation de λ est $1/\bar{x}_{10}$= 0.012. Cette valeur est cohérente avec la valeur trouvée à l'aide du graphe de probabilités, 0.013. On retrouvera ce principe d'estimation plus tard, sous le nom de méthode des moments.

2.3.1.2. Les valeurs extrêmes

La plus petite valeur $x_1^* = \min x_i$ et la plus grande valeur $x_n^* = \max x_i$ d'un échantillon sont évidemment des indications intéressantes. Leur moyenne $(x_1^* + x_n^*)/2$ est un indicateur de localisation.

En `R`, les commandes correspondantes sont `min(x)` et `max(x)`.

Pour les ampoules, $(x_1^* + x_n^*)/2$ = 128.35.

Problème : Les deux indicateurs que l'on vient de définir sont très sensibles aux valeurs extrêmes. En particulier, il arrive parfois qu'une série statistique présente des **valeurs aberrantes**, c'est à dire des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon. Par exemple, ce serait le cas si une durée de vie était égale à 0.01 ou 10000. En général, la présence d'une valeur aberrante est due à une erreur de saisie ou une erreur dans l'expérience ayant abouti à cette observation. Il faut alors l'éliminer avant d'effectuer l'analyse statistique. Il existe des méthodes de détection des valeurs aberrantes, mais il est souvent difficile de décider si une valeur est aberrante ou pas. Aussi est-il important de disposer d'indicateurs qui ne soient pas trop sensibles aux valeurs aberrantes. Or la moyenne est très sensible : si une des observations est extrêmement grande, elle va tirer la moyenne vers le haut. La médiane empirique est un indicateur de localisation construit pour être insensible aux valeurs aberrantes.

2.3.1.3. La médiane empirique

La **médiane empirique** de l'échantillon, notée $\tilde{x}_n$ ou $\tilde{q}_{n,1/2}$, est un réel qui partage l'échantillon ordonné en deux parties de même effectif. La moitié des observations sont inférieures à $\tilde{x}_n$ et l'autre moitié lui sont supérieures. Il y a donc une chance sur deux pour qu'une observation soit inférieure à la médiane, et évidemment une chance sur deux pour qu'une observation soit supérieure à la médiane.

Si n est impair, la médiane empirique est la valeur située au centre de l'échantillon ordonné : $\tilde{x}_n = x^*_{\frac{n+1}{2}}$.

Si n est pair, n'importe quel nombre compris entre $x^*_{\frac{n}{2}}$ et $x^*_{\frac{n}{2}+1}$ vérifie la définition de la médiane. Par convention, on prend en général le milieu de cet intervalle : $\tilde{x}_n = \left(x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}\right)/2$.

La commande `R` pour la médiane empirique est `median(x)`.

L'expression de la médiane montre bien que c'est un indicateur qui n'est pas sensible aux valeurs aberrantes. Pour l'illustrer, considérons les deux échantillons suivants :

1 3 5 8 10 1 3 5 8 10000

La médiane empirique est $\tilde{x}_5 = x^*_3 = 5$ pour les deux échantillons, alors que la moyenne empirique vaut 5.4 pour le premier échantillon et 2003.4 pour le deuxième. La moyenne est fortement influencée par la valeur aberrante 10000 du deuxième échantillon, alors que la médiane ne l'est pas du tout.

Dans l'exemple des ampoules, $\tilde{x}_{10} = (57.1 + 67.3)/2 = 62.2$. On constate que la médiane est ici nettement inférieure à la moyenne : la durée de vie moyenne est de 83.1 h, et pourtant une ampoule sur deux tombera en panne avant 62.2 h de fonctionnement. Cette propriété est caractéristique des distributions non symétriques dites "à queues lourdes" : un petit nombre d'ampoules auront une durée de vie nettement supérieure à la majeure partie des autres. C'est ce qu'on avait déjà observé sur l'histogramme, et qui peut se remarquer directement sur les données.

Le même phénomène se produit si la variable étudiée est le salaire des français. En 2022, pour un travail à temps plein, le salaire net mensuel moyen était de 2630 €, alors que le salaire net mensuel médian était de 2091 € (source `www.insee.fr`). Un français sur deux travaillant à temps plein touchait donc moins de 2091 € par mois, mais un petit nombre de personnes ayant un fort salaire, cela fait remonter la moyenne. Notons également que le seuil de pauvreté est défini comme 60% du revenu médian, ce qui concerne 9 millions de personnes en France.

On constate donc que la moyenne et la médiane empiriques sont deux résumés de l'échantillon dont la connaissance simultanée peut être riche d'enseignements.

Quand la distribution est symétrique, moyenne et médiane empiriques sont proches (pour une variable aléatoire de loi symétrique, l'espérance et la médiane théoriques sont égales).

2.3.1.4. Caractérisation des indicateurs de localisation

Un indicateur de localisation c est fait pour résumer au mieux à lui seul l'ensemble des observations. L'erreur commise en résumant l'observation x_i par c peut être quantifiée par une distance ou un écart entre ces deux valeurs : $d(x_i, c)$. L'erreur moyenne commise sur tout l'échantillon est $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$. Un bon indicateur de localisation doit minimiser cette erreur globale. L'indicateur c optimal est obtenu en annulant la dérivée de e par rapport à c.

- Si on choisit l'écart quadratique, $e = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$. La valeur de c qui minimise cette erreur est obtenue en annulant la dérivée de e par rapport à c :

$$\frac{\partial e}{\partial c} = -\frac{2}{n}\sum_{i=1}^{n}(x_i - c) = -2(\bar{x}_n - c)$$

  qui vaut 0 pour $c = \bar{x}_n$. La moyenne empirique est donc la valeur qui résume le mieux l'échantillon au sens dit "des moindres carrés".
- Si on choisit $e = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$, on obtient $c = \tilde{x}_n$.
- Si on choisit $e = \frac{1}{n}\sup_{i=1}^{n}|x_i - c|$, on obtient $c = (x_1^* + x_n^*)/2$.

Il est donc justifié d'utiliser ces trois quantités comme indicateurs de localisation.

2.3.2 Indicateurs de dispersion ou de variabilité

Pour exprimer les caractéristiques d'un échantillon, il est nécessaire de compléter les indicateurs de localisation par des indicateurs de dispersion, qui mesureront la variabilité des données.

Par exemple, le tableau 2.6 donne les températures mensuelles moyennes, en degrés Celsius, à New-York et à San Francisco, calculées sur une période de 30 ans.

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

TABLE 2.6 – températures mensuelles moyennes à New-York et à San Francisco

La température annuelle moyenne est de $12.5°$ à New-York et de $13.7°$ à San Francisco. En se basant uniquement sur ces moyennes, on pourrait croire que les climats de ces deux villes sont similaires. Or il est clair que la différence de température entre l'hiver et l'été est beaucoup plus forte à New-York qu'à San Francisco. Pour le déceler, il suffit de calculer un indicateur qui exprime la variabilité des observations.

Or, d'après la section précédente, l'erreur moyenne commise en résumant l'échantillon par un indicateur de localisation c est $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$. e exprime bien la variabilité de l'échantillon autour de c. On pourra donc construire des indicateurs de dispersion à partir de e en considérant différentes distances.

2.3.2.1. Variance et écart-type empiriques

Si on choisit la distance euclidienne, on a vu que $c = \bar{x}_n$. L'indicateur de dispersion correspondant est donc $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. Il est appelé **variance empirique** de l'échantillon, et mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

Il est facile de vérifier que $s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$.

L'**écart-type empirique** de l'échantillon est la racine carrée de la variance empirique : $s_n = \sqrt{s_n^2}$. Il s'exprime dans la même unité que les données, ce qui rend son interprétation plus facile que celle de la variance. Ainsi, l'écart-type des températures annuelles est de $8.8°$ à New-York et de $3°$ à San Francisco, ce qui exprime bien la différence de variabilité des températures entre les deux villes.

Cependant, la variabilité doit toujours se comparer à la valeur moyenne. En effet, une variabilité de $10°$ n'a pas le même sens si la température moyenne de référence est $12°$ ou $10000°$. Des données présentent une forte variabilité si l'écart-type est grand par rapport à la moyenne.

Aussi on définit le **coefficient de variation empirique** de l'échantillon par

$$cv_n = \frac{s_n}{\bar{x}_n}$$

L'intérêt de cet indicateur est qu'il est sans dimension. Une pratique empirique courante est de considérer que l'échantillon possède une variabilité significative si $cv_n > 0.15$. Si $cv_n \leq 0.15$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de tout l'échantillon.

Dans nos exemples, on obtient :

	$\bar{x}_n$	s_n^2	s_n	cv_n
ampoules	83.15	5540.2	74.4	0.89
t° New-York	12.5	77.7	8.8	0.70
t° San Francisco	13.7	8.9	3.0	0.22

On remarque donc une très forte variabilité des deux premiers échantillons et une variabilité assez faible du troisième.

En `R`, la commande `var(x)` donne $s'^2_n = \frac{n}{n-1} s_n^2$ au lieu de s_n^2. C'est aussi ce que l'on a sur les calculatrices dotées de fonctionnalités statistiques. On en verra l'explication au chapitre suivant. $s'_n = \sqrt{s'^2_n}$ est donné en `R` par `sd(x)` (standard deviation). Il n'y a pas de commande prédéfinie pour le coefficient de variation empirique.

Remarque 1 : $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$ évoque $Var[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - \left[E[X]\right]^2$. Les similitudes dans les noms et les formules suggèrent que la variance empirique est très liée à la variance de la loi de probabilité de la variable aléatoire sous-jacente. On reviendra sur ce point au chapitre suivant.

Remarque 2 : En finance, la variabilité d'une série de données est appelée **volatilité**. L'étude de la volatilité est fondamentale dans les analyses de risque financier.

2.3.2.2. Les quantiles empiriques

Les **quantiles empiriques** sont des valeurs qui partagent l'échantillon ordonné en un certain nombre de parties de même effectif.

- s'il y a 2 parties, on retrouve la médiane empirique $\tilde{x}_n$.
- s'il y a 4 parties, on parle de **quartiles**, notés $\tilde{q}_{n,1/4}$, $\tilde{q}_{n,1/2}$ et $\tilde{q}_{n,3/4}$. On a $\tilde{q}_{n,1/2}$= $\tilde{x}_n$.
- s'il y a 10 parties, on parle de **déciles**, notés $\tilde{q}_{n,1/10}, \ldots, \tilde{q}_{n,9/10}$.
- s'il y a 100 parties, on parle de **centiles**, notés $\tilde{q}_{n,1/100}, \ldots, \tilde{q}_{n,99/100}$.
- etc...

Plus généralement, les **quantiles empiriques** de l'échantillon $x_1, \ldots, x_n$ sont définis par :

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

où $\lfloor x \rfloor$ désigne la partie entière de x.

Pour $p = 1/2$, on retrouve bien l'expression de la médiane empirique $\tilde{x}_n$.

Dans l'exemple des ampoules, on n'a que 10 données, donc seuls les quartiles ont un sens. On connaît déjà la médiane empirique $\tilde{q}_{n,1/2} = \tilde{x}_n = 62.2$. On obtient $\tilde{q}_{n,1/4} = x^*_3 = 24.3$ et $\tilde{q}_{n,3/4} = x^*_8 = 118.4$.

Les quantiles empiriques sont très utilisés pour décrire des phénomènes concernant les extrémités des échantillons :

— En finance, la **value at risk** (VaR) est la plus utilisée des mesures de risque de marché. Elle représente la perte potentielle maximale d'un investisseur sur la valeur d'un portefeuille d'actifs, compte-tenu d'un horizon de détention et d'un niveau de confiance donnés. Par exemple, quand on dit qu'un portefeuille a une VaR de -3 M€ à 95% pour un horizon mensuel, cela signifie que l'on estime que ce portefeuille a 95% de chances de ne pas se déprécier de plus de 3 M€ en un mois. La VaR est donc ici le quantile d'ordre 5% de la distribution des rendements de ce portefeuille en un mois.

— Dans l'industrie pétrolière, les réserves sont classées en 3 catégories P10, P50 et P90, selon la probabilité qu'elles ont de pouvoir être exploitées dans le futur. Cela correspond aux quantiles d'ordre 10%, 50% et 90% de la loi du débit de pétrole du puits.

Par ailleurs, $[\tilde{q}_{n,1/4}, \tilde{q}_{n,3/4}]$ est un intervalle qui contient la moitié la plus centrale des observations. Sa largeur $\tilde{q}_{n,3/4} - \tilde{q}_{n,1/4}$ est un indicateur de dispersion, appelé **distance inter-quartiles**, qui est insensible aux valeurs aberrantes. Dans l'exemple des ampoules, elle vaut 94.1 h. On définit de la même manière des distances inter-déciles, inter-centiles,...

En `R`, la commande `quantile(x,p)` donne une version du quantile empirique d'ordre p légèrement différente de celle décrite ici (mais pour $p = 1/2$, on retrouve bien la médiane empirique) :

$$\begin{cases} x^*_{(n-1)p+1} & \text{si } (n-1)p \text{ est entier} \\ (1-q)x^*_{\lfloor (n-1)p \rfloor + 1} + q x^*_{\lfloor (n-1)p \rfloor + 2} & \text{sinon} \end{cases}$$

où $q = (n-1)p - \lfloor (n-1)p \rfloor$.

```
>  quantile(x,1/4)
  25%
27.15
```
(alors que $\tilde{q}_{n,1/4} = 24.3$)

La commande `summary(x)` donne en une seule fois les minimum, premier quartile, médiane, moyenne, troisième quartile et maximum de l'échantillon.

```
>  summary(x)
 Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
 5.40   27.15   62.20  83.15  111.70  251.30
```

Remarque : Puisqu'on considère les observations $x_1, \ldots, x_n$ comme des réalisations de variables aléatoires $X_1, \ldots, X_n$, toutes les quantités définies dans ce chapitre sont elles-mêmes des réalisations de variables aléatoires :

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

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

Chapitre 3

Estimation ponctuelle

3.1 Introduction

Dans ce chapitre, on suppose que les données $x_1, \ldots, x_n$ sont n réalisations indépendantes d'une même variable aléatoire sous-jacente X. Il est équivalent de supposer que $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi. Nous adopterons ici la seconde formulation, qui est plus pratique à manipuler.

Les techniques de statistique descriptive, comme l'histogramme ou le graphe de probabilités, permettent de faire des hypothèses sur la nature de la loi de probabilité des X_i. Des techniques statistiques plus sophistiquées, les tests d'adéquation, permettent de valider ou pas ces hypothèses.

On supposera ici que ces techniques ont permis d'adopter une famille de lois de probabilité bien précise (par exemple, loi normale, loi de Poisson, etc.) pour la loi des X_i, mais que la valeur du ou des paramètres de cette loi est inconnue.

On notera θ le paramètre inconnu. Le problème traité dans ce chapitre est celui de l'**estimation** du paramètre θ. Comme on l'a déjà dit, il s'agit de donner, au vu des observations $x_1, \ldots, x_n$, une approximation ou une évaluation de θ que l'on espère la plus proche possible de la vraie valeur inconnue. On pourra proposer une unique valeur vraisemblable pour θ (**estimation ponctuelle**, dans ce chapitre) ou un ensemble de valeurs vraisemblables (**estimation ensembliste** ou **région de confiance**, dans le chapitre suivant).

On notera $F(x;\theta)$ la fonction de répartition des X_i. Pour les variables aléatoires discrètes on notera $P(X = x;\theta)$ les probabilités élémentaires, et pour les variables aléatoires continues on notera $f(x;\theta)$ la densité. Par exemple, quand X est de loi exponentielle $\exp(\lambda)$, on aura $F(x;\lambda) = 1 - e^{-\lambda x}$ et $f(x;\lambda) = \lambda e^{-\lambda x}$.

3.2 Méthodes d'estimation

Il existe de nombreuses méthodes pour estimer un paramètre θ. Par exemple, nous avons déjà vu des estimations graphiques à partir des graphes de probabilité. Nous avons aussi utilisé le principe qu'une probabilité peut s'estimer par une proportion.

Dans cette section, nous ne nous intéressons qu'aux deux méthodes d'estimation les

plus usuelles, la méthode des moments et la méthode du maximum de vraisemblance.

Mais il faut d'abord définir précisément ce que sont une estimation et surtout un estimateur.

3.2.1 Définition d'un estimateur

Pour estimer θ on ne dispose que des données $x_1, \ldots, x_n$, donc une estimation de θ sera une fonction de ces observations.

Définition 5 *Une* **statistique** t *est une fonction des observations* $x_1, \ldots, x_n$ *:*

$$\begin{array}{rl} t: & \mathbb{R}^n \to \mathbb{R}^m \\ & (x_1, \ldots, x_n) \mapsto t(x_1, \ldots, x_n) \end{array}$$

Par exemple, $\bar{x}_n = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, x_1^*, $(x_1, x_3 + x_4, 2\ln x_6)$ sont des statistiques.

Puisque les observations $x_1, \ldots, x_n$ sont des réalisations des variables aléatoires $X_1, \ldots, X_n$, la quantité calculable à partir des observations $t(x_1, \ldots, x_n)$ est une réalisation de la variable aléatoire $t(X_1, \ldots, X_n)$. Et on retrouve par exemple le fait que $\bar{x}_n = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ est une réalisation de $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Pour simplifier les écritures, on note souvent $t_n = t(x_1, \ldots, x_n)$ et $T_n = t(X_1, \ldots, X_n)$. Par abus, on donne le même nom de statistique aux deux quantités, mais dans une perspective d'estimation, on va nommer différemment t_n et T_n.

Définition 6 *Un* **estimateur** *d'une grandeur* θ *est une statistique* T_n *à valeurs dans l'ensemble des valeurs possibles de* θ*. Une* **estimation** *de* θ *est une réalisation* t_n *de l'estimateur* T_n*.*

Un estimateur est donc une variable aléatoire, alors qu'une estimation est une valeur déterministe. Dans l'exemple des ampoules, l'estimateur de λ est $1/\bar{X}_n$ et l'estimation de λ est 0.012.

3.2.2 La méthode des moments

3.2.2.1. L'estimateur des moments (EMM)

C'est la méthode la plus naturelle, que nous avons déjà utilisée sans la formaliser. L'idée de base est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc...

Si le paramètre à estimer est l'espérance de la loi des X_i, alors on peut l'estimer par la moyenne empirique de l'échantillon. Autrement dit, si $\theta = E[X]$, alors l'**estimateur** de θ **par la méthode des moments (EMM)** est $\tilde{\theta}_n = \bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Plus généralement, pour $\theta \in \mathbb{R}$, si $E[X] = \varphi(\theta)$, où φ est une fonction inversible, alors l'estimateur de θ par la méthode des moments est $\tilde{\theta}_n = \varphi^{-1}\left(\bar{X}_n\right)$.

De la même manière, on estime la variance de la loi des X_i par la variance empirique de l'échantillon $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$.

Plus généralement, si la loi des X_i a deux paramètres θ_1 et θ_2 tels que $(E[X], Var[X])$ = $\varphi(\theta_1, \theta_2)$, où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont $\left(\tilde{\theta}_{1n}, \tilde{\theta}_{2n}\right) = \varphi^{-1}\left(\bar{X}_n, S_n^2\right)$.

Ce principe peut naturellement se généraliser aux moments de tous ordres, centrés ou non centrés : $E\left[(X - E[X])^k\right]$ et $E\left[X^k\right]$, $k \geq 1$.

3.2.2.2. Exemples

Exemple 1 : loi de Bernoulli

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$, $E[X] = p$. Donc l'estimateur de p par la méthode des moments est $\tilde{p}_n = \bar{X}_n$. Cet estimateur n'est autre que la proportion de 1 dans l'échantillon. On retrouve donc le principe d'estimation d'une probabilité par une proportion.

Exemple 2 : loi exponentielle

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi exponentielle $\exp(\lambda)$, $E[X] = 1/\lambda$. Donc l'estimateur de λ par la méthode des moments est $\tilde{\lambda}_n = 1/\bar{X}_n$.

Exemple 3 : loi normale

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(m, \sigma^2)$, $E[X] = m$ et $Var[X] = \sigma^2$, donc les estimateurs de m et σ^2 par la méthode des moments sont $\tilde{m}_n = \bar{X}_n$ et $\tilde{\sigma}_n^2 = S_n^2$.

Exemple 4 : loi gamma

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi gamma $G(a, \lambda)$, $E[X] = a/\lambda$ et $Var[X] = a/\lambda^2$. On en déduit facilement que :

$$\lambda = \frac{E[X]}{Var[X]} \quad \text{et} \quad a = \frac{[E[X]]^2}{Var[X]}$$

Donc les EMM de a et λ sont :

$$\tilde{\lambda}_n = \frac{\bar{X}_n}{S_n^2} \quad \text{et} \quad \tilde{a}_n = \frac{\bar{X}_n^2}{S_n^2}$$

Remarque : L'usage veut que la même notation $\tilde{\theta}_n$ désigne à la fois l'estimateur de θ (variable aléatoire) et l'estimation correspondante (réalisation de cette variable aléatoire sur l'expérience considérée). Par exemple, dans le cas de la loi exponentielle, $\tilde{\lambda}_n$

désigne aussi bien $1/\bar{X}_n$ que $1/\bar{x}_n$. Il faudra prendre garde à ne pas confondre les deux notions.

3.2.3 La méthode du maximum de vraisemblance

3.2.3.1. La fonction de vraisemblance

Définition 7 *Quand les observations sont toutes discrètes ou toutes continues, on appelle* **fonction de vraisemblance** *(ou plus simplement vraisemblance) pour l'échantillon $x_1, \ldots, x_n$, la fonction du paramètre θ :*

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n; \theta) & \textit{si les } X_i \textit{ sont discrètes} \\ \\ f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) & \textit{si les } X_i \textit{ sont continues} \end{cases}$$

Dans tous les exemples que nous traiterons ici, les X_i sont indépendantes et de même loi. Dans ce cas, la fonction de vraisemblance s'écrit :

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n} P(X_i = x_i; \theta) = \prod_{i=1}^{n} P(X = x_i; \theta) & \text{si les } X_i \text{ sont discrètes} \\ \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} f(x_i; \theta) & \text{si les } X_i \text{ sont continues} \end{cases}$$

Remarque : La probabilité et la densité utilisées dans cette définition sont des fonctions des observations $x_1, \ldots, x_n$, dépendant du paramètre θ. A l'inverse, la fonction de vraisemblance est considérée comme une fonction de θ dépendant des observations $x_1, \ldots, x_n$, ce qui permet, par exemple, de dériver cette fonction par rapport à θ.

3.2.3.2. Exemple introductif

Dans cet exemple, $n = 1$. On considère que l'on sait que X_1 est de loi binomiale $\mathcal{B}(15, p)$, avec p inconnu. On observe $x_1 = 5$ et on cherche à estimer p. La fonction de vraisemblance est :

$$\mathcal{L}(p; 5) = P(X_1 = 5; p) = \binom{15}{5} p^5 (1-p)^{15-5}$$

C'est la probabilité d'avoir observé un 5 quand la valeur du paramètre est p. Calculons-là pour quelques valeurs de p.

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mathcal{L}(p; 5)$	0.01	0.10	0.21	0.19	0.09	0.02	0.003	10^{-4}	$2\,10^{-7}$

On tire de cette table que quand $p = 0.8$, c'est-à-dire quand X_1 est de loi $\mathcal{B}(15, 0.8)$, il n'y a qu'une chance sur 10000 d'observer $x_1 = 5$. En revanche, il y a 21% de chances

d'observer un 5 quand $p = 0.3$. Il est donc beaucoup plus vraisemblable que p soit égal à 0.3 plutôt qu'à 0.8. En suivant ce raisonnement, on aboutit à dire que la valeur la plus vraisemblable de p est celle pour laquelle la probabilité d'observer un 5 est maximale. C'est donc la valeur de p qui maximise la fonction de vraisemblance.

Pour la calculer, on peut annuler la dérivée de la vraisemblance. Mais on remarque que la vraisemblance est un produit. Comme il est plus commode de maximiser (ou de dériver) une somme qu'un produit, on utilise le fait que la valeur qui rend maximale une fonction rend aussi maximal son logarithme. On va donc plutôt maximiser le logarithme de la fonction de vraisemblance, qu'on appelle la **log-vraisemblance**. Pour notre exemple, la log-vraisemblance vaut :

$$\ln \mathcal{L}(p; x_1) = \ln \binom{15}{x_1} + x_1 \ln p + (15 - x_1) \ln(1 - p)$$

Sa dérivée est :

$$\frac{\partial}{\partial p} \ln \mathcal{L}(p; x_1) = \frac{x_1}{p} - \frac{15 - x_1}{1 - p} = \frac{x_1 - 15p}{p(1 - p)}$$

qui s'annule pour $p = \dfrac{x_1}{15} = \dfrac{5}{15} = \dfrac{1}{3}$. Donc la valeur la plus vraisemblable de p est $\dfrac{1}{3}$. La vraisemblance maximale est $\mathcal{L}(\dfrac{1}{3}; 5) = 21.4\%$.

3.2.3.3. L'estimateur de maximum de vraisemblance (EMV)

En suivant le raisonnement précédent, pour n quelconque, il est logique de dire que la valeur la plus vraisemblable de θ est la valeur pour laquelle la probabilité d'observer $x_1, \ldots, x_n$ est la plus forte possible. Cela revient à faire comme si c'était l'éventualité la plus probable qui s'était produite au cours de l'expérience.

Définition 8 *L'***estimation de maximum de vraisemblance** *de θ est la valeur $\hat{\theta}_n$ de θ qui rend maximale la fonction de vraisemblance $\mathcal{L}(\theta; x_1, \ldots, x_n)$. L'***estimateur de maximum de vraisemblance (EMV)** *de θ est la variable aléatoire correspondante.*

Comme dans l'exemple, dans la plupart des cas, la fonction de vraisemblance s'exprime comme un produit. Donc $\hat{\theta}_n$ sera en général calculé en maximisant la log-vraisemblance :

$$\hat{\theta}_n = \arg\max_{\theta} \ln \mathcal{L}(\theta; x_1, \ldots, x_n)$$

Quand $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ et que toutes les dérivées partielles ci-dessous existent, $\hat{\theta}_n$ est solution du système d'équations appelées **équations de vraisemblance** :

$$\forall j \in \{1, \ldots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; x_1, \ldots, x_n) = 0$$

A priori, une solution de ce système d'équations pourrait être un minimum de la vraisemblance. Mais on peut montrer que la nature d'une fonction de vraisemblance fait que c'est bien un maximum que l'on obtient.

Il est fréquent que le système des équations de vraisemblance n'ait pas de solution explicite. Dans ce cas, on le résoud par des méthodes numériques, comme la méthode de Newton-Raphson. En `R`, la maximisation numérique peut se faire à l'aide de la commande `optim`.

3.2.3.4. Exemples

Exemple 1 : loi de Bernoulli

Si les X_i sont de loi $\mathcal{B}(p)$, on a :

$$P(X_i = x_i; p) = \left\{ \begin{array}{ll} p & \text{si } x_i = 1 \\ 1-p & \text{si } x_i = 0 \end{array} \right. = p^{x_i} (1-p)^{1-x_i}$$

Donc la fonction de vraisemblance est :

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

D'où $\ln \mathcal{L}(p; x_1, \ldots, x_n) = \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p)$.

Alors $\dfrac{\partial}{\partial p} \ln \mathcal{L}(p; x_1, \ldots, x_n) = \dfrac{\sum_{i=1}^{n} x_i}{p} - \dfrac{n - \sum_{i=1}^{n} x_i}{1-p} = \dfrac{\sum_{i=1}^{n} x_i - np}{p(1-p)}$, qui s'annule pour $p = \dfrac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$. Par conséquent, l'EMV de p est $\hat{p}_n = \bar{X}_n$.

Exemple 2 : loi exponentielle

Si les X_i sont de loi $\exp(\lambda)$, la fonction de vraisemblance est :

$$\mathcal{L}(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

D'où $\ln \mathcal{L}(\lambda; x_1, \ldots, x_n) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$.

Alors $\dfrac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda; x_1, \ldots, x_n) = \dfrac{n}{\lambda} - \sum_{i=1}^{n} x_i$, qui s'annule pour $\lambda = \dfrac{n}{\sum_{i=1}^{n} x_i} = \dfrac{1}{\bar{x}_n}$.

Par conséquent, l'EMV de λ est $\hat{\lambda}_n = \dfrac{1}{\bar{X}_n}$.

Exemple 3 : loi normale

Si les X_i sont de loi $\mathcal{N}(m, \sigma^2)$, la fonction de vraisemblance est :

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

D'où $\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{n}{2}\ln\sigma^2 - \dfrac{n}{2}\ln 2\pi - \dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$.

On doit annuler les dérivées partielles de ce logarithme par rapport à m et σ^2. On a :

- $\dfrac{\partial}{\partial m}\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i - m) = \dfrac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - nm\right)$, qui s'annule pour $m = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$.
- $\dfrac{\partial}{\partial \sigma^2}\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2$, qui s'annule pour $\sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$.

$\hat{m}_n$ et $\hat{\sigma}_n^2$ sont les valeurs de m et σ^2 qui vérifient les deux conditions en même temps. On a donc $\hat{m}_n = \bar{X}_n$ et $\hat{\sigma}_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = S_n^2$.

Exemple 4 : loi gamma

Si les X_i sont de loi gamma $G(a, \lambda)$, la fonction de vraisemblance est :

$$\mathcal{L}(a, \lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i; a, \lambda) = \prod_{i=1}^{n} \frac{\lambda^a}{\Gamma(a)}\, e^{-\lambda x_i}\, {x_i}^{a-1} = \frac{\lambda^{na}}{[\Gamma(a)]^n}\, e^{-\lambda\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} {x_i}^{a-1}$$

D'où $\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = na\ln\lambda - n\ln\Gamma(a) - \lambda\sum_{i=1}^{n} x_i + (a-1)\sum_{i=1}^{n}\ln x_i$.

On doit annuler les dérivées partielles de ce logarithme par rapport à a et λ. On a :

- $\dfrac{\partial}{\partial \lambda}\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = \dfrac{na}{\lambda} - \sum_{i=1}^{n} x_i$ qui s'annule pour $\lambda = \dfrac{na}{\sum_{i=1}^{n} x_i} = \dfrac{a}{\bar{x}_n}$.
- $\dfrac{\partial}{\partial a}\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = n\ln\lambda - n\dfrac{\Gamma'(a)}{\Gamma(a)} + \sum_{i=1}^{n}\ln x_i$.

En remplaçant λ par $a/\bar{x}_n$, on obtient que $\hat{a}_n$ est solution de l'équation implicite :

$$n\ln a - n\ln\bar{X}_n - n\frac{\Gamma'(a)}{\Gamma(a)} + \sum_{i=1}^{n}\ln X_i = 0$$

Il n'y a pas d'expression explicite de $\hat{a}_n$. Cette équation est à résoudre par des méthodes numériques. Une fois $\hat{a}_n$ déterminé, on en déduit $\hat{\lambda}_n = \hat{a}_n / \bar{X}_n$.

Remarque 1 : Dans les trois premiers exemples, la méthode des moments et la méthode du maximum de vraisemblance donnent les mêmes résultats. Ce n'est le cas que pour quelques lois de probabilité parmi les plus élémentaires. En fait, dans la plupart des cas, les deux méthodes fournissent des estimateurs différents.

C'est le cas de la loi gamma. On a deux estimateurs différents pour chaque paramètre. On doit donc se demander quel est le meilleur d'entre eux. Cela amène à se poser la question de la qualité et de l'optimalité d'un estimateur, ce qui fait l'objet de la section suivante.

Remarque 2 : On pourrait croire au vu de ces exemples que le calcul des estimateurs des moments est beaucoup plus simple que celui des estimateurs de maximum de vraisemblance. Mais ce n'est pas vrai en général.

3.3 Qualité d'un estimateur

En toute généralité, θ peut-être un paramètre à plusieurs dimensions, mais on supposera dans toute cette section et dans la suivante que θ est un réel. Cela signifie par exemple que, quand X est de loi normale $\mathcal{N}(m, \sigma^2)$, on s'intéressera séparément à la qualité des estimateurs de m et de σ^2. Les estimateurs T_n considérés ici seront donc des variables aléatoires réelles. Pour $\theta \in \mathbb{R}^d$, $d \geq 2$, toutes les notions de ces sections sont généralisables, mais la complexité des résultats augmente notablement. Par exemple, la notion de variance est remplacée par celle de matrice de covariance.

3.3.1 Estimateur sans biais et de variance minimale (ESBVM)

Un estimateur T_n de θ sera un bon estimateur s'il est suffisamment proche, en un certain sens, de θ. Il faut donc définir une mesure de l'écart entre θ et T_n. On appelle cette mesure le **risque** de l'estimateur. On a intérêt à ce que le risque d'un estimateur soit le plus petit possible.

Par exemple, les risques $T_n - \theta$, $|T_n - \theta|$, $(T_n - \theta)^2$ expriment bien un écart entre T_n et θ. Mais comme il est plus facile d'utiliser des quantités déterministes que des quantités aléatoires, on s'intéresse en priorité aux espérances des quantités précédentes. En particulier :

Définition 9

- *Le* **biais** *de T_n est $E[T_n - \theta] = E[T_n] - \theta$.*
- *Le* **risque quadratique** *ou* **erreur quadratique moyenne** *est :*

$$EQM(T_n) = E\left[(T_n - \theta)^2\right]$$

Dans le cas du biais, le risque peut être nul :

Définition 10 *Un* **estimateur** T_n *de* θ *est* **sans biais** *si et seulement si* $E[T_n] = \theta$. *Il est* **biaisé** *si et seulement si* $E[T_n] \neq \theta$.

Le biais mesure une erreur systématique d'estimation de θ par T_n. Par exemple, si $E[T_n] - \theta < 0$, cela signifie que T_n aura tendance à sous-estimer θ.

L'erreur quadratique moyenne s'écrit :

$$\begin{aligned} EQM(T_n) &= E\left[(T_n - \theta)^2\right] = E\left[(T_n - E[T_n] + E[T_n] - \theta)^2\right] \\ &= E\left[(T_n - E[T_n])^2\right] + 2E\left[T_n - E[T_n]\right]\left[E[T_n] - \theta\right] + E\left[(E[T_n] - \theta)^2\right] \\ &= Var[T_n] + \left[E[T_n] - \theta\right]^2 \\ &= \text{Variance de l'estimateur + carré de son biais} \end{aligned}$$

Si T_n est un estimateur sans biais, $EQM(T_n) = Var[T_n]$. On a donc intérêt à ce qu'un estimateur soit sans biais et de faible variance. Par ailleurs, on en déduit immédiatement que de deux estimateurs sans biais, le meilleur est celui qui a la plus petite variance.

La variance d'un estimateur mesure sa variabilité. Si l'estimateur est sans biais, cette variabilité est autour de θ. Si on veut estimer correctement θ, il ne faut pas que cette variabilité soit trop forte.

En pratique, si on observe plusieurs jeux de données similaires, on obtient une estimation de θ pour chacun d'entre eux. Alors si l'estimateur est de faible variance, ces estimations seront toutes proches les unes des autres, et s'il est sans biais leur moyenne sera très proche de θ.

Dans l'exemple des niveaux de bruit vu en TD, on a estimé le niveau de bruit moyen m par la moyenne empirique des $n = 20$ mesures effectuées $\bar{x}_n = 64.2$. Si on fait 20 autres mesures, on obtiendra une nouvelle valeur de cette moyenne. Ces deux valeurs sont deux estimations différentes de l'espérance m de la loi, deux réalisations de la même variable aléatoire $\bar{X}_n$. $\bar{X}_n$ est l'estimateur de m. Si on répète plusieurs fois cette expérience, les différentes moyennes obtenues doivent être toutes proches les unes des autres si l'estimateur est de faible variance. Si l'estimateur est sans biais, ces valeurs seront centrées sur le vraie valeur (inconnue) de m.

Enfin, il est logique de s'attendre à ce que, plus la taille des données augmente, plus on a d'information sur le phénomène aléatoire observé, donc meilleure sera l'estimation. En théorie, avec une observation infinie, on devrait pouvoir estimer θ sans aucune erreur. On peut traduire cette affirmation par le fait que le risque de l'estimateur T_n doit tendre vers 0 quand la taille n de l'échantillon tend vers l'infini. Cela revient à dire que l'estimateur T_n doit converger, en un certain sens, vers θ.

Il s'agit en fait d'étudier la convergence de la suite de variables aléatoires $\{T_n\}_{n\geq 1}$ vers la constante θ. Il existe plusieurs types de convergence de suites de variables aléatoires, décrites dans la section suivante. Ici on s'intéresse à la convergence en moyenne quadratique : l'estimateur T_n converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne tend vers 0 quand n tend vers l'infini. Si T_n est sans biais, il sera convergent en moyenne quadratique si et seulement si sa variance tend vers 0 quand n tend vers l'infini.

Finalement, on considèrera que le meilleur estimateur possible de θ est un **estimateur sans biais et de variance minimale (ESBVM)**. Un tel estimateur n'existe pas forcément.

Il existe des méthodes pour déterminer directement un ESBVM dans certains cas. Elles sont basées sur des techniques (exhaustivité, complétion), qui seront abordées dans le cours de Statistique Inférentielle Avancée. Dans le cadre de ce cours, on pourra parfois montrer facilement qu'un estimateur est un ESBVM en utilisant la quantité d'information de Fisher, définie plus loin.

Remarque : Ce n'est pas parce que T_n est un bon estimateur de θ que $\varphi(T_n)$ est un bon estimateur de $\varphi(\theta)$. Par exemple, il est fréquent d'avoir $E[T_n] = \theta$ et $E[\varphi(T_n)] \neq \varphi(\theta)$.

3.3.2 Convergences, théorème central-limite, loi des grands nombres

La plus forte des convergences de suites de variables aléatoires est la convergence presque sûre. Ce concept nécessite d'avoir défini une variable aléatoire comme une application mesurable d'un espace probabilisé dans un autre. Une suite de variables aléatoires $\{X_n\}_{n\geq 1}$ **converge presque sûrement** vers la variable aléatoire X si et seulement si $P\left(\left\{\omega; \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$.

Une suite de variables aléatoires $\{X_n\}_{n\geq 1}$ **converge en probabilité** vers la variable aléatoire X si et seulement si $\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$.

Une suite de variables aléatoires $\{X_n\}_{n\geq 1}$ **converge en loi** vers la loi de probabilité de fonction de répartition F si et seulement si $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ en tout point x où F est continue. Cela signifie que, quand n est grand, la loi de probabilité de X_n est approximativement la loi de fonction de répartition F.

Comme on l'a vu plus haut, une convergence très utile en statistique est la convergence en moyenne quadratique ou dans L^2. L^2 est l'ensemble des variables aléatoires réelles X telles que $E[X^2] < \infty$. Une suite de variables aléatoires $\{X_n\}_{n\geq 1}$ de L^2 **converge en moyenne quadratique** vers la variable aléatoire X si et seulement si $\lim_{n\to\infty} E\left[|X_n - X|^2\right] = 0$.

On montre que la convergence presque sûre entraîne la convergence en probabilité, qui elle-même entraîne la convergence en loi. On montre également que la convergence en moyenne quadratique entraîne la convergence en probabilité. Mais il n'y a pas de lien entre la convergence en moyenne quadratique et la convergence presque sûre.

Deux résultats fondamentaux en probabilités et statistique sont le théorème central-limite et la loi des grands nombres.

Théorème 1 Théorème Central-Limite *: Soit $\{X_n\}_{n\geq 1}$ une suite de variables aléatoi-res réelles indépendantes et de même loi, d'espérance $E[X]$ et d'écart-type $\sigma[X] = \sqrt{Var[X]}$ finis.*

Pour tout $n \geq 1$, on pose :

$$Z_n = \frac{\sum_{i=1}^{n} X_i - nE[X]}{\sqrt{nVar[X]}} = \sqrt{n}\,\frac{\bar{X}_n - E[X]}{\sigma[X]}$$

Alors la suite $\{Z_n\}_{n\geq 1}$ converge en loi vers la loi normale centrée-réduite, ce qui s'écrit :

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

Concrètement, cela signifie que la loi de toute variable aléatoire égale à la somme d'un nombre "suffisamment grand" de variables aléatoires indépendantes et de même loi est approximativement une loi normale. Plus précisément, pour n grand, $\sum_{i=1}^{n} X_i$ est approximativement de loi $\mathcal{N}\left(nE[X], nVar[X]\right)$. Ce qui est remarquable, c'est que ce résultat est vrai quelle que soit la loi des X_i.

De très nombreux phénomènes naturels sont la résultante d'un grand nombre de phénomènes élémentaires identiques, indépendants et additifs ce qui justifie l'importance (et le nom) de la loi normale.

Théorème 2 Loi forte des grands nombres *: Soit $\{X_n\}_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes et de même loi, d'espérance $E[X]$. Soit $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors la suite $\{\bar{X}_n\}_{n\geq 1}$ converge presque sûrement vers $E[X]$, ce qui s'écrit :*

$$\bar{X}_n \xrightarrow{ps} E[X]$$

Concrètement, cela signifie que quand on fait un très grand nombre d'expériences identiques et indépendantes, la moyenne des réalisations de la variable aléatoire à laquelle on s'intéresse tend vers l'espérance de sa loi. Ce résultat permet de justifier l'idée naturelle d'estimer une espérance par une moyenne empirique et une probabilité par une proportion.

3.3.3 Quantité d'information, efficacité d'un estimateur

La quantité d'information de Fisher est un outil précieux pour évaluer la qualité d'un estimateur. Elle n'est définie que sous certaines conditions de régularité. Ces conditions sont trop fastidieuses pour être écrites ici, mais sont vérifiées par la plupart des lois de probabilité usuelles.

Définition 11 *Pour $\theta \in \mathbb{R}$, si la loi des observations vérifie les conditions de régularité, on appelle* **quantité d'information** *(de Fisher) sur θ apportée par l'échantillon $x_1, \ldots, x_n$, la quantité :*

$$\mathcal{I}_n(\theta) = Var\left[\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta; X_1, \ldots, X_n)\right]$$

On montre que l'on a également :

$$\mathcal{I}_n(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln\mathcal{L}(\theta; X_1, \ldots, X_n)\right]$$

L'intérêt de la quantité d'information de Fisher est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur sans biais de θ. Ce résultat s'exprime sous la forme de la propriété suivante :

Propriété 1 Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR) *: Si la loi des observations vérifie les conditions de régularité, alors pour tout estimateur T_n de θ, on a :*

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

Ce résultat est particulièrement intéressant pour les estimateurs sans biais. En effet, si T_n est un estimateur sans biais de θ, alors $E[T_n] = \theta$, donc $Var[T_n] \geq \frac{1}{\mathcal{I}_n(\theta)}$.

La quantité $\frac{1}{\mathcal{I}_n(\theta)}$ est appelée la **borne de Cramer-Rao**. L'inégalité FDCR dit donc que la variance d'un estimateur sans biais quelconque de θ est forcément supérieure à cette borne.

Définition 12 *On appelle* **efficacité** *d'un estimateur T_n la quantité :*

$$Eff(T_n) = \frac{\left[\frac{\partial}{\partial\theta}E[T_n]\right]^2}{\mathcal{I}_n(\theta)Var\,[T_n]}$$

On a $0 \leq Eff(T_n) \leq 1$.
T_n est dit un estimateur **efficace** *si et seulement si $Eff(T_n) = 1$.*
T_n est dit **asymptotiquement efficace** *si et seulement si $\lim_{n\to+\infty} Eff(T_n) = 1$.*

- Si T_n est un estimateur sans biais de θ, $Eff(T_n) = \frac{1}{\mathcal{I}_n(\theta)Var[T_n]}$.
- Si un estimateur sans biais est efficace, sa variance est égale à la borne de Cramer-Rao, donc c'est forcément un ESBVM.
- Il est possible qu'il n'existe pas d'estimateur efficace de θ. Alors, s'il existe un ESBVM de θ, sa variance est strictement supérieure à la borne de Cramer-Rao.
- Si la valeur de la borne de Cramer-Rao est très grande, il est impossible d'estimer correctement θ car tous les estimateurs sans biais possibles auront une forte variance.

On peut donc juger de la qualité d'un estimateur sans biais en calculant son efficacité.

La définition de la quantité d'information ci-dessus est une définition générale, applicable quelle que soit la nature des variables aléatoires observées. Quand celles-ci sont indépendantes et de même loi, il est facile de voir que $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$. Par exemple, pour des variables aléatoires continues de densité f :

$$\begin{aligned}\mathcal{I}_n(\theta) &= Var\left[\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X_1,\ldots,X_n)\right] = Var\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n}f(X_i;\theta)\right] \\ &= Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(X_i;\theta)\right] = Var\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right] \\ &= \sum_{i=1}^{n}Var\left[\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right] = n\mathcal{I}_1(\theta)\end{aligned}$$

On peut remarquer que le calcul des dérivées de la fonction de vraisemblance est utile à la fois pour la détermination de l'estimateur de maximum de vraisemblance et pour le calcul de la quantité d'information.

3.4 Propriétés des EMM et des EMV

3.4.1 Propriétés des estimateurs des moments

Si $\theta = E[X]$, alors l'EMM de θ est $\tilde{\theta}_n = \bar{X}_n$. La justification de cette méthode est la loi des grands nombres, qui dit que $\bar{X}_n$ converge presque sûrement vers $E[X]$. Donc, si $\theta = E[X]$, $\bar{X}_n$ est un estimateur de θ convergent presque sûrement. Autrement dit, si on a beaucoup d'observations, on peut estimer une espérance par une moyenne empirique.

On peut en fait montrer facilement que $\bar{X}_n$ est un bon estimateur de $\theta = E[X]$, sans utiliser la loi des grands nombres :

$$E[\bar{X}_n] = E\Big[\frac{1}{n}\sum_{i=1}^{n}X_i\Big] = \frac{1}{n}\sum_{i=1}^{n}E[X_i] = \frac{1}{n}\,n\theta = \theta$$

Donc $\bar{X}_n$ est un estimateur sans biais de $\theta = E[X]$.

La variance de $\bar{X}_n$ est :

$$Var[\bar{X}_n] = Var\Big[\frac{1}{n}\sum_{i=1}^{n}X_i\Big] = \frac{1}{n^2}\sum_{i=1}^{n}Var[X_i] = \frac{Var[X]}{n}$$

car les X_i sont indépendantes, donc la variance de leur somme est égale à la somme de leurs variances, qui sont toutes égales à $Var[X]$. $Var[\bar{X}_n]$ tend vers 0 quand n tend vers l'infini. Par conséquent :

Propriété 2 *La moyenne empirique $\bar{X}_n$ est un estimateur sans biais et convergent en moyenne quadratique de $E[X]$.*

On considère maintenant l'estimation de la variance de la loi des X_i par la variance empirique de l'échantillon $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$.

Déterminons le biais de cet estimateur.

$$\begin{aligned} E[S_n^2] &= E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\Big] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}_n^2] = E[X^2] - E[\bar{X}_n^2] \\ &= Var[X] + E[X]^2 - Var[\bar{X}_n] - E[\bar{X}_n]^2 \\ &= Var[X] + E[X]^2 - \frac{Var[X]}{n} - E[X]^2 = \Big(1 - \frac{1}{n}\Big)Var[X] \\ &= \frac{n-1}{n}\, Var[X] \neq Var[X] \end{aligned}$$

Donc, contrairement à ce qu'on pourrait croire, la variance empirique S_n^2 n'est pas un estimateur sans biais de $Var[X]$. Cet estimateur n'est qu'asymptotiquement sans biais.

En revanche, on voit que $E\left[\frac{n}{n-1}\, S_n^2\right] = \frac{n}{n-1}E[S_n^2] = Var[X]$. On pose donc $S'^2_n = \frac{n}{n-1}\, S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. S'^2_n est appelée **variance estimée** de l'échantillon. Le résultat précédent montre que c'est un estimateur sans biais de $Var[X]$.

Par ailleurs, on montre que

$$Var[S'^2_n] = \frac{1}{n(n-1)}\left[(n-1)E\left[(X - E[X])^4\right] - (n-3)Var[X]^2\right]$$

qui tend vers 0 quand n tend vers l'infini. Par conséquent :

Propriété 3 *La* **variance estimée** $S'^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ *est un estimateur sans biais et convergent en moyenne quadratique de* $Var[X]$.

C'est pour cela que la commande `var(x)` en `R` donne la variance estimée, et non pas la variance empirique de l'échantillon `x`.

On peut montrer également que S'^2_n et S_n^2 convergent toutes les deux presque sûrement vers $Var[X]$.

Remarque 1 : On n'a pas de résultat général sur la qualité de S_n comme estimateur de l'écart-type de la loi, $\sigma[X] = \sqrt{Var[X]}$. A priori, ni S_n ni S'_n ne sont des estimateurs sans biais de $\sigma[X]$.

Remarque 2 : Le simple exemple de la variance montre qu'un estimateur des moments n'est pas forcément sans biais. On peut montrer qu'un EMM est asymptotiquement sans biais et convergent presque sûrement.

3.4.2 Propriétés des estimateurs de maximum de vraisemblance

Un estimateur de maximum de vraisemblance n'est pas forcément unique (la vraisemblance peut avoir plusieurs maxima), ni sans biais, ni de variance minimale, ni efficace. Mais il possède d'excellentes propriétés asymptotiques, pour peu que la loi des observations vérifie les conditions de régularité déjà évoquées pour la quantité d'information.

Propriété 4 *Si les X_i sont indépendants et de même loi dépendant d'un paramètre réel θ, cette loi vérifiant les conditions de régularité, on a :*

- *$\hat{\theta}_n$ converge presque sûrement vers θ.*
- *$\sqrt{\mathcal{I}_n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, ce qui signifie que, quand n est grand, $\hat{\theta}_n$ est approximativement de loi $\mathcal{N}\left(\theta, \dfrac{1}{\mathcal{I}_n(\theta)}\right)$. On en déduit que $\hat{\theta}_n$ est asymptotiquement gaussien, sans biais (son espérance tend vers θ) et efficace (sa variance tend vers la borne de Cramer-Rao $1/\mathcal{I}_n(\theta)$). Cette propriété peut aussi s'écrire :*

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

- *Si $\hat{\theta}_n$ est l'EMV de θ, alors $\varphi(\hat{\theta}_n)$ est l'EMV de $\varphi(\theta)$. De plus, si φ est dérivable, on a :*

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

  Ce résultat est connu sous le nom de **méthode delta**. *Quand n est grand, $\varphi(\hat{\theta}_n)$ est donc approximativement de loi $\mathcal{N}\left(\varphi(\theta), \dfrac{\varphi'(\theta)^2}{\mathcal{I}_n(\theta)}\right)$.*

- *En général, l'EMV est meilleur que l'EMM au sens où $Var[\hat{\theta}_n] \leq Var[\tilde{\theta}_n]$. C'est au moins vrai asymptotiquement.*

Le fait que l'EMV soit asymptotiquement sans biais et efficace fait que, si on a beaucoup de données, on est pratiquement certains que la méthode du maximum de vraisemblance est la meilleure méthode d'estimation possible. C'est pourquoi cette méthode est considérée comme globalement la meilleure et est utilisée de préférence à toute autre méthode, y compris celle des moments.

3.4.3 Exemples

Exemple 1 : loi de Bernoulli

L'EMM et EMV de p est $\hat{p}_n = \bar{X}_n$. On sait que $\bar{X}_n$ est un estimateur sans biais de $E[X]$. Or l'espérance de la loi $\mathcal{B}(p)$ est p, donc $\hat{p}_n$ est un estimateur sans biais de p.

On sait aussi que $Var[\bar{X}_n] = \dfrac{Var[X]}{n} = \dfrac{p(1-p)}{n}$, donc $\hat{p}_n$ est convergent en moyenne quadratique.

La quantité d'information est :

$$\begin{aligned}
\mathcal{I}_n(p) &= Var\left[\frac{\partial}{\partial p}\ln \mathcal{L}(p; X_1, \ldots, X_n)\right] = Var\left[\frac{\sum_{i=1}^{n} X_i - np}{p(1-p)}\right] = \frac{Var\left[\sum_{i=1}^{n} X_i\right]}{p^2(1-p)^2} \\
&= \frac{np(1-p)}{p^2(1-p)^2} \text{ car } \sum_{i=1}^{n} X_i \text{ est de loi binomiale } \mathcal{B}(n,p) \\
&= \frac{n}{p(1-p)}
\end{aligned}$$

On a donc $Var[\hat{p}_n] = \dfrac{1}{\mathcal{I}_n(p)}$, ce qui prouve que $\hat{p}_n$ est un estimateur efficace. Par conséquent, $\hat{p}_n$ est un ESBVM de p.

Exemple 2 : loi normale

Les EMM et EMV de m et σ^2 sont $\tilde{m}_n = \bar{X}_n$ et $\tilde{\sigma}_n^2 = S_n^2$. On sait qu'il vaut mieux estimer σ^2 par ${S'}_n^2$. Il est facile de montrer que $\bar{X}_n$ est un ESBVM de m. ${S'}_n^2$ est également un ESBVM de σ^2, mais la démonstration est moins immédiate.

L'EMV de $\sigma = \sqrt{\sigma^2}$ est $S_n = \sqrt{S_n^2}$. ${S'}_n^2$ est un ESBVM de σ^2, mais ni ${S'}_n$ ni S_n ne sont des ESBVM de σ (ce ne sont même pas des estimateurs sans biais). On montre qu'en fait, un ESBVM de σ est $\sqrt{\dfrac{n-1}{2}}\dfrac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\,{S'}_n$.

Chapitre 4

Intervalles de confiance

4.1 Problématique et définition

Jusqu'à présent, on a estimé un paramètre θ par une unique valeur $\hat{\theta}_n$ (estimation ponctuelle). Si l'estimateur $\hat{\theta}_n$ possède de bonnes propriétés (sans biais, variance minimale, efficacité), on peut s'attendre à ce que $\hat{\theta}_n$ soit proche de la vraie valeur de θ. Cependant, il est très peu probable que $\hat{\theta}_n$ soit exactement égal à θ. En particulier, si la loi de $\hat{\theta}_n$ est continue, on est certains que $P(\hat{\theta}_n = \theta) = 0$.

Dans l'exemple des niveaux de bruit vu en TD, on a estimé m par $\hat{m}_n = 64.2$. Mais il est bien évidemment possible que m ("le vrai") soit égal à 63 ou 66.

Par conséquent, plutôt que d'estimer θ par la seule valeur $\hat{\theta}_n$, il semble raisonnable de proposer un ensemble de valeurs vraisemblables pour θ, qu'il est logique de prendre proches de $\hat{\theta}_n$. Cet ensemble de valeurs est appelé **estimation ensembliste** ou **région de confiance**. Dire que toutes les valeurs de cet ensemble sont vraisemblables pour θ, c'est dire qu'il y a une forte probabilité que θ appartienne à cet ensemble.

On supposera dans ce chapitre que $\theta \in \mathbb{R}$, donc la région de confiance sera un intervalle (on parle parfois de "fourchette"). Quand $\theta \in \mathbb{R}^d$, avec $d \geq 2$, la région de confiance est en général un ellipsoïde.

Définition 13 *Un* **intervalle de confiance** *de* **seuil** *(ou* **niveau de signification***)* $\alpha \in [0, 1]$ *pour un paramètre* θ*, est un intervalle aléatoire* I *tel que* $P(\theta \in I) = 1 - \alpha$*.*

α est la probabilité que le paramètre θ n'appartienne pas à l'intervalle I, c'est à dire la probabilité que l'on se trompe en affirmant que $\theta \in I$. C'est donc une probabilité d'erreur, qui doit être assez petite. Les valeurs usuelles de α sont 10%, 5%, 1%, etc.

Remarque fondamentale : Les intervalles de confiance suscitent souvent des erreurs d'interprétation et des abus de langage. La raison essentielle de ce problème est la suivante.

Dans l'écriture $P(\theta \in I)$, θ est une grandeur inconnue mais non aléatoire. Ce sont les bornes de l'intervalle I qui sont aléatoires. Posons $I = [Z_1, Z_2]$. Z_1 et Z_2 sont des variables aléatoires. Soient z_1 et z_2 les réalisations de Z_1 et Z_2 pour une expérience donnée.

À titre indicatif, prenons l'exemple des niveaux de bruit, pour lequel $\theta = m$. Admettons que z_1=60 et z_2= 68. Il est correct de dire une phrase du type : "m a 95% de chances d'être compris entre Z_1 et Z_2", mais il est incorrect de dire : "m a 95% de chances d'être compris entre 60 et 68". En effet, dans cette dernière écriture, il n'y a rien d'aléatoire. m est ou n'est pas dans l'intervalle [60, 68]. La probabilité que m soit compris entre 60 et 68 est donc 0 ou 1, mais pas 95%.

En fait, si on recommence 100 fois l'expérience, on aura 100 réalisations du couple (Z_1, Z_2), et donc 100 intervalles de confiance différents. En moyenne, m sera dans 95 de ces intervalles.

Par conséquent, il vaut mieux dire : "on a une confiance de 95% dans le fait que m soit compris entre 60 et 68".

Le problème à régler est de trouver un procédé pour déterminer un intervalle de confiance pour un paramètre θ. Il semble logique de proposer un intervalle de confiance centré sur un estimateur performant $\hat{\theta}_n$, c'est-à-dire de la forme $I = \left[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon\right]$. Il reste alors à déterminer ε de sorte que :

$$P(\theta \in I) = P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha$$

Mais cette démarche ne va pas toujours aboutir. En effet, α est un réel fixé à l'avance qui ne doit pas dépendre de θ. ε ne doit pas non plus dépendre de θ pour que l'intervalle soit utilisable. Par conséquent, on ne peut déterminer un ε vérifiant l'égalité ci-dessus que si la loi de probabilité de $\hat{\theta}_n - \theta$ ne dépend pas de θ, ce qui n'est pas toujours le cas.

Si cet intervalle de confiance est petit, l'ensemble des valeurs vraisemblables pour θ est resserré autour de $\hat{\theta}_n$. Si l'intervalle de confiance est grand, des valeurs vraisemblables pour θ peuvent être éloignées de $\hat{\theta}_n$. Donc un intervalle de confiance construit à partir d'un estimateur permet de mesurer la précision de cet estimateur.

Pour trouver un intervalle de confiance, il existe plusieurs méthodes. La plus efficace consiste à chercher une **fonction pivotale**, c'est à dire une variable aléatoire fonction à la fois du paramètre θ et des observations $X_1, \ldots, X_n$, dont la loi de probabilité ne dépende pas de θ. Dans la suite de ce chapitre, nous allons illustrer cette méthodologie par des exemples, en déterminant des intervalles de confiance pour :

- la moyenne et la variance dans un échantillon de loi normale ;
- une proportion, c'est-à-dire le paramètre d'un échantillon de loi de Bernoulli.

4.2 Intervalles de confiance pour les paramètres de la loi normale

Pour un échantillon de loi normale, il va s'avérer utile d'utiliser le théorème de Fisher suivant.

Théorème de Fisher. *Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors, en posant $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, $S_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ et $S'^2_n = \dfrac{n}{n-1}S_n^2$, on a :*

- $\sum_{i=1}^{n} X_i$ *est de loi* $\mathcal{N}(nm, n\sigma^2)$.
- $\bar{X}_n$ *est de loi* $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.
- $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - m)^2$ *est de loi* χ^2_n.
- $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{nS_n^2}{\sigma^2}$ *est de loi* χ^2_{n-1}.
- $\bar{X}_n$ *et* S_n^2 *sont indépendantes.*
- $\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n} = \sqrt{n-1}\,\frac{\bar{X}_n - m}{S_n}$ *est de loi de Student* $St(n-1)$.

4.2.1 Intervalle de confiance pour la moyenne

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(m, \sigma^2)$, on sait que l'ESBVM de m est $\bar{X}_n$. La première idée est donc de chercher un intervalle de confiance pour m de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$. Conformément à ce qui précède, le problème revient, pour α fixé, à chercher ε tel que $P\left(|\bar{X}_n - m| \leq \varepsilon\right) = 1 - \alpha$.

Puisque $\bar{X}_n$ est de loi $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$, $U = \frac{\bar{X}_n - m}{\sqrt{\sigma^2/n}} = \sqrt{n}\,\frac{\bar{X}_n - m}{\sigma}$ est de loi $\mathcal{N}(0,1)$.

Alors $P(|\bar{X}_n - m| \leq \varepsilon) = P\left(|U| \leq \frac{\sqrt{n}\,\varepsilon}{\sigma}\right) = 1 - P\left(|U| > \frac{\sqrt{n}\,\varepsilon}{\sigma}\right) = 1 - \alpha$. Or la table 2 de la loi normale donne la valeur u_α telle que $P(|U| > u_\alpha) = \alpha$. Par conséquent, $\frac{\sqrt{n}\,\varepsilon}{\sigma} = u_\alpha$, donc $\varepsilon = \frac{\sigma}{\sqrt{n}}\,u_\alpha$. D'où le résultat :

Propriété 5 *Un intervalle de confiance de seuil* α *pour le paramètre* m *de la loi* $\mathcal{N}(m, \sigma^2)$ *est :*

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

Le problème est que cet intervalle n'est utilisable que si on connaît la valeur de σ. Or, dans la pratique, on ne connaît jamais les vraies valeurs des paramètres. Une idée naturelle est alors de remplacer σ par un estimateur, par exemple S'_n.

Mais si on fait cela, $P\left(m \in \left[\bar{X}_n - \frac{S'_n}{\sqrt{n}}\,u_\alpha, \bar{X}_n + \frac{S'_n}{\sqrt{n}}\,u_\alpha\right]\right) = P\left(\left|\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n}\right| \leq u_\alpha\right)$ $\neq 1 - \alpha$ car $\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n}$ n'est pas de loi $\mathcal{N}(0,1)$. Donc $\left[\bar{X}_n - \frac{S'_n}{\sqrt{n}}\,u_\alpha, \bar{X}_n + \frac{S'_n}{\sqrt{n}}\,u_\alpha\right]$ n'est pas un intervalle de confiance de seuil α pour m.

Mais le théorème de Fisher dit que $\sqrt{n}\,\dfrac{\bar{X}_n - m}{S'_n}$ est de loi de Student $St(n-1)$. On peut donc écrire $P\left(|\bar{X}_n - m| \leq \varepsilon\right) = P\left(|Y| \leq \sqrt{n}\,\dfrac{\varepsilon}{S'_n}\right) = 1 - P\left(|Y| > \sqrt{n}\,\dfrac{\varepsilon}{S'_n}\right)$, où Y est une variable aléatoire de loi $St(n-1)$. Or la table de la loi de Student donne la valeur $t_{n-1,\alpha}$ telle que $P\left(|Y| > t_{n-1,\alpha}\right) = \alpha$. Par conséquent, $\sqrt{n}\,\dfrac{\varepsilon}{S'_n} = t_{n-1,\alpha}$, donc $\varepsilon = \dfrac{S'_n}{\sqrt{n}}\,t_{n-1,\alpha}$. D'où le résultat :

Propriété 6 *Un intervalle de confiance de seuil α pour le paramètre m de la loi $\mathcal{N}(m, \sigma^2)$ est :*

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

Dans l'exemple des niveaux de bruit, on a n = 20, $\bar{x}_n$ = 64.2 et s'_n = 5.15. Pour α = 5%, la table de la loi de Student donne $t_{19,0.05}$ = 2.093. On en déduit qu'un intervalle de confiance de seuil 5% pour le niveau de bruit moyen est [61.8, 66.7].

Interprétation : La meilleure estimation possible du niveau de bruit moyen est 64.2 db. De plus, on a une confiance de 95% dans le fait que ce niveau de bruit moyen est compris entre 61.8 db et 66.7 db.

En `R`, u_α est obtenu par la commande `qnorm(1-alpha/2)` et $t_{n,\alpha}$ par la commande `qt(1-alpha/2,n)`. Pour obtenir l'intervalle de confiance, on procède donc de la façon suivante.

```
>  bruit <- c(54.8, 55.4, 57.7, 59.6, 60.1, 61.2, 62.0, 63.1,
   63.5, 64.2, 65.2, 65.4, 65.9, 66.0, 67.6, 68.1, 69.5, 70.6,
   71.5, 73.4)
>  n <- length(bruit)
>  alpha <- 0.05
>  mean(bruit)-sd(bruit)*qt(1-alpha/2,n-1)/sqrt(n)
[1] 61.82992
>  mean(bruit)+sd(bruit)*qt(1-alpha/2,n-1)/sqrt(n)
[1] 66.65008
```

Pour une raison qui n'apparaitra qu'au chapitre suivant, on peut aussi avoir directement l'intervalle de confiance à l'aide de la commande `t.test`. Dans la réponse à cette commande, l'intervalle est donné sous le nom de `95 percent confidence interval`.

```
>  t.test(bruit, conf.level=0.95)

        One Sample t-test

data: bruit
t = 55.7889, df = 19, p-value <  2.2e-16
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 61.82992 66.65008
sample estimates:
mean of x
    64.24
```

Remarque 1 : Rien n'oblige à prendre un intervalle de confiance du type $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$ (intervalle de confiance bilatéral). Tout intervalle I tel que $P(m \in I) = 1 - \alpha$ convient. Par exemple, les intervalles de confiance unilatéraux $\left[\bar{X}_n - \frac{S'_n}{\sqrt{n}}\, t_{n-1,2\alpha}, +\infty\right[$ et $]-\infty, \bar{X}_n + \frac{S'_n}{\sqrt{n}}\, t_{n-1,2\alpha}\right]$ sont aussi des intervalles de confiance de seuil α pour m, qui fournissent des bornes inférieure et supérieure pour l'estimation de m. Pour l'exemple :

```
>  mean(bruit)+sd(bruit)*qt(1-alpha,n-1)/sqrt(n)
[1] 66.23107
```

signifie qu'on a une confiance de 95 % dans le fait que le niveau de bruit moyen est inférieur à 66.2 db.

Remarque 2 : La largeur de l'intervalle de confiance est $2\frac{S'_n}{\sqrt{n}}\, t_{n-1,\alpha}$. La table de la loi de Student permet de constater que c'est une fonction décroissante en n comme en α, ce qui est logique. En effet, plus on a d'observations, plus on a d'informations, donc plus l'incertitude sur le paramètre diminue et plus l'intervalle de confiance est étroit. On retrouve le fait qu'un intervalle de confiance petit signifie qu'on estime le paramètre avec précision. D'autre part, plus α est petit, moins on veut prendre de risques de se tromper en disant que m est dans l'intervalle, donc plus on aura tendance à prendre des intervalles larges. À la limite, on ne prend aucun risque ($\alpha = 0$) en proposant comme intervalle de confiance $\mathbb{R}$ tout entier !

En pratique, un intervalle de confiance trop large n'a aucun intérêt (ça n'apporte pas grand chose d'avoir une forte confiance dans le fait que m est compris entre 1 et 10000), donc il faut parfois accepter un risque d'erreur relativement fort pour obtenir un intervalle de confiance utilisable.

Remarque 3 : La variable aléatoire $\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n}$ est une fonction des observations $X_1, \ldots, X_n$ et du paramètre m pour lequel on recherche un intervalle de confiance, dont la loi de probabilité ne dépend pas des paramètres du modèle m et σ^2. C'est ce qu'on a appelé une **fonction pivotale** et c'est ce que nous utiliserons à partir de maintenant pour construire des intervalles de confiance.

4.2.2 Intervalle de confiance pour la variance

Conformément à ce qui précède, on recherche une fonction pivotale, c'est à dire une fonction des observations $X_1, \ldots, X_n$ et de σ^2, dont la loi de probabilité ne dépend ni de m ni de σ^2. Une telle fonction est donnée par le théorème de Fisher : $\frac{nS_n^2}{\sigma^2}$ est de loi

χ^2_{n-1}.

On a donc, quels que soient les réels a et b, $0 < a < b$:

$$\begin{aligned} P\left(a \leq \frac{nS_n^2}{\sigma^2} \leq b\right) &= P\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right) \quad \text{d'une part} \\ &= F_{\chi^2_{n-1}}(b) - F_{\chi^2_{n-1}}(a) \quad \text{d'autre part.} \end{aligned}$$

Il y a une infinité de façons possibles de choisir a et b de sorte que cette probabilité soit égale à $1-\alpha$. La façon la plus usuelle de procéder est d'"équilibrer les risques", c'est-à-dire de prendre a et b tels que $F_{\chi^2_{n-1}}(b) = 1 - \frac{\alpha}{2}$ et $F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$.

La table de la loi du χ^2 donne la valeur $z_{n,\alpha}$ telle que, quand Z est une variable aléatoire de loi χ^2_n, alors $P(Z > z_{n,\alpha}) = 1 - F_{\chi^2_n}(z_{n,\alpha}) = \alpha$.

Alors, pour $b = z_{n-1,\alpha/2}$ et $a = z_{n-1,1-\alpha/2}$, on a bien $P\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right) = 1-\alpha$. D'où le résultat :

Propriété 7 *Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(m, \sigma^2)$ est :*

$$\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S'^2_n}{z_{n-1,\alpha/2}}, \frac{(n-1)S'^2_n}{z_{n-1,1-\alpha/2}}\right]$$

Dans l'exemple des niveaux de bruit, on a $n = 20$ et $s'^2_n = 26.5$. Pour $\alpha = 5\%$, on obtient $z_{19,0.025} = 32.85$ et $z_{19,0.975} = 8.91$. On en déduit qu'un intervalle de confiance de seuil 5% pour la variance du niveau de bruit est [15.3, 56.6].

On constate que cet intervalle de confiance est très large : l'estimation de la variance est moins précise que celle de la moyenne.

En `R`, $z_{n,\alpha}$ est obtenu par la commande `qchisq(1-alpha,n)`.

```
>   (n-1)*var(bruit)/qchisq(1-alpha/2,n-1)
[1] 15.33675
>   (n-1)*var(bruit)/qchisq(alpha/2,n-1)
[1] 56.57071
```

Remarque 1 : $P(a \leq \sigma^2 \leq b) = P\left(\sqrt{a} \leq \sigma \leq \sqrt{b}\right)$, donc un intervalle de confiance de seuil α pour l'écart-type σ est :

$$\left[\sqrt{\frac{n}{z_{n-1,\alpha/2}}}\, S_n, \sqrt{\frac{n}{z_{n-1,1-\alpha/2}}}\, S_n\right]$$

Remarque 2 : L'intervalle de confiance est de la forme $[\varepsilon_1 S_n^2, \varepsilon_2 S_n^2]$ avec $\varepsilon_1 < 1$ et $\varepsilon_2 > 1$ et non pas de la forme $[S_n^2 - \varepsilon, S_n^2 + \varepsilon]$. En fait, si on cherche un intervalle de confiance pour σ^2 de la forme $[S_n^2 - \varepsilon, S_n^2 + \varepsilon]$, la démarche ne va pas aboutir, et on ne peut pas le savoir à l'avance. C'est l'intérêt des fonctions pivotales, qui imposent d'elles-mêmes la forme de l'intervalle de confiance.

4.3 Intervalle de confiance pour une proportion

Le problème connu sous le nom d'"intervalle de confiance pour une proportion" est en fait le problème de la détermination d'un intervalle de confiance pour le paramètre p de la loi de Bernoulli, au vu d'un échantillon $X_1, \ldots, X_n$ de cette loi. Il s'agit donc de l'exemple 1 du chapitre précédent. On a montré que l'ESBVM de p est $\hat{p}_n = \bar{X}_n$.

Nous allons illustrer le problème traité à l'aide d'un exemple issu du contexte des sondages. Une élection oppose deux candidats A et B. Un institut de sondage interroge 800 personnes sur leurs intentions de vote. 420 déclarent voter pour A et 380 pour B. Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra le candidat A le jour de l'élection. Pour être dans le cadre de modélisation annoncé, il faut supposer que les n personnes interrogées ont des votes indépendants et que la probabilité qu'une personne choisie au hasard vote pour A est p. Notons que cela ne signifie pas qu'un électeur vote au hasard, c'est le choix d'une personne dans la population qui est aléatoire.

On pose :

$$x_i = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ personne interrogée déclare voter pour A} \\ 0 & \text{sinon} \end{cases}$$

Alors X_i est bien de loi $\mathcal{B}(p)$ et les X_i sont indépendantes. On remarque qu'on ne connait pas (heureusement !) le détail des votes des 800 personnes interrogées, mais cela n'est pas nécessaire puisque seul le nombre de personnes ayant voté pour A suffit pour estimer p : l'ESBVM de p est $\hat{p}_n = \bar{x}_n = 420/800 = 52.5\%$.

L'institut de sondage estime donc que le candidat A va gagner l'élection. Pour évaluer l'incertitude portant sur cette estimation, A demande un intervalle de confiance de seuil 5% pour p.

Pour déterminer directement un intervalle de confiance pour p, il faudrait trouver une fonction pivotale, c'est à dire une fonction des X_i et de p dont la loi ne dépende pas de p. On sait que si $X_1, \ldots, X_n$ sont indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$, alors $T = \sum_{i=1}^{n} X_i = n\hat{p}_n$ est de loi binomiale $\mathcal{B}(n, p)$. Mais la loi binomiale n'est pas facile à manipuler, donc ce résultat ne permet pas d'en déduire une fonction pivotale simple. On montre le résultat suivant :

Propriété 8 *Un intervalle de confiance exact de seuil α pour p est :*

$$\left[\frac{1}{1 + \dfrac{n-T+1}{T} f_{2(n-T+1),2T,\alpha/2}}, \frac{1}{1 + \dfrac{n-T}{T+1} f_{2(n-T),2(T+1),1-\alpha/2}} \right]$$

où les $f_{\nu_1,\nu_2,\alpha}$ sont des quantiles de la loi de Fisher-Snedecor.

En `R`, $f_{\nu_1,\nu_2,\alpha}$ est obtenu par la commande `qf(1-alpha,nu1,nu2)`. Quelques uns de ces quantiles sont aussi donnés dans les tables de la loi de Fisher-Snedecor.

Si on ne dispose pas de logiciel, cet intervalle n'est pas facile à utiliser car il nécessite l'emploi de nombreuses tables. C'est pourquoi on utilise souvent un intervalle de confiance approché, basé sur l'approximation de la loi binomiale par la loi normale.

En effet, comme n est très grand, on peut appliquer le théorème central-limite sur la loi des X_i et dire que $\sqrt{n}\,\dfrac{\bar{X}_n - E(X)}{\sqrt{Var(X)}} = \sqrt{n}\,\dfrac{\hat{p}_n - p}{\sqrt{p(1-p)}} = \dfrac{T-np}{\sqrt{np(1-p)}}$ est approximativement de loi $\mathcal{N}(0,1)$, ce qui fournit la fonction pivotale cherchée.

On écrit alors $P\left(\left|\dfrac{T-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha\right) \approx 1-\alpha$. Pour en déduire un intervalle de confiance, il suffit d'écrire $\left|\dfrac{T-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$ sous la forme $Z_1 \leq p \leq Z_2$. On a :

$$\left|\frac{T-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha \Leftrightarrow \frac{(T-np)^2}{np(1-p)} \leq u_\alpha^2 \Leftrightarrow p^2\left(n+u_\alpha^2\right) - p\left(2T+u_\alpha^2\right) + \frac{T^2}{n} \leq 0$$

Ce trinôme en p est toujours positif sauf entre ses racines. Donc ses deux racines sont les bornes de l'intervalle de confiance cherché :

$$\left[\frac{\dfrac{T}{n} + \dfrac{u_\alpha^2}{2n} - u_\alpha\sqrt{\dfrac{u_\alpha^2}{4n^2} + \dfrac{T(n-T)}{n^3}}}{1+\dfrac{u_\alpha^2}{n}}, \frac{\dfrac{T}{n} + \dfrac{u_\alpha^2}{2n} + u_\alpha\sqrt{\dfrac{u_\alpha^2}{4n^2} + \dfrac{T(n-T)}{n^3}}}{1+\dfrac{u_\alpha^2}{n}}\right]$$

Pour les valeurs usuelles de α et pour n grand, on peut négliger u_α^2 par rapport à n. En écrivant $\hat{p}_n = T/n$, on obtient un résultat final très simple. Puisque ce résultat utilise le théorème central-limite, donc n'est valable que quand n est suffisamment grand, l'intervalle obtenu porte le nom d'intervalle de confiance asymptotique :

Propriété 9 *Un intervalle de confiance asymptotique de seuil α pour p est :*

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

Dans l'exemple du sondage, $n = 800$, $t = 420$ et $\hat{p}_n = 52.5\%$. Avec `R`, on trouve $f_{762,840,0.025} = 1.1486$ et $f_{760,842,0.975} = 0.8702$. On obtient alors comme intervalle de confiance exact [0.4897, 0.5601] :

```
>  1/(1+(n-t+1)*qf(1-alpha/2,2*(n-t+1), 2*t)/t)
[1] 0.4897328
>  1/(1+(n-t)*qf(alpha/2,2*(n-t),2*(t+1))/(t+1))
[1] 0.5600823
```

On retrouve ce résultat avec la commande `binom.test` :

```
>  binom.test(t,n,conf.level=1-alpha)

        Exact binomial test

data: t and n
number of successes = 420, number of trials = 800,
p-value = 0.1679
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.4897328 0.5600823
sample estimates:
probability of success
             0.525
```

Pour $\alpha = 5\%$, $u_{0.05} = 1.96$. L'intervalle de confiance asymptotique de seuil 5% est alors [0.4904, 0.5596] :

```
>  pchap<-t/n
>  pchap-qnorm(1-alpha/2)*sqrt(pchap*(1-pchap)/n)
[1] 0.4903957
>  pchap+qnorm(1-alpha/2)*sqrt(pchap*(1-pchap)/n)
[1] 0.5596043
```

On constate que les deux intervalles sont extrêmement proches. C'est souvent le cas, ce qui fait que l'intervalle asymptotique est très largement utilisé.

En arrondissant, on conclut que l'on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Le problème est que cet intervalle de confiance n'est pas entièrement situé au-dessus de 50%. Il semble donc possible que, malgré l'estimation de 52.5%, le candidat A soit battu. On voit donc que ce qui importe dans cette situation, ce n'est pas vraiment d'estimer p, mais de déterminer si on peut admettre avec une confiance raisonnable que p est supérieur à 50%. C'est, entre autres, l'objet de la théorie des tests d'hypothèses, qui sera présentée au chapitre suivant.

Une autre possibilité pour résoudre le problème est de déterminer à quelle condition l'intervalle de confiance pour p sera entièrement au-dessus des 50%. Il s'agit donc de réduire la taille de l'intervalle de confiance. Si on prend l'intervalle asymptotique, sa largeur est $2u_\alpha \sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}}$. Donc, pour diminuer cette largeur, on peut, au choix, diminuer u_α ou augmenter n (en supposant qu'en augmentant n, on ne modifiera pas beaucoup la valeur de $\hat{p}_n$).

Diminuer u_α, c'est augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu. On retrouve ce qui a déjà été dit : pour obtenir des intervalles de confiance exploitables, il faut parfois accepter un risque d'erreur assez élevé.

Augmenter n, c'est augmenter le nombre de personnes interrogées. On peut même,

à α fixé, déterminer n de façon à ne pas dépasser une certaine largeur pour l'intervalle de confiance.

On sait que $\forall p \in [0,1], p(1-p) \leq 1/4$. Donc $2u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$. Par conséquent, si on détermine n tel que $\frac{u_\alpha}{\sqrt{n}} < l$, c'est à dire $n > \frac{u_\alpha^2}{l^2}$, on est sûr que la largeur de l'intervalle de confiance sera inférieure à l.

Pour $\alpha = 5\%$ et $n = 800$, $\frac{u_\alpha}{\sqrt{n}} = \frac{1.96}{\sqrt{800}} \approx 7\%$. La précision sur l'estimation de p est donc, avec une confiance de 95%, de plus ou moins 3.5%. C'est bien ce qu'on a constaté sur l'intervalle de confiance : $[49\%, 56\%]$. Si on veut, avec le même niveau de confiance, avoir une précision inférieure à 1%, il faudra interroger au moins $\frac{u_\alpha^2}{l^2} = \frac{1.96^2}{0.01^2} = 38416$ personnes. C'est rarement le cas dans les sondages, pour lesquels le nombre de personnes interrogées est en général de l'ordre de 1000.

En conclusion, il faut toujours tenir compte du nombre de personnes interrogées pour interpréter les résultats d'un sondage. C'est pour cela qu'il est obligatoire de préciser ce nombre quand les résultats du sondage sont publiés. Et il faut se méfier des conclusions péremptoires données à partir d'un échantillon de 1000 personnes.

Chapitre 5

Tests d'hypothèses

5.1 Introduction : le problème de décision

Dans tous les domaines, de l'expérimentation scientifique à la vie quotidienne, on est amené à prendre des décisions sur une activité risquée au vu de résultats d'expériences ou d'observation de phénomènes dans un contexte incertain. Par exemple :

- *informatique* : au vu des résultats des tests d'un nouveau logiciel, on doit décider si ce logiciel est suffisamment fiable et performant pour être mis en vente.
- *essais thérapeutiques* : décider si un nouveau traitement médical est meilleur qu'un ancien au vu du résultat de son expérimentation sur des malades.
- *finance* : au vu du marché, décider si on doit ou pas se lancer dans une opération financière donnée.
- *santé* : trancher sur la nocivité ou non des OGM ou des antennes de téléphonie mobile, décider s'il faut vacciner toute une population contre le covid.
- *justice* : décider si l'accusé est innocent ou coupable à partir des informations acquises pendant le procès.

Dans chaque cas, le **problème de décision** consiste à trancher, au vu d'observations, entre une hypothèse appelée **hypothèse nulle**, notée H_0, et une autre hypothèse dite **hypothèse alternative**, notée H_1. En général, on suppose qu'une et une seule de ces deux hypothèses est vraie. Un **test d'hypothèses** est une procédure qui permet de choisir entre ces deux hypothèses.

Dans un problème de décision, deux types d'erreurs sont possibles :

- **erreur de première espèce** : décider que H_1 est vraie alors que H_0 est vraie.
- **erreur de seconde espèce** : décider que H_0 est vraie alors que H_1 est vraie.

Les conséquences de ces deux erreurs peuvent être d'importances diverses. En général, une des erreurs est plus grave que l'autre :

- *informatique* : si on conclut à tort que le logiciel n'est pas assez fiable et performant, on engagera des dépenses inutiles pour le tester et l'analyser et on risque de se faire souffler le marché par la concurrence ; si on décide à tort qu'il est suffisamment fiable et performant, on va mettre en vente un produit qui ne satisfera pas la clientèle, ce qui peut coûter cher en image de marque comme en coût de

maintenance.

- *essais thérapeutiques* : on peut adopter un nouveau traitement moins efficace, voire pire que l'ancien, ou se priver d'un nouveau traitement plus efficace que l'ancien.
- *finance* : si on décide à tort que l'on peut lancer l'opération, on risque de perdre beaucoup d'argent ; si on décide à tort de ne pas lancer l'opération, on peut se priver d'un bénéfice important.
- *santé* : on peut dépenser des milliards d'euros en vaccins inutiles ou subir une pandémie grave à large échelle.
- *justice* : on peut condamner un innocent ou acquitter un coupable.

A toute décision correspond une probabilité de décider juste et une probabilité de se tromper :

- la probabilité de l'erreur de première espèce, qui est la probabilité de rejeter à tort H_0, est notée α et est appelée **seuil** ou **niveau de signification** du test. C'est la même terminologie que pour les intervalles de confiance, ce qui n'est pas un hasard, comme nous le verrons plus loin. Dans certains contextes, cette probabilité est appelée **risque fournisseur**.
- la probabilité de l'erreur de deuxième espèce est notée $1-\beta$ et est parfois appelée **risque client**.
- β est la probabilité de décider H_1 ou de rejeter H_0 à raison. Elle est appelée **puissance** du test.
- $1-\alpha$ est parfois appelée **niveau de confiance** du test.

Le tableau 5.1 résume simplement le rôle de ces probabilités.

Vérité / Décision	H_0	H_1
H_0	$1-\alpha$	$1-\beta$
H_1	α	β

TABLE 5.1 – probabilités de bonne et mauvaise décision dans un test d'hypothèses

L'idéal serait évidemment de trouver une procédure qui minimise les deux risques d'erreur en même temps. Malheureusement, on montre qu'ils varient en sens inverse, c'est-à-dire que toute procédure diminuant α va en général augmenter $1-\beta$ et réciproquement. Dans la pratique, on va donc considérer que l'une des deux erreurs est plus importante que l'autre, et tâcher d'éviter que cette erreur se produise. Il est alors possible que l'autre erreur survienne. Par exemple, dans le cas du procès, on fait en général tout pour éviter de condamner un innocent, quitte à prendre le risque d'acquitter un coupable.

On va choisir H_0 et H_1 de sorte que l'erreur que l'on cherche à éviter soit l'erreur de première espèce. Mathématiquement cela revient à se fixer la valeur du seuil du test α. Plus la conséquence de l'erreur est grave, plus α sera choisi petit. Les valeurs usuelles de α sont 10%, 5%, 1%, ou beaucoup moins. Le **principe de précaution** consiste à limiter au maximum la probabilité de se tromper, donc à prendre α très petit.

On appelle **règle de décision** une règle qui permette de choisir entre H_0 et H_1 au vu des observations $x_1, \ldots, x_n$, sous la contrainte que la probabilité de rejeter à tort H_0 est égale à α fixé. Une idée naturelle est de conclure que H_0 est fausse si il est très peu probable d'observer $x_1, \ldots, x_n$ quand H_0 est vraie.

Par exemple, admettons que l'on doive décider si une pièce est truquée ou pas au vu de 100 lancers de cette pièce. Si on observe 90 piles, il est logique de conclure que la pièce est truquée et on pense avoir une faible probabilité de se tromper en concluant cela. Mais si on observe 65 piles, que conclure ?

On appelle **région critique** du test, et on note W, l'ensemble des valeurs des observations $x_1, \ldots, x_n$ pour lesquelles on rejettera H_0. La région critique est souvent déterminée à l'aide du bon sens. Sinon, on utilisera une fonction pivotale ou des théorèmes d'optimalité. W dépend du seuil α et est déterminée a priori, indépendamment de la valeur des observations. Ensuite, si les observations appartiennent à W, on rejette H_0, sinon on ne la rejette pas.

Remarque : il vaut mieux dire "*ne pas rejeter* H_0" que "*accepter* H_0". En effet, si on rejette H_0, c'est que les observations sont telles qu'il est très improbable que H_0 soit vraie. Si on ne rejette pas H_0, c'est qu'on ne dispose pas de critères suffisants pour pouvoir dire que H_0 est fausse. Mais cela ne veut pas dire que H_0 est vraie. Un test permet de dire qu'une hypothèse est très probablement fausse ou seulement peut-être vraie. Par exemple, si on n'a pas de preuve qu'un accusé est coupable, cela ne veut pas forcément dire qu'il est innocent (et réciproquement).

Par conséquent, dans un problème de test, il faut choisir les hypothèses H_0 et H_1 de façon à ce que ce qui soit vraiment intéressant, c'est de rejeter H_0.

Récapitulons l'ensemble de la démarche à suivre pour effectuer un test d'hypothèses :

1. Choisir H_0 et H_1 de sorte que ce qui importe, c'est de ne pas se tromper en rejetant H_0.
2. Se fixer α selon la gravité des conséquences de l'erreur de première espèce.
3. Déterminer la région critique W.
4. Regarder si les observations se trouvent ou pas dans W.
5. Prendre la décision, c'est-à-dire conclure au rejet ou au non-rejet de H_0.

Pour le même problème de décision, plusieurs tests (c'est-à-dire plusieurs régions critiques) de même seuil sont souvent possibles. Dans ce cas, le meilleur de ces tests est celui qui minimisera la probabilité de l'erreur de seconde espèce, c'est à dire celui qui maximisera la puissance β. Le meilleur des tests possibles de seuil fixé est le **test le plus puissant**. Il arrive que l'on puisse le déterminer, mais pas toujours.

Dans de nombreux cas, les hypothèses d'un test peuvent se traduire sur la valeur d'un paramètre d'une loi de probabilité. Les tests de ce type sont appelés **tests paramétriques**. Dans l'exemple de l'élection, le problème est de trancher entre les deux hypothèses "$p \leq 1/2$" et "$p > 1/2$". Les tests qui ne portent pas sur la valeur d'un paramètre sont appelés **tests non paramétriques**. Il en existe de tous les types.

On commencera ce chapitre par formaliser le problème de test paramétrique quand l'observation est un échantillon d'une loi de probabilité. Puis on traitera le cas de la loi normale et de la loi de Bernoulli. On terminera ce chapitre en présentant le plus célèbre des tests d'hypothèses, le test du χ^2.

5.2 Formalisation du problème de test paramétrique sur un échantillon

Comme précédemment, les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi, dépendant d'un paramètre inconnu θ. On supposera que θ est un réel. Si θ est un paramètre vectoriel, on fera des tests sur chacune de ses composantes. Par exemple, on fera des tests sur la moyenne de la loi normale, puis des tests sur la variance, mais pas sur les deux en même temps.

Une **hypothèse** est **simple** si elle est du type "$\theta = \theta_0$", où θ_0 est un réel fixé. Une **hypothèse** est **composite** ou **multiple** si elle est du type "$\theta \in A$" où A est une partie de $\mathbb{R}$ non réduite à un élément.

5.2.1 Tests d'hypothèses simples

Un **test d'hypothèses simples** est un test dans lequel les hypothèses nulle et alternative sont simples toutes les deux. C'est donc un test du type

$$H_0 : \text{"}\theta = \theta_0\text{"} \text{ contre } H_1 : \text{"}\theta = \theta_1\text{"}.$$

Un tel test est un cas d'école : il permet de dire laquelle des deux valeurs θ_0 et θ_1 est la plus vraisemblable au vu des observations. Mais il ne prend pas en compte la possibilité que θ ne soit égal ni à θ_0 ni à θ_1. Pour cela, il faudra faire un test d'hypothèses composites.

Le seuil du test est la probabilité de rejeter à tort H_0, c'est à dire la probabilité que les observations soient dans la région critique quand la vraie valeur de θ est θ_0 :

$$\alpha = P\left((X_1, \ldots, X_n) \in W; \theta_0\right)$$

La puissance du test est la probabilité de rejeter à raison H_0, c'est à dire la probabilité que les observations soient dans la région critique quand la vraie valeur de θ est θ_1 :

$$\beta = P\left((X_1, \ldots, X_n) \in W; \theta_1\right)$$

5.2.2 Tests d'hypothèses composites

Un **test d'hypothèses composites** est un test dans lequel l'une au moins des deux hypothèses est composite. Les tests les plus usuels sont du type :

- **test bilatéral** : H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$" (seule H_1 est composite).
- **tests unilatéraux** : H_0 : "$\theta \leq \theta_0$" contre H_1 : "$\theta > \theta_0$" ou H_0 : "$\theta \geq \theta_0$" contre H_1 : "$\theta < \theta_0$" (H_0 et H_1 sont composites).

On pourrait aussi imaginer des tests du type H_0 : "$\theta \in [\theta_1, \theta_2]$" contre H_1 : "$\theta < \theta_1$ ou $\theta > \theta_2$". Toutes les variantes sont envisageables. Dans tous ces exemples, H_0 et H_1 sont complémentaires : des deux hypothèses, l'une est forcément vraie. C'est ce cas qui est important en pratique.

Quand une hypothèse est composite, la notion de puissance est à repréciser. En effet, β a été définie comme la probabilité de rejeter H_0 à raison, c'est à dire de rejeter H_0 quand H_1 est vraie. Or, dans les exemples ci-dessus, il y a une infinité de valeurs de θ pour lesquelles H_1 est vraie. Donc la puissance du test doit dépendre de la vraie valeur (inconnue) de θ, ce qui nous amène à redéfinir la puissance et le seuil d'un test :

Définition 14 *La* **puissance** *d'un test portant sur la valeur d'un paramètre réel θ est la fonction de θ définie par :*

$$\begin{array}{rl} \beta : & \mathbb{R} \to [0,1] \\ & \theta \mapsto \beta(\theta) = P\left((X_1, \ldots, X_n) \in W; \theta\right) \end{array}$$

Le **seuil** *du test est* $\alpha = \underset{H_0}{Sup}\, \beta(\theta)$.

$\beta(\theta)$ est la probabilité de rejeter H_0 quand la vraie valeur du paramètre est θ. $\alpha = \underset{H_0}{Sup}\, \beta(\theta)$ est la probabilité maximale de rejeter H_0 alors que H_0 est vraie, c'est à dire la plus forte probabilité de rejeter à tort H_0. Par exemple, pour un test bilatéral, $\alpha = \beta(\theta_0)$, et pour le premier test unilatéral présenté, $\alpha = \underset{\theta \leq \theta_0}{Sup}\, \beta(\theta)$.

Une fois H_0 et H_1 déterminées et α fixé, il faut construire la région critique W. Pour comprendre comment déterminer une région critique, nous allons détailler dans la section suivante la construction d'un test sur la moyenne d'une loi normale, à partir d'un exemple introductif.

5.3 Tests sur la moyenne d'une loi normale

5.3.1 Exemple introductif : essais thérapeutiques

Pour apaiser un certain type de maux de tête, on a l'habitude de traiter les malades avec un médicament A. Une étude statistique a montré que la durée de disparition de la douleur chez les malades traités avec A était une variable aléatoire de loi normale $\mathcal{N}(m_0, \sigma_0^2)$, avec $m_0 = 30$ mn et $\sigma_0 = 5$ mn. Un laboratoire pharmaceutique a conçu un nouveau médicament B et désire tester son efficacité. Pour cela, le nouveau médicament a été administré à n malades cobayes, et on a mesuré la durée de disparition de la douleur pour chacun d'entre eux : $x_1, \ldots, x_n$. Une étude de statistique descriptive sur ces données a amené les pharmacologues à considérer que cette durée était une variable aléatoire de loi normale $\mathcal{N}(m, \sigma^2)$.

Remarque : En toute rigueur, on ne devrait pas modéliser une durée (positive) par une variable aléatoire qui, comme pour la loi normale, peut prendre des valeurs négatives. En pratique, on peut le faire quand, pour les lois considérées, la probabilité que la variable soit négative est négligeable.

L'effet du nouveau médicament se traduit facilement sur la valeur de la durée moyenne de disparition de la douleur :

- "$m = m_0$" : le médicament B a en moyenne le même effet que le médicament A.
- "$m < m_0$" : le médicament B est en moyenne plus efficace que le médicament A.
- "$m > m_0$" : le médicament B est en moyenne moins efficace que le médicament A.

Nous reviendrons ultérieurement sur l'interprétation de la valeur de l'écart-type σ en termes d'efficacité du médicament.

Pour savoir s'il faut commercialiser B, il faut trancher entre ces 3 hypothèses. L'important est de ne pas se tromper si on décide de changer de médicament : il est préférable de conserver un médicament moins performant que le nouveau que d'adopter un médicament moins performant que l'ancien. Il faut donc que l'hypothèse "$m < m_0$" corresponde au rejet de H_0.

Par conséquent, nous allons tester H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" au vu de n réalisations indépendantes $x_1, \ldots, x_n$ de la loi $\mathcal{N}(m, \sigma^2)$.

5.3.2 Première idée

Puisque $\bar{X}_n$ est l'ESBVM de m, une première idée est de conclure que $m < m_0$ si et seulement si $\bar{x}_n < m_0$: la durée moyenne de disparition de la douleur sur les malades traités avec B est plus petite que ce qu'elle est sur les malades traités avec A. Cela revient à proposer comme région critique du test :

$$W = \{(x_1, \ldots, x_n); \bar{x}_n < m_0\}$$

Si $\bar{x}_n$ est beaucoup plus petit que m_0, il est en effet très probable que B soit plus efficace que A. Mais si $\bar{x}_n$ est proche de m_0 tout en étant plus petit, on risque de se tromper si on affirme que $m < m_0$. La probabilité de cette erreur, qui n'est autre que le risque de première espèce α, est très facile à calculer :

$$\begin{aligned} \alpha &= \underset{H_0}{Sup}\, \beta(m) = \underset{m \geq m_0}{Sup}\, P\left(\bar{X}_n < m_0; m\right) \\ &= \underset{m \geq m_0}{Sup}\, P\left(\sqrt{n}\,\frac{\bar{X}_n - m}{\sigma} < \sqrt{n}\,\frac{m_0 - m}{\sigma}; m\right) = \underset{m \geq m_0}{Sup}\, \phi\left(\sqrt{n}\,\frac{m_0 - m}{\sigma}\right) \end{aligned}$$

où ϕ est la fonction de répartition de la loi normale centrée-réduite. En effet, comme on l'a déjà vu, si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors $\bar{X}_n$ est de loi $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$ et $\sqrt{n}\,\frac{\bar{X}_n - m}{\sigma}$ est de loi $\mathcal{N}(0,1)$.

$\phi(u)$ est une fonction croissante de u, donc $\beta(m) = \phi\left(\sqrt{n}\,\frac{m_0 - m}{\sigma}\right)$ est une fonction décroissante de m. Par conséquent, $\alpha = \underset{m \geq m_0}{Sup}\, \beta(m) = \beta(m_0) = \phi(0) = 1/2$.

Il y a donc une chance sur deux de se tromper si on décide que B est plus efficace que A quand $\bar{x}_n < m_0$. C'est évidemment beaucoup trop.

5.3.3 Deuxième idée

On voit qu'il faut en fait rejeter H_0 quand $\bar{x}_n$ est *significativement plus petit* que m_0. Cela revient à prendre une région critique de la forme :

$$W = \{(x_1, \ldots, x_n); \bar{x}_n < l_\alpha\}, \text{ où } l_\alpha < m_0$$

La borne l_α dépend du seuil α que l'on s'est fixé. Moins on veut risquer de rejeter à tort H_0, plus α sera petit, et plus l_α sera petit. Le sens de l'expression *significativement plus petit* est lié à la valeur de α.

Un calcul analogue au précédent montre que :

$$\alpha = \underset{H_0}{Sup}\, \beta(m) = \underset{m \geq m_0}{Sup}\, P\left(\bar{X}_n < l_\alpha; m\right) = \underset{m \geq m_0}{Sup}\, \phi\left(\sqrt{n}\,\frac{l_\alpha - m}{\sigma}\right) = \phi\left(\sqrt{n}\,\frac{l_\alpha - m_0}{\sigma}\right)$$

On obtient donc $\sqrt{n}\,\dfrac{l_\alpha - m_0}{\sigma} = \phi^{-1}(\alpha)$, d'où $l_\alpha = m_0 + \dfrac{\sigma}{\sqrt{n}}\,\phi^{-1}(\alpha) = m_0 - \dfrac{\sigma}{\sqrt{n}}\,u_{2\alpha}$, avec les notations habituelles pour les quantiles de la loi normale.

En conclusion, on a :

Propriété 10 *Un test de seuil α de H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" est déterminé par la région critique :*

$$W = \left\{(x_1, \ldots, x_n); \bar{x}_n < m_0 - \frac{\sigma}{\sqrt{n}}\,u_{2\alpha}\right\}$$

5.3.4 Troisième idée

La région critique proposée ci-dessus pose un problème déjà rencontré à propos des intervalles de confiance : ce test est inutilisable si on ne connaît pas la vraie valeur de σ, ce qui est toujours le cas en pratique. Pour pallier cet inconvénient, on utilise la même procédure que pour les intervalles de confiance : on remplace σ par son estimateur S'_n, ce qui nécessite de remplacer la loi normale par la loi de Student.

Rappelons en effet que si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}(m, \sigma^2)$, alors $\sqrt{n}\,\dfrac{\bar{X}_n - m}{S'_n}$ est de loi $St(n-1)$. Alors, à partir d'une région critique de la forme $W = \{(x_1, \ldots, x_n); \bar{x}_n < l_\alpha\}$, on obtient :

$$\begin{aligned}\alpha &= \underset{H_0}{Sup}\, \beta(m) = \underset{m \geq m_0}{Sup}\, P\left(\bar{X}_n < l_\alpha; m\right) = \underset{m \geq m_0}{Sup}\, P\left(\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n} < \sqrt{n}\,\frac{l_\alpha - m}{S'_n}; m\right) \\ &= \underset{m \geq m_0}{Sup}\, F_{St(n-1)}\left(\sqrt{n}\,\frac{l_\alpha - m}{S'_n}\right) = F_{St(n-1)}\left(\sqrt{n}\,\frac{l_\alpha - m_0}{S'_n}\right)\end{aligned}$$

D'où $\sqrt{n}\,\dfrac{l_\alpha - m_0}{S'_n} = F^{-1}_{St(n-1)}(\alpha) = -t_{n-1,2\alpha}$, avec les notations habituelles pour les quantiles de la loi de Student. Finalement, $l_\alpha = m_0 - \dfrac{S'_n}{\sqrt{n}}\,t_{n-1,2\alpha}$.

En conclusion, on a :

Propriété 11 *Un test de seuil α de H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" est déterminé par la région critique :*

$$W = \left\{ (x_1, \ldots, x_n); \bar{x}_n < m_0 - \frac{s'_n}{\sqrt{n}} t_{n-1,2\alpha} \right\}$$

Remarque : La région critique peut aussi s'écrire :

$$W = \left\{ (x_1, \ldots, x_n); \sqrt{n} \frac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha} \right\}$$

Cette forme met en évidence l'utilisation de la variable aléatoire $\sqrt{n} \dfrac{\bar{X}_n - m_0}{S'_n}$ qui n'est autre que la fonction pivotale déjà vue à l'occasion des intervalles de confiance. C'est cette forme que l'on conservera dans la suite.

5.3.5 Exemple

Avec le médicament A, la durée moyenne de disparition de la douleur était 30 mn. On a administré le médicament B à 12 malades et relevé les durées de disparition de la douleur suivants :

25 28 20 32 17 24 41 28 25 30 27 24

La moyenne empirique de ces données est $\bar{x}_n$= 26.75 et l'écart-type estimé est $s'_n = 6.08$.

On décide de ne commercialiser B que si on est sûr à 95% qu'il est plus efficace que A. Cela revient donc à faire un test de H_0 : "$m \geq 30$" contre H_1 : "$m < 30$" au seuil $\alpha = 5\%$.

On voit qu'il s'agit finalement de déterminer si 26.75 est suffisamment inférieur à 30 pour que l'on puisse conclure que le médicament B réduit vraiment la durée de disparition de la douleur.

D'après ce qui précède, on rejettera H_0 si $\sqrt{n} \dfrac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha}$.

Or $\sqrt{n} \dfrac{\bar{x}_n - m_0}{s'_n} = \sqrt{12} \dfrac{26.75 - 30}{6.08} = -1.853$ et $t_{n-1,2\alpha} = t_{11,0.1} = 1.796$.

$-1.853 < -1.796$, donc les observations sont dans la région critique. On rejette donc H_0, ce qui signifie que l'on conclut que B est plus efficace que A, avec moins de 5% de chances de se tromper. Par conséquent, on peut lancer la commercialisation du médicament B.

5.3.6 La p-valeur

On voit ici le rôle fondamental du seuil α. Si on avait pris $\alpha = 1\%$, on aurait eu $t_{11,0.02} = 2.718$. Comme $-1.853 > -2.718$, on n'aurait pas rejeté H_0, donc on n'aurait pas adopté le médicament B.

Ce phénomène est normal : se fixer un seuil α petit revient à éviter au maximum d'adopter à tort le médicament B. Or un bon moyen de ne pas prendre ce risque, c'est

de conserver le médicament A. Le test de seuil $\alpha = 0$ consiste à conserver le médicament A quelles que soient les observations : la probabilité de rejeter à tort H_0 est nulle quand on ne rejette jamais H_0 ! En pratique, plus α est petit, moins on aura tendance à rejeter H_0. D'une certaine façon, cela signifie que le principe de précaution conduit au conservatisme...

Il est donc fondamental de bien savoir évaluer les risques et de choisir α en connaissance de cause.

Cet exemple avec $\alpha = 1\%$ permet également de comprendre la nuance entre "ne pas rejeter H_0" et "accepter H_0" : on va conclure que rien ne prouve que B est plus efficace que A, mais on ne va évidemment pas conclure que A est plus efficace que B.

La remarque précédente met en évidence l'existence d'un seuil critique α_c tel que pour tout seuil α supérieur à α_c, on rejettera H_0, et pour tout seuil α inférieur à α_c, on ne rejettera pas H_0. Cette valeur α_c est appelée la **p-valeur**.

α_c vérifie $\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} = -t_{n-1,2\alpha_c}$. Sur l'exemple, la table de la loi de Student permet de constater que $-t_{11,0.05} = -2.201 < -1.853 < -1.796 = -t_{11,0.1}$. On en déduit que $5\% < 2\alpha_c < 10\%$, d'où $2.5\% < \alpha_c < 5\%$.

Pour calculer exactement la p-valeur, on écrit :

$$-t_{n-1,2\alpha_c} = F^{-1}_{St(n-1)}\left(\frac{2\alpha_c}{2}\right) = F^{-1}_{St(n-1)}\left(\alpha_c\right) \Longrightarrow \alpha_c = F_{St(n-1)}\left(\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n}\right)$$

et on obtient ici $\alpha_c = 0.04547$.

La réponse à un test de seuil fixé est binaire : on rejette ou on ne rejette pas H_0. Fournir une p-valeur est une réponse plus riche puisqu'elle permet de connaître le résultat du test pour n'importe quel choix du seuil. C'est pourquoi le traitement des tests d'hypothèses par les logiciels de statistique consiste à fournir des p-valeurs.

En `R`, la commande permettant d'effectuer un test sur la moyenne d'une loi normale est `t.test`. L'option `alternative` permet de préciser lequel du test bilatéral et des deux tests unilatéraux on choisit. Sur l'exemple, on obtient :

```
>  medic<-c(25,28,20,32,17,24,41,28,25,30,27,24)
>  t.test(medic,alternative="less",mu=30)

      One Sample t-test

data: medic
t = -1.8526, df = 11, p-value = 0.04547
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
     -Inf 29.90056
sample estimates:
mean of x
    26.75
```

La p-valeur est ici $\alpha_c = 4.5\%$. Cela signifie que, pour tout seuil supérieur à 4.5% (c'est le cas de 5%), on rejettera H_0, donc on conclura que B est plus efficace que A,

et pour tout seuil inférieur à 4.5% (c'est le cas de 1%), on ne rejettera pas H_0, donc on conclura que B n'est pas plus efficace que A.

De manière générale, la p-valeur peut être comprise comme étant la probabilité, sous l'hypothèse nulle, que la statistique de test soit encore plus extrême que ce qui a été observé. Si cette probabilité est forte, il n'y a pas de raison de douter de la véracité de H_0. Mais si elle est faible, on peut en douter. Donc plus la p-valeur est petite, moins on prend de risque en rejetant H_0.

5.3.7 Remarques

Remarque 1 : Pour des raisons de symétrie, un test de "$m \leq m_0$" contre "$m > m_0$" aura pour région critique

$$W = \left\{ (x_1, \ldots, x_n); \sqrt{n}\, \frac{\bar{x}_n - m_0}{s'_n} > t_{n-1,2\alpha} \right\}$$

Remarque 2 : Pour le test bilatéral de H_0 : "$m = m_0$" contre H_1 : "$m \neq m_0$", le bon sens veut que l'on rejette H_0 si $\bar{x}_n$ est significativement éloigné de m_0. On prendra donc une région critique du type $W = \{(x_1, \ldots, x_n); |\bar{x}_n - m_0| > l_\alpha\}$. Alors, comme précédemment on obtient :

$$\begin{aligned} \alpha &= \underset{m=m_0}{Sup}\, P\left(\left|\bar{X}_n - m_0\right| > l_\alpha; m\right) = P\left(\left|\bar{X}_n - m_0\right| > l_\alpha; m_0\right) \\ &= P\left(\left|\sqrt{n}\, \frac{\bar{X}_n - m_0}{S'_n}\right| > \sqrt{n}\, \frac{l_\alpha}{S'_n}; m_0\right) \end{aligned}$$

On en déduit que $\sqrt{n}\, \dfrac{l_\alpha}{S'_n} = t_{n-1,\alpha}$, d'où $l_\alpha = \dfrac{S'_n}{\sqrt{n}}\, t_{n-1,\alpha}$. On obtient donc comme région critique :

$$W = \left\{ (x_1, \ldots, x_n); |\bar{x}_n - m_0| > \frac{s'_n}{\sqrt{n}}\, t_{n-1,\alpha} \right\} = \left\{ (x_1, \ldots, x_n); \left|\sqrt{n}\, \frac{\bar{x}_n - m_0}{s'_n}\right| > t_{n-1,\alpha} \right\}$$

Remarque 3 : Pour alléger les écritures, on écrit souvent une région critique en omettant l'expression $(x_1, \ldots, x_n)$;, ce qui donne par exemple $W = \left\{ \sqrt{n}\, \dfrac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha} \right\}$. Mais il faut toujours garder à l'esprit que la région critique est l'ensemble des valeurs des observations pour lesquelles on rejettera H_0.

5.3.8 Les tests de Student

Finalement, on dispose d'une procédure permettant d'effectuer le test bilatéral et les deux tests unilatéraux portant sur la moyenne de la loi normale. Ces trois tests sont connus sous le nom de **tests de Student**.

Propriété 12 *: Tests de Student sur la moyenne d'une loi normale.*

- *Test de "$m \leq m_0$" contre "$m > m_0$" :* $W = \left\{ \sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} > t_{n-1,2\alpha} \right\}$.
- *Test de "$m \geq m_0$" contre "$m < m_0$" :* $W = \left\{ \sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha} \right\}$.
- *Test de "$m = m_0$" contre "$m \neq m_0$" :* $W = \left\{ \left| \sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} \right| > t_{n-1,\alpha} \right\}$.

5.4 Lien entre tests d'hypothèses et intervalles de confiance

Dans le test bilatéral, on rejette l'hypothèse "$m = m_0$" à condition que $\left| \sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} \right| > t_{n-1,\alpha}$. Or :

$$\begin{aligned}
\left| \sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n} \right| > t_{n-1,\alpha} &\Leftrightarrow \bar{x}_n - m_0 < -\frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \text{ ou } \bar{x}_n - m_0 > +\frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \\
&\Leftrightarrow m_0 < \bar{x}_n - \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \text{ ou } m_0 > \bar{x}_n + \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \\
&\Leftrightarrow m_0 \notin \left[\bar{x}_n - \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha}, \bar{x}_n + \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \right]
\end{aligned}$$

Cet intervalle n'est autre que l'intervalle de confiance usuel pour la moyenne de la loi normale, vu au chapitre 4. On rejette donc "$m = m_0$" si m_0 n'appartient pas à cet intervalle. Il y a donc un lien étroit entre les tests d'hypothèses et les intervalles de confiance.

C'est logique : on a une confiance $1-\alpha$ dans le fait que m appartient à l'intervalle de confiance. Si m_0 n'appartient pas à cet intervalle, il est vraiment douteux que $m = m_0$. On a même une confiance $1-\alpha$ dans le fait que $m \neq m_0$. On peut donc construire un test d'hypothèses sur la valeur d'un paramètre à partir d'un intervalle de confiance pour ce paramètre.

Or, pour construire un tel intervalle, on a eu besoin d'une fonction pivotale. Par conséquent, pour construire un test paramétrique, il suffit de connaître une fonction pivotale. Dans le cas de la moyenne de la loi normale, la fonction pivotale est $\sqrt{n}\,\dfrac{\bar{X}_n - m}{S'_n}$.

La dualité entre intervalles de confiance et tests d'hypothèses fait que, en `R`, la commande `t.test` permet à la fois d'effectuer un test et d'obtenir un intervalle de confiance sur la moyenne de la loi normale.

Ainsi, la commande `t.test(x,conf.level=0.95)` effectue par défaut le test de "$m = 0$" contre "$m \neq 0$", et donne un intervalle de confiance pour m au seuil 5%. Dans l'exemple des niveaux de bruit, on obtient le résultat déjà vu au chapitre précédent :

```
>  t.test(bruit, conf.level=0.95)

        One Sample t-test

data: bruit
```

```
t = 55.7889, df = 19, p-value <  2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 61.82992 66.65008
sample estimates:
mean of x
    64.24
```

On retrouve bien que l'intervalle de confiance de seuil 5% pour m est [61.82992, 66.65008]. Etant donné que 0 n'est pas, et de loin, dans cet intervalle, l'hypothèse "$m = 0$" est très largement rejetée ce qui se traduit par une p-valeur extrêmement faible : $2.2\,10^{-16}$.

5.5 Procédure pour construire un test d'hypothèses

Finalement, le plus simple pour construire un test d'hypothèses portant sur la valeur d'un paramètre θ est de se fier à son bon sens. Si on connaît un estimateur $\hat{\theta}_n$ de θ, on procèdera de la façon suivante :

- Test de "$\theta \leq \theta_0$" contre "$\theta > \theta_0$" : on rejette H_0 si $\hat{\theta}_n$ est "trop grand". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n > l_\alpha\right\}$$
- Test de "$\theta \geq \theta_0$" contre "$\theta < \theta_0$" : on rejette H_0 si $\hat{\theta}_n$ est "trop petit". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n < l_\alpha\right\}$$
- Test de "$\theta = \theta_0$" contre "$\theta \neq \theta_0$" : on rejette H_0 si $\left|\hat{\theta}_n - \theta_0\right|$ est "trop grand" ou bien si $\hat{\theta}_n$ est "soit trop grand, soit trop petit". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n < l_{1,\alpha} \text{ ou } \hat{\theta}_n > l_{2,\alpha}\right\}, \text{ avec } l_{1,\alpha} < l_{2,\alpha}$$

Pour déterminer $l_\alpha, l_{1,\alpha}, l_{2,\alpha}$, il faut écrire $\alpha = \underset{H_0}{Sup}\, P\left((X_1, \ldots, X_n) \in W; \theta\right)$. Par exemple, dans le premier cas, $\alpha = \underset{\theta \leq \theta_0}{Sup}\, P\left(\hat{\theta}_n > l_\alpha\right)$. Pour pouvoir calculer $P\left(\hat{\theta}_n > l_\alpha\right)$, il faut utiliser une fonction pivotale.

Malheureusement, cette procédure de bon sens ne permet pas toujours de résoudre le problème. C'est le cas par exemple quand la loi de probabilité de $\hat{\theta}_n$ sous H_0 est complexe et qu'on ne peut pas trouver de fonction pivotale. D'autre part, le test obtenu par cette approche n'est pas forcément optimal, au sens où il peut en exister de plus puissants.

Il existe en fait des méthodes statistiques sophistiquées permettant de répondre à ces deux problèmes. Le résultat le plus important est le théorème de Neyman-Pearson. Mais ces procédures débordent du cadre de ce cours et ne seront pas évoquées ici.

Le principe "non-rejet $\neq$ acceptation" est à comprendre différemment pour les tests unilatéraux et bilatéraux. Pour les tests unilatéraux, la différence est franche : "B n'est pas plus efficace que A" est différent de "B est moins efficace que A". Pour les tests bilatéraux, H_0 est une hypothèse simple, donc accepter H_0 revient à choisir le modèle correspondant ($\theta = \theta_0$) pour le phénomène étudié. Or tous les modèles sont faux, car ce ne sont que des approximations de la réalité. Ne pas rejeter H_0 consiste à considérer que le modèle correspondant n'est pas absurde. Donc on peut l'adopter, ce qui revient en quelque sorte à "accepter" H_0, au sens où le modèle sous-jacent n'est pas trop mauvais.

5.6 Tests sur la variance d'une loi normale

On suppose ici que les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi normale $\mathcal{N}(m, \sigma^2)$. On souhaite tester par exemple H_0 : "$\sigma^2 \leq \sigma_0^2$" contre H_1 : "$\sigma^2 > \sigma_0^2$".

En suivant la démarche présentée ci-dessus, puisque l'ESBVM de σ^2 est ${S'_n}^2$, il est naturel de rejeter H_0 si ${S'_n}^2$ est "trop grand", donc de considérer une région critique de la forme $W = \left\{ {s'_n}^2 > l_\alpha \right\}$. Pour calculer $\alpha = \underset{H_0}{Sup}\, P\left({S'_n}^2 > l_\alpha \right)$, on utilise la fonction pivotale $\dfrac{(n-1){S'_n}^2}{\sigma^2}$, qui est de loi χ^2_{n-1}. On obtient :

$$\begin{aligned} \alpha &= \underset{\sigma^2 \leq \sigma_0^2}{Sup}\, P({S'_n}^2 > l_\alpha) = \underset{\sigma^2 \leq \sigma_0^2}{Sup}\, P\left(\frac{(n-1){S'_n}^2}{\sigma^2} > \frac{(n-1)l_\alpha}{\sigma^2} \right) \\ &= \underset{\sigma^2 \leq \sigma_0^2}{Sup} \left[1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma^2} \right) \right] = 1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma_0^2} \right) \end{aligned}$$

D'où $l_\alpha = \dfrac{\sigma_0^2}{n-1} F^{-1}_{\chi^2_{n-1}}(1-\alpha) = \dfrac{\sigma_0^2}{n-1} z_{n-1,\alpha}$, et la région critique du test est :

$$W = \left\{ {s'_n}^2 > \frac{\sigma_0^2}{n-1} z_{n-1,\alpha} \right\} = \left\{ \frac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha} \right\}$$

On aboutirait au même résultat en partant d'un intervalle de confiance de seuil α pour σ^2 du type $[a, +\infty[$.

Finalement, on obtient :

Propriété 13 *Tests sur la variance d'une loi normale :*

- *Test de "$\sigma^2 \leq \sigma_0^2$" contre "$\sigma^2 > \sigma_0^2$" :* $W = \left\{ \dfrac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha} \right\}$.
- *Test de "$\sigma^2 \geq \sigma_0^2$" contre "$\sigma^2 < \sigma_0^2$" :* $W = \left\{ \dfrac{(n-1){s'_n}^2}{\sigma_0^2} < z_{n-1,1-\alpha} \right\}$.
- *Test de "$\sigma^2 = \sigma_0^2$" contre "$\sigma^2 \neq \sigma_0^2$" :*

$$W = \left\{ \frac{(n-1){s'_n}^2}{\sigma_0^2} < z_{n-1,1-\alpha/2} \text{ ou } \frac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha/2} \right\}$$

Dans l'exemple de l'essai thérapeutique, la variance mesure la variabilité de l'effet du médicament. La variabilité est faible si l'effet du médicament est à peu près le même pour tout le monde, et elle est forte si les effets peuvent être très différents d'un individu à un autre. On a évidemment intérêt à avoir une variabilité assez faible pour bien contrôler les effets d'un traitement. Cette variabilité se traduit sur la variance de la loi normale qui modélise la durée de disparition de la douleur chez les malades traités.

Avec le médicament A, l'écart-type était $\sigma_0 = 5$ mn, ce qui signifie que, pour 95% des malades, la douleur disparaît entre $m_0 - 2\sigma_0 = 20$ mn et $m_0 + 2\sigma_0 = 40$ mn. Avec le médicament B, on estime σ par s'_n = 6.08 mn. La variabilité du second médicament est-elle significativement supérieure à celle du premier ?

C'est un test de "$\sigma \leq 5$" contre "$\sigma > 5$" , évidemment identique au test de "$\sigma^2 \leq 25$" contre "$\sigma^2 > 25$". La région critique est $W = \left\{ \dfrac{(n-1)s'^{\,2}_n}{\sigma_0^2} > z_{n-1,\alpha} \right\}$.

Au seuil α = 5%, on a $z_{11,5\%} = 19.68$. Et $\dfrac{(n-1)s'^{\,2}_n}{\sigma_0^2} = \dfrac{11 \times 6.08^2}{25} = 16.25$.

Comme $16.25 < 19.68$, on n'est pas dans la région critique, donc on ne rejette pas H_0 : on n'a pas de preuves suffisantes pour conclure que la variabilité de l'effet de B est supérieure à celle de A. La différence entre 6.08 et 5 n'est pas significative au seuil choisi.

La p-valeur est obtenue en écrivant :

$$z_{n-1,\alpha_c} = F^{-1}_{\chi^2_{n-1}}(1-\alpha_c) = 16.25 \implies \alpha_c = 1 - F_{\chi^2_{11}}(16.25) = 13.2\%.$$

Donc même au seuil 10%, on ne rejettera pas H_0.

5.7 Tests sur une proportion

On suppose ici que les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$. On sait que $T = \sum_{i=1}^{n} X_i$ est de loi binomiale $\mathcal{B}(n,p)$. On souhaite faire des tests sur la valeur de p.

Pour construire ces tests, on peut partir de l'intervalle de confiance exact pour p vu au chapitre 4. Mais compte-tenu de sa complexité, on se contentera de l'intervalle de confiance asymptotique : $\dfrac{T - np}{\sqrt{np(1-p)}}$ est approximativement de loi $\mathcal{N}(0,1)$, ce qui fournit la fonction pivotale (asymptotique) cherchée et permet de donner directement les tests sur une proportion :

Propriété 14 *Tests asymptotiques sur une proportion :*

- *Test de "$p \leq p_0$" contre "$p > p_0$" :* $W = \left\{ \dfrac{t - np_0}{\sqrt{np_0(1-p_0)}} > u_{2\alpha} \right\}$.
- *Test de "$p \geq p_0$" contre "$p < p_0$" :* $W = \left\{ \dfrac{t - np_0}{\sqrt{np_0(1-p_0)}} < -u_{2\alpha} \right\}$.

- *Test de "$p = p_0$" contre "$p \neq p_0$" :* $W = \left\{ \left| \frac{t - np_0}{\sqrt{np_0(1-p_0)}} \right| > u_\alpha \right\}$.

Dans l'exemple du sondage du chapitre 4, on a interrogé n = 800 personnes et t = 420 d'entre elles ont déclaré vouloir voter pour A. On a donc estimé le pourcentage p de voix qu'obtiendra le candidat A par $\hat{p}_n = 420/800 = 52.5\%$. Mais on a vu qu'un intervalle de confiance de seuil 5% pour ce pourcentage est [49%, 56%], dont une partie est située sous les 50%.

En fait, la seule chose qui intéresse le candidat A, c'est de savoir s'il va être élu ou pas. Il s'agit donc de faire un test dans lequel le rejet de H_0 correspond à l'élection de A. Par conséquent, on va tester "$p \leq 1/2$" contre "$p > 1/2$".

$\frac{t - np_0}{\sqrt{np_0(1-p_0)}} = \frac{420 - 800/2}{\sqrt{800/4}} = 1.414$. Au seuil 5%, $u_{0.1} = 1.645$. $1.414 < 1.645$, donc on n'est pas dans la région critique, donc on ne rejette pas H_0 : on ne peut pas affirmer que A sera élu avec moins de 5% de chances de se tromper.

La p-valeur du test asymptotique est la valeur α_c de α telle que $u_{2\alpha_c} = \phi^{-1}(1-\alpha_c) = \frac{t - np_0}{\sqrt{np_0(1-p_0)}} = 1.414$. On a donc $\alpha_c = 1 - \phi(1.414) = 7.86\%$.

En R, on peut effectuer le test exact grâce à la commande binom.test. On obtient sur l'exemple du sondage :

```
>  binom.test(420, 800, p=0.5, alternative="greater")

      Exact binomial test

data: 420 and 800
number of successes = 420, number of trials = 800,
p-value = 0.08395
alternative hypothesis: true probability of success is greater
than 0.5
95 percent confidence interval:
  0.4953009 1.0000000
sample estimates:
probability of success
         0.525
```

La p-valeur du test exact est 8.39 %, ce qui est bien cohérent avec la valeur donnée par le test asymptotique.

En conclusion, si on décide de conclure, au vu du sondage, que le candidat A sera élu, on a au pire 8.39% de chances de se tromper. Tout ce qui vient d'être dit n'est évidemment valable que si les résultats du sondage sont bien représentatifs de ce qui se passera le jour de l'élection, ce qui est loin d'être certain.

5.8 Le test du χ^2

Nous terminons ce chapitre par une présentation du plus célèbre des tests d'hypothèses, le test du χ^2.

Exemple introductif. On jette un dé 300 fois. On obtient les résultats suivants :

face obtenue	1	2	3	4	5	6
nombre de lancers	42	43	56	55	43	61

Peut-on en conclure que le dé est équilibré ?

Une idée naturelle est de dire que, si le dé est équilibré, on devrait avoir à peu près 300/6 = 50 fois chaque face. Si le résultat s'éloigne trop de 50 sur quelques unes des faces, on peut douter du fait que le dé est équilibré. Or on observe 61 fois la face 6 et 42 fois la face 1 : est-ce trop ou trop peu pour un dé équilibré ? On peut donc penser à rejeter l'hypothèse que le dé est équilibré si la "distance" entre les vecteurs $(42, 43, 56, 55, 43, 61)$ et $(50, 50, 50, 50, 50, 50)$ est "trop grande". Il reste à choisir une distance appropriée.

Plus généralement, on s'intéresse à une expérience qui a k issues possibles. On sait que, sous une certaine hypothèse H_0, les probabilités d'apparition de ces k issues sont respectivement $p_1, \ldots, p_k$ (avec $\sum_{j=1}^{k} p_j = 1$). On fait n expériences identiques et indépendantes et on compte les nombres n_j de fois où l'issue j s'est produite. On a forcément $\sum_{j=1}^{k} n_j = n$. Le problème est de décider si l'observation de $n_1, \ldots, n_k$ est compatible avec l'hypothèse H_0 que les probabilités des issues sont $p_1, \ldots, p_k$.

Dans l'exemple, $k = 6$, $\forall j \in \{1, \ldots, 6\}, p_j = \dfrac{1}{6}$ et $n = 300$.

Sous H_0, on s'attend à observer en moyenne np_j fois l'issue j (50 fois chaque face dans l'exemple). Il s'agit donc de déterminer si les n_j sont significativement proches ou éloignés des np_j. On peut alors penser à une région critique de la forme :

$$W = \left\{ \sum_{j=1}^{k} (n_j - np_j)^2 > l_\alpha \right\}$$

Pour déterminer l_α, il faut connaitre la loi de probabilité sous H_0 de $\sum_{j=1}^{k} (N_j - np_j)^2$, ou d'une variable aléatoire analogue.

Il est clair que, pour tout j, N_j est de loi binomiale $\mathcal{B}(n, p_j)$, mais les N_j ne sont pas indépendantes. En effet, puisque $\sum_{j=1}^{k} N_j = n$, si on connait $N_1, \ldots, N_{k-1}$, on connait N_k avec certitude.

Pour tout k-uplet d'entiers $(n_1, \ldots, n_k)$ tels que $\sum_{j=1}^{k} n_j = n$, on a :

$$\begin{aligned} P(N_1 = n_1, \ldots, N_k = n_k) &= P(\text{ sur les } n \text{ expériences, on a eu } n_1 \text{ fois l'issue 1, ..., } \\ &\quad n_k \text{ fois l'issue } k) \\ &= \binom{n}{n_1} p_1^{n_1} \binom{n-n_1}{n_2} p_2^{n_2} \ldots \binom{n-n_1-\ldots-n_{k-1}}{n_k} p_k^{n_k} \\ &= \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} \end{aligned}$$

On dit que le vecteur $(N_1, \ldots, N_k)$ est de loi **multinomiale** $\mathcal{M}(n; p_1, \ldots, p_k)$. Le test du χ^2 est basé sur le théorème suivant :

Propriété 15 . **Théorème de Pearson** *: Si $(N_1, \ldots, N_k)$ est de loi $\mathcal{M}(n; p_1, \ldots, p_k)$, alors :*

$$\Delta_n^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \xrightarrow{\mathcal{L}} \chi^2_{k-1}$$

Par conséquent, on peut construire un test consistant à rejeter l'hypothèse que les vraies probabilités des issues sont $p_1, \ldots, p_k$ si Δ_n^2 est "trop grand".

Définition 15 *: On appelle* **test du khi-deux** *le test de H_0 : "les probabilités des k issues sont $p_1, \ldots, p_k$" contre $H_1 = \bar{H}_0$ défini par la région critique :*

$$W = \left\{ \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} > z_{k-1,\alpha} \right\}$$

On constate que la région critique n'a pas tout à fait la forme proposée, mais elle s'en rapproche.

Dans l'exemple du dé, l'hypothèse que le dé est équilibré s'écrit H_0 : "$\forall j, p_j = \frac{1}{6}$". Alors, la statistique de test vaut $\delta_n^2 = \frac{(42-50)^2}{50} + \ldots + \frac{(61-50)^2}{50} = 6.88$.

Au seuil $\alpha = 5\%$, $z_{5,0.05} = 11.07$. 6.88 est nettement inférieur à 11.07, donc on ne rejette pas H_0 : rien n'indique que le dé n'est pas équilibré.

En `R`, la commande pour le test du χ^2 est `chisq.test` :

```
>  de<-c(42,43,56,55,43,61)
>  chisq.test(de,p=rep(1/6,6))

     Chi-squared test for given probabilities

data: de
X-squared = 6.88, df = 5, p-value = 0.2297
```

La p-valeur est de 23%, ce qui est assez élevé : même en s'autorisant 20% de chances de se tromper, on ne rejetterait pas H_0. Par conséquent, pour un dé équilibré, il n'est

pas du tout improbable d'observer 61 fois la face 6, même si on s'attend à ne l'observer en moyenne que 50 fois.

Remarque : Le résultat du théorème de Pearson n'est qu'asymptotique, donc il n'est valable que pour n suffisamment grand. En pratique, on considère que l'on peut effectuer le test si pour tout j, $n_j \geq 5$, ce qui est le cas dans l'exemple.

Chapitre 6

La régression linéaire

6.1 Introduction

Jusqu'à maintenant, on a considéré dans ce cours que les observations étaient unidimensionnelles, c'est-à-dire que les variables aléatoires étudiées étaient à valeurs dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. Cela signifie que l'on ne mesure qu'une seule variable sur chaque individu. Or il existe de nombreuses situations où il est intéressant de mesurer plusieurs variables sur chaque individu. Dans ce cas, les observations sont des réalisations de vecteurs aléatoires, dont les composantes sont dépendantes. Il faut alors faire une analyse statistique multidimensionnelle.

On ne traitera ici que le cas de données bidimensionnelles. Cela signifie que l'on observe les réalisations de couples aléatoires $(X_1, Y_1), \dots (X_n, Y_n)$. On supposera que ces couples sont indépendants et de même loi. Le problème est alors l'étude de cette loi, ce qui nécessite de se poser la question de la **dépendance** entre X_i et Y_i.

Un problème de **régression** consiste à chercher une fonction f telle que pour tout i, Y_i soit approximativement égale à $f(X_i)$. Le cas le plus simple est celui de la **régression linéaire simple**, où on cherche f de la forme $f(x) = \beta_1 x + \beta_0$. Pour estimer β_1 et β_0, on utilise la fameuse **méthode des moindres carrés**.

Exemple : Pour tester la performance du système de freinage d'une voiture, on la fait rouler jusqu'à atteindre une vitesse x (en mètres par seconde), à laquelle on freine. On mesure alors la distance de freinage y (en mètres). On fait l'expérience pour n vitesses différentes $x_1, \dots, x_n$ et on mesure les n distances de freinage correspondantes $y_1, \dots, y_n$. Toutes les mesures sont faites avec le même véhicule et le même pilote, et sont supposées indépendantes. On obtient le tableau 6.1 :

numéro de mesure i	1	2	3	4	5	6	7	8
vitesse x_i	5	10	15	20	25	30	35	40
distance de freinage y_i	3.42	5.96	31.14	41.76	74.54	94.92	133.78	169.16

TABLE 6.1 – Vitesse et distance de freinage d'une voiture

Quel modèle de dépendance entre la distance de freinage et la vitesse peut-on proposer ? Une dépendance affine est-elle raisonnable ? Peut-on estimer la distance de freinage d'une voiture lancée à 50 m/s ? Avec quelle précision ?

6.2 Le modèle de régression linéaire simple

On dispose de données bidimensionnelles qui sont n couples (x_i, y_i). y_i est supposée être la réalisation d'une variable aléatoire Y_i. A priori, x_i pourrait être aussi la réalisation d'une variable aléatoire X_i. En fait, on peut montrer que les méthodes développées pour le cas où les x_i sont supposés constants suffisent à traiter le cas où les x_i sont des réalisations de variables aléatoires. Aussi, dans tout ce chapitre, nous supposerons que les x_i sont des constantes connues.

Faire une **régression de** y **sur** x, c'est étudier le type de dépendance entre y et x. y peut dépendre de x, mais aussi de quantité d'autres facteurs qu'on ne connaît pas forcément. L'hypothèse du modèle de régression est que les effets de x et des autres facteurs s'ajoutent.

Définition 16 *: Le* **modèle de régression** *de Y sur x est défini par :*

$$Y = f(x) + \varepsilon$$

où
- *Y est la* **variable à expliquer** *ou variable expliquée.*
- *x est la* **variable explicative** *ou prédicteur ou régresseur.*
- *ε est l'erreur de prévision de Y par $f(x)$ ou* **résidu**.

Dans l'exemple, il est clair que la distance de freinage y dépend de la vitesse x, mais pas seulement : l'état de la route, la météo, la nervosité du conducteur, peuvent influer sur la distance de freinage. La vitesse à laquelle on freine peut être contrôlée par le conducteur, donc on peut supposer que les x_i sont des constantes connues. En revanche, même quand on connait x, la distance de freinage n'est pas prévisible avec certitude à l'avance. Donc il est logique de supposer que les y_i sont des réalisations de variables aléatoires Y_i. Il faut donc exprimer le fait que les variables aléatoires Y_i dépendent des x_i et d'un certain nombre d'autres facteurs imprévisibles et non mesurés, dont on suppose que l'effet s'ajoute à celui des x_i. Comme les données consistent en plusieurs observations de Y, obtenues pour différentes valeurs de x, le modèle s'écrit :

$$\forall i \in \{1, \ldots, n\},\ Y_i = f(x_i) + \varepsilon_i$$

Il faut alors préciser la forme de f et la loi de probabilité du vecteur des résidus ε_i. Le modèle de régression linéaire simple est obtenu en supposant que f est une fonction affine de x et que les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2. Cela revient à supposer que les n mesures sont indépendantes, effectuées dans les mêmes conditions (même loi) et que les effets des facteurs autres que le prédicteur s'équilibrent ($E[\varepsilon_i] = 0$).

Définition 17 *: Le* **modèle de régression linéaire simple** *ou* **modèle linéaire simple** *est défini par :*

$$\forall i \in \{1, \ldots, n\},\ Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

où β_0 et β_1 sont des paramètres réels inconnus, et les résidus ε_i sont indépendants, de même loi, centrés et de variance σ^2.

Dans un premier temps, on ne donne pas plus de précisions sur la loi des résidus. On verra que l'on peut estimer β_0, β_1 et σ^2 sans connaître cette loi. Usuellement, on suppose que les résidus sont de loi normale. Le modèle est alors un **modèle linéaire gaussien**.

L'espérance et la variance de Y_i sont :

$$\begin{aligned} E[Y_i] &= E\Big[\beta_1 x_i + \beta_0 + \varepsilon_i\Big] = \beta_1 x_i + \beta_0 \\ Var[Y_i] &= Var\Big[\beta_1 x_i + \beta_0 + \varepsilon_i\Big] = Var[\varepsilon_i] = \sigma^2 \end{aligned}$$

σ^2 mesure le bruit ou le poids des facteurs autres que x. Plus σ^2 est élevé, plus Y_i fluctue autour de $\beta_1 x_i + \beta_0$. Inversement, pour $\sigma^2 = 0$, les points (x_i, Y_i) sont parfaitement alignés : Y_i n'est plus aléatoire.

Les problèmes statistiques qui se posent sont :

- L'*estimation* de β_1, β_0 et σ^2, ponctuelle et par intervalle de confiance.
- La construction de *tests d'hypothèses* portant sur β_1, β_0 et σ^2.
- La *prévision* de y connaissant x.
- La *validation* du modèle : la liaison entre la vitesse et la distance de freinage est-elle bien affine ?

Remarque fondamentale : Dans un modèle de régression linéaire simple, la liaison entre x et y n'est pas linéaire, mais affine. Par abus de langage, on dit souvent que y dépend linéairement de x. Mais en fait, ce qui est important, c'est que y dépende linéairement du couple (β_1, β_0). Ainsi, contrairement aux apparences, les modèles suivants sont bien des modèles linéaires :

1. $Y_i = \beta_1 \ln x_i + \beta_0 + \varepsilon_i$
2. $Y_i = \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i$
3. Modèle de Cobb-Douglas : $Y_i = \beta_0 x_i^{\beta_1} \varepsilon_i$ (car $\ln Y_i = \beta_1 \ln x_i + \ln \beta_0 + \ln \varepsilon_i$)
4. Modèle logistique : $Y_i = \dfrac{e^{(\beta_1 x_i + \beta_0 + \varepsilon_i)}}{1 + e^{(\beta_1 x_i + \beta_0 + \varepsilon_i)}}$ (car $\ln \dfrac{Y_i}{1 - Y_i} = \beta_1 x_i + \beta_0 + \varepsilon_i$)

En revanche, le modèle $Y_i = \beta_1 + e^{\beta_0 x_i} + \varepsilon_i$ n'est pas linéaire.

6.3 Estimation par la méthode des moindres carrés

La première chose à faire est de dessiner le nuage des points (x_i, y_i), $\forall i \in \{1, \ldots, n\}$, pour déterminer le type de liaison pouvant exister entre x et y. A priori, n'importe quel type de liaison est possible, par exemple celles présentées dans la figure 6.1.

En `R`, la fonction `plot` permet de dessiner ce nuage. Sur l'exemple, on obtient la figure 6.2 :

```
>  vitesse<-c(5,10,15,20,25,30,35,40)
```

```
> freinage<-c(3.42,5.96,31.14,41.76,74.54,94.92,133.78,169.16)
> plot(vitesse,freinage)
```

FIGURE 6.1 – Exemple de différentes liaisons possibles entre x et y

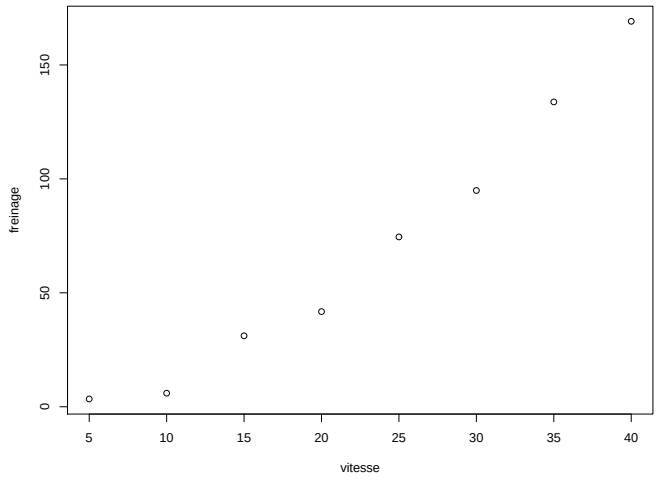


FIGURE 6.2 – Vitesse et distance de freinage : nuage de points

A première vue, l'hypothèse de liaison affine peut être retenue pour ces données. En fait, on verra plus tard qu'il existe des méthodes statistiques permettant de juger de la pertinence de cette hypothèse plus précisément que par une simple impression visuelle.

Le problème est maintenant de trouver la droite "la plus proche" de ce nuage, en un certain sens. La méthode la plus couramment utilisée est la **méthode des moindres carrés**, qui consiste à prendre la droite pour laquelle la somme des carrés des distances verticales des points à la droite est minimale (voir figure 6.3). Remarquons que ce n'est pas la distance euclidienne usuelle qui est utilisée.

La somme des carrés des distances verticales des points du nuage à la droite d'équation $y = \beta_1 x + \beta_0$ est $\sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$. Il faut trouver les valeurs de β_1 et β_0 qui minimisent cette quantité.

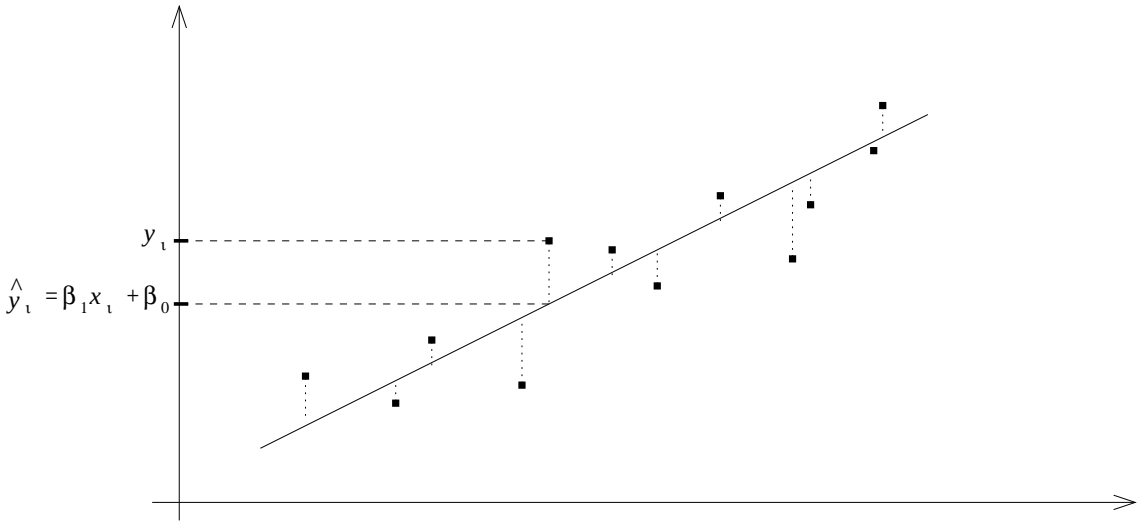


FIGURE 6.3 – La droite des moindres carrés

Il revient au même de minimiser la quantité $\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$, appelée **erreur quadratique moyenne**. C'est en effet une mesure de l'erreur moyenne que l'on commet en estimant y_i par $\hat{y}_i = \beta_1 x_i + \beta_0$: $\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. C'est ce qui justifie le nom de "moindres carrés" pour cette méthode.

Pour minimiser δ^2, il suffit d'annuler les dérivées partielles de δ^2 par rapport à β_1 et β_0 :

$$\begin{aligned}
\frac{\partial \delta^2}{\partial \beta_1} &= -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - \beta_1 x_i - \beta_0) = -2\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \beta_0 \frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
\frac{\partial \delta^2}{\partial \beta_0} &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0) = -2\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i - \beta_0\right]
\end{aligned}$$

On pose :

$$\begin{aligned}
\bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i && \text{moyenne empirique des } x_i \\
\bar{y}_n &= \frac{1}{n}\sum_{i=1}^{n} y_i && \text{moyenne empirique des } y_i \\
s_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2 && \text{variance empirique des } x_i \\
s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}_n^2 && \text{variance empirique des } y_i \\
c_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) && \text{covariance empirique entre les } x_i \text{ et les } y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n \\
r_{xy} &= \frac{c_{xy}}{s_x s_y} && \text{coefficient de corrélation linéaire} \\
& && \text{empirique entre les } x_i \text{ et les } y_i
\end{aligned}$$

$\bar{x}_n$ et s_x^2 sont des constantes. $\bar{y}_n$, s_y^2, c_{xy} et r_{xy} sont des réalisations de variables aléatoires, que l'on note respectivement $\bar{Y}_n$, S_Y^2, C_{xY} et R_{xY}.

r_{xy} est la version empirique du coefficient de corrélation linéaire $\rho(X, Y)$ et possède des propriétés équivalentes :

- $r_{xy} \in [-1, 1]$
- $r_{xy} = +1 \Longleftrightarrow$ les points (x_i, y_i) sont alignés sur une droite de pente positive.
- $r_{xy} = -1 \Longleftrightarrow$ les points (x_i, y_i) sont alignés sur une droite de pente négative.
- si y ne dépend pas de x, r_{xy} doit être proche de 0. Réciproquement, si r_{xy} est proche de 0, alors il n'y a pas de dépendance affine entre x et y, mais il est possible qu'il existe une dépendance non affine.

En `R`, r_{xy} est donné par `cor(x,y)`. La commande `cov(x,y)` ne donne pas c_{xy} mais $\frac{n}{n-1} c_{xy}$, pour les mêmes raisons qui font que `var(x)` donne s'^2_x au lieu de s_x^2.

Revenant à la minimisation de δ^2, on obtient :

$$\frac{\partial \delta^2}{\partial \beta_0} = 0 \Longrightarrow \bar{y}_n - \beta_1 \bar{x}_n - \beta_0 = 0$$

Par conséquent, la droite des moindres carrés passe par le barycentre du nuage, c'est à dire le point de coordonnées $(\bar{x}_n, \bar{y}_n)$.

En prenant en compte $\beta_0 = \bar{y}_n - \beta_1 \bar{x}_n$, on obtient :

$$\frac{\partial \delta^2}{\partial \beta_1} = 0 \Longrightarrow c_{xy} + \bar{x}_n \bar{y}_n - \beta_1 (s_x^2 + \bar{x}_n^2) - (\bar{y}_n - \beta_1 \bar{x}_n)\bar{x}_n = 0$$

D'où $c_{xy} - \beta_1 s_x^2 = 0$, et donc finalement :

Propriété 16 *Les estimateurs des moindres carrés de β_1 et β_0 sont :*

$$\widehat{\beta}_1 = \frac{C_{xY}}{s_x^2} \qquad \textit{et} \qquad \widehat{\beta}_0 = \bar{Y}_n - \frac{C_{xY}}{s_x^2} \bar{x}_n$$

L'**équation de la droite des moindres carrés** est alors :

$$y = \widehat{\beta}_1 x + \widehat{\beta}_0 = \bar{y}_n + \frac{c_{xy}}{s_x^2}(x - \bar{x}_n)$$

C'est ce résultat qu'on utilise pour estimer graphiquement des paramètres à partir d'un graphe de probabilités.

L'erreur quadratique moyenne minimale est :

$$\begin{aligned} \delta^2_{min} &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0\right)^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \bar{y}_n - \frac{c_{xy}}{s_x^2}(x_i - \bar{x}_n)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 - 2\frac{c_{xy}}{s_x^2}\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n) + \frac{{c_{xy}}^2}{s_x^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \end{aligned}$$

$$
\begin{aligned}
&= s_y^2 - 2\frac{c_{xy}}{s_x^2}c_{xy} + \frac{{c_{xy}}^2}{s_x^4}s_x^2 = s_y^2 - \frac{{c_{xy}}^2}{s_x^2} = s_y^2 - \frac{r_{xy}^2 s_x^2 s_y^2}{s_x^2} \\
&= s_y^2(1 - r_{xy}^2)
\end{aligned}
$$

On en déduit en particulier que $r_{xy} \in [-1, +1]$ et que $\delta_{min}^2 = 0$ si et seulement si $r_{xy} = \pm 1$, c'est-à-dire si et seulement si les points du nuage sont alignés. C'est logique.

Il reste à montrer maintenant que la droite des moindres carrés fournit une réponse à notre problème, c'est à dire que $\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des estimateurs de β_1 et β_0 de bonne qualité dans le modèle linéaire simple. En fait, il s'avère que ce sont les meilleurs estimateurs linéaires possibles.

Propriété 17 *$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des estimateurs sans biais de β_1 et β_0.*

Démonstration

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{C_{xY}}{s_x^2} = \frac{1}{s_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n) \\
&= \frac{1}{ns_x^2}\left[\sum_{i=1}^{n}(x_i - \bar{x}_n)Y_i - \bar{Y}_n \underbrace{\sum_{i=1}^{n}(x_i - \bar{x}_n)}_{0}\right] = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)Y_i
\end{aligned}
$$

D'où

$$
\begin{aligned}
E[\widehat{\beta}_1] &= \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)E[Y_i] = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)(\beta_1 x_i + \beta_0) \\
&= \frac{\beta_1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)x_i + \frac{\beta_0}{ns_x^2}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x}_n)}_{0} \\
&= \frac{\beta_1}{s_x^2}\left[\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}_n^2\right] = \beta_1
\end{aligned}
$$

Donc $\widehat{\beta}_1$ est un estimateur sans biais de β_1.

$$
\begin{aligned}
E[\widehat{\beta}_0] &= E[\bar{Y}_n - \widehat{\beta}_1\bar{x}_n] = \frac{1}{n}\sum_{i=1}^{n}E[Y_i] - \beta_1\bar{x}_n \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_1 x_i + \beta_0) - \beta_1\bar{x}_n = \beta_1\bar{x}_n + \beta_0 - \beta_1\bar{x}_n = \beta_0
\end{aligned}
$$

Donc $\widehat{\beta}_0$ est un estimateur sans biais de β_0. ■

De même, on montre que :

Propriété 18

$$
Var[\widehat{\beta}_1] = \frac{\sigma^2}{ns_x^2}, \qquad Var[\widehat{\beta}_0] = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}_n^2}{s_x^2}\right)
$$

$$
Cov(\widehat{\beta}_1, \widehat{\beta}_0) = -\frac{\sigma^2\bar{x}_n}{ns_x^2}, \qquad Cov(\widehat{\beta}_1, \bar{Y}_n) = 0
$$

On en déduit en particulier que $\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des estimateurs convergents en moyenne quadratique et asymptotiquement non corrélés.

La principale propriété des estimateurs des moindres carrés est que, non seulement ils sont sans biais, mais, de plus, ils sont optimaux au sens suivant :

Propriété 19 *:* **Théorème de Gauss-Markov.** *$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont les estimateurs de β_1 et β_0 sans biais et de variance minimale (ESBVM) parmi tous les estimateurs sans biais linéaires (qui s'écrivent comme des combinaisons linéaires des Y_i) .*

Nous avons estimé β_1 et β_0, il reste maintenant à estimer la variance des résidus σ^2. Puisque, pour tout i, $\sigma^2 = Var[\varepsilon_i] = Var[Y_i - \beta_1 x_i - \beta_0]$, il est naturel d'estimer σ^2 par la variance empirique des $Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$.

Définition 18

- *Les variables aléatoires $E_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$ sont appelées les* **résidus empiriques**.
- *La variance empirique des résidus empiriques est notée $S^2_{Y|x}$ et est appelée* **variance résiduelle**.

On a :

$$\begin{aligned} S^2_{Y|x} &= \frac{1}{n}\sum_{i=1}^{n} E_i^2 - \bar{E}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0\right)^2 - \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)\right]^2 \\ &= \delta^2_{min} - [\underbrace{\bar{Y}_n - \widehat{\beta}_1 \bar{x}_n - \widehat{\beta}_0}_{0}]^2 \text{ (la moyenne des résidus empiriques est nulle)} \\ &= \delta^2_{min} = S^2_Y(1 - R^2_{xY}) \end{aligned}$$

Donc la variance résiduelle n'est rien d'autre que l'erreur quadratique moyenne minimale.

Dans le cas d'un échantillon, la variance empirique est un estimateur biaisé de la variance de l'échantillon. Pour la débiaiser, on la multiplie par $\frac{n}{n-1}$. Ici, on a deux échantillons. On peut montrer qu'alors la variance résiduelle est un estimateur biaisé de σ^2, et que, pour la débiaiser, il faut la multiplier par $\frac{n}{n-2}$. D'où finalement :

Propriété 20 $\widehat{\sigma}^2 = \frac{n}{n-2} S^2_{Y|x} = \frac{n}{n-2} S^2_Y(1 - R^2_{xY}) = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2$ *est un estimateur sans biais de σ^2.*

On n'a pas de résultat particulier sur la variance de cet estimateur dans le cas général.

Remarque : Il est important de noter que toutes les propriétés énoncées dans cette section sont valables *quelle que soit la loi des résidus ε_i*. Quand on rajoute une hypothèse

sur cette loi, on peut donner des précisions sur la loi des estimateurs, leur qualité, construire des intervalles de confiance et des tests d'hypothèses sur les paramètres du modèle. C'est ce qu'on fera dans la section suivante en supposant les résidus de loi normale.

Revenons maintenant à l'exemple sur la liaison entre vitesse et distance de freinage. Les indicateurs statistiques sont :

$$\bar{x}_n = 22.5, \quad \bar{y}_n = 69.33, \quad s_x^2 = 131.25, \quad s_y^2 = 3172.54, \quad c_{xy} = 632.31, \quad r_{xy} = 0.9799.$$

Le fait que r_{xy} soit très proche de 1 indique une forte corrélation linéaire positive, ce qui se voit clairement sur le nuage de points.

Les estimations des paramètres du modèle de régression linéaire simple sont données par :

$$\widehat{\beta}_1 = \frac{c_{xy}}{s_x^2} = 4.82, \quad \widehat{\beta}_0 = \bar{y}_n - \frac{c_{xy}}{s_x^2}\bar{x}_n = -39.06, \quad \widehat{\sigma}^2 = \frac{n}{n-2}s_y^2(1 - r_{xy}^2) = 168.4.$$

La droite des moindres carrés a donc pour équation $y = 4.82x - 39.06$. On peut la superposer au nuage des points grâce à la commande `R abline` et on obtient la figure 6.4.

```
> beta1chapeau<-cov(vitesse,freinage)/var(vitesse)
> beta1chapeau
[1] 4.817619
> beta0chapeau<-mean(freinage)-beta1chapeau*mean(vitesse)
> beta0chapeau
[1] -39.06143
> sigma2chapeau<-7/6*var(freinage)*(1-cor(vitesse,freinage)^2)
> sigma2chapeau
[1] 168.3939
> plot(vitesse,freinage)
> abline(beta0chapeau,beta1chapeau)
```

On peut alors facilement prévoir la distance de freinage d'une voiture lancée à 50 m/s : $4.82 * 50 - 39.06 = 201.9$ m.

6.4 Le modèle linéaire simple gaussien

6.4.1 Définition du modèle et estimation des paramètres

On suppose maintenant que la loi des résidus est une loi normale. Cette hypothèse est souvent naturelle, par exemple quand les résidus représentent du bruit ou une erreur de mesure.

Définition 19 *Le modèle de régression linéaire simple gaussien est défini par :*

$$\forall i \in \{1, \ldots, n\}, \ Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

où les ε_i sont des variables aléatoires indépendantes et de même loi normale centrée de variance σ^2, $\mathcal{N}(0, \sigma^2)$.

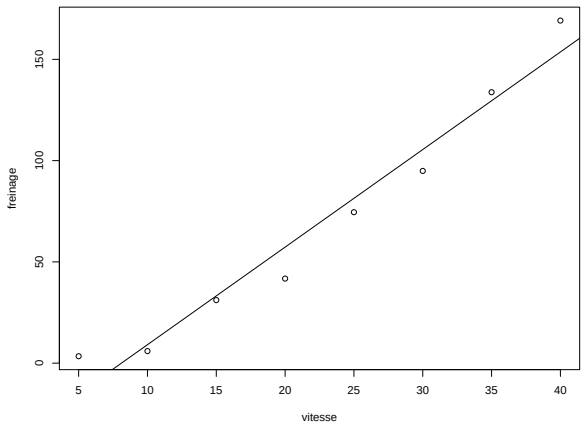


FIGURE 6.4 – Vitesse et distance de freinage : droite des moindres carrés

Alors les Y_i sont indépendantes et de lois de probabilité respectives $\mathcal{N}(\beta_1 x_i + \beta_0, \sigma^2)$. Cela permet d'en déduire les lois de probabilité des estimateurs des paramètres :

Propriété 21

- $\widehat{\beta}_1$ *est de loi* $\mathcal{N}\left(\beta_1, \dfrac{\sigma^2}{n s_x^2}\right)$.
- $\widehat{\beta}_0$ *est de loi* $\mathcal{N}\left(\beta_0, \dfrac{\sigma^2}{n}(1 + \dfrac{\overline{x}_n^2}{s_x^2})\right)$.
- $\dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0\right)^2$ *est de loi* $\mathcal{X}^2_{n-2}$.
- $\widehat{\sigma}^2$ *est indépendant de* $\bar{Y}_n, \widehat{\beta}_1$ *et* $\widehat{\beta}_0$.
- $\widehat{\beta}_1$, $\widehat{\beta}_0$ *et* $\widehat{\sigma}^2$ *sont les ESBVM de* β_1, β_0 *et* σ^2.

Les résultats sur $\widehat{\beta}_1$ et $\widehat{\beta}_0$ se démontrent facilement en utilisant le fait que toute combinaison linéaire de variables aléatoires indépendantes et de lois normales (les Y_i) est une variable aléatoire de loi normale. Les résultats sur $\widehat{\sigma}^2$ sont plus complexes à démontrer et peuvent se comprendre comme une généralisation du théorème de Fisher.

Enfin, on savait déjà que $\widehat{\beta}_1$, $\widehat{\beta}_0$ et $\widehat{\sigma}^2$ étaient des estimateurs sans biais et convergents de β_1, β_0 et σ^2. Le dernier résultat de cette propriété dit que, dans le cas gaussien, ces estimateurs sont en plus optimaux. Intuitivement, cela signifie que, dans le modèle linéaire gaussien, les meilleurs estimateurs linéaires sont les meilleurs estimateurs tout court.

6.4.2 Maximum de vraisemblance

Propriété 22 *Les estimateurs de maximum de vraisemblance de* β_1, β_0 *et* σ^2 *sont* $\widehat{\beta}_1$, $\widehat{\beta}_0$ *et* $\dfrac{n-2}{n}\widehat{\sigma}^2$.

Démonstration : La fonction de vraisemblance associée à l'observation $y_1 \dots y_n$ est :

$$\mathcal{L}(\beta_1, \beta_0, \sigma^2; y_1 \dots y_n) = f_{(Y_1,\dots,Y_n)}(y_1 \dots y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(y_i - \beta_1 x_i - \beta_0)^2}{2\sigma^2}} = \frac{1}{\sigma^n(\sqrt{2\pi})^n}\, e^{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2}$$

D'où $\ln \mathcal{L}(\beta_1, \beta_0, \sigma^2; y_1, \dots, y_n) = -\dfrac{n}{2}\ln\sigma^2 - n\ln\sqrt{2\pi} - \dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2.$

Les valeurs de β_1 et β_0 qui maximisent la log-vraisemblance minimisent

$$\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2 = n\delta^2$$

Par conséquent, les estimateurs de maximum de vraisemblance de β_1 et β_0 sont les estimateurs des moindres carrés $\widehat{\beta}_1$ et $\widehat{\beta}_0$.

Quant à σ^2, on écrit :

$$\frac{\partial}{\partial\sigma^2}\ln\mathcal{L}(\beta_1, \beta_0, \sigma^2; y_1, \dots, y_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

qui s'annule pour $\sigma^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$.

Donc l'estimateur de maximum de vraisemblance de σ^2 est $\dfrac{1}{n}\sum\limits_{i=1}^{n}(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2 = \delta^2_{min} = S^2_{Y|x} = \dfrac{n-2}{n}\widehat{\sigma}^2$. ■

6.4.3 Intervalles de confiance et tests d'hypothèses

On sait estimer ponctuellement β_1, β_0 et σ^2. On souhaite maintenant en donner des intervalles de confiance et effectuer des tests d'hypothèses sur la valeur de ces paramètres.

On sait que $U = \dfrac{\widehat{\beta}_1 - \beta_1}{\sigma} s_x\sqrt{n}$ est de loi $\mathcal{N}(0,1)$, que $Z = \dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2}$ est de loi $\mathcal{X}^2_{n-2}$, et que ces deux variables aléatoires sont indépendantes. Alors la définition de la loi de Student permet d'en déduire que :

$$\sqrt{n-2}\frac{U}{\sqrt{Z}} = \sqrt{n-2}\frac{\dfrac{\widehat{\beta}_1 - \beta_1}{\sigma}s_x\sqrt{n}}{\sqrt{\dfrac{(n-2)\widehat{\sigma}^2}{\sigma^2}}} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}}s_x\sqrt{n} \text{ est de loi } St(n-2).$$

De la même façon, on obtient :

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}}\frac{s_x\sqrt{n}}{\sqrt{s_x^2 + \overline{x}_n^2}} \text{ est de loi } St(n-2).$$

On en déduit les intervalles de confiance suivants :

Propriété 23

- *Un intervalle de confiance de seuil α pour β_1 est :*

$$\left[\widehat{\beta}_1 - \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}}, \widehat{\beta}_1 + \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}}\right]$$

- *Un intervalle de confiance de seuil α pour β_0 est :*

$$\left[\widehat{\beta}_0 - \frac{t_{n-2,\alpha}\widehat{\sigma}\sqrt{s_x^2+\bar{x}_n^2}}{s_x\sqrt{n}}, \widehat{\beta}_0 + \frac{t_{n-2,\alpha}\widehat{\sigma}\sqrt{s_x^2+\bar{x}_n^2}}{s_x\sqrt{n}}\right]$$

- *Un intervalle de confiance de seuil α pour σ^2 est :*

$$\left[\frac{(n-2)\widehat{\sigma}^2}{z_{n-2,\frac{\alpha}{2}}}, \frac{(n-2)\widehat{\sigma}^2}{z_{n-2,1-\frac{\alpha}{2}}}\right]$$

Dans l'exemple, choisissons pour seuil $\alpha = 10\%$. On a $t_{6,0.1} = 1.943$, $z_{6,0.05} = 12.59$ et $z_{6,0.95} = 1.64$. On obtient donc :

$$IC(\beta_1) = [4.04, 5.60],\ IC(\beta_0) = [-58.71, -19.41],\ IC(\sigma^2) = [80.2, 617.8]$$

Les intervalles de confiance pour β_0 et σ^2 sont larges, ce qui traduit le fait que ces paramètres sont plutôt mal estimés, essentiellement à cause du faible nombre de données. En revanche, β_1 semble assez bien estimé.

La dualité entre intervalles de confiance et tests d'hypothèses fait que l'on peut très facilement construire des tests bilatéraux sur les paramètres à partir des intervalles précédents. Par exemple, pour tester H_0 : "$\beta_1 = b$" contre "$\beta_1 \neq b$", on rejettera H_0 au seuil α si et seulement si b n'est pas dans l'intervalle de confiance de même seuil pour β_1. On obtient :

$$\begin{aligned} b \notin \left[\widehat{\beta}_1 - \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}}, \widehat{\beta}_1 + \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}}\right] &\iff b < \widehat{\beta}_1 - \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}} \text{ ou } b > \widehat{\beta}_1 + \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}} \\ &\iff \widehat{\beta}_1 - b < -\frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}} \text{ ou } \widehat{\beta}_1 - b > \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}} \\ &\iff |\widehat{\beta}_1 - b| > \frac{\widehat{\sigma} t_{n-2,\alpha}}{s_x\sqrt{n}} \end{aligned}$$

On rejette "$\beta_1 = b$" ssi $\widehat{\beta}_1$ est "trop éloigné" de b. On en déduit la région critique du test. De la même manière, on obtient les régions critiques des tests similaires sur β_0 et σ^2 :

Propriété 24 *Tests d'hypothèses bilatéraux sur β_1, β_0 et σ^2.*

- *Test de seuil α de "$\beta_1 = b$" contre "$\beta_1 \neq b$" :*

$$W = \left\{ \left| \frac{\widehat{\beta}_1 - b}{\widehat{\sigma}} \right| s_x \sqrt{n} > t_{n-2,\alpha} \right\}$$

- *Test de seuil α de "$\beta_0 = b$" contre "$\beta_0 \neq b$" :*

$$W = \left\{ \left| \frac{\widehat{\beta}_0 - b}{\widehat{\sigma}} \right| \frac{s_x \sqrt{n}}{\sqrt{s_x^2 + \bar{x}_n^2}} > t_{n-2,\alpha} \right\}$$

- *Test de seuil α de "$\sigma = \sigma_0$" contre "$\sigma \neq \sigma_0$" :*

$$W = \left\{ \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} < z_{n-2,1-\frac{\alpha}{2}} \quad \textit{ou} \quad \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} > z_{n-2,\frac{\alpha}{2}} \right\}$$

De la même façon, on peut construire des tests unilatéraux. Par exemple, on rejettera H_0 : "$\beta_1 \leq b$" au profit de "$\beta_1 > b$" ssi $\widehat{\beta}_1 - b$ est significativement grand, ce qui donne pour région critique $W = \left\{ \frac{\widehat{\beta}_1 - b}{\widehat{\sigma}} s_x \sqrt{n} > t_{n-2,2\alpha} \right\}$. De même, un test de "$\beta_1 \geq b$" contre "$\beta_1 < b$" aura pour région critique $W = \left\{ \frac{\widehat{\beta}_1 - b}{\widehat{\sigma}} s_x \sqrt{n} < -t_{n-2,2\alpha} \right\}$.

Parmi les autres hypothèses intéressantes à tester figure évidemment celle qui fonde le modèle : y a-t-il vraiment une dépendance affine entre x et y ? Jusqu'à maintenant, on n'a jugé de la pertinence du modèle de régression linéaire que sur un critère graphique subjectif, en appréciant le degré d'alignement des points du nuage des (x_i, y_i). Il est important de proposer un critère statistique objectif.

Pour cela, on remarque que, si les points (x_i, y_i) sont parfaitement alignés, alors r_{xy} est égal à ± 1. Inversement, si r_{xy} est proche de 0, on peut rejeter l'hypothèse de liaison affine. Il est donc naturel de construire un test de pertinence de la régression linéaire qui consiste à accepter la liaison affine si r_{xy} est "suffisamment proche" de ± 1, ou si r_{xy} est "suffisamment éloigné" de 0.

En pratique, la question que l'on se pose est : *à partir de quelle valeur de r_{xy} peut-on admettre l'hypothèse de liaison affine entre les variables ?* 90 % ? 95 % ? 99 % ? En fait, cette borne va dépendre de la taille de l'échantillon et du risque d'erreur que l'on accepte de prendre.

L'équation de la droite de régression est :

$$y = \bar{y}_n + \frac{c_{xy}}{s_x^2}(x - \bar{x}_n) = \bar{y}_n + \frac{r_{xy} s_y}{s_x}(x - \bar{x}_n)$$

Donc, si $r_{xy} \approx 0$ alors $y \approx \bar{y}_n =$ constante. Or, dans le modèle, y n'est constant que si $\beta_1 = 0$.

Par conséquent, on peut tester l'absence de corrélation entre les x_i et les y_i en testant "$\beta_1 = 0$" contre "$\beta_1 \neq 0$". La région critique d'un tel test est :

$$W = \left\{ \frac{|\widehat{\beta}_1|}{\widehat{\sigma}} s_x \sqrt{n} > t_{n-2,\alpha} \right\}$$

On peut l'écrire de manière plus parlante en faisant intervenir le coefficient de corrélation linéaire empirique.

En effet, $\widehat{\beta}_1 = \dfrac{C_{xY}}{s_x^2} = R_{xY}\dfrac{S_Y}{s_x}$ et $\widehat{\sigma}^2 = \dfrac{n}{(n-2)}S_Y^2(1-R_{xY}^2)$.

Par conséquent :

$$\frac{\widehat{\beta}_1}{\widehat{\sigma}}s_x\sqrt{n} = \frac{R_{xY}S_Y}{S_Y\sqrt{1-R_{xY}^2}}\sqrt{n-2} = \frac{R_{xY}}{\sqrt{1-R_{xY}^2}}\sqrt{n-2}$$

Par ailleurs, si T est de loi $St(n-2)$, alors T^2 est de loi de Fisher-Snedecor $F(1,n-2)$. Donc $P(|T| > t_{n-1,2\alpha}) = 1-\alpha = P(T^2 > t_{n-1,2\alpha}^2) = P(T^2 > f_{1,n-2,\alpha})$, ce qui implique que $t_{n-2,\alpha}^2 = f_{1,n-2,\alpha}$. On en déduit que :

Propriété 25 *Sous H_0 : "$\beta_1 = 0$", $\dfrac{R_{xY}}{\sqrt{1-R_{xY}^2}}\sqrt{n-2}$ est de loi $St(n-2)$, et $\dfrac{(n-2)R_{xY}^2}{1-R_{xY}^2}$ est de loi $F(1,n-2)$.*

On peut alors réécrire la région critique en remarquant que :

$$\begin{aligned}
\frac{|\widehat{\beta}_1|}{\widehat{\sigma}}s_x\sqrt{n} > t_{n-2,\alpha} &\iff \frac{|r_{xy}|}{\sqrt{1-r_{xy}^2}}\sqrt{n-2} > t_{n-2,\alpha} \iff \frac{(n-2)r_{xy}^2}{1-r_{xy}^2} > f_{1,n-2,\alpha} \\
&\iff (n-2)r_{xy}^2 > f_{1,n-2,\alpha}(1-r_{xy}^2) \\
&\iff r_{xy}^2(n-2+f_{1,n-2,\alpha}) > f_{1,n-2,\alpha} \iff r_{xy}^2 > \frac{f_{1,n-2,\alpha}}{n-2+f_{1,n-2,\alpha}}
\end{aligned}$$

D'où :

Propriété 26 Test de pertinence de la régression = *test de seuil α de H_0 : "$\beta_1 = 0$" contre H_1 : "$\beta_1 \neq 0$" :*

$$W = \left\{\frac{(n-2)r_{xy}^2}{1-r_{xy}^2} > f_{1,n-2,\alpha}\right\} = \left\{r_{xy}^2 > \frac{f_{1,n-2,\alpha}}{n-2+f_{1,n-2,\alpha}}\right\}$$

Cette région critique est bien de la forme attendue : on accepte l'hypothèse de liaison affine ssi r_{xy} est significativement proche de ± 1.

Dans l'exemple, $\dfrac{(n-2)r_{xy}^2}{1-r_{xy}^2} = 144.7$. La table de la loi de Fisher-Snedecor donne $f_{1,6,0.05} = 5.99$ et $f_{1,6,0.01} = 13.8$. Même au seuil 1%, on est très largement dans la région critique, donc on conclut que la régression linéaire est ici très pertinente.

Le nom de "test de pertinence de la régression" est abusif : on teste en fait si, parmi toutes les droites $y = \beta_1 x + \beta_0$, la droite constante $y = \beta_0$ est plausible ou pas.

6.5 Etude complète de l'exemple en R

En R, la commande permettant d'effectuer une régression linéaire de y sur x est `lm(y~x)`. Le résultat d'une régression est donné grâce à la commande `summary`. Sur l'exemple, on obtient :

```
> regvf<-lm(freinage~vitesse)
> summary(regvf)

Call:
lm(formula = freinage ~ vitesse)

Residuals:
    Min      1Q  Median      3Q     Max
-15.531  -7.766  -2.609   7.048  18.393

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.0614    10.1113  -3.863  0.00833 **
vitesse       4.8176     0.4005  12.030    2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.98 on 6 degrees of freedom
Multiple R-Squared: 0.9602,     Adjusted R-squared: 0.9536
F-statistic: 144.7 on 1 and 6 DF,  p-value: 2.002e-05
```

La colonne `Estimate` donne les estimations des moindres carrés de β_0 et β_1, $\hat{\beta}_0 = -39.06$ et $\hat{\beta}_1 = 4.82$.

La colonne `Std.error` donne les valeurs de $\dfrac{\hat{\sigma}\sqrt{s_x^2 + \bar{x}_n^2}}{s_x\sqrt{n}}$ et $\dfrac{\hat{\sigma}}{s_x\sqrt{n}}$, ce qui permet de déterminer des intervalles de confiance pour β_0 et β_1.

La colonne `t value` donne les valeurs de $\dfrac{|\hat{\beta}_0|}{\hat{\sigma}}\dfrac{s_x\sqrt{n}}{\sqrt{s_x^2 + \bar{x}_n^2}}$ et $\dfrac{|\hat{\beta}_1|}{\hat{\sigma}}s_x\sqrt{n}$, ce qui permet d'effectuer les tests de "$\beta_0 = 0$" contre "$\beta_0 \neq 0$" et "$\beta_1 = 0$" contre "$\beta_1 \neq 0$".

La colonne `Pr(>|t|)` donne les p-valeurs de ces tests. Dans l'exemple, ces p-valeurs sont très faibles, donc les hypothèses "$\beta_0 = 0$" et "$\beta_1 = 0$" sont largement rejetées. C'est logique puisque 0 n'appartient pas aux intervalles de confiance déterminés pour β_0 et β_1. Dans les lignes correspondantes, plus il y a d'étoiles plus le rejet est fort.

La `Residual standard error` est $\hat{\sigma}$, ce qui permet de retrouver $\hat{\sigma}^2 = 12.98^2 = 168.4$.

Le `Multiple R-Squared` est r_{xy}^2, ce qui permet de faire le test de pertinence de

la régression. La `F-statistic` est la statistique de ce test, $\frac{(n-2)r_{xy}^2}{1-r_{xy}^2}$. On retrouve qu'elle vaut 144.7. La `p-value` fournie est la p-valeur de ce test. Elle est très faible ($2\,10^{-5}$), donc on conclut bien que la régression linéaire est pertinente sur notre exemple.

Enfin, les commandes `plot(vitesse,freinage)` puis `lines(vitesse, fitted.values (regvf))` permettent de retrouver la figure représentant le nuage de points et la droite des moindres carrés.

Le modèle de régression linéaire simple gaussien semble donc à première vue satisfaisant pour l'exemple. Cependant, on s'aperçoit que ce modèle prévoit une distance de freinage négative pour toute vitesse inférieure à 8.1 m/s! D'autre part, la forme du nuage peut évoquer plus un polynôme qu'une droite, et des raisons physiques incitent à penser que la distance de freinage est plutôt une fonction quadratique de la vitesse. Enfin, il est obligatoire que la distance de freinage correspondant à une vitesse nulle soit zéro.

Tous ces arguments amènent à penser que le modèle $Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ pourrait être avantageusement remplacé par le modèle $Y_i = \beta_1 x_i^2 + \beta_0 x_i + \varepsilon_i$. C'est encore un modèle linéaire, qui se traite de façon similaire au précédent. Nous n'avons pas le temps d'étudier théoriquement ce modèle, mais il est facile de le mettre en oeuvre grâce à `R`. On obtient sur l'exemple :

```
> v2<-vitesse^2
> regvf2<-lm(freinage~v2+vitesse-1)
> summary(regvf2)

Call:
lm(formula = freinage ~ v2 + vitesse - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5569 -3.0400 -0.9151  2.7337  5.5614

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
v2      0.100497   0.007826  12.842 1.37e-05 ***
vitesse 0.246712   0.256589   0.962    0.373
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.54 on 6 degrees of freedom
Multiple R-Squared: 0.9981,     Adjusted R-squared: 0.9974
F-statistic:  1545 on 2 and 6 DF,  p-value: 7.275e-09
```

On a donc $\hat{\beta}_1 = 0.1005$, $\hat{\beta}_0 = 0.2467$ et $\hat{\sigma}^2 = 4.54^2 = 20.51$. La parabole d'équation $y = 0.1005x^2 + 0.2467x$ peut être appelée *parabole des moindres carrés* et est représentée sur la figure 6.5 :

```
> plot(vitesse,freinage)
> lines(vitesse,fitted.values(regvf2))
```

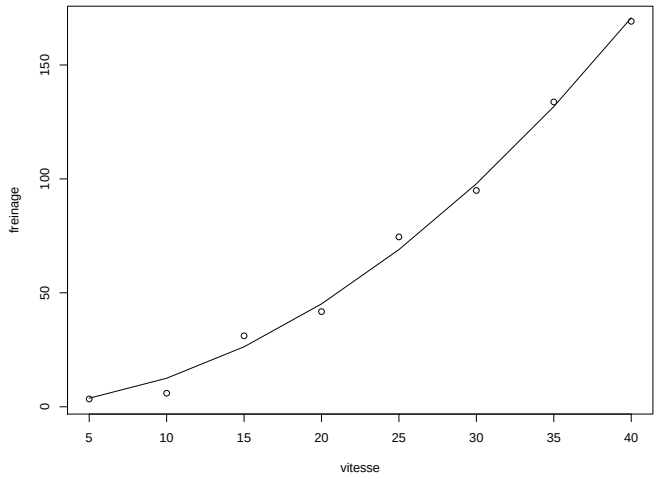


FIGURE 6.5 – Vitesse et distance de freinage, parabole des moindres carrés

Le coefficient de corrélation linéaire empirique est $r_{xy} = \sqrt{0.9981} = 0.99905$. Il est nettement plus proche de 1 que celui du modèle précédent, qui valait 0.9799. De la même façon, la p-valeur du test de pertinence de la régression vaut $7.3\,10^{-9}$, qui est nettement plus petite que celle que l'on avait obtenue dans le modèle précédent, $2\,10^{-5}$. Ces deux arguments montrent que le nouveau modèle est bien meilleur que le précédent.

La prévision de distance de freinage à la vitesse de 50 m/s est maintenant de $0.1005 * 502 + 0.2467 * 50 = 263.6$ m, alors qu'elle était de 201.9 m pour le modèle précédent. Cette importante différence peut avoir de grandes conséquences pratiques et met en évidence l'importance du choix d'un bon modèle de régression.

La couverture de ce polycopié, reproduite ci-dessous, représente un exemple de régression polynomiale d'ordre 3.

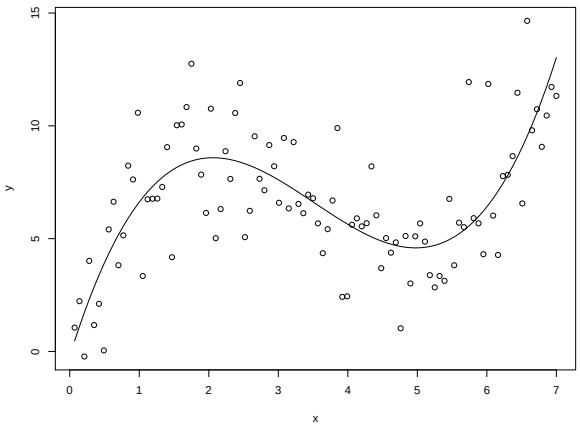


FIGURE 6.6 – Régression polynomiale d'ordre 3

Chapitre 7

Annexe A : Bases de probabilités pour la statistique

Cette annexe énonce quelques résultats de base du calcul des probabilités utiles pour la statistique. Les notions sont présentées sans aucune démonstration. Les détails ont été vus dans le cours de Probabilités Appliquées du premier semestre, ou seront développés dans les cours de deuxième et troisième année.

7.1 Variables aléatoires réelles

7.1.1 Loi de probabilité d'une variable aléatoire

Mathématiquement, une variable aléatoire est définie comme une application mesurable. On se contentera ici de la conception intuitive suivante.

Une **variable aléatoire** est une grandeur dépendant du résultat d'une expérience aléatoire, c'est-à-dire non prévisible à l'avance avec certitude. Par exemple, on peut dire que la durée de vie d'une ampoule électrique ou le résultat du lancer d'un dé sont des variables aléatoires. Pour une expérience donnée, ces grandeurs prendront une valeur donnée, appelée réalisation de la variable aléatoire. Si on recommence l'expérience, on obtiendra une réalisation différente de la même variable aléatoire.

On ne s'intéresse ici qu'aux **variables aléatoires réelles**, c'est-à-dire à valeurs dans $\mathbb{R}$ ou un sous-ensemble de $\mathbb{R}$. On note traditionnellement une variable aléatoire par une lettre majuscule (X) et sa réalisation par une lettre minuscule (x).

Le calcul des probabilités va permettre de calculer des grandeurs comme la durée de vie moyenne d'une ampoule ou la probabilité d'obtenir un 6 en lançant le dé. Ces grandeurs sont déterminées par la **loi de probabilité** de ces variables aléatoires.

Il y a plusieurs moyens de caractériser la loi de probabilité d'une variable aléatoire. La plus simple est la fonction de répartition.

On appelle **fonction de répartition** de la variable aléatoire X la fonction

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & F_X(x) = P(X \leq x) \end{array}$$

F_X est croissante, continue à droite, telle que $\lim_{x\to-\infty} F_X(x) = 0$ et $\lim_{x\to+\infty} F_X(x) = 1$. Elle permet de calculer la probabilité que X appartienne à n'importe quel intervalle de $\mathbb{R}$:

$$\forall (a,b) \in \mathbb{R}^2,\ a < b,\ P(a < X \leq b) = F_X(b) - F_X(a)$$

Les variables aléatoires peuvent être classées selon le type d'ensemble dans lequel elles prennent leurs valeurs. Dans la pratique, on ne s'intéressera qu'à deux catégories : les variables aléatoires discrètes et les variables aléatoires continues (ou à densité).

7.1.2 Variables aléatoires discrètes et continues

Une **variable aléatoire** X est dite **discrète (v.a.d.)** si et seulement si elle est à valeurs dans un ensemble E fini ou dénombrable. On peut noter $E = \{x_1, x_2, ...\}$.

Exemples :

- Face obtenue lors du lancer d'un dé : $E = \{1, 2, 3, 4, 5, 6\}$.
- Nombre de bugs dans un programme : $E = \mathbb{N}$.

La loi de probabilité d'une v.a.d. X est entièrement déterminée par les probabilités élémentaires $P(X = x_i), \forall x_i \in E$.

La fonction de répartition de X est alors $F_X(x) = P(X \leq x) = \sum\limits_{x_i \leq x} P(X = x_i)$.

Une **variable aléatoire** X est dite **continue (v.a.c.)** si et seulement si sa fonction de répartition F_X est continue et presque partout dérivable. Sa dérivée f_X est alors appelée densité de probabilité de X, ou plus simplement **densité** de X. Une v.a.c. est forcément à valeurs dans un ensemble non dénombrable.

Exemples :

- Appel de la fonction Random d'une calculatrice : $E = [0, 1]$.
- Durée de bon fonctionnement d'un système : $E = \mathbb{R}^+$.

On a alors $\forall (a,b) \in \mathbb{R}^2,\ a < b,\ P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$.

Plus généralement, $\forall B \subset \mathbb{R}, P(X \in B) = \int_B f_X(x)\,dx$. Donc la densité détermine entièrement la loi de probabilité de X.

f_X est une fonction positive telle que $\int_{-\infty}^{+\infty} f_X(x)\,dx = P(X \in \mathbb{R}) = 1$.

Connaissant la loi de X, on est souvent amenés à déterminer celle de $Y = \varphi(X)$. Quand X est discrète, il suffit d'écrire $P(Y = y) = P\left(\varphi(X) = y\right)$. Si φ est inversible, on obtient $P(Y = y) = P\left(X = \varphi^{-1}(y)\right)$. Quand X est continue, on commence par déterminer la fonction de répartition de Y en écrivant $F_Y(y) = P(Y \leq y) = P\left(\varphi(X) \leq y\right)$, puis on en déduit sa densité par dérivation. Quand φ est inversible, on obtient la **formule du changement de variable** :

$$f_Y(y) = \frac{1}{\left|\varphi'\left(\varphi^{-1}(y)\right)\right|} f_X\left(\varphi^{-1}(y)\right)$$

Remarque : Il existe des lois de probabilité de variables aléatoires réelles qui ne sont ni discrètes ni continues. Par exemple, si X est la durée de bon fonctionnement d'un

système qui a une probabilité non nulle p d'être en panne à l'instant initial, on a $\lim_{x\to 0^-} F_X(x) = 0$ (une durée ne peut pas être négative) et $F_X(0) = P(X \leq 0) = P(X = 0) = p$. Par conséquent F_X n'est pas continue en 0. La loi de X ne peut donc pas être continue, et elle n'est pas non plus discrète puisqu'elle est à valeurs dans $\mathbb{R}^+$. Ce type de variable aléatoire ne sera pas étudié dans ce cours.

7.1.3 Moments et quantiles d'une variable aléatoire réelle

Si X est une variable aléatoire discrète, son **espérance mathématique** est définie par :

$$E[X] = \sum_{x_i \in E} x_i P(X = x_i)$$

Si X est une variable aléatoire continue, son espérance mathématique est définie par :

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

Concrètement, $E[X]$ est ce qu'on s'attend à trouver comme moyenne des résultats obtenus si on répète l'expérience un grand nombre de fois. Par exemple, si on lance une pièce de monnaie 10 fois, on s'attend à trouver en moyenne 5 piles.

Plus généralement, on peut s'intéresser à l'espérance mathématique d'une fonction de X :

- Si X est une v.a.d., $E[\varphi(X)] = \sum_{x_i \in E} \varphi(x_i) P(X = x_i)$.
- Si X est une v.a.c., $E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f_X(x)\, dx$.

Ce résultat permet de calculer l'espérance de $\varphi(X)$ sans avoir à déterminer entièrement sa loi.

Deux espérances de ce type sont particulièrement utiles :

- Si X est une v.a.d., sa **fonction génératrice** est définie par $G_X(z) = E\left[z^X\right] = \sum_{x_i \in E} z^{x_i} P(X = x_i)$.
- Si X est une v.a.c., sa **fonction caractéristique** est définie par $\phi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} f_X(x)\, dx$.

Au même titre que la fonction de répartition et la densité, les fonctions génératrices et caractéristiques définissent entièrement les lois de probabilité concernées.

Soit k un entier naturel quelconque. Le **moment d'ordre** k de X est $E\left[X^k\right]$ et le **moment centré d'ordre** k est $E\left[(X - E(X))^k\right]$.

De tous les moments, le plus important est le moment centré d'ordre 2, appelé aussi **variance**. La variance de X est $Var[X] = E\left[(X - E(X))^2\right]$, qui se calcule plus facilement sous la forme $Var[X] = E[X^2] - [E[X]]^2$.

L'**écart-type** de X est $\sigma[X] = \sqrt{Var[X]}$.

La variance et l'écart-type sont des indicateurs de la dispersion de X : plus la variance de X est petite, plus les réalisations de X seront concentrées autour de son espérance.

Le **coefficient de variation** de X est $CV[X] = \dfrac{\sigma[X]}{E[X]}$. C'est également un indicateur de dispersion, dont l'avantage est d'être sans dimension. Il permet de comparer les dispersions de variables aléatoires d'ordres de grandeur différents ou exprimées dans des unités différentes. En pratique, on considère que, quand $CV[X]$ est inférieur à 15%, l'espérance peut être considérée comme un bon résumé de la loi.

Soit $p \in]0,1[$. Le **quantile d'ordre** p (ou p**-quantile**) de la loi de X est tout réel q_p vérifiant $P(X < q_p) \leq p \leq P(X \leq q_p)$.

- Si F est continue et strictement croissante (donc inversible), on a simplement $P(X < q_p) = P(X \leq q_p) = F_X(q_p) = p$, d'où $q_p = F_X^{-1}(p)$.
- Si F_X est constante égale à p sur un intervalle $[a,b]$, n'importe quel réel de $[a,b]$ est un quantile d'ordre p. En général, on choisit de prendre le milieu de l'intervalle : $q_p = \dfrac{a+b}{2}$.
- Si F_X est discontinue en q et telle que $\lim\limits_{x \to q^-} F_X(x) < p \leq F_X(q)$, alors $q_p = q$.

Les tables fournies donnent les quantiles les plus usuels des lois normale, du chi-deux, de Student et de Fisher-Snedecor.

7.2 Vecteurs aléatoires réels

On ne s'intéressera ici qu'aux vecteurs aléatoires $(X_1, \ldots, X_n)$ constitués de n variables aléatoires réelles toutes discrètes ou toutes continues.

7.2.1 Loi de probabilité d'un vecteur aléatoire

La loi d'un vecteur aléatoire $(X_1, \ldots, X_n)$ est déterminée par sa fonction de répartition :

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right)$$

Si les X_i sont discrètes, cette loi est aussi déterminée par les probabilités élémentaires $P\left(X_1 = x_1, \ldots, X_n = x_n\right)$.

Si les X_i sont continues, la densité de $(X_1, \ldots, X_n)$ est définie, si elle existe, par :

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)$$

On a alors $\forall B \subset \mathbb{R}^n, P\left((X_1, \ldots, X_n) \in B\right) = \int \ldots \int_B f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$.

Les variables aléatoires $X_1, \ldots, X_n$ sont (mutuellement) **indépendantes** si et seulement si :

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)$$

Pour des variables discrètes cela donne $P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$.

Et pour des variables continues, $f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

Concrètement, l'indépendance signifie que la valeur prise par l'une des variables n'a aucune influence sur la valeur prise par les autres.

7.2.2 Espérance et matrice de covariance d'un vecteur aléatoire

L'**espérance mathématique** d'un vecteur aléatoire est le vecteur des espérances mathématiques de ses composantes : $E[(X_1, \ldots, X_n)] = (E[X_1], \ldots, E[X_n])$.

L'équivalent de la variance en dimension n est la **matrice de covariance** du vecteur $(X_1, \ldots, X_n)$, notée $K_{(X_1,\ldots,X_n)}$ ou K, dont les coefficients sont donnés par

$$k_{ij} = Cov(X_i, X_j), \forall (i,j) \in \{1, ..., n\}^2$$

$Cov(X_i, X_j)$ est la covariance des variables aléatoires X_i et X_j et est définie par :

$$Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])] = E[X_i X_j] - E[X_i]E[X_j]$$

Pour $i = j$, $Cov(X_i, X_i) = E[X_i^2] - [E[X_i]]^2 = Var[X_i]$.

Pour $i \neq j$, la covariance de X_i et X_j traduit le degré de corrélation entre ces deux variables. En particulier, si X_i et X_j sont indépendantes, $Cov(X_i, X_j) = 0$ (mais la réciproque est fausse). Par conséquent, si $X_1, \ldots, X_n$ sont indépendantes, leur matrice de covariance K est diagonale.

Le **coefficient de corrélation linéaire** entre X_i et X_j est $\rho(X_i, X_j) = \dfrac{Cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}$.

On montre que :

- $\rho(X_i, X_j) \in [-1, +1]$.
- $\rho(X_i, X_j) = +1 \Leftrightarrow X_i = aX_j + b$, avec $a > 0$ et $b \in \mathbb{R}$.
- $\rho(X_i, X_j) = -1 \Leftrightarrow X_i = -aX_j + b$, avec $a > 0$ et $b \in \mathbb{R}$.
- si $\rho(X_i, X_j) > 0$, X_i et X_j sont corrélées positivement, ce qui signifie qu'elles varient dans le même sens. Par exemple, X_i et X_j peuvent être la taille et le poids d'individus pris au hasard.
- si $\rho(X_i, X_j) < 0$, X_i et X_j sont corrélées négativement, ce qui signifie qu'elles varient en sens contraire. Par exemple, X_i et X_j peuvent être l'âge et la résistance d'un matériau.

- si $\rho(X_i, X_j) = 0$, il n'y a pas de corrélation linéaire entre X_i et X_j. Cela ne signifie pas que X_i et X_j sont indépendantes. Il peut éventuellement y avoir une corrélation non linéaire.

L'espérance mathématique est linéaire : si X et Y sont des variables aléatoires et a, b et c des réels, alors $E[aX + bY + c] = aE[X] + bE[Y] + c$.

En revanche, la variance n'est pas linéaire : si X et Y sont des variables aléatoires et a, b et c des réels, alors $Var[aX + bY + c] = a^2Var[X] + 2abCov(X, Y) + b^2Var[Y]$.
Si X et Y sont indépendantes, $Cov(X_i, X_j) = 0$, donc $Var[aX + bY + c] = a^2Var[X] + b^2Var[Y]$. En particulier, la variance de la somme de variables aléatoires indépendantes est égale à la somme des variances de ces variables. Mais ce résultat est faux si les variables ne sont pas indépendantes.

7.3 Lois de probabilité usuelles

Les tables de lois de probabilité fournies donnent notamment, pour les lois les plus usuelles, les probabilités élémentaires ou la densité, l'espérance, la variance, et la fonction génératrice ou la fonction caractéristique. On présente dans cette section quelques propriétés supplémentaires de quelques unes de ces lois.

7.3.1 Loi binomiale

Une variable aléatoire K est de loi binomiale $\mathcal{B}(n, p)$ si et seulement si elle est à valeurs dans $\{0, 1, ..., n\}$ et $P(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}$.

Le nombre de fois où, en n expériences identiques et indépendantes, un évènement de probabilité p s'est produit, est une variable aléatoire de loi $\mathcal{B}(n, p)$.

La loi de Bernoulli $\mathcal{B}(p)$ est la loi $\mathcal{B}(1, p)$.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{B}(m, p)$, alors $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{B}(nm, p)$. En particulier, la somme de n v.a. indépendantes et de même loi $\mathcal{B}(p)$ est de loi $\mathcal{B}(n, p)$.

7.3.2 Loi géométrique

Une variable aléatoire K est de loi géométrique $\mathcal{G}(p)$ si et seulement si elle est à valeurs dans $\mathbb{N}^*$ et $P(K = k) = p\,(1 - p)^{k-1}$.

Dans une suite d'expériences identiques et indépendantes, le nombre d'expériences nécessaires pour que se produise pour la première fois un évènement de probabilité p, est une variable aléatoire de loi $\mathcal{G}(p)$.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{G}(p)$, alors $\sum_{i=1}^{n} X_i$ est de loi binomiale négative $\mathcal{BN}(n, p)$.

7.3.3 Loi de Poisson

Une variable aléatoire K est de loi de Poisson $\mathcal{P}(\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{N}$ et $P(K=k) = e^{-\lambda}\,\dfrac{\lambda^k}{k!}$.

Pour $n \geq 50$ et $p \leq 0.1$, la loi binomiale $\mathcal{B}(n,p)$ peut être approchée par la loi de Poisson $\mathcal{P}(np)$. On dit que la loi de Poisson est la loi des évènements rares : loi du nombre de fois où un événement de probabilité très faible se produit au cours d'un très grand nombre d'expériences identiques et indépendantes.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{P}(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{P}(n\lambda)$.

7.3.4 Loi exponentielle

Une variable aléatoire X est de loi exponentielle $exp(\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{R}^+$ et $f_X(x) = \lambda\, e^{-\lambda x}$.

La loi exponentielle est dite sans mémoire : $\forall (t,x) \in \mathbb{R}^{+^2}, P(X > t+x | X > t) = P(X > x)$.

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $exp(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi gamma $G(n,\lambda)$.

7.3.5 Loi gamma et loi du chi-2

Une variable aléatoire X est de loi gamma $G(a,\lambda)$ si et seulement si elle est à valeurs dans $\mathbb{R}^+$ et $f_X(x) = \dfrac{\lambda^a}{\Gamma(a)}\, e^{-\lambda x} x^{a-1}$. Les propriétés de la fonction gamma sont rappelées sur les tables.

La loi $G(1,\lambda)$ est la loi $exp(\lambda)$.

La loi $G\left(\dfrac{n}{2}, \dfrac{1}{2}\right)$ est appelée loi du chi-2 à n degrés de liberté, notée χ^2_n.

Si X est de loi $G(a,\lambda)$ et α est un réel strictement positif, alors αX est de loi $G\left(a, \dfrac{\lambda}{\alpha}\right)$.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $G(\alpha,\lambda)$ et $G(\beta,\lambda)$, alors $X+Y$ est de loi $G(\alpha+\beta,\lambda)$. En particulier, si X et Y sont indépendantes et de lois respectives χ^2_n et χ^2_m, alors $X+Y$ est de loi χ^2_{n+m}.

7.3.6 Loi normale

Une variable aléatoire X est de loi normale $\mathcal{N}(m,\sigma^2)$ si et seulement si elle est à valeurs dans $\mathbb{R}$ et $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-m)^2}{2\sigma^2}}$.

Si X est de loi $\mathcal{N}(m, \sigma^2)$, alors $aX+b$ est de loi $\mathcal{N}(am+b, a^2\sigma^2)$. En particulier, $\dfrac{X-m}{\sigma}$ est de loi $\mathcal{N}(0,1)$.

$P(X \in [m-\sigma, m+\sigma]) = 68.3\%$. $\quad$ $P(X \in [m-2\sigma, m+2\sigma]) = 95.4\%$.
$P(X \in [m-3\sigma, m+3\sigma]) = 99.7\%$.

Si X est de loi $\mathcal{N}(0,1)$, alors X^2 est de loi χ^2_1.

Si X et Y sont des variables aléatoires indépendantes de lois respectives $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$, alors $aX + bY$ est de loi $\mathcal{N}\left(am_1 + bm_2, a^2\sigma_1^2 + b^2\sigma_2^2\right)$.

7.3.7 Lois de Student et de Fisher-Snedecor

Soit U une variable aléatoire de loi $\mathcal{N}(0,1)$ et X une variable aléatoire de loi χ^2_n. Si U et X sont indépendantes, alors $\sqrt{n}\,\dfrac{U}{\sqrt{X}}$ est de loi de Student à n degrés de liberté $St(n)$.

Soit X une variable aléatoire de loi χ^2_n et Y une variable aléatoire de loi χ^2_m. Si X et Y sont indépendantes, alors $\dfrac{mX}{nY}$ est de loi de Fisher-Snedecor $F(n,m)$.

Ces deux définitions entrainent que si T est de loi $St(n)$, alors T^2 est de loi $F(1,n)$.

Les lois de Student et de Fisher-Snedecor sont toujours utilisées par l'intermédiaire de tables ou à l'aide d'un logiciel de statistique. Il n'est donc pas nécessaire de donner l'expression de leur densité.

Chapitre 8

Annexe B : Lois de probabilité usuelles

8.1 Caractéristiques des lois usuelles

8.1.1 Variables aléatoires réelles discrètes

Dans le tableau ci-dessous, on suppose $n \in \mathbb{N}^*$, $p \in]0,1[$ et $\lambda \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Probabilités	$E[X]$	$Var[X]$	Fonction caractéristique $\varphi_X(t) = E[e^{itX}]$
Bernoulli $\mathcal{B}(p)$	$P(X=0) = 1-p$ $P(X=1) = p$	p	$p(1-p)$	$1-p+pe^{it}$
Binomiale $\mathcal{B}(n,p)$	$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \mathbb{1}_{\{0,\dots,n\}}(k)$	np	$np(1-p)$	$(1-p+pe^{it})^n$
Binomiale négative $\mathcal{BN}(n,p)$	$P(X=k) = \binom{k-1}{n-1} p^n (1-p)^{k-n} \mathbb{1}_{\{n,\dots\}}(k)$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	$\left(\frac{pe^{it}}{1-(1-p)e^{it}}\right)^n$
Poisson $\mathcal{P}(\lambda)$	$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}\, \mathbb{1}_{\mathbb{N}}(k)$	λ	λ	$e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$P(X=k) = p(1-p)^{k-1} \mathbb{1}_{\mathbb{N}^*}(k)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$
Hypergéométrique $\mathcal{H}(N,m,n)$ $(m,n) \in \{1,\dots,N\}^2$	$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \mathbb{1}_{\{0,\dots,\min(m,n)\}}(k)$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	

8.1.2 Variables aléatoires réelles continues

La fonction Gamma est définie pour $a > 0$ par $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$.

$$\text{On a : } \forall\, n \in \mathbb{N}^*, \quad \Gamma(n) = (n-1)!\,, \quad \Gamma(1) = 1\,, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\,,$$

$$\forall\, a \in]1, +\infty[\,,\ \Gamma(a) = (a-1)\Gamma(a-1)\,.$$

Dans le tableau ci dessous, $[a, b] \subset \mathbb{R}$, $m \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $\lambda \in \mathbb{R}_+^*$, $\alpha \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$

Loi et Symbole $X \rightsquigarrow$	Densité	$E[X]$	$Var[X]$	Fonction caractéristique $\varphi_X(t) = E[e^{itX}]$
Loi Uniforme $\mathcal{U}[a, b]$	$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
Loi Normale $\mathcal{N}(m, \sigma^2)$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \mathbb{1}_{\mathbb{R}}(x)$	m	σ^2	$e^{itm - \frac{\sigma^2 t^2}{2}}$
Loi Exponentielle $\exp(\lambda) = G(1, \lambda)$	$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\left(1 - \frac{it}{\lambda}\right)^{-1}$
Loi Gamma $G(\alpha, \lambda)$	$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\left(1 - \frac{it}{\lambda}\right)^{-\alpha}$
Loi du Chi-deux $\chi_n^2 = G(\frac{n}{2}, \frac{1}{2})$	$f_X(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \mathbb{1}_{\mathbb{R}_+}(x)$	n	$2n$	$(1 - 2it)^{-\frac{n}{2}}$
Première loi de Laplace	$f_X(x) = \frac{1}{2} e^{-\lvert x \rvert} \mathbb{1}_{\mathbb{R}}(x)$	0	2	$\frac{1}{1+t^2}$

La fonction Beta est définie pour $a > 0$ et $b > 0$ par

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

Dans le tableau suivant, on suppose $a \in \mathbb{R}_+^*$, $b \in \mathbb{R}_+^*$ et $\eta \in \mathbb{R}_+^*$, $\beta \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Densité	$E[X]$	$Var[X]$
Loi Beta de 1ère espèce $\beta_1(a,b)$	$f_X(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Loi Beta de 2ème espèce $\beta_2(a,b)$	$f_X(x) = \frac{1}{\beta(a,b)} \frac{x^{a-1}}{(1+x)^{a+b}} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{a}{b-1}$ si $b > 1$	$\frac{a(a+b-1)}{(b-1)^2(b-2)}$ si $b > 2$
Loi de Weibull $\mathcal{W}(\eta,\beta)$	$f_X(x) = \frac{\beta}{\eta^\beta} x^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\eta\Gamma(1+\frac{1}{\beta})$	$\eta^2\left[\Gamma(1+\frac{2}{\beta}) - \Gamma(1+\frac{1}{\beta})^2\right]$

8.1.3 Vecteurs aléatoires dans $\mathbb{N}^d$ et dans $\mathbb{R}^d$

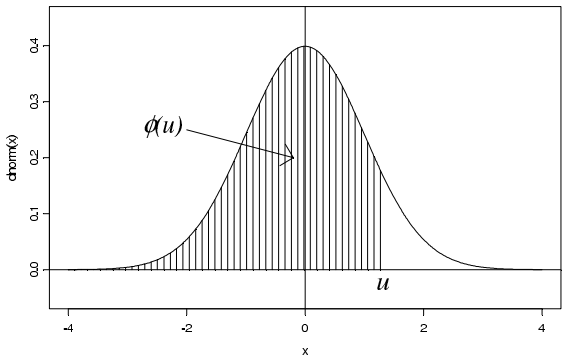
Dans le tableau suivant, on a :
$n \in \mathbb{N}^*$, $p = (p_1, p_2, \ldots, p_d) \in]0,1[^d$, $\sum_{i=1}^d p_i = 1$ et $k = (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d$, $\sum_{i=1}^d k_i = n$.
$m \in \mathbb{R}^d$ et $\Sigma \in M_{d,d}$.

Loi et Symbole $X \rightsquigarrow$	Probabilités ou Densité	$E[X]$	Matrice de covariance	Fonction Caractéristique
Loi Multinomiale $\mathcal{M}_d(n,p)$	$P(X=k) = \frac{n!}{k_1!\ldots k_d!} p_1^{k_1} p_2^{k_2} \ldots p_d^{k_d} \mathbb{1}_{\mathbb{N}^d}(k)$	np	$c_{i,i} = np_i(1-p_i)$ $c_{i,j} = -np_ip_j$, $i \neq j$	$\left[\sum_{i=1}^d p_i z_i\right]^n$
Loi normale $\mathcal{N}_d(m,\Sigma)$	$f_X(x) = \frac{1}{\sqrt{det\Sigma}(\sqrt{2\pi})^d} e^{-\frac{1}{2}{}^t(x-m)\Sigma^{-1}(x-m)}$	m	Σ	$e^{i{}^t m t - \frac{1}{2}{}^t t \Sigma t}$

8.2 Tables de lois

8.2.1 Table 1 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0,1)$, la table donne la valeur de $\phi(u) = P(U \leq u)$. En R, la commande correspondante est `pnorm(u)`.



u	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

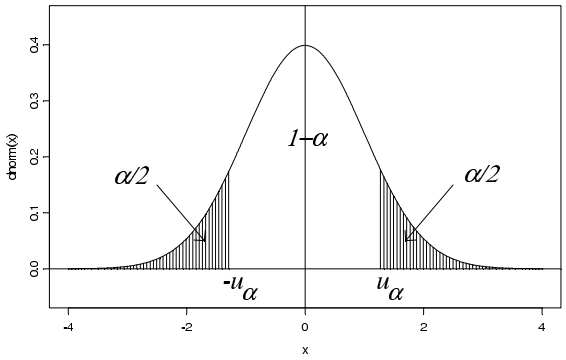
Lecture de la table : $\phi(1.25) = \phi(1.2 + 0.05) = 0.8944$.

Grandes valeurs de u

u	3.0	3.5	4.0	4.5
$\phi(u)$	0.9987	0.99977	0.999968	0.999997

8.2.2 Table 2 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0,1)$ et α un réel de $[0,1]$, la table donne la valeur de $u_\alpha = \phi^{-1}\left(1-\frac{\alpha}{2}\right)$ telle que $P(|U| > u_\alpha) = \alpha$. En R, la commande correspondante est `qnorm(1-alpha/2)`.



α	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	$+\infty$	2.5758	2.3263	2.1701	2.0537	1.96	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

Lecture de la table : $u_{0.25} = u_{0.2+0.05} = 1.1503$.

Petites valeurs de α

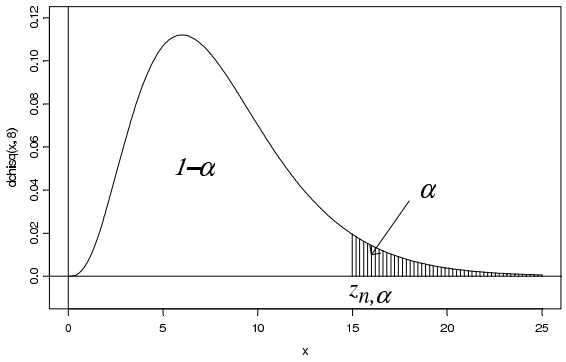
α	0.002	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_α	3.0902	3.2905	3.8906	4.4171	4.8916	5.3267	5.7307	6.1094

$$\text{Pour } p < \frac{1}{2}, \phi^{-1}(p) = -u_{2p}\,.$$

$$\text{Pour } p \geq \frac{1}{2}, \phi^{-1}(p) = u_{2(1-p)}\,.$$

8.2.3 Table de la loi du χ^2

X étant une variable aléatoire de loi du χ^2 à n degrés de libertés et α un réel de $[0,1]$, la table donne la valeur de $z_{n,\alpha} = F^{-1}_{\chi^2_n}(1-\alpha)$ telle que $P(X > z_{n,\alpha}) = \alpha$. En R, la commande correspondante est `qchisq(1-alpha, n)`.



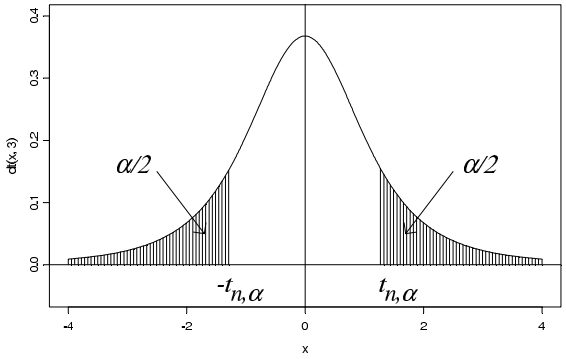
n \ α	0.995	0.990	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00004	0.0002	0.001	0.004	0.02	0.06	0.15	0.45	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.6	13.82
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70

Pour $n > 30$, on admet que $z_{n,\alpha} \approx \frac{1}{2}\left(u_{2\alpha} + \sqrt{2n-1}\right)^2$ si $\alpha < \frac{1}{2}$

et $z_{n,\alpha} \approx \frac{1}{2}\left(\sqrt{2n-1} - u_{2(1-\alpha)}\right)^2$ si $\alpha \geq \frac{1}{2}$.

8.2.4 Table de la loi de Student

X étant une variable aléatoire de loi $St(n)$ et α un réel de $[0,1]$, la table donne la valeur de $t_{n,\alpha} = F^{-1}_{St(n)}\left(1-\frac{\alpha}{2}\right)$ telle que $P(|X| > t_{n,\alpha}) = \alpha$. En R, la commande correspondante est `qt(1-alpha/2,n)`. Pour $n = +\infty$, $t_{+\infty,\alpha} = u_\alpha$.



$_n{}^\alpha$	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$+\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

8.2.5 Tables de la loi de Fisher-Snedecor

X étant une variable aléatoire de loi $F(\nu_1, \nu_2)$, les tables donnent les valeurs de $f_{\nu_1,\nu_2,\alpha} = F^{-1}_{F(\nu_1,\nu_2)}(1-\alpha)$ telles que $P(X > f_{\nu_1,\nu_2,\alpha}) = \alpha$ pour $\alpha = 5\%$ et $\alpha = 1\%$. En R, la commande correspondante est `qf(1-alpha, nu1, nu2)`. $f_{\nu_2,\nu_1,\alpha} = \dfrac{1}{f_{\nu_1,\nu_2,1-\alpha}}$.

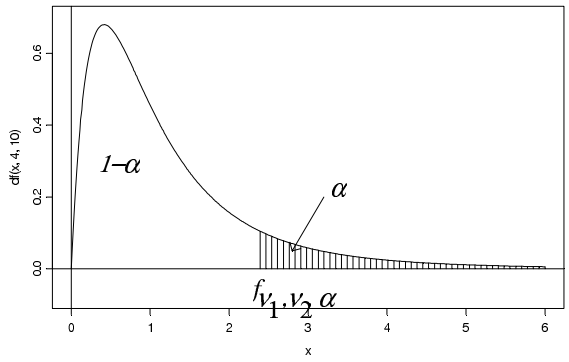


Table 1 : $\alpha = 5\%$.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	241.9	243.9	246.5	248	249	251.1	252.2	253	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.40	19.41	19.43	19.45	19.45	19.47	19.48	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.69	8.66	8.64	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.84	5.80	5.77	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.60	4.56	4.53	4.46	4.43	4.41	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.92	3.87	3.84	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.49	3.44	3.41	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.20	3.15	3.12	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.99	2.94	2.90	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.83	2.77	2.74	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.70	2.65	2.61	2.53	2.49	2.46	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.60	2.54	2.51	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.51	2.46	2.42	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.44	2.39	2.35	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.38	2.33	2.29	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.33	2.28	2.24	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.29	2.23	2.19	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.25	2.19	2.15	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.21	2.16	2.11	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.18	2.12	2.08	1.99	1.95	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.16	2.10	2.05	1.96	1.92	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.13	2.07	2.03	1.94	1.89	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.11	2.05	2.01	1.91	1.86	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.09	2.03	1.98	1.89	1.84	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.07	2.01	1.96	1.87	1.82	1.78	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.99	1.93	1.89	1.79	1.74	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.90	1.84	1.79	1.69	1.64	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.03	1.95	1.85	1.78	1.74	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.82	1.75	1.70	1.59	1.53	1.48	1.39
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.77	1.70	1.65	1.54	1.48	1.43	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.75	1.68	1.63	1.52	1.45	1.39	1.28
$+\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.64	1.57	1.52	1.39	1.32	1.24	1.00

Table 2 : $\alpha = 1\%$.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	16	20	24	40	60	100	$+\infty$
1	4052	4999	5403	5625	5764	5859	5928	5981	6056	6106	6170	6209	6235	6287	6313	6334	6366
2	98.5	99.0	99.17	99.25	99.3	99.33	99.36	99.37	99.4	99.42	99.44	99.45	99.46	99.47	99.48	99.49	99.5
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.23	27.05	26.83	26.69	26.60	26.41	26.32	26.24	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.55	14.37	14.15	14.02	13.93	13.75	13.65	13.58	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.05	9.89	9.68	9.55	9.47	9.29	9.20	9.13	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.52	7.40	7.31	7.14	7.06	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.28	6.16	6.07	5.91	5.82	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.48	5.36	5.28	5.12	5.03	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.92	4.81	4.73	4.57	4.48	4.41	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.52	4.41	4.33	4.17	4.08	4.01	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.21	4.10	4.02	3.86	3.78	3.71	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	3.97	3.86	3.78	3.62	3.54	3.47	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.78	3.66	3.59	3.43	3.34	3.27	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.94	3.80	3.62	3.51	3.43	3.27	3.18	3.11	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.80	3.67	3.49	3.37	3.29	3.13	3.05	2.98	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.69	3.55	3.37	3.26	3.18	3.02	2.93	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.59	3.46	3.27	3.16	3.08	2.92	2.83	2.76	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.51	3.37	3.19	3.08	3.00	2.84	2.75	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.43	3.30	3.12	3.00	2.92	2.76	2.67	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.37	3.23	3.05	2.94	2.86	2.69	2.61	2.54	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.31	3.17	2.99	2.88	2.80	2.64	2.55	2.48	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.26	3.12	2.94	2.83	2.75	2.58	2.50	2.42	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.21	3.07	2.89	2.78	2.70	2.54	2.45	2.37	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.17	3.03	2.85	2.74	2.66	2.49	2.40	2.33	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.13	2.99	2.81	2.70	2.62	2.45	2.36	2.29	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.98	2.84	2.66	2.55	2.47	2.30	2.21	2.13	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.80	2.66	2.48	2.37	2.29	2.11	2.02	1.94	1.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.70	2.56	2.38	2.27	2.18	2.01	1.91	1.82	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.63	2.50	2.31	2.20	2.12	1.94	1.84	1.75	1.60
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.55	2.42	2.23	2.12	2.03	1.85	1.75	1.65	1.49
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.50	2.37	2.19	2.07	1.98	1.80	1.69	1.60	1.43
$+\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.32	2.18	2.00	1.88	1.79	1.59	1.47	1.36	1.00

8.3 Exemples de représentations de probabilités et de densités

8.3.1 Lois discrètes

Loi binomiale

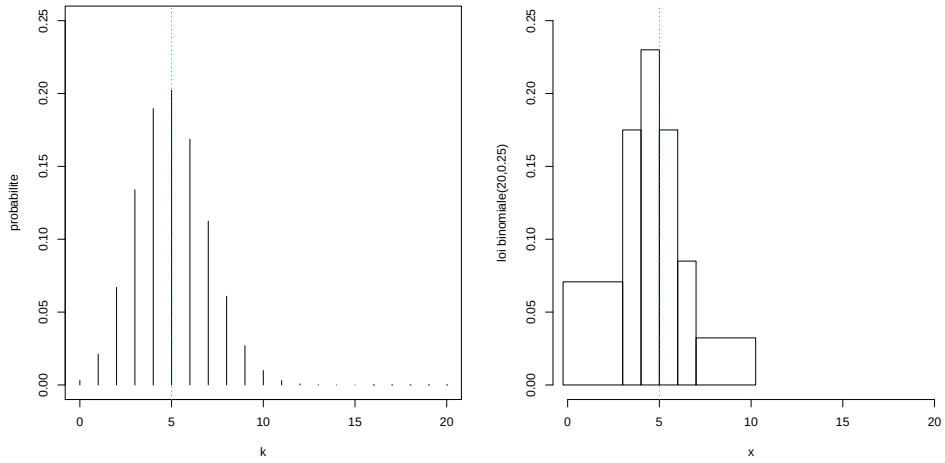


FIGURE 8.1 – Loi binomiale $\mathcal{B}(20, 0.25)$

La figure ci-dessus concerne la loi binomiale $\mathcal{B}(20, 0.25)$.

La partie gauche représente les probabilités élémentaires $P(X = k) = \binom{20}{k}\ 0.25^k\ 0.75^{20-k}$ pour $k \in \{0, \ldots, 20\}$, quand X est une variable aléatoire de loi $\mathcal{B}(20, 0.25)$.

La partie droite est un histogramme d'un échantillon de taille 200 simulé selon la loi $\mathcal{B}(20, 0.25)$.

Le trait vertical représente l'espérance de la loi $\mathcal{B}(20, 0.25)$.

Les figures suivantes présentent des représentations similaires pour les lois $\mathcal{B}(200, 0.25)$, $\mathcal{B}(20, 0.10)$ et $\mathcal{B}(200, 0.10)$.

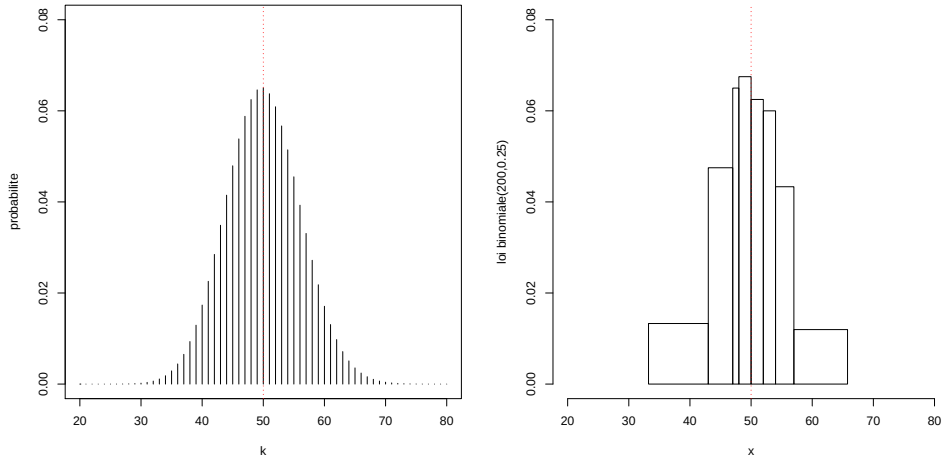


FIGURE 8.2 – Loi binomiale $\mathcal{B}(200, 0.25)$

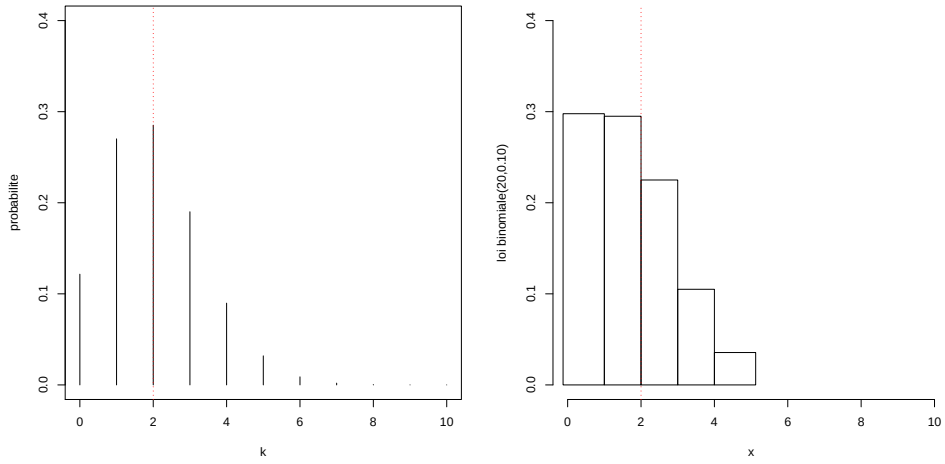


FIGURE 8.3 – Loi binomiale $\mathcal{B}(20, 0.10)$

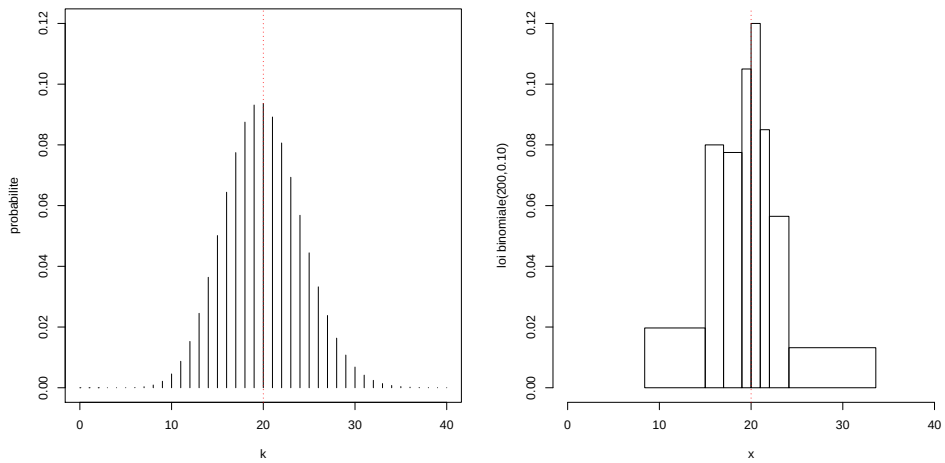


FIGURE 8.4 – Loi binomiale $\mathcal{B}(200, 0.10)$

Loi géométrique

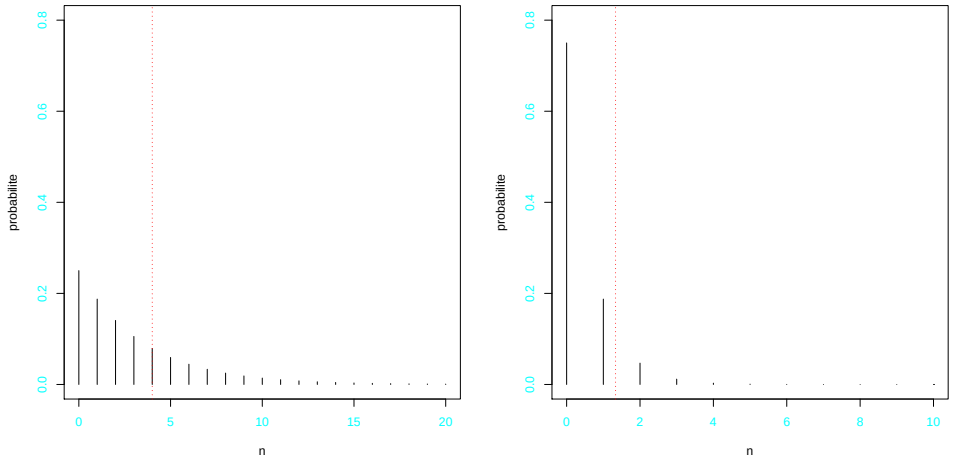


FIGURE 8.5 – Lois géométriques $\mathcal{G}(0.25)$ et $\mathcal{G}(0.75)$

Loi de Poisson

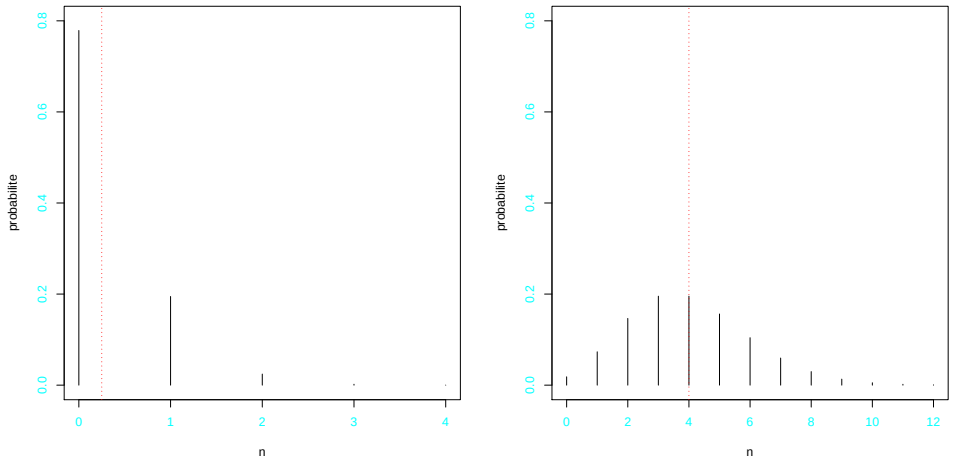


FIGURE 8.6 – Lois de Poisson $\mathcal{P}(0.25)$ et $\mathcal{P}(4)$

8.3.2 Lois continues

Loi normale

La figure ci-dessous présente la densité des lois normales $\mathcal{N}(0,1)$, $\mathcal{N}(2,2)$ et $\mathcal{N}(2,0.7)$.

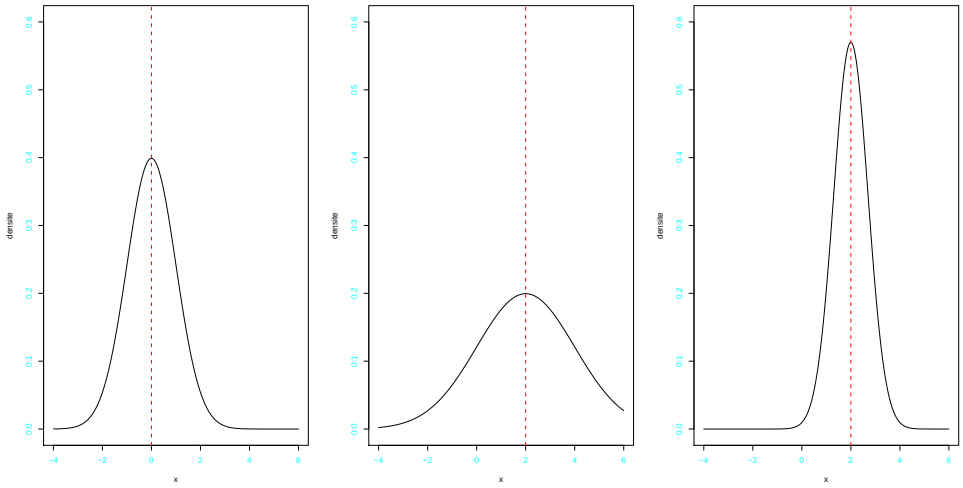


FIGURE 8.7 – Lois normales $\mathcal{N}(0,1)$, $\mathcal{N}(2,2)$ et $\mathcal{N}(2,0.7)$

Les figures suivantes présentent les densités d'autres lois de probabilité continues.

Loi exponentielle

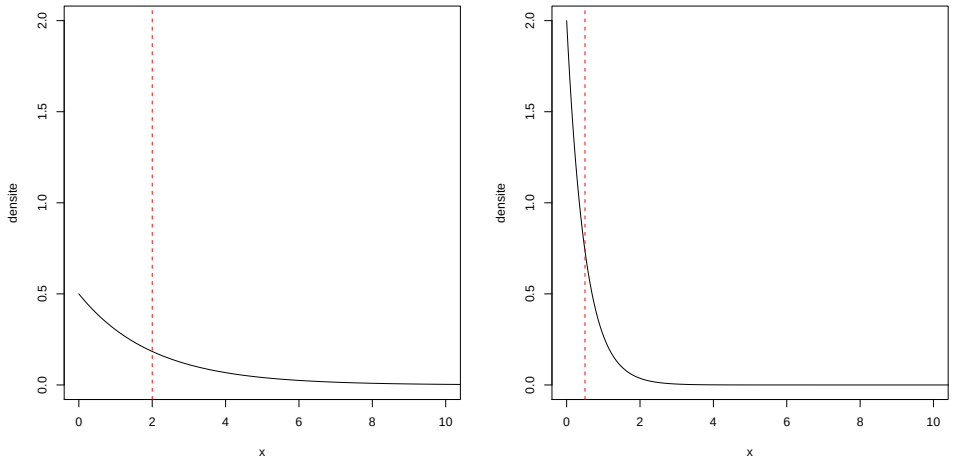


FIGURE 8.8 – Lois exponentielles $\mathcal{E}xp(0.5)$ et $\mathcal{E}xp(2)$

Loi gamma

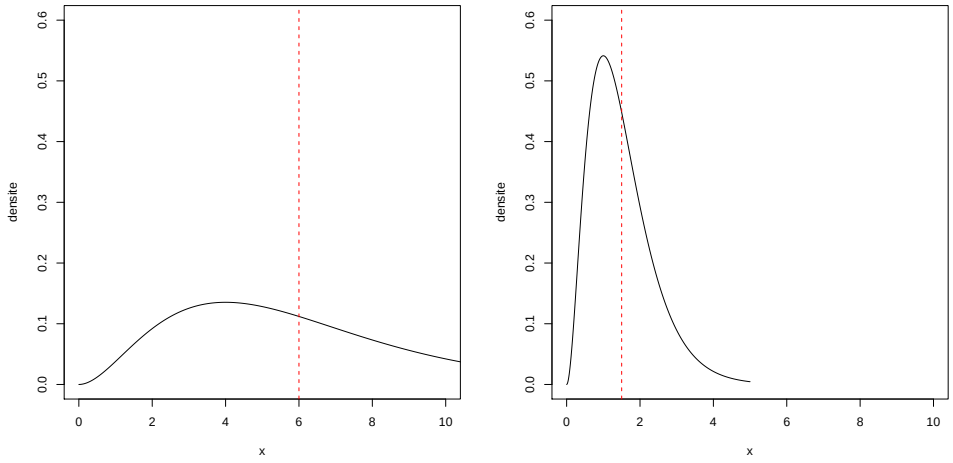


FIGURE 8.9 – Lois gamma $\mathcal{G}(3, 0.5) = \chi^2_6$ et $\mathcal{G}(3, 2)$

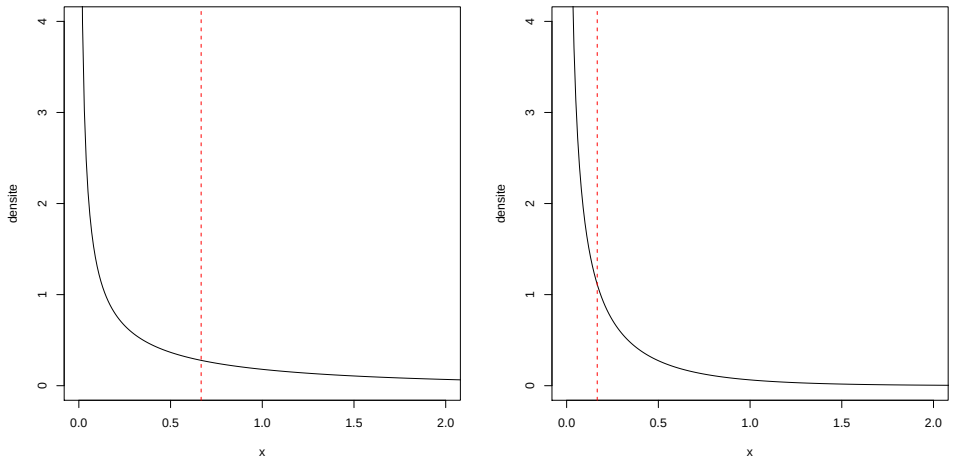


FIGURE 8.10 – Lois gamma $\mathcal{G}(1/3, 0.5) = \chi^2_{1/6}$ et $\mathcal{G}(1/3, 2)$

Loi de Student

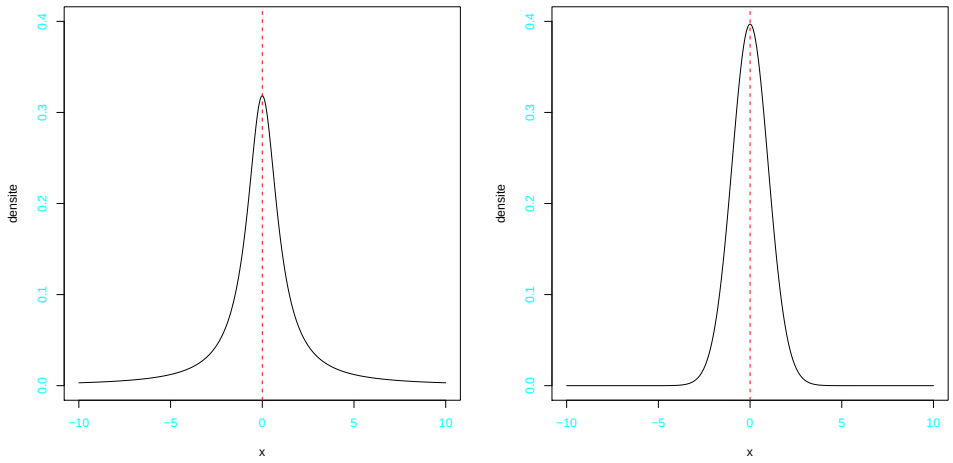


FIGURE 8.11 – Lois de Student $St(1)$ et $St(50)$

Loi de Fisher-Snedecor

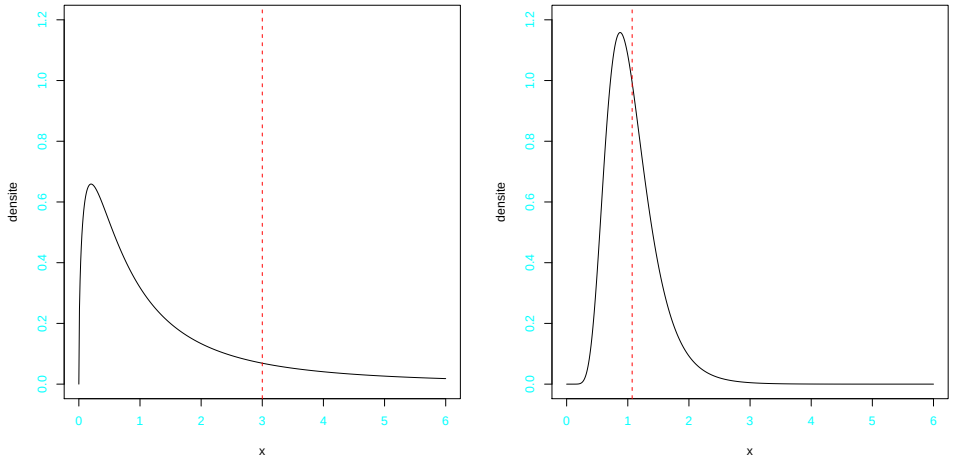


FIGURE 8.12 – Lois de Fisher-Snedecor $F(3,3)$ et $F(30,30)$

Loi de Weibull

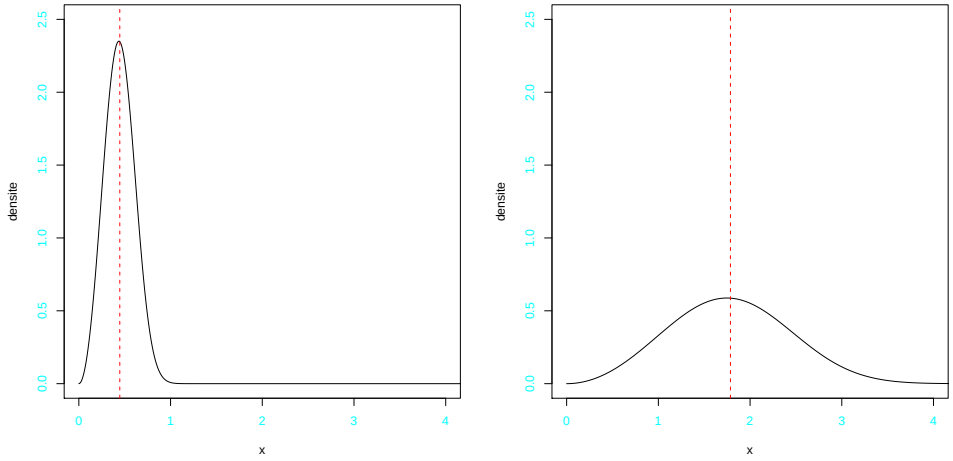


FIGURE 8.13 – Lois de Weibull $\mathcal{W}(0.5, 3)$ et $\mathcal{W}(2, 3)$

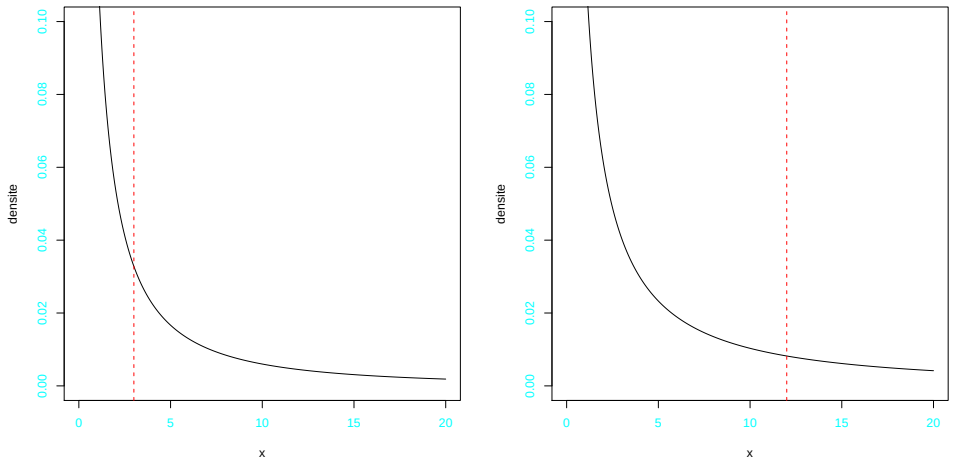


FIGURE 8.14 – Lois de Weibull $\mathcal{W}(0.5, 1/3)$ et $\mathcal{W}(2, 1/3)$

Lois beta

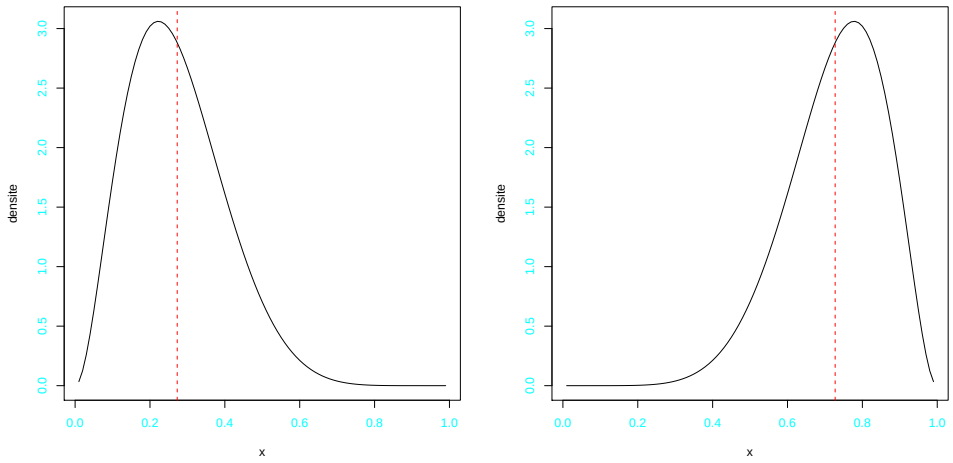


FIGURE 8.15 – Lois beta de première espèce $\beta_1(3,8)$ et $\beta_1(8,3)$

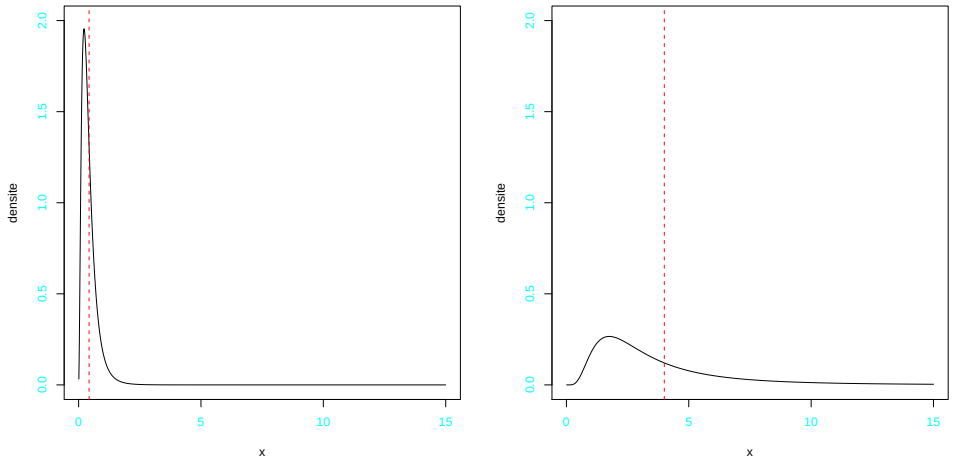


FIGURE 8.16 – Lois beta de deuxième espèce $\beta_2(3,8)$ et $\beta_2(8,3)$

Chapitre 9

Annexe C : Introduction à R

Ce chapitre fournit une introduction élémentaire à R. Pour plus de détails, voir les liens présentés sur Chamilo.

9.1 Les bases de R

R est un logiciel libre dédié au traitement de données et leur analyse statistique. Il contient une collection d'outils pour la statistique, un environnement graphique et un langage de programmation orienté objet.

Nous utiliserons R par le biais de l'environnement de développement RStudio. Un de ses principaux intérêts est l'utilisation du langage de balise RMarkdown (Rmd), qui permet la création de cahiers de travail (notebooks). Les notebooks sont des interfaces de programmation interactive, intégrant à la fois du texte en LaTeX et du code R, exportables en html et pdf. Les compte-rendus de TP et la partie R de l'examen seront rédigés au format Rmd.

Les objets créés en R (données, résultats, fonctions,...) sont stockées dans le répertoire .RData créé par défaut. Le résultat d'une procédure statistique peut être ainsi réutilisé lors de différentes sessions. Il est donc important de créer un répertoire pour chaque projet statistique effectué en R.

L'historique d'une session R est conservé dans le fichier .Rhistory.

Techniquement, R est un langage fonctionnel. Les commandes élémentaires sont constituées d'expressions et d'affectations. Par exemple :

```
> 2 + 5
[1] 7
> a <- c(9,3,7,5)
> a
[1] 9 3 7 5
> a + 3
[1] 12 6 10 8
> a[2:4]
[1] 3 7 5
```

```
> a>6
[1] TRUE FALSE TRUE FALSE
> a[a>6]
[1] 9 7
> which(a>6)
[1] 1 3
```

R peut être complété en écrivant de nouvelles fonctions. Voici un exemple où l'on souhaite calculer la statistique $\texttt{stat.log}(x_1, \ldots, x_n) = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i$ où $\forall i, x_i > 0$. On pourra définir une fonction de la façon suivante :

```
> stat.log <- function(x)
{
n <- length(x)
s <- 0
for (i in (1:n)) { s <- s + log(x[i]) }
-s/n
}
```

La fonction `stat.log` pourra être désormais utilisée comme une fonction standard de R. Par exemple :

```
> stat.log(a)
[1] -1.712796
```

Les propriétés de R permettent en fait de faire le même calcul beaucoup plus simplement :

```
> -mean(log(a))
[1] -1.712796
```

9.2 Commandes pour les deux premiers TD en R

Pour tracer un histogramme des données `x` dont l'aire est égale à 1, les bornes des classes sont données par le vecteur `bornes`, et les plages de valeurs des abscisses par le vecteur `xlim` :

```
histx <- hist(x, prob=T, breaks=bornes, xlim=xlim, ...)
```

Pour un histogramme à classes de même effectif, les bornes des classes peuvent être calculées comme des quantiles empiriques, à l'aide d'une commande du type :

```
breaks <- c(a0, quantile(x,seq(1,k-1)/k),ak)
```

La droite de régression linéaire sur le nuage des points d'abcisses `abs` et d'ordonnées `ord` est obtenue à l'aide de :

```
reg <- lm(ord~abs)
```

La pente de la droite des moindres carrés est donnée par `reg$coefficient[2]` et

l'ordonnée à l'origine par `reg$coefficient[1]`.

Pour tracer la droite obtenue, l'une des commandes suivantes pourra être utilisée : `lines(abs, fitted.values(reg))` ou `abline(reg)`.

9.3 Quelques commandes utiles de R

`help(mean)`	aide sur la commande `mean`
`x <- c(3,14,15,9)`	crée un vecteur ligne $x = (3, 14, 15, 9)$
`n <- length(x)`	taille du vecteur x
`sum(x^2)`	$\sum_i x_i^2$
`mean(x)`	moyenne empirique de l'échantillon x
`round(x)`	valeurs de x arrondies à l'entier le plus proche
`seq(from=1,to=10,by=2)`	séquence ($1 + 2k$; k entier, $1 + 2k \leq 10$)
`rep(x,3)`	concaténation de 3 répliques du vecteur x
`solve(a,b)`	solution du système linéaire $ax = b$
`diag(x)`	matrice diagonale de diagonale x
`var(x)`	variance estimée ${s'_n}^2$
`sqrt(x)`	racine carrée de x, élément par élément.
`summary(x)`	moyenne, médiane, quartiles et valeurs extrêmes
`hist(x)`	histogramme de x
`sort(x)`	tri de x par valeurs croissantes
`qqnorm(x)`	graphe de probabilités pour la loi normale
`plot(x,y)`	trace le nuage de points $\{(x_i, y_i)\}_i$
`abline(b,a)`	superpose au graphique précédent la droite d'équation $y = ax + b$
`points(x,z)`	superpose au graphique précédent le nuage de points $\{(x_i, z_i)\}_i$
`lines(x,z)`	superpose au graphique précédent la ligne polygonale reliant les points $\{(x_i, z_i)\}_i$
`lm(y~x)`	régression linéaire de y sur x
`lm(y~x)$coefficients[2]`	pente de la droite de régression
`lm(y~x)$coefficients[1]`	ordonnée à l'origine de la droite de régression
`lines(x,fitted.values(lm(y~x))`	superpose au graphique précédent la droite de régression
`postscript("nom.eps")`	redirection de la sortie graphique vers le fichier nom.eps
`dev.off()`	termine la redirection graphique vers un fichier
`par()`	contrôle des paramètres graphiques
`par(mfcol=c(2,1))`	graphique à 2 lignes et 1 colonnes
`cat("bonjour","\ n")`	imprime à l'écran le mot `bonjour` et retourne à la ligne
`source("nom.R")`	charge les commandes `R` contenues dans le fichier nom.R dans `R`
`if, else`	structure de contrôle ou d'itération
`for, while, repeat`	...

9.4 Lois de probabilité usuelles en R

Toutes les lois de probabilité usuelles ont été implémentées en R. Chaque loi est identifiée par une abréviation :

- loi binomiale : `binom`
- loi de Poisson : `pois`
- loi géométrique : `geom`. Attention, la commande `geom` concerne en fait la loi de $X - 1$, où X est de loi géométrique.
- loi exponentielle : `exp`
- loi gamma : `gamma`
- loi du chi 2 : `chisq`
- loi normale : `norm`.
- loi de Student : `t`
- loi de Fisher-Snedecor : `f`
- Loi uniforme : `unif`
- Loi beta de première espèce : `beta`
- Loi de Cauchy : `cauchy`
- Loi hypergéométrique : `hyper`
- Loi log-normale : `lnorm`
- Loi logistique : `logis`
- Loi négative binomiale : `nbinom`
- Loi de Weibull : `weibull`
- Loi de Wilcoxon : `wilcox`

Pour chaque loi, 4 fonctions sont disponibles, identifiées par un préfixe :

- Probabilités élémentaires pour les v.a.d. ou densité pour les v.a.c. : `d`
- Fonction de répartition : `p`
- Quantiles : `q`
- Simulation : `r`

Une commande R pour une loi de probabilité est constituée d'un préfixe suivi de l'abréviation de la loi. Les paramètres dépendent de la loi choisie.

Exemples :

- `pnorm(u)` donne la fonction de répartition de la loi normale centrée-réduite $\mathcal{N}(0,1)$ au point u, $\phi(u)$. On retrouve la table 1 de la loi normale.

```
> pnorm(0.61)
[1] 0.7290691
```

- `dnorm(x, m, σ)` donne la densité de la loi normale $\mathcal{N}(m, \sigma^2)$ (et non pas $\mathcal{N}(m, \sigma)$) au point x.

```
> dnorm(1.2,2,5)
```

```
[1] 0.07877367
```

- `qnorm(p)` donne le quantile d'ordre p de la loi $\mathcal{N}(0,1)$, $\phi^{-1}(p)$. On retrouve la table 2 de la loi normale en prenant $p = 1 - \alpha/2$.

```
>  qnorm(1-0.05/2)
[1] 1.959964
```

- `rnorm(n, m, σ)` simule un échantillon de taille n de la loi $\mathcal{N}(m, \sigma^2)$.

```
>  rnorm(10, 20, 1)
[1] 21.63128 20.16724 17.21667 18.76593 20.48102 20.46236 20.41822
[8] 19.91344 21.19312 19.89164
```

- `dbinom(k, n, p)` donne $P(K = k)$ quand K est de loi binomiale $\mathcal{B}(n, p)$.

```
>  dbinom(3,5,0.2)
[1] 0.0512
```

- `rpois(n, λ)` simule un échantillon de taille n de la loi de Poisson $\mathcal{P}(\lambda)$.

```
>  rpois(15,4)
[1] 8 3 2 1 6 6 7 5 3 3 4 4 6 1 1
```

- `qchisq(p,n)` donne le quantile d'ordre p de la loi du chi 2 χ^2_n. On retrouve la table de la loi du chi 2 en prenant $p = 1 - \alpha$.

```
>  qchisq(1-0.05,20)
[1] 31.41043
```

- `qt(p,n)` donne le quantile d'ordre p de la loi de Student $St(n)$. On retrouve la table de la loi de Student en prenant $p = 1 - \alpha/2$.

```
>  qt(1-0.3/2,12)
[1] 1.083211
```

- `qf(p,ν1,ν2)` donne le quantile d'ordre p de la loi de Fisher-Snedecor $F(\nu_1, \nu_2)$. On retrouve la table de la loi de Fisher-Snedecor en prenant $p = 1 - \alpha$.

```
>  qf(1-0.05,8,22)
[1] 2.396503
```

9.5 Principaux tests d'hypothèses en R

`t.test(x,...)`	test de Student sur l'espérance d'une loi normale
`binom.test()`	test sur une proportion
`var.test(x,y,...)`	test de Fisher sur la variance de 2 échantillons gaussiens indépendants
`t.test(x,y,...)`	test de Student sur l'espérance de 2 échantillons gaussiens indépendants
`prop.test()`	test de comparaison de proportions
`chisq.test(x,...)`	test du χ^2 sur les probabilités d'évènements et tables de contingence
`ks.test(x,...)`	test de Kolmogorov-Smirnov sur un ou deux échantillons
`wilcox.test(x,...)`	test de Wilcoxon-Mann-Whitney sur un ou deux échantillons

9.6 Graphiques dans R

9.6.1 Graphique simple

Le script suivant en R permet de tracer un nuage de 100 points dont les coordonnées sont des variables aléatoires indépendantes et de même loi normale centrée-réduite $\mathcal{N}(0,1)$, et de le sauvegarder au format postscript dans le fichier "rnorm.ps".

```
postscript("rnorm.ps")
plot(rnorm(100),rnorm(100))
dev.off()
```

Les instructions suivantes permettent d'insérer cette figure dans un document Latex et de pouvoir la référencer sous le nom de figure 9.1.

```
\begin{figure}[htbp]
\begin{center}
  % Requires \usepackage{graphicx}
  \includegraphics[width=8 cm, angle=270]{rnorm.ps}\\
  \caption{{\it  Utilisation de rnorm}}\label{rnorm}
  \end{center}
\end{figure}
```

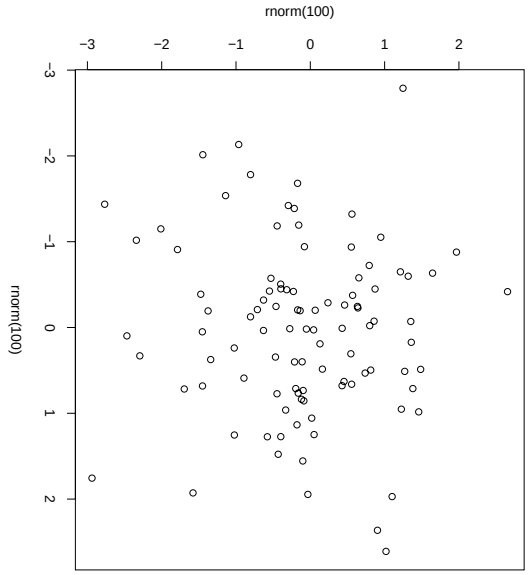


FIGURE 9.1 – *Utilisation de rnorm*

9.6.2 Autres fonctions graphiques

`abline(h=u)`	ajoute une droite d'équation y=u.
`abline(v=u)`	ajoute une droite d'équation x=u.
`legend(x,y,legend,...)`	ajoute une légende d'utilisation très flexible
`text(x,y,labels,...)`	ajoute du texte dans un graphe
`axis(side,at, labels..)`	ajoute un axe au graphique
`arrows(x0,y0,x1,y1,...)`	dessine des flèches
`symbols(x,y,....)`	dessine des cercles, des carrés, ...
`box(...)`	ajoute une boite
`polygon(x,y)`	ajoute un polygone

voir aussi `image(),pairs(),persp()`,...

9.6.3 Paramétrage de la commande plot

Le script suivant :

```
postscript("graphesR.ps")
x<- seq(-2*pi,2*pi,0.05)
y <- sin(x)
par(mfrow=c(2,2))
plot(x,y,xlab=expression(lambda),ylab=expression(plain(sin) * lambda))
plot(x,y,type="l", main="trait continu")
plot(x[seq(5,1000,by=5)],y[seq(5,1000,by=5)], type="b",axes=F)
plot(x,y,type="n", ylim=c(-2,1))
text(0,0.05,"Divers paramétrages de la fonction plot")
dev.off()
```

permet d'obtenir la figure 9.2.

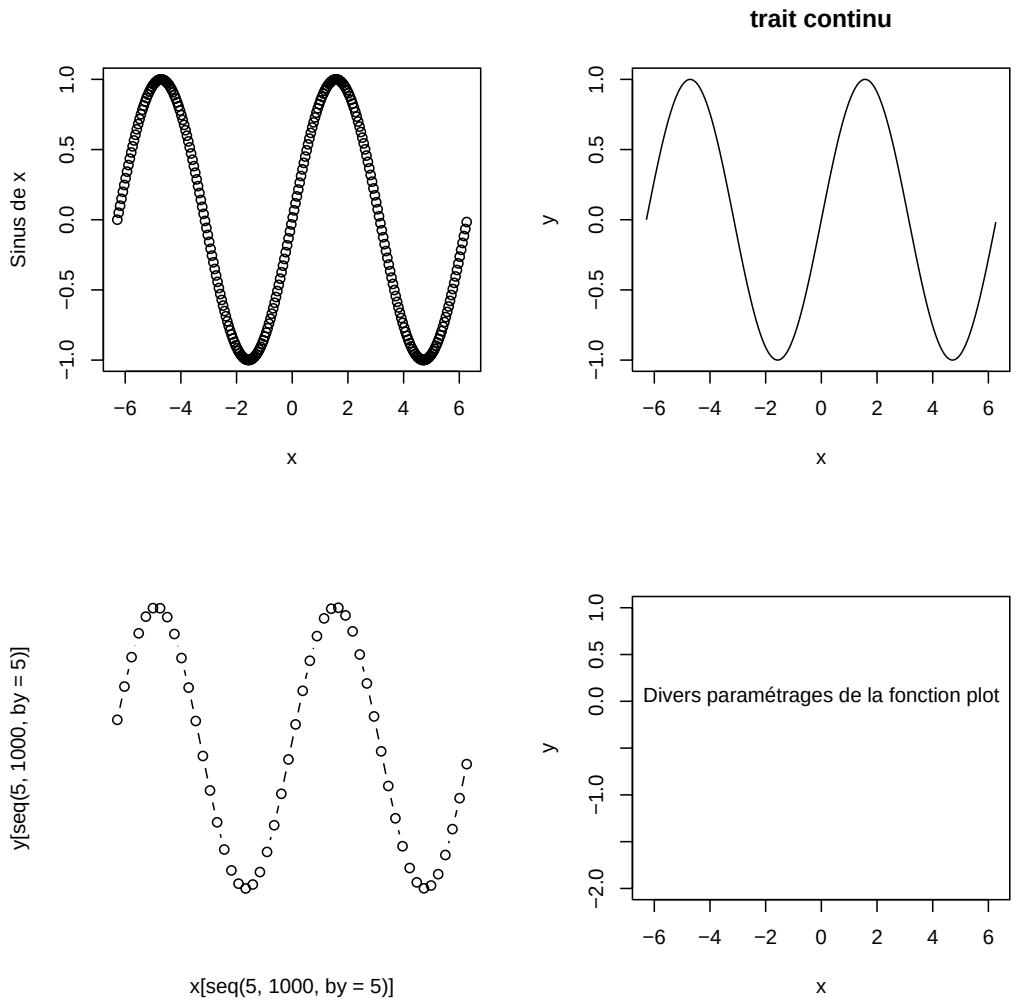


FIGURE 9.2 – R permet de créer plusieurs types de graphiques

Bibliographie

[1] DAGNELIE P., *Statistique théorique et appliquée*, 3ème édition, De Boeck Université, 2015.

[2] DALGAARD P., *Introductory Statistics with R*, 2ème édition, Springer, 2008.

[3] LAFAYE DE MICHEAUX P., DROUILHET R. LIQUET B., *Le logiciel R*, 2ème édition, Springer, 2014.

[4] MONTGOMERY D.C., RUNGER G.C., *Applied Statistics and Probability for Engineers*, 5ème édition, Wiley, 2013.

[5] MORGENTHALER S., *Introduction à la statistique*, 3ème édition, Presses polytechniques et universitaires romandes, 2007.

[6] SAPORTA G., *Probabilités, analyse de données et statistique*, 3ème édition, Technip, 2011.