Cours de Statistique et Probabilités

FC

Contents

Chapter 1

Préambule

Ce support de cours est largement inspiré du livre de Lafaye de Micheaux et al. **?** ainsi que du cours de Statistique de Bernard Ycart de l'Université de Grenoble Alpes.

Chapter 2

Décrire une ou plusieurs séries de valeurs

2.1 Les graphiques

Si la série de valeurs est qualitative on fera un diagramme en barres : voir un exemple avec la figure 1.

Si la série de valeurs est quantitative :

- valeurs discrètes : diagramme en bâtons (voir figure 2)
- valeurs continues : histogramme (voir figure 3)

Pour les variables quantitatives, on peut aussi représenter la fonction de répartition (empirique) notée $\hat{F}(x)$: pour cela on calcule pour chaque point de l'axe des x ainsi : (voir exemple figure 4)

$$\hat{F}(x) = \frac{\text{nombre de valeurs dans la série} \leqslant x}{n}$$

2.2 Résumés numériques

2.2.1 Résumés de position d'une distribution

2.2.1.1 Le ou les modes

Les modes sont les valeurs de la variable X qui apparaissent le plus fréquemment. Il peuvent se calculer pour une variable de n'importe quel type, bien que pour une variable continue, on parle de classe modale.

2.2.1.2 La médiane

La médiane d'une série statistique est la valeur m_e de la variable X qui partage cette série statistique en deux parties (inférieure et supérieure à m_e) de même effectif. Cette quantité ne se calcule pas sur des variables purement qualitatives. Pour la calculer, on distingue deux cas :

- L'effectif total N est impair, alors m_e est la valeur située à la position $\frac{N+1}{2}$
- L'effectif total N est pair, alors m_e est n'importe quelle valeur entre $\frac{N}{2}$ et $\frac{N}{2}+1$.

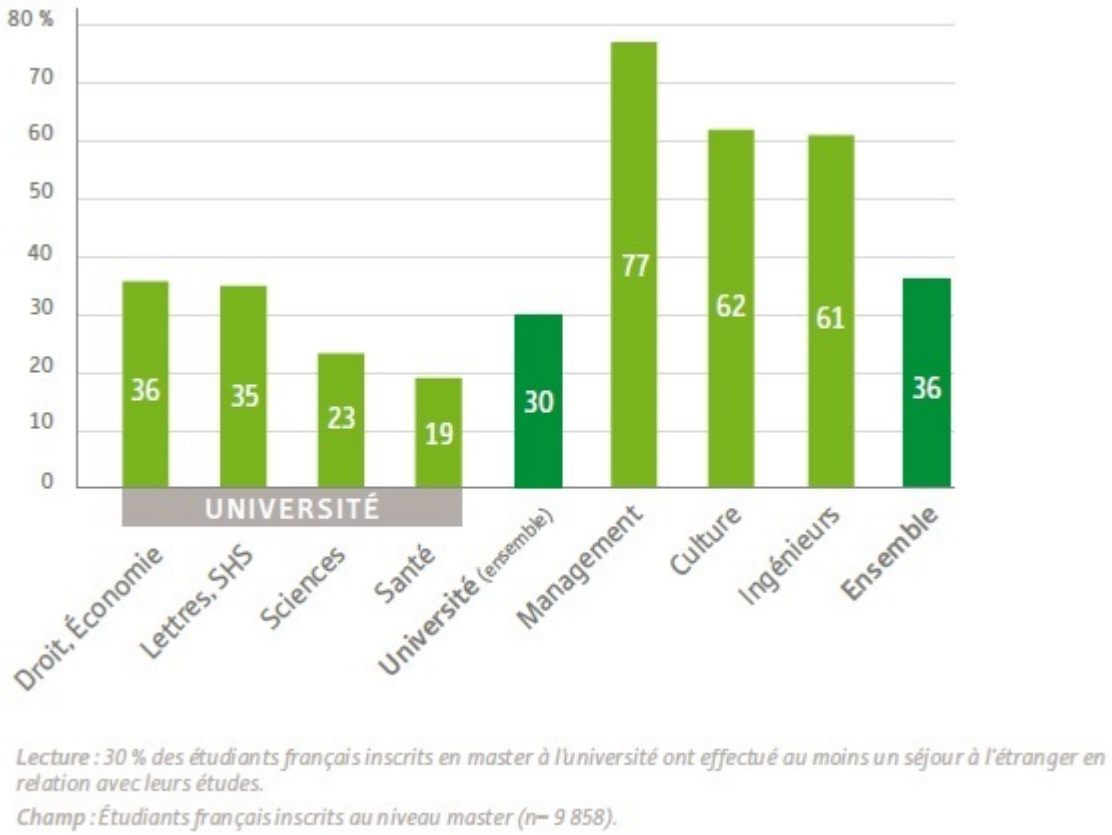


Figure 2.1: figure 1 : Diagramme en barres (source : Observatoire le vie étudiante)

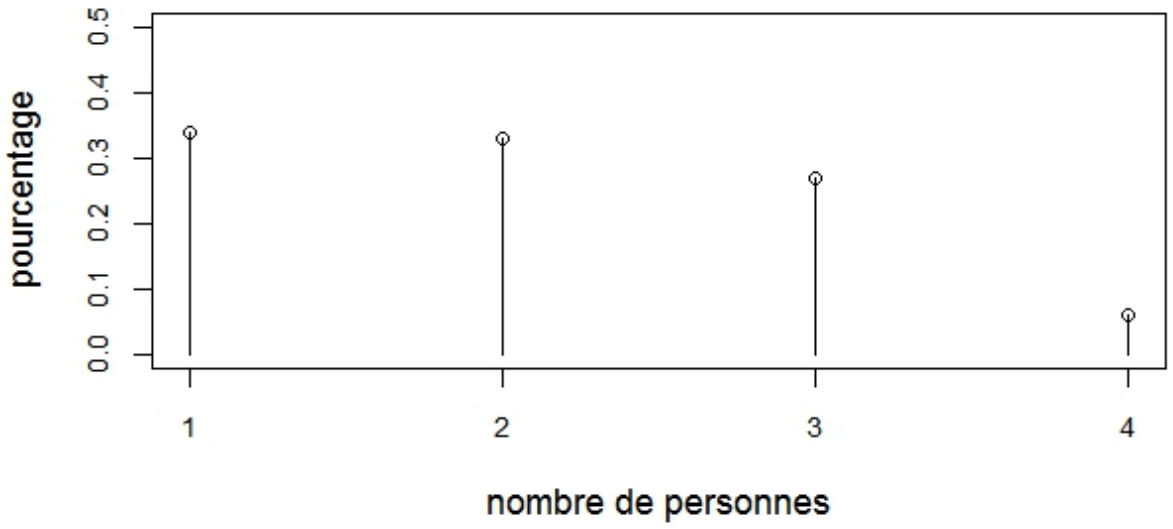


Figure 2.2: figure 2 : Diagramme en bâtons : nombre de personnes par ménage en Rhône-Alpes au 01/01/2011 (source : INSEE)

2.2.1.3 La moyenne

Elle se calcule uniquement sur des variables quantitatives via la fonction mean().

2.2.1.4 Les fractiles

Le fractile d'ordre p $(0 < p < 1)$ est la valeur q_p de la variable X qui coupe l'échantillon en deux portions, l'une ayant un nombre d'éléments (inférieurs à q_p) égal à $p\%$ du nombre total d'éléments et l'autre à $(1-p)\%$ étant supérieurs à q_p. Il ne se calcule pas pour des variables purement qualitatives. Si on prend $p = 0.5$, on retrouve la définition de la médiane.

2.2.2 Résumé de dispersion d'une distribution

Ces résumés peuvent être calculés uniquement pour des variables quantitatives. Les principales sont :

- Variance σ^2 de la population.
- l'écart type est la racine carrée de la variance.
- Coefficient de variation $c_v = \frac{\sigma}{\mu}$

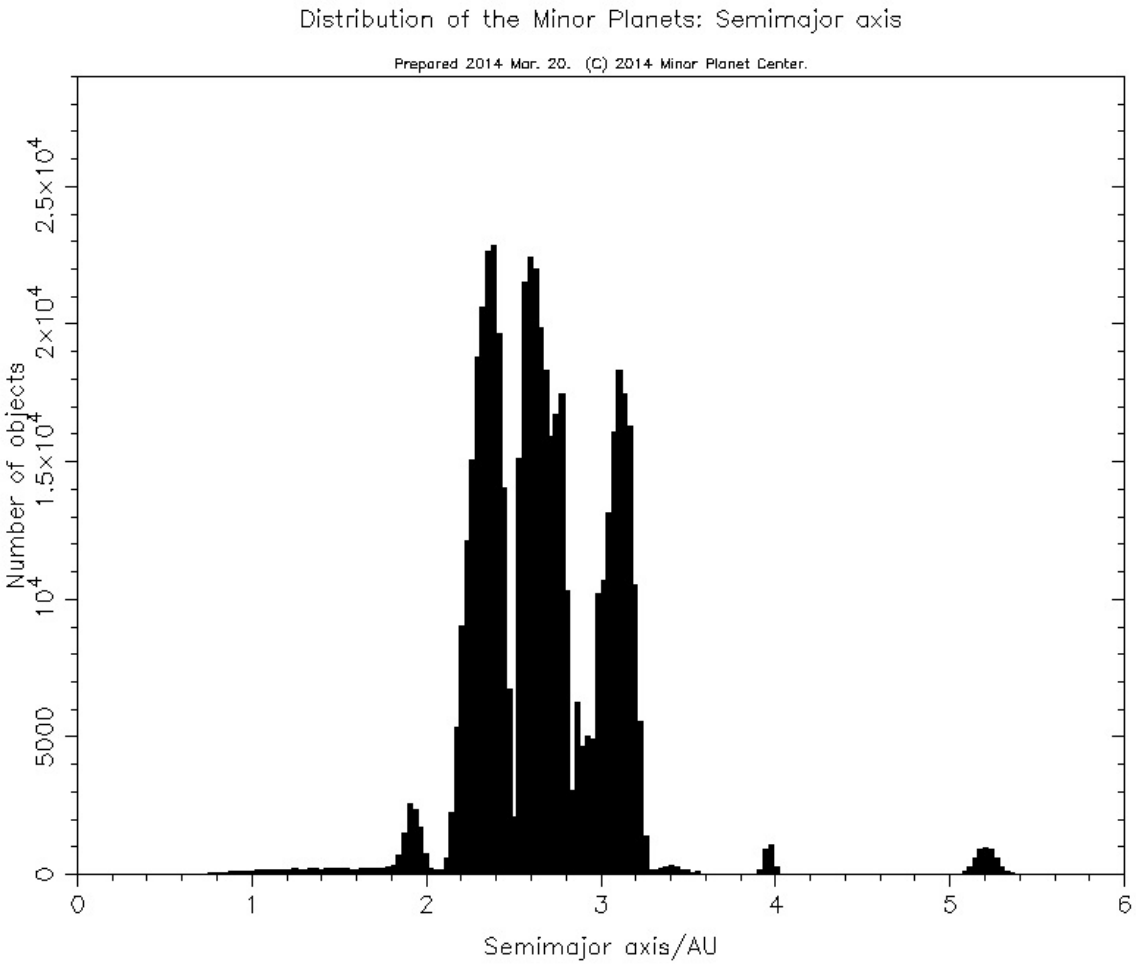


Figure 2.3: figure 3 : Histogramme : demi grands axes des orbites d'astéroïdes (source : Minor Planet Center)

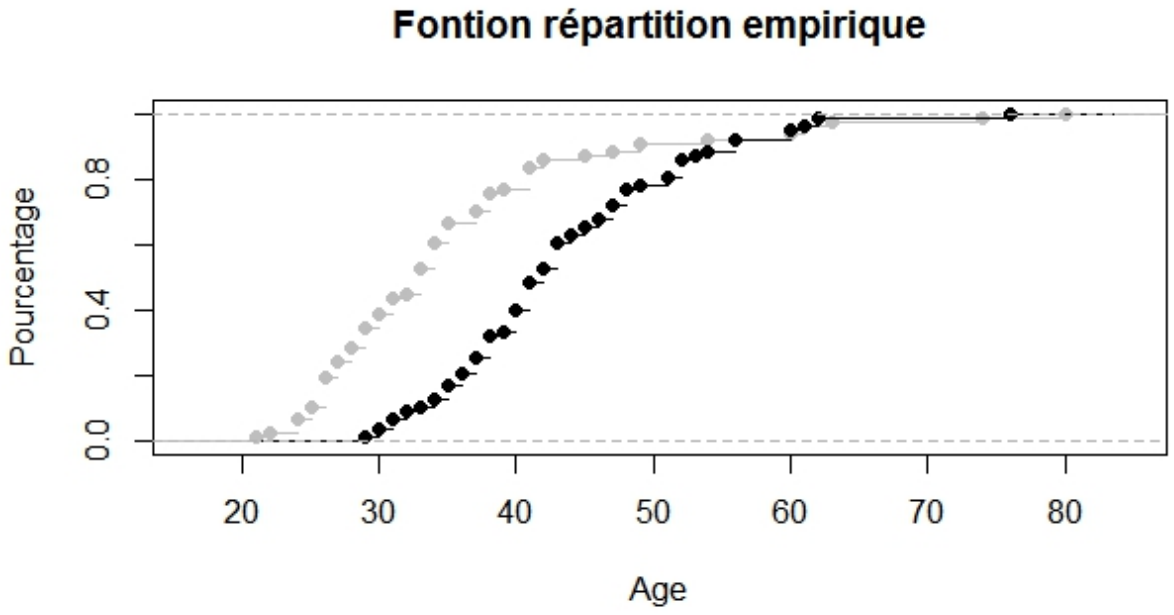


Figure 2.4: figure 4 : Fonction de répartition des âges des acteurs (en noir) et actrices (en gris) ayant reçu l'oscar du meilleur acteur depuis 1929 (source : Journal of Statistics Education)

Chapter 3

Probabilités

3.1 Densité et fonction de répartition d'une variable quantitative continue

La variable aléatoire X associée à une fonction f donnée et définie sur $\mathbb{R}$ représente le fait de tirer un nombre au hasard avec la probabilité suivante :

$$\text{Proba}(X \leq t) = \int_{-\infty}^{t} f(x)dx = F(t)$$

où t est un réel fixé.

Naturellement, cette écriture n'a de sens que si :

1. f est une fonction positive sur $\mathbb{R}$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1.$

f est appelée "**densité**" de X.

Cette probabilité est notée $F(t)$: F, vue comme une fonction de t définie sur $\mathbb{R}$, est appelée **fonction de répartition** de X. La valeur $F(t)$ peut être vue comme l'aire de la surface délimitée par la demi-droite $]-\infty, t]$, la droite $y = t$ et la courbe représentative de f.

- L'espérance de X (appelée aussi moyenne de X) correspond à la valeur suivante :

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\,dx.$$

- La variance de X est :

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - \mathbb{E}[X]^2.$$

3.2 Exemple : la loi normale

On appelle loi normale la loi d'une variable aléatoire réelle continue X dont la densité s'écrit :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

où μ est la moyenne de X et σ^2 est la variance de X. On dit que X suit la loi Normale de moyenne μ et de variance σ^2 et on note $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$.

Si $\mu = 0$, on dit que X est centrée.

Si $\sigma^2 = 1$, on dit que X est réduite.

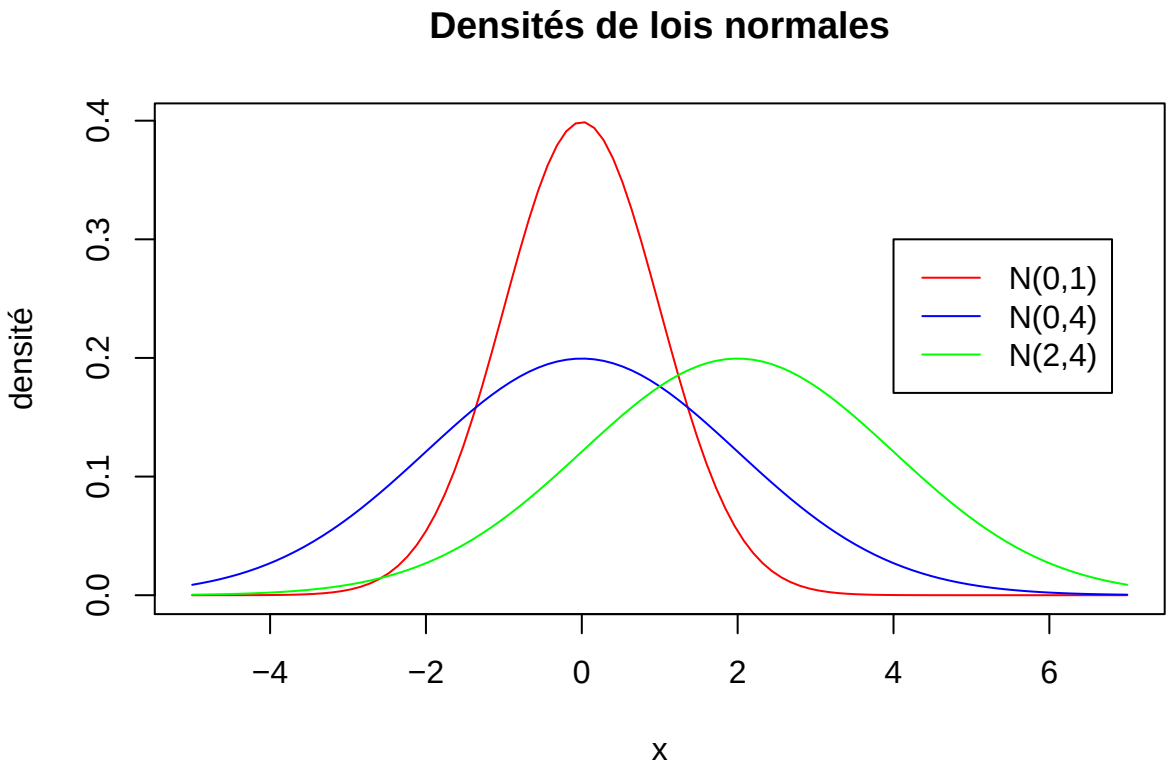
Une propriété importante sur la loi normale est la suivante :

Theorem 3.1. *Si* $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ *alors* $\dfrac{X - \mu}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$.

Remarque : Attention ! Cette propriété nous dit que pour centrer et réduire une loi normale, il faut lui retrancher sa moyenne et **diviser par l'écart type (et non pas la variance)**.

Les figures suivantes nous donnent des exemples de densité de différentes lois normales.

```
curve(dnorm(x,0,1),from = -5,to = 7,col="red",main="Densités de lois normales",ylab="densité")
curve(dnorm(x,0,2),from = -5,to = 7,col="blue",add = TRUE)
curve(dnorm(x,2,2),from = -5,to = 7,col="green",add = TRUE)
legend(4,0.3,legend=c("N(0,1)","N(0,4)","N(2,4)"),col = c("red","blue","green"),lty=1)
```



3.3 Loi forte des grands nombres

3.4 Théorème central limite

Le théorème central limite dit que, si un grand nombre de variables aléatoires indépendantes ayant la même loi sont ajoutées, leur somme suit approximativement une loi normale.

1. Pour des échantillons distribués suivant une loi binomiale, la loi binomiale $B(n, p)$ se comporte comme la loi normale $\mathcal{N}(np, np(1-p))$ pour n grand.

2. Si $\{X_i\}_{i=1}^{\infty}$ est une suite de variables aléatoires indépendantes de même loi et de moyenne μ et de variance σ^2, alors $\bar{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$ suit approximativement une loi normale $\mathcal{N}(\mu, \sigma^2/n)$ pour n grand
3. ou, $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$, variable centrée réduite issue de $\bar{X}_n$, suit approximativement une loi normale $\mathcal{N}(0,1)$ pour n grand.

Chapter 4

Estimation

4.1 Introduction

En probabilités, on travaille avec une loi connue. En statistique, cette loi est inconnue.

Le statisticien travaille sur des données (notes de qualité de pièces produites dans une usine, données météorologiques, résultats d'expériences médicales ou physiques,...). Il le fait à la demande d'un interlocuteur qui a des attentes plus ou moins précises. Ces attentes peuvent être de plusieurs types :

- extraire des résumés pertinents des données,
- répondre à une question comme "le réchauffement climatique est-il réel ?",
- prendre une décision comme la mise sur le marché d'un nouveau médicament,
- effectuer une prévision, par exemple sur le résultat d'une élection qui aura lieu prochainement,...

Il élabore un modèle et construit des outils pour répondre aux questions de son interlocuteur dans ce modèle. Il doit bien sûr garder un sens critique vis à vis du modèle qu'il a construit. Il est bien sûr crucial pour le statisticien d'estimer les paramètres au vu des données dont il dispose et d'avoir une idée de la précision de cette estimation. On introduit tout d'abord les estimateurs puis on verra enfin comment évaluer la précision des estimateurs au travers d'intervalles de confiance.

En résumé, voici les étapes de la statistique inférentielle :

1. Observation d'une variable X sur un groupe d'individus choisis d'une façon aléatoire et indépendante dans la population totale.
2. On obtient des observations $x_1, \ldots, x_n$, réalisations de variables aléatoires indépendantes et de même loi $X_1, \ldots, X_n$. On fait une étude descriptive de $x_1, \ldots, x_n$ (histogramme, moyenne, ...).
3. Au vu de l'étude descriptive, trouver une loi de probabilité acceptable pour les variables $X_1, \ldots, X_n$.
4. Inférence statistique : utiliser $x_1, \ldots, x_n$ pour estimer les paramètres du modèle et en déduire des propriétés sur la population totale.

4.2 Statistique et estimateur

1. Pour un paramètre inconnu, un estimateur est une fonction des données, prenant des valeurs proches de ce paramètre.

1 Avant que les données ne soient collectées, l'estimateur est une variable aléatoire 2 Une fois les données collectées, l'estimation est la valeur de l'estimateur pour ces données.

2. Estimer un paramètre θ inconnu, c'est donc trouver une statistique $T = \tau(X_1, \ldots, X_n)$ dont on pense que la valeur observée $\tau(x_1, \ldots, x_n)$ sera probablement "suffisamment proche" de la valeur inconnue θ.

Dans ce cas, T sera appelé estimateur de θ, et, $\tau(x_1, \ldots, x_n)$ sera une estimation de θ. (valeur numérique).

1. Le biais de T est la différence entre l'espérance de T et la vraie valeur (inconnue) de θ : Biais=$\mathbb{E}[T] - \theta$.
2. L'erreur quadratique est l'espérance des carrés des différences : QE= $\mathbb{E}[(T-\theta)^2]$.

L'estimateur T est :

- sans biais si le biais est nul (Les valeurs de T sont centrées sur la vraie valeur)
- asymptotiquement sans biais si le biais tend vers 0 quand la taille de l'échantillon tend vers l'infini.
- consistant si la probabilité de s'éloigner de la valeur à estimer de plus de ϵ (ϵ petit) tend vers 0 quand la taille de l'échantillon augmente.

Voici maintenant quelques exemples standards d'estimateurs

1. La fréquence empirique d'un évènement est un estimateur sans biais consistant de la probabilité de cet évènement.
2. La moyenne empirique d'un échantillon est un estimateur sans biais consistant de l'espérance théorique de ces variables :
$$T(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
3. La variance empirique notée S_n^2 d'un échantillon (lorsque la moyenne est inconnue) est
$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Cet estimateur est biaisé. On peut montrer que

$$\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$$

Ainsi, on obtient un estimateur sans biais en multipliant la variance empirique par $n/(n-1)$ où n désigne la taille de l'échantillon, noté ${S'_n}^2$:

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

C'est cette dernière quantité qui est donnée dans le logiciel R via la fonction var(). Si l'on veut calculer la variance empirique d'un échantillon sous le logiciel R, il faudra donc faire le nécessaire : par exemple faire une nouvelle fonction que l'on pourra appeler var.pop().

4.3 Estimation par intervalle de confiance

Lorsque l'on estime un paramètre θ, on veut avoir une idée de la précision de l'estimation effectuée. C'est le rôle des intervalles de confiance.

Problème :

Peut-on trouver deux statistiques T_1 et T_2 telles que

$$p(T_1 \leq \theta \leq T_2) = 1 - \alpha$$

avec $0 < \alpha < 1$ fixé ? ou encore peut-on trouver deux statistiques T_1 et T_2 de manière à ce qu'on ait beaucoup de chance de trouver le paramètre inconnu entre ces deux statistiques ?

1. L'intervalle $[T_1, T_2]$ est un intervalle aléatoire appelé intervalle de confiance.

2. α est le risque d'erreur. Le paramètre α représente la probabilité que l'intervalle $[T_1(X_1,\dots,X_n), T_2(X_1,\dots,X_n)]$ ne contienne pas le paramètre inconnu θ. En affirmant que $[T_1, T_2]$ contient θ, on se trompe en moyenne 100α fois sur 100.
3. $(1-\alpha)$ est appelé niveau de confiance ou coefficient de sécurité.

4.3.1 Intervalles de confiance pour une moyenne

4.3.1.1 Cas d'un échantillon gaussien

On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. On rappelle que la moyenne empirique et que la variance empirique sont données par

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

1. Si σ^2 est connue, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

   où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

2. Si σ^2 est inconnue, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre $n-1$.

4.3.1.2 Cas d'un échantillon non gaussien, mais de grande taille

Pour de grands échantillons, sans hypothèse de normalité, un intervalle de confiance de niveau $1-\alpha$ pour la moyenne μ est

$$\left[\bar{X} - u_{1-\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + u_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

4.3.2 Intervalle de confiance pour une variance

On se place dans le cas où X suit une loi normale, $\mathcal{N}(\mu, \sigma^2)$.

Un intervalle de confiance de niveau $1-\alpha$ pour la variance σ^2 est

$$\left[\frac{nS^2}{q^{n-1}_{1-\alpha/2}}; \frac{nS^2}{q^{n-1}_{\alpha/2}}\right] = \left[\frac{(n-1)(S')^2}{q^{n-1}_{1-\alpha/2}}; \frac{(n-1)(S')^2}{q^{n-1}_{\alpha/2}}\right]$$

où $q^{n-1}_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi du chi-2 de paramètre $n-1$ et $q^{n-1}_{\alpha/2}$ son quantile d'ordre $\alpha/2$.

4.3.3 Intervalle de confiance pour une proportion

On suppose que l'on est en présence d'un échantillon de grande taille (en pratique $n \geq 30$). Un intervalle de confiance de niveau $(1-\alpha)$ pour une proportion p inconnue est

$$\left[\bar{X} - u_{1-\alpha/2}\sqrt{(\frac{\hat{X}(1-\hat{X})}{n})}; \hat{X} + u_{1-\alpha/2}\sqrt{(\frac{\hat{X}(1-\hat{X})}{n})}\right].$$

où n est la taille de l'échantillon, $\bar{X}$ la fréquence empirique et $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Chapter 5

Tests

Bibliography