Statistique inférentielle

PROF. ARMEL YODÉ

Table des matières

Chapitre 1

Rappels et compléments

1.1 Loi normale

Théorème 1.1.1. *Soient $X_1,\dots,X_n$ des variables aléatoires indépendantes identiquement distribuées de loi normale $\mathcal{N}(m,\sigma^2)$ avec $\mu\in\mathbb{R}$ et $\sigma^2>0$. Posons*

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

Alors nous avons :

1. S_n^2 *et* $\overline{X}_n$ *sont indépendantes*
2. $\overline{X}_n \hookrightarrow \mathcal{N}\left(m,\frac{\sigma^2}{n}\right)$.
3. $\frac{(n-1)S_n^2}{\sigma^2} \hookrightarrow \chi^2(n-1)$ *(loi de Khi-deux à $n-1$ degrés de liberté).*
4. $\frac{\sqrt{n}\left(\overline{X}_n-m\right)}{S_n} \hookrightarrow T(n-1)$ *(loi de Student à $n-1$ degrés de liberté).*

1.2 Convergences

On considère une suite de variables aléatoires réelles $(X_n)_{n\geq 1}$ définies sur le même espace probabilisé $(\Omega,\mathcal{A},\mathbb{P})$. On note F_{X_n} la fonction de répartition de X_n et F_X celle de X.

Définition 1.2.1. *On dit que la suite $(X_n)_{n\geq 1}$ converge en loi vers la variable aléatoire X et on note $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$ si*

$$\lim_{n\to+\infty} F_{X_n}(x)=F_X(x)$$

en tout point x où F_X est continue.

Définition 1.2.2. *On dit que la suite $(X_n)_{n\geq 1}$ converge en probabilité vers un réelle a et on note $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$, si quelque soit $\varepsilon>0$,*

$$\lim_{n\to+\infty}\mathbb{P}\{|X_n-X|\geq\varepsilon\}=0.$$

Remarque 1.2.1. *La convergence en probabilité implique la convergence en loi :*

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X \Longrightarrow X_n \xrightarrow[n\to+\infty]{\mathscr{L}} X.$$

Cependant si $X = a$ où a est une constant, alors il y a équivalence entre les deux modes de convergence

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} a \Longleftrightarrow X_n \xrightarrow[n\to+\infty]{\mathscr{L}} a.$$

Proposition 1.2.1. *Soit $a \in \mathbb{R}$. Si* $\begin{cases} \mathbb{E}(X_n) \xrightarrow[n\to+\infty]{} a \\ Var(X_n) \xrightarrow[n\to+\infty]{} 0 \end{cases}$ *alors* $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} a$.

Démonstration. En appliquant l'inégalité de Markov, nous aobtenons

$$\begin{aligned} 0 \le \mathbb{P}\Big(\Big|X_n - a\Big| > \varepsilon\Big) &\le \frac{1}{\varepsilon^2}\mathbb{E}\Big(\Big|X_n - a\Big|^2\Big) \\ &= \frac{1}{\varepsilon^2}\Big(\mathbb{E}\Big(X_n - \mathbb{E}(X_n)\Big)^2 + \Big(\mathbb{E}(X_n) - a\Big)^2\Big) \\ &= \frac{1}{\varepsilon^2}\Big(Var\Big(X_n\Big) + \Big(\mathbb{E}(X_n) - a\Big)^2\Big). \end{aligned}$$

Ainsi, nous remarquons que si $\begin{cases} \mathbb{E}(X_n) \xrightarrow[n\to+\infty]{} a \\ \mathrm{Var}(X_n) \xrightarrow[n\to+\infty]{} 0 \end{cases}$, alors d'après le Théorème des gendarmes

$$\lim_{n\to+\infty} \mathbb{P}\{|X_n - X| \ge \varepsilon\} = 0.$$

□

1.3 Théorèmes limites

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes identiquement distribuées de moyenne m et de variance $\sigma^2 > 0$. Posons

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

Nous nous intéressons à deux résultats importants concernant la moyenne empirique $\overline{X}_n$ de variables aléatoires indépendantes identiquement distribuées.

1.3.1 Lois des grands nombres

Théorème 1.3.1. *Soient $(X_n)_{n\ge1}$ une suite de variables aléatoires réelles indépendantes identiquement distribuées telles que $\mathbb{E}(X_1) = m < +\infty$ et $Var(X_n) = \sigma^2$. Alors, nous avons*

$$\overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} m.$$

Démonstration. Nous utilisons la proposition 1.2.1. En effet, nous avons $\mathbb{E}(\overline{X}_n) = m$ et $Var(\overline{X}_n) = \dfrac{\sigma^2}{n}$.

□

Ce résultat signifie que lorsque n devient grand, la moyenne empirique $\overline{X}_n$ se réduit "presque" à la moyenne théorique m.

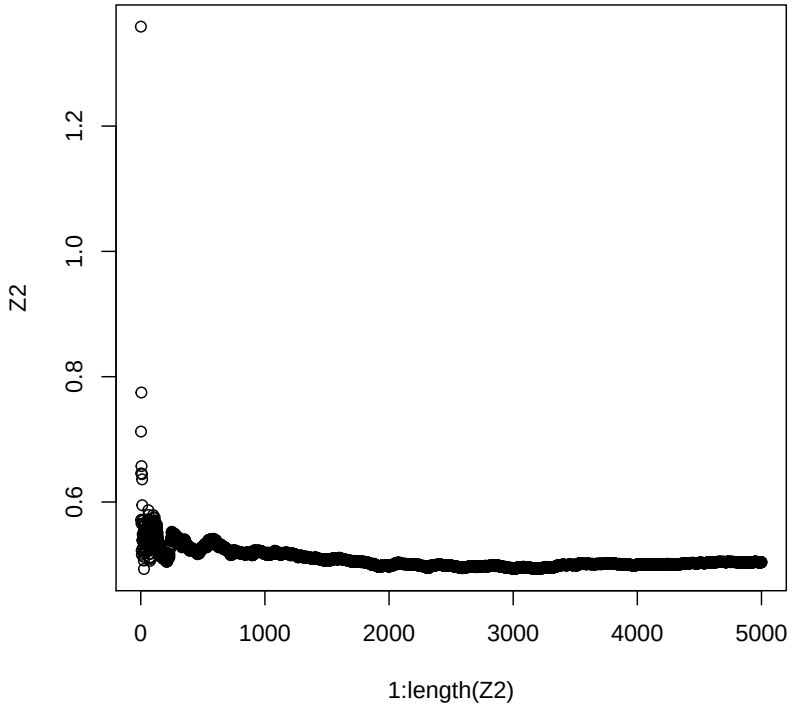
Illustration de la loi des grands nombres

Dans cet exemple $(X_n)_{n\geq 1}$ est une suite de variables aléatoires indépendantes identiquement de loi exponentielle $\mathscr{E}(2)$. D'après la loi des grands nombres

$$\overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{2}.$$

Ce résultat est illustré par le graphique ci-dessous.

```
> Z=rexp(5000,2)
> Z1=cumsum(Z)
> Z2=rep(0,5000)
> for(i in 1:5000){
+   Z2[i]=Z1[i]/i
+ }
> plot(1:length(Z2),Z2)
```



1.3.2 Théorème Central limite

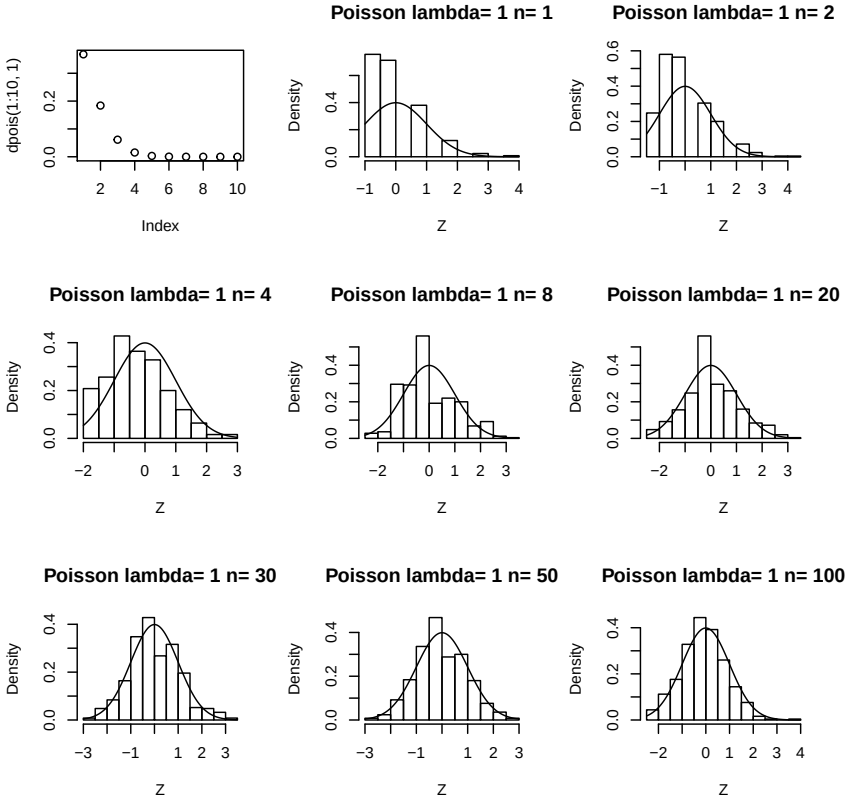
Le théorème central limite permet d'étudier la convergence en loi de la moyenne empirique $\overline{X}_n$.

Théorème 1.3.2. *Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes identiquement distribuées telles que $\mathbb{E}(X_1) = m < +\infty$ et $\sigma^2 = var(X_1) > 0$. Alors, nous avons*

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1) \quad \Longleftrightarrow \quad \sqrt{n}(\overline{X}_n - m) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,\sigma^2).$$

Illustration du théorème central limite

```
> par(mfrow=c(3,3))
> plot(dpois(1:10,1))
> n=c(1,2,4,8,20,30,50,100)
> for (i in 1:length(n))
+ {
+ X=rpois(500*n[i],1)
+ A=matrix(X,ncol=500)
+ M=apply(A,2,mean)
+ Z=sqrt(n[i])*(M-1)
+ hist(Z,freq=FALSE,main=paste("Poisson lambda=",1,"n=", n[i]))
+ curve(dnorm, add=TRUE)
+ }
```



Autrement dit, quand n est assez grand $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$ converge vers la loi normale centrée réduite $\mathcal{N}(0,1)$, c'est à dire que la moyenne empirique $\overline{X}_n$ suit approximativement une loi normale $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$. En pratique, l'approximation est fréquemment réalisée dès que $n \geq 30$.

1.3.3 Théorème de Slusky et généralisation

Théorème 1.3.3. *Soient X_n et Y_n deux suites de variables aléatoires telles que :*

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$$

$$Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} c$$

où c est une constante. Alors

$$X_n + Y_n \xrightarrow[n\to+\infty]{\mathscr{L}} X + c$$

$$X_n Y_n \xrightarrow[n\to+\infty]{\mathscr{L}} cX$$

$$\frac{X_n}{Y_n} \xrightarrow[n\to+\infty]{\mathscr{L}} \frac{X}{c} \quad si\ c \neq 0.$$

L'on peut généraliser ces résultats. Quelle condition doit vérifier une fonction g pour que $g(X_n)$ converge en loi (ou en probabilité) vers $g(X)$ dès que X_n converge en loi (ou en probabilité) ves X. Le résultat suivant permet de répondre à cette question.

Théorème 1.3.4. *Soit g est une fonction continue. Alors*

- $X_n \xrightarrow[n\to+\infty]{\mathscr{L}} X \implies g(X_n) \xrightarrow[n\to+\infty]{\mathscr{L}} g(X)$.
- $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X \implies g(X_n) \xrightarrow[n\to+\infty]{\mathbb{P}} g(X)$.

1.3.4 La méthode delta

Si

$$\sqrt{n}(Y_n - y) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, \sigma_y^2),$$

quelle est la loi asymptotique de la variable aléatoire $\sqrt{n}(g(Y_n) - g(y))$? C'est à dire,

$$\sqrt{n}(g(Y_n) - g(y)) \xrightarrow[n\to+\infty]{\mathscr{L}} \quad ?$$

Quelles sont les conditions sur la fonction g ? La méthode delta permet de répondre à ce type de préoccupations.

Théorème 1.3.5. *Si la suite de variables aléatoires (Y_n) est asymptotiquement normale, telle qu'il existe y et σ_y^2 avec*

$$\sqrt{n}(Y_n - y) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, \sigma_y^2)$$

et si g est une fonction de classe $\mathscr{C}^1$ alors $g(Y_n)$ est asymptotiquement normal

$$\sqrt{n}(g(Y_n) - g(y)) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, \sigma_y^2 (g'(y))^2).$$

Chapitre

2 Modélisation statistique

On souhaite étudier X un caractère sur une population donnée. On supposera que le caracère X est quantitatif.

2.1 Echantillonnage

Exemple 2.1.1. *Une entreprise de l'industrie textile souhaite étudier le poids et la taille des ivoiriens et ivoiriennes de plus de* 18 *ans (population) afin d'ajuster au mieux ses produits à la morphologie de ses clients.*

Pour mener à bien cette étude, l'entreprise a deux solutions : le recensement ou l'échantillonnage.

Recensement : il consiste à mesurer le caractère X de facon exhaustive pour tous les individus de la population. Le recensement n'est bien évidemment applicable que lorsque la taille de la population étudiée est relativement faible.

Dans le cas où la taille de la population est grande, il faut recourir à l'échantillonnage. L'échantillonnage se définit comme la méthode de construction d'un échantillon.

Echantillon : c'est un sous-ensemble de la population ; le nombre d'individus sélectionnés dans l'échantillon correspond à la taille de l'échantillon, notée n ; on parle alors de n-échantillon.

Quel est l'intérêt de constituer un échantillon ? L'idée est d'étudier le caractère pour les individus sélectionnés dans l'échantillon afin d'en tirer de l' information sur ce caractère pour l'ensemble de la population. Par conséquent, d'un côté la taille n de l'échantillon doit être suffisamment importante pour que l'on puisse obtenir une information fiable sur la population, mais d'un autre côté elle doit être la plus petite possible afin de limiter le coût de l'enquête.

Une question se pose alors : comment choisir les individus qui composent l'échantillon ? On distingue deux grandes méthodes d'échantillonnage. La première repose sur un choix déterministe des individus. On parle dans ce cas d'échantillon déterministe (ou certain) : les individus de l'échantillon ne sont pas choisis au hasard. En pratique la méthode la plus utilisée est celle de l'échantillonnage aléatoire.

Echantillon aléatoire : c'est un échantillon dont les individus sont tirés au hasard parmi la population. Le tirage de l'échantillon peut se faire avec remise (un même individu de la population peut apparaître plusieurs fois dans l'échantillon) ou sans remise (chaque individu de la population ne peut apparaître qu'une seule fois dans l'échantillon).
On considère deux situations différentes conduisant à un échantillon :

- la répétition d'une expérience aléatoire

  Exemple 2.1.2. *On lance n fois une pièce. On note*

  $$X_i = \begin{cases} 1 & \text{si le lancer } i \text{ est pile} \\ 0 & \text{si lancer } i \text{ est face.} \end{cases}$$

  S'il s'agit de la même pièce et qu'on ne modifie pas la manière dont on lance, alors on peut dire que les X_i sont indépendantes et identiquement distribuées de loi commune la loi de Bernoulli $\mathcal{B}(1,\theta)$. Le paramètre θ représente la probabilité du succès, c'est à dire la probabilité d'obtenir pile.

- la considération d'un échantillon au sein d'une population

  Exemple 2.1.3. *Deux candidats Kouko et Yao sont en présence d'une élection. n personnes sont tirées au hasard parmi les électeurs et interrogées sur leurs intentions de vote. On note*

  $$X_i = \begin{cases} 1 & \text{si l'individu } i \text{ vote Kouko} \\ 0 & \text{si l'individu } i \text{ vote Yao.} \end{cases}$$

  Les valeurs observées sont considérées comme étant les réalisations de variables aléatoires $X_1,\ldots,X_n$ indépendantes et identiquement distribuées selon la distribution finale des voix, c'est à dire la loi de Bernoulli $\mathcal{B}(1,\theta)$. Le paramètre θ représente la probabilité du succès, c'est à dire la probabilité de voter pour Kouko.

2.2 Modèles statistiques

Soit X une variable aléatoire réelle (discrète ou continue) dont la loi de probabilité $\mathbb{P}_\theta$ dépend d'un paramètre inconnu θ.

Définition 2.2.1. *On appelle modèle statistique la donnée d'une famille de lois de probabilité $\left\{\mathbb{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d\right\}$; Θ est appelé espace des paramètre.*

Définition 2.2.2. *Un échantillon de X de taille n est un n-uplet $(X_1,\ldots,X_n)$ de variables aléatoires indépendantes de même loi que X.*

Remarque 2.2.1. *Attention ! Il ne faut pas confondre l'échantillon aléatoire (collection de variables aléatoires indiquées par une lettre majuscule) et la réalisation de cet échantillon (notée avec des lettres minuscules) :*

$$\text{Echantillon :} \quad (X_1,\ldots,X_n)$$

$$\text{Réalisation :} \quad (x_1,\ldots,x_n)$$

Définition 2.2.3. *On appelle statistique toute variable aléatoire ne dépendant que de l'échantillon $(X_1,\ldots,X_n)$.*

Remarque 2.2.2. *Une statistique est un résumé de l'échantillon.*

La statistique inférentielle a pour objectif d'avoir des informations sur le paramètre inconnu θ en se basant sur l'échantillon $(X_1,\dots,X_n)$. On part de l'échantillon pour avoir une meilleure connaissance de la population.

Si X est une variable aléatoire réelle, alors on note :
— $f(x,\theta)$ si X est une variable aléatoire à densité
— $f(x,\theta)=\mathbb{P}_\theta(X=x)$ si X est une variable aléatoire discrète.

Exemple 2.2.1. *1. Modèle de Bernouilli :* $\left\{\mathscr{B}(1,\theta),\theta\in\Theta=]0,1[\subset\mathbb{R}\right\}$ *:*

$$f(x,\theta)=\mathbb{P}_\theta(X=x)=\theta(1-\theta)1_{\{0,1\}}(x).$$

2. Modèle gaussien : $\left\{\mathscr{N}(\mu,\sigma^2),\theta=(\mu,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+^*\subset\mathbb{R}^2\right\}$ *:*

$$f(x,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big).$$

3. Modèle exponentiel : $\left\{\mathscr{E}(\theta),\theta\in\Theta=\mathbb{R}_+^*\subset\mathbb{R}\right\}$ *:*

$$f(x,\theta)=\theta e^{-\theta x}1_{\mathbb{R}^+}(x).$$

4. Modèle de Poisson : $\left\{\mathscr{P}(\theta),\theta\in\Theta=\mathbb{R}_+^*\subset\mathbb{R}\right\}$ *:*

$$f(x,\theta)=e^{-\theta}\frac{\theta^x}{x!}1_{\mathbb{N}}(x).$$

Définition 2.2.4. *Le* ***support*** *de* $\mathbb{P}_\theta$ *est l'ensemble* $\{x:f(x,\theta)>0\}$.

Définition 2.2.5. *Si toutes les lois* $\mathbb{P}_\theta$, $\theta\in\Theta$ *ont un support commun alors le modèle est dit* ***homogène****. Cela signifie que pour chaque* $\theta\in\Theta$, $\left\{x:f(x,\theta)>0\right\}$ *ne dépend pas de* θ.

Exemple 2.2.2. *1. Le modèle de Bernouilli est un modèle homogène car son support* $\left\{0,1\right\}$ *est indépendant de* θ.

2. Le modèle uniforme $\{\mathscr{U}_{[0,\theta]},\theta>0\}$ *n'est pas homogène. En effet, la densité de la loi uniforme sur* $[0,\theta]$ *étant* $f(x,\theta)=\dfrac{1}{\theta}1_{[0,\theta]}(x)$*, son support* $[0,\theta]$ *dépendant du paramètre.*

Définition 2.2.6. *Le modèle statistique* $\{\mathbb{P}_\theta,\theta\in\Theta\}$ *est* ***identifiable*** *lorsque l'application* $\theta\longmapsto\mathbb{P}_\theta$ *est injective.*

Exercice 2.2.1. *Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes. On interroge* n *votants, n étant considéré comme petit devant le nombre total de votants, et on récolte les nombres* n_A *et* n_B *de voix pour* A *et* B *respectivement* $(n_A+n_B=n$*, en ne tenant pas compte des votes blancs ou nuls pour simplifier).*

1. *Décrire l'observation associée à cette expérience et le modèle statistique engendré par cette observation.*
2. *Montrer que le modèle statistique engendré par cette observation est identifiable. Exhiber sa vraisemblance.*

Chapitre

3 Exhaustivité

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de probabilité dépendant d'un paramètre inconnu $\theta \in \mathbb{R}$.

3.1 Vraisemblance

Définition 3.1.1. *On appelle vraisemblance d'un échantillon $(X_1,\dots,X_n)$ la fonction définie par*

$$\begin{aligned} L(x_1,\dots,x_n,\cdot) : \Theta &\to \mathbb{R}^+ \\ \theta &\mapsto L(x_1,\dots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta). \end{aligned}$$

Exemple 3.1.1. *Soit l'échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,\theta)$ avec $\theta \in]0,1[$. X_1 suit une loi de Bernouilli $\mathscr{B}(1,\theta)$ si*

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}1_{\{0,1\}}(x) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{si } x \in \{0,1\} \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance est

$$\begin{aligned} L(x_1,\dots,x_n,\theta) &= \prod_{i=1}^{n} f(x_i,\theta) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}1_{\{0,1\}}(x_i) \\ &= (1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i} 1_{\{0,1\}^n}(x_1,\dots,x_n) \\ &= \begin{cases} (1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i} & \text{si } (x_1,\dots,x_n) \in \{0,1\}^n \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

Exemple 3.1.2. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi exponentielle $\mathscr{E}(\theta)$ avec $\theta > 0$. X_1 suit la loi exponentielle $\mathscr{E}(\theta)$ si*

$$f(x,\theta) = \theta e^{-\theta x}1_{\mathbb{R}_+^*}(x) = \begin{cases} \theta e^{-\theta x} & \text{si } x \in \mathbb{R}_+^* \\ 0 & \text{sinon} \end{cases}$$

La vraisemblance est

$$\begin{aligned} L(x_1,\dots,x_n,\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} 1_{\mathbb{R}_+^*}(x_i) \\ &= \theta^n e^{-\theta \sum_{i=1}^n x_i} 1_{(\mathbb{R}_+^*)^n}(x_1,\dots,x_n). \\ &= \begin{cases} \theta^n e^{-\theta \sum_{i=1}^n x_i} & si\ (x_1,\dots,x_n) \in (\mathbb{R}_+^*)^n \\ 0 & sinon \end{cases} \end{aligned}$$

Exemple 3.1.3. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathcal{N}(m,\sigma^2)$ avec $m \in \mathbb{R}$ et $\sigma > 0$. X_1 suit la loi normale $\mathcal{N}(m,\sigma^2)$ si*

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

La vraisemblance est

$$\begin{aligned} L(x_1,\dots,x_n,m,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2} \\ &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-m)^2} \end{aligned}$$

Exercice 3.1.1.

1. *On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de Poisson $\mathcal{P}(\theta)$ avec $\theta > 0$. Ecrire la vraisemblance.*
2. *On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi uniforme $\mathcal{U}([0,\theta]$ avec $\theta > 0$. Ecrire la vraisemblance.*
3. *On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi Gamma $\mathcal{G}(\alpha,\rho)$ avec $\alpha,\rho > 0$. Ecrire la vraisemblance.*

3.2 Exhaustivité

Un échantillon nous apporte une certaine information sur le paramètre θ. Lorsque l'on résume cet échantillon par une statistique, il s'agit de ne pas perdre cette information. Une statistique qui conserve l'information contenue dans l'échantillon sera dite exhaustive.

Définition 3.2.1. *La statistique $T(X_1,\dots,X_n)$ est dite exhaustive pour θ si la loi conditionnelle de $(X_1,\dots,X_n)$ sachant $T(X_1,\dots,X_n)$ ne dépend pas de θ.*

Le théorème ci-dessus appelé théorème de factorisation permet de trouver une statistique exhaustive ou de justifier qu'une statistique est exhaustive.

Théorème 3.2.1. *La statistique $T(X_1,\dots,X_n)$ est exhaustive pour θ si et seulement si la vraisemblance peut se factoriser sous la forme*

$$L(x_1,\dots,x_n,\theta) = g(T(x_1,\dots,x_n),\theta)h(x_1,\dots,x_n).$$

Exemple 3.2.1. *Soit l'échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathcal{B}(1,\theta)$ avec $\theta \in]0,1[$.*
La vraisemblance est

$$L(x_1,\dots,x_n,\theta) = (1-\theta)^n \Big(\frac{\theta}{1-\theta}\Big)^{\sum_{i=1}^n x_i} 1_{\{0,1\}^n}(x_1,\dots,x_n)$$

Nous avons

$$g\Big(\sum_{i=1}^{n} x_i, \theta\Big) = (1-\theta)^n \Big(\frac{\theta}{1-\theta}\Big)^{\sum_{i=1}^n x_i}$$
$$h(x_1,\dots,x_n) = 1_{\{0,1\}^n}(x_1,\dots,x_n).$$

Grâce au théorème de factorisation, on déduit que la statistique $\sum_{i=1}^n X_i$ *est exhaustive pour* θ.

Exemple 3.2.2. *Soit un échantillon* $(X_1,\dots,X_n)$ *issu d'une loi exponentielle* $\mathscr{E}(\theta)$ *avec* $\theta > 0$. *La vraisemblance est*

$$L(x_1,\dots,x_n,\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i} 1_{(\mathbb{R}_+^*)^n}(x_1,\dots,x_n).$$

Nous avons

$$g\Big(\sum_{i=1}^{n} x_i, \theta\Big) = \theta^n e^{-\theta \sum_{i=1}^n x_i}$$
$$h(x_1,\dots,x_n) = 1_{(\mathbb{R}_+^*)^n}(x_1,\dots,x_n).$$

Grâce au théorème de factorisation, on déduit que la statistique $\sum_{i=1}^n X_i$ *est exhaustive pour* θ.

Exemple 3.2.3. *Soit un échantillon* $(X_1,\dots,X_n)$ *issu d'une loi normale* $\mathscr{N}(m,\sigma^2)$ *avec* $m \in \mathbb{R}$ *connue et* $\sigma > 0$ *inconnue. La vraisemblance est*

$$L(x_1,\dots,x_n,\sigma^2) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-m)^2}$$

Nous avons

$$g\Big(\sum_{i=1}^{n} (x_i-m)^2, \sigma^2\Big) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-m)^2}$$
$$h(x_1,\dots,x_n) = 1.$$

Grâce au théorème de factorisation, on déduit que la statistique $\sum_{i=1}^n (x_i-m)^2$ *est exhaustive pour* σ^2.

Exercice 3.2.1.

1. *On considère un échantillon* $(X_1,\dots,X_n)$ *issu d'une loi de Poisson* $\mathscr{P}(\theta)$ *avec* $\theta > 0$. *Déterminer une statistique exhaustive pour* θ.
2. *On considère un échantillon* $(X_1,\dots,X_n)$ *issu d'une loi uniforme* $\mathscr{U}([0,\theta])$ *avec* $\theta > 0$. *Déterminer une statistique exhaustive pour* θ.
3. *On considère un échantillon* $(X_1,\dots,X_n)$ *issu d'une loi normale* $\mathscr{N}(m,\sigma^2)$ *avec* $m \in \mathbb{R}, \sigma^2 > 0$. *Déterminer une statistique exhaustive pour* (m,σ^2).

Chapitre

4 Information de Fisher

4.1 Définition et propriétés

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de probabilité $\mathbb{P}_\theta$ admettant une densité ou de fonction de masse $f(\cdot,\theta)$ avec $\theta \in \Theta \subset \mathbb{R}$. On note

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

la vraisemblance de l'échantillon. Pour mesurer l'information contenue dans un échantillon $(X_1,\dots,X_n)$, Ronald Aylmer Fisher (1890-1962) a défini la quantité ci-dessous.

Définition 4.1.1. *On appelle information de Fisher au point θ apportée par l'échantillon $(X_1,\dots,X_n)$ la quantité*

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \ln(L(X_1,\dots,X_n,\theta))}{\partial\theta}\right)^2\right]$$

La proposition ci-dessus donne quelques propriétés de l'information de Fisher.

Proposition 4.1.1. *Nous avons :*

1. $I_n(\theta) \geq 0,\ \forall\theta \in \Theta$.
2. *Si X et Y sont indépendantes de lois respectives $\mathbb{P}_\theta$ et $\mathbb{Q}_\theta$. Notons $I_X(\theta)$, $I_Y(\theta)$ et $I_{(X,Y)}(\theta)$ les informations de Fisher au point θ respectivement apportées par X, Y, et (X,Y). Alors, nous avons :Alors, nous avons :*

   $$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

   Comme conséquence, l'information de Fisher $I_n(\theta)$ au point θ fournie par l'échantillon $(X_1,\dots,X_n)$ vérifie

   $$I_n(\theta) = nI_{X_1}(\theta)$$

   où $I_{X_1}(\theta)$ l'information de Fisher au point θ fournie par X_1.
3. *$T(X_1,\dots,X_n)$ est exhaustive$\Longleftrightarrow I_n(\theta) = I_T(\theta)\quad \forall\theta \in \Theta$ où $I_T(\theta)$ est l'information de Fisher au point θ fournie par $T(X_1,\dots,X_n)$. Cette propriété permet donc d'établir l'exhaustivité d'une statistique.*

Théorème 4.1.1. *Si le support de X_1 ne dépend pas de θ et si la vraisemblance $\theta \mapsto L(x_1,\dots,x_n,\theta)$ est deux fois dérivable, alors*

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ln(L(X_1,\dots,X_n,\theta))}{\partial\theta^2}\right].$$

4.2 Exemples

Exemple 4.2.1. *Soit l'échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,\theta)$ avec $\theta\in]0,1[$. Le support de la loi de Bernouilli $\{0,1\}$ est indépendant de θ. La vraisemblance*

$$L(x_1,\dots,x_n,\theta)=(1-\theta)^n\Big(\frac{\theta}{1-\theta}\Big)^{\sum_{i=1}^n x_i}1_{\{0,1\}^n}(x_1,\dots,x_n)$$

Pour tout $(x_1,\dots,x_n)\in\{0,1\}^n$, $L(x_1,\dots,x_n,\theta)>0$ et $\theta\mapsto L(x_1,\dots,x_n,\theta)$ est deux fois dérivable. La log-vraisemblance est donc

$$\ln L(x_1,\dots,x_n,\theta)=\sum_{i=1}^n x_i\ln(\theta)+(n-\sum_{i=1}^n x_i)\ln(1-\theta)$$
$$\frac{\partial^2\ln L(x_1,\dots,x_n,\theta)}{\partial\theta^2}=\frac{-\sum_{i=1}^n x_i}{\theta^2}-\frac{n-\sum_{i=1}^n x_i}{(1-\theta)^2}$$

Ainsi, nous avons :

$$I_n(\theta)=-\mathbb{E}_\theta\left[\frac{\partial^2\ln L(X_1,\dots,X_n,\theta)}{\partial\theta^2}\right]=\frac{n}{\theta(1-\theta)}.$$

Exemple 4.2.2. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathscr{N}(m,\sigma^2)$ avec $m\in\mathbb{R}$ et $\sigma>0$. La vraisemblance est*

$$L(x_1,\dots,x_n,m)=\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-m)^2}$$

Le support de la loi normale est $\mathbb{R}$ qui est indépendant de m. De plus la vraisemblance $m\mapsto L(x_1,\dots,x_n,m)$ est infiniment dérivable. La log-vraisemblance est :

$$\ln\Big(L(x_1,\dots,x_n,m)\Big)=-n\ln(\sqrt{2\pi}\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-m)^2.$$

$$\frac{\partial^2\ln L(x_1,\dots,x_n,m)}{\partial m^2}=-\frac{n}{\sigma^2}.$$

Ainsi, nous avons :

$$I_n(m)=-\mathbb{E}_m\left[\frac{\partial^2\ln L(X_1,\dots,X_n,m)}{\partial m^2}\right]=\frac{n}{\sigma^2}.$$

On en déduit que l'information est d'autant plus grande que la variance est plus petite.

Exercice 4.2.1. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathscr{N}(m,\sigma^2)$ avec $m\in\mathbb{R}$ et $\sigma>0$. Déterminer l'information de Fisher au point σ^2 fournie par $(X_1,\dots,X_n)$.*

Exercice 4.2.2. *Soit X une variable aléatoire suivant une loi gamma $\Gamma(a,\rho)$. Nous disposons de $(X_1,\dots,X_n)$, un échantillon aléatoire de taille n de loi parente X. Déterminer l'information de Fisher pour ρ fournie par $(X_1,\dots,X_n)$.*

Chapitre

5 Estimateurs

5.1 Principe général de l'estimation

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de probabilité $\mathbb{P}_\theta$ où $\theta \in \Theta \subset \mathbb{R}$ est inconnu. L'objectif est d'estimer θ en se basant sur l'échantillon $(X_1,\dots,X_n)$.

Définition 5.1.1. *Un estimateur $\widehat{\theta}_n$ du paramètre θ est une statistique*

$$\widehat{\theta}_n = T(X_1,\dots,X_n)$$

à valeurs dans un domaine acceptable pour θ.

- Si $(x_1,\dots,x_n)$ est une observation de $(X_1,\dots,X_n)$, $T(x_1,\dots,x_n)$ est appelée estimation de θ.
- Il faut faire la distinction entre l'estimateur de θ (qui est une variable aléatoire réelle) et l'estimation de θ qui est une grandeur numérique.

Bien évidemment, cette statistique $T(X_1,\dots,X_n)$ n'est pas choisie au hasard ! L'idée est de trouver une statistique de sorte à fournir une bonne estimation du paramètre d'intérêt θ.

Exemple 5.1.1. *Supposons que les variables aléatoires $(X_1,\dots,X_n)$ un échantillon issu d'une loi de moyenne m et de variance σ^2.*

- *La moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est un "bon" estimateur de la moyenne m. On verra dans la suite ce qu'en entend par "bon estimateur".*
- *La variance empirique $V_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (X_i - \overline{X}_n)^2$ est un estimateur de la variance σ^2.*

La question est de savoir ce qu'est un "bon estimateur". Quelles propriétés doit satisfaire un estimateur pour être considéré comme "bon" ? Nous devons distinguer deux cas suivant la taille d'échantillon n :

- propriétés à distance finie (pour n fixé)
- propriétés asymptotiques (pour $n \to +\infty$).

5.1.1 Propriétés à distance finie

5.1.1.1 Loi exacte

Définition 5.1.2. *La loi à distance finie (ou loi exacte) d'un estimateur correspond à la loi valable pour toute valeur de la taille de l'échantillon $n \in \mathbb{N}$.*

Remarque 5.1.1. *En dehors du modèle gaussien, il est souvent difficile de déterminer la loi exacte des estimateurs.*

5.1.1.2 Biais

Définition 5.1.3. *Le biais d'un estimateur $\widehat{\theta}_n$ de θ est défini par*

$$b_n(\theta) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta = \mathbb{E}_\theta(\widehat{\theta}_n - \theta).$$

Le biais de l'estimateur est la moyenne des écarts systématiques entre $\widehat{\theta}_n$ et θ. L'absence d'un écart systématique entre $\widehat{\theta}_n$ et θ se traduit par un biais nul.

Définition 5.1.4. *Un estimateur $\widehat{\theta}_n$ de θ est dit sans biais lorsque pour tout $\theta \in \Theta$*

$$\mathbb{E}_\theta(\widehat{\theta}_n) = \theta.$$

Dans le cas contraire, l'estimateur $\widehat{\theta}_n$ est dit biaisé.

Exercice 5.1.1. *On considère un échantillon $(X_1,\ldots,X_n)$ issu d'une loi de moyenne m et de variance σ^2 inconnues. Montrer que :*

- *$\overline{X}_n$ est un estimateur sans biais de m.*

  Le biais de $\overline{X}_n$ est donné par

  $$b(m) = \mathbb{E}_m(\overline{X}_n) - m$$

  $$\mathbb{E}_m(\overline{X}_n) = \mathbb{E}_m\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_m(X_i)$$

  Comme $\mathbb{E}_m(X_1) = \ldots = \mathbb{E}_m(X_n) = m$ alors nous pouvons ecrire

  $$\mathbb{E}_m(\overline{X}_n) = \mathbb{E}_m\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_m(X_i) = \frac{1}{n}\sum_{i=1}^n m = \frac{nm}{n} = m$$

  Finalement, nous obtenons

  $$\mathbb{E}_m(\overline{X}_n) = m \Longleftrightarrow b(m) = \mathbb{E}_m(\overline{X}_n) - m = 0 \qquad \forall m \in \mathbb{R}.$$

- *La variance empirique $V_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ est un estimateur biaisé de σ^2. En déduire que $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ est un estimateur sans biais de σ^2*

  $$\mathbb{E}_{\sigma^2}(V_n^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

  Cependant

  $$\mathbb{E}_{\sigma^2}(V_n^2) = \frac{n-1}{n}\sigma^2 \longrightarrow \sigma^2$$

  $$\sigma^2 = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \frac{n}{n-1}\mathbb{E}_{\sigma^2}(V_n^2) = \mathbb{E}_{\sigma^2}\Big(\frac{n}{n-1}V_n^2\Big) = \mathbb{E}_{\sigma^2}(S_n^2).$$

Exercice 5.1.2. *On considère un échantillon $(X_1,\ldots,X_n)$ issu d'une loi densité :*

$$f(x,\theta) = \frac{x}{\theta^2} e^{-x/\theta} 1_{\mathbb{R}_+^*}(x) \quad \theta > 0.$$

On cherche un estimateur sans biais de θ et on considère $\overline{X}_n$ comme un premier essai. Montrer que $\overline{X}_n$ est biaisé, et montrez comment modifier cet estimateur pour en obtenir un qui ne l'est pas.

Comme $\mathbb{E}_\theta(X_1) = \ldots = \mathbb{E}_\theta(X_n)$

$$\mathbb{E}_\theta(\overline{X}_n) = \mathbb{E}_\theta\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta(X_i) = \frac{n\mathbb{E}_\theta(X_1)}{n} = \mathbb{E}_\theta(X_1) = \int_{-\infty}^{+\infty} xf(x,\theta)dx.$$

$$= \int_{-\infty}^{+\infty} x\frac{x}{\theta^2}e^{-x/\theta}1_{\mathbb{R}_+^*}(x)dx = \frac{1}{\theta^2}\int_0^{+\infty} x^2e^{-x/\theta}dx = \frac{1}{\theta^2}\lim_{u\to+\infty}\int_0^u x^2e^{-x/\theta}dx$$

Exercice 5.1.3. *Deux recherches indépendantes font état d'échantillonnages effectués auprès d'une même population. Les seules données présentées sont les moyennes $\overline{X}_1$ et $\overline{X}_2$ et les tailles des échantillons n_1 et n_2. Déterminer la valeur k telle que $k(\overline{X}_1 - \overline{X}_2)^2$ est un estimateur sans biais de la variance σ^2 de la population.*

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \Longrightarrow \mathbb{E}(X^2) = var(X) + (\mathbb{E}(X))^2$$

5.1.1.3 Risque quadratique

On mesure la précision d'un estimateur par son risque quadratique.

Définition 5.1.5. *Pour un estimateur $\widehat{\theta}_n$ de θ, le risque quadratique est défini par*

$$\begin{aligned} R(\widehat{\theta}_n,\theta) &= \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2 \\ &= var_\theta(\widehat{\theta}_n) + (b_n(\theta))^2 \end{aligned}$$

Définition 5.1.6. *Soient $\widehat{\theta}_n$ et $\widetilde{\theta}_n$ deux estimateurs de θ. On dit que $\widehat{\theta}_n$ est préférable à $\widetilde{\theta}_n$ si*

$$R(\widehat{\theta}_n,\theta) \le R(\widetilde{\theta}_n,\theta) \quad \forall\theta\in\Theta \Longleftrightarrow R(\widehat{\theta}_n,\theta) - R(\widetilde{\theta}_n,\theta) \le 0 \quad \theta\in\Theta.$$

Un estimateur optimal au sens du risque quadratique est l'estimateur qui a le plus petit risque quadratique pour toute valeur de $\theta\in\Theta$. Il est souvent difficile, voire impossible, de trouver un estimateur optimal.

Remarque 5.1.2. *Pour un estimateur sans biais $\widehat{\theta}_n$ de θ, le risque quadratique est défini par*

$$R(\widehat{\theta}_n,\theta) = var_\theta(\widehat{\theta}_n)$$

Définition 5.1.7. *Soient $\widehat{\theta}_n$ et $\widetilde{\theta}_n$ deux estimateurs sans biais de θ. On dit que $\widehat{\theta}_n$ est préférable à $\widetilde{\theta}_n$ si*

$$var_\theta(\widehat{\theta}_n) \le var_\theta(\widetilde{\theta}_n) \quad \forall\theta\in\Theta \Longleftrightarrow var_\theta(\widehat{\theta}_n) - var_\theta(\widetilde{\theta}_n) \le 0 \quad \theta\in\Theta.$$

Exercice 5.1.4. *On considère un échantillon $(X_1,\ldots,X_n)$ issu d'une loi uniforme $\mathcal{U}([0,\theta])$. considérons les deuxestimateurs suivants : $\widehat{\theta}_1 = 2\overline{X}_n$ et $\widehat{\theta}_2 = \max(X_1,\ldots,X_n)$.*

1. *Montrer que $\widehat{\theta}_1$ est un estimateur sans biais de θ.*
2. *Montrer que $\widehat{\theta}_2$ est un estimateur biaisé de θ ; déterminer son biais ; déterminec c tel que $\widehat{\theta}_3 = c\widehat{\theta}_2$ soit un estimateur sans biais de θ.*
3. *Déterminer la variance de $\widehat{\theta}_1$ et la variance de $\widehat{\theta}_3$ et dites lequel des deux estimateurs est meilleur.*

   Etudier le signe de la fonction suivante

   $$\theta \mapsto var_\theta(\widehat{\theta}_1) - var_\theta(\widehat{\theta}_3)$$

   sur l'espace $\Theta = \mathbb{R}_+^*$

5.1.1.4 Borne de Cramer-Rao

Le résultat suivant indique que le risque quadratique d'un estimateur sans biais (i.e. sa variance) ne peut être inférieure à une certaine borne qui dépend de l'information de Fisher.

Théorème 5.1.1. *On suppose que l'information de Fisher sur θ apportée par $(X_1,\dots,X_n)$ existe et est strictement positive pour tout θ. Soit $\widehat{\theta}_n$ un estimateur sans biais de θ. Alors nous avons*

$$var_\theta(\widehat{\theta}_n) \geq \frac{1}{I_n(\theta)} \qquad \forall \theta \in \Theta.$$

La borne $BRC(\theta) = \dfrac{1}{I_n(\theta)}$ est appelée borne de Cramer-Rao.

Remarque 5.1.3. *Si $\widehat{\theta}_n$ est un estimateur sans biais de $h(\theta)$ alors*

$$var_\theta(\widehat{\theta}_n) \geq \frac{(h'(\theta))^2}{I_n(\theta)}.$$

Dans ce cas, la borne de Cramer-Rao pour l'estimation sans biais de $h(\theta)$ est :

$$BCR(\theta) = \frac{(h'(\theta))^2}{I_n(\theta)}.$$

Exemple 5.1.2. *Soit l'échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathcal{B}(1,\theta)$ avec $\theta \in]0,1[$. L'information de Fisher est*

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

Ainsi la borne de Cramer-Raopour l'estimation sans biais de θ est :

$$BCR(\theta) = \frac{1}{I_n(\theta)} = \frac{\theta(1-\theta)}{n}.$$

Exemple 5.1.3. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathcal{N}(m,\sigma^2)$ avec $m \in \mathbb{R}$ inconnue et $\sigma > 0$ connue. L'information de Fisher est*

$$I_n(m) = \frac{n}{\sigma^2}.$$

Ainsi la borne de Cramer-Rao pour l'estimation sans biais de m est :

$$BCR(m) = \frac{1}{I_n(m)} = \frac{\sigma^2}{n}.$$

Définition 5.1.8. *Un estimateur $\widehat{\theta}_n$ de θ est dit efficace si*

- *$\widehat{\theta}_n$ est sans biais*
- *$var_\theta\left(\widehat{\theta}_n\right) = BCR(\theta)$.*

Exercice 5.1.5. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathcal{N}(m,\sigma^2)$ avec $m \in \mathbb{R}$ inconnue et $\sigma > 0$ connue. Montrer que $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur efficace de m.*

Il suffit de montrer que

— $\mathbb{E}_m(\overline{X}_n) = m$

— $var_m\left(\overline{X}_n\right) = \dfrac{1}{I_n(m)} = \dfrac{\sigma^2}{n}$

$$var_m\left(\overline{X}_n\right) = var_m\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2} var_m\Big(\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\Big(\sum_{i=1}^{n} var_m(X_i)\Big)$$

car les varibales $X_1,\dots,X_n$ sont indépendantes.

5.1.2 Propriétés asymptotiques

5.1.2.1 Convergence ou consistance

Définition 5.1.9. *Un estimateur $\widehat{\theta}_n$ de θ est dit asymptotiquement sans biais lorsque pour tout θ,*

$$\mathbb{E}_\theta(\widehat{\theta}_n) \xrightarrow[n\to+\infty]{} \theta.$$

Définition 5.1.10. *$\widehat{\theta}_n$ est un estimateur convergent (ou consistant) de θ si*

$$\widehat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta \quad \text{lorsque } n \to +\infty$$

c'est à dire

$$\forall \varepsilon > 0 \quad \lim_{n\to+\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0.$$

Interprétation : La convergence est une des propriétés les plus importantes pour un estimateur. On a la garantie qu'à un rang n assez grand et avec grande probabilité, $\widehat{\theta}_n$ soit proche du paramètre θ.

Exercice 5.1.6. *Considerons un échantillon $(X_1,\dots,X_n)$ issu d'une loi de moyenne m et variance $\sigma^2 > 0$. Montrer la moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur convergent de m.*

Solution 1. *loi des grands nombres.*

Solution 2. *Inégalité de Bienaymé-Tchebythcev + Théorème des gendarmes. Pour tout $\varepsilon > 0$*

$$0 \leq \mathbb{P}_m\left(\left|\overline{X}_n - m\right| > \varepsilon\right) = \mathbb{P}_m\left(\left|\overline{X}_n - \mathbb{E}_m(\overline{X}_n)\right| > \varepsilon\right) \leq \frac{var_m(\overline{X}_n)}{\varepsilon^2}$$

Solution 3. *Comme $\mathbb{E}_m(\overline{X}_n) = m$, il suffit de montrer que $var_m(\overline{X}_n) \longrightarrow 0$ pour conclure.*

5.1.2.2 Normalité asymptotique

Définition 5.1.11. *Un estimateur $\widehat{\theta}_n$ de θ est dit asymptotiquement normal si*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,\sigma_\theta^2) \quad n \to +\infty$$

où σ_θ^2 est à déterminer.

Interprétation : La normalité asymptotique est une propriété plus précise qui indique que la fluctuation de l'estimateur autour de θ est approximativement normale.

Exemple 5.1.4. *Considerons un échantillon $(X_1,\dots,X_n)$ issu d'une loi de moyenne m et variance $\sigma^2 > 0$. Montrer la moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur asymptotiquement normal de m.*

Solution. *Le Théorème Central Limite permet de répondre à cette question.*

Exemple 5.1.5. *Considerons un échantillon $(X_1,\dots,X_n)$ issu d'une loi de moyenne m et variance $\sigma^2 > 0$. Montrer la moyenne empirique $\overline{X}_n^2 = \left(\dfrac{1}{n}\sum_{i=1}^{n} X_i\right)^2$ est un estimateur asymptotiquement normal de m^2.*

Solution. *Pour repondre à la question, on utlise la delta-method.*

- *D'après le Théorème Central Limite, nous avons*

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

- *En posant* $g(x) = x^2$, $g'(x) = 2x$ *et* g *est classe* $\mathscr{C}^1$ *sur* $\Theta = \mathbb{R}$ *et on a*

$$\sqrt{n}(\overline{X}_n^2 - m^2) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, \sigma^2 \times (2m)^2) = \mathcal{N}(0, 4m^2\sigma^2).$$

C'est à dire

$$\sqrt{n}(\overline{X}_n^2 - m^2) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, 4m^2\sigma^2)$$

Chapitre

6 Méthodes d'estimation

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de probabilité $\mathbb{P}_\theta$ avec θ inconnu.

6.1 Méthode des moments

Principe de la méthode :

— Trouver des fonctions g et q telles que

$$\mathbb{E}(g(X_1) = q(\theta). \tag{6.1.1}$$

Il faudrait choisir de préférence q bijective.

— Remplacer dans (6.1.1), la moyenne théorique par la moyenne empirique :

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = q(\theta) \tag{6.1.2}$$

— Résoudre (6.1.2) ; si q est bijective alors l'estimateur par la méthode des moments est donné par :

$$\widehat{\theta}_n = q^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\Big).$$

Exemple 6.1.1. *Considérons l'échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,\theta)$ avec $\theta \in]0,1[$.*

1. *Etape 1 : $\mathbb{E}_\theta(X_1) = \theta$; $g(X_1) = X_1$ et $q(\theta) = \theta$.*
2. *Etape 2 : $\overline{X}_n = \theta$.*
3. *Etape 3 : On conclut que l'estimateur est $\widehat{\theta}_n = \overline{X}_n$*

Exemple 6.1.2. *Considérons l'échantillon $(X_1,\dots,X_n)$ issu d'une loi exponentielle $\mathscr{E}(\theta)$ avec $\theta > 0$.*

1. *Etape 1 : $\mathbb{E}_\theta(X_1) = \frac{1}{\theta}$; $g(X_1) = X_1$ et $q(\theta) = \frac{1}{\theta}$. est bijective.*
2. *Etape 2 : $\overline{X}_n = \frac{1}{\theta}$.*
3. *Etape 3 : On conclut que l'estimateur est $\widehat{\theta}_n = \frac{1}{\overline{X}_n}$*

Exemple 6.1.3. *Considérons l'échantillon $(X_1,\dots,X_n)$ issu d'une loi exponentielle $\mathscr{E}(\theta)$ avec $\theta > 0$.*

1. *Etape 1 :*

$$\mathbb{E}_\theta(X_1^2) = var_\theta(X_1) + (E_\theta(X_1))^2 = \frac{1}{\theta^2} + \frac{1}{\theta^2}$$

$g(x) = x^2$ *et* $q(\theta) = \frac{2}{\theta^2}$ *est bijective.*

2. *Etape 2 :* $\frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{2}{\theta^2}$.

3. *Etape 3 :*

$$\theta = \sqrt{\frac{2}{\frac{1}{n}\sum_{i=1}^n X_i^2}}$$

On conclut que l'estimateur est

$$\widehat{\theta}_n = \sqrt{\frac{2}{\frac{1}{n}\sum_{i=1}^n X_i^2}}$$

Exercice 6.1.1. *Pendant une année, un assureur a enregistré les montants de sinistres suivants*

$$\{500, 1000, 1500, 2500, 4500\}.$$

Il décide de modéliser ces données par une loi Log-normale(μ, σ^2). *En utilisant la méthode des moments, estimer les paramètres* μ *et* σ^2. *Calculer ensuite la probabilité d'avoir un sinistre supérieur à 4 500.*

Les montants sont en milliers de francs.

Exercice 6.1.2. *Soit* $(X_1, \ldots, X_n)$ *un échantillon d'une population de loi uniforme sur* $[\theta, 1]$. *Déterminer par la méthode des moments l'estimateur de* θ. *Etudier ses propriétés.*

Exercice 6.1.3. *Soit* $(X_1, \ldots, X_n)$ *un échantillon d'une population de loi gamma* $\Gamma(2, \rho)$ *avec* ρ *inconnu. Déterminer par la méthode des moments l'estimateur de* ρ. *Etudier ses propriétés.*

6.2 Methode du maximum de vraisemblance

La vraisemblance de l'échantillon $(X_1, \ldots, X_n)$ est donnée par

$$L_n(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta).$$

Dans le cas d'une loi discrète

$$L_n(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i).$$

Pour un échantillon de taille 1

$$L_1(x, \theta) = \mathbb{P}_\theta(X_1 = x).$$

Principe de la méthode : Choisir comme estimateur la statistique $\widehat{\theta}_n$, la valeur de θ qui maximise la vraisemblance $L_n(X_1, \ldots, X_n, \theta)$:

Définition 6.2.1. $\widehat{\theta}_n$ *est un estimateur du maximum de vraisemblance de* θ *si*

$$\forall \theta \in \Theta \quad L_n(X_1, \ldots, X_n, \widehat{\theta}_n) \geq L_n(X_1, \ldots, X_n, \theta).$$

La recherche d'un maximum de la vraisemblance n'est pas forcément réduite à un simple calcul des zéros de la dérivée de L. Cependant, ce cas étant le plus fréquent, il est logique de poser les deux hypothèses suivantes :

— le support $X(\Omega)$ ne dépend pas de θ.

— la vraisemblance L est deux fois continûment dérivable par rapport θ.

Alors $\widehat{\theta}_n$ est solution du système :

$$\begin{cases} \dfrac{\partial L_n(X_1,\ldots,X_n,\theta)}{\partial\theta}(\hat{\theta}_n)=0 \\ \\ \dfrac{\partial^2 L_n(X_1,\ldots,X_n,\theta)}{\partial\theta^2}(\hat{\theta}_n)<0. \end{cases}$$

Puisque la fonction logarithme est croissante, vu la forme de L, il est aussi aisé d'utiliser le logarithme de la vraisemblance si $f(x,\theta)>0$, $\forall x\in X(\Omega)$, $\forall\theta$. Un estimateur du maximum de vraisemblance maximise le logarithme de la vraisemblance $L_n(X_1,\ldots,X_n,\theta)$:

$$\ln(L_n(X_1,\ldots,X_n,\theta))=\sum_{i=1}^{n}\ln(f(X_i,\theta).$$

Un estimateur du maximum de vraisemblance $\hat{\theta}_n$ est alors solution du système

$$\begin{cases} \dfrac{\partial \ln(L_n(X_1,\ldots,X_n,\theta))}{\partial\theta}(\hat{\theta}_n)=0 \\ \\ \dfrac{\partial^2 \ln(L_n(X_1,\ldots,X_n,\theta))}{\partial\theta^2}(\hat{\theta}_n)<0. \end{cases}$$

Proposition 6.2.1. *Si $T(X_1,\ldots,X_n)$ est une statistique exhaustive pour θ, l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ en dépend.*

Proposition 6.2.2. *Si $\widehat{\theta}_n$ est un estimateur du maximum de vraisemblance de θ alors $h(\widehat{\theta}_n)$ est un estimateur du maximum de vraisemblance de $h(\theta)$.*

Exemple 6.2.1. *Soit l'échantillon $(X_1,\ldots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,\theta)$ avec $\theta\in]0,1[$. La vraisemblance de $(x_1,\ldots,x_n)$ issu d'une loi de Bernouilli est :*

$$\begin{aligned} L(x_1,\ldots,x_n,\theta)&=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}1_{\{0,1\}}(x_i) \\ &=(1-\theta)^n\Big(\frac{\theta}{1-\theta}\Big)^{\sum_{i=1}^n x_i}1_{\{0,1\}^n}(x_1,\ldots,x_n). \end{aligned}$$

Pour tout $(x_1,\ldots,x_n)\in\{0,1\}^n$, la log-vraisemblance est donnée

$$\begin{aligned} \ln L(x_1,\ldots,x_n,\theta)&=\sum_{i=1}^{n}x_i\ln(\theta)+(n-\sum_{i=1}^{n}x_i)\ln(1-\theta) \\ \frac{\partial \ln L(x_1,\ldots,x_n,\theta)}{\partial\theta}&=\frac{\sum_{i=1}^n x_i}{\theta}-\frac{n-\sum_{i=1}^n x_i}{(1-\theta)}=0\Longleftrightarrow\theta=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}_n \\ \frac{\partial^2 \ln L(x_1,\ldots,x_n,\theta)}{\partial\theta^2}(\overline{x}_n)&=\frac{-n\overline{x}_n}{\overline{x}_n^2}-\frac{n-n\overline{x}_n}{(1-\overline{x}_n)^2}<0. \end{aligned}$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\hat{\theta}_n=\overline{X}_n.$$

Etude des propriétés de $\hat{\theta}_n$.

1. *D'après la loi des grands nombres, $\overline{X}_n$ est un estimateur convergent de θ.*
2. *D'après le Théorème Central limite $\overline{X}_n$ est asymptotiquement normal :*

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

3. *$\overline{X}_n$ est un estimateur efficace de θ.*

Exemple 6.2.2. *Soit un échantillon $(X_1,\dots,X_n)$ issu d'une loi exponentielle de paramètre $\theta > 0$. La vraisemblance de $(x_1,\dots,x_n)$ est*

$$\begin{aligned} L(x_1,\dots,x_n,\theta) &= \prod_{i=1}^{n} \theta \exp(-\theta x_i) \mathbb{1}_{\mathbb{R}_+^*}(x_i) \\ &= \theta^n \exp\Big(-\theta \sum_{i=1}^{n} x_i\Big) 1_{(\mathbb{R}_+^*)^n}(x_1,\dots,x_n). \end{aligned}$$

Pour tout $(x_1,\dots,x_n) \in (\mathbb{R}_+^*)^n$, on a

$$\ln(L(x_1,\dots,x_n,\theta)) = n\ln(\theta) - \theta \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L(x_1,\dots,x_n,\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \Longleftrightarrow \theta = \frac{1}{\overline{x}_n}$$

$$\frac{\partial^2 \ln L(x_1,\dots,x_n,\theta)}{\partial \theta^2}\Big(\frac{1}{\overline{x}_n}\Big) = -n\overline{x}_n^2 < 0.$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\hat{\theta}_n = \frac{1}{\overline{X}_n}.$$

Pour montrer que $\hat{\theta}_n$ est biaisé (ou sans biais), il faut calculer

$$\mathbb{E}\Big(\frac{1}{\overline{X}_n}\Big) = \mathbb{E}\Big(\frac{n}{\sum_{i=1}^n X_i}\Big) = n \times \mathbb{E}\Big(\frac{1}{\sum_{i=1}^n X_i}\Big)$$

Comme les variables X_i sont indépendantes et de même loi $\mathscr{E}(\theta) = \Gamma(1,\theta)$, on en déduit que

$$\sum_{i=1}^{n} X_i \rightsquigarrow \Gamma(n,\theta).$$

Si $X \rightsquigarrow \Gamma(a,\theta)$, $Y \rightsquigarrow \Gamma(b,\theta)$ et X et Y sont indépendantes alors

$$X + Y \rightsquigarrow \Gamma(a+b,\theta)$$

Posons $Z = \sum_{i=1}^{n} X_i$, nous avons

$$Z \rightsquigarrow \Gamma(n,\theta) \Longleftrightarrow f_Z(z,\theta) = \frac{\theta^n}{\Gamma(n)} z^{n-1} e^{-\theta z} 1_{\mathbb{R}_*^+}(z)$$

Finalement

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\overline{X}_n}\right) &= \mathbb{E}\left(\frac{n}{\sum_{i=1}^n X_i}\right) \\
&= n \times \mathbb{E}\left(\frac{1}{\sum_{i=1}^n X_i}\right) \\
&= n \times \mathbb{E}\left(\frac{1}{Z}\right) \quad Z = \sum_{i-1}^{n} X_i \\
&= \int_{-\infty}^{+\infty} \frac{1}{z} f_Z(z,\theta) dz \\
&= \frac{\theta^n}{\Gamma(n)} \int_0^{+\infty} z^{n-2} e^{-\theta z} dz \\
&= \frac{\theta^n}{\Gamma(n)} \int_0^{+\infty} z^{(n-1)-1} e^{-\theta z} dz \\
&= \frac{\theta^n}{\Gamma(n)} \times \frac{\Gamma(n-1)}{\theta^{n-1}}
\end{aligned}$$

Utiliser la formule suivante :

$$\begin{aligned}
&\frac{\Gamma(a)}{\rho^a} = \int_0^{+\infty} x^{a-1} e^{-\rho x} dx \\
&\Gamma(n) = (n-1)\Gamma(n-1) \qquad n \text{ entier} \geq 1 \\
&\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx.
\end{aligned}$$

Après les calculs, on obtiendra

$$\mathbb{E}\left(\frac{1}{\overline{X}_n}\right) = \frac{n}{n-1}\theta \neq \theta.$$

Etude des propriétés de $\hat{\theta}_n$.

1. *D'après la loi des grands nombres, on a :*

$$\overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{\theta}.$$

Comme, l'application $x \mapsto \frac{1}{x}$ est continue sur $\mathbb{R}_+^*$, alors

$$\frac{1}{\overline{X}_n} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta.$$

2. *D'après le Théorème Central limite $\overline{X}_n$ est asymptotiquement normal :*

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\theta}\right) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(0, \frac{1}{\theta^2}\right).$$

Comme, l'application $g : x \mapsto \frac{1}{x}$ est dérivable sur $\mathbb{R}_+^*$ et $g'(x) = -\frac{1}{x^2}$, on obtient par la delta-méthode :

$$\sqrt{n}(g(\overline{X}_n) - g(1/\theta)) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(0, \frac{1}{\theta^2}(g'(1/\theta))^2\right).$$

c'est à dire

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \theta\right) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, \theta^2).$$

3. $\hat{\theta}_n$ *est un estimateur biaisé de* θ*. Il ne peut donc pas être efficace.*

Exercice 6.2.1. *Soit X la variable aléatoire représentant le montant d'un sinistre. On suppose* $X \mapsto \mathcal{E}(\lambda)$*. Pour des contrats d'assurance comportant une franchise forfaitaire de 100 $ et une limite supérieure de 3 000 $, les montants de sinistres suivants ont été payés par l'assureur :*

$$\{100, 200, 250, 425, 515, 630, 1000, 1500, 2900, 2900\}.$$

Estimer le montant espéré d'un sinistre par la méthode du maximum de vraisemblance.

Exercice 6.2.2. *Soit* $(X1, \ldots, X_n)$ *un échantillon issu d'une population de loi géométrique de paramètre* p*. Déterminer par la méthode du maximum de vraisemblance l'estimateur de* p*. Etudier ses propriétés.*

Exercice 6.2.3. *Soit* $(X1, \ldots, X_n)$ *un échantillon issu d'une population de loi uniforme* $\mathcal{U}([0, \theta])$ *avec* $\theta > 0$*. Déterminer par la méthode du maximum de vraisemblance l'estimateur de* θ*. Etudier ses propriétés.*

Chapitre

7 Estimation par intervalle de confiance

En estimation ponctuelle, on ne propose qu'une seule valeur pour le paramètre d'intérêt. Il n'y a quasiment aucune chance que cette valeur soit la vraie valeur. L'objectif de ce chapitre est de proposer une fourchette de valeurs possibles, tout un intervalle, ni trop gros, pour qu'il soit assez informatif, ni trop petit, pour qu'on soit raisonnablement sûr qu'il contienne la vraie valeur.

7.1 Introduction

Définition 7.1.1. *Soit $\alpha \in]0,1[$; on appelle intervalle de confiance pour le paramètre θ de niveau de confiance égale à $1-\alpha$, un intervalle aléatoire $I(X_1,\dots,X_n) \subset \Theta$ tel que*

$$\mathbb{P}_\theta(I(X_1,\dots,X_n) \ni \theta) = 1-\alpha.$$

Définition 7.1.2. *On dira que un intervalle aléatoire $I(X_1,\dots,X_n)$ est un intervalle de confiance pour le paramètre θ de niveau de confiance asymptotique égale à $1-\alpha$ si*

$$\lim_{n\to+\infty} \mathbb{P}_\theta(I(X_1,\dots,X_n) \ni \theta) = 1-\alpha.$$

Lorsque

$$I(X_1,\dots,X_n) = [T_n^*(X_1,\dots,X_n), T_n^{**}(X_1,\dots,X_n)]$$

où $T_n^*(X_1,\dots,X_n)$ et $T_n^{**}(X_1,\dots,X_n)$ sont des statistiques à valeurs dans Θ, on parle d'intervalle de confiance bilatéral. Dans le cas où

$$I(X_1,\dots,X_n) = [T_n^*(X_1,\dots,X_n), +\infty[$$

ou

$$I(X_1,\dots,X_n) =]-\infty, T_n^*(X_1,\dots,X_n)],$$

on parle d'intervalle de confiance unilatéral.

Remarque 7.1.1. *Dans l'univers des échantillons possibles, pour une proportion au moins $1-\alpha$ d'entre eux, on obtient un intervalle qui contient θ.*

Remarque 7.1.2. *A α fixé, l'intervalle de confiance est d'autant meilleur que sa longueur est petite.*

Remarque 7.1.3. *On doit comprendre un intervalle de confiance de niveau* $1-\alpha$ *comme un intervalle aléatoire qui a une probabilité* $1-\alpha$ *de contenir le vrai parametre* θ*.*

Définition 7.1.3. *Soit* X *une variable aléatoire réelle de fonction de répartition* $F(x) = \mathbb{P}(X \leq x)$*. Pour* $\alpha \in]0,1[$*, on appelle quantile (ou fractile) d'ordre* α *de la loi de* X *le nombre*

$$q_\alpha = \inf\{x \in \mathbb{R}, F(x) \geq \alpha\}.$$

Lorsque la fonction de répartition F est continue et strictement croissante, elle est inversible d'inverse F^{-1} et pour tout $\alpha \in]0,1[$, on a $q_\alpha = F^{-1}(\alpha)$.

7.2 Construction d'un intervalle de confiance

1. Construction de la fonction pivot (ou pivotale)
2. Détermination des constantes
3. Pivotement

7.2.1 Fonction pivotale

Définition 7.2.1. *On appelle fonction pivotale pour* θ *toute fonction de l'échantillon et de* θ*,* $\phi(X_1,\ldots,X_n,\theta)$ *dont la loi ne dépend pas de* θ*.*

Définition 7.2.2. *Une fonction asymptotiquement pivotale pour* θ *est une variable aléatoire,* $\phi(X_1,\ldots,X_n,\theta)$ *qui converge en loi vers une variable aléatoire dont la loi ne dépend pas de* θ*.*

7.2.2 Construction d'un intervalle de confiance bilateral

7.2.2.1 Méthode non asymptotique

1. Soit $\phi(X_1,\ldots,X_n,\theta)$ une fonction pivotale pour θ.
2. Pour un seuil $\alpha \in]0,1[$ fixé, soient q_1 et q_2 tels que

   $$\mathbb{P}_\theta\Big[q_1 \leq \phi(X_1,\ldots,X_n,\theta) \leq q_2\Big] = 1-\alpha$$

   c'est à dire

   $$\mathbb{P}_\theta\Big[\phi(X_1,\ldots,X_n,\theta) \leq q_1\Big] = \alpha_1$$
   $$\mathbb{P}_\theta\Big[\phi(X_1,\ldots,X_n,\theta) \geq q_2\Big] = \alpha_2$$

   avec $\alpha_1 + \alpha_2 = \alpha$.
3. La double inéquation

   $$q_1 \leq h(X_1,\ldots,X_n,\theta) \leq q_2 \tag{7.2.1}$$

   peut se résoudre (ou "pivoter") en θ selon

   $$T_1(X_1,\ldots,X_n) \leq \theta \leq T_2(X_1,\ldots,X_n),$$

   on en déduit immédiatement un intervalle de confiance bilatéral pour θ de niveau de confiance $1-\alpha$.

7.2.2.2 Méthode asymptotique

- Soit T_n un estimateur de θ tel que

$$\frac{T_n-\theta}{s_n(\theta)} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

où $s_n(\theta)$ est une fonction continue de θ.

- Si la fonction $\dfrac{T_n-\theta}{s_n(\theta)}$ pivote pour isoler θ, on obtient l'intervalle de confiance approchée.
- Sinon T_n étant convergeant, moyennant la continuité de s_n (quelque soit n), on obtient

$$\frac{T_n-\theta}{s_n(T_n)} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1).$$

Le pivotement est alors immédiat.

Remarque 7.2.1. *Pour les intervalles de confiance unilatéraux, on utilise la méthode ci-dessus.*

7.2.3 Densité de probabilité unimodale

Définition 7.2.3. *Une densité de probabilité f sur $\mathbb{R}$ est unimodale autour d'un mode s'il existe x^* un mode tel que f croissante sur $]-\infty,x^*]$ et f décroissante sur $[x^*,+\infty[$.*

Proposition 7.2.1. *Soit f une densité unimodale et $[a,b]$ unintervalle satisfaisant*

i) $\displaystyle\int_a^b f(x)dx = 1-\alpha$

ii) $f(a)=f(b)>0$

iii) $a\le x^*\le b$ *où x^* est le mode de f.*

Alors $[a,b]$ est l'intervalle le plus court parmi tous les intervalles satisfaisant i).

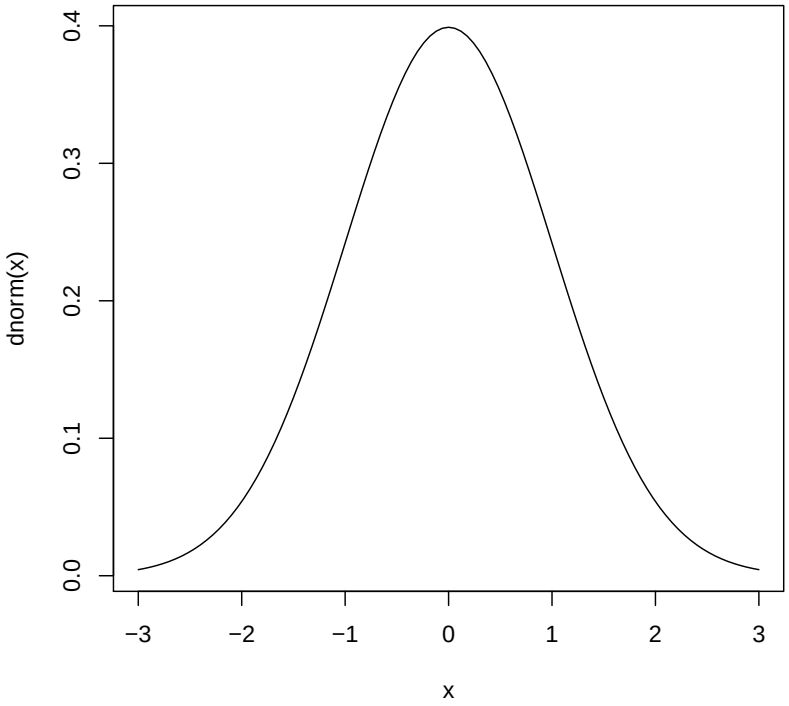
Exemple 7.2.1. *1. La loi normale centrée-réduite. L'intervalle le plus court est de la forme $[-b,b]$ où $b=z_{1-\frac{\alpha}{2}}$ est le quantile d'odre $1-\dfrac{\alpha}{2}$ de $\mathcal{N}(0,1)$.*
En effet, $f(a)=f(b)\Leftrightarrow a=-b$ et si $X\hookrightarrow\mathcal{N}(0,1)$

$$\mathbb{P}(-b\le X\le b)=1-\alpha \Leftrightarrow \mathbb{P}(X\le b)=1-\frac{\alpha}{2}.$$

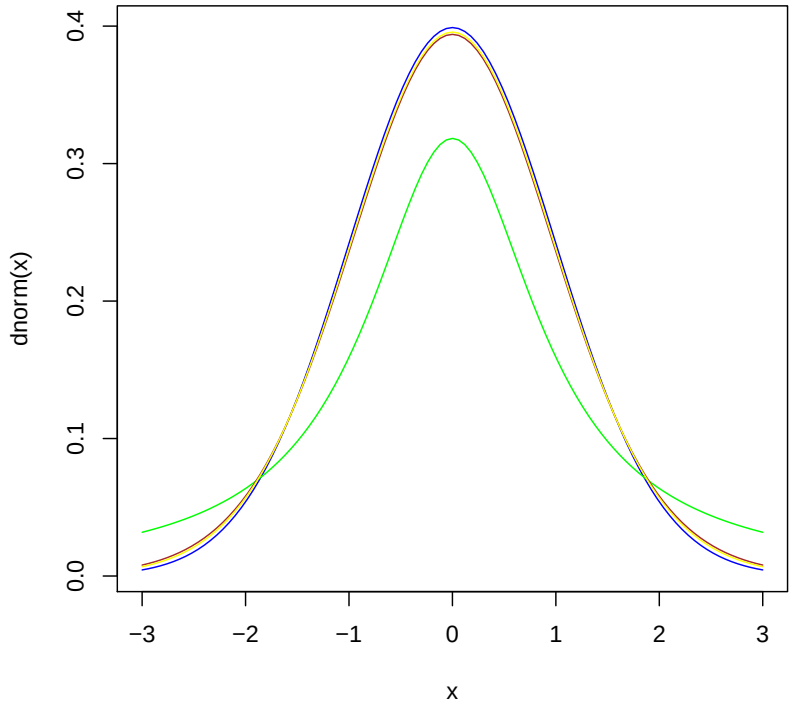
2. La loi de Student $T(n)$. L'intervalle le plus court est de la forme $[-b,b]$ où $b=t^{(n)}_{1-\frac{\alpha}{2}}$ est le quantile d'odre $1-\dfrac{\alpha}{2}$ de $T(n)$.

Loi normale centrée-réduite

```
> curve(dnorm(x),-3,3)
```


dnorm(x)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

Loi de Student



Proposition 7.2.2. *Nous avons le résultat suivant :*

$$T(n) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

7.3 Exemples

7.3.1 Intervalle de confiance pour la moyenne d'une loi normale

Considérons un échantillon $(X_1,\ldots,X_n)$ issu d'une loi normale $\mathcal{N}(\mu,\sigma^2)$ avec $\theta=(\mu,\sigma^2)$.

Si $X \hookrightarrow \mathcal{N}(\mu,\sigma^2)$ alors $$\frac{X-m}{\sigma} \hookrightarrow \mathcal{N}(0,1)$$

1. **σ^2 connue et estimation de μ.** Nous savons que $\overline{X}_n$ est un estimateur efficace de μ. De plus

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma} = \frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \hookrightarrow \mathcal{N}(0,1).$$

Par suite $\dfrac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}$ est une fonction pivot. Ainsi, nous obtenons

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

soit

$$\mathbb{P}\left(\overline{X}_n - \frac{\sigma z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{\sigma z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) = 1-\alpha$$

i.e.

L'intervalle de confiance de niveau $1-\alpha$ de la moyenne μ lorsque σ^2 est connue est

$$\left[\overline{\mathbf{X}}_{\mathbf{n}} - \mathbf{z}_{\mathbf{1-\frac{\alpha}{2}}}\frac{\sigma}{\sqrt{\mathbf{n}}}, \overline{\mathbf{X}}_{\mathbf{n}} + \mathbf{z}_{\mathbf{1-\frac{\alpha}{2}}}\frac{\sigma}{\sqrt{\mathbf{n}}}\right]$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0,1)$

Remarque 7.3.1. *On appelle marge d'erreur la quantité*

$$ME = z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Taille d'échantillon. Fixons $\varepsilon > 0$. Nous cherchons à choisir une taille d'échantillon telle que $ME \leq \varepsilon$. Ainsi, on cherche la taille n d'échantillon tel que

$$|\mu - \bar{X}_n| \leq z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \varepsilon$$

c'est à dire

$$n \geq \frac{\sigma^2 z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2}.$$

2. σ^2 inconnue et estimation de μ. Nous avons le résultat suivant

$$\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{S} \hookrightarrow T(n-1) \quad \text{avec} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2.$$

Cette variable aléatoire est une fonction pivotale pour μ. De plus la densité de la loi de Student vérifie les hypothèses de la Proposition 7.2.1. Ainsi,

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{S} \leq t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté. Il s'ensuit que

$$\mathbb{P}\left(\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

L'intervalle de confiance pour μ de niveau $1-\alpha$ lorsque σ^2 est inconnue est

$$\left[\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté $\mathcal{T}(n-1)$

Nous remarquons que $|\mu - \bar{X}_n| \leq t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$.

Remarque 7.3.2. *On appelle marge d'erreur la quantité*

$$ME = t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}.$$

Taille d'échantillon. Fixons $\varepsilon > 0$. Nous cherchons à choisir une taille d'échantillon telle que $ME \leq \varepsilon$. Ainsi, on cherche la taille n d'échantillon tel que

$$|\mu - \bar{X}_n| \leq t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \varepsilon$$

c'est à dire

$$n \geq \frac{S^2 t^2_{1-\frac{\alpha}{2}}}{\varepsilon^2}.$$

7.3.2 Intervalle de confiance pour la variance d'une loi normale

1. **μ connue et estimation de σ^2.** Nous savons que $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ est un bon estimateur de σ^2. On déduit alors que

$$\frac{nV^2}{\sigma^2} \hookrightarrow \chi^2(n).$$

Ainsi, nous avons

$$\mathbb{P}\left(a \leq \frac{nV^2}{\sigma^2} \leq b\right) = 1-\alpha$$

$$\mathbb{P}\left(\frac{nV^2}{\sigma^2} < a\right) + \mathbb{P}\left(\frac{nV^2}{\sigma^2} > b\right) = \alpha.$$

Ainsi $a = \chi^{(n)}_{\alpha_2}$ et $b = \chi^{(n)}_{1-\alpha_1}$ avec $\alpha_1 + \alpha_2 = \alpha$. On déduit que

L'intervalle de confiance de niveau $1-\alpha$ pour σ^2 lorsque la moyenne μ est connue est :

$$\left[\frac{nV^2}{\chi^{(n)}_{1-\alpha_1}}, \frac{nV^2}{\chi^{n}_{\alpha_2}}\right].$$

2. **μ inconnue et estimation de σ^2.** Nous avons

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

Ainsi, nous avons

$$\mathbb{P}\left(q_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq q_2\right) = 1-\alpha$$

$$\mathbb{P}\Big[\frac{(n-1)S^2}{\sigma^2} < q_1\Big] + \mathbb{P}\Big[\frac{(n-1)S^2}{\sigma^2} > q_2\Big] = \alpha.$$

Ainsi $q_1 = \chi^{(n-1)}_{\alpha_2}$ et $q_2 = \chi^{(n-1)}_{1-\alpha_1}$ avec $\alpha_1 + \alpha_2 = \alpha$. On déduit que

L'intervalle de confiance de niveau $1-\alpha$ pour σ^2 lorsque la moyenne μ est inconnue est :

$$\Big[\frac{(n-1)S^2}{\chi^{(n-1)}_{1-\alpha_2}}, \frac{(n-1)S^2}{\chi^{(n-1)}_{\alpha_1}}\Big].$$

7.3.3 Intervalle de confiance pour une proportion

On considère un échantillon $(X_1,\ldots,X_n)$ issu de la loi de Bernouilli $\mathscr{B}(1,p)$, $p \in]0,1[$. D'après le Théorème Central limite, nous avons :

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

On remplace alors le numérateur $\sqrt{p(1-p)}$ et $\sqrt{\overline{X}_n(1-\overline{X}_n)}$ et on obtient toujours

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

Pour n assez grand,

$$\begin{aligned}
&\mathbb{P}\Big[-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \leq z_{1-\frac{\alpha}{2}}\Big] \\
&= \mathbb{P}\Big[\overline{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p \leq \overline{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\Big] \\
&= 1-\alpha.
\end{aligned}$$

où $z_{1-\frac{\alpha}{2}}$ est quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée-réduite.

L'intervalle de confiance pour la proportion p de niveau de confiance $1-\alpha$ est :

$$\Big[\overline{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\Big]$$

La marge d'erreur est donc

$$ME = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq z_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}$$

car pour tout $x \in [0,1]$, on a

$$\sqrt{x(1-x)} \leq \frac{1}{2}.$$

Pour déterminer la taille n telle que $ME \leq \varepsilon$, il suffit donc de résoudre

$$z_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \leq \varepsilon.$$

Ce qui nous donne alors

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{2\varepsilon}\right)^2.$$

Exercice 7.3.1. *Une compagnie prélève un échantillon de 50 chèques parmi les 2 500 reçus en une journée donnée. On suit le parcours des chèques jusqu'au moment de leur dépôt dans le compte de la compagnie. On constate que 18 des 50 chèques ont mis plus de 5 jours à être déposés.*

1. *Déterminer un intervalle de confiance à 95% pour la proportion p de chèques dont le délai (entre la réception et le dépôt) excède 5 jours.*
2. *Déterminez un intervalle de confiance à 95% pour le nombre de chèques dont le délai excède 5 jours.*
3. *Supposons qu'on veuille faire un échantillonnage sur les chèques de l'année entière (au nombre de 650 000). À un niveau de 95%, quelle est la taille de l'échantillon qu'il faudrait prélever dans les conditions suivantes (vous prendrez pour p l'estimation que vous obtenez avec l'échantillon que vous venez de prélever) ?*
   (a) *si on accepte une marge de 2% dans l'estimation de la proportion ;*
   (b) *si on accepte une marge d'erreur relative (voir le numéro précédent) de 5% de la proportion réelle ;*
   (c) *si on accepte une marge d'erreur de 10 000 chèques dans l'estimation du nombre de chèques qui accusent un délai de plus de 5 jours.*

7.3.4 Intervalle de confiance pour la moyenne d'une loi quelconque

On considère un échantillon $(X_1,\ldots,X_n)$ issu d'une loi de probabilité admettant une moyenne m et une variance σ^2. D'après le Théorème central limite, nous avons le résultat suivant :

$$\frac{\sqrt{n}(\overline{X}_n - m)}{S_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

L'intervalle de confiance pour m de niveau asymptotique $1-\alpha$ est donné par

$$\left[\overline{X}_n - z_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}\right]$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de $\mathcal{N}(0,1)$. Les approximations ci-dessus sont valables si la taille de l'échantillon est suffisamment grande ($n \geq 30$)

Chapitre 8

Généralités sur les tests d'hypothèses

8.1 Principe des tests

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi $\mathbb{P}_\theta$ avec $\theta \in \Theta$. Soient Θ_0 et Θ_1 deux sous-ensembles de Θ tels que $\Theta = \Theta_0 \cup \Theta_1$ et $\Theta_0 \cap \Theta_1 = \emptyset$. Soientles hypothèses :

$H_0 : \theta \in \Theta_0$

$H_1 : \theta \in \Theta_1$

L'hypothèse H_0 est appelée hypothèse nulle et H_1, hypothèse alternative. Une hypothèse est dite simple si elle est réduite à un singléton. Les deux hypothèses sont telles que une et une seule est vraie.

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses à partir des résultats d'un échantillon. La décision consiste à choisir H_0 ou H_1. Il y a quatre cas qui sont reproduits dans le tableau ci-dessous

	H_0 vraie	H_1 vraie
H_0 décidée	Bonne décision	Erreur de deuxième espèce
H_1 décidée	Erreur de première espèce	Bonne décision

Exemple 8.1.1. ***Contrôle de qualité.*** *Une machine produit des pièces classées soit "bonnes" codées par 0, soit "défectueuses" codées par 1. Le nombre de pièces fabriquées étant gigantesque et l'examen de chaque pièce étant relativement coûteux, on ne peut évaluer la qualité de sa production que sur un lot de taille n faible au regard de la production. On observe alors ce lot de n pièces et on note $(x_1,\dots,x_n)$ les observations.*
Modélisation : *on suppose que x_i est la réalisation d'une variable aléatoire X_i de loi de Bernouilli $\mathcal{B}(1,p)$, $p \in]0,1[$; nous faisons les hypothèses suivantes :*

- $X_1,\dots,X_n$ ***sont indépendantes :*** *on admet que des petites variations aléatoires pouvant influer sur la qualité des pièces ne se repercutent pas d'une pièce à une autre.*
- $X_1,\dots,X_n$ ***sont identiquement distribuées :*** *on admet que la production a été stable durant la période d'observation ; cette stabilité est caractérisée par la constance de la probabilité p pour chaque pièce produite d'être défectueuse.*

Nous considérons le problème de test de H_0 : ***la machine est aux normes*** *contre H_1 :* ***la machine n'est pas aux normes.***

- *Erreur de première espèce : décider que la machine n'est pas aux normes alors qu'en réalité elle est aux normes : dépenses inutiles de réparation ou de changement de matériels.*
- *Erreur de deuxième espèce : décider que la machine est aux normes alors qu'en réalité elle n'est pas aux normes : production de mauvaises pièces pouvant aboutir à un mécontentement de la clientèle, voire à des problèmes de sécurité.*

Définition 8.1.1. *On appelle test une statistique $\psi(X_1,\dots,X_n)$ à valeurs dans $\{0,1\}$ telle que*

$$\psi(X_1,\dots,X_n)=0 \Longrightarrow \text{on accepte } H_0$$
$$\psi(X_1,\dots,X_n)=1 \Longrightarrow \text{on accepte } H_1.$$

Définition 8.1.2. *On appelle région critique la région d'acceptation de l'hypothèse alternative H_1 :*

$$W=\Big\{(X_1,\dots,X_n):\psi(X_1,\dots,X_n)=1\Big\}.$$

Un test est caractérisé par sa région critique.

Définition 8.1.3. *On appelle risque de première espèce du test $\psi(X_1,\dots,X_n)$ la probabilité de l'erreur de première espèce :*

$$\begin{aligned}\alpha_\psi : \Theta_0 &\longrightarrow [0,1]\\ \theta &\longmapsto \mathbb{P}_\theta(W).\end{aligned}$$

Définition 8.1.4. *On appelle niveau du test $\psi(X_1,\dots,X_n)$ la quantité*

$$\sup_{\theta\in\Theta}\alpha_\psi(\theta).$$

Le test $\psi(X_1,\dots,X_n)$ est dit de niveau $\alpha\in(0,1)$ si

$$\sup_{\theta\in\Theta}\alpha_\psi(\theta)=\alpha.$$

Remarque 8.1.1. *Le niveau du test est le plus gros risque de première espèce possible.*

Définition 8.1.5. *On appelle risque de deuxième espèce du test $\psi(X_1,\dots,X_n)$ la probabilité de l'erreur de deuxième espèce :*

$$\begin{aligned}\beta_\psi : \Theta_1 &\longrightarrow [0,1]\\ \theta &\longmapsto \mathbb{P}_\theta(\overline{W}).\end{aligned}$$

L'idéal serait de diminuer les deux risques d'erreur en même temps. Malheureusement, on montre qu'ils varient en sens inverse. Dans la pratique des tests statistiques, il est de règle de se fixer α, ce qui fait jouer à H_0 un rôle prééminent.
Un test est déterminé par sa région critique W. La région critique dépend du niveau α et d'une statistique appelée variable de décision. Pour la déterminer, il est indispensable de connaître la loi de la variable de décision sous l'hypothèse H_0. Lorsque $(x_1,\dots,x_n)$ sont des valeurs observées de cet échantillon,

- si $(x_1,\dots,x_n)\in W$, alors on rejette H_0 et on accepte H_1 ;
- si $(x_1,\dots,x_n)\notin W$, alors on accepte H_0 et on rejette H_1.

Définition 8.1.6. *On appelle puissance du test $\psi(X_1,\dots,X_n)$ la probabilité d'accepter H_1 quand H_1 est vraie :*

$$\begin{aligned}\gamma_\psi : \Theta_1 &\longrightarrow [0,1]\\ \theta &\longmapsto \mathbb{P}_\theta(W).\end{aligned}$$

La puissance
- croît avec le niveau de signification α.
- croît avec la taille del'échantillon
- dépend de la région critique.

Remarque 8.1.2. *Nous avons $\forall\theta\in\Theta_1, \gamma_\psi(\theta) = 1-\beta_\psi(\theta)$.*

Remarque 8.1.3. *Un bon test est un test qui, pour un niveau α donné, maximise la puissance.*

Définition 8.1.7. *Un test $\psi(X_1,\dots,X_n)$ est sans biais lorsque la puissance du test est supérieure au niveau α sur Θ_1 :*

$$\gamma(\theta)\geq\alpha \quad \forall\theta\in\Theta_1.$$

8.2 Etapes des tests

1. Etape préliminaire : modélisation du problème.
2. Formulation des hypothèses H_0 et H_1.
3. Choix du seuil du test α.
4. Choix d'une statistique de test T_n, dont on connaît la loi sous H_0
5. Etude du comportement de T_n sous H_1 et déduction de la forme de la zone critique.
6. Calcul de cette zone pour le niveau α fixé puis confrontation aux données ; et / ou calcul de la p-valeur du test sur les données
7. Conclusion statistique : conservation ou rejet de l'hypothèe de départ H_0 et commentaire éventuel sur la p-valeur.
8. Conclusion stratégique : décision que l'on va prendre une fois éclairé par le résultat statistique.

8.3 Construction d'un test d'hypothèses

Pour construire un test d'hypothèses portant sur la valeur d'un paramètre θ, l'on peut se fier au bon sens. Si on connaît un estimateur $\hat{\theta}_n$ de θ, on pourrait procéder de la façon suivante : soit θ_0 une valeur possible de θ.

- Test de $H_0 : \theta\leq\theta_0$ contre $H_1 : \theta>\theta_0$.
  On rejette H_0 si $\hat{\theta}_n$ est "trop grand" i.e. la région critique est

$$W = \{\hat{\theta}_n - \theta_0 > l_\alpha\}.$$

- Test de $H_0 : \theta\geq\theta_0$ contre $H_1 : \theta<\theta_0$.
  On rejette H_0 si $\hat{\theta}_n$ est "trop petit" i.e. la région critique est

$$W = \{\hat{\theta}_n - \theta_0 < l_\alpha\}.$$

- Test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.
  On rejette H_0 si $\left|\hat{\theta}_n - \theta_0\right|$ est "trop grand" i.e. la région critique est

$$W = \left\{\left|\hat{\theta}_n - \theta_0\right| > l_\alpha\right\}.$$

- Test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$.
  - $W = \{\hat{\theta}_n > l_\alpha\}$ si $\theta_1 > \theta_0$
  - $W = \{\hat{\theta}_n < l_\alpha\}$ si $\theta_1 < \theta_0$.

Pour déterminer l_α, il faut résoudre l'équation $\mathbb{P}_{\theta_0}(W) = \alpha$.

8.4 La p-value

En pratique, plutôt que de calculer la région critique en fonction de α, on préfère donner un seuil critique de α^* appelée p-value, qui est telle que

- si $\alpha^* < \alpha$, on rejette H_0
- si $\alpha < \alpha^*$, on accepte H_0.

Les logiciels statistiques calculent et présentent les p-valeurs qui sont difficiles à obtenir sans moyen de calcul approprié.

Chapitre

9 Test d'hypothèse simple contre hypothèse simple

9.1 Théorème de Neyman-Pearson

Notons $L(X_1,\dots,X_n,\theta)$ la vraisemblance de l'échantillon $(X_1,\dots,X_n)$. Soient θ_0 et θ_1 deux éléments de Θ tels que $\theta_0 \neq \theta_1$. L'objectif est de tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ au seuil α.

Théorème 9.1.1. *Pour tout $\alpha \in]0,1[$, il existe une constante $k_\alpha \in \mathbb{R}_+$ telle que le meilleur test au niveau α a pour région critique*

$$W = \left\{(X_1,\dots,X_n) : \frac{L(X_1,\dots,X_n,\theta_1)}{L(X_1,\dots,X_n,\theta_0)} > k_\alpha\right\}.$$

La constante k_α est déterminé par l'équation

$$\mathbb{P}_{\theta_0}[W] = \mathbb{P}_\theta\left(\frac{L(X_1,\dots,X_n,\theta_1)}{L(X_1,\dots,X_n,\theta_0)} > k_\alpha\right) = \alpha.$$

9.2 Exemples

9.2.1 Test sur une proportion

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,p)$ avec $p \in]0,1[$ inconnue. On veut tester $H_0 : p = 1/4$ contre $H_1 : p = 1/2$ au seuil $\alpha = 0.05$.
Le rapport de vraisemblance est : Nous avons

$$\frac{L(X_1,\dots,X_n,p_1)}{L(X_1,\dots,X_n,p_0)} = \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{\sum_{i=1}^n X_i}\left(\frac{1-p_1}{1-p_0}\right)^n$$

Si $p_0 = 1/4$ et $p_1 = 1/2$, nous obtenons

$$\frac{L(X_1,\dots,X_n,1/2)}{L(X_1,\dots,X_n,1/4)} = 3^{\sum_{i=1}^n X_i}(2/3)^n$$

Ainsi, nous avons

$$\begin{aligned}\frac{L(X_1,\dots,X_n,1/2)}{L(X_1,\dots,X_n,1/4)} > k_\alpha &\Leftrightarrow 3^{\sum_{i=1}^n X_i}(2/3)^n > k_\alpha \\ &\Leftrightarrow \sum_{i=1}^n X_i > \frac{\ln\left((\frac{3}{2})^n k_\alpha\right)}{\ln(3)} = K_\alpha.\end{aligned}$$

D'après Neyman-Pearson, la région critique est de la forme :

$$W = \left\{ \sum_{i=1}^{n} X_i > K_\alpha \right\}$$

La constante K_α est déterminée par

$$\mathbb{P}_{1/4}\left[\sum_{i=1}^{n} X_i > K_\alpha \right] = \alpha.$$

Si la taille de l'échantillon est suffisamment grand $(n > 30)$, nous pouvons utiliser le Théoréme Central Limite qui permet d'approximer la loi de $\sum_{i=1}^{n} X_i$ par la loi normale $\mathcal{N}(np, np(1-p))$.

Détermination de K_α. Sous H_0, $\sum_{i=1}^{n} X_i$ suit approximativement la loi normale $\mathcal{N}\left(\frac{n}{4}, \frac{3n}{16}\right)$ et nous avons

$$\begin{aligned}
\alpha &= \mathbb{P}_{1/4}\left[\sum_{i=1}^{n} X_i > K_\alpha \right] \\
&= \mathbb{P}_{1/4}\left[\frac{\sum_{i=1}^{n} X_i - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} > \frac{K_\alpha - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \right] \\
&= 1 - \mathbb{P}_{1/4}\left[\frac{\sum_{i=1}^{n} X_i - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \leq \frac{K_\alpha - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \right] \\
&= 1 - \Phi\left(\frac{K_\alpha - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \right).
\end{aligned}$$

Ce qui implique

$$\Phi\left(\frac{K_\alpha - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \right) = 1 - \alpha.$$

Soit $u_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de $\mathcal{N}(0,1)$. Alors, nous avons

$$u_{1-\alpha} = \frac{K_\alpha - \frac{n}{4}}{\sqrt{\frac{3n}{16}}} \Longleftrightarrow K_\alpha = u_{1-\alpha}\sqrt{\frac{3n}{16}} + \frac{n}{4}.$$

La région critique du test optimal est :

$$W = \left\{ (X_1, \ldots, X_n) : \sum_{i=1}^{n} X_i > u_{1-\alpha}\sqrt{\frac{3n}{16}} + \frac{n}{4} \right\}.$$

Sous l'alternative $\sum_{i=1}^{n} X_i$ suit la loi $\mathcal{N}\left(\frac{n}{2}, \frac{n}{4}\right)$. La puissance du test est donnée par

$$\begin{aligned}\gamma_n &= \mathbb{P}_{1/2}\left[\sum_{i=1}^{n} X_i > K_\alpha\right] \\ &= \mathbb{P}_{1/2}\left[\frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{\frac{n}{4}}} > \frac{K_\alpha - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right] \\ &= 1 - \Phi\left(\frac{\sqrt{3}u_{1-\alpha} - \sqrt{n}}{2}\right)\end{aligned}$$

On remarque que $\lim_{n\to+\infty} \gamma_n = 1$. On dit que le test est asymptotiquement puissant :

Exercice 9.2.1. *On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,p)$ avec $p \in]0,1[$. Tester $H_0 : p = 1/2$ contre $H_1 : p = 1/4$ au seuil $\alpha = 0.05$.*

9.2.2 Test sur la moyenne d'un échantillon gaussien

On considère un échantillon $(X_1,\dots,X_n)$ issu d'une loi normale $\mathcal{N}(m,\sigma^2)$ avec $m \in \mathbb{R}$ inconnue et $\sigma^2 > 0$. On veut tester $H_0 : m = m_0$ contre $H_1 : m = m_1$ au niveau $\alpha = 0.05$ avec $m_1 > m_0$.

Exercice 9.2.2. *Tester $H_0 : m = m_0$ contre $H_1 : m = m_1$ au niveau $\alpha = 0.05$ avec $m_1 < m_0$.*

Exercice 9.2.3. *Tester $H_0 : m = m_0$ contre $H_1 : m = m_1$ au niveau $\alpha = 0.05$ avec $m_1 > m_0$.*

Chapitre

10 Tests de Student : un échantillon

10.1 Introduction

On appelle test de Student un test de comparaison de la moyenne dans un échantillon gaussien, c'est à dire un échantillon $(X_1,\dots,X_n)$ issu de la loi normale $\mathcal{N}(m,\sigma^2)$. Soit m_0 une valeur possible de m. La moyenne empirique $\overline{X}_n$ est un estimateur efficace de m.
Deux résultats importants :

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(m,\frac{\sigma^2}{n}\right) \Longleftrightarrow \frac{\sqrt{n}\left(\overline{X}_n - m\right)}{\sigma} \hookrightarrow \mathcal{N}(0,1).$$

$$\frac{\sqrt{n}\left(\overline{X}_n - m\right)}{S_n} \hookrightarrow \mathcal{T}(n-1)$$

qui est la loi de Student à $n-1$ dégrés de liberté avec

$$S_n = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2}.$$

10.2 $H_0 : m \leq m_0$ contre $H_1 : m > m_0$

10.2.1 On suppose que la variance σ^2 est connue.

En se référant à la Section 8.3, nous obtenons une première forme de la région critique

$$W = \left\{\overline{X}_n - m_0 > l_\alpha\right\},$$

où la constante l_α est déterminée par (le test étant de niveau α)

$$\mathbb{P}_{m_0}\left(\overline{X}_n - m_0 > l_\alpha\right).$$

Sous l'hypothèse H_0,

$$\overline{X}_n \hookrightarrow \mathcal{N}\left(m_0, \frac{\sigma^2}{n}\right) \Longleftrightarrow \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} \hookrightarrow \mathcal{N}(0,1).$$

Ce qui implique alors

$$\mathbb{P}_{m_0}\left(\frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} > \frac{\sqrt{n}l_\alpha}{\sigma}\right) = \alpha.$$

Ainsi, on en déduit que

$$\frac{\sqrt{n}l_\alpha}{\sigma} = q_{1-\alpha} \Leftrightarrow l_\alpha = \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $\mathcal{N}(0,1)$.

La région critique au niveau α du test $H_0 : m \leq m_0$ contre $H_1 : m > m_0$ lorsque σ^2 est connue est

$$\begin{aligned} W &= \left\{\overline{X}_n - m_0 > \frac{\sigma}{\sqrt{n}} q_{1-\alpha}\right\} \\ &= \left\{\frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} > q_{1-\alpha}\right\} \end{aligned} \tag{10.2.1}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée-réduite.

Remarque 10.2.1. *On accepte H_1 au niveau α lorsque la différence $\overline{X}_n - m_0$ est significative, c'est à dire strictement supérieure à $\frac{\sigma}{\sqrt{n}} q_{1-\alpha}$.*

Exercice 10.2.1. *Une marque de tablettes de chocolat annonce que ses tablettes contiennent une teneur en cacao supérieure à 430 g par kg. On effectue un contrôle de qualité sur un échantillon de 10 tablettes et on obtient les teneurs suivantes en g/kg : 505.1 423.5 462.0 391.9 412.1 487.2 439.0 434.1 441.1 474.2. On admet que chaque mesure suit une loi normale $\mathcal{N}(m, \sigma^2)$.*

1. *Ecrire le modèle et les hypothèses du test qu'on veut faire.*
2. *On admet dans un premier temps (au vu de contrôles antérieurs) que $\sigma = 24$. Que peut-on conclureau niveau $\alpha = 0.05$?*

Solution 10.2.1.

1. — *Soit X_i la teneur en cacao en g/kg de la tablette i. La variable aléatoire X_i suit une loi normale $\mathcal{N}(m, \sigma^2)$. On dispose d'un échantillon $(X_1, \ldots, X_{10})$ issu d'une loi normale $\mathcal{N}(m, \sigma^2)$.*
   — *Le modèle statistique est donc $\left\{\mathcal{N}(m, \sigma^2) : (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\right\}$*
   — *$H_0 : m \leq 430$ contre $H_1 : m > 430$.*
2. *Au niveau $\alpha = 0.05$, la région critique du test est :*

$$W = \left\{\frac{\sqrt{10}\left(\overline{X}_{10} - 430\right)}{24} > q_{0.95}\right\}$$

où $q_{0.95} = 1.644$ est le quantile d'ordre 0.95 de la loi normale centrée-réduite. Par suite, nous obtenons :

$$W = \left\{\frac{\sqrt{10}\left(\overline{X}_{10} - 430\right)}{24} > 1.644\right\}$$

Puisque

$$\overline{x}_{10} = \frac{1}{10}(505.1+423.5+462.0+391.9+412.1+487.2+439.0+434.1+441.1+474.2) = 447.02$$

et

$$\frac{\sqrt{10}(447.02-430)}{24} = 2.243 > 1.644,$$

on accepte H_1 au niveau $\alpha = 0.05$. Ainsi, on peut conclure que les tablettes de cette marque contiennent une teneur en cacao supérieure à 430 g par kg.

10.2.2 On suppose σ^2 est inconnue

Nous allons remplacer dans (10.2.1), σ par par l'écart-type empirique modifié S_n.

La région critique au niveau α du test $H_0 : m \leq m_0$ contre $H_1 : m > m_0$ lorsque σ^2 est inconnue est

$$W = \left\{ \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{S_n} > t_{1-\alpha,n-1} \right\}$$

où $t_{1-\alpha,n-1}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté $\mathcal{T}(n-1)$.

Exercice 10.2.2. *Une marque de tablettes de chocolat annonce que ses tablettes contiennent une teneur en cacao supérieure à 430 g par kg. On effectue un contrôle de qualité sur un échantillon de 10 tablettes et on obtient les teneurs suivantes en g/kg : 505.1 423.5 462.0 391.9 412.1 487.2 439.0 434.1 441.1 474.2. On admet que chaque mesure suit une loi normale $\mathcal{N}(m,\sigma^2)$. Que peut-on conclure au niveau $\alpha = 0.05$?*

Solution 10.2.2. *Au niveau $\alpha = 0.05$, nous voulons tester $H_0 : m \leq 430$ contre $H_1 : m > 430$. La région critique du test est :*

$$W = \left\{ \frac{\sqrt{10}\left(\overline{X}_{10} - 430\right)}{S_{10}} > t_{0.95,9} \right\}$$

où $t_{0.95,9} = 1.833$ est le quantile d'ordre 0.95 de la loi de Student à 9 degrés de liberté. Par suite, nous obtenons :

$$W = \left\{ \frac{\sqrt{10}\left(\overline{X}_{10} - 430\right)}{35} > 1.833 \right\}$$

Puisque

$$\overline{x}_{10} = \frac{1}{10}(505.1+423.5+462.0+391.9+412.1+487.2+439.0+434.1+441.1+474.2) = 447.02$$

et

$$\frac{\sqrt{10}(447.02-430)}{35} = 1.5378 < 1.833,$$

on rejette H_1 au niveau $\alpha = 0.05$. Ainsi, on peut conclure que les tablettes de cette marque ne contiennent pas une teneur en cacao supérieure à 430 g par kg.

10.3 $H_0 : m \geq m_0$ contre $H_1 : m < m_0$

10.3.1 On suppose que la variance σ^2 est connue.

La région critique au niveau α du test $H_0 : m \geq m_0$ contre $H_1 : m < m_0$ lorsque σ^2 est connue est

$$W = \left\{ \overline{X}_n < m_0 + \frac{\sigma}{\sqrt{n}} q_\alpha \right\} = \left\{ \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} < q_\alpha \right\} \tag{10.3.1}$$

où q_α est le quantile d'ordre α de la loi normale centrée-réduite.

Exercice 10.3.1. *Le département de contrôle de la qualité d'une entreprise détermine que le poids moyen net d'une boîte de céréales ne devrait pas être inférieur à* $200\ g$. *L'expérience a montré que les poids sont approximativement distribués normalement avec un écart-type de* $15\ g$. *Un échantillon de* 15 *boîtes prélevé aléatoirement sur la ligne de production donne un poids moyen de* $195\ g$. *Cela est-il suffisant pour pouvoir affirmer que le poids moyen des boîtes est inférieur à* $200\ g$ *?*

Solution 10.3.1.

1. *Tester* $H_0 : m \geq 200$ *contre* $H_1 : m < 200$ *au niveau* $\alpha = 0.05$
2. *Au niveau* $\alpha = 0.05$, *la région critique du test est*

$$W = \left\{ \overline{X}_{15} < 200 + \frac{15}{\sqrt{15}} q_{0.05} \right\}$$

   où $q_{0.05} = -q_{0.95} = -1.644$ *est le quantile d'ordre* 0.05 *de la loi normale centrée-réduite.* $200 - \frac{15}{\sqrt{15}} * 1.64 = 193.65$
3. *Puisque* $195 > 193.65$, *on accepte* H_0. *Même si* $\bar{x} < 200\ g$, *il n'y a pas d'éléments significatifs indiquant que le poids moyen des boites est inférieure à* $200\ g$.

10.3.2 On suppose que la variance σ^2 est inconnue.

La région critique au niveau α du test $H_0 : m \geq m_0$ contre $H_1 : m < m_0$ lorsque σ^2 est inconnue est

$$W = \left\{ \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{S_n} < t_{\alpha,n-1} \right\} \tag{10.3.2}$$

où $t_{\alpha,n-1}$ est le quantile d'ordre α de la loi de Student à $n-1$ degrés de liberté $\mathcal{T}(n-1)$.

Exercice 10.3.2. *Le département de contrôle de la qualité d'une entreprise détermine que le poids moyen net d'une boîte de céréales ne devrait pas être inférieur à* $200\ g$. *L'expérience a montré que les poids sont approximativement distribués normalement. Un échantillon de* 15 *boîtes prélevé aléatoirement sur la ligne de production donne un poids moyen de* $195\ g$ *avec un écart-type estimé égal à* $15\ kg$.. *Cela est-il suffisant pour pouvoir affirmer que le poids moyen des boîtes est inférieur à* $200\ g$ *?*

Solution 10.3.2. *1. Tester $H_0 : m \geq 200$ contre $H_1 : m < 200$ au niveau $\alpha = 0.05$*

2. Au niveau $\alpha = 0.05$, la région critique du test est

$$W = \left\{ \frac{\sqrt{15}\left(\overline{X}_{15} - 200\right)}{S_{15}} < t_{0.05,14} \right\}$$

où $t_{0.05,14} = -1.761$ est le quantile d'ordre 0.05 de la loi de Student à 14 degrés de liberté ($\mathcal{T}(14)$).

3. Puisque $\frac{\sqrt{15}(195-200)}{15} = -1.291 > -1.761$, on accepte H_0.Au niveau $\alpha = 0.05$, il n'y a pas d'éléments significatifs indiquant que le poids moyen des boites est inférieure à 200 g.

10.4 $H_0 : m = m_0$ contre $H_1 : m \neq m_0$

La région critique au niveau α du test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ lorsque σ^2 est connue est

$$W = \left\{ \left| \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} \right| > q_{1-\frac{\alpha}{2}} \right\} \tag{10.4.1}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée-réduite.

Exercice 10.4.1. *Une entreprise de vente par correspondance demande un montant fixe pour les frais d'envoi, indépendamment du poids du colis. Une étude réalisée il y a quelques années a montré que le poids moyen d'un colis était de 17.5 kg avec un écart-type de 3.6 kg. La comptabilité soupçonne que le poids moyen est maintenant différent de 17.5 kg. Un échantillon aléatoire de 100 colis est prélevé et fournit un poids moyen de $\bar{x} = 18.4$ kg. On suppose que les poids des colis sont distribués normalement. Que conclure au niveau $\alpha = 0.05$*

Solution 10.4.1. *1. Nous voulons tester l'hypothèse $H_0 : m = 17.5$ contre $H_1 : m \neq 17.5$ au niveau $\alpha = 0.05$.*

2. Au niveau $\alpha = 0.05$, la région critique du test est

$$\begin{aligned} W &= \left\{ \left| \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{\sigma} \right| > q_{0.975} \right\} \\ &= \left\{ \overline{X}_n < m_0 - \frac{\sigma}{\sqrt{n}} q_{0.975} \right\} \cup \left\{ \overline{X}_n > m_0 + \frac{\sigma}{\sqrt{n}} q_{0.975} \right\} \end{aligned}$$

où $q_{0.975} = 1.96$ est le quantile d'ordre 0.975 de la loi normale centrée-réduite.

$$\begin{aligned} m_0 + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} &= 17.5 + \frac{3.6}{\sqrt{100}} * 1.96 = 18.2056 \\ m_0 - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} &= 17.5 - \frac{3.6}{\sqrt{100}} * 1.96 = 16.7944 \end{aligned}$$

3. Puisque $\bar{x} > 18.2056$, on rejette H_0 i.e le poids moyen des colis a changé.

10.4.1 On suppose que la variance σ^2 est inconnue.

La région critique au niveau α du test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ lorsque σ^2 est inconnue est

$$W = \left\{ \left| \frac{\sqrt{n}\left(\overline{X}_n - m_0\right)}{S_n} \right| > t_{1-\frac{\alpha}{2},n-1} \right\} \tag{10.4.2}$$

où $t_{1-\frac{\alpha}{2},n-1}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté $\mathcal{T}(n-1)$.

Exercice 10.4.2. *Une entreprise de vente par correspondance demande un montant fixe pour les frais d'envoi, indépendamment du poids du colis. Une étude réalisée il y a quelques années a montré que le poids moyen d'un colis était de* 17.5 *kg. La comptabilité soupçonne que le poids moyen est maintenant différent de* 17.5 *kg. Un échantillon aléatoire de* 100 *colis est prélevé et fournit un poids moyen de* $\bar{x} = 18.4$ *kg avec un écat-type estimé égal à* 3.6*. On suppose que les poids des colis sont distribués normalement. Que conclure au niveau* $\alpha = 0.05$

Solution 10.4.2.

1. *Nous voulons tester l'hypothèse* $H_0 : m = 17.5$ *contre* $H_1 : m \neq 17.5$ *au niveau* $\alpha = 0.05$.
2. *Au niveau* $\alpha = 0.05$*, la région critique du test est :*

$$W = \left\{ \left| \frac{\sqrt{100}\left(\overline{X}_{100} - 17.5\right)}{S_{100}} \right| > t_{0.975,99} \right\}$$

où $t_{0.975,100} = 1.9842$ *est le quantile d'ordre* 0.975 *de la loi de Student à* 99 *degrés de liberté* $\mathcal{T}(99)$.

3. *Puisque* $\dfrac{\sqrt{100}(18.4-17.5)}{3.6} = 2.5 > 1.9842$*, on rejette* H_0 *i.e le poids moyen des colis a changé.*

Chapitre

11 Tests de Student : deux échantillons

11.1 Introduction

Soient P_1 et P_2 deux populations. On étudie un caractère (rendement, chiffre d'affaire, seuil de perception, etc.) sur ces deux populations. Le caractère a pour espérance m_1 et pour variance σ_1^2 dans la population P_1 et a pour espérance m_2 et pour variance σ_2^2 dans la population P_2. Pour des raisons techniques, on supposera que le caractère est distribué selon une loi normale. On dispose alors de deux échantillons $(X_1,\ldots,X_{n_1})$ et $(Y_1,\ldots,Y_{n_2})$ issus respectivement de P_1 et P_2, tels que X_i et Y_j sont indépendantes :

- $(X_1,\ldots,X_{n_1})$ est issu de $\mathcal{N}(m_1,\sigma_1^2)$
- $(Y_1,\ldots,Y_{n_2})$ est issu de $\mathcal{N}(m_2,\sigma_2^2)$.

Dans cette section, on comparera les moyennes et les variances des deux échantillons. Les moyennes empiriques, variances empiriques modifiées des deux échantillons sont notées respectivement $\overline{X}_{n_1}$, S_1^2, $\overline{Y}_{n_2}$ et S_2^2.

Exemple 11.1.1. *Deux groupes d'étudiants de tailles respectives $n_1 = 25$ et $n_2 = 31$ ont suivi le même cours de statistique et passe le même examen. Les moyennes et écarts-types empiriques des notes obtenues dans les deux groupes sont respectivement :*

	moyenne	*Variance S^2*
Groupe 1	*12.8*	*3.4*
Groupe 2	*11.3*	*2.9*

On suppose que les notes sont reparties dans les deux groupes selon des lois normales et qu'elles sont toutes independantes. Peut-on considérer que le premier groupe est meilleur que le deuxième, c'est-à-dire qu'un point et demi d'écart entre les moyennes est significatif d'une différence de niveau ? La procédure à suivre consiste à tester d'abord l'égalité des variances, puis l'égalité des moyennes.

Exemple 11.1.2. *Deux variétés de blé ont été cultivées chacune sur 8 parcelles ($n_1 = n_2 = 8$). Les rendements observés (en quintaux/hectare) sont regroupés dans le tableau ci-dessus :*

	moyenne	*variance σ^2*
Echantillon 1	*80.0*	*1.00*
Echantillon 2	*81.5*	*1.00*

Si l'on considère que les **16** *parcelles, la variété* **2** *présente en moyenne un rendement supérieur (de* $1.5q/ha$*) à celui de la variété* **1**. *Peut-on généraliser ce résultat ? Autrement dit, la différence observée (de* $1.5q/ha$*) doit être considérée comme une conséquence d'un rendement moyen différent selon la variété ou, au contraire, est-il fortuit ? Selon un autre point de vue, la question peut être posée ainsi : la différence de moyenne obervée doit être imputée au hasard (c'est-à-dire à la variété "naturelle" dite aussi "résiduelle" pour exprimer que l'on ne sait l'expliquer par la statistique) ?*

11.2 Test de Fisher de comparaison des variances

Comparer les variances des deux échantillons revient à résoudre par exemple le problème de test suivant : $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$ est

$$W = \left\{ \frac{S_1^2}{S_2^2} < f^*_{\frac{\alpha}{2}} \right\} \cup \left\{ \frac{S_1^2}{S_2^2} > f^*_{1-\frac{\alpha}{2}} \right\}$$

où $f^*_{\frac{\alpha}{2}}$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Fisher à (n_1-1, n_2-1) degrés de liberté, $f^*_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Fisher à (n_1-1, n_2-1) degrés de liberté et

$$S_{n_1} = \left(\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_i - \overline{X}_{n_1} \right)^2 \right)^{1/2}$$

$$S_{n_2} = \left(\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_i - \overline{Y}_{n_2} \right)^2 \right)^{1/2}.$$

11.3 Test de Student de comparaison des moyennes

On désire maintenant comparer les moyennes. Le test d'égalité des moyennes est :

$$H_0 : m_1 = m_2 \text{ contre } H_0 : m_1 \neq m_2.$$

Lorsque H_0 est vraie, on observe très rarement une parfaite égalité des moyennes. La question est donc de savoir à partir de quel écart de moyenne va-t-on choisir H_1 ?
La région critique est de la forme

$$W = \left\{ \left| \overline{X}_{n_1} - \overline{Y}_{n_2} \right| > l_\alpha \right\}.$$

Pour déterminer l_α, l'on a besoin de la loi de $\overline{X}_{n_1} - \overline{Y}_{n_2}$ sous l'hypothèse H_0. Nous savons que

$$\overline{X}_{n_1} \hookrightarrow \mathcal{N}\left(m_1, \frac{\sigma_1^2}{n_1} \right)$$

$$\overline{Y}_{n_2} \hookrightarrow \mathcal{N}\left(m_2, \frac{\sigma_2^2}{n_2} \right).$$

Comme ces deux variables sont indépendantes, on en déduit que

$$\overline{X}_{n_1} - \overline{Y}_{n_2} \hookrightarrow \mathcal{N}\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right).$$

Ainsi nous avons

$$V = \frac{(\overline{X}_{n_1} - \overline{Y}_{n_2}) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \hookrightarrow \mathcal{N}(0,1).$$

Par suite, sous H_0, nous obtenons

$$V = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \hookrightarrow \mathcal{N}(0,1).$$

11.3.1 Résolution du test lorsque les variances connues

$$W = \left\{ \left| \overline{X}_{n_1} - \overline{Y}_{n_2} \right| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

Exemple 11.3.1. *Revenons à l'exemple 11.1.2. Les variances sont connues,* $\sigma_1^2 = \sigma_2^2 = 1$, $n_1 = n_2 = 8$ *et les rendements moyens observés* $\bar{x}_8 = 80q/h$ *et* $\bar{y}_8 = 81.5q/h$. *On suppose que le seuil du test est* $\alpha = 0.05$. *De ce fait,* $u_{0.975} = 1.96$ *Nous avons donc*

$$u_{0.975}\sqrt{\frac{1}{8} + \frac{1}{8}} = 0.98 \quad \bar{x}_8 - \bar{y}_8 = -1.5 < -0.98.$$

Nous décidons donc de rejeter H_0. *La variété* 2 *a un rendement moyen différent de celui de la variété* 1.

11.3.2 Résolution du test lorsque les variances sont inconnues

Posons

$$Z = \frac{(n_1 - 1)S_{n_1}^2}{\sigma_1^2} + \frac{(n_2 - 1)S_{n_2}^2}{\sigma_2^2}.$$

Comme $\dfrac{(n_1 - 1)S_{n_1}^2}{\sigma_1^2} \hookrightarrow \chi^2(n_1 - 1)$ et $\dfrac{(n_2 - 1)S_{n_2}^2}{\sigma_2^2} \hookrightarrow \chi^2(n_2 - 1)$ et que ces deux variables sont indépendantes, nous obtenons $Z \hookrightarrow \chi^2(n_1 + n_2 - 2)$. De plus, les variables aléatoires Z et V sont indépendantes. Par la définition de la loi de Student, nous déduisons que

$$T_{n_1,n_2} = \frac{V}{\sqrt{\frac{Z}{n_1+n_2-2}}} = \frac{\sqrt{n_1 + n_2 - 2}(\overline{X}_{n_1} - \overline{Y}_{n_2}) - (m_1 - m_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\left(\frac{(n_1-1)S_{n_1}^2}{\sigma_1^2} + \frac{(n_2-1)S_{n_2}^2}{\sigma_2^2}\right)}} \hookrightarrow \mathcal{T}(n_1 + n_2 - 2).$$

Sous l'hypothèse $H_0 : m_1 = m_2$, nous avons

$$T_{n_1,n_2} = \frac{\sqrt{n_1 + n_2 - 2}(\overline{X}_{n_1} - \overline{Y}_{n_2})}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\left(\frac{(n_1-1)S_{n_1}^2}{\sigma_1^2} + \frac{(n_2-1)S_{n_2}^2}{\sigma_2^2}\right)}} \hookrightarrow \mathcal{T}(n_1 + n_2 - 2).$$

On note que lorsque n_1 et n_2 sont grands, le caractère gaussien des observations n'est plus requis, et que T_{n_1,n_2} suit approximativement, sous H_0, une loi $\mathcal{N}(0,1)$..

Supposons que $\sigma_1^2 = \sigma_2^2$.

Si le test de Fisher accepte l'égalité des variances (H_0), nous avons

$$T_{n_1,n_2} = \sqrt{\frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}} \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{(n_1-1)S_{n_1}^2 + (n_2-1)S_{n_2}^2} \hookrightarrow \mathcal{T}(n_1+n_2-2)$$

La région critique au niveau $\alpha \in]0,1[$ est

$$W = \left\{ \left| T_{n_1,n_2} \right| > t_{1-\frac{\alpha}{2},n_1+n_2-2} \right\}$$

où $t_{1-\frac{\alpha}{2},n_1+n_2-2}$ est le quantile d'odre $1-\frac{\alpha}{2}$ de la loi de Student $\mathcal{T}(n_1+n_2-2)$.

Supposons que $\sigma_1^2 \neq \sigma_2^2$.

A priori, si le test de Fisher rejette l'égalité des variances, on ne peut pas appliquer le test. On estime séparément σ_1^2 et σ_2^2 par leurs estimateurs S_1^2 et S_2^2. Posons

$$T_{n_1,n_2} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}}.$$

Sous H_0, $T_{n_1,n_2} \approx T([\nu])$

$$\nu = \frac{\left(\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}\right)^2}{\frac{S_{n1}^4}{n_1^2(n_1-1)} + \frac{S_{n2}^4}{n_2^2(n_2-1)}}.$$

La région critique au niveau $\alpha \in]0,1[$ est

$$W = \left\{ \left| T_{n_1,n_2} \right| > q_{1-\frac{\alpha}{2}} \right\}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'odre $1-\frac{\alpha}{2}$ de la loi de Student $[\nu]$ degrés de liberté.

Chapitre

12 Tests de comparaison des proportions

12.1 Test sur la valeur d'une proportion

Soient un échantillon $(X_1,\dots,X_n)$ issu d'une loi de Bernouilli $\mathscr{B}(1,p)$ et p_0 une valeur possible de p. Nous savons que $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur efficace de p. De plus, d'après le théorème central-limite, pour n assez grand, nous avons l'approximation en loi suivante

$$\frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{p(1-p)}} \hookrightarrow \mathscr{N}(0,1).$$

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : p \leq p_0$ contre $H_1 : p > p_0$ est :

$$W = \left\{ \overline{X}_n > \sqrt{\frac{p_0(1-p_0)}{n}} q_{1-\alpha} + p_0 \right\}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de loi normale centrée-réduite $\mathscr{N}(0,1)$.

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : p \geq p_0$ contre $H_1 : p < p_0$ est :

$$W = \left\{ \overline{X}_n < \sqrt{\frac{p_0(1-p_0)}{n}} q_{\alpha} + p_0 \right\}$$

où q_{α} est le quantile d'ordre α de loi normale centrée-réduite $\mathscr{N}(0,1)$.

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ est :

$$W = \left\{ \overline{X}_n < p_0 - \sqrt{\frac{p_0(1-p_0)}{n}} q_{1-\frac{\alpha}{2}} \right\} \cup \left\{ \bar{X}_n > p_0 + \frac{\sqrt{p_0(1-p_0)}}{n} q_{1-\frac{\alpha}{2}} \right\}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de loi normale centrée-réduite $\mathscr{N}(0,1)$.

12.2 Test de comparaison de deux proportions

Le problème se pose quand on veut comparer deux populations selon un critère qui est une proportion :

- Comparer les performances deux machines au vu de la proportion de pièces défectueuses qu'elles produisent.
- Comparer les proportions de soulards à Yopougon et Cocody pour vérifier les idées reçues.

Mathematiquement, on a une première population de taille n_1 et une seconde de taille n_2. On veut comparer les deux population selon un critère. On note X_i et Y_i les variables aléatoires définies respectivement par

$$X_i = \begin{cases} 1 & \text{si le } i\text{ème individu de la population 1 présente la caractéristique} \\ 0 & \text{sinon} \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{si le } i\text{ème individu de la population 2 présente la caractéristique} \\ 0 & \text{sinon.} \end{cases}$$

On note p_1 la probabilité qu'un individu de la population 1 possède la caractéristique et p_2 la probabilité qu'un individu de la population 2 possède la caractéristique. On souhaite comparer p_1 et p_2. On suppose que

— $X_1,\dots,X_{n_1}$ sont indépendantes
— $Y_1,\dots,Y_{n_2}$ sont indépendantes
— $(X_1,\dots,X_{n_1})$ et $(Y_1,\dots,Y_{n_2})$ sont indépendants.

Alors $\sum_{i=1}^{n_1} X_i$ suit la loi binomiale $\mathscr{B}(n_1,p_1)$ et $\sum_{i=1}^{n_2} Y_i$ suit la loi binomiale $\mathscr{B}(n_2,p_2)$.
On se contentera ici de supposer que les tailles d'échantillons sont suffisamment grandes pour que l'on puisse faire l'approximation de la loi binomiale par la loi normale :

— $n_1p_1 > 5$, $n_1(1-p_1) > 5$,
— $n_2p_2 > 5$ et $n_2(1-p_2) > 5$.

Alors on peut considérer que $\sum_{i=1}^{n_1} X_i$ et $\sum_{i=1}^{n_2} Y_i$ sont des variables aléatoires indépendantes et approximativement de lois normales, respectivement $\mathscr{N}(n_1p_1, n_1p_1(1-p_1))$ et $\mathscr{N}(n_2p_2, n_2p_2(1-p_2))$.

Comme les estimateurs optimaux de p_1 et p_2 sont respectivement $\overline{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et $\overline{Y}_{n_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$, la région critique du test

$$H_0 : p_1 = p_2 \text{ contre } H_1 : p_1 \neq p_2$$

est donnée par

$$W = \left\{ \left| \overline{X}_{n_1} - \overline{Y}_{n_2} \right| > l_\alpha \right\}$$

où l_α est déterminé par l'équation

$$\mathbb{P}_{H_0}(W) = \alpha.$$

Sous les conditions ci-dessus, nous avons alors

$$\overline{X}_{n_1} \hookrightarrow \mathscr{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$$

$$\overline{Y}_{n_2} \hookrightarrow \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

Comme $\overline{X}_{n_1}$ et $\overline{Y}_{n_2}$ sont indépendantes, nous déduisons que

$$\overline{X}_{n_1} - \overline{Y}_{n_2} \hookrightarrow \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Sous $H_0 : p_1 = p_2 = p$, nous avons

$$\overline{X}_{n_1} - \overline{Y}_{n_2} \hookrightarrow \mathcal{N}\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

et

$$\overline{X}_{n_1} - \overline{Y}_{n_2} \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \hookrightarrow \mathcal{N}(0,1).$$

Comme p est inconnu, en remplaçant p par son estimateur $\hat{p} = \dfrac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}$ le résultat ci-dessus reste approximativement vrai. En posant

$$\hat{\sigma} = \sqrt{\frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\left(1 - \frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

sous l'hypothèse nulle H_0 la statistique

$$U = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\hat{\sigma}} \hookrightarrow \mathcal{N}(0,1).$$

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : p_1 \leq p_2$ contre $H_1 : p_1 > p_2$ est :

$$W = \left\{U > q_{1-\alpha}\right\}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de loi normale centrée-réduite $\mathcal{N}(0,1)$.

Au niveau $\alpha \in]0,1[$, a région critique du test $H_0 : p_1 \geq p_2$ contre $H_1 : p_1 < p_2$ est :

$$W = \left\{U < q_{\alpha}\right\}$$

où q_{α} est le quantile d'ordre α de loi normale centrée-réduite $\mathcal{N}(0,1)$.

Au niveau $\alpha \in]0,1[$, la région critique du test $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$ est :

$$W = \left\{|U| > q_{1-\frac{\alpha}{2}}\right\}.$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de loi normale centrée-réduite $\mathcal{N}(0,1)$.

Exercice 12.2.1. *La machine* 1 *a produit* 96 *pièces dont* 12 *défectueuses. La machine* 2 *a produit* 55 *pièces dont* 10 *défectueuses. Peut-on en conclure que la machine* 1 *est significativement plus performante que la machine* 2 *?*

Exercice 12.2.2. *Dans un sondage réalisé entre le 18 avril 2012 sur 2552 personnes, l'Ifop demande "si dimanche prochain se déroulait le second tour de l'élection présidentielle, pour lequel des candidats suivants y aurait-il le plus de chances que vous votiez". 54% des personnes interrogées ont choisi Hollande contre 46% Sarkozy. Dans un sondage du CSA, le 17 avril 2012, à la question "Si le second tour de l'élection présidentielle de 2012 avait lieu dimanche prochain et que vous aviez le choix entre les deux candidats suivants, pour lequel y aurait-il le plus de chances que vous votiez ?", 58% des 886 personnes interrogées avaient choisi Hollande contre 42% Sarkozy.*

1. *Y a-t-il une différence significative entre ces deux résultats ?*
2. *Même question si l'on considère les sondages de la semaine précédente : le 16 avril 2012, l'Ifop publiait les scores de 55.5%-44.5% sur 1808 interrogés et le CSA trouvait les scores de 57%-43% sur 886 interrogées.*
3. *Donner la p-valeur des tests asymptotiques précédents.*

Exercice 12.2.3. *Une étude des décisions rendues par des jurys dans des cas de vols par effraction où l'accusé était de race noire a révélé les faits suivants : parmi les 28 cas où les victimes étaient de race noire, l'accusé a été trouvé coupable dans 12 cas ; parmi les 36 cas où la victime était de race blanche, l'accusé a été trouvé coupable dans 23 cas. Peut-on conclure que les jurys ont une plus forte tendance à déclarer coupables ceux qui sont accusés d'avoir commis des vols contre des Blancs ?*

Chapitre

13 Tests du χ^2

13.1 Test d'adéquation à une loi donnée

13.1.1 Cas d'une loi discrète

On observe une variable aléatoire discrète X susceptible de prendre k valeurs

$$a_1,\dots,a_k.$$

On note $P=(p_1,\dots,p_k)$ le vecteur des probabilités définies par

$$p_j=\mathbb{P}(X=a_j),\qquad j\in\{1,\dots,k\}.$$

On suppose que le vecteur P est inconnu. Soit $P^*=(p_1^*,\dots,p_k^*)$ un vecteur de probabilités connu ($\sum_{j=1}^{k} p_j^*=1$). On veut resoudre le problème de test suivant :

$$H_0:P=P^*\qquad\text{contre}\qquad H_1:P\neq P^*.$$

Pour $j=1,\dots,k$, on note

$$\widehat{p}_j=\frac{N_j}{n}$$

la fréquence empirique de a_j ; N_j représente le nombre d'obervations de la modalité a_j dans l'échantillon observé de taille n. Le vecteur des fréquences empiriques est

$$\widehat{P}=(\widehat{p}_1,\dots,\widehat{p}_k).$$

Définition 13.1.1. *On appelle distance du χ^2, la quantité*

$$T_n=n\sum_{j=1}^{k}\frac{(\hat{p}_j-p_j^*)^2}{p_j^*}=\sum_{j=1}^{k}\frac{(N_j-np_j^*)^2}{np_j^*}.$$

T_n mesure l'écart entre les effectifs observés et les effectifs "théoriques" sous l'hypothèse H_0

Au niveau $\alpha\in]0,1[$, la région critique du test

$$W=\left\{T_n>\chi^2_{1-\alpha,k-1}\right\}$$

où $\chi^2_{1-\alpha,k-1}$ est le quantile d'ordre $1-\alpha$ de la loi de khi-deux $\chi(k-1)$ à $k-1$ degrés de liberté.

Remarque 13.1.1. *En pratique, ce test marche bien si $n \geq 30$ et $np_j^* \geq 5$ pour tout j. Si cette condition n'est pas satisfaite, on peut regrouper les valeurs de a_j pour lesquelles p_j^* est trop faible.*

Exercice 13.1.1. *Lors de cent lancers d'un dé à six faces, on observe les résultats suivants :*

x	1	2	3	4	5	6
Effectif observé	20	13	17	12	23	15
Effectif théorique	100/6	100/6	100/6	100/6	100/6	100/6

Tester au niveau 5% l'hypothèse H_0 ={*le dé n'est pas pipé*} *contre l'hypothèse H_1=*{*le dé est pipé*}*.*

Solution : *Posons $P^* = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. Il s' agit ici de tester au niveau 5%*

$$H_0 : P = P^* \quad \text{contre} \quad H_1 : P \neq P^*.$$

Pour tout $j = 1, \ldots, k$, nous avons $100 \times p_j^* \geq 5$ et la taille $n = 100 \geq 30$. Les conditions d'utlisation du test sont respectées.
Au niveau 5%, la région critique du test est

$$W = \left\{ T_{100} > \chi^2_{0.95,5} \right\}$$

où

$$T_{100} = \sum_{j=1}^{6} \frac{(N_j - 100 \times p_j^*)^2}{100 \times p_j^*}.$$

Comme $T_{100} =$ et $\chi^2_{0.95,5} = 11.0705$

13.1.2 Cas d'une loi continue

On observe $X_1, \ldots, X_n$ i.i.d. de même loi issue d'une loi P inconnue, continue. Etant donnée P^* une loi continue, on considère le problème de test d'hypothèses suivant

$$H_0 : P = P^* \qquad \text{contre} \qquad H_1 : P \neq P^*.$$

Dans cette situation, on doit partitionner $\mathbb{R}$ en k classes A_j, $j = 1, \ldots, k$. Pour appliquer les mêmes idées que plus haut, d'une part, k doit être assez grand pour que les lois discrètes, c'est-à-dire $\{p_j = P(A_j)\}$ et $\{p_j^* = P^*(A_j)\}$, soient assez proches des lois continues P et P^*. D'autre part, les probabilités $P(A_j)$ doivent être suffisamment grandes, pour que l'approximation asymptotique soit valable.

13.2 Test d'adéquation à une famille de lois

On veut tester si la loi de probabilité inconnue $P = (p_1, \ldots, p_k)$ sur $\{a_1, \ldots, a_k\}$ est égale à une loi $P^*(\theta) = (p_1^*(\theta), \ldots, p_k^*(\theta))$, où $\theta = (\theta_1, \ldots, \theta_s)$ est inconnu. On considère donc le problème de test suivant

$$H_0 : P = P^*(\theta) \qquad \text{contre} \quad H_1 : P \neq P^*(\theta).$$

1. Comme précédemment, nous avons

$$T_n(\theta) = \sum_{j=1}^{k} \frac{(N_j - np_j^*(\theta))^2}{np_j^*(\theta)}$$

mais la quantité $T_n(\theta)$ n'est plus une statistique car θ est inconnu.

2. On estime θ par l'estimateur du maximum de vraisemblance $\hat{\theta}_n$. On obtient

$$T_n(\hat{\theta}_n) = \sum_{j=1}^{k} \frac{(N_j - np_j^*(\hat{\theta}_n))^2}{np_j^*(\hat{\theta}_n)}.$$

Sous H_0,nous avons

$$T_n(\hat{\theta}_n) \xrightarrow{\mathscr{L}} \chi^2(k-s-1).$$

Au niveau $\alpha \in]0,1[$, la région critique du test

$$W = \left\{ T_n(\hat{\theta}_n) > \chi^2_{1-\alpha,k-s-1} \right\}$$

où $\chi^2_{1-\alpha,k-s-1}$ est le quantile d'ordre $1-\alpha$ de la loi de khi-deux $\chi(k-s-1)$ à $k-s-1$ degrés de liberté.

Exercice 13.2.1. *En se référant aux dates de début du pontificat (dates de consécration) et de fin (par décès, démission ou inaptitude), la durée d'exercice de chacun des* 265 *précédents papes (excepté François) a été calculée en nombre d'années. Les résultats groupés en cinq tranches sont présentés dans le tableau suivant :*

Pontificat	*Nombre de papes*
moins d'une année	*46*
1 an - 5 ans	*76*
5 ans - 10 ans	*68*
10 ans -20 ans	*63*
20 ans et plus	*12*

Que penser, au seuil de signification de 5%*, de l'hypothèse selon laquelle la distribution du pontificat des papes serait une distribution exponentielle ?*

13.3 Test d'indépendance

On observe un couple (X,Y) à valeurs dans $\{c_1,\dots,c_r\} \times \{d_1,\dots,d_s\}$ et on veut tester si Y et Z sont indépendantes. On considère un échantillon de taille $((X_1,Y_1),\dots,(X_n,Y_n))$ de même loi que (X,Y).

$$X \text{ et } Y \text{ sont indépendantes} \iff N_{ij} = \frac{N_{i\bullet}N_{\bullet j}}{n}$$

où

$$N_{i\bullet} = \sum_{j=1}^{s} N_{ij} \qquad N_{\bullet j} = \sum_{i=1}^{r} N_{ij}.$$

La statistique de test est définie par

$$T_n = \sum_{j=1}^{r}\sum_{l=1}^{s} \frac{\left(N_{jl} - \frac{N_{j\bullet}N_{\bullet l}}{n}\right)^2}{\frac{N_{j\bullet}N_{\bullet l}}{n}}.$$

Sous l'hypothèse H_0, la statistique T_n converge en loi vers $\chi^2((r-1)(s-1))$.

Au niveau $\alpha \in]0,1[$, la région critique du test

$$W = \left\{ T_n > \chi^2_{1-\alpha,(r-1)(s-1)} \right\}$$

où $\chi^2_{1-\alpha,(r-1)(s-1)}$ est le quantile d'ordre $1-\alpha$ de la loi de khi-deux $\chi(r-1)(s-1))$ à $(r-1)(s-1)$ degrés de liberté.

Exercice 13.3.1. *Une enquête sur l'influence de la ceinture de sécurité a donné les résultats suivants : sur 10.779 conducteurs ayant subit un accident l'enquête rapporte les effectifs dans le tableau qui suit selon la gravité et le port au non de la ceinture de sécurité :*

Nature des blessures	*Port de la ceinture*	*Pas de ceinture*
Graves ou fatales	*5*	*141*
Blessures sérieuses	*25*	*330*
Peu ou pas de blessures	*1229*	*9049*

La ceinture de sécurité a-t'elle une influence sur la gravité des blessures lors d'un accident ?

Chapitre

14 Exercices avec solutions

Voici quelques indications concernant la fiche de TD stat 4. Laissez les étudiants exprimer leurs talents au tableau. Cette fiche ne doit en aucun cas se retrouver dans les mains des étudiants.

Exercice 1. Afin de mieux gérer les demandes de crédits de ses clients, un directeur d'agence bancaire réalise une étude relative à la durée de traitement des dossiers, supposée suivre une distribution normale. Un échantillon de 30 dossiers a donné :

Durée de taitement (en jours)	$[0,10[$	$[10,20[$	$[20,30[$	$[30,40[$	$[40,50[$	$[50,60[$
Effectif	3	6	10	7	3	1

1. Déterminer les estimateurs de la moyenne m et de la variance σ^2 par la méthode du maximum de vraisemblance. Etudier leurs propriétés.

   La vraisemblance de l'échantillon est :

$$\begin{aligned} L(m,\sigma^2,X_1,\dots,X_n) &= \prod_{i=1}^{n} f(m,\sigma^2,X_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2\sigma^2}(X_i-m)^2\Big) \\ &= \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-m)^2\Big) \end{aligned}$$

   La méthode du maximum de vraisemblance consiste à trouver la valeur de (m,σ^2) qui maximise la vraisemblance. Il s'agit ici de maximiser une fonction à deux variables à valeurs réelles. Comme la fonction $x \mapsto \ln(x)$ est croissante, nous avons

$$\begin{aligned} (\widehat{m}_n,\widehat{\sigma_n^2}) &= \arg \max_{(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} \ln\Big(L(m,\sigma^2,X_1,\dots,X_n)\Big) \\ &= \arg \max_{(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} \ln\Big(L(m,\sigma^2,X_1,\dots,X_n)\Big). \end{aligned}$$

   Pour des raisons de simplicité de calcul, on utilise en général la log-vraisemblance. En effet dériver une somme est moins périlleux que dériver un produit. Deux méthodes à expliquer aux étudiants :

— ***Méthode 1 :*** *Maximiser une fonction à deux variables à valeurs réelles, c'est à dire, résoudre le problème de maximisation :*

$$\max_{(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} \ln\Big(L(m,\sigma^2,X_1,\ldots,X_n)\Big).$$

— ***Méthode 2 :*** *Fixer σ^2 et résoudre*

$$\widehat{m}_n = \arg\max_{m\in\mathbb{R}} \ln\Big(L(m,\sigma^2,X_1,\ldots,X_n)\Big).$$

Nous avons alors $\forall \sigma^2 > 0$

$$\ln\Big(L(m,\sigma^2,X_1,\ldots,X_n)\Big) \leq \ln\Big(L(\widehat{m}_n,\sigma^2,X_1,\ldots,X_n)\Big).$$

Puis, résourdre

$$\widehat{\sigma_n^2} = \arg\max_{m\in\mathbb{R}} \ln\Big(L(\widehat{m}_n,\sigma^2,X_1,\ldots,X_n)\Big).$$

Nous obtenons alors pour tout $(m,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*$:

$$\ln\Big(L(m,\sigma^2,X_1,\ldots,X_n)\Big) \leq \ln\Big(L(\widehat{m}_n,\widehat{\sigma_n^2},X_1,\ldots,X_n)\Big).$$

On obtient :

$$\widehat{m}_n = \overline{X}_n \qquad \widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2.$$

Attention : en ce qui concerne la variance, il faut dériver par rapport à σ^2 et non par rapport à σ.

Propriétés des estimateurs : *Il existe deux types de propriétés : non asymptotiques et asymptotiques.*

Intéressons nous à $\widehat{\mathbf{m}}_\mathbf{n} = \dfrac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{X_i}$:

Propriétés non asymptotiques

— $\mathbb{E}(\widehat{m}_n) = m \Rightarrow \widehat{m}_n$ *est un estimateur sans biais de* m.
— *la variance de* $\widehat{m}_n$ *est*

$$\mathbb{V}(\widehat{m}_n) = \frac{\sigma^2}{n}.$$

— *L'information de Fisher apportée par l'échantillon* $(X_1,\ldots,X_n)$ *sur le paramètre* m *est :*

$$I_n(m) = -\mathbb{E}\Big(\frac{\partial^2 \ln\Big(L(m,\sigma^2,X_1,\ldots,X_n)\Big)}{\partial m^2}\Big) = \frac{n}{\sigma^2}.$$

— $\widehat{m}_n$ *est un estimateur sans biais de* m *et* $\mathbb{V}(\widehat{m}_n) = \dfrac{\sigma^2}{n} = \dfrac{1}{I_n(m)} \Rightarrow \widehat{m}_n$ *est un estimateur efficace de* m.

Propriétés asymptotiques

— $\widehat{m}_n$ *est un estimateur convergent de* m. *On peut le montrer de deux manières :*
  — *soit la définition en utilisant* ***l'inégalité de Bienaymé-Tchebithev***
  — *soit par la* ***loi des grands nombres*** *:*
  — *soit en montrant que* $\mathbb{E}(\widehat{m}_n) \longrightarrow m$ *et* $\mathbb{V}(\widehat{m}_n) \longrightarrow 0$.

— $\widehat{m}_n$ *est un estimateur asymptotiquement normal, c'est à dire,*

$$\sqrt{n}(\widehat{m}_n - m) \xrightarrow{loi} \mathcal{N}(0, \sigma^2).$$

On le montre en utilisant ***le Théorème Cenral Limite*** *qui permet d'étudier le comportement asymptotique de la moyenne empirique pour des variables $X_1, \ldots, X_n$ i.i.d. de moyenne m et de variance $\sigma^2 > 0$.*

Intéressons nous $\widehat{\sigma^2}_{\mathbf{n}} = \dfrac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{X_i} - \overline{\mathbf{X}}_{\mathbf{n}})^{\mathbf{2}}$.

Propriétés non asymptotiques

— $\mathbb{E}(\widehat{\sigma_n^2}) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2 \Rightarrow \widehat{\sigma_n^2}$ *est un estimateur biaisé de* σ^2.

— $\widehat{\sigma_n^2}$ *est un estimateur biaisé de* $\sigma^2 \Rightarrow \widehat{\sigma_n^2}$ *n'est pas un estimateur efficace de* σ^2. *(****Pas la peine de calculer l'information de Fisher et la borne de Cramer-Rao, la condition sans biais n'étant pas vérifiée.****)*

Propriétés asymptotiques

— $\mathbb{E}(\widehat{\sigma_n^2}) = \dfrac{n-1}{n}\sigma^2 \longrightarrow \sigma^2 \Rightarrow \widehat{\sigma_n^2}$ *est un estimateur asymptotiquement sans biais de* σ^2.

— *Vérifier que la variance* $\mathbb{V}(\widehat{\sigma_n^2}) \longrightarrow 0$ *pour assurer la convergence de* $\widehat{\sigma_n^2}$ *vers* σ^2.

— ***Pas la peine d'établir la normalité asymptotique ; c'est un peu compliqué pour eux je crois ! Si vous trouvez simple, faites moi signe !***

2. Donner les estimations ponctuelles de la moyenne m et de la variance σ^2.
Utiliser les centres des intervalles pour faire les estimations :

$$\overline{X}_{30} = \frac{1}{n}\sum_{i=1}^{30} c_i = \frac{1}{30}\sum_{j=1}^{6} n_j c_j.$$

$$\widehat{\sigma_{30}^2} = \frac{1}{30}\sum_{i=1}^{30}(c_i - \overline{X}_{30})^2 = \frac{1}{30}\sum_{j=1}^{6} n_j (c_j - \overline{X}_{30})^2.$$

3. Donner une estimation de m par intervalle de confiance au seuil de risque 5%.
D'après le cours, l'intervalle de confiance pour m de niveau 0.95 est

$$\left[\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{0.975}^{(n-1)}, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{0.975}^{(n-1)}\right]$$

où $t_{0.975}^{(n-1)}$ est le quantile d'ordre 0.975 de la loi de Student à $n-1$ degrés de liberté et

$$S_n = \sqrt{\frac{n}{n-1}\widehat{\sigma_n^2}}.$$

4. Au seuil de 5%, tester l'hypothèse $H_0 : m = 30$ contre $H_1 : m < 30$. Que pouvez-vous conclure ?
La région critique du test au seuil $\alpha = 0.05$ est :

$$W = \left\{(X_1, \ldots, X_{30}) : \frac{\sqrt{30}(\overline{X}_{30} - 30)}{S_{30}} < t_{0.05}^{(29)}\right\}.$$

Rappel du cours : Considérons un échantillon $(X_1, \ldots, X_n)$ issu de la loi normale $\mathcal{N}(m, \sigma^2)$. Si σ^2 est connue :

Hypothèses	Région Critique
$H_0 : m \leq m_0$ vs $H_1 : m > m_0$	$W = \left\{(X_1,\ldots,X_n) : \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{\sigma} > q_{1-\alpha}\right\}$
$H_0 : m \geq m_0$ vs $H_1 : m < m_0$	$W = \left\{(X_1,\ldots,X_n) : \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{\sigma} < q_{\alpha}\right\}$
$H_0 : m = m_0$ vs $H_1 : m \neq m_0$	$W = \left\{(X_1,\ldots,X_n) : \left\lvert\dfrac{\sqrt{n}(\overline{X}_n - m_0)}{\sigma}\right\rvert > q_{1-\frac{\alpha}{2}}\right\}$

Si σ^2 est inconnue :

Hypothèses	Région Critique
$H_0 : m \leq m_0$ vs $H_1 : m > m_0$	$W = \left\{(X_1,\ldots,X_n) : \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S_n} > t_{1-\alpha}^{(n-1)}\right\}$
$H_0 : m \geq m_0$ vs $H_1 : m < m_0$	$W = \left\{(X_1,\ldots,X_n) : \dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S_n} < t_{\alpha}^{(n-1)}\right\}$
$H_0 : m = m_0$ vs $H_1 : m \neq m_0$	$W = \left\{(X_1,\ldots,X_n) : \left\lvert\dfrac{\sqrt{n}(\overline{X}_n - m_0)}{S_n}\right\rvert > t_{1-\frac{\alpha}{2}}^{(n-1)}\right\}$

Exercice 2. La société "Votre santé" est une entreprise de vente par correspondance de produits de beauté dits "naturels". Elle gère un fichier de **350000** clients et propose chaque mois une offre promotionnelle accompagnée d'un cadeau. Le taux de réponse à cette offre est généralement de **15%**, la marge moyenne par réponse de **340** fcfa. Mlle Claire, nouvellement en charge de ce fichier, a retenu comme cadeau un abonnement gratuit de six mois, au mensuel "Votre beauté Madame". Elle pense que cela pourrait augmenter le taux de réponse à la prochaine offre ; toutefois cette proposition ne serait rentable que si le taux de réponse dépassait les **17.5%** (avec la même marge moyenne évidemment). Elle envisage de tester la réalité de ces hypothèses sur un échantillon de clientes. La précision voulue pour son estimation est de l'ordre de **2%**.

1. Quelle taille d'échantillon doit-elle choisir afin d'atteindre la précision voulue (avec un niveau de confiance de **0.95**) ?

— ***Modélisation (à ne jamais oublier !)***
  — *Population : les 350 000 clients*
  — *Echantillon : Soit X_i la variable aléatoire définie par :*

$$X_i = \begin{cases} 1 & \text{si le ième client achète} \\ 0 & \text{sinon} \end{cases}$$

Nous $(X_1,\ldots,X_n)$ est un échantillon issu de la loi de Bernouilli $\mathcal{B}(1,p)$ où p s'interprète comme la proportion des clients qui acheteraient si l'offre se généralisait à l'ensemble des clients.

— ***Détermination de n. Exposer les deux methodes et privilégier ici la deuxième car on a une idée de $\overline{X}_n$.***
  — ***Méthode pessimiste (majoration de l'écart-type) :*** *L'intervalle de confiance de niveau $1-\alpha$ est donné par*

$$\left[\overline{X}_n - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right] \subset \left[\overline{X}_n - q_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}\right]$$

puisque $\sqrt{\overline{X}_n(1-\overline{X}_n)} \leq \frac{1}{2}$. La marge d'erreur est donc :

$$ME = q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq q_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}.$$

Nous déterminons n *tel que*

$$q_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}} \leq 0.02 \Rightarrow n \geq \left(\frac{q_{1-\frac{\alpha}{2}}}{0.04}\right)^2 = 2401.$$

— ***Méthode plus optimiste (on pense que le taux de réponse sera proche du taux habituel qui est*** 15%***) :*** *L'intervalle de confiance de niveau* $1-\alpha$ *est donné par*

$$\left[\overline{X}_n - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]$$

avec sans doute $\overline{X}_n(1-\overline{X}_n)$ *sans doute proche de son ancienne* $0.15(1-0.15)$*. Nous déterminons alors* n *tel que*

$$ME = q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} = q_{1-\frac{\alpha}{2}}\sqrt{\frac{0.15(1-0.15)}{n}} \leq 0.02$$

$$\Rightarrow n \geq 1224.51 \Rightarrow n = 1225.$$

2. Les résultats d'un sondage sur un échantillon de 1225 clientes vous sont donnés en annexe.

   Donner une estimation par intervalle au niveau 0.95 du pourcentage p de réponses positives attendues à l'offre.

3. Mlle Claire se propose de procéder au test d'hypothèses suivant $H_0 : p = 17.5\%$ contre $H_0 : p > 17.5\%$. Expliquer pourquoi elle envisage ce test. Calculer la p-value. Qu'en concluez-vous ?

   Si elle rejette son H_0 *pour* H_1 *alors, elle saura que* $p_0 > 17.5\%$ *et que la nouvelle promotion est à étudier. En effet, on apprend réellement d'un test lorsque* H_0 *est rejetée !*
   La région critique du test est :

$$W = \left\{(X_1,\ldots,X_{1225}) : \frac{\sqrt{1225}(\overline{X}_{1225} - 0.175)}{\sqrt{0.175(1-0.175)}} > q_{1-\alpha}\right\}$$

Hypothèses	Région Critique
$H_0 : p \leq p_0$ vs $H_1 : p > p_0$	$W = \left\{(X_1,\ldots,X_n) : \frac{\sqrt{n}(\overline{X}_n - p_0)}{\sqrt{p_0(1-p_0)}} > q_{1-\alpha}\right\}$
$H_0 : p \geq p_0$ vs $H_1 : p < p_0$	$W = \left\{(X_1,\ldots,X_n) : \frac{\sqrt{n}(\overline{X}_n - p_0)}{\sqrt{p_0(1-p_0)}} < q_{\alpha}\right\}$
$H_0 : p = p_0$ vs $H_1 : p \neq p_0$	$W = \left\{(X_1,\ldots,X_n) : \left\lvert\frac{\sqrt{n}(\overline{X}_n - p_0)}{\sqrt{p_0(1-p_0)}}\right\rvert > q_{1-\frac{\alpha}{2}}\right\}$

4. Mlle Claire pense que les nouveaux clients (inscrits depuis moins de 6 mois) ont un taux de réponse inférieur aux anciens. Confirmer ou infirmer cette hypothèse.

   Nous allons faire un test de comparaison des proportions pour répondre à la question
   — ***Modélisation (à ne jamais oublier !)*** *Soient les variables aléatoires définies par :*

$$Y_i = \begin{cases} 1 & \text{si le ième ancien client a répondu} \\ 0 & \text{sinon} \end{cases}$$

$$Z_i = \begin{cases} 1 & \text{si le ième ancien client a répondu} \\ 0 & \text{sinon} \end{cases}$$

On dispose ainsi de deux échantillons : $(Y_1,\ldots,Y_{850})$ *issu d'une loi de Bernouilli* $\mathscr{B}(1,p_{anciens})$ *et* $(Z_1,\ldots,Z_{375})$ *issu d'une loi de Bernouilli* $\mathscr{B}(1,p_{nou})$ *où* p_{anc} *et* p_{nou} *représentent respectivement les taux de réponses dans les deux populations.*

— *Nous considérons le problème de*

$$H_0 : p_{an} = p_{nou} \text{ contre } H_1 : p_{anc} > p_{nou}$$

La variable de décision est

$$T \frac{\overline{Y}_{850} - \overline{Z}_{375}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{850}+\frac{1}{375})}}$$

où

$$\widehat{p} = \frac{850 \times \overline{Y}_{850} + 375 \times \overline{Z}_{375}}{850+375}.$$

La région critique du test est :

$$W = \left\{ T > q_{1-\alpha} \right\}.$$

— *Pour* $\alpha = 0.05$ $q_{0.95} = 1.64$ *et* $t = 2.13$. *On voit que* $2.13 > 1.64$. *Ainsi, au niveau* $\alpha = 0.05$, *nous acceptons* H_1, *c'est à dire que les anciens sont plus recptifs que les nouveaux.*

Théorème 14.0.1. *Posons*

$$\widehat{p} = \frac{n_1 \overline{X}_{n_1} + n_2 \overline{X}_{n_2}}{n_1+n_2}.$$

- *La région critique du test* $H_0 : p_1 \leq p_2$ *contre* $H_1 : p_1 > p_2$ *est :*

$$W = \left\{ \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} > q_{1-\alpha} \right\}.$$

- *La région critique du test* $H_0 : p_1 \geq p_2$ *contre* $H_1 : p_1 < p_2$ *est :*

$$W = \left\{ \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} < q_{\alpha} \right\}.$$

- *La région critique du test* $H_0 : p_1 = p_2$ *contre* $H_1 : p_1 \neq p_2$ *est :*

$$W = \left\{ \left| \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \right| > q_{1-\frac{\alpha}{2}} \right\}.$$

5. Il s'agit dans cette question de déterminer un intervalle de confiance au niveau 0.95 de la marge de la campagne promotionnelle. Peut-on considérer que la marge moyenne attendue de cette campagne sera la même que pour les campagnes précédentes. On posera cette alternative sous forme de test.

— ***Modélisation :*** *Pour chacune des 258 commandes, soit M_i la variable aléatoire qui donne la marge réalisée pour la commande i. Pour faire simple, nous allons supposer que $(M_1,\dots,M_{258})$ est un échantillon issu d'une loi normale $\mathcal{N}(m,\sigma^2)$. Ici, m et σ^2 sont inconnues.*

— *Au niveau $\alpha = 0.05$, nous considérons le problème de test de*

$$H_0 : m = 340 \text{ contre } H_1 : m < 340$$

La région critique du test est

$$W = \left\{ \frac{\sqrt{258}(\overline{M}_{258} - 340)}{S_{258}} < t_\alpha^{(257)} \right\}$$

où $t_\alpha^{(257)}$ est le quantile d'ordre 0.05 de la loi de Student à 257 degrés de liberté. On peut utiliser la table de la loi normale centrée réduite car la loi de Student converge vers la loi normale $\mathcal{N}(0,1)$ lorsque le nombre de degrés de liberté $n \to +\infty$ ($n > 30$ en pratique.)

— *On a $t = -0.97$ et $t_\alpha^{(257)} = -1.65$. Nous avons donc $-0.97 > -1.65$. Nous en déduisons qu'au niveau 5%, on conerve H_0, c'est à dire en moyenne, la marge ne diffère pas significativement de 340.*

Annexe : résultats du sondage

	Nouveaux clients	Anciens clients
Nombre d'individus	1225	850
Nombre de réponses	258	193

Marge totale	Marge moyenne	Ecart-type de la marge
8 514 000	33 000	16 500

Exercice 3. On considère un échantillon $(X_1,\dots,X_n)$ issu de la loi exponentielle $\mathcal{E}(\theta)$ avec $\theta > 0$ inconnu.

1. Déterminer l'estimateur $\widehat{\theta}_n$ par la methode du maximum de vraisemblance.

   La vraisemblance de $(X_1,\dots,X_n)$ est

$$\begin{aligned} L(X_1,\dots,X_n,\theta) &= \prod_{i=1}^{n} \theta \exp(-\theta X_i) 1_{\mathbb{R}_+^*}(X_i) \\ &= \theta^n \exp\Big(-\theta \sum_{i=1}^{n} X_i\Big) 1_{(\mathbb{R}_+^*)^n}(X_1,\dots,X_n). \end{aligned}$$

   Pour tout $(X_1,\dots,X_n) \in (\mathbb{R}_+^*)^n$, on a

$$\ln(L(X_1,\dots,X_n,\theta)) = n\ln(\theta) - \theta \sum_{i=1}^{n} X_i$$

$$\frac{\partial \ln L(X_1,\ldots,X_n,\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} X_i = 0 \Longleftrightarrow \theta = \frac{1}{\overline{X}_n}$$

$$\frac{\partial^2 \ln L(X_1,\ldots,X_n,\theta)}{\partial \theta^2}\Big(\frac{1}{\overline{X}_n}\Big) = -n\overline{X}_n^2 < 0.$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\hat{\theta}_n = \frac{1}{\overline{X}_n}.$$

2. Montrer que $\widehat{\theta}_n$ peut être obtenu par la methode des moments.
   Nous avons
   $$\mathbb{E}(X_1) = \frac{1}{\theta} \Rightarrow \overline{X}_n = \frac{1}{\theta} \Rightarrow \theta = \frac{1}{\overline{X}_n}$$

3. Déterminer les propriétés asymptotiques de $\widehat{\theta}_n$.

   (a) D'après la loi des grands nombres, on a :

   $$\overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{\theta}.$$

   Comme, l'application $x \mapsto \dfrac{1}{x}$ est continue sur $\mathbb{R}_+^*$, alors

   $$\frac{1}{\overline{X}_n} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta.$$

   (b) D'après le Théorème Central limite $\overline{X}_n$ est asymptotiquement normal :

   $$\sqrt{n}\Big(\overline{X}_n - \frac{1}{\theta}\Big) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\Big(0, \frac{1}{\theta^2}\Big).$$

   Comme, l'application $g : x \mapsto \frac{1}{x}$ est dérivable sur $\mathbb{R}_+^*$ et $g'(x) = -\frac{1}{x^2}$, on obtient par la delta-méthode :

   $$\sqrt{n}(g(\overline{X}_n) - g(1/\theta)) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\Big(0, \frac{1}{\theta^2}(g'(1/\theta))^2\Big).$$

   c'est à dire

   $$\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \theta\Big) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,\theta^2).$$

4. Montrer que $\widehat{\theta}_n$ est un estimateur biaisé de θ. En déduire un estimateur $\widetilde{\theta}_n$ sans biais de θ.
   Montrer que
   $$\mathbb{E}(\widehat{\theta}_n) \neq \theta.$$
   Utiliser la linéarité de l'espérance pour tirer $\widetilde{\theta}_n$.

5. L'estimateur $\widetilde{\theta}_n$ est-il efficace ?
   Je crois que $\widetilde{\theta}_n$ n'est pas efficace malgré qu'il soit sans biais. Mais il faut vérifier que la variance :
   $$\mathbb{V}(\widetilde{\theta}_n) > BCR(\theta),$$
   où $BCR(\theta)$ est la borne de Cramer-Rao.

Exercice 4. Pour 30 femmes et 20 hommes, on a observé le salaire mensuel. Les résultats mesurés en euros sont ci-dessous :

Salaire des femmes

1955	1764	1668	1441	1970	1795	1716	1911	1660	2001
1744	1676	1695	1652	1626	1698	1656	1739	1789	1716
1684	1445	1646	1617	1630	1440	1850	1252	1493	1537

Salaire des hommes

2283	2010	1970	2019	1941	2024	2046	1962	1948	2071
2108	1880	2008	2119	2030	2014	1919	1837	2094	2169

Au seuil de 5%, le salaire moyen des hommes est-il significativement supérieur à celui des femmes ?

Il s'agit ici de faire un test de comparaison des moyennes dans un échantillon gaussien.

— $(X_1,\dots,X_{n_1})$ *est issu de* $\mathcal{N}(m_1,\sigma_1^2)$
— $(Y_1,\dots,Y_{n_2})$ *est issu de* $\mathcal{N}(m_2,\sigma_2^2)$.
— $(X_1,\dots,X_{n_1})$ *et* $(Y_1,\dots,Y_{n_2})$ *sont indépendants.*

Problème : *tester* $H_0 : m_1 = m_2$ *contre* $H_1 : m_1 \neq m_2$ *au niveau* α.
La variable de décision dépend du fait que les variances σ_1^2 *et* σ_2^2 *soient égales ou non. Il faut donc commencer par comparer les variances :*

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ contre } H_1 : \sigma_1^2 \neq \sigma_2^2$$

La région critique au niveau α *est donnée*

$$W = \left\{ \frac{S_1^2}{S_2^2} > f_{1-\frac{\alpha}{2}} \right\} \cup \left\{ \frac{S_1^2}{S_2^2} < f_{\frac{\alpha}{2}} \right\}$$

où f_β *est le quantile d'odre* β *de la loi de Fisher avec* $n_1 - 1$ *et* $n_2 - 1$ *degrés de liberté. La région critique du test au niveau* α *:*

$$W = \left\{ |T| > t_{1-\frac{\alpha}{2}}(m) \right\}$$

où

$$m = n_1 + n_2 - 2 \text{ si } \sigma_1 = \sigma_2$$

et

$$m = \frac{\left(\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2} \right)^2}{\frac{S_{n_1}^4}{n_1^2(n_1-1)} + \frac{S_{n_2}^4}{n_2^2(n_2-1)}} \text{ si } \sigma_1 \neq \sigma_2.$$

Année Universitaire 2018-2019
Examen (2 heures)
Enseignant : Prof. YODE Armel

Exercice 1. Une enquête concernant l'utilisation des cartes bancaires (CB) a été effectuée en septembre 2005 auprès des personnes agées de 18 ans. Les résultats (partiels) de cette enquête sont présentés dans le tableau ci-dessous :

Description	Effectif
Personnes interrogées	501
Porteurs de CB	433
ayant effectué au moins un achat par CB	400
ayant effectué au moins un achat par CB sur Internet	144

Dans la suite, on s'intéresse à la proportion p **de personnes ayant effectué un achat par CB sur Internet parmi celles qui ont effectué au moins un achat par CB.**

1. Donner le modèle théorique permettant l'étude de p : population, échantillon, variable aléatoire, loi.

   - *La population étudiée est l'ensemble des clients ayant effectué au moins un achat par CB.*
   - *On dispose d'un échantillon de taille* 400 *issu de cette population.*
   - *Soit* X_i *la variable aléatoire définie par :*

   $$X_i = \begin{cases} 1 & \text{si le client } i \text{ a effectué au moins un achat par CB sur intenet} \\ 0 & \text{sinon} \end{cases}$$

   X_i *suit une loi de Bernouilli* $\mathscr{B}(1,p)$*. De plus les variables aléatoires* $X_1,\ldots,X_n$ *sont indépendantes.*

2. Donner un estimateur $\widehat{p}$ de p par la méthode du maximum de vraisemblance. Etudier les propriétés de l'estimateur $\widehat{p}$.

   La vraisemblance de l'échantillon $(X_1,\ldots,X_n)$ *est :*

$$\begin{aligned} L(p,X_1,\ldots,X_n) &= \prod_{i=1}^{n} f(X_i,p) \\ &= \prod_{i=1}^{n} p^{X_i}(1-p)1_{\{0,1\}} \\ &= (1-p)^n\Big(\frac{p}{1-p}\Big)^{\sum_{i=1}^n X_i} 1_{\{0,1\}^n} \end{aligned}$$

   Pour tout $p \in]0,1[$, $(X_1,\ldots,X_n) \in \{0,1\}^n$, $L(p,X_1,\ldots,X_n) > 0$ *et*

$$\ln\Big(L(p,X_1,\ldots,X_n)\Big) = n\ln(1-p) - \sum_{i=1}^{n} X_i \ln\Big(\frac{p}{1-p}\Big)$$

La log-vraisembleance est

$$\ln L(X_1,\ldots,X_n,p) = \sum_{i=1}^{n} X_i \ln(p) + (n - \sum_{i=1}^{n} X_i)\ln(1-p)$$

Condition du premier ordre

$$\frac{\partial \ln L(X_1,\ldots,X_n,p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{n} X_i}{(1-p)} = 0 \Longleftrightarrow p = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

Condition du deuxième ordre

$$\frac{\partial^2 \ln L(X_1,\ldots,X_n,p)}{\partial p^2}(\overline{X}_n) = \frac{-n\overline{X}_n}{\overline{X}_n^2} - \frac{n - n\overline{X}_n}{(1-\overline{X}_n)^2} < 0.$$

L'estimateur du maximum de vraisemblance de p est donné par

$$\widehat{p}_n = \overline{X}_n.$$

Étude des propriétés asymptotiques de $\widehat{p}_n$.

(a) D'après la loi des grands nombres, $\overline{X}_n$ est un estimateur convergent de p.

(b) D'après le Théorème Central limite $\overline{X}_n$ est asymptotiquement normal :

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

Étude des propriétés non asymptotiques de $\widehat{p}_n$.

(a) $E(\widehat{p}) = p$

(b) L'information de Fisher est :

$$I_n = -\mathbb{E}\Big(\frac{\partial^2 \ln L(X_1,\ldots,X_n,p)}{\partial p^2}\Big) = \frac{n}{p(1-p)}.$$

La borne de Cramer-Rao est donc :

$$BCR(p) = \frac{p(1-p)}{n}.$$

$\overline{X}_n$ est un estimateur efficace de p car $\widehat{p}_n$ est sans biais et

$$var(\widehat{p}_n) = \frac{p(1-p)}{n} = BCR(p).$$

3. Donner une estimation de p.

   Une estimation de p est $\dfrac{144}{400} = 0.36$

4. Calculer un intervalle de confiance de niveau de confiance 95% pour p.

   L'intervalle de confiance pour p de niveau $1-\alpha$ est :

$$\Big[\overline{X}_n - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\Big] =$$

$$\Big[0.36 - 1.96\sqrt{\frac{0.36(1-0.36)}{400}}, 0.36 + 1.96\sqrt{\frac{0.36(1-0.36)}{400}}\Big] = [0.313, 0.407]$$

5. Si on suppose constant le pourcentage de personnes interrogées ayant effectué au moins un achat par CB sur Internet, quelle devrait être la taille de l'échantillon pour connaitre p à 3% près (avec un niveau de confiance de 95%) ?

   Nous avons

$$\left|p-\overline{X}_n\right| \leq q_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq 0.03 \Rightarrow n \geq \frac{q^2_{1-\frac{\alpha}{2}}\overline{X}_n(1-\overline{X}_n)}{(0.03)^2}$$

$$\Rightarrow n \geq \frac{(1.96)^2 * 0.36(1-0.36)}{(0.03)^2} = 983.44 \Rightarrow n = 984.$$

6. En janvier 2005, une enquête similaire évaluait à 32% la part de personnes ayant effectué au moins un achat par CB sur Internet parmi celles ayant effectué au moins un achat par CB.

   (a) Les données de l'enquête de septembre 2005 permettent-elles de conclure à une augmentation significative de la part de personnes utilisant leur CB sur Internet, en prenant un risque de première espèce de 1% ?

   Il s'agit ici de tester $H_0 : p \leq 0.32$ *contre* $H_1 : p > 0.32$ *au seuil* $\alpha = 0.01$. *La région critique est donc*

$$W = \left\{\frac{\sqrt{400}(\widehat{p}_n - 0.32)}{\sqrt{0.32 * 0.68}} > q_{0.99}\right\}$$

   où $q_{0.99} = 2.33$ *est le quantile d'ordre* 0.99 *de la loi normale centrée réduite. Comme*
   $\frac{\sqrt{400}(\widehat{p}_n - 0.32)}{\sqrt{0.32 * 0.68}} = 1.714 < 2.33$, *alors au seuil de* 1%, *les données de septembre 2005 ne permettent pas de conclure àune augmentation significative de la part des personnes utilisant leur CB sur internet.*

   (b) Quelle est la puissance du test lorsque $p = 34\%$?

   La puissance du test au point $p = 0.34$ *est donée par :*

$$\gamma(3) = \mathbb{P}_{34}\left(\frac{\sqrt{400}(\widehat{p}_n - 0.32)}{\sqrt{0.32 * 0.68}} > 2.33\right)$$

$$= \mathbb{P}\left(\widehat{p}_{400} > 2.33\sqrt{\frac{0.32 * 0.68}{400}} + 0.32\right)$$

   Sous l'hypothèse H_1, $\widehat{p}_n \sim \mathcal{N}\left(0.34, \frac{0.34 * 0.66}{400}\right) \Leftrightarrow \frac{\sqrt{400}(\widehat{p}_n - 0.34)}{\sqrt{0.34 * 0.66}} \sim \mathcal{N}(0,1)$. *Ainsi, nous obtenons :*

$$\gamma(3) = \mathbb{P}_{0.34}\left(\frac{\sqrt{400}(\widehat{p}_n - 0.34)}{\sqrt{0.34 * 0.66}} > \sqrt{\frac{400}{0.34 * 0.66}}\left[2.33\sqrt{\frac{0.32 * 0.68}{400}} + 0.32 - 0.34\right]\right)$$

Exercice 2. On considère un échantillon $(X_1, \ldots, X_n)$ issu de la loi exponentielle $\mathcal{E}\left(\frac{1}{\theta}\right)$ avec $\theta > 0$ inconnu.

1. Déterminer l'estimateur $\widehat{\theta}_n$ par la methode du maximum de vraisemblance.

La vraisemblance est :

$$\begin{aligned} L(X_1,\ldots,X_n,\theta) &= \prod_{i=1}^{n} f(X_i,\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta}\exp\Big(-\frac{1}{\theta}X_i\Big)1_{\mathbb{R}_+^*} \\ &= \frac{1}{\theta^n}\exp\Big(-\frac{1}{\theta}\sum_{i=1}^{n}X_i\Big)1_{\mathbb{R}_+^{*n}} \end{aligned}$$

Pour tout $(X_1,\ldots,X_n)\in\mathbb{R}_+^{*n}$, $\theta>0$

$$L(X_1,\ldots,X_n,\theta) = \frac{1}{\theta^n}\exp\Big(-\frac{1}{\theta}\sum_{i=1}^{n}X_i\Big) > 0.$$

Alors, nous avons :

$$\ln\Big(L(X_1,\ldots,X_n,\theta)\Big) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n}X_i$$

Condition du premier ordre :

$$\frac{\partial \ln\Big(L(X_1,\ldots,X_n,\theta)\Big)}{\partial\theta} = 0 \Longrightarrow \theta = \overline{X}_n.$$

Condition du second ordre :

$$\frac{\partial^2 \ln\Big(L(X_1,\ldots,X_n,\theta)\Big)}{\partial\theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n}X_i$$

Comme $\dfrac{n}{\overline{X}_n^2} - \dfrac{2n}{\overline{X}_n^2} < 0$ *alors l'EMV est* $\widehat{\theta}_n = \dfrac{1}{n}\sum_{i=1}^{n}X_i$

2. Vérifier que $\widehat{\theta}_n$ peut être obtenu par la méthode des moments.
3. L'estimateur $\widehat{\theta}_n$ est-il efficace ?
4. Déterminer les propriétés asymptotiques de $\widehat{\theta}_n$.
5. Déterminer les propriétés asymptotiques de $\widehat{\theta}_n^2$.

Exercice 3. Une étude a été réalisée sur le cancer de la gorge. Pour cela, une population de **1000** personnes a été interrogée. les résultats obtenus sont donnés dans le tableau de contingences suivant :

	Atteint du cancer de la gorge	Non atteint du cancer de la gorge
Fumeur	344	258
Non fumeur	160	238

Doit-on rejeter au niveau **5%** l'hypothèse d'indépendance des deux caract'eres : $\boldsymbol{X}$=(être fumeur) et $\boldsymbol{Y}$=(être atteint du cancer de la gorge).

Exercice 4. Sur deux groupes de même taille **9** malades, on expérimente les effets d'un nouveau médicament. On observe les résultats suivants :

Groupe 1	15	18	17	20	21	18	17	15	19
Groupe 2	12	16	17	18	17	15	18	14	16

1. Comparer au niveau **5%** les variances des deux populations
2. Comparer au niveau **5%** les moyennes des deux populations