Magistère de Pharmacologie, UNSA.

Une brève introduction aux Statistiques

Julien Arino
INRIA Sophia Antipolis
Julien.Arino@sophia.inria.fr

Février – Mars 2000

Table des matières

1. Rappels de probabilités

1.1 Probabilité

Ensemble des possibles – On appelle ensemble des évènements possibles (ou ensemble des possibles, ou univers des possibles), et l'on note Ω, l'espace de toutes les réalisations possibles d'une expérience. Un élément de Ω est un *évènement*.

Exemple – Jet d'un dé à six faces: $\Omega = \{1,2,3,4,5,6\}$. Taille d'un être humain: $\Omega = [0.3; 3]$ mètres (à la louche).

Probabilité – Une probabilité est une fonction P de l'ensemble $\mathcal{P}(\Omega)$ des parties de Ω, à valeurs dans [0,1], vérifiant les propriétés suivantes:

1. $P(\emptyset) = 0$.
2. $P(\Omega) = 1$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Remarque– $A \cap B$ correspond à la réalisation simultanée de A et B, tandis que $A \cup B$ correspond à la réalisation de l'un au moins des deux évènements A ou B.

Espace probabilisé – On appelle *espace probabilisé*, et l'on note (Ω,P), la donnée d'un ensemble Ω, ensemble des évènements possibles, et d'une loi de probabilité P.

Propriété 1.1 *Soit P une probabilité sur $\mathcal{P}(\Omega)$. Alors si l'on note $\bar{A}$ le complémentaire de A dans Ω (*i.e. *$A \cap \bar{A} = \emptyset$ et $\Omega = A \cup \bar{A}$), on a*

$$P(\bar{A}) = 1 - P(A)$$

Indépendance – Deux évènements A et B sont dits *indépendants* ssi

$$P(A \cap B) = P(A)P(B)$$

1.2 Variables aléatoires

Variable aléatoire – Une variable aléatoire est une fonction X définie sur Ω. Dans la suite, on notera v.a. les variables aléatoires. On dit qu'une v.a. est *discrète* si l'ensemble de ses valeurs est dénombrable, sinon on parle de v.a. *continue*.

> Dénombrable ne signifie pas nécessairement fini, mais "énumérable". Ainsi l'ensemble $\mathbb{N}$ des entiers naturels est infini dénombrable. Par contre, tout intervalle (non réduit à un point) de $\mathbb{R}$ n'est pas dénombrable.

Exemple – Jet d'un dé à six faces: la v.a. X "résultat du jet" est discrète. Taille d'un être humain: la v.a. X "taille d'un tre humain" est continue.

Fonction de répartition – On appelle *fonction de répartition* de la v.a. X, la fonction $F : \mathbb{R} \to [0,1]$, qui à tout $x \in \mathbb{R}$ associe la probabilité que $X \leq x$.

Une fonction de répartition est une fonction monotone croissante, telle que

$$\lim_{x \to -\infty} F(x) = 0 \text{ et } \lim_{x \to +\infty} F(x) = 1 \tag{1.1}$$

Densité – On appelle *densité de probabilité* de la v.a. continue X, la fonction $f : \mathbb{R} \to [0,1]$, qui à tout $x \in \mathbb{R}$ associe sa probabilité. Une densité est telle que

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{1.2}$$

> A la différence de la fonction de répartition, qui est définie pour tout type de v.a., la densité de probabilité n'est définie que pour les v.a. continues. On parle d'ailleurs parfois de v.a. à densité plutôt que de v.a. continues.

Dans le cas de lois continues, la fonction de répartition et la densité de probabilité sont liées (F est une primitive de f):

$$\forall x \in \mathbb{R},\ F(x) = \int_{-\infty}^{x} f(t)dt \tag{1.3}$$

1.3 Espérance mathématique, variance

Espérance mathématique – L'*espérance mathématique* est la "somme" des valeurs possibles de la loi de probabilité, pondérée par la probabilité de chaque

point. Elle est définie dans chacun des cas possibles, v.a. discrètes et v.a. continues.

Variance – La *variance* $Var(X)$ d'une variable aléatoire X est l'espérance de la variable aléatoire Y définie par $Y = X - E(X)$.

$$Var(X) = \frac{1}{n}\sum (X - \mu)^2$$

L'ecart-type de la v.a. X est la racine carrée de la variance de X.

La variance et l'écart type mesurent la dispersion autour de la moyenne. On note souvent σ^2 la variance et σ l'écart type.

L'espérance est à valeurs dans $\mathbb{R}$. La variance et l'ecart type, par contre, sont à valeurs dans $\mathbb{R}^+$.

1.3.1 Cas des v.a. discrètes

Dans le cas d'une v.a. discrète X, l'espérance est donnée par

$$E(X) = \sum_{x \in \Omega} xP(X = x)$$

Exemple – Lancer d'un dé à six faces, non pipé. Chacune des faces est équiprobable, donc $P(i) = P(X = i) = 1/6$ pour $i = 1, \ldots, 6$. Soit X la v.a. "résultat du lancer". L'espérance mathématique de X est donc

$$\begin{aligned} E(X) &= \sum_{x \in \Omega} xP(X = x) \\ &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + \cdots + 6 \cdot P(X = 6) \\ &= (1 + 2 + \cdots + 6)1/6 \\ &= 21/6 = 3.5 \end{aligned}$$

Pour calculer la variance, on va utiliser la formule (1.4): $Var(X) = E(X^2) - E(X)^2$. Il nous faut donc calculer $E(X^2)$.

$$\begin{aligned} E(X^2) &= \sum_{x \in \Omega} x^2 P(X = x) \\ &= 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) + \cdots + 6^2 \cdot P(X = 6) \\ &= (1 + 2^2 + \cdots + 6^2)1/6 \\ &= 91/6 \end{aligned}$$

Par conséquent, $Var(X) = 91/6 - (21/6)^2 = 105/36 \simeq 2.92$, et $\sigma_X \simeq 1.71$.

1.3.2 Cas des v.a. continues

Dans le cas d'une v.a. continue X, l'espérance est donnée par

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

Exemple – Soit X une v.a. continue, exponentielle de paramètre θ (voir plus loin). X admet donc une densité de probabilité $f(x)$ donnée par

$$f(x) = \begin{cases} \theta e^{-\theta x} & \text{pour } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

L'espérance mathématique de X est donc

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xf(x)dx \\ &= \int_{0}^{+\infty} x\theta e^{-\theta x}dx \\ &= \theta \int_{0}^{+\infty} xe^{-\theta x}dx \end{aligned}$$

En intégrant par parties, on obtient donc

$$\begin{aligned} E(X) &= \theta\Big(-\frac{1}{\theta}\big[xe^{-\theta x}\big]_0^{+\infty} + \frac{1}{\theta}\int_0^{+\infty} e^{-\theta x}dx\Big) \\ &= \int_0^{+\infty} e^{-\theta x}dx \\ &= -\frac{1}{\theta}\big[e^{-\theta x}\big]_0^{+\infty} \\ &= -\frac{1}{\theta}(0-1) = \frac{1}{\theta} \end{aligned}$$

Pour calculer la variance, on procède de la mme faon que dans le cas discret, il faut donc calculer $E(X^2)$, qui est donnée par

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \theta \int_0^{+\infty} x^2 e^{-\theta x}dx$$

que l'on intègre par parties On obtient finalement que $Var(X) = \frac{1}{\theta^2}$.

1.3.3 Quelques propriétés utiles

Propriété 1.2 *Soit X une v.a., a et $b \in \mathbb{R}$, alors:*

$$E(aX + b) = aE(X) + b$$

(on dit que l'espérance mathématique est un opérateur linéaire*).*

	V.a. discrètes	V.a. continues
Nature de Ω	Discret	Continu
F. Répartition	$F(k) = \sum_{i=1}^{k} P(X = i)$	$F(x) = \int_{-\infty}^{x} f(t)dt$
Espérance	$E(X) = \sum_{k=1}^{n} x_k P(X = k)$	$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$
Variance	$V(X) = \sum_{k=1}^{n} (x_k - \mu)^2 P(X = k)$	$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$

TAB. 1.1 – *Rappel des différentes propriétés des v.a. discrètes et continues*

Propriété 1.3 *Si X et Y sont deux v.a. indépendantes, alors*

$$E(XY) = E(X)E(Y)$$

Si X et Y sont deux v.a. non indépendantes, alors on ne peut rien dire à priori sur $E(XY)$. Par contre, que X et Y soient indépendantes ou non, puisque l'espérance est linéaire, on a toujours

$$E(X + Y) = E(X) + E(Y)$$

Propriété 1.4 *Soit X une v.a., a et $b \in \mathbb{R}$, alors:*

$$Var(aX + b) = a^2 Var(X)$$

Par ailleurs, on peut aussi écrire la variance sous la forme suivante:

$$Var(X) = E(X^2) - E(X)^2 \tag{1.4}$$

Soit X et Y deux v.a. indépendantes, alors

$$Var(X + Y) = Var(X) + Var(Y)$$

La forme (1.4) de la variance sera souvent utilisée pour simplifier les calculs.

1.4 Variables aléatoires discrètes

Dénomination	Loi de probabilité	Moyenne	Variance
Bernoulli $B(p)$	$P(X = x_1) = p$ et $P(X = x_2) = 1 - p$	p	$p(1-p)$
Binomiale $B(n,p)$	$P(X = k) = C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
Loi de Poisson $\lambda > 0$	$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ

TAB. 1.2 – *Principales lois discrètes*

1.4.1 Loi de Bernoulli

On dit qu'une v.a. X suit la loi de Bernoulli de paramètre p lorsqu'elle peut prendre les valeurs 1 ou 0 (succès ou échec) avec les probabilités respectives p et $1-p$.

La loi de Bernoulli décrit par exemple l'issue du jet d'une pièce de monnaie, avec dans ce cas $p = 0.5$.

1.4.2 Loi binomiale

Une loi binomiale de paramètre (n,p) est la somme de n variables de Bernoulli de paramètre p. Elle modélise le nombre de succès parmi n épreuves indépendantes (tirage au sort avec remise), où $P(\text{succès}) = p$ pour chaque tirage.

1.4.3 Loi de Poisson

Cette distribution est appropriée lorsque l'on étudie la fréquence d'un évènement dont la probabilité est infime mais qui se produit quand même parce qu'il possède un très grand nombre d'occasions de se réaliser. Elle est utile par exemple dans le domaine biomédical pour décrire l'apparition de phénomènes accidentels, tels les mutations.

1.5 Variables aléatoires continues

1.5.1 Loi uniforme

On dit qu'une v.a. X suit une loi uniforme sur l'intervalle $[a,b]$ lorsque X peut prendre n'importe qu'elle valeur de l'intervalle $[a,b]$ avec une densité de probabilité constante $\frac{1}{b-a}$.

1.5.2 Loi normale

La loi normale, ou loi de Laplace-Gauss, est une loi fondamentale dans le calcul des probabilités et dans les statistiques.

Théorème 1.1 *Soit X une v.a. suivant une loi normale de paramètres (μ,σ^2). Alors la v.a. $Y = \frac{X-\mu}{\sigma}$ suit une loi normale centrée réduite.*

Remarque– On note aussi $Y \rightsquigarrow \mathcal{N}(0,1)$.

Ce théorème permet d'étudier des lois normales de paramètres (μ,σ^2), puisque la seule loi qui soit tabulée est la loi normale centrée réduite.

Dénomination	Loi de probabilité	Moyenne	Variance
Uniforme	$f(x) = \begin{cases} \frac{1}{b-a} \text{ pour } x \in [a,b] \\ 0 \text{ sinon} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normale centrée réduite	$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	0	1
Normale	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	μ	σ^2
Exponentielle θ	$f(x) = \begin{cases} \theta e^{-\theta x} & \text{pour } x \geq 0 \\ 0 & \text{sinon} \end{cases}$	$\frac{1}{\theta}$	$\frac{1}{\theta^2}$
Chi deux à ν d.d.l	$f(x) = \begin{cases} \frac{e^{-x/2} x^{\nu-2}}{2^{\nu/2}\Gamma(\nu/2)} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$	ν	2ν
Student à ν d.d.l	$f(x) = \frac{(1+\frac{x^2}{\nu})^{-\frac{1}{2}(\nu+1)}}{\sqrt{\nu}\beta(\frac{1}{2},\frac{\nu}{2})}$	0 pour $\nu \geq 2$	$\frac{\nu}{\nu-2}$ pour $\nu \geq 3$

TAB. 1.3 – *Principales lois continues*

La place prépondérante de la loi de Laplace-Gauss découle du résultat suivant, appelé théorème central limite ou théorème de la limite centrale.

Théorème 1.2 (TCL) *Soit $X_1, \ldots, X_n$ une suite de v.a. indépendantes et identiquement distribuées (i.i.d.,* ie *de mme moyenne μ et variance σ^2). Alors la somme $S_n = \sum_{i=1}^{n} X_i$ suit une loi normale $\mathcal{N}(n\mu, n\sigma^2)$.*

Ce théorème affirme que la répétition un certain nombre de fois d'une expérience aléatoire suit une loi normale. En pratique, on considère que cette approximation devient valable pour un nombre d'expériences $n > 30$.

Remarque– Il découle directement du TCL que

$$\frac{S_n}{n} \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

En effet,

$$\begin{aligned} E(\frac{S_n}{n}) &= \frac{1}{n} E(S_n) \\ &= \frac{1}{n} E(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

et

$$\begin{aligned} Var(\frac{S_n}{n}) &= \frac{1}{n^2} Var(S_n) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

1.5.3 Loi log-normale

Loi log-normale – On dit qu'une v.a. X suit une loi *log-normale* si la v.a. $Y = \ln X$ suit une loi normale

La loi log-normale est l'une des lois les plus usitées dans le domaine des sciences de la vie. En effet, là o la loi normale autorise des valeurs négatives, la loi log-normale ne permet que des valeurs positives. Par ailleurs, elle permet de profiter de tous les résultats relatifs à la loi normale, modulo un simple changement de variables.

1.5.4 Loi exponentielle

La loi exponentielle est le plus souvent utilisée pour décrire des probabilités qui vont aller en diminuant avec le temps. L'exemple le plus connu est le temps d'attente d'un bus Dans le domaine biomédical, son application la plus connue est l'étude de données de *survie*.

1.6 Lecture des tables

La plupart des lois de probabilité sont tabulées. Le plus souvent, c'est leurs fonctions de répartition qui sont représentées.

De part la place prépondérante qu'elle occupe dans le calcul des probabilités, la loi normale centrée réduite est très souvent utilisée en statistiques. Par conséquent, dans tout ce qui suit, nous noterons $\Phi(t)$ sa fonction de répartition.

1.7 Exercices

Exercice 1.1 *Soit X une v.a. suivant une loi normale $\mathcal{N}(10,4)$. Exprimer sous une forme permettant l'utilisation des tables:*

1. $P(X \geq c)$
2. $P(a \leq X \leq b)$

Exercice 1.2 *Soit X une v.a. gaussienne telle que:*

$$P(X > 15) = 0.15866 \text{ et } P(X < 8) = 0.34458$$

Déterminer $E(X)$ et σ.

Exercice 1.3 *Soit X une v.a. continue de loi uniforme sur [0;100]. Calculer:*

1. $P(X = 20)$.
2. $P(0 < X < 20)$.
3. $E(X)$.
4. $Var(X)$.

Exercice 1.4 *Dans une cage de 30 souris, 20 sont blanches et 10 sont grises. On injecte des cellules cancéreuses aux souris. On suppose que la survie des souris suit une loi exponentielle, de paramètre respectivement $\lambda_b = 0.14$ et $\lambda_g = 0.1$. Calculer:*

1. *La moyenne de survie pour les deux types de souris.*
2. *La probabilité pour qu'une souris grise vive 12 mois.*
3. *La probabilité pour qu'une souris blanche vive 12 mois.*

Exercice 1.5 *La taille moyenne des individus d'une population est de 1.75 mètres, avec un écart-type de 7cm. Quelle est la probabilité:*

1. *qu'un individu mesure plus de 2 mètres?*

2. *qu'un individu mesure entre 1.60 et 1.70 mètre?*
3. *qu'un individu mesure entre 1.65 et 1.80 mètre?*

1.8 Problèmes supplémentaires

Problème 1.1 *Etablir les espérances mathématiques des diverses lois données dans ce chapitre (hormis Chi-deux et Student, trop ardues).*

Problème 1.2 *Etablir la formule 1.4:* $Var(X) = E(X^2) - E(X)^2$.

1.9 Correction des exercices

Exercice 1.1– On utilise le Théorème 1.1: la loi Y définie par $Y = \frac{X-\mu}{\sigma} = \frac{X-10}{2}$ suit une loi normale centrée réduite $\mathcal{N}(0,1)$. Par conséquent:

1.

$$\begin{aligned} P(X > c) &= 1 - P(X \leq c) \\ &= 1 - P(\frac{X-10}{2} \leq \frac{c-10}{2}) \\ &= 1 - \Phi(\frac{c-10}{2}) \end{aligned}$$

2.

$$\begin{aligned} P(a < X < b) &= P(a-10 < X-10 < b-10) \\ &= P(\frac{a-10}{2} < \frac{X-10}{2} < \frac{b-10}{2}) \\ &= P(\frac{X-10}{2} < \frac{b-10}{2}) - P(\frac{X-10}{2} < \frac{a-10}{2}) \\ &= \Phi(\frac{b-10}{2}) - \Phi(\frac{a-10}{2}) \end{aligned}$$

Exercice 1.2– On utilise le Théorème 1.1: la loi Y définie par $Y = \frac{X-E(X)}{\sigma}$ suit une $\mathcal{N}(0,1)$. On réécrit les deux contraintes de faon à faire apparaitre Y, en n'oubliant pas qu'une inégalité reste identique si, des deux cotés du signe d'inégalité, on fait les mmes opérations.

$$\begin{aligned} P(X > 15) &= P\big(X - E(X) > 15 - E(X)\big) \\ &= P\big(\frac{X-E(X)}{\sigma} > \frac{15-E(X)}{\sigma}\big) \\ &= P\big(Y > \frac{15-E(X)}{\sigma}\big) \\ &= 1 - P\big(Y < \frac{15-E(X)}{\sigma}\big) \\ &= 1 - \Phi\big(\frac{15-E(X)}{\sigma}\big) \end{aligned}$$

Par conséquent, on doit avoir $1 - \Phi\big(\frac{15-E(X)}{\sigma}\big) = 0.15866$. On cherche donc sur la table de la loi de Gauss la valeur telle que

$$\Phi\big(\frac{15-E(X)}{\sigma}\big) = 0{,}84134$$

Par conséquent, on doit avoir $\frac{15-E(X)}{\sigma} = 1$, soit $15 - E(X) = \sigma$.

De la mme faon, on a

$$\begin{aligned} P(X<8) &= P\big(X-E(X)<8-E(X)\big) \\ &= P\big(\frac{X-E(X)}{\sigma}<\frac{8-E(X)}{\sigma}\big) \\ &= P\big(Y<\frac{8-E(X)}{\sigma}\big) \end{aligned}$$

Par conséquent, on doit avoir $\Phi\big(\frac{8-E(X)}{\sigma}\big)=0.34458$. Les tables ne présente que les valeurs de la loi de Gauss pour des $x\geq 0$, soit des probabilités supérieures à 0.5. Ici, nous devons donc utiliser la symétrie de la loi, c'est à dire que $\Phi(x)=1-\Phi(-x)$. Donc $8-E(X)=-0.4016\sigma$.

On a donc à résoudre le système suivant:

$$\left|\begin{array}{l} 15-E(X)=\sigma \\ 8-E(X)=-0.4016\sigma \end{array}\right.$$

En écrivant ces deux termes en fonction de $E(X)$, on a $15-\sigma=8+0.4016\sigma$, soit $\sigma=7/1.4016=4.99\simeq 5$, et, en utilisant la première équation, $E(X)=15-\sigma=10$.

Exercice 1.3–

1. Piège! Pour une loi continue, la probabilité d'un point isolé est égale à zéro. En effet, dans notre exemple:

$$\begin{aligned} P(X=20) &= \int_{20}^{20}\frac{dx}{100} \\ &= \frac{1}{100}[x]_{20}^{20} \\ &= \frac{1}{100}(20-20) \\ &= 0 \end{aligned}$$

2. On calcule

$$\begin{aligned} P(0<X<20) &= \int_{0}^{20}\frac{dx}{100} \\ &= \frac{1}{100}[x]_{0}^{20} \\ &= \frac{1}{100}(20-0) \\ &= \frac{1}{5} \end{aligned}$$

3. On utilise le tableau 1.5: $E(X)=\frac{0+100}{2}=50$.
4. On utilise le tableau 1.5: $Var(X)=\frac{(100-0)^2}{12}=833.33$.

Exercice 1.4–

1. Soit X_b la v.a. "durée de survie d'une souris blanche", X_g la v.a. "durée de survie d'une souris grise. On utilise le tableau 1.5: $E(X_b) = \frac{1}{\lambda_b} = 7.14$ et $E(X_g) = 10$.
2. On doit calculer:

$$\begin{aligned} P(X_g \geq 12) &= 1 - P(X_g < 12) \\ &= 1 - \int_0^{12} \lambda_g e^{-\lambda_g x} dx \\ &= 1 - \lambda_g \int_0^{12} e^{-\lambda_g x} dx \\ &= 1 - \lambda_g [-\frac{1}{\lambda_g} e^{-\lambda_g x}]_0^{12} \\ &= 1 + (e^{-12\lambda_g} - e^0) \\ &= e^{-12\lambda_g} \\ &= 0.301 \end{aligned}$$

3. De la même manière, on a $P(X_b \geq 12) = e^{-12\lambda_b} = 0.187$.

Exercice 1.5– En l'absence de précisions sur la loi de probabilité que suit la taille d'un individu, on va faire l'*hypothèse de normalité*: la taille X suit une loi de Laplace-Gauss, de paramètres (175,49). Ici, plusieurs approches sont possibles (mais doivent donner les mêmes résultats) selon la table de la loi de Gauss que l'on utilise:

- Sur la table I, qui donne $P = P(X \geq u)$ et $Q = P(X \leq u)$, on doit procéder "à l'envers": on cherche dans la table la valeur la plus proche, et on lit le P ou le Q correspondant, dans les marges.
- Sur la table II (page 300 de *Statistiques descriptives et décisionnelles*), on peut lire directement la probabilité. C'est celle-ci que nous utiliserons ici.

1. On cherche $P(X > 200)$. On centre et on réduit:

$$\begin{aligned} P(X > 200) &= P(X - 175 > 25) \\ &= P(\frac{X - 175}{7} > \frac{25}{7}) \\ &= 1 - P(\frac{X - 175}{7} \leq \frac{25}{7}) \\ &= 1 - \Phi(\frac{25}{7}) = 1 - \Phi(3.57) \end{aligned}$$

On cherche dans la table (où $\Pi(t) = \Phi(t)$). On doit lire dans les grandes valeurs de t (et le résultat est alors utilisable directement, puisque la table des grandes valeurs donne $1 - \Phi(t)$). On lit la colonne 0.6 de la ligne $t = 3$. Par conséquent,

$$P(X > 200) \simeq 159 \cdot 10^{-6}$$

soit encore $1.59 \cdot 10^{-4}$ (environ 1.6 pour 10000).

2. On cherche maintenant $P(160 < X < 170)$. On centre et on réduit:

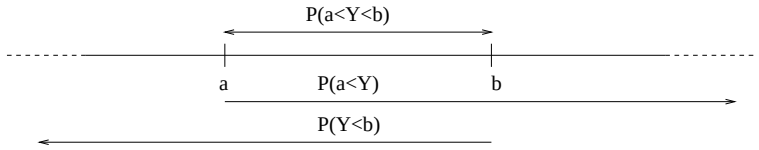
$$\begin{aligned} P(160 < X < 170) &= P(-15 < X - 175 < -5) \\ &= P(\frac{-15}{7} < \frac{X-175}{7} < \frac{-5}{7}) \end{aligned}$$

On va utiliser la propriété suivante:

Propriété 1.5 *Soit Y une v.a., P sa loi de probabilité et $F(t)$ sa fonction de répartition. Alors*

$$P(a < Y < b) = F(b) - F(a)$$

Ceci se montre facilement: $P(a < Y < b) = P(a < Y) \cap P(Y < b) = P(a < Y) + P(Y < b) - P[(a < Y) \cup (Y < b)]$ d'après la propriété 3 d'une probabilité. Or, puisque $a < b$, on a $P[(a < Y) \cup (Y < b)] = 1$ (on regardera la figure ci-dessus pour s'en convaincre, $\cup$ consistant en la réunion des deux demi-droites $P(a < Y)$ et $P(Y < b)$), et par conséquent,



$P(a < Y < b) = 1 - P(Y < a) + P(Y < b) - 1 = P(Y < b) - P(Y < a) = F(b) - F(a)$.
Par conséquent:

$$P(160 < X < 170) = \Phi(\frac{-5}{7}) - \Phi(\frac{-15}{7})$$

La table ne donne les probabilités que pour les valeurs positives de t. Mais la symétrie de la loi de Gauss permet d'écrire que $\Phi(-t) = 1 - \Phi(t)$ pour tout t (et en particulier pour $t < 0$). Par conséquent,

$$\begin{aligned} P(160 < X < 170) &= 1 - \Phi(\frac{5}{7}) - (1 - \Phi(\frac{15}{7})) \\ &= \Phi(\frac{15}{7}) - \Phi(\frac{5}{7}) \\ &= \Phi(2.14) - \Phi(0.71) \\ &= 0.98382 - 0.76115 \\ &= 0.22267 \end{aligned}$$

3. On fait le même raisonnement que dans le point précédent:

$$P(165 < X < 180) = P(\frac{165-175}{7} < \frac{X-175}{7} < \frac{180-175}{7})$$

$$\begin{aligned} &= \Phi(\frac{5}{7}) - \Phi(\frac{-10}{7}) \\ &= 0.76115 - (1 - \Phi(1.43)) \\ &= 0.76115 - 1 + 0.92364 \\ &= 0.68469 \end{aligned}$$

2. Introduction aux statistiques

2.1 Variabilité

Un problème qui se pose lorsque l'on est confronté à des mesures expérimentales est celui de la *variabilité*. Cette variabilité peut être de plusieur types:

Variabilité interne. Propre à la population que l'on considère. Dans le cas de mesures biologiques, elle peut être *intra individuelle*, c'est à dire liée à l'évolution avec le temps et la situation de l'objet de l'étude, ou *inter individuelle*, liée aux différences biologiques entre les individus d'une même population.

Variabilité externe. Dans le cas de mesures de populations biologiques, ce type de variabilité a deux sources: les *instruments* de mesure, qui sont plus ou moins précis, et l'*observateur*, qui est, lui aussi, plus ou moins précis

L'objet de la *statistique* est de dégager des *lois* de la connaissance des mesures, en prennant en compte la variabilité, et de préciser la validité de ces lois.

Un concept important en statistiques est la notion de *loi vraie*. On suppose que la variable aléatoire suit une loi dite vraie, donnée par la nature. Le but des statistiques est de connatre cette loi, en se basant sur l'observation d'éléments issus de cette loi.

2.2 Langage

Modèle statistique – Un *modèle statistique* est la donnée d'un ensemble Ω et d'une famille $(P_\theta)_{\theta\in\Theta}$ de lois de probabilité sur cet ensemble.

Echantillon – On appelle *échantillon* de taille n de X, ou encore n-échantillon de X, une suite de n variables aléatoires $X_1,\ldots,X_n$, indépendantes et de même loi que X.

Réalisation d'un échantillon – Les valeurs prises par l'échantillon, notées

$x_1, \ldots, x_n$, sont appelées *réalisation* de l'échantillon.

Statistique – On appelle *statistique* toute fonction d'un n-échantillon. Une statistique est donc une variable aléatoire fonction de $X_1, \ldots, X_n$.

Par exemple, la moyenne empirique est une statistique.

2.3 Statistique descriptive

Cette branche de la statistique s'occupe de la description des données. Nous traiterons rapidement cette section, la statistique déscriptive devrait en effet vue comme un moyen de se familiariser avec les données, et non comme un moyen d'analyse.

On suppose que pour chaque *individu* d'une *population*, on mesure la valeur d'un *caractère*. On est souvent obligé de regrouper les données en classes, de façon à les rendre plus synthétiques. Toutefois, cette opération fait perdre de l'*information*.

Dans la suite, nous supposerons que l'on a procédé à un tel regroupement, en k classes. On a donc une *table des effectifs*, qui donne, pour chaque valeur du caractère, le nombre d'individus qui le possèdent à un tel niveau:

Valeur du caractère	x_1	x_2	$\ldots$	x_k
Nombre d'individus	n_1	n_2	$\ldots$	n_k

Pour simplifier, nous noterons $n = \sum_{i=1}^{k} n_i$ la taille de l'échantillon.

Fréquence relative – La *fréquence relative* du caractère x_i dans la population est donnée par

$$f_i = \frac{n_i}{n} \tag{2.1}$$

Table des effectifs cumulés – A la table des effectifs, on peut associer une table des *effectifs cumulés* qui donne, pour chaque valeur du caractère, le nombre d'individus ayant ce caractère ou l'un des précédents.

Bien sur, il faut pour cela que le caractère soit de nature *ordinale* (*i.e.* que l'on puisse le classer) et non de nature *cardinale*. Ainsi, établir une table des effectifs cumulés des tailles d'un être humain a un sens, alors qu'établir une table des effectifs cumulés de la couleur d'une voiture en a moins

On obtient une table ayant l'apparence suivante:

Valeur du caractère	x_1	x_2	$\ldots$	x_k
Nombre d'individus	n_1	n_2	$\ldots$	n_k
Effectif cumulé	$c_1 = n_1$	$c_2 = n_1 + n_2$	$\ldots$	$c_k = \sum_{i=1}^{k} n_i$

Fonction de répartition empirique – La *fonction de répartition empirique* est calculée en utilisant la table des effectifs cumulés:

$$F(x_i) = \frac{c_i}{n} \tag{2.2}$$

On a bien sur

$$F(x_i) = \sum_{j=1}^{i} f_i$$

2.3.1 Caractéristiques de position d'une distribution

Centiles – On appelle α_i-*centile* d'une distribution la valeur du caractère tel que $F(x_i) = \frac{\alpha_i}{100}$.

Quartiles – On appelle *quartiles* les 25 et 75-centiles d'une distribution.

Médiane – La *médiane* correspond au $50^{\text{ème}}$ centile d'une distribution (*i.e.* $\alpha_i = 0.5$). C'est donc la valeur centrale de la distribution. Elle est calculée en cherchant le i tel que $F(x_i) = 0.5$.

> Dans le cas continu, si le point cherché tombe "entre deux classes", on utilisera une interpolation linéaire (*i.e.* que l'on relie les deux points, on trouve l'intersection avec la droite $y = \alpha_i$, et on en déduit la valeur de x_i), ou bien la moyenne des deux classes.

Mode – Le *mode* d'une distribution est la (les) valeur(s) la plus fréquente dans la distribution.

Moyenne expérimentale – La *moyenne expérimentale*, ou *moyenne empirique* d'un échantillon est donnée par:

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

Remarque– Si la distribution est unimodale et symétrique, alors la médiane, le mode et la moyenne expérimentale coïncident. La réciproque est fausse en général.

2.3.2 Caractéristiques de dispersion d'une distribution

Variance empirique – La variance (empirique) d'un échantillon est la moyenne des carrés des écarts à la moyenne:

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

Intervalle interquartile – L'*intervalle interquartile* est la différence entre le troisième et le premier quartile.

Remarque– L'intervalle interquartile contient donc environ 50% des observations. De la même manière, on définit aussi l'intervalle interdécile, qui comprend environ 90% des observations, et l'intervalle intercentile qui contient lui environ 99% des observations.

Coefficient de dissymétrie – Le *coefficient de dissymétrie* se calcule à partir de la médiane M_e, du premier quartile Q_1 et du troisième quartile Q_3, par la formule suivante:

$$\frac{(Q_3 - M_e) - (M_e - Q_1)}{Q_3 - Q_1}$$

Coefficient de variation – On appelle *coefficient de variation* d'une distribution le quotient de l'écart-type par la moyenne (empirique):

$$V = \frac{s}{\bar{x}}$$

Souvent, on exprime ce coefficient en pourcentage, en utilisant donc

$$V = \frac{100s}{\bar{x}}$$

Le coefficient de variation permet de réduire les effets de taille. Supposons par exemple que l'on mesure (en centimètres) la taille moyenne d'un groupe de baleines et d'un groupe de souris. Le calcul de la variance (et donc de l'écart-type) se base sur les carrés des écarts à la moyenne. Il va donc amplifier les différences pour les baleines (un écart de 30 centimètres de la moyenne va donner 900 dans le calcul de la variance). Si l'on veut pouvoir comparer les variations dans ces deux populations, on doit donc restreindre cet effet.

2.4 Données groupées – Données non groupées

Le tableau suivant donne les différences de calcul de certaines des quantités qui ont été présentées, selon que les données ont été ou non regroupées en classes.

Quantité	Données non groupées	Données groupées
effectif	n	$\sum_{i=1}^{k} F(x_i)$
somme des observations	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{k} x_i F(x_i)$
somme des carrés	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{k} x_i^2 F(x_i)$
moyenne	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{x} = \frac{\sum xF(x)}{\sum F(x)}$
variance	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$	$s^2 = \frac{\sum [(x-\bar{x})^2 F(x)]}{\sum F(x)-1}$

2.5 Exercices

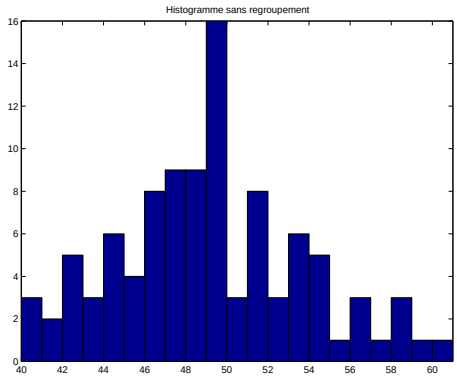
Exercice 2.1 *On étudie la valeur d'un caractère sur un échantillon de 100 individus. On obtient les valeurs suivantes:*

61	*50*	*59*	*44*	*50*	*48*	*41*	*46*	*50*	*55*
52	*51*	*43*	*54*	*48*	*49*	*44*	*56*	*50*	*42*
55	*50*	*54*	*60*	*41*	*52*	*54*	*50*	*47*	*43*
58	*49*	*50*	*57*	*48*	*47*	*40*	*48*	*46*	*45*
50	*50*	*52*	*55*	*47*	*48*	*52*	*53*	*50*	*52*
57	*47*	*49*	*48*	*45*	*54*	*48*	*45*	*52*	*48*
45	*50*	*44*	*49*	*45*	*42*	*43*	*50*	*51*	*47*
43	*55*	*59*	*46*	*46*	*49*	*54*	*54*	*47*	*55*
43	*48*	*47*	*50*	*47*	*49*	*49*	*50*	*53*	*50*
57	*59*	*51*	*52*	*53*	*50*	*49*	*52*	*45*	*49*

1. *Donner la table des effectifs et des fréquences de cet échantillon.*
2. *Donner la table des effectifs cumulés et des fréquences cumulées de cet échantillon.*
3. *Tracer la courbe des fréquences cumulées.*
4. *Déterminer la moyenne, le mode de cet échantillon.*
5. *Déterminer la variance de cet échantillon.*
6. *Déterminer la médiane, les quartiles (25 et 75èmes centiles) de cet échantillon.*
7. *Cette distribution est-elle symétrique?*
8. *Donner l'intervalle interquartile.*
9. *Déterminer le coefficient de dissymétrie de cet échantillon.*
10. *Déterminer le coefficient de variation de cet échantillon.*

2.6 Correction des exercices

Exercice 2.1– Ce cas est très simple, puisque les fréquences s'obtiennent directement en divisant par 100 les valeurs dans les différentes classes. Si on fait l'histogramme de l'échantillon, on a:



1. On calcule la somme des valeurs: 4966, et on classe les différentes valeurs:

Valeur	40	41	42	43	44	45	46	47	48	49
Nombre	1	2	2	5	3	6	4	8	9	9
Fréquence	0.01	0.02	0.02	0.05	0.03	0.06	0.04	0.08	0.09	0.09

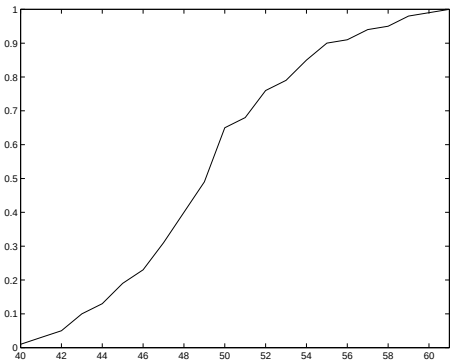
50	51	52	53	54	55	56	57	58	59	60	61
16	3	8	3	6	5	1	3	1	3	1	1
0.16	0.03	0.08	0.03	0.06	0.05	0.01	0.03	0.01	0.03	0.01	0.01

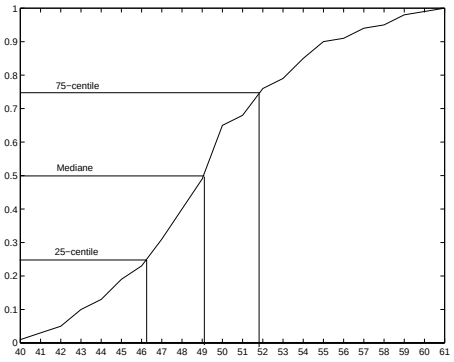
2. Effectifs cumulés et fréquences cumulées:

Valeur	40	41	42	43	44	45	46	47	48	49
Nombre	1	3	5	10	13	19	23	31	40	49
Fréquence	0.01	0.03	0.05	0.1	0.13	0.19	0.23	0.31	0.4	0.49

50	51	52	53	54	55	56	57	58	59	60	61
65	68	76	79	85	90	91	94	95	98	99	100
0.65	0.68	0.76	0.79	0.85	0.9	0.91	0.94	0.95	0.98	0.99	1

3. On obtient la figure suivante:
4. Moyenne: 49.66 (directement avec la somme). Mode: 50 (d'après le tableau).
5. Variance: 20.7721 (écart type: 4.5576).
6. Médiane: 50 (d'après la figure). Quartiles: 47 et 52 (d'après la figure). Attention: on arrondit au supérieur. Il s'agit en effet de valeurs discrètes des caractères.





7. Moyenne= 49.66, mode=50 et médiane=50. On ne peut pas pour autant conclure sur la symétrie, il suffit de regarder l'histogramme pour s'en convaincre.
8. L'intervalle interquartile est donné par $Q_3 - Q_1 = 5$: 5 classes regroupent environ 50% des observations.
9. On calcule

$$\frac{(Q_3 - M_e) - (M_e - Q_1)}{Q_3 - Q_1} = \frac{(52 - 50) - (50 - 47)}{52 - 47} = -\frac{1}{5}$$

   On peut donc conclure que l'échantillon observé contient légèrement plus de valeurs à gauche qu'à droite de la médiane. Donc la distribution n'est pas symétrique.
10. On calcule $100s/\bar{x} = 455.76/49.66 = 9.18\%$.

3. Estimation

L'estimation est un aspect important du travail statistique. Elle consiste à essayer de déduire la *vraie* valeur d'une caractéristique (d'un paramètre) d'une population, au vu d'un échantillon non nécessairement exhaustif de cette population.

Considérons une culture bactérienne. L'expérimentateur aimerait avoir une information sur la taille moyenne des bactéries. Plusieurs problèmes se posent:

- Il ne peut pas de faon réaliste mesurer la taille (*paramètre*) de toutes les bactéries de sa culture (*population*).
- Par conséquent, il prend un *échantillon* de la population, et mesure la taille des bactéries de cet échantillon.

Au vu de ses mesures, il peut calculer une estimation de la taille, dont il peut supposer qu'elle représente la taille moyenne des cellules de sa population (*estimation ponctuelle*). Toutefois, puisque l'expérimentateur n'a pas mesuré toutes les bactéries, et qu'il sait que la variabilité de la taille des cellules est importante, quelle confiance peut-il accorder à cette estimation? Le but de l'estimation *par intervalles de confiance* est de donner une information sur le risque pris lors de l'estimation.

3.1 Estimation ponctuelle

Soit X une v.a. suivant une loi de probabilité d'espérance mathématique $\mu = E(X)$ et de variance $\sigma^2 = E\left[(X-\mu)^2\right]$. On procède à un échantillonage, i.e que l'on constitue n v.a. $X_1,\ldots,X_n$ indépendantes et de même loi que X par répétition n fois de l'expérience de façon indépendante. On obtient une réalisation $x_1,\ldots,x_n$ des v.a. $X_1,\ldots,X_n$.

Estimateur – Soit $X_1,\ldots,X_n$ un n-échantillon, $(x_1,\ldots,x_n)$ une réalisation de $(X_1,\ldots,X_n)$. La variable aléatoire $T = \varphi(X_1,\ldots,X_n)$ est un *estimateur*, le réel $\varphi(x_1,\ldots,x_n)$ est une *estimation*, la fonction φ étant une *statistique*.

Lorsque la loi de la v.a. étudiée dépend d'un paramètre inconnu θ, une statistique prennant ses valeurs dans l'ensemble des valeurs possibles de θ s'appelle un estimateur de θ. La valeur $\hat{\theta}$ qu'elle prendra pour une réalisation du n-échantillon sera l'estimation correspondante de θ.

Remarque– L'estimateur étant une statistique, c'est une variable aléatoire.

Estimateur sans biais – Un estimateur est dit *sans biais* si son espérance mathématique est égale au paramètre qu'il estime. En d'autres termes, si $\hat{\theta}$ est un estimateur de θ, c'est un estimateur sans biais de θ si et seulement si

$$E(\hat{\theta}) = \theta$$

Estimateur biaisé – Si $E(\hat{\theta}) \neq \theta$, alors on dit que l'estimateur est biaisé. La quantité $E(\hat{\theta}) - \theta$ est alors appelé le *biais* de l'estimateur.

3.1.1 Estimateurs ponctuels usuels

Proportion – Pour estimer la proportion p d'individus d'une population possédant une certaine caractéristique, à partir d'un n-échantillon, on compte le nombre k d'individus de l'échantillon possédant la caractéristique. L'estimateur de la proportion est alors $\hat{p} = \frac{k}{n}$. C'est un estimateur sans biais de p.

Moyenne – L'estimateur usuel de la moyenne est $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. C'est un estimateur sans biais de μ.

Remarque– Dans certains calculs, on se servira de la propriété suivante (admise):

$$Var(\bar{X}) = \frac{1}{n} Var(X)$$

Variance – Deux estimateurs sont habituellement utilisés pour la variance. Le premier, défini par

$$\hat{s}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

est un estimateur sans biais de σ^2. Le second, défini par

$$\hat{s}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

est un estimateur biaisé de σ^2, asymptotiquement sans biais (*i.e.* que pour une réalisation de grande taille, le biais de l'estimation tend vers 0).

Justification

Les deux estimateurs de la variance diffèrent seulement par le terme en facteur devant la somme. Calculons donc $\sum(X_i - \bar{X})^2$:

$$\sum(X_i - \bar{X})^2 \quad = \quad \sum[(X_i - \mu) + (\mu - \bar{X})]^2$$

$$\begin{aligned} &= \sum[(X_i-\mu)^2 + 2(X_i-\mu)(\mu-\bar{X}) + (\mu-\bar{X})^2] \\ &= \sum(X_i-\mu)^2 + 2\sum(X_i-\mu)(\mu-\bar{X}) + \sum(\mu-\bar{X})^2 \end{aligned}$$

Or

$$\begin{aligned} 2\sum(X_i-\mu)(\mu-\bar{X}) &= 2(\mu-\bar{X})\sum(X_i-\mu) \\ &= 2(\mu-\bar{X})[\sum X_i - n\mu] \\ &= 2(\mu-\bar{X})[n\bar{X}-n\mu] \\ &= 2(\mu-\bar{X})(\bar{X}-\mu) \end{aligned}$$

Donc

$$\begin{aligned} \sum(X_i-\bar{X})^2 &= \sum(X_i-\mu)^2 - 2n(\mu-\bar{X})^2 + n(\mu-\bar{X}) \\ &= \sum(X_i-\mu)^2 - n(\mu-\bar{X})^2 \end{aligned}$$

Calculons maintenant $E(\sum(X_i-\bar{X})^2)$ en utilisant cette dernière expression.

$$\begin{aligned} E(\sum(X_i-\bar{X})^2) &= E[\sum(X_i-\mu)^2 - n(\mu-\bar{X})^2] \\ &= E[\sum(X_i-\mu)^2] - E[n(\mu-\bar{X})^2] \\ &= \sum E[(X_i-\mu)^2] - nE[(\mu-\bar{X})^2] \\ &= \sum \sigma^2 - nVar(\bar{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2 \end{aligned}$$

On peut donc maintenant calculer très facilement l'espérance mathématique des deux estimateurs de la variance:

$$\begin{aligned} E(\hat{s}_1^2) &= E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2] \\ &= \frac{1}{n-1}E[\sum_{i=1}^{n}(X_i-\hat{\mu})^2] \\ &= \frac{1}{n-1}(n-1)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

qui est donc un estimateur sans biais de la variance, et

$$E(\hat{s}_2^2) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2]$$

$$\begin{aligned} &= \frac{1}{n}E[\sum_{i=1}^{n}(X_i - \hat{\mu})^2] \\ &= \frac{1}{n}(n-1)\sigma^2 \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

qui est un estimateur biaisé de la variance. Mais c'est un estimateur asymptotiquement sans biais, ce que l'on peut montrer de deux manières:

- Soit l'on considère la limite quand n devient grand (tend vers l'infini) de $E(\hat{s}_2^2)$:

  $$\lim_{n\to\infty} E(\hat{s}_2^2) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

  (en se rappelant que la limite en ∞ d'une fraction polynomiale est donnée par la limite en ∞ du quotient des termes de plus haut degré du numérateur et du dénominateur –ici n dans les deux cas, donc $\frac{n}{n} = 1$–).
- Soit on calcule le biais de l'estimateur:

  $$\begin{aligned} E(\hat{s}_2^2) - \sigma^2 &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= (\frac{n-1}{n} - 1)\sigma^2 \\ &= -\frac{\sigma^2}{n} \end{aligned}$$

  et l'on calcule sa limite quand $n \to \infty$, qui vaut 0. Puisque le biais tend vers 0 quand la taille de l'échantillon augmente, c'est que l'erreur que l'on commet devient de plus en plus petite. Donc l'estimateur $\hat{s}_2^2$ est asymptotiquement sans biais.

Pour un échantillon de taille modeste, on utilisera de préférence $\hat{s}_1^2$, alors que pour des échantillons de taille conséquente (50), on utilisera celui que l'on veut.

3.2 Estimation par intervalles

De faon à faire apparatre la dépendance des lois de probabilité de leurs paramètres, nous les noterons par la suite P_θ.

Intervalle de fluctuation – On appelle *intervalle de fluctuation* (IF) au *niveau* α (ou au *risque* $1-\alpha$) d'une valeur observée x d'un caractère, l'intervalle I centré en μ tel que

$$P_\theta(x \in I) = \alpha \tag{3.1}$$

Un intervalle de fluctuation comprend donc, avec une probabilité supérieure à α, les valeurs possibles de la réalisation.

On se donne la probabilité α d'une erreur (en général, $\alpha = 0.05$ ou 0.01). Cette erreur est souvent appelée *niveau de signification*, ou encore *seuil de signification*.

Intervalle de confiance – On appelle *intervalle de confiance* (IC) au *niveau* (de *sécurité*) $1 - \alpha$ du paramètre θ, l'intervalle I tel que

$$P_\theta(X \in I) = 1 - \alpha \tag{3.2}$$

Un intervalle de confiance comprend donc, avec une probabilité supérieure à $1 - \alpha$, les valeurs possibles du paramètre θ.

Afin de savoir quelle loi de probabilité utiliser selon les cas possibles, on se réferera à la table 3.2, qui donne, selon le paramètre que l'on cherche à estimer et la loi de la v.a., l'estimateur à utiliser.

Exemple – On observe la variable X: nombre d'interventions journalières de pompiers d'une (petite) caserne, et on obtient l'échantillon de taille n suivant:

Nombre d'interventions/jour	Nombre de jours
0	13
1	26
2	28
3	18
4	11
5	4
6	0

On fait l'hypothèse que la v.a. considérée suit une loi de Poisson de paramètre λ. Construire un intervalle de confiance pour ce paramètre, au niveau $\alpha = 0.05$.

On calcule: $\bar{X} = 2$ et $s_2^2 = 1.76$. La distribution d'une loi de Poisson est donnée par:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

et le paramètre λ est aussi l'espérance de la loi.

Pour construire l'intervalle de confiance, ce que l'on cherche à faire c'est trouver T_1 et T_2 tels que:

$$P[T_1 < \lambda < T_2] \geq 1 - \alpha$$

On va utiliser une approximation normale (on le fera le plus souvent possible, lorsque la taille de l'échantillon le permet, *i.e.* est supérieure à 30). Ainsi, on pourra lire les valeurs sur la table de la loi normale. On a

$$\frac{\bar{X} - E(X)}{\sqrt{Var(\bar{X})}} \rightsquigarrow \mathcal{N}(0,1)$$

avec $Var(\bar{X}) = Var(X)/n$.

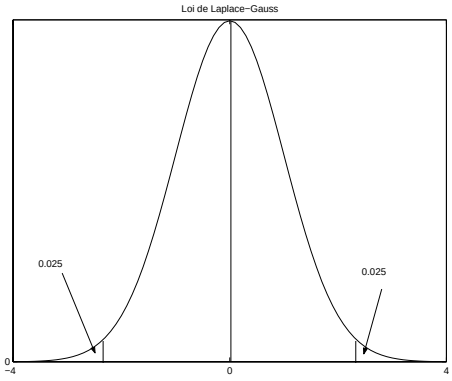
On réécrit cette quantité:

$$\begin{aligned} \frac{\bar{X} - E(X)}{\sqrt{Var(\bar{X})}} &= \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \\ &= \sqrt{n}\frac{\bar{X} - \mu}{\sqrt{\lambda}} \end{aligned}$$

Ce que l'on cherche est donc les u_1 et u_2 tels que

$$P[u_1 \leq \sqrt{n}\frac{\bar{X} - \mu}{\sqrt{\lambda}} \leq u_2] = 1 - \alpha$$

Ceci est représenté sur la figure suivante:



Or la loi $\mathcal{N}(0,1)$ est une loi symétrique, centrée en 0. On va donc utiliser cette propriété:

$$\begin{aligned} &P[u_1 \leq \sqrt{n}\frac{\bar{X} - \mu}{\sqrt{\lambda}} \leq u_2] = 1 - \alpha \\ &P[\left|\sqrt{n}\frac{\bar{X} - \mu}{\sqrt{\lambda}}\right| \leq u_3] = 1 - \alpha \\ &P[\left|\sqrt{n}\frac{\bar{X} - \mu}{\sqrt{\lambda}}\right| \leq 1.96] = 0.95 \\ &P[\left|\sqrt{n}(\bar{X} - \mu)\right| \leq 1.96\sqrt{\lambda}] = 0.95 \end{aligned}$$

Après quelques calculs, on trouve

$$IC_{5\%} = \left[\bar{X} - \frac{1.96^2}{200} - \sqrt{(\bar{X} + \frac{1.96^2}{200})^2 - \bar{X}^2}\,;\bar{X} - \frac{1.96^2}{200} + \sqrt{(\bar{X} + \frac{1.96^2}{200})^2 - \bar{X}^2}\right]$$

soit [1.74 ; 2.3].

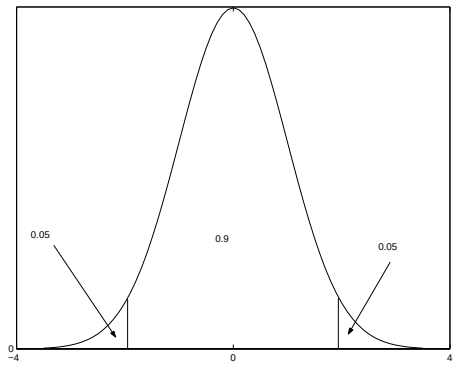
Exemple – Soit $x_1,\ldots,x_n$ la réalisation de n v.a. i.i.d. $X_1,\ldots,X_n$ de loi $\mathcal{N}(\mu,\sigma^2)$. On sait que $\sigma^2 = 9$. Donner un IC à 90% pour μ, basé sur l'estimation $\bar{x} = 19.3$ et avec $n = 16$.
On regarde la table 3.2: l'estimateur de μ suit une loi $\mathcal{N}(\bar{x},\frac{\sigma^2}{n})$. Donc on a

$$\sqrt{n}\frac{\bar{x}-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$$

On doit donc trouver T_1 et T_2 tels que:

$$P[T_1 \leq \sqrt{n}\frac{\bar{x}-\mu}{\sigma} \leq T_2] = 0.9$$

Comme $\mathcal{N}(0,1)$ est symétrique, on aura donc $T_2 = -T_1$, et on cherche alors sur la table la probabilité pour que la loi normale prenne une valeur égale à 0.95 (*i.e* à $1-\frac{\alpha}{2}$, voir la figure suivante). On y lit 1.6449.



Par conséquent on a:

$$\begin{aligned}
P[-1.6449 \leq \sqrt{n}\frac{\bar{x}-\mu}{\sigma} \leq 1.6449] &= 0.9\\
P[-1.6449\sigma \leq \sqrt{n}(\bar{x}-\mu) \leq 1.6449\sigma] &= 0.9\\
P[-1.6449\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.6449\frac{\sigma}{\sqrt{n}}] &= 0.9\\
P[-1.6449\frac{\sigma}{\sqrt{n}} - \bar{x} \leq \mu \leq 1.6449\frac{\sigma}{\sqrt{n}} - \bar{x}] &= 0.9\\
P[\bar{x} - 1.6449\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.6449\frac{\sigma}{\sqrt{n}}] &= 0.9
\end{aligned}$$

Dans cette expression, on remplace les valeurs connues ($\bar{x}$, σ et n), et on obtient finalement:

$$IC_{10\%} = [18.06\ ;\ 20.53]$$

Exemple – On veut déterminer un IC au niveau de sécurité 0.95 de la proportion p inconnue de toxicomanes dans une population. Un échantillon de taille $n = 200$ a été prélevé et a donné 10 toxicomanes.

On considère la v.a. X: nombre de personnes toxicomanes dans le n-échantillon prélevé. X suit une loi binomiale de paramètres (200,p) (on note $\mathcal{B}(200,p)$).

On fait une approximation normale, en utilisant la formule suivante (valable pour $n > 30$, $np \geq 5$ et $np(1-p) \geq 5$):

$$Z_n = \frac{X - np}{\sqrt{np(1-p)}} \rightsquigarrow \mathcal{N}(0,1) \tag{3.3}$$

Par conséquent, on cherche à déterminer ϵ_α tel que

$$P(|Z_n| \leq \epsilon_\alpha) = 1 - \alpha = 0.95$$

On trouve 1.96 dans la table de la loi normale centrée réduite. Encore une fois: on lit la table pour la valeur 0.975, car ce qu'on veut, en fait, c'est "retrancher" l'erreur $\alpha = 0.05$ à 1, et puisque la loi normale est symétrique, cela implique que l'on enlève deux fois 0.025 (comme cela est représenté sur la figure précédente).

On utilise ensuite la formule approchée suivante (qui ne sera pas démontrée ici):

$$IC = \left[\frac{X}{n} - \frac{\epsilon_\alpha}{\sqrt{n}} \sqrt{\frac{X}{n}(1 - \frac{X}{n})} \; ; \; \frac{X}{n} + \frac{\epsilon_\alpha}{\sqrt{n}} \sqrt{\frac{X}{n}(1 - \frac{X}{n})} \right] \tag{3.4}$$

On a donc

$$\begin{aligned} IC &= [\frac{10}{200} - \frac{1.96}{\sqrt{200}} \sqrt{\frac{10}{200}\frac{190}{200}} \; ; \; \frac{10}{200} + \frac{1.96}{\sqrt{200}} \sqrt{\frac{10}{200}\frac{190}{200}}] \\ &= [\frac{1}{20} - 0.14 \sqrt{\frac{1}{20}\frac{19}{20}} \; ; \; \frac{1}{20} + 0.14 \sqrt{\frac{1}{20}\frac{19}{20}}] \\ &= [0.0192 \; ; \; 0.0808] \end{aligned}$$

3.3 Exercices

Exercice 3.1 *Le nombre de plaquettes par mm^3 de sang peut être considéré comme une v.a. de moyenne 300 u/mm^3 et d'écart-type 75 u/mm^3.*

1. *Donner l'IF à 95% du nombre de plaquettes.*
2. *Quelle taille d'échantillon faudrait-il tirer au sort pour que l'IF au risque 5% de M_n (moyenne du nombre de plaquettes pour un échantillon de n sujets) ait une largeur de 10 u/mm^3 ?*

Exercice 3.2 *On observe la valeur d'un caractère sur un échantillon de 35 individus. On obtient les valeurs suivantes: 15, 6, 21, 0, 27, 9, 8, 23, 16, 26, 7, 28, 28, 13, 18, 29, 2, 17, 5, 25, 19, 24, 11, 1, 2, 0, 11, 6, 22, 22, 15, 25, 14, 24, 5.*

Déterminer un intervalle de confiance à 95% de la moyenne de ce caractère.

Paramètre à estimer	Loi de la v.a. X		Estimateur ponctuel du paramètre	Effectif de l'échantillon	Loi de l'estimateur
Moyenne μ	Loi normale	σ connu	$\bar{x}$	n quelconque	$\mathcal{N}(\bar{x}, \frac{\sigma^2}{n})$
		σ inconnu estimé par s_1		n quelconque	Student $(n-1)$ d.d.l $s(\bar{x}, \frac{s_1}{\sqrt{n}})$
	Loi inconnue	σ inconnu estimé par s_1		$n > 30$	$\mathcal{N}(\bar{x}, \frac{s_1^2}{n})$
				$n < 30$	inconnue
Proportion p	Loi parente de la loi binomiale	σ inconnu estimé par $s = \sqrt{\frac{f(1-f)}{n-1}}$	$f = \frac{k}{n}$	n quelconque	approximation normale
Ecart-type σ	Loi normale	μ connu	$\sigma_1^2 = \frac{\sum (x_i - \mu)^2}{n}$	n quelconque	$(n-1)\frac{s_1^2}{\sigma_1^2}$: du χ^2 à $n-1$ d.d.l.
		μ inconnu estimé par $\bar{x}$	$\sigma_2^2 = \frac{\sum (x_i - \bar{x})^2}{n}$	n quelconque	$(n-1)\frac{s_1^2}{\sigma_2^2}$: du χ^2 à $n-1$ d.d.l.

TAB. 3.1 – *Principaux estimateurs et leurs lois de probabilité*

Exercice 3.3 *Reprendre le deuxième exemple d'intervalle de confiance dans le cours (loi normale), et donner:*

1. *Un IC à 90% si* $n = 100$.
2. *Un IC à 95%, puis à 99%, avec* $n = 16$.

Exercice 3.4 *Une population contient 12% d'individus malades.*

1. *Quel est l'intervalle de fluctuation (IF) à 95% du nombre de malades pour un échantillon de 100 malades?*
2. *Quel est l'IF à 99% du nombre de malades pour un échantillon de 100 malades?*
3. *Donner une justification "verbale" de la différence constatée.*
4. *Que deviennent ces IF si le nombre d'individus de l'échantillon est égal à 200?*

3.4 Correction des exercices

Exercice 3.1–

1. Ici, les renseignements donnés ne sont pas suffisants pour obtenir une forme "numérique" de l'IF. On se contente de préparer son écriture. Soit X la v.a. "nombre de plaquettes par mm^3 de sang".

$$\begin{aligned}
&P(a < X < b) = 0.95\\
&P(a - 300 < X - 300 < b - 300) = 0.95\\
&P(\frac{a-300}{75} < \frac{X-300}{75} < \frac{b-300}{75}) = 0.95
\end{aligned}$$

   Maintenant, si l'on suppose que la v.a. X suit une loi normale, on peut conclure, en remplaçant a et b par respectivement -1.96 et 1.96 Mais sinon, on en reste là.

2. Le nombre moyen de plaquettes pour un échantillon de n sujets est tel que

$$M_n \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

   Construisons un IF de M_n au niveau 0.95:

$$\begin{aligned}
&P(-\epsilon_\alpha \leq \frac{M_n - \mu}{\sigma}\sqrt{n} \leq \epsilon_\alpha) = 0.95\\
&P(-\epsilon_\alpha\sigma \leq (M_n - \mu)\sqrt{n} \leq \epsilon_\alpha\sigma) = 0.95\\
&P(-\epsilon_\alpha\frac{\sigma}{\sqrt{n}} \leq M_n - \mu \leq \epsilon_\alpha\frac{\sigma}{\sqrt{n}}) = 0.95\\
&P(\mu - \epsilon_\alpha\frac{\sigma}{\sqrt{n}} \leq M_n \leq \mu + \epsilon_\alpha\frac{\sigma}{\sqrt{n}}) = 0.95
\end{aligned}$$

   On cherche un IF de largeur 10. On doit donc avoir

$$\mu + \epsilon_\alpha\frac{\sigma}{\sqrt{n}} - (\mu - \epsilon_\alpha\frac{\sigma}{\sqrt{n}}) = 10$$

   Ce qui s'écrit:

$$\frac{2\epsilon_\alpha\sigma}{\sqrt{n}} = 10$$

   On resoud cette expression en fonction de n:

$$\begin{aligned}
10 = \frac{2\epsilon_\alpha\sigma}{\sqrt{n}} \quad &\Leftrightarrow \quad \sqrt{n}10 = 2\epsilon_\alpha\sigma\\
&\Leftrightarrow \quad \sqrt{n} = \frac{2\epsilon_\alpha\sigma}{10}\\
&\Leftrightarrow \quad n = \left(\frac{2\epsilon_\alpha\sigma}{10}\right)^2
\end{aligned}$$

   Le risque $\alpha = 0.05$ donne $\epsilon_\alpha = 1.96$. On remplace les valeurs connues ($\mu = 300$, $\sigma = 75$) et on obtient finalement que $n = 865$ (en arrondissant au supérieur pour être sûr d'avoir la largeur de l'IF).

Exercice 3.2– On calcule la moyenne et la variance empiriques de l'échantillon: $m = 524/35 = 14.97$ et

$$s_1^2 = \frac{1}{n-1}\Big(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\Big) = 85.91$$

Puisque $n > 30$, on suppose que la moyenne empirique suit une $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, où μ et σ^2 sont respectivement la moyenne et la variance vraies. On estime la variance par la variance empirique.

L'IC au risque 5% (ou au niveau 95%) est donc donné par:

$$14.97 - 1.96\sqrt{\frac{85.91}{35}} \leq \mu \leq 14.97 + 1.96\sqrt{\frac{85.91}{35}}$$

soit [11.9 ; 18.04].

Exercice 3.3– On reprend ici la formule qui a été obtenue dans l'exemple, en l'adaptant selon les cas.

1. Ici, on peut reprendre directement, en remplaçant $n = 10$ par $n = 100$ dans les calculs. On obtient donc l'intervalle de confiance suivant:
$$IC_{10\%} = [19.3 - 1.6449\frac{3}{10}\ ;\ 19.3 + 1.6449\frac{3}{10}] = [18.8\ ;\ 19.8]$$
2. Ici, on doit consulter à nouveau la table de la loi normale. Il ne faut pas oublier que du fait de la symétrie de la loi normale, on cherche la valeur telle que $1 - \frac{\alpha}{2}$ soit égal au risque. Pour $\alpha = 0.05$, on trouve la valeur 1.96 (certainement la valeur la plus usitée en statistiques), pour $\alpha = 0.01$ on trouve 2.5758. On réécrit ensuite la formule de l'exemple en remplaçant 1.6449 par ces valeurs, et on obtient:
$$IC_{5\%} = [17.83\ ;\ 20.77]$$
et
$$IC_{1\%} = [17.37\ ;\ 21.23]$$

Exercice 3.4–

1. La taille de l'échantillon est suffisament grande (¿30), le pourcentage observé p_0 suit approximativement une loi de Gauss de moyenne $p = 0.12$ et de variance $\sigma = \frac{p(1-p)}{n} = 0.001056$. Les conditions $np \geq 5$ et $n(1-p) \geq 5$ sont bien vérifiées.
La valeur lue dans la table pour $1 - \alpha = 0.05$ est 1.96, l'IF est donc défini par
$$\frac{|p_0 - p|}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96$$
soit $|p_0 - 0.12| \leq 1.96\sqrt{0.001056} = 0.064$. L'IF est donc
$$0.06 \leq p_0 \leq 0.18$$

2. La raisonnement est le même, si ce n'est que l'on remplace ici 1.96 par la valeur correspondant à $1-\alpha$=0.01, *i.e.* 2.576, et on obtient l'IF:

$$IF_{99\%} = [0.04\ ;\ 0.2]$$

3. L'IF est plus large lorsque le risque est plus petit. En effet, en choisissant un niveau de risque plus faible, on doit englober plus de valeurs, de manière à être plus sur de ne pas se tromper.

La forme de l'intervalle de confiance pour la moyenne d'une loi normale est très utilisé, nous le rappelons donc ici sous forme de résultat.

Théorème 3.1 *Soit $X_1,\ldots,X_n$ n v.a. i.i.d de loi $\mathcal{N}(\mu,\sigma^2)$. Supposons que la variance σ^2 soit connue. Alors l'intervalle de confiance de μ au niveau $1-\alpha$ est donné par:*

$$IC_\alpha = [\bar{x} - \epsilon_\alpha \frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + \epsilon_\alpha \frac{\sigma}{\sqrt{n}}] \tag{3.5}$$

où la valeur ϵ_α est lue sur la table de la loi normale centrée réduite, en cherchant la valeur telle que $P(\mathcal{N}(0,1) > \epsilon_\alpha) = 1 - \frac{\alpha}{2}$.

4. Tests d'hypothèses

Revenons à l'exemple introductif du chapitre précédent. L'expérimentateur a estimé la taille des bactéries de sa culture. La question qu'il peut se poser à présent est: les bactéries de la culture ont-elles une taille significativement différente de celle connue dans la littérature? Ou encore: disposant de deux cultures différentes, chacune correspondant à un milieu nutritif différent, peut-on conclure que la taille moyenne des bactéries dans les deux cultures est différente? C'est à ce genre de questions que répond la théorie des tests.

Les mathématiques nécessaires à la construction effective des tests d'hypothèses sont ardues. Nous nous contenterons donc ici de donner des procédures de test, sans donner leur justification théorique. La littérature regorge par ailleurs de tests divers et variés, adaptés à de très nombreuses situations. De même, de nombreux tests sont implémentés dans les logiciels de statistiques. Mais leur utilisation "en aveugle", sans aucune compréhension des processus sous-jacents, n'est pas une bonne chose, puisqu'elle conduira très souvent à des erreurs, aussi bien dans les conditions d'utilisation que dans l'interpretation des résultats. C'est pourquoi nous nous bornerons ici aux tests les plus simples.

4.1 Définitions

Nous supposons une v.a. X, de loi de probabilité $\mathcal{L}(\theta)$ (*i.e.* dont la loi a pour paramètres $\theta \in \Theta$).

Test d'hypothèse – Réaliser un test, c'est décider laquelle des deux hypothèses suivantes:

$$\begin{array}{l} \mathbf{H_0} : \ \theta \in \Theta_0 \\ \mathbf{H_1} : \ \theta \in \Theta_1 \end{array} \tag{4.1}$$

est vérifiée, où $\Theta_0 \cap \Theta_1 = \emptyset$ et $\Theta_0 \cup \Theta_1 = \Theta$.

Remarque– Le rle de l'hypothèse $\mathbf{H_0}$ et de l'hypthèse $\mathbf{H_1}$ ne sont pas symétriques. Le choix de $\mathbf{H_0}$ et $\mathbf{H_1}$ est donc important.

Hypothèse nulle – On appelle *hypothèse nulle*, l'hypothèse $\mathbf{H_0}$. Elle correspond à ce qui est naturellement admis, concernant les valeurs des paramètres.

Hypothèse alternative – On appelle *hypothèse alternative*, l'hypothèse $\mathbf{H_1}$. Elle correspond à ce que l'on cherche à mettre en valeur grâce au test.

De mme, l'interprétation du résultat d'un test n'est pas le mme selon l'hypothèse que l'on considère:

$\mathbf{H_0}$ Accepter cette hypothèse ne "mène à rien". Si le test conduit à accepter $\mathbf{H_0}$, on dit qu'il est *non significatif*.

$\mathbf{H_1}$ Si le test conduit à rejeter $\mathbf{H_0}$ et donc à accepter $\mathbf{H_1}$, on dit qu'il est *significatif*.

Exemple – Si l'expérimentateur constate (intuitivement) que ses bactéries semblent plus grandes que toutes celles de la mme espèce qu'il a cultivé jusqu'à présent, il formulera son test de la faon suivante:

$\mathbf{H_0}$: les bactéries de cet échantillon ont la taille habituelle.
$\mathbf{H_1}$: les bactéries de cet échantillon sont plus grandes que d'habitude.

Il procède à un test d'hypothèse. Si ce test le conduit à accepter $\mathbf{H_0}$, le test est non significatif. La seule conclusion à laquelle il puisse parvenir est que, malgré ce qu'il lui semblait, il n'a pu mettre en évidence une différence entre la taille des bactéries de sa culture et celles qu'il a cultivé auparavant. Si, par contre, le test le conduit à rejeter $\mathbf{H_0}$, et donc à accepter $\mathbf{H_1}$, il peut conclure à une différence significative des tailles.

La notion d'hypothèse nulle et d'hypothèse alternative est assez difficile à comprendre, mais est fondamentale. Le paragraphe suivant [1] explique la raison d'un tel choix.

> **Le raisonnement statistique déductif, une réduction à l'absurde**
>
> La faon dont un statisticien procède pour répondre à la sorte de question considérée dans la section précédente peut paratre inutilement compliquée à première vue: pourquoi, en effet, commencer par poser l'hypothèse principale [nulle] $\mathbf{H_0}$ alors que c'est en fait l'hypothèse contraire [alternative] $\mathbf{H_1}$ qui correspond à ce que l'on souponne, à ce qu'on se demande, à ce qu'on souhaiterait découvrir? - C'est parce que la sorte de raisonnement qu'on se trouve à faire dans ce genre de contexte est une *réduction à l'absurde*.
>
> Ce qui permet de découvrir du nouveau dans la recherche scientifique, c'est l'observation d'une différence qu'on ne pouvait pas prévoir à partir des connaissances antérieures. Toutefois, étant donné la variabilité des phénomènes biologiques, comme la stature, on peut s'attendre à ce qu'une observation particulière diffère sensiblement de la moyenne paramétrique d'une population connue mme si elle provient bien de cette population. On doit donc *passer au crible* les apparences de différences que la réalité nous présente et ne retenir que celles qui sont assez importantes.

1. P. Jolicoeur, *Introduction à la biométrie*, Décarie - Masson, 1991

> C'est pourquoi on commence par poser l'hypothèse principale [nulle], suivant laquelle l'observation proviendrait de la population déjà connue mme si elle diffère de sa moyenne paramétrique. La différence observée $(X - \mu_A)$ serait alors due purement à la variation individuelle. S'il s'avère ensuite que la probabilité d'obtenir une telle observation est trop faible suivant l'hypothèse principale [nulle], on considère cette dernière comme invraisemblable ("absurde") et on la rejette, mais c'est alors la réalité elle-mme, par l'intermédiaire des données analysées, qui motive la conclusion à laquelle on parvient. Il faut noter que, lorsqu'on conserve l'hypothèse principale [nulle], on ne peut pas considérer qu'on l'a confirmée mais uniquement qu'on n'a pas réussi à l'infirmer.

Hypothèse simple – On dit qu'une hypothèse est *simple* si elle correspond à une seule valeur du paramètre θ.

Hypothèse composée – On dit qu'une hypothèse est *composée* (ou *composite*) si elle correspond à un ensemble de valeurs de θ.

Exemple – $\theta = 0.5$ est une hypothèse simple, alors que $\theta < 0.5$, $\theta > 0.5$ et $\theta \neq 0.5$ sont des hypothèses composées.

Test au seuil $1-\alpha$ – Supposons que $\Theta_0 \cap \Theta_1 = \emptyset$ et $\Theta_0 \cup \Theta_1 = \Theta$. Réaliser un test au niveau de risque α (ou au seuil de signification α), c'est décider laquelle parmi les deux hypothèses suivantes est vérifiée:

$$\begin{array}{l} \mathbf{H_0} : \ \theta \in \Theta_0 \\ \mathbf{H_1} : \ \theta \in \Theta_1 \end{array} \tag{4.2}$$

en faisant en sorte que la probabilité d'accepter l'hypothèse $\mathbf{H_0}$ si c'est elle qui est vérifiée, soit égale à $1 - \alpha$, ce que l'on note

$$P(\theta \in \Theta_0 | \mathbf{H_0}) = 1 - \alpha \tag{4.3}$$

ou encore

$$P_{\mathbf{H_0}}(\theta \in \Theta_0) = 1 - \alpha$$

Puissance d'un test – La puissance d'un test, que l'on note $1 - \beta$, indique sa capacité à détecter une différence si elle existe. Elle correspond donc à la probabilité de rejeter $\mathbf{H_0}$ alors qu'elle est fausse. Cette probabilité se note donc

$$P_{\mathbf{H_1}}(\theta \in \Theta_1) = 1 - \beta \tag{4.4}$$

Risque de première espèce – On appelle *risque de première espèce*, ou erreur de type I, la probabilité de rejeter $\mathbf{H_0}$ alors qu'elle est vraie.

Remarque– Il s'agit donc de la probabilité de de conclure que $\theta \in \Theta_1$ sous l'hypothèse $\mathbf{H_0}$, ce qui s'écrit $P_{\mathbf{H_0}}(\theta \in \Theta_1)$. Or dans la définition, nous avons fait l'hypothèse que $\{\Theta_0,\Theta_1\}$ est une partition de Θ. Par conséquent, la probabilité que $P_{\mathbf{H_0}}(\theta \in \Theta_1)$ est le complémentaire dans Θ de $P_{\mathbf{H_0}}(\theta \in \Theta_0)$, soit α.

Risque de deuxième espèce – On appelle *risque de deuxième espèce*, ou erreur de type II, et l'on note en général β, la probabilité d'accepter $\mathbf{H_0}$ alors qu'elle est fausse.

Remarque– De la même manière que l'erreur de type I, l'erreur de type II se définit en fonction de β, par $\beta = P_{\mathbf{H_1}}(\theta \in \Theta_0)$.

Remarque– En général, on spécifie le niveau d'un test, ou bien sa puissance. Les valeurs des erreurs de type I et II découlent ensuite de ce choix.

Ces différentes notions sont résumées dans le tableau suivant.

		Situation réelle	
		$\mathbf{H_0}$ vraie	$\mathbf{H_0}$ fausse
Décision	$\mathbf{H_0}$ acceptée	Conclusion correcte $P = (1-\alpha)$	Erreur de type II $P = \beta$
	$\mathbf{H_0}$ rejetée	Erreur de type I $P = \alpha$ (seuil de signification)	Conclusion correcte $P = (1-\beta)$ (puissance)

4.2 Construction de tests

4.2.1 Conduite d'un test

Pour effectuer un test d'hypothèse, on procède comme suit:

- On commence par décider des hypothèses $\mathbf{H_0}$ et $\mathbf{H_1}$ convenables pour le test.
- De ces hypothèses, on déduit la statistique de test, sous l'hypothèse $\mathbf{H_0}$.
- Sous cette hypothèse, on regarde l'échantillon: obéit-il à la loi que l'on vient de choisir?
- Si c'est le cas, alors on accepte $\mathbf{H_0}$, le test est non significatif.
- Si par contre, l'échantillon ne semble pas "coller" à cette loi, on a mis en évidence une différence: on rejette l'hypothèse $\mathbf{H_0}$. Le test a mis en évidence une différence significative.

Remarque– Dans la plupart des cas, si la taille de l'échantillon est suffisante, on peut utiliser une loi normale, ce qui est possible en vertu du TCL.

4.2.2 Relation avec l'estimation par intervalles

Afin de faire apparaitre les liens qui existent entre les intervalles (de confiance et de fluctuation) et les tests d'hypothèses, considérons le petit exemple suivant.

Soit $X_1, \ldots, X_n$ un n-échantillon de X, v.a. de loi $\mathcal{N}(\mu,\sigma^2)$. On suppose que σ^2 est connue. On estime μ par $\bar{X}$. On veut tester les hypothèses suivantes, au seuil $\alpha = 0.05$.

$$\begin{array}{l} \mathbf{H_0}: \ \mu = \mu_0 \\ \mathbf{H_1}: \ \mu \neq \mu_0 \end{array} \tag{4.5}$$

On se place alors sous l'hypothèse $\mathbf{H_0}$, *i.e.* que l'on suppose que la statistique du test ($\bar{X}$) suit une loi $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$.

Dans ces conditions, si l'on construit un intervalle de confiance au niveau $\alpha = 0.05$ de μ, on va disposer d'une information que l'on pourrait résumer ainsi: "on a 95% de chances que μ (la vraie valeur du paramètre) soit dans l'intervalle $[T_1; T_2]$", où T_1 et T_2 sont donnés par lecture de la table de la loi de Gauss (cf. Théorème 3.1):

$$IC_{5\%} = [\bar{X} - \epsilon_\alpha \frac{\sigma}{\sqrt{n}} \ ; \ \ \bar{X} + \epsilon_\alpha \frac{\sigma}{\sqrt{n}}]$$

où ϵ_α, lu dans la table, vaut 1.96 pour un niveau de 0.05.

Si l'hypothèse $\mathbf{H_0}$ est vérifiée, alors μ_0 doit se trouver dans cet intervalle. Donc le test est significatif si $\mu_0 \notin IC_{5\%}$.

Remarque– Il ne sera pas toujours possible de construire un test de cette manière. Pour utiliser ce raisonnement, il faut en effet que le test soit *bilatéral* (ou *bilatère*), *i.e.* de la forme $\mu_0 \neq \mu$), et non *unilatère* (de la forme $\mu_0 > \mu$ ou $\mu_0 < \mu$).

Remarque– Par analogie avec les intervalles de confiance, on comprend donc que plus le niveau de signification d'un test est faible (plus α est petit), moins grandes sont les chances de rejeter $\mathbf{H_0}$.

4.2.3 Comparaison d'une proportion observée à une proportion théorique

Soit p_0 la proportion théorique (supposée connue) et p_1 la proportion observée sur un échantillon de taille n. On teste les hypothèses suivantes:

$$\begin{array}{l} \mathbf{H_0}: \ \pi = p_0 \\ \mathbf{H_1}: \ \pi \neq p_0 \end{array}$$

Sous l'hypothèse $\mathbf{H_0}$, le nombre d'individus dans l'échantillon présentant le caractère devrait tre $C_1 = np_0$, et le nombre d'individus ne le présentant pas $C_2 = n(1-p_0)$. Soit O_1 et O_2 les nombres de personnes dans l'échantillon présentant et ne présentant pas le caractère, *i.e.* les analogues de C_1 et C_2 en utilisant la proportion observée p_1.

Ainsi, on a les quantités suivantes, si "Caractère" correspond aux individus présentant le caractère et "!Caractère" correspond aux individus ne présentant pas le caractère:

	Sous $\mathbf{H_0}$ (théorie)	Observation
Caractère	$C_1 = np_0$	$O_1 = np_1$
!Caractère	$C_2 = n(1-p_0)$	$O_2 = n(1-p_1)$

Statistique du test

La quantité

$$\chi^2_{Obs} = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2}$$

suit une loi du χ^2 à 1 degré de liberté (d.d.l.).

Règle d'acceptation

On accepte $\mathbf{H_0}$ si la probabilité pour cette quantité d'tre supérieure à sa valeur résultant de l'observation est inférieure au risque choisi, *i.e.* si

$$\chi^2_{Obs} < \epsilon_\alpha$$

où ϵ_α est lue sur l'une des deux tables de la loi du chi-deux fournies, en cherchant $P(\chi^2_1 > \epsilon_\alpha) = \alpha$. Autrement dit:

- Si $\chi^2_{Obs} \geq \epsilon_\alpha$ alors on rejette $\mathbf{H_0}$ et donc on accepte $\mathbf{H_1}$.
- Si $\chi^2_{Obs} < \epsilon_\alpha$ alors on ne rejette pas $\mathbf{H_0}$ et donc le test est non significatif.

Conditions d'application

Ce test n'est valide que si les quantités observées sont supérieures à 5.

4.2.4 Comparaison de deux proportions observées

Soit π_1 et π_2 deux proportions les proportions théoriques d'un caractère dans deux populations. On observe deux échantillons de tailles respectives n_1 et n_2. On veut tester les hypothèses suivantes:

$$\mathbf{H_0}: \ \pi_1 = \pi_2$$
$$\mathbf{H_1}: \ \pi_1 \neq \pi_2$$

Cas des petits échantillons

Dans le cas de petits échantillons, on appliquera un test d'indépendance, développé dans la section 4.2.7, page 50. On dressera alors un tableau avec deux entrées en ligne et deux entrées en colonnes, et la statistique suivra une loi du χ^2 à 1 d.d.l. Pour appliquer ce test, il faut que $n_i p_i \geq 10$ et $n_i(1-p_i) \geq 10$ pour $i = 1{,}2$.

Statistique du test, grands échantillons

Soit p_1 et p_2 les estimateurs respectifs de π_1 et π_2, alors Z est la statistique de notre test:

$$Z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

où

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Sous l'hypothèse $\mathbf{H_0}$, Z suit une loi normale centrée réduite et s'écrit:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Règle d'acceptation, grands échantillons

- Si $\mid z \mid \geq \epsilon_\alpha$, rejet de $\mathbf{H_0}$.
- Si $\mid z \mid < \epsilon_\alpha$, test non significatif.

où la valeur ϵ_α est ici (grands échantillons) lue sur la table de la $\mathcal{N}(0,1)$.

4.2.5 Comparaison d'une moyenne observée à une moyenne théorique

Il s'agit ici du test que nous avons introduit lorsque nous avons fait le lien entre tests et intervalles de confiance.

4.2.6 Comparaison de deux moyennes observées

Soient $\bar{x}_1$ et $\bar{x}_2$ les moyennes observées sur deux échantillons d'effectifs respectifs n_1 et n_2, et de variance respectives s_{11}^2 et s_{12}^2. On veut tester les hypothèses suivantes:

$$\mathbf{H_0}: \ \mu_1 = \mu_2$$
$$\mathbf{H_1}: \ \mu_1 \neq \mu_2$$

Statistique du test, grands échantillons

Si les deux échantillons sont grands, on peut considérer que sous $\mathbf{H_0}$, $\bar{X}_1 - \bar{X}_2$ suit une $\mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, que l'on approximera par $\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}$ (avec s l'estimateur non biaisé de la variance). En d'autres termes, la statistique est Z, définie par:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

Sous $\mathbf{H_0}$, Z s'exprime de la manière suivante:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

et Z suit alors une loi Normale centrée réduite.

Règle de décision, grands échantillons

- Si $\mid z \mid \geq \epsilon_\alpha$ alors on rejette $\mathbf{H_0}$
- Si $\mid z \mid < \epsilon_\alpha$ alors on ne rejette pas $\mathbf{H_0}$, le test est non significatif.

où ϵ_α est lu sur la table de la $\mathcal{N}(0,1)$.

Par intervalles de confiance

On calcule l'IC au niveau $1-\alpha$. On accepte $\mathbf{H_0}$ si $\mu_1 - \mu_2 = 0 \in IC_\alpha$, on la rejette sinon.

4.2.7 Test d'indépendance de deux populations

Soient X et Y deux v.a. qualitatives (possédant un nombre fini de valeurs). On veut tester

$\mathbf{H_0}$: X et Y indépendantes.
$\mathbf{H_1}$: X et Y non indépendantes.

On fait alors n observations et on indique dans chaque case (a_i,b_j) du tableau suivant l'effectif n_{ij} des observations où a_i (les p valeurs possibles de X) et b_j (les q valeurs possibles de Y) sont observés en même temps.

Dans ce tableau, on fait aussi apparaître les sommes *marginales* (notées $n_{i,\bullet}$ et $n_{\bullet,j}$).

	b_1	$\cdots$	b_j	$\cdots$	b_q	
a_1	$n_{1,1}$		$n_{1,j}$		$n_{1,q}$	$n_{1,\bullet}$
$\vdots$						
a_i	$n_{i,1}$		$n_{i,j}$		$n_{i,q}$	$n_{i,\bullet}$
$\vdots$						
a_p	$n_{p,1}$		$n_{p,j}$		$n_{p,q}$	$n_{p,\bullet}$
	$n_{\bullet,1}$		$n_{\bullet,j}$		$n_{\bullet,q}$	n

On construit alors le tableau analogue au précédent, mais correspondant aux effectifs *espérés* sous $\mathbf{H_0}$. Ceci se fait en faisant le produit des marges du tableau des observés, et en divisant par l'effectif total, comme cela est schématisé dans le tableau (abrégé) suivant:

	$\cdots$	b_j	$\cdots$	
$\vdots$				
a_i		$e_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n}$		$n_{i,\bullet}$
$\vdots$				
		$n_{\bullet,j}$		n

Statistique du test

La statistique

$$D_n = \sum_{i=1}^{i=p}\sum_{j=1}^{j=q} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}} = \sum_{i=1}^{i=p}\sum_{j=1}^{j=q} \frac{n_{i,j}^2}{e_{i,j}} - n$$

suit une loi du chi-deux à $(p-1)(q-1)$ d.d.l.

Règle d'acceptation

On calcule d_n, valeur numérique de D_n dans la réalisation considérée. On rejette ensuite $\mathbf{H_0}$ si $d_n > \epsilon_\alpha$, où ϵ_α est lue sur la table du χ^2, en cherchant $P(\chi^2_{(p-1)(q-1)} > \epsilon_\alpha) = \alpha$. Pour résumer, en fonction du ϵ_α trouvé:

- Si $d_n \leq \epsilon_\alpha$ on ne peut pas rejeter $\mathbf{H_0}$. Le test est non significatif.
- Si $d_n > \epsilon_\alpha$, on rejette $\mathbf{H_0}$, on accepte donc $\mathbf{H_1}$ au niveau de risque α. Le test conduit à rejeter l'hypothèse d'indépendance de X et Y.

Remarque– Si d_n est petit, c'est que l'écart entre les valeurs observées et les valeurs théoriques (sous l'hypothèse d'indépendance) est petite. Par ailleurs, plus la taille du tableau augmente, plus les fluctuations possibles théoriquement (sous l'hypothèse d'indépendance) augmentent, ce que l'on constate en lisant une table du χ^2 pour un α donné et pour des d.d.l. croissants.

Exemple – Un examen a eu lieu en trois sessions en parallèle. Les copies ont été notées par trois enseignants différents i = 1,2,3, avec cinq notes possibles j = 1,2,3,4,5. On a obtenu le tableau suivant (ou les effectifs marginaux ont été calculés):

	1	2	3	4	5	
1	19	26	35	18	4	$n_{1,\bullet} = 102$
2	14	28	36	21	5	$n_{2,\bullet} = 104$
3	8	26	38	22	7	$n_{3,\bullet} = 101$
	$n_{\bullet,1} = 41$	$n_{\bullet,2} = 80$	$n_{\bullet,3} = 109$	$n_{\bullet,4} = 61$	$n_{\bullet,5} = 16$	$n = 307$

Peut-on dire qu'il y a un lien entre l'enseignant et la note?

On va donc tester les hypothèses suivantes:

$$\begin{aligned} \mathbf{H_0}: &\ X \text{ est indépendante de } Y \\ \mathbf{H_1}: &\ X \text{ n'est pas indépendante de } Y \end{aligned}$$

On commence par etablir le tableau des effectifs espérés sous $\mathbf{H_0}$. Pour cela, on recopie les marges du tableau des effectifs observés, et on remplit chaque case intérieure par le produit des marges divisé par le nombre total.

	1	2	3	4	5	
1	13.62	26.58	36.21	20.27	5.32	$n_{1,\bullet} = 102$
2	13.89	27.1	36.93	20.66	5.42	$n_{2,\bullet} = 104$
3	13.49	26.32	35.86	20.07	5.26	$n_{3,\bullet} = 101$
	$n_{\bullet,1} = 41$	$n_{\bullet,2} = 80$	$n_{\bullet,3} = 109$	$n_{\bullet,4} = 61$	$n_{\bullet,5} = 16$	$n = 307$

Maintenant, on fait la somme ligne par ligne (ou colonne par colonne) des quotients des nombres observés au carré par les nombres espérés:

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{5}\frac{n_{i,j}^2}{e_{i,j}} &= \frac{19^2}{13.62}+\frac{26^2}{26.58}+\frac{35^2}{36.21}+\frac{18^2}{20.27}+\frac{4^2}{5.32}+\frac{14^2}{13.89}+\frac{28^2}{27.1}\\ &\quad+\frac{36^2}{36.93}+\frac{21^2}{20.66}+\frac{5^2}{5.42}+\frac{8^2}{13.49}+\frac{26^2}{26.32}+\frac{38^2}{35.86}+\frac{22^2}{20.07}\\ &\quad+\frac{7^2}{5.26}\\ &= 26.51+25.43+33.83+15.98+3.0+14.11+28.93+35.09\\ &\quad+21.35+4.61+4.74+25.68+40.27+24.12+9.32\\ &= 312.97\end{aligned}$$

Par conséquent, la statistique $D_n = \sum_{i=1}^{i=p}\sum_{j=1}^{j=q}\frac{n_{i,j}^2}{e_{i,j}} - n$ prend pour cette réalisation la valeur $d_n = 312.97 - 307 = 5.97$.

On regarde maintenant la table de la loi du χ^2, pour $\alpha = 0.05$ et à $(p-1)(q-1) = 8$ d.d.l. On y lit

$$P(\chi_8^2 > \epsilon_\alpha) = \alpha = 0.05 \Leftrightarrow \epsilon_\alpha = 15.5$$

Par conséquent, puisque $d_n = 5.97 \not> 15.5$, on ne peut pas rejeter l'hypothèse $\mathbf{H_0}$ selon laquelle la correction est indépendante du correcteur.

4.3 Exercices

Exercice 4.1 *Soit X une v.a. et x une réalisation de X. X peut suivre deux lois dont les densités de probabilité sont:*

$$f_1(x) = \left| \begin{array}{l} 1 \text{ pour } x \in [0;1] \\ 0 \text{ sinon} \end{array} \right. \qquad f_2(x) = \left| \begin{array}{l} 0 \text{ pour } x < 0 \\ 0.5 \text{ pour } x \in [0;0.5] \\ 1.5 \text{ pour } x \in [0.5;1] \\ 0 \text{ pour } x > 1 \end{array} \right.$$

On considère le test suivant:

$$\begin{array}{l} \mathbf{H_0}: X \text{ est distribuée selon la loi 1.} \\ \mathbf{H_1}: X \text{ est distribuée selon la loi 2.} \end{array}$$

On définit la région de rejet de $\mathbf{H_0}$ par $\{x \geq 0.5\}$. Calculer:

1. *Le risque de première espèce.*
2. *Le risque de deuxième espèce.*
3. *La puissance du test.*

Exercice 4.2 *Sur* 4000 *naissances, on relève* 2065 *garons. Tester l'hypothèse selon laquelle la probabilité d'un garon à la naissance est $\frac{1}{2}$, avec le seuil* 0.05 *puis* 0.01.

Exercice 4.3 *On veut tester le fait que dans une certaine population, il y a 50% d'hommes et 50% de femmes. On prend un échantillon de 10000 personnes, et l'on trouve 48.8% d'hommes. Réaliser ce test en utilisant les intervalles de confiance.*

Exercice 4.4 *Après correction d'un examen, on a tiré au hasard deux échantillons de 50 étudiants. Le premier groupe a suivi des séances de TD le matin (am), le second l'après-midi (pm). Les résultats dans chacun des deux groupes sont les suivants:*

Groupe	*Recalés*	m	s_1^2
am	*18*	*12*	*16.7*
pm	*8*	*11.2*	*4.5*

1. *Tester l'hypothèse selon laquelle le pourcentage de recalés est différent entre les deux groupes.*
2. *Tester la différence (ou non différence) des moyennes, au risque 5%.*
3. *Les conclusion changent-elles si le risque est différent?*

Exercice 4.5 *La proportion de naissances prématurées est habituellement de 5%. Sur 170 accouchements de femmes de plus de 35 ans, on a observé 16 naissances prématurées. En utilisant deux méthodes différentes, décider si cette proportion est compatible avec la proportion habituelle, au niveau* $1 - \alpha = 0.95$.

4.4 Correction des exercices

Exercice 4.1– Se référer au tableau des types de risques.

1. Le risque de première espèce est la probabilité de rejeter $\mathbf{H_0}$ sachant que c'est $\mathbf{H_0}$ qui est vraie. Par conséquent, ce risque est donné par

$$P(\text{rejet de } \mathbf{H_0}|\mathbf{H_0})$$

Nous avons défini la région de rejet de $\mathbf{H_0}$ par $\{x \geq 0.5\}$. On cherche donc

$$P(X \geq 0.5|\mathbf{H_0})$$

Or, sous l'hypothèse $\mathbf{H_0}$, X a la densité $f_1(x)$. La probabilité pour que $X \geq 0.5$ est donc:

$$P(X \geq 0.5) = 1 - P(X < 0.5) = 1 - \int_0^{0.5} dt = 1 - [t]_0^{0.5} = 0.5$$

La probabilité d'une erreur de type I est donc de 0.5.

2. Le risque de deuxième espèce est la probabilité de ne pas rejeter $\mathbf{H_0}$ alors que l'on devrait (*i.e.* alors que $\mathbf{H_1}$ est vraie). Il nous faut donc calculer

$$P(\text{non rejet de } \mathbf{H_0}|\mathbf{H_1})$$

ce qui s'exprime aussi

$$P(X < 0.5|\mathbf{H_1})$$

Si on est sous l'hypothèse $\mathbf{H_1}$, c'est que la densité de X est donnée par $f_2(x)$. On calcule donc $P(X < 0.5)$ en utilisant f_2:

$$P(X < 0.5) = \int_0^{0.5} 0.5dt = [0.5t]_0^{0.5} = 0.25$$

La probabilité d'une erreur de type II est donc de 0.25.

3. La puissance d'un test est la probabilité de rejeter $\mathbf{H_0}$ alors que $\mathbf{H_1}$ est vraie. On doit donc calculer

$$P(\text{rejet de } \mathbf{H_0}|\mathbf{H_1})$$

On est donc sous $\mathbf{H_1}$, et la densité de X est donnée par f_2. Quelle est alors la probabilité de rejeter $\mathbf{H_0}$, *i.e.* la probabilité que $X \geq 0.5$? On calcule:

$$P(X \geq 0.5) = 1 - P(X < 0.5) = 1 - 0.25 = 0.75$$

Exercice 4.2– Soit p_0 la proportion théorique de garçcons, on définit les hypothèses suivantes:

$$\mathbf{H_0} : \pi = p_0 = \tfrac{1}{2}$$
$$\mathbf{H_1} : \pi \neq p_0 = \tfrac{1}{2}$$

Soit $p_1 = \frac{2065}{4000}$ la proportion observée de garcon sur un échantillon de taille $n = 4000$.

Sous l'hypothèse H_0, le nombre d'individus dans l'échantillon présentant le caractère (garçon) devrait être $C_1 = np_0$, et le nombre d'individus ne le présentant pas $C_2 = n(1 - p_0)$.

Soient O_1 et O_2 les nombres de personnes dans l'échantillon présentant et ne présentant pas le caractère, *i.e.* les analogues de C_1 et C_2 en utilisant la proportion observée p_1.

	Sous $\mathbf{H_0}$ (théorie)	Observation
Garcon	$C_1 = np_0 = \frac{4000}{2}$	$O_1 = np_1 = 2065$
Fille	$C_2 = n(1 - p_0) = \frac{4000}{2}$	$O_2 = n(1 - p_1) = 1935$

Statistique du test:

$$\chi^2_{Obs} = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2}$$

suit une loi du χ^2 à 1 degré de liberté (d.d.l.).

La réalisation de notre statistique est:

$$\chi^2_{Obs} = \frac{(2065 - 2000)^2}{2000} + \frac{(1935 - 2000)^2}{2000} = 4.22$$

Règle d'acceptation:
On accepte $\mathbf{H_0}$ si la probabilité pour cette quantité d'tre supérieure à sa valeur résultant de l'observation est inférieure au risque choisi, *i.e.* si

$$\chi^2_{Obs} < \epsilon_\alpha$$

où $P(\chi^2_1 \geq \epsilon_\alpha) = 1 - \alpha$ est lu sur l'une des deux tables de la loi du chi-deux fournies. Autrement dit:

- Si $\chi^2_{Obs} \geq \epsilon_\alpha$ alors on rejette $\mathbf{H_0}$ et donc on accepte $\mathbf{H_1}$.
- Si $\chi^2_{Obs} < \epsilon_\alpha$ alors on ne rejette pas $\mathbf{H_0}$ et donc le test est non significatif.

Réalisation du test:

- Pour $\alpha = 0.05$, $\epsilon_\alpha = 3.84$: $4.22 \geq 3.84$ alors $\mathbf{H_0}$ est rejettée (et donc $\mathbf{H_1}$ est acceptée). Le test est significatif, on peut conclure qu'au niveau de risque 5%, la proportion de garçons et de filles est différente.
- Pour $\alpha = 0.01$, $\epsilon_\alpha = 6.63$: $4.22 \ngeq 6.63$ alors le test est non significatif. On ne peut donc pas conclure à une différence de proportions.

Exercice 4.3– On va utiliser l'approximation d'une loi binomiale par une loi normale, que nous rappelons dans l'encadré suivant.

Soit X suivant une loi binomiale de paramètres (n,p). Alors si n est suffisament grand, $X \rightsquigarrow \mathcal{N}(np,np(1-p))$.

– On va réaliser le test suivant:

$\mathbf{H_0}$: La proportion d'hommes dans la population est $\pi = 0.5$
$\mathbf{H_1}$: La proportion d'hommes dans la population est $\pi \neq 0.5$

– L'estimation de π sur notre échantillon de taille $n = 10000$ est $p_1 = 0.488$.
– L'IC à 95% de π (proportion d'hommes) est donné par

$$IC_{95\%}(\pi) = [p_1 - 1.96\sqrt{\frac{p_1(1-p_1)}{n}} \; ; \; p_1 + 1.96\sqrt{\frac{p_1(1-p_1)}{n}}]$$

soit

$$[0.478 \; ; \; 0.498]$$

– $p_0 = \frac{1}{2} \notin IC_{95\%}(\pi)$ donc l'hypothèse $\mathbf{H_0}$ est rejetée et l'hypothèse $\mathbf{H_1}$ est acceptée.
– Au risque $\alpha = 0.01$, $p_0 \in IC_{99\%}(\pi) = [0.475 \; ; \; 0.501]$ donc le test est non significatif.

Exercice 4.4–

1. Ici, il s'agit d'un test de comparaison de proportions observées. On veut tester

$$\mathbf{H_0} : \pi_{am} = \pi_{pm}$$
$$\mathbf{H_1} : \pi_{am} \neq \pi_{pm}$$

Statistique du test:
La statistique de notre test est:

$$Z = \frac{p_{am} - p_{pm} - (\pi_{am} - \pi_{pm})}{\sqrt{p(1-p)(\frac{1}{n_{am}} + \frac{1}{n_m})}}$$

où

$$p = \frac{n_{am}p_{am} + n_{pm}p_{pm}}{n_{am} + n_{pm}}$$

Sous l'hypothèse $\mathbf{H_0}$, Z suit une loi normale centrée réduite et s'écrit:

$$Z = \frac{p_{am} - p_{pm}}{\sqrt{p(1-p)(\frac{1}{n_{am}} + \frac{1}{n_m})}}$$

Maintenant, on calcule les proportions observées:

$$p_{am} = \frac{18}{50} = 0.36 \quad p_{pm} = \frac{8}{50} = 0.16$$

et

$$p = \frac{n_{am}p_{am} + n_{pm}p_{pm}}{n_{am} + n_{pm}} = \frac{18+8}{100} = 0.26$$

on trouve que $|z| = 0.2/0.08 = 2.5$.
Règles d'acceptation:
– Si $\mid z \mid \geq \epsilon_\alpha$, rejet de $\mathbf{H_0}$.

– Si $\mid z \mid < \epsilon_\alpha$, test non significatif.

Réalisation du test:

– Pour le risque $\alpha = 0.05$: $|z| = 2.5 \geq 1.96$ donc on rejette H_0.
– Pour le risque $\alpha = 0.01$: $|z| = 2.5 \ngeq 2.57$ donc on ne rejette pas H_0, le test est non significatif.

2. Ici, il s'agit d'un test de comparaison de moyennes observées:

$$\mathbf{H_0} : \mu_{am} = \mu_{pm}$$
$$\mathbf{H_1} : \mu_{am} \neq \mu_{pm}$$

Statistique du test

$$Z = \frac{\bar{X}_{am} - \bar{X}_{pm} - (\mu_{am} - \mu_{pm})}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

Sous H_0 (si n_1 et n_2 sont grands), on a

$$Z = \frac{\bar{X}_{am} - \bar{X}_{pm}}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

alors Z suit une loi Normale centrée réduite.

Régle d'acceptation:

– Si $\mid z \mid \geq \epsilon_\alpha$ alors on rejette $\mathbf{H_0}$
– Si $\mid z \mid < \epsilon_\alpha$ alors on ne rejette pas $\mathbf{H_0}$, le test est non significatif.

Réalisation du test:

Calcul de la réalisation z de Z sous $\mathbf{H_0}$:

$$z = \frac{12 - 11.2}{\sqrt{\frac{4.5}{50} + \frac{16.7}{50}}} = 1.23$$

Donc pour le risque $\alpha = 0.05$: $1.23 \ngeq 1.96$. Par conséquent, le test est non significatif.

3. Bien sur. On a déja vu dans la première partie que le test conduisait à rejeter $\mathbf{H_0}$ à $\alpha = 0.05$, alors qu'à $\alpha = 0.01$ il était non significatif.
Reprenons le deuxième test. On a vu qu'au niveau $\alpha = 0.05$, le test est non significatif. On va chercher dans la table de la loi normale la première valeur permettant de rejeter $\mathbf{H_0}$. Pour cela, il faut que le z que nous avons calculé, 1.23, soit supérieur ou égal à $\epsilon_{\frac{\alpha}{2}}$ (puisqu'on est en train de lire $|z|$, on utilise donc encore la symétrie de la loi normale).
La lecture de la table de la page 300 de *Statistiques descriptives et décisionnelles*, à la ligne 1.2 et la colonne 0.03 donne 0.89065. Si l'on lit l'autre table, on cherche la première valeur supérieure à 1.23, soit 1.2319 dans la ligne 0.89, colonne 0.001 (table lue dans le "sens de Q").
Ainsi, à tout niveau de risque supérieur à (1-0.89)/2=0.065 (approximativement), on va rejeter $\mathbf{H_0}$, et le test sera significatif.

Exercice 4.5– Dans les deux cas, on veut tester les hypothèses:

$\mathbf{H_0}$: La proportion de naissances prématurées est $\pi = 0.05$
$\mathbf{H_1}$: La proportion de naissances prématurées est $\pi \neq 0.05$

– On va construire un IF au niveau 95% du nombre d'accouchements prématurés observés:

$$IF_{95\%} = [170(0.05 - 1.96\sqrt{\frac{0.95 \times 0.05}{170}}) \ ; \ 170(0.05 + 1.96\sqrt{\frac{0.95 \times 0.05}{170}})]$$

soit $[2.93 \ ; \ 14.07] \simeq [3 \ ; \ 14]$. Par conséquent, puisque 16 n'appartient pas à cet intervalle, on doit rejeter $\mathbf{H_0}$.
Au niveau 95%, la proportion de naissances prématurées dans l'échantillon considéré n'est pas égal à 5%.

– On va utiliser le test de comparaison d'une moyenne observée à une moyenne théorique. On construit donc le tableau suivant:

	Prématurés	A terme	Total
Observés	16	154	170
Théoriques	$170 \times 0.05 = 8.5$	$170 \times 0.95 = 161.5$	170

Maintenant, calculons χ^2_O:

$$\chi^2_O = \frac{(16 - 8.5)^2}{8.5} + \frac{(161.5 - 154)^2}{161.5} = 6.97$$

Lisons maintenant la valeur de la loi du Chi-deux à $\nu = 1$ d.d.l., au niveau 5%. Elle vaut 3.84. Par conséquent, puisque $\chi^2_O > 3.841$, on rejette $\mathbf{H_0}$.

5. Méthodes non paramétriques

Dans tout ce qui a précédé, nous avons appliqué des méthodes dites *paramétriques*. En effet, nous avons toujours supposé que les processus étudiés suivaient des lois de probabilité bien définies (binomiales, normales, etc.), dont nous cherchions à connaître les paramètres. Les seuls composantes inconnues des distributions de probabilité sous-jacentes à une réalisation sont supposées être les valeurs vraies des paramètres. Toutefois, ce genre d'hypothèse peut tre contraignant.

Il existe des méthodes qui contournent ce problème. Ces méthodes sont dites *non paramétriques*, puisqu'à la différence des précédentes, elles ne font aucune hypothèse sur les processus sous-jacents.

Les méthodes que nous introduirons ici portent sur la comparaison de deux échantillons, issus de deux populations dont on cherche à savoir si elles sont ou non différentes (on parle d'*inférence portant sur deux populations*).

5.1 Test de Wilcoxon (observations non couplées)

Soient F et G les fonctions de répartition de deux variables indépendantes X et Y. On considère deux échantillons: $X_1, \ldots, X_n$ de X et $Y_1, \ldots, Y_m$ de Y. On ne suppose rien *a priori* sur F et G. On définit de manière intuitive le test suivant:

$\mathbf{H_0}$: $F = G$.
$\mathbf{H_1}$: Les Ysont "plus grandes" en général que les X.

Il nous faut définir une "mesure" de la grandeur relative. On utilisera donc la notion suivante. On dira que Y est *stochastiquement* plus grande que X ssi:

$$\forall \xi \in \mathbb{R},\ P(Y \geq \xi) \geq P(X \geq \xi)$$

Ce qui est équivalent à:

$$\forall \xi \in \mathbb{R},\ G(\xi) \geq F(\xi)$$

On observe une réalisation $x_1, \ldots, x_n$ de $X_1, \ldots, X_n$ et une réalisation $y_1, \ldots, y_m$ de $Y_1, \ldots, Y_m$.

Statistique du test

On classe les valeurs observées $x_1,\ldots,x_n,y_1,\ldots,y_m$ par ordre croissant de valeurs. Puis on omet les indices, ne retenant que le fait qu'une valeur provient de la réalisation de X ou de Y. La statistique de Wilcoxon est la somme T des rangs occupés par les lettres X.

Exemple – Supposons que l'on ait observé les données suivantes:
- Réalisation de l'échantillon issu du X: 5, 1.2, 6.4, 11.2.
- Réalisation de l'échantillon issu du Y: 8.5, 7.1, 1.5, 11.4, 11.6.

On classe alors les données par ordre croissant, en indiquant dans un premier temps leur position dans leur réalisations respectives:

$$\omega = x_2 y_3 x_1 x_3 y_2 y_1 x_4 y_4 y_5$$

Ensuite, on réécrit ω en "oubliant" les indices:

$$\omega = xyxxyyxyy$$

Ici la réalisation de la statistique de Wilcoxon est donc donnée par la somme des positions de x dans ω, soit $t = 1 + 3 + 4 + 7 = 15$.

La loi de T lorsque $\mathbf{H_0}$ est vérifiée est obtenue de la façon suivante:
- Le nombre des épreuves ω possibles est

$$\left(\begin{array}{c} m+n \\ n \end{array} \right) = C^n_{m+n} = \frac{(m+n)!}{n!(m+n-n)!} = \frac{(m+n)!}{n!m!}$$

- La probabilité $P(T \leq t)$ (où t est la valeur obtenue sur la réalisation considérée) est donnée par:

$$P(T \leq t) = \frac{\#\{\text{cas favorables à } T \leq t\}}{C^m_{n+m}}$$

Lorsque cela est possible, on dénombrera ces cas favorables. Sinon, on utilisera l'approximation normale définie par la suite (grands échantillons).

Remarque– Soient $n \in \mathbb{N}$ et $p < n$, $p \in \mathbb{N}$. On rappelle que la quantité

$$C^p_n = \frac{n!}{p!(n-p)!}$$

définit le nombre de *combinaisons* possibles de p éléments indifférenciés dans n cases. Cette quantité vérifie notamment les propriétés suivantes:
- $C^0_n = C^n_n = 1$.
- $C^p_n = C^{n-p}_n$.

Exemple – Dans le cas de l'exemple précédent, on cherche donc toutes les combinaisons possibles de 4 lettres X et 5 lettres X telles que la somme des positions de X dans le ω résultant soit inférieure ou égale à 15. Par exemple, $\omega = xxxxyyyyy$ donne $T = 1 + 2 + 3 + 4 = 10$ qui est un cas favorable.

Règle de rejet

Le test est de la forme:

$$\text{Rejet de } \mathbf{H_0} \Leftrightarrow P(T \leq t) < \alpha$$

Statistique du test, grands échantillons

Lorsque m et n sont grands (en pratique, supérieurs à 8), T est à peu près gaussienne, avec:

$$E(T) = \frac{m(m+n+1)}{2} \text{ et } Var(T) = \frac{mn(m+n+1)}{12}.$$

Règle d'acceptation, grands échantillons

On calculera donc

$$P\left(\mathcal{N}\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right) \leq t\right)$$

et on rejetera $\mathbf{H_0}$ si cette quantité est inférieure à α.

5.2 Test de Wilcoxon (observations couplées)

Soient X et Y deux variables aléatoires définies sur une même catégorie d'épreuves telles que $Z = Y - X$ soit une loi continue. On veut tester:

$\mathbf{H_0}$: Z est une variable aléatoire symétrique (même loi que $-Z$).
$\mathbf{H_1}$: Z est stochastiquement supérieure à $-Z$.

Remarque– Ce que l'on cherche donc à savoir, c'est si $P(Z) = P(-Z)$. Si ce n'est pas le cas (*i.e.* si la loi de Z est asymétrique), c'est donc (selon les hypothèses formulées) que $P(Z) \geq P(-Z)$, ce qui veut dire que $P(Y - X) \geq P(X - Y)$, soit $Y \geq X$.

Pour cela, on fait n observations indépendantes $(X_1,Y_1),\ldots,(X_n,Y_n)$ du couple (X,Y), on range les $Z_i = Y_i - X_i$ par ordre de valeurs absolues croissantes, et l'on ne retient que leurs signes. L'épreuve observée est donc une suite de n signes + ou −, par exemple $\omega = (- - - + - + + + +)$.

Soit W la somme des rangs des signes − (ici $W = 1 + 2 + 3 + 5 = 11$).

Statistique du test, petits échantillons

La statistique de W est déterminée de la façon suivante

$$P(W \leq w) = \frac{\#\{\text{cas favorables à } W \leq w\}}{2^n}$$

Règle de rejet

Le test est de la forme:

$$\text{Rejet de } \mathbf{H_0} \Leftrightarrow P(W \leq w) < \alpha$$

Lorsque n n'est pas trop grand, la loi de W sous $\mathbf{H_0}$ s'obtient en comptant les cas favorables.

Statistique du test, grands échantillons

Lorsque n est assez grand, W est à peu près gaussienne, avec:

$$E(W) = \frac{n(n+1)}{4} \text{ et } Var(W) = \frac{n(n+1)(2n+1)}{24}.$$

Règle d'acceptation, grands échantillons

On calculera donc

$$P\left(\mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right) \leq w\right)$$

et on rejetera $\mathbf{H_0}$ si cette quantité est inférieure à α.

5.3 Test des signes (observations couplées)

Soient X et Y deux variables aléatoires définies sur une même catégorie d'épreuves, telles que $P(X = Y) = 0$. On veut tester:

$$\begin{array}{l} \mathbf{H_0}: \ P(Y > X) = P(X > Y) = \frac{1}{2} \\ \mathbf{H_1}: \ P(Y > X) > \frac{1}{2} \end{array}$$

Ce que l'on cherche donc à mettre en évidence, c'est une disymétrie des distributions de X et Y. Plus précisemment, on veut savoir si on a une certaine chance que les mesures issues de Y soient plus grandes que celles issues de X.

Conduite du test

Pour réaliser ce test, on fait n observations indépendantes $(X_1,Y_1),\ldots,(X_n,Y_n)$ du couple (X,Y), et on note V le nombre de couples tels que $Y_i < X_i$. Le test est de la forme:

$$\text{Rejet de } \mathbf{H_0} \Leftrightarrow P(V \leq v) < \alpha$$

Si $\mathbf{H_0}$ est réalisée, la plus grande valeur possible de $P(V \leq v)$ est obtenue en supposant que V obéisse à la loi binomiale $B(n; \frac{1}{2})$.

Remarque– On rappelle (tableau 1.4 page 9) que la distribution d'une loi binomiale $\mathcal{B}(n,p)$ est donnée par:

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

Ici, il nous faudra donc calculer

$$\begin{aligned} P(V \leq v) &= \sum_{i=0}^{v} C_n^i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} \\ &= \sum_{i=0}^{v} C_n^i \left(\frac{1}{2}\right)^n \\ &= \left(\frac{1}{2}\right)^n \sum_{i=0}^{v} C_n^i \end{aligned}$$

Seulement, on ne peut pas raisonnablement aller plus loin. Lorsque v n'est pas trop grand, on pourra faire la somme (assez rapide en faisant des simplifications). Sinon on utilisera donc la table fournie (*Table I. Test des signes*), qui donne une information sur la valeur maximale de v (pour un n donné) pouvant conduire à un rejet de $\mathbf{H_0}$.

Remarque– Le test des signes rend compte de la remarque intuitive suivante: supposons que l'on mesure une caractéristique sur des données appariées (couplées), mais dans des laboratoires différents. Alors les appareils utilisés pour faire les mesures peuvent être étalonnés différemment, et la seule information vraiment significative est le signe de la différence entre un couple de mesures.

5.4 Cas des ex-aequo

Nous ne traiterons pas ici ces cas, plus compliqués. Les hypothèses généralement faites sont que la présence d'ex-aequo signifie l'imprécision des appareils de mesure (les lois étant supposées continues, on ne devrait pas avoir de valeurs égales dans un échantillon). On résoud ces indéterminations en conduisant deux fois les tests. On juge alors le test impossible si les deux tests conduisent à deux décisions différentes.

5.5 Exercices

Exercice 5.1 *On tente un traitement médical sur certaines personnes du mme ge atteintes d'une grave maladie cardiaque. On note les durées pendant lesquelles ces personnes ont encore vécu ainsi que les durées de vie d'autres malades du mme ge non traités. On a relevé les résultats suivants:*

Traités	*1.2*	*6.3*	*6.5*	*7.8*	*11.2*	*15.6*
Non traités	*0.4*	*3.5*	*4.8*	*6.7*		

Tester l'hypothèse selon laquelle le traitement ne prolonge pas la durée de vie d'un malade (seuil 0.05).

Exercice 5.2 *Deux échantillons indépendants de taille 20 de deux v.a. X et Y ont conduit aux résultats suivants:*

X : 147, 193, 238, 22, 252, 143, 178, 209, 259, 263, 226, 179, 253, 262, 181, 169, 210, 233, 248, 194.

Y : 240, 254, 192, 157, 168, 170, 207, 222, 201, 215, 217, 243, 172, 183, 197, 241, 182, 163, 173, 167.

Avec le seuil 5%, tester avec le test de Wilcoxon l'hypothèse selon laquelle X et Y ont même loi.

Exercice 5.3 *On considère l'hypothèse selon laquelle la densité de l'écorce d'un chêne liège est la même sur le côté nord et sur le côté sud d'un arbre. Pour cela on découpe des cubes de liège de même dimensions sur chaque côté nord et chaque côté sud de 20 arbres. Les poids observés sont les suivants:*

Arbre	*Nord*	*Sud*	*Arbre*	*Nord*	*Sud*
1	*68.3*	*72.5*	*11*	*32.2*	*31.9*
2	*60.1*	*56.0*	*12*	*63.3*	*58.1*
3	*52.2*	*55.8*	*13*	*54.2*	*52.7*
4	*41.7*	*39.2*	*14*	*47.0*	*46.2*
5	*32.0*	*31.4*	*15*	*91.9*	*90.2*
6	*30.9*	*35.5*	*16*	*56.1*	*55.4*
7	*39.3*	*39.2*	*17*	*79.6*	*75.1*
8	*42.0*	*41.1*	*18*	*81.2*	*86.6*
9	*37.7*	*43.3*	*19*	*78.4*	*75.3*
10	*33.5*	*31.7*	*20*	*46.6*	*43.8*

Au seuil $\alpha = 0.05$,

1. *Appliquer le test des signes.*
2. *Appliquer le test de Wilcoxon.*
3. *Pourquoi les résultats sont-ils différents?*

5.6 Correction des exercices

Exercice 5.1– Il faut ici faire un test de Wilcoxon. Comme le nombre de données n'est pas égal pour les traités et les non traités, on utilisera un test pour des données non couplées. Soit X la durée de vie qui reste à un malade non traité, Y la durée de vie restant à un malade traité. On veut donc tester les hypothèses suivantes:

$\mathbf{H_0}$: $X = Y$
$\mathbf{H_1}$: Y est stochastiquement supérieure à X

Ainsi, on espère faire apparaitre une différence (positive) pour les sujets traités. On rappelle que c'est la formulation "naturelle" d'un test d'hypothèse: $\mathbf{H_0}$ correspond à une position "conservatrice" ou "prudente". Ainsi ici on formule $\mathbf{H_0}$ en supposant que le traitement n'apporte rien, l'objectif étant d'éventuellement infirmer cette hypothèse.

On classe les durées de vie dans l'ordre croissant, et l'on remplace les valeurs numériques par X ou Y selon la classe à laquelle appartient le sujet. On obtient donc:

$$\omega = XYXXYXYYY$$

La somme des rangs des X est donc:

$$T = 1 + 3 + 4 + 7 = 15.$$

On a $m = 4$ et $n = 6$. Pour procéder à un test de Wilcoxon, il nous faut donc maintenant:

- Compter le nombre d'épreuves favorables à $T \leq 15$.
- Etablir le nombre total d'épreuves possibles.

Pour commencer, comptons donc le nombre de cas possibles avec $m = 4$ et tels que $T \leq 15$. Elles sont (1,2,3,4), (1,2,3,5), (1,2,3,6), (1,2,3,7), (1,2,3,8), (1,2,3,9), (1,2,4,5), (1,2,4,6), (1,2,4,7), (1,2,4,8), (1,2,5,6), (1,2,5,7), (1,3,4,5), (1,3,4,6), (1,3,4,7) [qui correspond à l'epreuve observée], (1,3,5,6), (2,3,4,5) et (2,3,4,6), soit 16 cas possibles.

Le nombre total d'épreuves possibles est donné quant à lui par le nombre total de faons de placer 4 lettres X dans 10 cases, soit $C_{10}^4 = 210$.

Dans le cas o $\mathbf{H_0}$ est vérifiée, on a donc

$$P(T \leq 15) = \frac{18}{210} = 0.086$$

On accepte donc $\mathbf{H_0}$ à tout seuil inférieur à 0.086 (en particulier pour $\alpha = 0.05$), tandis qu'on le refuse pour tout seuil supérieur à 0.086.

On ne peut donc pas rejeter l'hypothèse que le traitement ne prolonge pas la vie d'un malade.

Exercice 5.2– On classe les 40 résultats par valeurs croissantes, et on ne tient compte que des lettres X et Y.

On obtient donc le résultat suivant:

$$\begin{array}{rl} \omega = & XXYYYYXYYYXXXYYYXXYYY \\ & XXYYYXXXXYYYXXXYXXX \end{array}$$

La somme des rangs de X est $T = 1 + 2 + 7 + \ldots + 39 + 40 = 462$. On fait une approximation gaussienne. Si $\mathbf{H_0}$ est réalisée, T est une v.a. gaussienne d'espérance $\frac{20\times 41}{2} = 410$. La probabilité de trouver une valeur de T aussi petite est donc supérieure à $\frac{1}{2}$, et on accepte $\mathbf{H_0}$.

Exercice 5.3– On va noter X le poids d'un cube extrait côté sud, Y le poids d'un cube extrait côté nord. Notons $Z = Y - X$.

1. Test des signes: on va tester les hypotèses

$$\begin{array}{l} \mathbf{H_0}: \ P(Y > X) = P(X > Y) = \frac{1}{2} \\ \mathbf{H_1}: \ P(Y > X) > \frac{1}{2} \end{array}$$

Z prend 5 fois une valeur négative. Si $\mathbf{H_0}$ est réalisée, le nombre de fois V où Z est négative obéit à une $\mathcal{B}(20, \frac{1}{2})$. Ici, le calcul n'est pas trop ardu:

$$\begin{aligned} P(V \leq 5) &= \left(\frac{1}{2}\right)^{20} \sum_{i=0}^{5} C_{20}^{i} \\ &= \frac{1}{1048576}\left(C_{20}^{0} + C_{20}^{1} + C_{20}^{2} + C_{20}^{3} + C_{20}^{4} + C_{20}^{5}\right) \\ &= \frac{1}{1048576}\left(1 + \frac{20!}{1!(20-1)!} + \frac{20!}{2!(20-2)!} + \frac{20!}{3!(20-3)!}\right. \\ &\qquad \left. + \frac{20!}{4!(20-4)!} + \frac{20!}{5!(20-5)!}\right) \\ &= \frac{1}{1048576}\left(1 + \frac{20!}{19!} + \frac{20!}{2!18!} + \frac{20!}{3!17!} + \frac{20!}{4!16!} + \frac{20!}{5!15!}\right) \\ &= \frac{1}{1048576}\left(1 + 20 + \frac{20\times 19}{2} + \frac{20\times 19\times 18}{3\times 2}\right. \\ &\quad \left. + \frac{20\times 19\times 18\times 17}{4\times 3\times 2} + \frac{20\times 19\times 18\times 17\times 16}{5\times 4\times 3\times 2}\right) \\ &= \frac{1}{1048576}\Big(1 + 20 + (10\times 19) + (10\times 19\times 6) \\ &\qquad + (5\times 19\times 3\times 17) + (19\times 3\times 17\times 16)\Big) \\ &= \frac{1}{1048576}\Big(1 + 20 + 190 + 1140 + 4845 + 15504\Big) \\ &= \frac{21700}{1048576} \\ &\simeq 0.0207 \end{aligned}$$

On a donc $P(V \leq 4) \simeq 0.02$. Cette valeur étant inférieure à 0.05, le test des signes conduit à rejeter $\mathbf{H_0}$.
Si l'on avait utilisé la table du test des signes, on aurait lu à la ligne $n = 20$ et dans la colonne $\alpha = 0.05$ la valeur $k = 5$. Ainsi, on aurait aussi pu conclure au rejet de $\mathbf{H_0}$.
Ainsi, on rejette l'hypothèse selon laquelle les densités d'ecorce sont égales des deux côtés. On a donc plus de chances de trouver de l'ecorce plus epaisse du côté Nord que du côté Sud.

2. Test de Wilcoxon: on va tester les hypotèses

$\mathbf{H_0}$: Z est une v.a. symétrique.
$\mathbf{H_1}$: Z n'est pas une v.a. symétrique.

On va reprendre le tableau des valeurs et indiquer pour chaque couple, au lieu du numéro de l'arbre (qui ne sert pas dans le test de Wilcoxon), la valeur de la différence $Y - X$, soit Nord-Sud:

Nord	Sud	$Y-X$	Nord	Sud	$Y-X$
68.3	72.5	-4.2	32.2	31.9	0.3
60.1	56.0	4.1	63.3	58.1	5.2
52.2	55.8	-3.6	54.2	52.7	1.5
41.7	39.2	2.5	47.0	46.2	0.8
32.0	31.4	0.6	91.9	90.2	1.7
30.9	35.5	-4.6	56.1	55.4	0.7
39.3	39.2	0.1	79.6	75.1	4.5
42.0	41.1	0.9	81.2	86.6	-5.4
37.7	43.3	-5.6	78.4	75.3	3.1
33.5	31.7	1.8	46.6	43.8	2.8

Rangées dans l'ordre des valeurs absolues croissantes, les valeurs observées de Z sont: 0.1, 0.3, 0.6, 0.7, 0.8, 0.9, 1.5, 1.7, 1.8, 2.5, 2.8, 3.1, -3.6, 4.1, -4.2, 4.5, -4.6, 5.2, -5.4 et -5.6.
On ne retient que les signes, l'épreuve observée est donc

$$\omega = (+ + + + + + + + + + + + - + - + - + - -)$$

La somme des rangs des signes - est donc ici 13+15+17+19+20=84. On doit donc maintenant calculer $P(W \leq 84$ sous l'hypothèse $\mathbf{H_0}$. Toutefois, le décompte des cas favorables est ici fastidieux. On va donc utiliser l'approximation gaussienne. Dans le cours, on a indiqué que

$$W \rightsquigarrow \mathcal{N}(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}$$

Donc on considère donc W comme une v.a. gaussienne, d'espérance $\frac{20\times 21}{4} = 105$ et de variance $\frac{20\times 21\times 41}{24} = 717.5$.
On a alors

$$P(W \leq 84) = P(\frac{W - 105}{\sqrt{717.5}} \leq \frac{84 - 105}{\sqrt{717.5}})$$

que l'on recherche sur la table de la loi normale centrée réduite (table de la page 300 de *Statistiques descriptives et décisionnelles*, en utilisant la note de bas de page): $\Phi(-0.78) = 1 - \Phi(0.78) = 1 - 0.7823 = 0.2177$.
Puisque cette valeur est supérieure à $\alpha = 0.05$, le test de Wilcoxon ne conduit pas à rejeter $\mathbf{H_0}$.
D'après le test de Wilcoxon, on ne peut pas rejeter l'hypothèse que la densité de l'écorce varie selon le côté (Nord ou Sud) de l'arbre.

3. Le nombre de signes - étant faible, le test des longueurs conduit à rejeter $\mathbf{H_0}$, mais le test de Wilcoxon tient compte du fait que ces signes - ont des rangs élevés, et conduit à penser que la valeur faible de V provient en fait du seul hasard.

6. Une étude détaillée

Dans les chapitres précédents, nous avons donné les bases nécessaires à la compréhension des statistiques. Bien entendu, ceci ne reste qu'un minuscule apperçu d'un domaine très vaste. Citons rapidement certains problèmes proches, que nous aurions pu également développer: l'estimation et les tests par maximum de vraisemblance; la régression; l'analyse de la variance; la théorie de tests de Neymann-Pearson, etc.

Ici, plutôt que d'étudier l'un de ces aspects, nous donnerons un exemple, en faisant le lien entre les diverses notions que nous avons introduites. Nous en profiterons pour donner quelques dernières notions, à mesure que leur utilisation paraîtra utile.

Dans cet exemple, nous allons nous intéresser à deux populations cellulaires, représentant en fait deux "états" différents de la même population. Il s'agit de travaux menés au laboratoire Analyse d'Images en Pathologie Cellulaire de l'Hopital Saint Louis, à Paris, par S. Portet et J. Vassy. Les données (très partielles) présentées ici le sont avec leur autorisation.

6.1 Préalables

Le cytosquelette

Le *cytosquelette* est un réseau protéique constituant le "squelette" de la cellule. Il est constitué d'un enchevètrement de filaments de plusieurs types, organisés en réseaux qui peuvent s'interconnecter. Chaque type de filament intervient dans le fonctionnement cellulaire. Les filaments (cytokeratine) qui nous intéressent ici, sont supposés préserver l'integrité cellulaire et participer a la transduction des signaux mécaniques extracellulaires par des variations architecturales.

Hypothèse biologique

Le cytosquelette transmet au noyau les informations mécaniques relatives aux conditions extracellulaires, en se réorganisant, en modifiant son architecture.

FIG. 6.1 – *Sous-images sur lesquelles sont estimés les paramètres. Image du haut: cellule sous une gravité de 1g. Image du bas: cellule sous micro-gravité (μg).*

FIG. 6.2 – *Cellule sous une gravité de 1g.*

Protocole expérimental

Afin de tester l'hypothèse, les cellules (issues de la même culture) ont été divisées en plusieurs groupes. Certains de ces groupes ont été fixés au sol (1g par la suite). D'autres ont été centrifugés. Enfin, plusieurs groupes ont pris place à bord d'une mission spatiale non habitée, et ont été fixés tout au long d'un séjour en orbite de 3 semaines (μg par la suite, avec différents instants).

Imagerie

De retour sur terre, les différentes cultures ont été photographiées grâce à un microscope confocal, réglé sur une magnification unique (les sous-images présentées représentent $83\mu m^2$) afin d'obtenir des images comparables. Diverses caractéristiques ont été extraites de ces images, de façon à se livrer à une analyse statistique.

Deux des six images sur lesquelles nous conduirons notre analyse sont présentées page 72. Il s'agit d'images extraites des deux photos de la page 72: cellules où les filaments de cytokératine ont été marqués. Bien que prises sous un grossissement exactement identique, on remarque que la photo 6.1 (μg) ne contient qu'une seule cellule, alors que la photo 6.1 (1g) en contient plusieurs. Ceci est dû à l'étalement de la cellule sous μg.

6.2 Paramètres considérés

Au total, 4 paramètres ont été calculés pour chacune des images. Les valeurs sont obtenues par utilisation de méthodes d'analyse d'images. Nous n'en considèrerons ici que trois:

FIG. 6.3 – *Cellule sous micro-gravité (μg).*

Périmètre

Sa valeur s'obtient en comptant le nombre de pixels formant le tour d'une région. Par conséquent, la valeur est entière.

Aire

La valeur de l'aire d'une région s'obtient en comptant le nombre de pixels la composant. Comme le périmètre, puisque l'on compte un nombre de pixels, la valeur de l'aire est entière.

Compacité

L'indice de compacité d'une région R_i est obtenu en utilisant la formule suivante:

$$C_i = \frac{4\pi A_i}{P_i^2}$$

Cet indice décrit l'aplatissement d'une région:

- S'il est proche de 0, on a affaire à une région très applatie ou découpée. En effet, considérons le cas d'une ellipse d'axe semimineur a et d'axe semimajeur b. Elle est d'aire πab et de circonférence $\simeq \pi\sqrt{2(a^2+b^2)}$. Par conséquent, son coefficient de compacité vaut

$$\begin{aligned} C_i &\simeq \frac{4\pi(\pi ab)}{2\pi^2(a^2+b^2)} \\ &\simeq \frac{2ab}{a^2+b^2} \end{aligned}$$

  Maintenant, si l'on fait tendre a vers 0 et b vers $+\infty$, on se rapproche d'une droite (l'ellipse devient de plus en plus aplatie et longue). Et on a $\lim C_i = 0$.
- S'il est proche de 1, la région est de forme sphérique. En effet, dans le cas extrème d'un disque, on a $C_i = 4\pi(\pi r^2)/(2\pi r)^2 = 1$.

6.3 Les données

Nous disposons des valeurs prises par les paramètres dans trois paires (sol et microgravité) de sous-images. Les paramètres étant relatifs à des régions des sous-images, leur nombre est égal au sein de chacune de ces dernières. Ainsi par exemple, dans la sous-image 1 de micro-gravité, il y a 16 valeurs de périmètre, d'aire et de compacité. Le nombre de paramètres estimés sur chacune des sous-images est résumé dans le tableau suivant, où la notion de paire est arbitraire

(puisqu'il ne s'agit pas de données appariées).

Paire	# paramètres, sol	# paramètres, microgravité
1	17	16
2	19	12
3	18	17
Total	54	45

Nous ne présenterons ici que l'un des jeux de données sur les trois que nous étudierons au total.

Périmètre, cellule au sol

48 74 55 88 71 121 92 101 94 107
66 102 60 74 58 56 42

Périmètre, cellule en vol

51 52 133 53 166 377 59 70 123 109
92 133 112 128 67 80

Aire, cellule au sol

83 259 136 283 187 521 333 389 362 495
203 371 129 246 143 135 88

Aire, cellule en vol

142 101 715 131 937 2542 132 185 485 374
267 578 486 673 147 320

Compacité, cellule au sol

0.4527 0.5944 0.5650 0.4592 0.4662 0.4472 0.4944 0.4792
0.5148 0.5433 0.5856 0.4481 0.4503 0.5645 0.5342 0.5410
0.6269

Compacité, cellule en vol

0.6861 0.4694 0.5079 0.6283 0.5860 0.4273 0.2248 0.4765
0.4744 0.4028 0.3956 0.3964 0.4106 0.4869 0.5162 0.4115

6.4 Description des données

Dans cette première étape, on cherche à se familiariser avec les données. On les étudie donc sous le regard des statistiques descriptives.

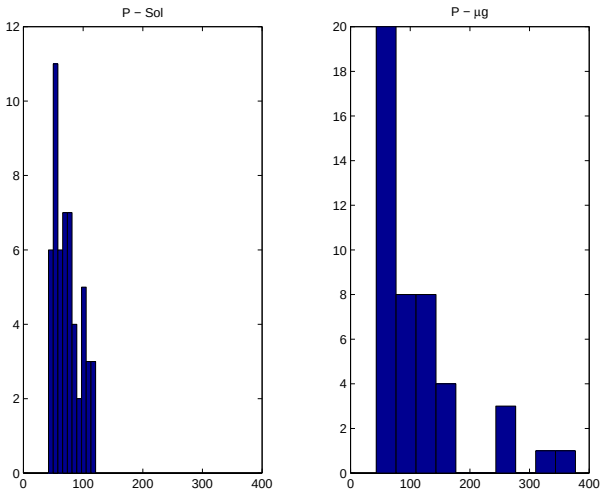


FIG. 6.4 – *Histogrammes du périmètre des zones du réseau filamenteux.*

Histogrammes et fréquences cumulées

La première idée qui vient est de représenter les données sous forme d'histogrammes. Nous allons seulement représenter les données agrégées. Ces histogrammes sont représentés dans les figures 6.4 et suivantes.

Paramètre	Sol	Microgravité
Périmètre	$\bar{P} = 73.2778$ $s_P^2 = 489.3365$	$\bar{P} = 110.4222$ $s_P^2 = 5956.7$
Aire	$\bar{A} = 232.3889$ $s_A^2 = 16580$	$\bar{A} = 535.5556$ $s_A^2 = 456570$
Compacité	$\bar{C} = 0.5192$ $s_C^2 = 0.0055$	$\bar{C} = 0.4771$ $s_C^2 = 0.0119$

La kurtosis

Le *degré d'aplatissement* (ou *kurtosis* en anglais, souvent aussi employé en français) est une mesure de la dispersion des données. Il est définie, pour une distribution donnée, par

$$k = \frac{E(x-\mu)^3}{\sigma^4}$$

La kurtosis d'une distribution normale est 3. Les valeurs supérieures à 3 indiquent donc un étalement plus grand des valeurs, tandis que les valeurs inférieures à 3 indiquent un resserrement des valeurs.

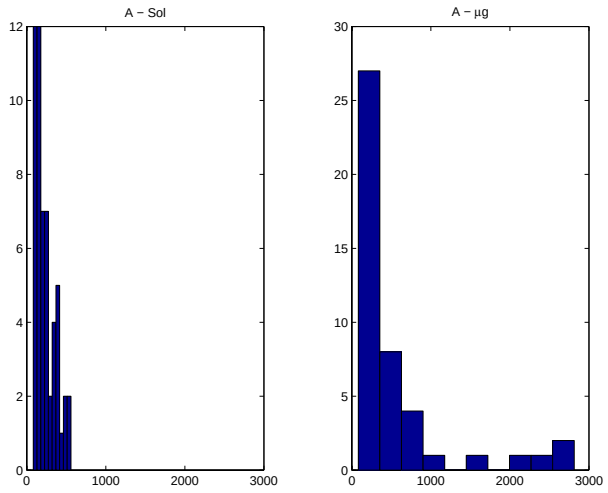


FIG. 6.5 – *Histogrammes de l'aire des zones du réseau filamenteux.*

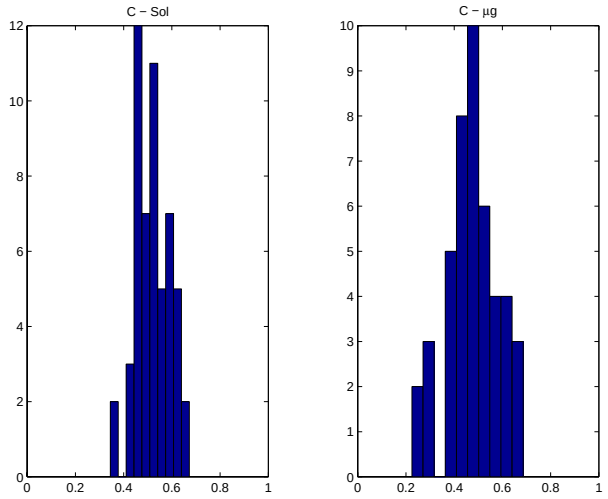


FIG. 6.6 – *Histogrammes de l'indice de compacité des zones du réseau filamenteux.*

Les valeurs calculées de la kurtosis pour nos données (en remplaçant μ et σ par leurs estimateurs $\bar{X}$ et s_1) sont:

Paramètre	Sol	Microgravité
Périmètre	2.1808	6.2689
Aire	2.7945	6.8386
Compacité	2.6023	2.8429

Skewness

Le degré de dissymétrie (*skewness* en anglais) est une mesure de la dissymétrie d'un échantillon. Il est défini par

$$y = \frac{E(x-\mu)^3}{\sigma^3}$$

Si il est négatif, cela signifie que les données sont plus distribuées à gauche de la moyenne empirique de l'échantillon, tandis qu'une valeur positive indique un épanchement à droite. Le degré de dissymétrie d'une distribution symétrique vaut 0.

Les valeurs calculées du degré de dissymétrie pour nos données sont:

Paramètre	Sol	Microgravité
Périmètre	0.5364	1.9037
Aire	0.8966	2.1973
Compacité	-0.0330	-0.1702

6.5 Un test de Wilcoxon

Soit $\mathcal{C}$ l'un des caractères (paramètres) considérés. Dans la suite, on notera indifféremment $\mathcal{C}_P$ ou P le périmètre, $\mathcal{C}_A$ ou A l'aire, et $\mathcal{C}_C$ ou C la compacité. Si l'on veut savoir si $\mathcal{C}$ diffère entre deux images (réalisations), il va nous falloir utiliser des tests pour des observations non couplées, puisque les réalisations sont de tailles différentes pour chacune des images. Même si les observations étaient de même taille au sol et en vol, nous n'utiliserions pas de test pour observations couplées: il ne s'agit pas en effet d'observations du même individu dans deux situations différentes.

6.5.1 Sur un jeu de données

Nous allons commencer par étudier le jeu de données 1. Soit X la v.a. "valeur du caractère $\mathcal{C}$ dans les données au sol", Y la v.a. "valeur du caractère $\mathcal{C}$ dans les données en vol".

On va tester les hypothèses

$\mathbf{H_0}$: $Y = X$
$\mathbf{H_1}$: Y est stochastiquement supérieure à X

Périmètre

On classe les éléments des deux réalisations dans l'ordre des valeurs croissantes, et on retient seulement le fait qu'un valeur provient de X ou de Y. On obtient:

$$xxyyyxxxyxxyyyxxxyxxyxxxxyyxyyyyyy$$

La somme des rangs de x est 244. On fait l'approximation normale: la statistique de Wilcoxon T suit une

$$\mathcal{N}(\frac{m(m+n+1)}{2},\frac{mn(m+n+1)}{12})$$

Ici, $n = 17$ et $m = 16$. Donc on va chercher la probabilité qu'une $\mathcal{N}(272{,}770.67)$ soit inférieure à 244.

$$\begin{aligned} P(\mathcal{N}(272{,}770.67) < 244) &= P(\frac{\mathcal{N}(272{,}770.67) - 280}{\sqrt{770.67}} < \frac{244-272}{\sqrt{770.67}}) \\ &= P(\mathcal{N}(0{,}1) < -1.009) \\ &\simeq \Phi(-1.01) \end{aligned}$$

On cherche sur la table, où l'on trouve que $\Phi(1.01) = 0.84375$, soit $\Phi(-1.01) = 1 - 0.84375 = 0.15625 \simeq 0.16$. Par conséquent, puisque $P(T \leq 265) \simeq 0.16 \not< 0.05$, on ne peut pas rejeter l'hypothèse $\mathbf{H_0}$ selon laquelle les deux lois sont équivalentes.

Aire

On conduit le même raisonnement sur les données de l'aire. On obtient la réalisation

$$xxyxyyxxyxyyxxxxyxyxxxyxyyxxyyyyy$$

où la somme des rangs de x vaut 250. L'approximation normale est la même que dans le test sur le périmètre. Adaptée aux valeurs de cette réalisation, on obtient finalement que

$$P(T \leq 250) \simeq 1 - 0.78524 = 0.21476$$

Par conséquent, on ne peut pas ici non plus rejeter l'hypothèse nulle $\mathbf{H_0}$ selon laquelle les aires sont de même loi.

Compacité

Au vu des sous-images, on a l'impression que le réseau cytosqueletique est plus "désorganisé" pour une cellule en microgravité que pour une cellule au sol. Ainsi, on s'attend à ce que l'indice de compacité soit plus élevé au sol qu'en vol.

Deux façons de procéder sont possibles: on inverse l'hypothèse alternative, ou bien on échange le rôle de X et de Y. Nous allons procéder de cette deuxième manière. Ainsi, X sera ici la v.a. compacité en vol, et Y la v.a. compacité au sol. Le test est alors le même que dans les cas précédents.

On trouve alors (en se rappelant qu'ici x représente une donnée en vol):

$$xxxxxxxyyyyyyyxxxyxyxyxyyyyyyyxyyxx$$

dans lequel la somme des rangs de x vaut 227.

On a $n = 16$ et $m = 17$, on fait l'approximation normale, et l'on trouve

$$P(T \leq 227) \simeq 1 - \Phi(2.23) = 1 - 0.98713 = 0.01287$$

Par conséquent, puisque $0.01287 < 0.05$, on rejette l'hypothèse $\mathbf{H_0}$. Le test est donc significatif (et le sera à tout risque supérieur à 1.2%). Le paramètre de compacité (pour la sous-image sol 1 et la sous-image μg 1) de la cellule au sol est supérieur au paramètre de compacité pour la cellule en vol.

6.5.2 Tests deux à deux

Pour être complet, il nous faut tester les relations 2 à 2 entre les réalisations au sol et en vol. Les résultats de ce calcul (fastidieux, 9 tests sur chacun des paramètres) sont résumés dans les tableaux qui suivent. Dans ces tableaux, les images sont notées I_i (où i = 1,2,3), et les cases contiennent la probabilité que la statistique de Wilcoxon soit inférieure à la somme des rangs de x trouvés. Si cette probabilité est à même de conduire au rejet de $\mathbf{H_0}$ au niveau de risque $\alpha = 0.05$, elle est notée en caractères gras.

Périmètre

		μg		
		I_1	I_2	I_3
Sol	I_1	$P(T \leq 244) = 0.16$	$P(T \leq 239) = 0.99$	$P(T \leq 277) = 0.24$
	I_2	$P(T \leq 266) = 0.23$	$P(T \leq 260) = 0.99$	$P(T \leq 306) = 0.39$
	I_3	$P(T \leq 268) = 0.38$	$P(T \leq 262) = 0.99$	$P(T \leq 301) = 0.43$

Aire

		μg		
		I_1	I_2	I_3
Sol	I_1	$P(T \leq 250) = 0.21$	$P(T \leq 234) = 0.99$	$P(T \leq 290) = 0.4$
	I_2	$P(T \leq 277) = 0.36$	$P(T \leq 260) = 0.99$	$P(T \leq 325) = 0.63$
	I_3	$P(T \leq 266) = 0.31$	$P(T \leq 250) = 0.99$	$P(T \leq 306) = 0.5$

Compacité

Ici, X correspond aux cellules en vol et Y aux cellules au sol. La significativité d'un test indique donc qu'on peut conclure que la compacité du réseau au sol

est supérieure à celle du réseau en vol.

		μg		
		I_1	I_2	I_3
sol	I_1	$\mathbf{P(T \leq 227) = 0.012}$	$\mathbf{P(T \leq 177) = 2.10^{-4}}$	$P(T \leq 255) = 0.072$
	I_2	$\mathbf{P(T \leq 208) = 8.10^{-6}}$	$\mathbf{P(T \leq 167) = 1.10^{-8}}$	$\mathbf{P(T \leq 231) = 7.10^{-5}}$
	I_3	$\mathbf{P(T \leq 257) = 0.022}$	$\mathbf{P(T \leq 197) = 3.10^{-4}}$	$P(T \leq 301) = 0.22$

6.5.3 Données agrégées

Pour terminer avec les tests de Wilcoxon, nous procédons à l'agrégation de données. On a donc 54 valeurs des paramètres au sol, et 45 valeurs en micro-gravité. On fait le test de Wilcoxon sur ces valeurs, et on trouve:

Périmètre $P(T \leq 2369) = 0.79$.

Aire $P(T \leq 2400) = 0.85$.

Compacité $P(T \leq 1952) = 1.10^{-7}$.

Ainsi, sur les données agrégées, les tests sur le périmètre et l'aire sont non significatifs, alors que le test sur la compacité est significatif (conduit au rejet de $\mathbf{H_0}$).

6.5.4 Interprétation de ce test

Il n'est pas possible de mettre en évidence une différence significative entre les tailles (périmètres et aires) des zones visibles sur les sous-images. Toutefois, lorque l'on calcule l'indice de compacité de ces mêmes zones, il apparait nettement une différence. Ainsi, le test de Wilcoxon sur les données agrégées conduit-il au rejet de l'hypothèse nulle (d'égalité des lois) au niveau de risque de 1.10^{-7}, soit 1.10^{-5}%!

Remarque– Cela montre que les apparences de différences peuvent être trompeuses. Si l'on se réfère au tableau présentant les moyennes et variances, on pourrait à première vue conclure à une différence du périmètre et de l'aire et à une similitude de la compacité. Ces tests viennent de nous montrer que c'est l'inverse qui se produit. Cela s'explique par la très forte variance des paramètres périmètre et aire et la faible variance du paramètre compacité.

6.6 Quelques considérations paramétriques

6.6.1 Intervalles de confiance pour la moyenne

Dans le cas des données sous-image par sous-image, nous ne pourrons rien dire sur la moyenne, car la taille de l'échantillon est toujours inférieure à 30. Par contre, pour les données agrégées, nous pouvons, d'après la Table 3.2, construire un IC de μ en utilisant une $\mathcal{N}(\bar{x}, \frac{s_1^2}{n})$. Les valeurs de la moyenne et de la variance empiriques ont été calculées dans la section 6.4.

Rappelons que 'IC au niveau $1-\alpha$ de la moyenne μ est donné par:

$$IC_\alpha = [\bar{x} - \epsilon_\alpha \frac{s_1}{\sqrt{n}} \; ; \; \bar{x} + \epsilon_\alpha \frac{s_1}{\sqrt{n}}]$$

où la valeur ϵ_α est lue sur la table de la loi normale centrée réduite, en cherchant la valeur telle que $P(\mathcal{N}(0,1) > \epsilon_\alpha) = 1 - \frac{\alpha}{2}$.

On obtient donc les intervalles de confiance suivants pour la moyenne μ, au niveau de risque $\alpha = 0.05$:

Paramètre	Sol	Microgravité
Périmètre	[67.38 ; 79.18]	[87.87 ; 132.97]
Aire	[198.04 ; 266.73]	[338.13 ; 732.98]
Compacité	[0.50 ; 0.54]	[0.44 ; 0.51]

6.6.2 Interprétation en terme de tests

On peut interpréter ce tableau de la façon suivante: au risque 5%, les seuls intervalles de confiance ayant une intersection non vide sont ceux relatifs à la compacité. Ainsi, si l'on procède à des tests (paramétriques) d'hypothèse on rejettera, au niveau de risque 5%, l'hypothèse de même moyenne pour le périmètre et l'aire, et le test ne sera pas significatif pour la compacité.

Ce résultat est tout à fait normal: la comparaison à laquelle on vient de se livrer suppose que les lois des deux échantillons que l'on compare sont les mêmes. Mais, en l'absence de vérifications plus complètes, rien ne prouve que tel est le cas.

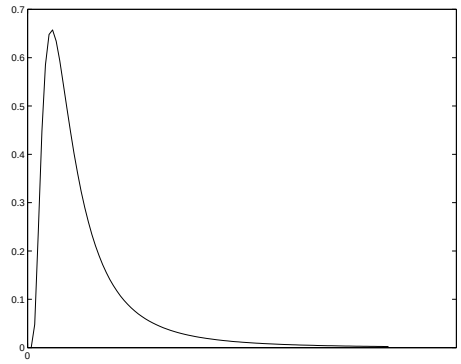


FIG. 6.7 – *Fonction de répartition d'une loi de Fisher.*

6.7 Conclusion

On ne peut pas vraiment conclure, sinon sur le fait que la distribution des indices de compacité est différentes entre le sol et la microgravité. Pourtant, le nombre de données semble satisfaisant. En fait, cela n'est pas vrai. Au regard des histogrammes, on constate la présence de "trous". Puis la kurtosis donne une piste: des valeurs aussi élevées que le sont celles du périmètre et de l'aire en microgravité indiquent la présence probable de *données aberrantes* (*outliers* en anglais). Ces valeurs peuvent traduire plusieurs situations:

- Il y a des erreurs de mesure importantes.
- On ne dispose pas d'assez de données, et la loi sous-jacente est par conséquent mal appréhendée.
- On est en présence d'un *mélange de lois*. Les données sont en fait issues de plusieurs lois sous jacentes.

Ici, il est évident que tout au moins dans le cas de la microgravité, il faudrait disposer de plus de valeurs du périmètre et de l'aire pour trancher ce problème.

Par exemple, dans le cas du périmètre, on se dit qu'il se pourrait bien que les valeurs soient distribuées selon une loi de Fisher, dont la fonction de répartition a l'aspect présenté figure 6.7. Il existe des tests capables de répondre à cette question. Mais la présence des "trous" dans les données ne permettra pas pour autant de trancher entre l'hypothèse d'un échantillon issu d'une telle loi, et l'hypothèse de valeurs aberrantes dues par exemple à une mauvaise mesure.

De manière plus générale, cet exemple présente indirectement certains des travers qui apparaissent parfois dans la pratique des statistiques. Nous avons obtenu une conclusion positive (pour la différence) en utilisant une méthode non paramétrique, et une conclusion apparament inverse en utilisant une méthode paramétrique. L'honneteté voudrait que l'on présente ces deux résultats. La rigueur voudrait elle que l'on cherche les causes de ces apparentes différences.

Mais bien souvent ce n'est pas le cas, et le seul résultat présenté est celui qui va dans le sens de ce que l'on cherchait à montrer. Pourtant, les statistiques permettent de justifier de façon rigoureuse ces contradictions apparentes, pour peu que l'on se donne la peine d'y consacrer un peu de reflexion.