Notes de cours de statistique mathématique élémentaire

Christian Léonard
Département de mathématiques et informatique. Université Paris Ouest Nanterre

Je n'ai pas inclus les illustrations, mais le cours reste lisible.

1

Variables aleatoires discretes

Alea jacta est. En lançant un dé, j'observe une quantité aléatoire susceptible de prendre les valeurs 1, 2, 3, 4, 5 ou 6. Si mon dé est honnête, j'ai une chance sur six d'obtenir chacune de ces valeurs. Nous dirons donc que la probabilité d'observer la valeur 4, par exemple, est $\frac{1}{6}$. Ce qui en notant X le résultat aléatoire du lancer de dé, s'écrit symboliquement : $P(X = 4) = \frac{1}{6}$. On a de même :

$$P(X = 1) = P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}.$$

La probabilité d'observer 3 ou 5 est égale à

$$\frac{\text{nombre d'événements favorables}}{\text{nombre d'événements possibles}} = \frac{\text{nombre d'éléments de } \{3,5\}}{\text{nombre d'éléments de } \{1,2,3,4,5,6\}} = \frac{2}{6} = \frac{1}{3}.$$

En d'autres termes, on a une chance sur trois d'observer soit 3, soit 5. Cette probabilité s'écrit symboliquement $P(X \in \{3,5\}) = \frac{1}{3}$. Remarquons que

$$P(X \in \{3,5\}) = P(X = 3) + P(X = 5)$$

puisque $P(X = 3) + P(X = 5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. De même, la probabilité d'obtenir une valeur paire est

$$P(X \in \{2,4,6\}) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Maintenant, on me propose le jeu suivant : si le dé prend sa valeur dans $\{1,2,3\}$, je gagne 1 franc ; s'il prend sa valeur dans $\{4,5\}$, je gagne 5 francs et s'il prend la valeur 6, je gagne 35 francs. En notant Y mon gain aléatoire (en francs), la probabilité de gagner 1 franc est

$$P(Y = 1) = P(X \in \{1,2,3\}) = \frac{3}{6} = \frac{1}{2},$$

de même $P(Y = 5) = P(X \in \{4,5\}) = \frac{2}{6} = \frac{1}{3}$ et $P(Y = 35) = P(X = 6) = \frac{1}{6}$.

Les comportements aléatoires de X et de Y sont entièrement décrits par les fonctions suivantes :

$$p_X(x) = P(X = x),\ x = 1,2,\ldots,6 \quad \text{et} \quad p_Y(y) = P(Y = y),\ y = 1,5,35,$$

c'est-à-dire : $p_X(1) = \cdots = p_X(6) = \frac{1}{6}$ et $p_Y(1) = \frac{1}{2}, p_Y(5) = \frac{1}{3}$ et $p_Y(35) = \frac{1}{6}$. Noter que

$$p_X(1) + \cdots + p_X(6) = 1 = 100\% \quad \text{et} \quad p_Y(1) + p_Y(5) + p_Y(35) = 1 = 100\%.$$

Or $p_X(1) + \cdots + p_X(6) = P(X \in \{1, \ldots, 6\})$ et $p_Y(1) + p_Y(5) + p_Y(35) = P(Y \in \{1, 5, 35\})$, de sorte que les égalités précédentes signifient que j'ai 100% de chance d'obtenir (j'obtiens à coup sûr) une valeur dans $\{1, \ldots, 6\}$ et de gagner soit 1 Fr, soit 5 Fr, soit 35 Fr.

Cet exemple motive les définitions suivantes.

On dit qu'une quantité aléatoire X susceptible de prendre un nombre fini : k, de valeurs numériques : $x_1, x_2, \ldots, x_k$ est une **variable aléatoire discrète**. *Son comportement aléatoire est décrit par la fonction*

$$p_X(x) = P(X = x), \ x = x_1, \ldots, x_k$$

qui satisfait les conditions

$$0 \leq p_X(x) \leq 1, \ x = x_1, \ldots, x_k \quad \textit{et} \quad p_X(x_1) + \cdots + p_X(x_k) = 1.$$

Cette fonction p_X est appelée la **loi de** X.

Dans l'exemple du dé, p_X et p_Y peuvent être représentés graphiquement à l'aide de barres :

Au lieu de considérer les événements de la form $(X = x)$, il sera pratique de s'intéresser à ceux de la forme $(X \leq x)$ où x parcourt l'ensemble des nombres réels. Avec notre dé, nous avons par exemple :

$$P(X \leq 1) = \frac{1}{6}, P(X \leq 4) = P(X \in \{1, 2, 3, 4\}) = \frac{4}{6}, P(X \leq 6) = \frac{6}{6} = 1$$

ainsi que

$$P(X \leq 1.2) = P(X = 1) = \frac{1}{6} \quad \text{et} \quad P(X \leq 0.5) = 0.$$

L'égalité $P(X \leq 0.5) = 0$ signifie qu'il y a une probabilité 0 (aucune chance) d'obtenir une face dont le numéro est inférieur à 0.5. de même :

$$P(Y \leq 1) = \frac{1}{2},$$
$$P(Y \leq 21.95) = P(Y \leq 5) = P(Y = 1) + P(Y = 5) = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} \quad \text{et}$$
$$P(Y \leq 100) = P(Y \leq 35) = P(Y = 1) + P(Y = 5) + P(Y = 35) = 1.$$

En notant ces probabilités cumulées $F_X(x) = P(X \leq x)$ et $F_Y(y) = P(Y \leq y)$, nous avons les représentations graphiques suivantes :

Dans le graphique de F_X, la hauteur des marches est $\frac{1}{6}$ alors que dans celui de F_Y, la hauteur de la marche située en $y = 1$ est $p_Y(1)$, celle de la marche située en $y = 5$ est $p_Y(5)$, celle de la marche située en $y = 35$ est $p_Y(35)$ et celle de la marche située en $y = 5.2$ est $P(Y = 5.2) = 0$: il n'y a pas de marche à cet endroit.

On pose la définition suivante : soit X une variable aléatoire discrète, la fonction

$$F_X(x) = P(X \leq x),\ x \in \mathbb{R}$$

est appelée la **fonction de répartition** *de X.*

Voici le mode de calcul de F_X. On ordonne les valeurs possibles de X par ordre croissant : $x_1 \leq x_2 \leq \cdots \leq x_k$. Si x est situé entre les $j^{\text{ème}}$ et $(j+1)^{\text{ème}}$ valeurs : $x_j \leq x < x_{j+1}$, alors $F_X(x) = p_X(x_1) + \cdots + p_X(x_{j-1}) + p_X(x_j)$. Si $x < x_1$, alors $F_X(x) = 0$ et si $x \geq x_k$, alors $F_X(x) = p_X(x_1) + \cdots + p_X(x_k) = 1$.

Remarquons qu'une fonction de répartition croît toujours de 0 à 1.

Soit A un ensemble de valeurs que X peut prendre. De deux choses l'une : soit X appartient à A, soit X n'appartient pas à A. Cette remarque se traduit symboliquement par :

$$P(X \in A) + P(X \notin A) = 100\% = 1.$$

On l'utilise souvent sous la forme : $P(X \notin A) = 1 - P(X \in A)$. En particulier, nous avons pour tout $x \in \mathbb{R}$:

$$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x).$$

Dans l'exemple du dé, nous avons $P(X > 4) = 1 - F_X(4) = 1 - \frac{4}{6} = \frac{2}{6}$. Il convient de faire attention et de distinguer $P(X > x)$ et $P(X \geq x)$. En effet, $P(X > 4) = P(X \in \{5, 6\})$ et $P(X \geq 4) = P(X \in \{4, 5, 6\}) = \frac{3}{6}$. De même, il faut distinguer $P(X < x)$ et $P(X \leq x)$.

La personne qui organise le jeu de dé (et qui se propose de me donner 1, 5 ou 35 francs), se demande en retour combien elle doit me faire payer la partie pour être bénéficiaire. Cette personne raisonne correctement de la manière suivante. La partie lui coûte 1 Fr avec la probabilité $P(Y = 1) = \frac{1}{2}$, 5 Fr

avec la probabilité $P(Y=5)=\frac{1}{3}$ et 35 Fr avec la probabilité $P(Y=35)=\frac{1}{6}$. Si un grand nombre de parties a lieu, à peu près 1 partie sur 2 (proportion $\frac{1}{2}$) lui coûtera 1 Fr, 1 partie sur 3 (proportion $\frac{1}{3}$) lui coûtera 2 Fr et 1 partie sur 6 (proportion $\frac{1}{6}$) lui coûtera 35 Fr. Donc, approximativement, en moyenne une partie lui coûtera

$$\frac{1}{2}\cdot(1\text{ Fr})+\frac{1}{3}\cdot(5\text{ Fr})+\frac{1}{6}\cdot(35\text{ Fr})=8\text{ Fr}.$$

C'est-à-dire qu'elle s'attend à payer en moyenne 8 Fr par partie. Une telle moyenne pondérée s'appelle l'espérance mathématique de Y. Si cette personne décide de proposer la partie à 10 Fr, elle s'attend à gagner en moyenne 2 Fr par partie. Mais comme je ne perdrai pas plus de 9 Fr par partie, il se peut je me laisse tenter par la possibilité du gain de 25 Fr.

La formule ci-dessus est un cas particulier de la formule générale de l'*espérance mathématique* $E(Y)$ d'une variable aléatoire Y de loi p_Y :

$$E(Y)=P(Y=y_1)\cdot y_1+\cdots+P(Y=y_k)\cdot y_k=p_Y(y_1)\cdot y_1+\cdots+p_Y(y_k)\cdot y_k$$

où $y_1,\ldots,y_k$ sont les valeurs prises par Y.

D'autre part, mon gain Y est fonction du résultat X du lancer du dé. Plus précisément, $Y=u(X)$ avec

$$u(x)=\begin{cases}1 & \text{si } x = 1,2,3\\ 5 & \text{si } x = 4,5\\ 35 & \text{si } x = 6\end{cases}$$

Il s'ensuit que nous devons avoir $E(Y)=E[u(X)]$, et si l'on prend pour $E[u(X)]$ la quantité

$$\begin{aligned}&P(X=1)\cdot u(1)+P(X=2)\cdot u(2)+\cdots+P(X=6)\cdot u(6)\\ =&\frac{1}{6}\cdot u(1)+\frac{1}{6}\cdot u(2)+\cdots+\frac{1}{6}\cdot u(6)\\ =&\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 35=8\end{aligned}$$

cette égalité est satisfaite.

*Ce qui nous suggère la définition générale de l'***espérance mathématique** *d'une fonction de* X :

$$\begin{aligned}E[u(X)]&=P(X=x_1)\cdot u(x_1)+\cdots+P(X=x_k)\cdot u(x_k)\\ &=p_X(x_1)\cdot u(x_1)+\cdots+p_X(x_k)\cdot u(x_k).\end{aligned}$$

Pour une variable aléatoire discrète générale X, $E(X)$ *s'appelle sa* **moyenne.** *Si* $E(X)=\mu$, *on définit la* **variance** *de* X *par*

$$Var(X)=E[(X-\mu)^2]=p_X(x_1)\cdot(x_1-\mu)^2+\cdots+p_X(x_k)\cdot(x_k-\mu)^2$$

*et l'***écart type** *de* X *est défini par*

$$\sigma=\sqrt{Var(X)}=\sqrt{E[(X-\mu)^2]}.$$

Par exemple, si X est la face du dé, nous avons

$$E(X) = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \cdots + \frac{1}{6}\cdot 6 = 7/2 = 3.5.$$

et

$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{6}\cdot(1-3.5)^2 + \frac{1}{6}\cdot(2-3.5)^2 + \cdots + \frac{1}{6}\cdot(6-3.5)^2 \\ &= 35/12 = 2.917\end{aligned}$$

et l'écart type est $\sigma = \sqrt{35/12} \simeq 1.708$.

On montre par le calcul que la variance de X est aussi égale à :

$$\mathrm{Var}(X) = E(X^2) - \Big(E(X)\Big)^2.$$

dans l'exemple précédent, on vérifie bien que

$$E(X^2) = \frac{1}{6}\cdot 1^2 + \frac{1}{6}\cdot 2^2 + \cdots + \frac{1}{6}\cdot 6^2 = 15.167,$$

de sorte que $\mathrm{Var}(X) = 15.167 - (3.5)^2 = 2.917$.

Un exemple important. Une des variables aléatoires les plus simples est X qui ne peut prendre que deux valeurs. On choisit souvent pour ces deux valeurs : 0 et 1. La variable aléatoire prend la valeur 1 avec la probabilité p où $0 \leq p \leq 1$, elle prend donc l'autre valeur : 0, avec la probabilité complémentaire : $1-p$. Sa loi est donc

$$p_X(1) = p \quad \text{et} \quad p_X(0) = 1-p.$$

On dit que X suit une **loi de Bernoulli** de paramètre p. Ce que l'on note

$$X \sim \mathcal{B}(p).$$

Calculons les moyenne et variance de X. Nous avons, $E(X) = p\cdot 1 + (1-p)\cdot 0 = p$ et $E(X^2) = p\cdot 1^2 + (1-p)\cdot 0^2 = p\cdot 1 + (1-p)\cdot 0 = p$, de sorte que $\mathrm{Var}(X) = E(X^2) - \Big(E(X)\Big)^2 = p - p^2 = p(1-p)$. On résume :

$$\text{si } X \sim \mathcal{B}(p), \quad \text{alors :} \quad E(X) = p \quad \text{et} \quad \mathrm{Var}(X) = p(1-p).$$

Exercices

1. Dans un hall de gare se tiennent 50 personnes : 11 ont des revenus très faibles (Classe 1), 19 ont des revenus assez faibles (Classe 2), 14 ont des revenus moyens (Classe 3) et 6 ont des revenus élevés (Classe 4). Une de ces personnes est interrogée au hasard. Soit X la variable aléatoire, à valeurs dans $\{1,2,3,4\}$, qui est égale à la classe de revenu de la personne interrogée.

Trouver la loi p_X et la fonction de répartition F_X de X.
En donner des représentations graphiques.

2. Deux des huit barrettes de mémoire de mon ordinateur sont défectueuses. Pour le réparer, je décide de retirer au hasard 2 barrettes et de les remplacer par des barrettes en bon état. Soit X le nombre de barrettes défectueuses qui se trouvent parmi les 2 barrettes que je viens de retirer.

Trouver la loi p_X et la fonction de répartition F_X de X.
En donner des représentations graphiques.

3. La loi p_X de X est donnée par $p_X(0) = 3/10, p_X(1) = 3/10, p_X(2) = 1/10$ et $p_X(3) = 3/10$. Calculer les moyenne, variance et écart-type de X.

4. On prend deux boules au hasard (sans remplacement) dans une urne qui contient 3 boules vertes et 5 boules rouges. Soit X le nombre de boules vertes qui viennent d'être tirées. Calculer les moyenne et variance de X.

5. Trouver la moyenne et la variance de la variable aléatoire dont la fonction de répartition est donnée par

$$F_X(x) = \begin{cases} 0 & \text{si } x < 10 \\ 1/4 & \text{si } 10 \leq x < 15 \\ 3/4 & \text{si } 15 \leq x < 20 \\ 1 & \text{si } 20 \leq x. \end{cases}$$

2

Variables aleatoires continues

Il existe des quantités aléatoires qui peuvent prendre une infinité de valeurs. Par exemple, si je joue à Pile ou Face jusqu'à ce que j'obtienne Pile pour la première fois, le nombre de tirages X qu'il me faut pour voir apparaître Pile une première fois peut prendre toutes les valeurs entières 1, 2, ... Même si la probabilité que X dépasse 1000000000 est très faible, il est tout de même possible que cet événement se produise.

Mais que penser de ma calculette qui possède un programme de tirage de nombres au hasard ? Ces nombres sont tirés entre 0 et 1 et l'on m'a dit que tous ces nombres ont la même probabilité d'être tirés. Soit X le nombre que me donne le programme de ma calculette. Je sais qu'à coup sûr $X \in [0,1]$, ce qui s'écrit symboliquement :

$$P(X \in [0,1]) = 100\% = 1.$$

($[0,1]$ désigne l'ensemble de tous les réels compris entre 0 et 1). Quelle est la probabilité que X prenne exactement la valeur 0.2 ? Puisque ce tirage ne favorise ni ne défavorise aucune valeur de $[0,1]$, je dois avoir

$$\begin{aligned} P(X = 0.2) &= \frac{\text{nombre de réels qui valent 0.2 parmi les réels de [0,1]}}{\text{nombre de réels de [0,1]}} \\ &= \frac{1}{\infty} \\ &= 0. \end{aligned}$$

De sorte que pour tout $x \in [0,1]$, $P(X = x) = 0$. On ne s'est pas trompé en me disant que toutes les valeurs sortent avec la même probabilité, mais ça ne m'avance pas pour calculer $P(X \in [0, \frac{1}{2}])$. Pourtant, il est clair que puisque $\frac{1}{2}$ est le milieu de [0,1], il y a autant de chance pour que X soit supérieur à $\frac{1}{2}$ que pour que X lui soit inférieur. On a donc $P(X \in [0, \frac{1}{2}]) = P(X \in [\frac{1}{2}, 1]) = 50\% = 0.5$.

Puisque 0.5 est la longueur des segments $[0, \frac{1}{2}]$ et $[\frac{1}{2}, 1]$, ceci nous suggére que le comportement aléatoire de X est décrit, pour tous $0 \leq a \leq b \leq 1$, par

$$P(X \in [a,b]) = \text{longueur de } [a,b] = b - a.$$

En particulier, en considérant des intervalles qui enserrent de plus en plus la valeur $x = 0.2$, nous avons

$$P(X \in [0.15, 0.25]) = 0.10 = 10\%$$
$$P(X \in [0.19, 0.21]) = 0.02 = 2\%$$
$$P(X \in [0.199, 0.201]) = 0.002 = 0.2\%$$
$$P(X = 0.2) = P(X \in [0.2, 0.2]) = 0.$$

Si le tirage de X est *uniforme* sur l'intervalle $[0, L]$, plutôt que sur $[0, 1]$, on doit bien sûr avoir $P(X \in [0, L]) = 100\% = 1$, et il est naturel de généraliser la formule : $\dfrac{\text{nombre d'événements favorables}}{\text{nombre d'événements possibles}}$ (pour ne privilégier ni ne défavoriser aucunes des valeurs de $[0, L]$), par :

$$P(X \in [a, b]) = \frac{\text{longueur de } [a, b]}{\text{longueur de } [0, L]} = \frac{b - a}{L}, \text{ pour tous } 0 \leq a \leq b \leq L.$$

En considérant la fonction

$$f_X(x) = \begin{cases} \frac{1}{L} & \text{si } x \in [0, L] \\ 0 & \text{si } x \notin [0, L] \end{cases}$$

l'interprétation graphique de la formule $P(X \in [a, b]) = \frac{b-a}{L}$ est la suivante :

La surface du rectangle hachuré est $(b - a) \cdot \frac{1}{L} = \frac{b-a}{L} = P(X \in [a, b])$. En particulier, la surface du rectangle pointillé est $L \cdot \frac{1}{L} = 1 = P(X \in [0, L])$.

La fonction f_X détermine le comportement du tirage aléatoire X **uniforme** sur $[0, L]$. Elle joue un rôle analogue à la loi p_X d'une variable aléatoire discrète.

Pour tout $x \in \mathbb{R}$, la quantité $F_X(x) = P(X \leq x)$ est donnée par

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x}{L} & \text{si } 0 \leq x \leq L \\ 1 & \text{si } x \geq L \end{cases}$$

puisque l'événement $X \leq x$ est impossible si $x \leq 0$, $X \leq x$ est toujours satisfait si $x \geq L$ et si $0 \leq x \leq L$, $P(X \leq x) = P(X \in [0, x]) = \frac{x-0}{L} = \frac{x}{L}$.

On appelle f_X la *densité* de la loi de X et F_X est sa *fonction de répartition.*

On peut généraliser cette façon de construire des quantités aléatoires, de la manière suivante. On se donne une fonction f positive, dont le graphe est tel que

la surface comprise entre le graphe de f et l'axe horizontal est égale à 1. On décrit alors le comportement d'une quantité aléatoire X par la formule

(1) $P(X \in [a,b]) =$ surface de $= \dfrac{\text{surface de}}{\text{surface de}}$, pour tous $a \leq b$

la dernière égalité ayant lieu puisque surface de $= 1$. En particulier, pour tout $x \in \mathbb{R}$

$$P(X = x_o) = P(X \in [x_o, x_o]) = \text{surface de} \qquad = 0$$

et

$$P(X \in \mathbb{R}) = P(X \in]-\infty, +\infty[) = \text{surface de} \qquad = 1.$$

Nous donnons maintenant quelques définitions.

Une quantité aléatoire X dont le comportement est décrit par (1) est appelée une **variable aléatoire continue.** *La fonction f est sa* **densité.** *Sa* **fonction de répartition** *F est définie, comme pour les variables aléatoires discrètes, par*

$$F(x) = P(X \leq x), \ x \in \mathbb{R}.$$

Nous avons donc $F(x_o) =$ surface de . Mathématiquement, la surface de est donnée par l'intégrale $\int_a^b f(x)\,dx$, donc

$$P(X \in [a,b]) = \int_a^b f(x)\,dx.$$

En particulier $F(x) = \int_{-\infty}^{x} f(t)\,dt$ et sa dérivée est $F'(x) = f(x)$.

Une formule très utile au sujet des fonctions de répartition des variables aléatoires continues, est celle-ci :

$$P(a \leq X \leq b) = F(b) - F(a), \quad a \leq b,$$

où X est une variable aléatoire continue de fonction de répartiton F. En effet,

surface de = surface de − surface de soit

$$\begin{array}{ccccc} P(a \leq X \leq b) & = & P(X \leq b) & - & P(X \leq a) \\ & = & F(b) & - & F(a) \end{array}$$

Si f est de la forme

alors pour tous $\alpha \leq a \leq b \leq \beta$, $P(X \in [a,b]) =$ surface de $= 0$. En d'autres termes, X ne peut pas prendre les valeurs x telles que $f(x) = 0$.

Il existe aussi des notions de **moyenne** de X : $E(X)$, et de **variance** de X : $\mathrm{Var}(X)$, lorsque X est une variable aléatoire continue.

Mathématiquement, les définitions de $E(X)$ et $\mathrm{Var}(X)$ sont

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

et en notant $E(X) = \mu$,

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx.$$

Dessinons le graphe de la densité f de X sur une plaque de bois régulière et découpons la partie comprise entre l'axe horizontal et f. Si l'on cherche à maintenir cet objet découpé en équilibre sur

une pointe en contact avec l'axe horizontal Ox, le seul endroit où l'on peut placer la pointe se situe en $x = E(X) = \mu$.

Attention ! Si l'on découpe cet objet en suivant la droite verticale passant par $E(X) = \mu$, les deux morceaux ainsi obtenus n'ont pas nécessairement la même masse.

Exemples. Soit X un tirage aléatoire uniforme sur $[0, 5]$, alors $E(X) = \frac{5}{2} = 2.5$:

On considère une variable aléatoire Y de densité : $f_Y = \begin{cases} 2y & \text{si } y \in [0,1] \\ 0 & \text{sinon} \end{cases}$. Alors, $E(Y) = \int y f_Y(y)\, dy = \int_0^1 2y^2\, dy = \frac{2}{3}$. Notons que $P(Y \leq \frac{2}{3}) = \int_0^{\frac{2}{3}} 2y\, dy = \frac{4}{9} \neq \frac{1}{2}$.

La variance de X et son **écart-type** $\sigma(X) = \sqrt{\mathrm{Var}(X)}$ sont des quantités qui mesurent la dispersion des valeurs possibles de X autour de sa moyenne. Considérons les quatre fonctions de densité suivantes

Nous avons : $\mathrm{Var}(X_1) < \mathrm{Var}(X_0)$, $\mathrm{Var}(X_2) > \mathrm{Var}(X_0)$ et $\mathrm{Var}(X_3) > \mathrm{Var}(X_0)$.

Exercices

1. Soit X une variable aléatoire distribuée uniformément sur $[-1, +1]$.

a) Donner la densité de X. Dessiner son graphe.

b) Calculer $E(X)$, $\mathrm{Var}(X)$ et $\sigma(X)$. *Indication :* $\int_a^b x^2\,dx = \frac{b^3-a^3}{3}$.

c) Mêmes questions lorsque X est une variable aléatoire distribuée uniformément sur $[-2, +2]$.

d) Mêmes questions lorsque X est une variable aléatoire distribuée uniformément sur $[-3, +3]$.

e) Comparer les résultats.

2. Soit X une variable aléatoire dont la densité est de la forme

$$f(x) = \begin{cases} c & \text{si } x \in [0.5, 1.5] \\ c & \text{si } x \in [3, 5] \\ 0 & \text{sinon} \end{cases}$$

a) Calculer c pour que f soit une densité. Représenter f graphiquement.

b) Calculer $E(X)$. *Indication :* On pourra dessiner le graphe de $x \mapsto xf(x)$ et calculer une surface.

c) Posons $\mu = E(X)$, calculer $P(X \geq \mu)$ et $P(X \leq \mu)$.

3

La loi normale

On dit qu'une variable aléatoire continue Z suit une **loi normale centrée réduite**, si sa densité est définie par

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}.$$

On note $\mathcal{N}(0,1)$ la loi normale centrée réduite ainsi que $X \sim \mathcal{N}(0,1)$ pour signifier que la variable aléatoire X suit la loi $\mathcal{N}(0,1)$.

La loi normale est une des lois les plus importantes pour les applications statistiques. Elle apparaît naturellement lorsqu'on observe des grands échantillons. Ce point sera détaillé lors de la Leçon 4 à l'occasion du Théorème de la Limite Centrale. La représentation de f_Z est

C'est la fameuse "courbe en cloche". On remarque qu'elle est symétrique par rapport à l'axe vertical et on en déduit que si $Z \sim \mathcal{N}(0,1)$, alors $P(Z \leq 0) = P(Z \geq 0) = \frac{1}{2}$ et $E(Z) = 0$.
On note Φ la fonction de répartition de $\mathcal{N}(0,1)$:

$$\Phi(t) = P(Z \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz, \quad z \in \mathbb{R}.$$

La surface hachurée dans la figure ci-dessus est $\Phi(z_o)$. Il n'est pas possible d'évaluer l'intégrale ci-dessus à l'aide d'une fonction usuelle. Par contre, des approximations numériques sont accessibles ; elles ont été tabulées dans la Table I (de valeurs numériques).

En raison de la symétrie de f_Z, on a : $\Phi(-t) = 1 - \Phi(t)$.

C'est pourquoi, seules les valeurs de $\Phi(t)$ pour $t \geq 0$ ont été considérées dans la Table I.

Exemple 1. Si $Z \sim \mathcal{N}(0,1)$, alors

$$\begin{aligned} P(0 \leq Z \leq 2) &= \Phi(2) - \Phi(0) = 0.9772 - 0.5000 = 0.4772, \\ P(1.25 \leq Z \leq 2.75) &= \Phi(2.75) - \Phi(1.25) = 0.9970 - 0.8944 = 0.1026 \quad \text{et} \\ P(-1.65 \leq Z \leq 0.70) &= \Phi(0.70) - \Phi(-1.65) = \Phi(0.70) - [1 - \Phi(1.65)] \\ &= 0.7580 - 1 + 0.9505 = 0.7085. \end{aligned}$$

Exemple 2. Si $Z \sim \mathcal{N}(0,1)$, trouver des constantes a, b et c telles que

$$P(0 \leq Z \leq a) = 0.4147, \quad P(Z > b) = 0.05 \quad \text{et} \quad P(|Z| \leq c) = 0.95.$$

Ces trois équations sont équivalentes à

$$P(Z \leq a) = 0.9147, \quad P(Z \leq b) = 0.95 \quad \text{et} \quad P(Z \leq c) = 0.975,$$

respectivement. On voit dans la Table I que $a = 1.37, b = 1.645$ et $c = 1.96$.

On peut montrer que si $Z \sim \mathcal{N}(0,1)$, alors

$$E(Z) = 0 \quad \text{et} \quad \text{Var}(Z) = 1.$$

Le $(0,1)$ de $\mathcal{N}(0,1)$ correspond à ces égalités. On généralise maintenant la définition de la loi normale centrée réduite.

Soit X une variable aléatoire continue qui peut s'écrire sous la forme

$$X = \mu + \sigma Z$$

où $\mu \in \mathbb{R}$ et Z suit un loi $\mathcal{N}(0,1)$. On dit alors que X suit une **loi normale** *de moyenne μ et de variance σ^2. Ce que l'on note : $X \sim \mathcal{N}(\mu, \sigma^2)$.*

On peut en effet montrer que dans ce cas : $E(X) = \mu$ et $\text{Var}(X) = \sigma^2$.

Il est clair que

$$\text{si} \quad X \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{alors :} \quad \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

On utilise cette remarque de la façon suivante. Soient $X \sim \mathcal{N}(\mu, \sigma^2)$ et $a \leq b$. Alors

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Exemple 3. Si $X \sim \mathcal{N}(3, 16)$, alors

$$\begin{aligned} P(4 \leq X \leq 8) &= P\left(\frac{4-3}{4} \leq \frac{X-3}{4} \leq \frac{8-3}{4}\right) \\ &= \Phi(1.25) - \Phi(0.25) = 0.8944 - 0.5987 = 0.2957, \\ P(0 \leq X \leq 5) &= P\left(\frac{0-3}{4} \leq Z \leq \frac{5-3}{4}\right) \\ &= \Phi(0.5) - \Phi(-0.75) = \Phi(0.5) + \Phi(0.75) - 1 = 0.4649 \quad \text{et} \\ P(-2 \leq X \leq 1) &= P\left(\frac{-2-3}{4} \leq Z \leq \frac{1-3}{4}\right) \\ &= \Phi(-0.5) - \Phi(-1.25) = 0.2029. \end{aligned}$$

Exemple 4. Si $X \sim \mathcal{N}(25, 36)$, on veut une constante c telle que

$$P(|X - 25| \leq c) = 0.9544.$$

On veut donc

$$P\left(-\frac{c}{6} \leq \frac{X-25}{6} \leq \frac{c}{6}\right) = 0.9544.$$

C'est-à-dire

$$\Phi\left(\frac{c}{6}\right) - \left[1 - \Phi\left(\frac{c}{6}\right)\right] = 0.9544,$$

soit

$$\Phi\left(\frac{c}{6}\right) = 0.9772.$$

La lecture de la Table I, nous permet de voir que $\Phi(2) = 0.9772$. Par conséquent, $c/6 = 2$ et $c = 12$.

Exercices

1. Si $Z \sim \mathcal{N}(0,1)$, trouver

a)	$P(0.53 < Z \leq 2.06)$	**b**)	$P(-0.79 \leq Z < 1.52)$
c)	$P(-2.63 \leq Z \leq -0.51)$	**d**)	$P(Z > -1.77)$
e)	$P(Z > 2.89)$	**f**)	$P(\|Z\| < 1.96)$
g)	$P(\|Z\| < 1)$	**h**)	$P(\|Z\| < 2)$

2. Un producteur de saucissons indique le poids 204 grammes sur ses produits. On suppose que la loi des poids de ces saucissons est $\mathcal{N}(213.7, 16)$. Soit X le poids d'un saucisson pris au hasard à la sortie de l'usine. Trouver $P(X < 204)$.

3. Si $X \sim \mathcal{N}(0.15, 0.25)$, trouver

a)	$P(0.53 < X \leq 2.06)$	**b**)	$P(-0.79 \leq X < 1.52)$
c)	$P(-2.63 \leq X \leq -0.51)$	**d**)	$P(X > -1.77)$
e)	$P(X > 2.89)$	**f**)	$P(\|X\| < 1.96)$
g)	$P(\|X\| < 1)$	**h**)	$P(\|X\| < 2)$

4

Les grands echantillons

Notion d'echantillon aleatoire. On observe un **échantillon aléatoire**, c'est-à-dire qu'on observe les valeurs $x_1, \ldots, x_n$ relatives à n individus. Ces données proviennent de variables aléatoires $X_1, \ldots, X_n$ ayant toutes la **même loi** et que l'on suppose **indépendantes** les unes des autres.

Dire que $X_1, \ldots, X_n$ ont la même loi, c'est dire que leurs fonctions de répartition sont égales : $F_{X_1}(x) = \cdots = F_{X_n}(x),\ \forall x$. On rappelle que $F_X(x) = P(X \leq x)$.

Dire que $X_1, \ldots, X_n$ sont indépendantes, signifie que la connaissance de $X_2 = 0.21$ (par exemple) n'apporte aucune information sur le comportement aléatoire des autres variables $X_1, X_3, X_4, \ldots$ Plus généralement, la connaissance de $X_2 = 0.21$ et $X_5 \geq 0$ (par exemple), n'apporte aucune information sur le comportement aléatoire des autres variables $X_1, X_3, X_4, X_6, \ldots$, etc. Expérimentalement, pour que $X_1, \ldots, X_n$ soient indépendantes, il faut que les individus $1, 2, \ldots, n$ n'aient pas d'influence mutuelle. Pour observer un échantillon, un enquêteur se gardera, après avoir interrogé un individu i (dont la réponse est $X_i = x_i$) de lui demander de lui recommander un ami (ou un ennemi, etc.) pour continuer son enquête. La procédure généralement requise pour fabriquer un échantillon est le tirage au sort des individus interrogés au sein d'une population. Plus la taille de l'échantillon tiré au hasard est grande, plus l'échantillon est représentatif de la population à étudier.

Si les variables aléatoires sont discrètes, l'indépendance de $X_1, \ldots, X_n$ se traduit mathématiquement par

$$P(X_1 = a_1 \text{ et } X_2 = a_2 \text{ et } \ldots \text{ et } X_n = a_n) = P(X_1 = a_1)P(X_2 = a_2) \cdots P(X_n = a_n)$$

où les $a_1, \ldots, a_n$ parcourent toutes les valeurs possibles de $X_1, \ldots, X_n$. Une propriété analogue existe pour les variables aléatoires continues.

On appelle **échantillon** *de taille n de la loi de X la donnée de n variables aléatoires réelles $X_1, \ldots, X_n$ indépendantes, ayant toutes la même loi qu'une variable aléatoire X donnée.*

Exemple 1. Par exemple, un échantillon de taille 25 de la loi $\mathcal{N}(-21, 15.2)$ est la donnée de variables aléatoires $X_1, \ldots, X_{25}$ indépendantes qui suivent toutes la loi $\mathcal{N}(-21, 15.2)$.

Exemple 2. (Proportion d'une catégorie d'individus). Un exemple important est celui du tirage "au hasard" (uniforme) dans une grande population d'individus dont une proportion p ($0 \leq p \leq 1$) appartient à une catégorie particulière (par exemple : sensibilité politique, chômeur, homme, femme, fumeur, salaire mensuel supérieur à 11000 francs, etc...) On tire au hasard 100 individus ($i = 1, \ldots, 100$) dans cette population. La variable X_i prend la valeur $x_i = 1$ si le $i^{\text{ème}}$ individu appartient à la catégorie étudiée ou la valeur $x_i = 0$ sinon. Puisque le tirage est uniforme, X_i suit une loi de Bernoulli de paramètre p, notée $\mathcal{B}(p)$ (voir la Leçon 1). Si les tirages sont indépendants, $X_1, \ldots, X_{100}$ est un échantillon de taille 100 de la loi $\mathcal{B}(p)$.

A part l'exemple que nous venons de considérer, dans la pratique on ne connaît pas, en général, avec précision la forme de la loi des X_i que l'on observe. Toutefois, il est possible d'estimer la moyenne $\mu := E(X_1) = \cdots = E(X_n)$ à l'aide de la **moyenne empirique observée**

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Une idée naturelle est de dire que μ ne doit pas être très éloignée de la moyenne empirique : $\mu \simeq \bar{x} = \dfrac{x_1 + \cdots + x_n}{n}$. Bien sûr, μ ne dépend pas de notre observation (c'est un paramètre théorique que l'on cherche à estimer) et une autre expérience qui nous aurait amenés à observer $X_1 = x'_1, \ldots, X_n = x'_n$, nous amènerait à la conclusion $\mu \simeq \dfrac{x'_1 + \cdots + x'_n}{n}$ de sorte qu'il est faux d'affirmer que μ vaut $\bar{x}$.

Heureusement, un résultat mathématique vient à notre secours.

Loi des Grands Nombres. *Soit un grand nombre n de variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi (un échantillonde taille n). Alors, avec une probabilité proche de 100 %, la variable aléatoire*

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

prend des valeurs proches de $\mu := E(X_1) = \cdots = E(X_n)$.

La proximité de $\bar{x}$ avec μ est d'autant plus grande que la taille n de l'échantillon est importante.

Dans l'exemple 1, on peut s'attendre à ce que la moyenne empirique observée $\bar{x} = \dfrac{x_1 + \cdots + x_{25}}{25}$ soit proche de la moyenne théorique $\mu = -21$.

Dans l'exemple 2, la moyenne empirique observée

$$\begin{aligned}\bar{x} &= \frac{x_1 + \cdots + x_{100}}{100} \\ &= \frac{\text{nombre de d'individus dans l'échantillon appartenant à la catégorie étudiée}}{\text{taille de l'échantillon}}\end{aligned}$$

est la proportion observée d'individus dans l'échantillon appartenant à la catégorie étudiée. On peut s'attendre à ce que cette proportion observée soit proche de la proportion $p = E(X)$ d'individus dans la population totale, appartenant à la de la catégorie étudiée.

Il existe un résultat mathématique plus précis que la loi des grands nombres ; il sera d'une importance capitale dans la suite de ce cours. C'est le Théorème de la Limite Centrale.

Théorème de la Limite Centrale. *Soit un grand nombre n de variables aléatoires indépendantes $X_1, \ldots, X_n$ et de même loi (un échantillonde taille n). On note μ et σ^2 les moyenne et variance commune de $X_1, \ldots, X_n$. Lorsque n est grand, la variable aléatoire $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ suit approximativement la loi normale $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$, même si les variables aléatoires ne sont pas normales.*

Sous les mêmes conditions, ce théorème peut aussi s'énoncer des deux manières suivantes.

- $X_1 + \cdots + X_n$ *suit approximativement la loi normale* $\mathcal{N}(n\mu, n\sigma^2)$, *ou*
- $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ *suit approximativement la loi normale* $\mathcal{N}(0, 1)$.

Dans la pratique, on considère qu'à partir de $n \geq 30$, n est suffisamment grand pour pouvoir appliquer l'approximation du Théorème de la Limite Centrale.

Loi binômiale. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes qui suivent une loi de Bernoulli de paramètre p : $\mathcal{B}(p)$ (voir la Leçon 1). On considère leur somme

$$S = X_1 + \cdots + X_n.$$

C'est une variable aléatoire qui prend ses valeurs dans l'ensemble $\{0, 1, \ldots, n\}$. Par définition, la loi de S est la **loi binômiale** de paramètres n, p que l'on note $\mathcal{B}(n, p)$. Un calcul de dénombrement nous donne, pour tout $0 \leq k \leq n$,

$$I\!P(S = k) = C_n^k p^k (1-p)^{n-k}$$

où $C_n^k = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 21}$ est le nombre de parties à k éléments dans un ensemble à n éléments.

En fait, lorsque n devient grand, ces quantités et surtout des quantités comme $I\!P(a \leq S \leq b)$ deviennent difficiles à calculer, même avec des calculatrices puissantes. Heureusement, le Théorème de la Limite Centrale va venir à notre secours, comme nous pourrons le constater dans l'exercice suivant.

Approximation normale d'une loi binômiale. Soit S une variable aléatoire de loi binômiale $\mathcal{B}(n, p)$. Par définition, ceci signifie que S peut s'écrire

$$S = X_1 + \cdots + X_n$$

où $X_1, \ldots, X_n$ sont des variables indépendantes de loi de Bernoulli de paramètre p (voir la Leçon 1). C'est-à-dire que X_i peut prendre les valeurs 0 ou 1 avec les probabilités $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$, où $0 \leq p \leq 1$. Lorsque n est grand (supérieur à 30 en pratique), on peut appliquer le Théorème de la Limite Centrale avec $\mu = E(X) = p$ et $\sigma^2 = \mathrm{Var}(X) = p(1-p)$. On obtient que $S = X_1 + \cdots + X_n$ suit approximativement la loi normale $\mathcal{N}(np, np(1-p))$. Pour calculer la probabilité $P(a \leq S \leq b)$ où a et b sont des entiers $0 \leq a \leq b \leq n$, on effectue l'approximation

suivante :

$$\begin{aligned} P\left(a \leq S \leq b\right) &= P\left(a - \frac{1}{2} \leq S \leq b + \frac{1}{2}\right) \\ &= P\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq \frac{S - np}{\sqrt{np(1-p)}} \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \\ &\simeq P\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \\ &= \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

où Φ désigne la fonction de répartition de la loi normale.
La première égalité ci-dessus s'appelle la correction de continuité. Dans l'exemple suivant, nous allons constater que dans certaines situations, cette approximation reste excellente même lorsque n est plus petit que 30 ($n = 10$ dans ce qui suit).

Exemple. On joue 10 fois à Pile ou Face. Soit S le nombre de Face obtenu. La loi de S est parfaitement calculable à l'aide de dénombrements. On obtient en particulier que

$$P(5 \leq S \leq 8) = p_S(5) + \cdots + p_S(8) = \frac{627}{1024} \simeq 0.6123.$$

On peut écrire $S = X_1 + \cdots + X_{10}$ où les $X_i = \begin{cases} 1 & \text{si Face} \\ 0 & \text{si Pile} \end{cases}$ sort au $i^{\text{ème}}$ lancer, de sorte que $\mu = E(X_i) = \frac{1}{2}$ et que $\mathrm{Var}(X_i) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$. Grâce au Théorème de la Limite Centrale, on sait que S suit approximativement une loi $\mathcal{N}(10\mu, 10\sigma^2) = \mathcal{N}(5, 5/2)$. Ce phénomène est illustré par la figure suivante.

Donc $\dfrac{S-5}{\sqrt{5/2}}$ suit approximativement une loi $\mathcal{N}(0,1)$ et en notant Z une variable aléatoire de loi $\mathcal{N}(0,1)$:

$$\begin{aligned} P(5 \leq S \leq 8) = P(4.5 \leq S \leq 8.5) &= P\left(\frac{4.5-5}{\sqrt{5/2}} \leq \frac{S-5}{\sqrt{5/2}} \leq \frac{8.5-5}{\sqrt{5/2}}\right) \\ &\simeq P(-0.32 \leq Z \leq 2.21) = \Phi(2.21) - \Phi(-0.32) = \Phi(2.21) - (1 - \Phi(0.32)) \\ &\simeq 0.9864 - (1 - 0.6255) = 0.6119 \end{aligned}$$

ce qui est très proche du résultat exact : 0.6123.
La première égalité $P(5 \leq S \leq 8) = P(4.5 \leq S \leq 8.5)$ s'appelle la *"correction pour la continuité" :* on déplace les bornes à mi-chemin entre l'événement étudié et son complément. Si on l'avait négligée, on aurait obtenu

$$\begin{aligned} P(5 \leq S \leq 8) &= P\left(\frac{5-5}{\sqrt{5/2}} \leq \frac{S-5}{\sqrt{5/2}} \leq \frac{8-5}{\sqrt{5/2}}\right) \\ &\simeq P(0 \leq Z \leq 1.90) = \Phi(1.90) - \Phi(0) = 0.4713 \end{aligned}$$

qui est une moins bonne approximation que la précédente.

Notons que si la variable aléatoire à approximer par une variable aléatoire normale est continue, on n'a pas besoin de la correction pour la continuité.

Exercices

1. Supposons que les poids des adultes (en kg) sont d'écart-type 12 kg. On prélève un échantillon de taille n pour estimer la moyenne inconnue μ de la population par la moyenne empirique $\overline{X}$. Quelle est la probabilité que l'écart entre $\overline{X}$ et μ soit supérieure à 5 kg si

a) $n = 12$ **b)** $n = 25$ **c)** $n = 35$ **d)** $n = 50$?

2. *Sondage :* On veut connaître la proportion p des gens qui, dans la population générale, sont en faveur d'une certaine proposition. Dans un échantillon de n personnes, on obtiendra X réponses favorables à la proposition en question. Notons $\hat{p} = X/n$ la proportion expérimentale des réponses favorables.

a) Si $n = 100$ et $p = 0.5$, déterminer $P(\hat{p} > 0.6)$.

b) Si $n = 100$ et $p = 0.4$, déterminer $P(\hat{p} > 0.5)$.

c) Si $n = 100$ et $p = 0.4$, déterminer approximativement c afin que $P(p - c < \hat{p} < p + c) \simeq 90\%$.

d) Si $n = 1000$ et $p = 0.4$, déterminer approximativement c afin que $P(p - c < \hat{p} < p + c) \simeq 90\%$.

3. Deux archers s'affrontent dans un concours de tir à l'arc. À chaque tir, Gaston a 50% de chance d'atteindre la cible. Légèrement plus habile, René atteint la cible avec une probabilité de 60%. Chacun tire 20 flèches. Calculer :

a) la probabilité que Gaston ait plus de 13 coups au but.

b) la probabilité que Gaston gagne le tournoi.

c) la probabilité que René gagne le tournoi.

d) la probabilité d'un match nul.

5

Estimation de la moyenne d'un grand echantillon

Un cas d'école. Dans un premier temps, on suppose que l'on observe un échantillon aléatoire $X_1, \ldots, X_n$ d'une loi (commune à $X_1, \ldots, X_n$) de la moyenne μ inconnue et de variance σ_o^2 connue. On cherche à **estimer** la moyenne μ à partir de l'observation $x_1, \ldots, x_n$ de notre échantillon. Si n est grand, la Loi des Grands Nombres nous permet d'affirmer qu'avec une grande probabilité μ n'est pas très éloigné de la moyenne empirique observée :

$$\mu \simeq \bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Bien sûr, μ ne dépend pas de notre observation (c'est un paramètre théorique que l'on cherche à estimer) et une autre expérience qui nous aurait amené à observer $X_1 = x_1', \ldots, X_n = x_n'$, nous amènerait à la conclusion $\mu \simeq \dfrac{x_1' + \cdots + x_n'}{n}$, de sorte qu'il est faux d'affirmer que μ vaut $\bar{x}$.
De manière à prendre en compte les fluctuations du hasard, nous allons estimer μ à l'aide d'un **intervalle de confiance** (une fourchette d'estimation). La technique mathématique repose sur le Théorème de la Limite Centrale qui énonce que si $X_1, \ldots, X_n$ est un échantillon d'une loi de moyenne μ et de variance σ_o^2, en posant

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

nous avons approximativement

$$Z_n := \frac{\overline{X} - \mu}{\sigma_o/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Ce résultat est faux si les $X_1, \ldots, X_n$ ne sont pas supposées indépendantes.
De ce fait, pour toute probabilité $(1-\alpha)$ $(0 \leq \alpha \leq 1)$, on peut trouver dans la Table I le nombre $z_{\frac{\alpha}{2}}$ tel que

$$\begin{aligned} P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\sigma_o/\sqrt{n}} \leq z_{\frac{\alpha}{2}}\right) &= P(-z_{\frac{\alpha}{2}} \leq Z_n \leq z_{\frac{\alpha}{2}}) \\ &\simeq P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha, \end{aligned}$$

où Z suit une loi normale $\mathcal{N}(0,1)$.

Des valeurs souvent utilisées sont

$$\begin{array}{ccccccccc} \alpha = 10\% & \rightarrow & 1-\alpha = 90\% & \text{et} & z_{\frac{\alpha}{2}} & = & z_{0,05} & \simeq & 1,645 \\ \alpha = 5\% & \rightarrow & 1-\alpha = 95\% & \text{et} & z_{\frac{\alpha}{2}} & = & z_{0,025} & \simeq & 1,960 \\ \alpha = 1\% & \rightarrow & 1-\alpha = 99\% & \text{et} & z_{\frac{\alpha}{2}} & = & z_{0,005} & \simeq & 2,576 \end{array}$$

Puisque $\alpha > 0$, les inégalités suivantes sont équivalentes

$$\begin{aligned} -z_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\sigma_o/\sqrt{n}} &\leq z_{\frac{\alpha}{2}} \\ -z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \leq \overline{X}-\mu &\leq z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \\ -\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \leq -\mu &\leq -\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \\ \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \geq \mu &\geq \overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \end{aligned}$$

Par conséquent

$$P(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}) \simeq P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha$$

ce qui s'écrit aussi

$$P([\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}] \ni p) \simeq 1-\alpha$$

et se traduit de la façon suivante. Avec une probabilité $(1-\alpha)$, la moyenne théorique μ se trouve dans l'intervalle aléatoire $[\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}, \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}]$. Une fois observé l'échantillon, la moyenne empirique $\bar{x}$ est connue.

Si la variance théorique $\sigma^2 = \sigma_o^2$ est aussi connue, alors l'intervalle observé

$$\left[\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}, \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma_o}{\sqrt{n}}\right]$$

*est un intervalle connu. C'est l'***intervalle de confiance** *pour μ avec* **coefficient de sécurité** $(1-\alpha)$.

Exemple 1. On suppose que des notes d'examen (sur 100), ont une loi de moyenne μ inconnue et d'écart-type $\sigma_o = 15$. Un échantillon de taille $n = 25$ est observé, on trouve $\bar{x} = 69,2$. Alors

$$\bar{x} \pm 1,645 \cdot \left(\frac{\sigma_o}{\sqrt{n}}\right) \quad \text{ou} \quad 69,2 \pm 1,645 \cdot \left(\frac{15}{\sqrt{25}}\right) \quad \text{ou} \quad [64.265, 74.135]$$

est un intervalle de confiance pour μ avec le coefficient de sécurité 90%.

Exemple 2. Soit $\bar{x}$ la moyenne empirique observée sur un échantillon de taille 16 d'une distribution (loi) normale $\mathcal{N}(\mu, 23.04)$. Un intervalle de confiance pour μ avec coefficient de sécurité 90% est

$$\left[\bar{x}-1.645 \cdot \sqrt{\frac{23.04}{16}}, \bar{x}+1.645 \cdot \sqrt{\frac{23.04}{16}}\right].$$

Pour une observation particulière de $\bar{x}$, cet intervalle contient ou ne contient pas la valeur inconnue μ. Toutefois, si un grand nombre de tels intervalles est (observé et) calculé, il reste vrai qu'à peu près 90% d'entre eux contiennent la moyenne μ.
Sur un ordinateur, 15 échantillons de taille16 d'une distribution (loi) normale $\mathcal{N}(5, 23.04)$ ont été simulés. Pour chacun de ces 15 échantillons, nous avons calculé l'intervalle de confiance pour μ avec coefficient de sécurité 90%, comme si la moyenne μ était inconnue. Sur la figure suivante sont représentés ces 15 intervalles : 13 d'entre eux (soit 86.7%) contiennent la moyenne $\mu = 5$.

Dans la pratique. Dans la pratique il n'y a aucune raison, si on ne connaît pas la moyenne μ, de connaître l'écart-type σ. Dans ce cas, l'intervalle de confiance obtenu plus haut, étant fonction de $\sigma = \sigma_o$, n'est pas accessible au calcul. Une fois de plus, c'est la grande taille n de l'échantillon qui va nous permettre de nous en sortir. En effet, la Loi des Grands Nombres nous permet d'estimer la variance σ^2 inconnue à l'aide des observations $X_1, \ldots, X_n$. Un estimateur naturel de la variance est la **variance empirique de l'échantillon**, déjà rencontré en Statistique Descriptive. Il est donné par

$$S^2 = \frac{1}{n-1}\Big[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\Big].$$

On note $s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$ sa valeur observée. De même, un estimateur naturel de

l'écart-type : l'**écart-type empirique de l'échantillon**, déjà rencontré en Statistique Descriptive est donné par

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\Big[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\Big]}.$$

On note $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}$ sa valeur observée. On peut montrer mathématiquement, que lorsque n est grand, l'écart-type empirique observé : s, est proche de l'écart-type théorique inconnu σ :

$$s \simeq \sigma.$$

Il est alors possible de remplacer dans la formule de l'intervalle de confiance trouvée plus haut, la valeur σ_o par la valeur observée : s, ce qui nous donne le résultat suivant.

Si les observations sont indépendantes et de même loi, l'intervalle observé

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$

*est l'***intervalle de confiance** *pour μ avec* **coefficient de sécurité** $(1-\alpha)$.
Important. En pratique, on considère que n est suffisamment grand, lorsque $n \geq 30$.

Ceci signifie à peu près, qu'avec une probabilité $1-\alpha$, l'intervalle de confiance $[\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}]$ contient la vraie valeur inconnue μ de la moyenne.

Nous terminons cette leçon en rappelant une formule bien pratique pour le calcul de la variance empirique

$$\begin{aligned} s^2 &= \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \\ &= \frac{(x_1)^2 + \cdots + (x_n)^2}{n-1} - \frac{n}{n-1}(\bar{x})^2. \end{aligned}$$

Par exemple, sur cinq observations $x_1 = 2.4,\ x_2 = 3.4,\ x_3 = 5.2,\ x_4 = -0.8,\ x_5 = 1.0$, on obtient $x_1 + \cdots + x_5 = 11.2$ et $(x_1)^2 + \cdots + (x_5)^2 = 46$. Ceci nous donne :
$\bar{x} = \frac{11.2}{5} = 2.24$,
$s^2 = \frac{(x_1)^2+\cdots+(x_5)^2}{4} - \frac{5}{4}(\bar{x})^2 = \frac{46}{4} - \frac{5}{4}(2.24)^2 = 5.228$ ou bien
$s^2 = \frac{(2.4-2.24)^2+(3.4-2.24)^2+(5.2-2.24)^2+(-0.8-2.24)^2+(1.0-2.24)^2}{4} = 5.228$, c'est-à-dire
$s = \sqrt{5.228} = 2.2865$.

Exercices

1. Un échantillon aléatoire de taille 28 nous donne $x_1+\cdots+x_{28} = 122.70$ ainsi que $x_1^2+\cdots+x_{28}^2 = 697.89$. Trouver des intervalles de confiance pour la moyenne avec le coefficient de sécurité

a) 99% **b)** 95% **c)** 90% **d)** 80%.

2. Trouver un intervalle de confiance pour μ avec coefficient de sécurité : 75%, pour les observations :

x_1	= 624	x_2	= 532	x_3	= 565	x_4	= 492
x_5	= 407	x_6	= 591	x_7	= 611	x_8	= 558
x_9	= 631	x_{10}	= 542	x_{11}	= 587	x_{12}	= 452
x_{13}	= 406	x_{14}	= 592	x_{15}	= 641	x_{16}	= 568
x_{17}	= 625	x_{18}	= 502	x_{19}	= 687	x_{20}	= 522

3. Une observation d'un échantillon de taille n nous donne $\bar{x} = 7.21$ et $s = 3.10$. On veut annoncer un intervalle de confiance pour μ avec coefficient de sécurité 99%. A partir de quelles valeurs de n, l'intervalle de confiance a-t'il une largeur inférieure à ± 0.1 ? Même question avec ± 0.01.

Réponse. On cherche n tel que : $\frac{z_{\frac{\alpha}{2}} \cdot s}{\sqrt{n}} \leq 0.1$. Soit $\sqrt{n} \geq \frac{z_{\frac{\alpha}{2}} \cdot s}{0.1}$. Donc, en élevant les deux membres de cette inégalité au carré : $n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{0.1} \right)^2$. Puisque $1 - \alpha = 99\%$, $\alpha/2$ vaut 0.5% et on lit dans la table I que $z_{\frac{\alpha}{2}} = 2.576$. Finalement, $n \geq \left(\frac{2,576 \cdot 3,10}{0,1} \right)^2 \simeq 6377$. Il faut donc un échantillon de taille au moins 6377 pour pouvoir annoncer un intervalle de confiance pour μ avec le coefficient de sécurité 99% et la précision ± 0.1.

Lorsqu'on cherche la précision ± 0.01, le même raisonnement nous amène à $n \geq \left(\frac{2,576 \cdot 3,10}{0,01} \right)^2 \simeq 637700$. Il faut donc un échantillon de taille au moins 637700 pour pouvoir annoncer un intervalle de confiance pour μ avec le coefficient de sécurité 99% et la précision ± 0.01.

6

Estimation d'une proportion

Nous cherchons à estimer la proportion d'une catégorie particulière d'individus (par exemple : sensibilité politique, chômeur, homme, femme, fumeur, salaire mensuel supérieur à 11000 francs, etc...) au sein d'une population totale (voir l'Exemple 2 de la Leçon 4, où cette question a déjà été abordée). Soit p ($0 \leq p \leq 1$) cette proportion qui nous est inconnue avec exactitude, à moins d'interroger toute la population. Pour l'estimer, nous tirons au hasard n individus dans la population totale, c'est-à-dire que nous effectuons n tirages indépendants et uniformes. On observe, dans cet échantillon, une proportion

$$\begin{aligned}\hat{p} &= \frac{\text{nombre d'individus dans l'échantillon appartenant à la catégorie étudiée}}{\text{taille de l'échantillon}} \\ &= \bar{x} = \frac{x_1 + \cdots + x_n}{n}\end{aligned}$$

où x_i est la réalisation d'une variable aléatoire X_i qui prend la valeur $X_i = 1$ si le $i^{\text{ème}}$ individu appartient à la catégorie étudiée ou la valeur $X_i = 0$ sinon. Puisque le tirage est uniforme, X_i suit une loi de Bernoulli de paramètre p, notée $\mathcal{B}(p)$ (voir la Leçon 1), où p est la vraie proportion à estimer. Puisque les tirages sont indépendants, $X_1, \ldots, X_n$ est un échantillon de taille n de la loi $\mathcal{B}(p)$. Lorsque n est grand, nous sommes dans les conditions d'application de la Loi des Grands Nombres qui affirme que l'observation $\hat{p} = \bar{x}$ est proche avec une grande probabilité de la moyenne théorique $\mu = E(X)$. Or, lorsque X suit une loi $\mathcal{B}(p)$, on a

$$E(X) = p \quad \text{et} \quad \text{Var}(X) = p(1-p).$$

Nous avons donc, lorsque n est grand, avec une grande probabilité :

$$\hat{p} \simeq p.$$

C'est-à-dire : la proportion observée sur l'échantillon est proche de la proportion de la catégorie considérée dans la population totale. Ce résultat est le principe de tous les sondages dont les médias sont si friands.

En fait, les résultats de la Leçon 5 nous permettent de donner un intervalle de confiance pour p. Nous savons que si $\mathrm{Var}(X) = \sigma_o$, $\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma_o}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma_o}{\sqrt{n}}\right]$ est un intervalle de confiance pour μ avec le coefficient de sécurité $(1-\alpha)$. Dans la situation présente, puisque $\sigma_o = \sqrt{p(1-p)}$, ceci signifie que $\left[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]$ est un intervalle de confiance pour p avec le coefficient de sécurité $(1-\alpha)$.

Malheureusement, les bornes de cet intervalle s'expriment à l'aide de la proportion p inconnue. Cet intervalle de confiance n'est donc pas calculable à l'aide de l'observation $\hat{p}$. Toutefois, nous avons vu que $p \simeq \hat{p}$, de sorte que $p(1-p) \simeq \hat{p}(1-\hat{p})$ et que l'intervalle $[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$ est proche du précédent. Par conséquent :

L'intervalle observé

$$\left[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

*est l'***intervalle de confiance** *pour la proportion p avec* **coefficient de sécurité** $(1-\alpha)$.
Important. Cet intervalle n'est valable que lorsque $n\hat{p} \geq 6$ et $n(1-\hat{p}) \geq 6$.

Exemple 1. Lors d'un sondage auprès de 500 personnes et portant sur leurs opinions politiques, 180 personnes se sont déclarées favorables au parti A. Estimer la proportion p des gens favorables au parti A au moyen d'un intervalle de confiance de coefficient de sécurité 90%.

Solution : On a $\hat{p} = 180/500 = 0.360$. Pour avoir $1-\alpha = 90\%$, il faut prendre $z_{\frac{\alpha}{2}} = 1.645$. Il ne reste plus qu'à employer la formule

$$\left(\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \left(0.360 \pm 1.645 \sqrt{\frac{0.36 \times 0.64}{500}}\right)$$
$$= (0.360 \pm 0.035) = [0.325, 0.395].$$

Remarque. Lorsqu'on estime un paramètre au moyen d'un intervalle de confiance, deux qualités espérées : **précision** et **sécurité**, sont en opposition. On ne peut améliorer l'une sans diminuer l'autre. Si l'on exige beaucoup de sécurité (risque α très petit), on obtiendra un intervalle de confiance plus large que si l'on se contente d'une sécurité plus raisonnable. Si l'on veut beaucoup de précision (intervalle étroit), il faudra "payer" cette précision par un risque d'erreur plus considérable. La seule façon d'obtenir à la fois une bonne précision et une grande sécurité est de ne pas lésiner sur la valeur de n, ce qui n'est pas toujours économique.

Exemple 2. Avec $n = 100$, on a obtenu $\hat{p} = 0.21$. Calculer les intervalles de confiance avec coefficient de sécurité 50%, 10%, 5%, 1% et 0.1% pour p.

Solution : Les cinq valeurs de α donnent des $z_{\frac{\alpha}{2}}$ qui valent respectivement : 0.674, 1.645, 1.960, 2.576 et 3.291. Les cinq intervalles de confiance sont présentés dans le tableau suivant.

$1-\alpha$	$z_{\frac{\alpha}{2}}$	**Intervalle de confiance**	**Longueur**
50%	0.674	$[0.18, 0.24]$	0.06
90%	1.645	$[0.14, 0.28]$	0.14
95%	1.960	$[0.13, 0.29]$	0.16
99%	2.576	$[0.11, 0.31]$	0.20
99.9%	3.291	$[0.08, 0.34]$	0.26

Lequel de ces cinq intervalles de confiance est le meilleur ? Assurément, un risque de $\alpha = 50\%$ est beaucoup trop fort et le premier intervalle n'est pas très satisfaisant. De même, un coefficient de sécurité de 99.9% paraît exagéré et rend l'intervalle de 30% plus large que celui obtenu avec $1 - \alpha = 99\%$. En général, on choisit α entre 1% et 10%, selon le contexte et l'importance relative de nos besoins en précision et en sécurité.

Exemple 3. Si l'on sait déjà que la valeur du paramètre p est voisine de 0.15, combien d'observations doit-on effectuer pour que l'intervalle de confiance de coefficient de sécurité 95% pour p soit de demi-longueur approximative 0.05 ? 0.02 ? 0.01 ?

Solution : Puisque $1 - \alpha = 95\%$, on doit prendre $z_{\frac{\alpha}{2}} = 1.960$. La demi-longueur : r, de l'intervalle de confiance sera donc $1.960\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. On ne sait pas à l'avance quelle sera la valeur de $\hat{p}$, mais on peut s'attendre à ce qu'il prenne une valeur voisine de p qu'on a supposé proche de 0.15. On a donc approximativement

$$r \simeq 1.960\sqrt{\frac{0.15 \times 0.85}{n}} = \frac{0.700}{\sqrt{n}}.$$

En exprimant n en fonction de r, on obtient $n \simeq 0.49/r^2$ et en donnant successivement à r les valeurs 0.05, 0.02 et 0.01 on obtient pour n les valeurs 196, 1 225 et 4 900.

En fait, il n'y a pas de raison en général pour supposé a priori que p est proche d'une valeur donnée à l'avance. C'est pourquoi, nous considérons le problème qui suit.

Exemple 4. Combien d'observations doit-on effectuer afin que, *quelle que soit la valeur de* p, l'intervalle de confiance de coefficient de sécurité 95% pour p soit de demi-longueur au plus 0.05 ? 0.03 ? 0.02 ? 0.01 ?

Solution : La demi-longueur de l'intervalle de confiance de coefficient de sécurité 95% est

$$1.960\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Or, la valeur maximale pour $\hat{p}(1-\hat{p})$ est $1/4$ (quand $\hat{p} = 1/2$). Quelle que soit la valeur de $\hat{p}$, la demi-longueur maximum de l'intervalle de confiance vaut $\ell_{\max} = 1.960/\sqrt{4n} = 0.98/\sqrt{n}$. Pour avoir $\ell \leq \ell_{\max}$, il faut prendre $n \geq (0.98/\ell_{\max})^2$. En donnant successivement à $\ell_{\max}$ les valeurs 0.05, 0.03, 0.02 et 0.01, on obtient $n \geq 385$, $n \geq 1068$, $n \geq 2401$ et $n \geq 9604$.

Exercices

1. Sur un échantillon tiré au hasard de 500 électeurs, 254 ont déclaré être favorables à une proposition gouvernementale et prévoient de voter oui pour cette proposition. Donner un intervalle de confiance, avec coefficient de sécurité 90%, pour la proportion p dans la population totale des électeurs favorables à cette proposition.

2. Un étudiant a pipé un dé en perçant des trous en deux points et en les remplissant d'un métal lourd. Pour estimer la probabilité p de sortir un quatre avec ce dé pipé, l'étudiant le lance 600 fois et observe un quatre 87 fois.

a) Construire un intervalle de confiance pour p avec coefficient de sécurité 90%.

b) Est-ce que l'étudiant a réussi a faire décroître la fréquence de sortie du quatre ?

c) Que se passe-t'il avec un coefficient de sécurité de 80% ?

3. Un grossiste en café souhaite savoir si une nouvelle marque plus chère a la préférence des consommateurs. Sur un échantillon de 90 consommateurs, 53 ont déclaré préférer la nouvelle marque. Donner un intervalle de confiance avec coefficient de sécurité 95%, pour le pourcentage des consommateurs qui préfèrent la nouvelle marque.

7

Du nouveau a Evry-la-Garenne ?

A la suite d'une enquête menée en 1980 à Évry-la-Garenne, il apparaissait que 50% des foyers avaient un revenu mensuel inférieur à 7.4 KFr (en franc constant). Une seconde enquête est menée en 1992 auprès de 10 foyers. Les revenus mensuels obtenus sont (en KFr) :

10.0 7.8 10.4 11.0 5.6 12.2 12.8 5.2 3.4 8.6

Peut-on affirmer que les revenus ont globalement augmenté depuis 1980 ?

Il ne s'agit plus d'estimer un paramètre inconnu (moyenne, proportion), mais de répondre par oui ou non à la question : *"Les revenus ont-ils augmenté dans l'ensemble ?"* Quelle que soit la réponse, elle sera susceptible d'être vraie ou fausse, dans la mesure où nous n'interrogeons pas tous les foyers d'Évry-la-Garenne. On peut même se douter qu'avec un aussi petit échantillon d'enquête, notre incertitude sera grande.

Notre but est, dans un premier temps, de préciser la question, et donc le type de réponse que nous allons y apporter et, dans un deuxième temps, de quantifier la probabilité de donner une réponse exacte.

Précision de la question. L'information que nous avons est que 50 % des foyers en 1980 avaient un revenu inférieur à 7.4. Nous allons donc essayer de savoir si la proportion des foyers dont le revenu est inférieur à 7.4 a diminué de 1980 à 1992. Pour cela, nous introduisons la notion mathématique de médiane d'une loi de variable aléatoire.

Soit X une variable aléatoire de densité f_X. Sa **médiane** *m est un nombre réel tel que $F_X(m) := P(X \leq m) = 50\%$ (voir la figure ci-dessus).*

Attention ! Il ne faut pas confondre médiane et moyenne. Par exemple,
si $f_X(x) = \begin{cases} \frac{1}{4} & \text{si } 0 \leq x \leq 1 \\ \frac{3}{4} & \text{si } 1 < x \leq 2 \\ 0 & \text{sinon} \end{cases}$, on a $F_X(t) = \begin{cases} 0 & \text{si } t \leq 0 \\ \frac{t}{4} & \text{si } 0 \leq t \leq 1 \\ \frac{1}{4} + \frac{3(t-1)}{4} & \text{si } 1 < t \leq 2 \\ 1 & \text{si } t \geq 2 \end{cases}$ et
$F_X(m) = 0.5 \Longleftrightarrow \dfrac{1}{4} + \dfrac{3(m-1)}{4} = \dfrac{1}{2} \Longleftrightarrow m = \dfrac{4}{3} = 1.333$. Alors que,
$\mu = E(X) = \displaystyle\int_0^1 \frac{x}{4}\, dx + \int_1^2 \frac{3x}{4}\, dx = \frac{5}{4} = 1.25.$

Revenons à Évry-la-Garenne, sa piscine et son terrain de camping. On note m la médiane de la répartition des revenus par foyer en 1992. Si rien n'a changé entre 1980 et 1992, alors m garde sa valeur de 1980, c'est-à-dire : $m = 7.4$. Si le revenu a globalement augmenté, alors $m > 7.4$ et dans la cas contraire $m < 7.4$. Le **test statistique** que nous allons construire va nous permettre de choisir parmi les deux hypothèses

$$\begin{array}{lcl} H_0 & : & m = 7.4 \\ H_1 & : & m > 7.4 \end{array}$$

laquelle a le plus chance d'être vraie, compte tenu des résultats de notre enquête.

Notons que nous supposons a priori que $m \geq 7.4$, c'est-à-dire que la médiane n'a pas pu décroître.

Une réponse statistique. Notons X le revenu d'un foyer tiré au hasard, ainsi que

$$p := P(X \leq 7.4).$$

Si H_0 est vraie, alors $p = \frac{1}{2}$. Si H_1 est vraie, alors $0 \leq p < \frac{1}{2}$.

Soient $X_1, \ldots, X_{10}$ les revenus des 10 foyers. Ces variables aléatoires sont indépendantes et de même loi inconnue. On considère les nouvelles variables aléatoires

$$Y_i = \begin{cases} 1 & \text{si } X_i \leq 7.4 \\ 0 & \text{si } X_i > 7.4 \end{cases}, \qquad i = 1, \ldots, 10.$$

Ce sont des variables aléatoires indépendantes qui suivent une loi de Bernoulli de paramètre $p = P(X \leq 7.4) : Y_i \sim \mathcal{B}(p),\ i = 1, \ldots, 10$. On en déduit que la variable aléatoire

$$U := Y_1 + \cdots + Y_{10} \sim \mathcal{B}(10, p)$$

suit une loi binômiale : $\mathcal{B}(10, p)$, où p est un paramètre inconnu (voir la Leçon 4, pour la loi binômiale).
En particulier,

$$\begin{array}{lll} \text{Si } H_0 \text{ est vraie, alors :} & U \sim \mathcal{B}(10, \frac{1}{2}). \\ \text{Si } H_1 \text{ est vraie, alors :} & U \sim \mathcal{B}(10, p), \quad 0 \leq p < \frac{1}{2}. \end{array}$$

En d'autres termes, sous H_0, le nombre de revenus inférieurs à 7.4 : U, a la même loi que le nombre de Pile en jouant 10 fois à Pile ou Face.

Si H_1 est vraie, on peut s'attendre à ce que la valeur observée : u, de U soit plus petite que les valeurs typiques de U sous H_0. Nous prenons donc une *règle de décision* de la forme suivante

$$\begin{array}{llll} \text{si on observe} & (u \leq c), & \text{alors :} & \text{on rejette } H_0 \quad (\textit{on accepte } H_1), \\ \text{si on observe} & (u \geq c+1), & \text{alors :} & \text{on ne rejette pas } H_0, \quad (\textit{on accepte } H_0) \end{array}$$

où c est un **seuil de décision** que nous allons déterminer en fonction du risque d'erreur que nous nous autorisons.

On cherche à "contrôler" la probabilité de se tromper en prenant notre décision.

Une première manière de se tromper est de prendre la décision de rejeter H_0, alors que H_0 est vraie. Avec notre règle de décision, cette erreur se produit lorsque, sous H_0, on observe l'évènement $(U \leq c)$. La probabilité d'une telle erreur est donc $P_{H_0}(U \leq c)$, c'est-à-dire la probabilité en jouant 10 fois à Pile ou Face d'observer Pile c fois ou moins.

Par exemple, si l'on choisit $c = 0, 1, 2$ ou 3, on lit dans la Table II de la loi binômiale, que

$$P_{H_0}(U \leq 0) = 0.0010, \quad P_{H_0}(U \leq 1) = 0.0107, \quad P_{H_0}(U \leq 2) = 0.0547, \quad P_{H_0}(U \leq 3) = 0.1719.$$

Une autre manière de se tromper est de prendre la décision de ne pas rejeter H_0, alors que H_1 est vraie. La probabilité d'une telle erreur est $P_{H_1}(U \geq c+1)$. Supposons que le paramètre inconnu p vaille effectivement $p = P_{H_1}(X \leq 7.4) = 0.30$. Dans ce cas, $U \sim \mathcal{B}(10, 0.3)$ et avec $c = 0, 1, 2$ ou 3, on lit dans la Table II que

$$P_{H_1}(U \leq 0) = 0.0282, \quad P_{H_1}(U \leq 1) = 0.1493, \quad P_{H_1}(U \leq 2) = 0.3828, \quad P_{H_1}(U \leq 3) = 0.6496,$$

d'où il vient les probabilités d'erreur correspondantes sont

$$\begin{array}{ll} P_{H_1}(U \geq 1) = 1 - 0.0282 = 0.9718 & P_{H_1}(U \geq 2) = 1 - 0.1493 = 0.8507 \\ P_{H_1}(U \geq 3) = 1 - 0.3828 = 0.6172 & P_{H_1}(U \geq 4) = 1 - 0.6496 = 0.3504. \end{array}$$

On peut faire un calcul analogue pour toutes les valeurs de p. On rassemble ces calculs pour $p = 0.1$ et $p = 0.3$ dans le tableau ci-dessous.

	$P_{H_0}(U \leq c)$	$P_{H_1}(U \geq c+1)$ $(p = 0.3)$	$P_{H_1}(U \geq c+1)$ $(p = 0.2)$	$P_{H_1}(U \geq c+1)$ $(p = 0.1)$
$\mathbf{c = 0}$	0.0010	0.9718	0.8926	0.6513
$\mathbf{c = 1}$	0.0107	0.8507	0.6242	0.2639
$\mathbf{c = 2}$	0.0547	0.6172	0.3222	0.0702
$\mathbf{c = 3}$	0.1719	0.3504	0.1209	0.0128

On se rend compte sur ce tableau, que si on cherche à rendre petit la probabilité d'erreur $P_{H_0}(U \leq c)$ en faisant décroître c, l'autre probabilité d'erreur $P_{H_1}(U \geq c+1)$ grandit. D'autre part, la probabilité

d'erreur $P_{H_1}(U \geq c+1)$ diminue à mesure que le paramètre p inconnu s'éloigne de $\frac{1}{2}$. Il semble, qu'un équilibre à peu près satisfaisant se trouve autour des valeurs de $P_{H_0}(U \leq c)$ proches de 5%.

On se donne une probabilité d'erreur α de l'ordre de 5% (par exemple $\alpha = 1\%$, 5% ou 10%) et on choisit une valeur entière c_α de c telle que $P_{H_0}(U \leq c_\alpha)$ soit proche de α et

$$P_{H_0}(U \leq c_\alpha) \leq \alpha.$$

Dans notre exemple, avec $\alpha = 6\%$, on choisit $c_\alpha = c_{0.06} = 2$. Notre **règle de décision** au **niveau** $\alpha = 6\%$ est donc :

si on observe $(u \leq 2)$, alors : on rejette H_0 (*on accepte* H_1),

si on observe $(u \geq 3)$, alors : on ne rejette pas H_0, (*on accepte* H_0)

Si on observe $u \leq 2$, on rejettera H_0 avec une probabilité inférieure à 6% de se tromper (par définition du niveau α).

Si on observe $u \leq 3$, on ne rejettera pas H_0 avec une probabilité de se tromper : $P_{H_1}(U \geq 3)$, qui dépend de la valeur de p. Nous l'avons calculée pour quelques valeurs de p, $0 \leq p < \frac{1}{2}$.

p	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05
$P_{H_1}(U \geq 3)$	0.9004	0.8327	0.7384	0.6172	0.4744	0.3222	0.1798	0.0702	0.0115

Ce qui nous donne la courbe

En prenant connaissance de cette courbe, les services sociaux d'Évry-la-Garenne ont décidé de mener une enquête plus sérieuse. Cette fois-ci, 100 foyers ont été consultés : sur ces 100 foyers, 35 ont un revenu inférieur à 7.4. Que conclure au niveau $\alpha = 5\%$?

On reprend la même démarche, mais cette fois-ci

$$U := Y_1 + Y_2 + \cdots + Y_{100} \sim \mathcal{B}(100, p) \quad \text{avec} \quad p = P(X \leq 7.4).$$

En particulier, sous H_0, $p = \frac{1}{2}$ et $U \sim \mathcal{B}(100, \frac{1}{2})$. On cherche $c = c_{0.05}$ tel que $P_{H_0}(U \leq c) \simeq 0.05$. L'approximation normale (voir la Leçon 4), nous permet de dire que U suit approximativement une loi $\mathcal{N}(100 \cdot \frac{1}{2}, 100 \cdot \frac{1}{2}(1 - \frac{1}{2})) = \mathcal{N}(50, 25)$, donc

$$\text{sous } H_0, \quad \frac{U - 50}{\sqrt{25}} = \frac{U - 50}{5} \sim \mathcal{N}(0, 1).$$

De ce fait, on prend c tel que

$$P_{H_0}(U \leq c+0.5) = P_{H_0}\left(\frac{U-50}{5} \leq \frac{c+0.5-50}{5}\right) = P\left(Z \leq \frac{c+0.5-50}{5}\right) = 0.05$$

où $Z \sim \mathcal{N}(0,1)$ et $+0.5$ est la correction de continuité. C'est-à-dire : $\frac{c-49.5}{5} = -1.645$, donc $c = 49,5 - 5\cdot(1,645) = 41,275$ qui n'est pas un entier, on lui préfère

$$c = 41$$

puisque $U \leq 41.275 \Longleftrightarrow U \leq 41$. La règle de décision au niveau 5% est

si on observe $(u \leq 41)$, alors : on rejette H_0 (*on accepte* H_1),

si on observe $(u \geq 42)$, alors : on ne rejette pas H_0.

Puisque nous avons observé $u = 35$, on rejette H_0 au niveau $\alpha = 5\%$.
Calculons maintenant les probabilités d'erreur

$$P_{H_1}(\text{ on accepte } H_0) = P_{H_1}(U \geq 42)$$

en fonction de $p = P(X \leq 7.4)$, $0 \leq p < \frac{1}{2}$. Puisque $U \sim \mathcal{B}(100,p)$, l'approximation normale nous permet d'avoir approximativement $U \sim \mathcal{N}(100\cdot p, 100\cdot p(1-p))$, d'où

$$\begin{aligned} P_{H_1}(U \geq 42) &= P_{H_1}\left(\frac{U-100\cdot p}{\sqrt{100\cdot p(1-p)}} \geq \frac{41.5-100\cdot p}{\sqrt{100\cdot p(1-p)}}\right) \\ &\simeq P\left(Z \geq \frac{41.5-100\cdot p}{10\cdot\sqrt{p(1-p)}}\right) = 1-\Phi\left(\frac{41.5-100\cdot p}{10\cdot\sqrt{p(1-p)}}\right) \end{aligned}$$

où $Z \sim \mathcal{N}(0,1)$ et Φ est sa fonction de répartition. Grâce à la Table I, on obtient

p	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05
$P_{H_1}(U \geq 42)$	0.7591	0.3797	0.0865	0.0060	$8\cdot10^{-5}$	10^{-5}	$\simeq 0$	$\simeq 0$	$\simeq 0$

Ce qui nous donne la courbe

On constate avec soulagement que les probabilités d'accepter H_0 à tort sont considérablement plus faibles avec une enquête menée auprès de 100 personnes, plutôt qu'auprès de 10.

La conclusion de notre test est la suivante : puisque nous avons observé 35 foyers sur 100 dont les revenus sont inférieurs à 7.4, compte tenu de notre règle de décision, nous acceptons H_1 au niveau $\alpha = 5\%$. Ceci signifie que la probabilité de nous tromper en rejetant H_0 est inférieure à 5%.

Si l'on avait observé 44 foyers sur 100 dont les revenus étaient inférieurs à 7.4, compte tenu de notre règle de décision, nous n'aurions pas rejeté H_0 au niveau $\alpha = 5\%$. Rapidement parlé, nous aurions accepté H_0. La probabilité de se tromper dans une telle situation dépend de la valeur du paramètre inconnu p : pour $p = 0.45$, elle est de 0.7591 ; pour $p = 0.40$, elle est de 0.3797 et pour $p = 0.35$, elle est de 0.0865. Par conséquent, la probabilité de cette erreur devient raisonnablement petite si p est plus petit que 0.35. Le problème est que lorsque $0.35 < p < 0.50$, on peut facilement ne pas rejeter H_0. En conclusion : **seul le rejet de $\mathbf{H_0}$ est significatif.**

8

Le test du signe

Le test que nous avons mis en place à la Leçon 7 s'appelle un test du signe. Nous le reprenons ici dans un cadre général. Ce test sert à décider si l'hypothèse $H_0 : m = m_o$ est vérifiée, où m est la médiane de la loi d'une variable aléatoire X continue et m_o est une valeur que l'on se donne ($m_o = 7.4$ dans l'exemple de la leçon précédente).

L'hypothèse H_0 est appelée l'**hypothèse nulle**. On peut considérer les trois **hypothèses alternatives** : $H_1 : m > m_o$, $H_1 : m < m_o$ ou bien $H_1 : m \neq m_o$. Chacune correspond à des règles de décision de forme différente. Considérons pour le moment, l'hypothèse alternative

$$H_1 : m > m_o,$$

comme dans l'exemple d'Évry-la-Garenne.

On considère un échantillon statistique de taille n : $X_1, \ldots, X_n$, de la loi inconnue d'une variable aléatoire X. A chacun des $X_i, i = 1, \ldots, n$, on associe le nombre

$$Y_i = \begin{cases} 1 & \text{si } X_i < m_o \\ 0 & \text{si } X_i \geq m_o \end{cases}$$

de sorte que

$$U = Y_1 + \cdots Y_n$$

est le nombre (aléatoire) des $X_i,\ i = 1, \ldots, n$ qui sont *plus petits* que m_o.

Les observations que l'on obtient sont $x_1, \ldots, x_n$, et on calcule le nombre $u = y_1 + \cdots + y_n$ des $x_i,\ i = 1, \ldots, n$ qui sont plus petits que m_o. Notre <u>règle de décision</u> est

si on observe $(u \leq c_\alpha)$, alors : on rejette H_0 (*on accepte* H_1),

si on observe $(u \geq c_\alpha + 1)$, alors : on ne rejette pas H_0,

où α est le <u>niveau du test</u> que nous nous imposons ($\alpha = 1\%$, 5%, 10%, etc...), c'est-à-dire la probabilité de rejeter H_0 à tort :

$$P_{H_0}(\text{rejeter } H_0) \simeq \alpha.$$

En d'autres termes, on calcule le **seuil** c_α de sorte que

$$P_{H_0}(U \leq c_\alpha) \simeq \alpha.$$

Ce calcule est basé sur la constatation qu'en notant

$$p = P(X < m_o)$$

le paramètre inconnu du problème, la variable aléatoire U suit une loi binômiale :

$$U \sim \mathcal{B}(n, p).$$

En particulier, sous H_0, par définition de la médiane $m = m_o$, nous avons $p = \frac{1}{2}$ et

$$\text{sous } H_0, \quad U \sim \mathcal{B}(n, \frac{1}{2}).$$

Si n est petit, on obtient c_α dans la Table II de la loi $\mathcal{B}(n, \frac{1}{2})$.
Si n est grand (n plus grand que 15), l'approximation normale est très bonne. Par conséquent, nous avons approximativement

$$\text{sous } H_0, \quad U \sim \mathcal{N}(\frac{n}{2}, \frac{n}{4});$$

ce qui nous permet de calculer

$$P_{H_0}(U \leq c) = Po(U \leq c + \frac{1}{2}) = P\left(\frac{U - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \leq \frac{(c + \frac{1}{2}) - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right) \simeq \Phi\left(\frac{2c + 1 - n}{\sqrt{n}}\right)$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Comme d'habitude, on note z_α le nombre qui satisfait

$$\Phi(z_\alpha) = 1 - \alpha.$$

On rappelle que pour $\alpha = 2.5\% : z_\alpha = z_{0.025} = 1.960$, pour $\alpha = 5\% : z_\alpha = z_{0.05} = 1.645$ et pour $\alpha = 10\% : z_\alpha = z_{0.10} = 1.282$.
Il satisfait aussi $\Phi(-z_\alpha) = \alpha$, puisque Φ est symétrique par rapport à zéro. De ce fait, la définition de $c_\alpha : P_{H_0}(U \leq c_\alpha) \simeq \alpha$ s'écrit aussi $\Phi\left(\frac{2c_\alpha + 1 - n}{\sqrt{n}}\right) \simeq \Phi(-z_\alpha)$. Donc, c_α est solution de l'équation $\frac{2c_\alpha + 1 - n}{\sqrt{n}} \simeq -z_\alpha$, soit

$$c_\alpha \simeq \frac{n}{2} - \frac{z_\alpha \sqrt{n} + 1}{2}.$$

Plus exactement, c_α est le plus grand entier inférieur à $\frac{n}{2} - \frac{z_\alpha \sqrt{n} + 1}{2}$.

Si l'on teste $H_0 : m = m_o$ contre

$$H_1 : m < m_o,$$

à chacun des $X_i, i = 1, \ldots, n$, on associe le nombre $Z_i = \begin{cases} 1 & \text{si } X_i > m_o \\ 0 & \text{si } X_i \leq m_o \end{cases}$ de sorte que $V = Z_1 + \cdots Z_n$ est le nombre (aléatoire) des $X_i,\ i = 1, \ldots, n$ qui sont *plus grands* que m_o.

Les observations que l'on obtient sont $x_1, \ldots, x_n$, et on calcule le nombre $v = z_1 + \cdots + z_n$ des $x_i,\ i = 1, \ldots, n$ qui sont plus grands que m_o. Notre règle de décision est

si on observe $(v \leq c_\alpha)$, alors : on rejette H_0 (*on accepte* H_1),

si on observe $(v \geq c_\alpha + 1)$, alors : on ne rejette pas H_0,

où α est le niveau du test et c_α est calculé comme précédemment.

Exemple 1. Soit X l'intervalle de temps en secondes entre deux appels téléphoniques à un standard. On teste $H_0 : m = 6.2$ contre $H_1 : m < 6.2$. L'observation d'un échantillon de taille $n = 8$ nous donne

$$6.8 \quad 5.7 \quad 6.9 \quad 5.3 \quad 4.1 \quad 3.8 \quad 1.7 \quad 6.0$$

On commence par "construire le test", c'est-à-dire par calculer la règle de décision en fonction du niveau désiré.
On s'impose le niveau $\alpha = 5\%$. Si V désigne le nombre aléatoire de valeurs de l'échantillon qui dépassent 6.2, sous H_0, V suit la loi $\mathcal{B}(8, \frac{1}{2})$ et la lecture de la Table II de $\mathcal{B}(8, \frac{1}{2})$ nous donne

$$P_{H_0}(V \leq 0) = 0.0039, \quad P_{H_0}(V \leq 1) = 0.0352, \quad P_{H_0}(V \leq 2) = 0.1445.$$

Par conséquent $c_{0.05} = 1$. Notre règle de décision au niveau 5% est donc :

si on observe $(v \leq 1)$, alors : on rejette H_0 (*on accepte* H_1),

si on observe $(v \geq 2)$, alors : on ne rejette pas H_0.

Puisqu'on observe $v = 2$ valeurs supérieures à 6.2, on ne rejette pas H_0 au niveau 5%.
C'est seulement pour des niveaux $\alpha \geq 14.45\%$ que l'on rejette H_0, à partir de nos observations.

Exemple 2. Pour tester les performances comparées de deux balles de golf de marque A et B, on demande à 6 joueurs expérimentés de frapper ces balles (3 frappent A avant B et 3 frappent B avant A). Pour chaque joueur, on note les longueurs L_A et L_B des trajectoires des deux balles.

Golfeur	$\mathbf{L_A}$	$\mathbf{L_B}$	$\mathbf{sgn(L_A - L_B)}$
1	265	252	+
2	272	276	−
3	246	243	+
4	260	246	+
5	274	275	−
6	263	246	+

Quelle est la meilleure balle ?

Avant tout, il convient de constater que les observations ne sont pas indépendantes. En effet, les deux longueurs L_{Ai} et L_{Bi} provenant d'un même joueur i sont corrélées. Par contre, les couples $(L_A, L_B)_i,\ i = 1, \ldots, 6$ sont indépendants les un des autres. En particulier, les différences

$D_i := (L_A - L_B)_i,\ i = 1, \ldots, 6$ sont indépendantes les unes des autres. On dit que les observations sont **appariées**.

Pour répondre à notre question, il faut se demander ce que sont les hypothèse nulle H_0 et alternative H_1. En notant m la médiane de la loi de $D := L_A - L_B$, on peut penser à $H_0 : m > 0$ contre $H_1 : m < 0$. Mais cela présuppose qu'il y a nécessairement une balle effectivement meilleure que l'autre, puisque la possibilité $m = 0$ n'est pas prise en compte. De plus, nous n'avons étudié que des hypothèses nulles de la forme simple $H_0 : m = m_o$, alors que $H_0 : m > 0$ est une hypothèse plus complexe (dite multiple). En fait, il faudrait pouvoir faire un test des trois hypothéses $H_0 : m = 0$, $H_1 : m > 0$ et $H_1' : m < 0$. Ce qui est assez délicat. Nous n'aborderons pas cette question, mais nous allons tester

$$H_0 : m = 0 \quad \text{contre} \quad H_1 : m \neq 0,$$

pour savoir s'il existe une différence significative entre les comportements des deux balles.

On est en présence d'un échantillon statistique de taille n (ici $n = 6$), de variables appariées $(X_i, Y_i), i = 1, \ldots, n$. On cherche à savoir si $H_0 : P(X < Y) = \frac{1}{2}$ ou bien $H_1 : P(X < Y) \neq \frac{1}{2}$. Pour cela on regarde les nouvelles variables aléatoires

$$D_i = X_i - Y_i, i = 1, \ldots, n.$$

Elles forment un échantillon de la loi de $D = X - Y$, de médiane m et les hypothèses du test se réécrivent

$$H_0 : m = 0 \quad \text{et} \quad H_1 : m \neq 0.$$

A chacun des $D_i, i = 1, \ldots, n$, on associe le nombre

$$Y_i = \begin{cases} 1 & \text{si } D_i < 0 \\ 0 & \text{si } D_i \geq 0 \end{cases}$$

de sorte que

$$U = Y_1 + \cdots Y_n$$

est le nombre (aléatoire) des $D_i,\ i = 1, \ldots, n$ qui sont *plus petits* que 0.

Les observations que l'on obtient sont $d_1, \ldots, d_n$, et on calcule le nombre $u = y_1 + \cdots + y_n$ des $d_i,\ i = 1, \ldots, n$ qui sont plus petits que 0. Notre règle de décision est

si on observe $\quad (u \leq c_{\frac{\alpha}{2}})$ ou $(u \geq n - c_{\frac{\alpha}{2}})$, $\quad$ alors : $\quad$ on rejette H_0,

si on observe $\quad (c_{\frac{\alpha}{2}} + 1 \leq u \leq n - c_{\frac{\alpha}{2}} - 1)$, $\quad$ alors : $\quad$ on ne rejette pas H_0,

où α est le niveau du test que nous nous imposons, c'est-à-dire la probabilité de rejeter H_0 à tort : $P_{H_0}(\text{rejeter } H_0) \simeq \alpha$ et $c_{\frac{\alpha}{2}}$ se calcule comme c_α (mais en remplaçant α par $\frac{\alpha}{2}$). En particulier, lorsque n est grand, nous avons

$$c_{\frac{\alpha}{2}} \simeq \frac{n}{2} - \frac{z_{\frac{\alpha}{2}}\sqrt{n} + 1}{2}.$$

Plus exactement, $c_{\frac{\alpha}{2}}$ est le plus grand entier inférieur à $\frac{n}{2} - \frac{z_{\frac{\alpha}{2}}\sqrt{n}+1}{2}$.

On rappelle que pour $\alpha = 5\% : z_{\frac{\alpha}{2}} = z_{0.025} = 1.960$, pour $\alpha = 10\% : z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$ et pour $\alpha = 20\% : z_{\frac{\alpha}{2}} = z_{0.10} = 1.282$.

La forme de cette règle de décision est basée sur la remarque de bon sens suivante : si $m = 0$, alors, il y autant de chance pour que la variable aléatoire D soit positive ou négative. Donc les valeurs typique de U (sous H_0) se situent autour de $\frac{n}{2}$. On rejettera H_0 si l'on observe une quantité u de valeurs négatives, significativement éloignée de $\frac{n}{2}$. Notons que ce test est symétrique : on rejette H_0 si l'on observe une quantité $v = n - u$ de valeurs positives, significativement éloignée de $\frac{n}{2}$. De plus, puisque $v + u = n$, on a

$$(u \leq c_{\frac{\alpha}{2}}) \text{ ou } (u \geq n - c_{\frac{\alpha}{2}}) \Longleftrightarrow (v \leq c_{\frac{\alpha}{2}}) \text{ ou } (v \geq n - c_{\frac{\alpha}{2}}) \quad \text{et}$$
$$(c_{\frac{\alpha}{2}} + 1 \leq u \leq n - c_{\frac{\alpha}{2}} - 1) \Longleftrightarrow (c_{\frac{\alpha}{2}} + 1 \leq v \leq n - c_{\frac{\alpha}{2}} - 1),$$

et la règle de décision est inchangée si l'on remplace u par v.

Appliquons ceci au test des balles de golf. La Table II de la loi $\mathcal{B}(6, \frac{1}{2})$ nous indique que

$$P_{H_0}(U \leq 0) = 0.0156, \quad P_{H_0}(U \leq 1) = 0.1094 \quad \text{et} \quad P_{H_0}(U \leq 2) = 0.3438.$$

Avec $\alpha = 5\%$, nous avons $c_{\frac{\alpha}{2}} = c_{0.025} = 0$. D'ailleurs, même avec un niveau de 20%, nous prenons encore $c_{0.10} = 0$. C'est-à-dire qu'avec ce niveau, on ne rejette H_0, que lorsque <u>toutes</u> les observations de $L_A - L_B$ sont positives ou bien <u>toutes</u> les observations de $L_A - L_B$ sont négatives.

On a obtenu $u = 2$ observations de $L_A - L_B$ négatives. Donc on ne rejette pas H_0 aux niveaux 5% et même 20% : *il n'y a pas de différence significative de comportement entre les deux balles* à ces niveaux de test.

Puisque $P_{H_0}(U \leq 2) = 0.3438$, on ne rejette H_0 avec nos observations qu'en prenant un niveau $\alpha \geq 2 \times 0.3438 = 0.6876$. Ce qui n'est pas raisonnable.

Exercices

1. Pour cet ensemble de données provenant d'un échantillon, tester $H_0 : m = 4.8$ contre $H_1 : m \neq 4.8$. On faira usage d'un niveau de confiance approximativement égal à 10%.

1.0	10.3	16.7	38.4	2.4
2.6	8.9	36.3	27.1	3.8
1.9	0.9	0.4	9.2	3.0

2. Une enquête est menée auprès de 514 paires de frères (non jumeaux). Il apparaît que pour 273 de ces paires, l'ainé a atteint un niveau d'étude plus élevé que le cadet. Y-a-t'il un effet de l'ordre de naissance sur la réussite dans les études ? Faire des tests de niveaux 5 et 10%.

3. Dans une expérience pédagogique à l'école primaire, 14 paires d'enfants sont choisies de façon à avoir, par paire, les mêmes capacités et le même milieu. On enseigne à lire à l'un d'eux par la méthode globale et à l'autre par la méthode analytique. On obtient les notes suivantes

Globale	66	69	70	62	64	62	72	76	78	64	73	80	67	74
Analytique	64	68	69	60	66	61	70	75	72	65	70	78	68	72

Y-a-t-il une différence de résultats entre les deux méthodes ?

4. On effectue sur 10 personnes deux numérations globulaires à deux dates différentes. Les résultats obtenus indiquent le nombre de globules rouges par mm^3, divisé par 100 000.

15 Janvier	:	46	42	51	42	40	54	49	46	47	47
2 Septembre	:	47	47	44	45	54	50	48	48	45	55

Y a-t-il évolution de la formule sanguine ?

5. Onze individus ont été traité avec le soporifique S et un produit inactif I. Pour chacun des 11 sujets, le temps de sommeil moyen après traitement a été enregistré. On a observé (en minutes)

Individu	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
S	560	470	580	570	550	480	460	540	620	550	620
I	590	530	430	360	430	570	490	480	380	400	350

Ces résultats permettent-ils d'affirmer que le soporifique S est efficace ?

6. 80 rats sont répartis en 40 paires d'individus de même poids. Dans chaque paire un rat est soumis à un régime A, l'autre à un régime B. 28 des rats A pèsent plus lourd que leurs compagnons. Les deux régimes sont-ils équivalents ?

7. On souhaite comparer deux médicaments sensés soulager la douleur post-opératoire. On a observé sur 16 patients dont 8 ont pris un médicament A habituel et les 8 autres un médicament B expérimental, les nombres suivants d'heures de soulagement

A	6,8	3,1	5,8	4,5	3,3	4,7	4,2	4,9
B	4,4	2,5	2,8	2,1	6,6	0,0	4,8	2,3

Que pensez-vous de la mise en place d'un test de l'existence d'une différence entre A et B ?

9

Le test du Khi-Deux d'ajustement

Le khi-2 (χ^2) est un test simple basé sur les différences entre effectifs observés et effectifs théoriques. Testons l'hypothèse nulle H_0 suivante : les naissances en Suède se répartissent uniformément tout au long de l'année. On dispose pour cela d'un échantillon observé de 88 naissances, groupées selon des saisons de longueurs variables : Printemps (avril-juin ; 91 jours), Eté (juillet-août ; 62 jours), Automne (septembre-octobre ; 61 jours), Hiver (novembre-mars ; 151 jours). Nous avons observé 26 naissances au printemps, ainsi que 21, 7 et 34 naissances en été, automne et hiver respectivement.

Sous H_0, on attend théoriquement un nombre de naissances proportionnel à la durée de la saison, c'est-à dire $88 \times \frac{91}{365} = 21.94$ naissances au printemps, ainsi que $88 \times \frac{62}{365} = 14.95$, $88 \times \frac{61}{365} = 14.71$ et $88 \times \frac{151}{365} = 36.40$ naissances en été, automne et hiver respectivement. Soit le tableau :

Saison	**Effectif observé**	**Effectif attendu sous $\mathbf{H}_o$**
Printemps : 1	26	21,94
Eté : 2	21	14,95
Automne : 3	7	14,71
Hiver : 4	34	36,40
Total	88	88

Faisons correspondre les indices 1, 2, 3 et 4 aux saisons : printemps, été, automne et hiver respectivement. On note $O_1 = 26$, $O_2 = 20$, $O_3 = 8$ et $O_4 = 34$ les effectifs observés correspondants, ainsi que $T_1 = 21,94$, $T_2 = 14,95$, $T_3 = 14,71$ et $T_4 = 36,40$ les effectifs attendus sous H_0 correspondants.

Une mesure de la distance entre les effectifs observés et théoriques (attendus sous H_0) devra prendre en compte les écarts $O_1 - T_1, \ldots, O_4 - T_4$. Pour avoir une idée de la taille globale de la distance, il ne sert à rien de faire la somme des écarts puisque : $(O_1 - T_1) + \cdots + (O_4 - T_4) = (O_1 + \cdots O_4) - (T_1 + \cdots T_4) = 88 - 88 = 0$. On résoud le problème en élevant au carré chaque écart : $(O - T)^2$. Puis pour prendre en compte son importance relative en considèrant $\frac{(O-T)^2}{T}$. Finalement, pour la distance entre les effectifs observés et attendus sous H_0, on prend la somme

de la contribution de toutes les classes :

$$\begin{aligned}\chi^2 &= \frac{(O_1 - T_1)^2}{T_1} + \frac{(O_2 - T_2)^2}{T_2} + \frac{(O_3 - T_3)^2}{T_3} + \frac{(O_4 - T_4)^2}{T_4} \\ &= \frac{(26 - 21,94)^2}{21,94} + \frac{(21 - 14,95)^2}{14,95} + \frac{(7 - 14,71)^2}{14,71} + \frac{(34 - 36,40)^2}{36,40} = 7,39\end{aligned}$$

Un χ^2 est positif et il ne vaut zéro que si les effectifs attendus sous H_0 et observés coïncident. Il sera d'autant plus grand que les écarts entre effectifs attendus sous H_0 et observés sont importants. Par conséquent, on aura tendance à rejeter H_0 lorsque la distance χ^2 observée : χ^2_{obs}, sera grande. La régle de décision sera de la forme

$$\text{rejeter } H_0 \text{ si } \chi^2_{obs} > c_\alpha$$

où c_α est une constante à déterminer selon le niveau α désiré.

Dans le cas présent, il y a 4 classes et on dira qu'il y a $4 - 1 = 3$ degrés de liberté. Le seuil c_α se lit dans une table du khi-2 à 3 degrés de liberté. On lit dans la table que $I\!P(\chi^2_3 > 7,875) = 1 - I\!P(\chi^2_3 \leq 7,815) = 1 - 0,95 = 0,05$ et que $I\!P(\chi^2_3 > 6,251) = 1 - I\!P(\chi^2_3 \leq 6,251) = 1 - 0,90 = 0,10$. Au niveau $\alpha = 0,05$, on prend donc $c_{0,05} = 7,815$ et au niveau $\alpha = 0,10$, on prend $c_{0,10} = 6,251$. On constate que notre distance observée χ^2_{obs} satisfait $6,251 < \chi^2_{obs} = 7,39 < 7,815$, par conséquent on rejette H_0 au niveau 10% et on accepte H_0 au niveau 5%.

De façon générale, soient r classes numérotées $1, 2, \ldots, r$. Elles sont représentées dans la population selon certaines proportions inconnues $p_1, \ldots, p_r$ respectivement (on a $p_1 + \cdots + p_r = 1$). On cherche à tester

$$H_0 : p_1 = \pi_1, p_2 = \pi_2, \ldots, p_r = \pi_r,$$

où $\pi_1, \ldots, \pi_r$ sont des proportions données telles que $\pi_1 + \cdots + \pi_r = 1$. Si on observe n individus, les effectifs attendus sous H_0 sont $T_i = n\pi_i$ pour les classes $i = 1, \ldots, r$ et le tableau des observations prend la forme suivante :

Classe	**Effectif observé**	**Effectif attendu sous $\mathbf{H}_o$**
1	O_1	$T_1 = n\pi_1$
2	O_2	$T_2 = n\pi_2$
$\vdots$	$\vdots$	$\vdots$
r	O_r	$T_r = n\pi_r$
Total	n	n

Dans l'exemple précédent, nous avions $r = 4$, $n = 88$, $\pi_1 = 91/365$, $\pi_2 = 62/365$, $\pi_3 = 61/365 =$ et $\pi_4 = 151/365$. La distance du χ^2 est donnée par

$$\chi^2 = \frac{(O_1 - T_1)^2}{T_1} + \cdots + \frac{(O_r - T_r)^2}{T_r} \tag{9.1}$$

que l'on note rapidement

$$\chi^2 = \sum \frac{(O-T)^2}{T}, \tag{9.2}$$

où la lettre grecque Σ (sigma) signifie "somme". Le nombre de degrés de liberté est

$$\text{d.d.l.} = r - 1, \tag{9.3}$$

ce qui signifie que l'on doit déterminer le seuil c_α à l'aide de la table de la loi du khi-2 à $(r-1)$ degrés de liberté :

$$I\!P(\chi^2_{r-1} > c_\alpha) = \alpha.$$

Exemple 1. Le Bureau de la statistique du gouvernement du Québec a dénombré 84 579 nouveau-nés dans la province en 1986. De ce nombre, 43 220 étaient des garçons et 41 359 des filles. En supposant que le sexe de nouveau-nés est déterminé au hasard (hypothèse H_0), on se serait attendu à avoir $84579 \times \frac{1}{2} = 42289,5$ garçons et autant de filles. On trouve

$$\begin{aligned}\chi^2_{obs} &= \frac{(43220 - 42289,5)^2}{42289,5} + \frac{(41359 - 42289,5)^2}{42289,5} \\ &= 40,95.\end{aligned}$$

On a $r = 2$, donc d.d.l. $= 1$, comme $I\!P(\chi^2_1 > 6,635) = 0,01$ et $40,95 > 6,635$, on rejette l'hypothèse H_0 avec un niveau de 1%.

Avec le même niveau, on ne rejette pas l'hypothèse nulle de 51% de garçons et de 49% de filles qui donnent des effectifs théoriques (attendus sous H_0) de $84579 \times 0,51 = 43135,29$ garçons et $84579 \times 0,49 = 41443,71$ filles, car alors

$$\begin{aligned}\chi^2_{obs} &= \frac{(43220 - 43135,29)^2}{43135,29} + \frac{(41359 - 41443,71)^2}{41443,71} \\ &= 0,34 \not> 6,635.\end{aligned}$$

Exemple 2. Voici les résultats obtenus par Mendel à la suite de croisements de pois hybrides quant à la forme (lisse ou ridée) et à la couleur :

Graines	**Jaunes**	**Vertes**	**Total**
Lisses	315	108	423
Ridées	101	32	133
Total	416	140	556

On veut tester l'hypothèse de la ségrégation mendélienne et de la recombinaison libre qui correspond à $H_0 : \pi(LJ) = 9/16, \pi(LV) = 3/16, \pi(RJ) = 3/16, \pi(RV) = 1/16$. Le tableau des effectifs théoriques sous H_0 est le suivant

Graines	**Jaunes**	**Vertes**	**Total**
Lisses	312,75	104,25	423
Ridées	104,25	34,75	133
Total	416	140	556

En effet, $556 \times \frac{9}{16} = 312,75$; $556 \times \frac{3}{16} = 104,25$ et $556 \times \frac{1}{16} = 34,75$.

On obtient $\chi^2_{obs} = \frac{(315-312,75)^2}{312,75} + \frac{(108-104,25)^2}{104,25} + \frac{(101-104,25)^2}{104,25} + \frac{(32-34,75)^2}{34,75} = 0,47$. Le nombre de degrés deliberté est $4-1=3$. Or, on a $I\!P(\chi^2_3 > 0,45) = 80\%$ et $I\!P(\chi^2_3 > 0,71) = 70\%$, de sorte qu'on accepte H_0 au niveau 70% et donc à tous les niveaux inférieurs.

Exemple 3. Le tableau suivant donne les effectifs de pois selon la couleur des fleurs (Pourpre ou Vermillon) et la forme du pollen (Allongé ou Rond) obtenus par Bateson en 1909 en croisant des pois hybrides. On veut tester l'hypothèse de la ségrégation mendélienne et de la recombinaison libre qui correspond à $H_0 : \pi(PA) = 9/16, \pi(PR) = 3/16, \pi(VA) = 3/16, \pi(VR) = 1/16$.

Classe	**Effectif observé**	**Effectif attendu sous $\mathbf{H}_o$**
PA	1528	$2132 \times 9/16 = 1199,25$
PR	106	$2132 \times 3/16 = 399,75$
VA	117	$2132 \times 9/16 = 399,75$
VR	381	$2132 \times 9/16 = 133,25$
Total	2 132	2 132

On trouve alors $\chi^2_{obs} = \frac{(1528-1199,25)^2}{1199,25} + \frac{(106-399,75)^2}{399,75} + \frac{(117-399,75)^2}{399,75} + \frac{(381-133,25)^2}{133,25} = 966,61$ et $I\!P(\chi^2_3 > 11,3) = 0,01$. On rejette donc l'hypothèse au niveau 1%.

Une règle de validité des tests du khi-2 est que les effectifs théoriques par classe soient tous supérieurs ou égaux à 5. Si ça n'est pas le cas, on regroupe certaines classes.

Exemple 4. D'après le document *Current Housing Reports* publié par le U.S. Bureau of the Census, la distribution des modes de chauffage de maison est

Chauffage	Gaz	Fuel	Electricité	LPG	Bois	Autre
Pourcentage	56,7	14,3	16,0	4,5	6,7	1,8

On a sélectionné au hasard 200 maisons construites après 1974. Nos observations donnent

Chauffage	Gaz	Fuel	Electricité	LPG	Bois	Autre
Fréquence	91	16	110	14	17	2

Peut-on au vu de cet échantillon conclure que la distribution du mode de chauffage des maisons construites après 1974 diffère de la distribution de l'ensemble des maisons américaines ? On prendra $\alpha = 0,05$.

Il y a 6 classes dans cette expérience statistique. Mais, on constate que l'effectif théorique de la classe"Autre" est $2OO \cdot 1,8\% = 3,6 < 5$, on doit donc la regrouper avec une autre. On prend une classe peu représentée, par exemple "Bois", et on crée la classe "Bois et autre". On a maintenant

$r = 5$ classes.

Classe	**Effectif observé**	**Effectif attendu sous $\mathbf{H}_o$**
Gaz	91	$200 \times 0,567 = 113,4$
Fuel	16	$200 \times 0,143 = 28,6$
Electricité	110	$200 \times 0,160 = 32$
LPG	14	$200 \times 0,045 = 9$
Bois et autre	19	$200 \times 0,085 = 17$
Total	200	200

On obtient $\chi^2_{obs} = \frac{(91-113,4)^2}{113,4} + \frac{(16-28,6)^2}{28,6} + \frac{(110-32)^2}{32} + \frac{(14-9)^2}{9} + \frac{(19-17)^2}{17} \geq \frac{(110-32)^2}{32} = 190,125$ qui est supérieur à $13,28$: seuil de niveau 1% pour la loi du khi-2 à $5-1=4$ degrés de liberté. On rejette donc, au niveau 1%, l'hypothèse H_0 de conservation du mode de chauffage domestique avant et après 1974. On la rejette donc à plus forte raison au niveau $\alpha = 0,05$.

Exercices

1. La distribution de 300 accouchements selon les jours de la semaine est donnée par le tableau de données suivant :

Jour	L	Ma	Me	J	V	S	D	Total
Effectif	50	42	47	42	44	40	35	300

Un administrateur d'hôpital vous demande de vérifier si les accouchements se répartissent uniformément. Répondez lui à l'aide d'un test de niveau 10%.

2. Dans une étude célèbre, des données ont été prélevées sur 6587 suicides en France. Voici la distribution des suicides selon le jour de la semaine :

Jour	L	Ma	Me	J	V	S	D	Total
Effectif	1001	1035	982	1033	905	737	894	6587

Tester au niveau 10% l'hypothèse selon laquelle les suicides se répartissent uniformément sur les jours de la semaine.

10

Le test du Khi-Deux d'independance

Contingence signifie dépendance, de sorte qu'un tableau de contingence est un tableau qui montre comment une caractéristique dépend d'une autre. Le tableau suivant montre, par exemple, comment le revenu Y (exprimé en milliers de \$) dépend de la région X, dans un échantillon de 400 familles américaines, en 1971.

Y : Revenu	0-5	5-10	10-15	15-	**Total**
X : Région					
Sud	28	42	30	24	124
Nord	44	78	78	76	276
Total	72	120	108	100	400

Dans le cas général, X peut prendre les r modalités $i = 1, 2, \ldots, r$ et Y les s modalités $j = 1, 2, \ldots, s$. Ici, $r = 2$, $i \in \{\text{Nord, Sud}\}$ et $s = 4$, $j \in \{\text{0-5, 5-10, 10-15, 15-}\}$. Soient $p_i^X = I\!P(X = i)$, $p_j^Y = I\!P(Y = j)$ et $p_{ij} = I\!P(X = i \text{ et } Y = j)$. Avec cette notation, la proportion des individus de la population appartenant à la classe i selon la variable X, est

$$p_i^X = p_{i\bullet} := p_{i1} + p_{i2} + \cdots + p_{is}, \quad \text{pour tous les } i = 1, \ldots, r.$$

De même, la proportion des individus de la population appartenant à la classe j selon la variable Y, est

$$p_j^Y = p_{\bullet j} := p_{1j} + p_{2j} + \cdots + p_{rj}, \quad \text{pour tous les } j = 1, \ldots, s.$$

Les variables X et Y sont indépendantes si

$$H_0 : p_{ij} = p_{i\bullet} \times p_{\bullet j}, \quad \text{pour tous les } i = 1, \ldots, r, j = 1, \ldots, s.$$

Le problème qu'on se propose de résoudre est celui du test de cette hypothèse d'indépendance à l'aide des résultats d'un échantillon de taille n extrait de la population.

Supposons qu'on observe n_{ij} individus appartenant à la cellule (i, j), il y a alors

$$n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{is}$$

individus appartenant à la classe i pour X, et

$$n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{rj}$$

Tableau 10.1. Tableau de contingence pour deux variables X et Y

Y : / **X**	1	2	...	j	...	s	**Total**
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet s}$	n

Tableau 10.2. Effectifs attendus sous l'hypothèse H_0 d'indépendance

Y : / **X**	1	2	...	j	...	s	**Total**
1	$\frac{n_{1\bullet}n_{\bullet 1}}{n}$	$\frac{n_{1\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{1\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{1\bullet}n_{\bullet s}}{n}$	$n_{1\bullet}$
2	$\frac{n_{2\bullet}n_{\bullet 1}}{n}$	$\frac{n_{2\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{2\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{2\bullet}n_{\bullet s}}{n}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\frac{n_{i\bullet}n_{\bullet 1}}{n}$	$\frac{n_{i\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet s}}{n}$	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$\frac{n_{r\bullet}n_{\bullet 1}}{n}$	$\frac{n_{r\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{r\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{r\bullet}n_{\bullet s}}{n}$	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet s}$	n

individus appartenant à la classe j pour Y. Le nombre total d'individus de l'échantillon est n et on a les égalités

$$n = \sum_i \sum_j n_{ij} = n_{1\bullet} + \cdots + n_{r\bullet} = n_{\bullet 1} + \cdots + n_{\bullet s}.$$

Présentées dans un tableau, ces informations forme le Tableau de contingence 10.1.

Les quantités $n_{i\bullet}$ et $n_{\bullet j}$ apparaissent aux extrêmes des rangées et des colonnes, ils sont appelés effectifs marginaux des variables X et$Y$. Pour un échantillon de taille n, on s'attend sous H_0 à observer pour la cellule (i, j) l'effectif

$$np_{i\bullet}p_{\bullet j} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}.$$

Cette situation est présentée dans le Tableau 10.2. Dans le tableau de contingence des revenus américains, les effectifs attendus sous l'hypothèse H_0 d'indépendance région/revenu sont :

	0-5	5-10	10-15	15-	**Total**
Sud	$\frac{124\times 72}{400} = 22,32$	$\frac{124\times 120}{400} = 37,2$	$\frac{124\times 108}{400} = 33,48$	$\frac{124\times 100}{400} = 31,0$	124
Nord	$\frac{276\times 72}{400} = 49,68$	$\frac{276\times 120}{400} = 82,8$	$\frac{276\times 108}{400} = 74,52$	$\frac{276\times 100}{400} = 69,0$	276
Total	72	120	108	100	400

Le Tableau 10.1 est celui des observations alors que le Tableau 10.2 est celui des effectifs théoriques (attendus sous H_0). On peut donc noter que l'observation de la cellule (i, j) est $O_{ij} = n_{ij}$ alors que sont analogue théorique est $T_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}$. Par analogie avec (9.1) et (9.2), la distance du khi-2 est donc

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \tag{10.1}$$

que l'on note rapidement

$$\chi^2 = \sum\sum \frac{(O - T)^2}{T}. \tag{10.2}$$

Comme lors de la Leçon 9, χ^2_{obs} est positif et il ne vaut zéro que si les effectifs attendus sous H_0 et observés coïncident. Il sera d'autant plus grand que les écarts entre effectifs attendus sous H_0 et observés sont importants. Par conséquent, on aura tendance à rejeter H_0 lorsque la distance χ^2_{obs} sera grande. La régle de décision sera de la forme

$$\text{rejeter } H_0 \text{ si } \chi^2_{obs} > c_\alpha$$

où c_α est une constante à déterminer selon le niveau α désiré.

Par contre contrairement à (9.3), pour le test d'indépendance le nombre de degrés de liberté n'est pas $rs-1$, mais il est égal à

$$\text{d.d.l.} = (r-1)(s-1), \tag{10.3}$$

ce qui signifie que l'on doit déterminer le seuil c_α à l'aide de la table de la loi du khi-2 à $(r-1)(s-1)$ degrés de liberté :

$$I\!P(\chi^2_{(r-1)(s-1)} > c_\alpha) = \alpha.$$

Dans le cas qui nous intéresse, nous avons d.d.l. $= (2-1)(4-1) = 3$ et

$$\chi^2_{obs} = \frac{(28-22,32)^2}{22,32} + \frac{(42-37,2)^2}{37,2} + \frac{(30-33,48)^2}{33,48} + \frac{(24-31,0)^2}{31,0}$$
$$+ \frac{(44-49,68)^2}{49,68} + \frac{(78-82,8)^2}{82,8} + \frac{(78-74,52)^2}{74,52} + \frac{(76-69,0)^2}{69,0} = 5,81.$$

Puisque

$$I\!P(\chi^2_3 > 6,251) = 0,10$$

et que

$$\chi^2_{obs} = 5,81 < 6,251,$$

on accepte H_0 au niveau 10%, et à plus forte raison au niveau 5%.

Exemple 1. On reprend les données de l'exemple 2 de la Leçon 9, des pois de Mendel.

Graines	Jaunes	Vertes	**Total**
Lisses	315	108	423
Ridées	101	32	133
Total	416	140	556

On veut tester l'indépendance des caractères "forme" et "couleur". Le tableau des effectifs théoriques est le suivant

Graines	Jaunes	Vertes	**Total**
Lisses	316,49	106,51	423
Ridées	99,51	33,49	133
Total	416	140	556

On obtient $\chi^2_{obs} = \frac{(315-316,49)^2}{316,49} + \frac{(108-106,51)^2}{106,51} + \frac{(101-99,51)^2}{99,51} + \frac{(32-33,49)^2}{33,49} = 0,116$. Le nombre de degrés deliberté est $(2-1)(2-1) = 1$. Or, on a $I\!P(\chi^2_1 > 0,116) = 66\%$, de sorte qu'on accepte

l'hypothèse H_0 d'indépendance des caractères "forme" et "couleur" au niveau 66% et donc à tous les niveaux inférieurs.

Exemple 2. Afin de savoir si *les mathématiciens sont philosophes*, on a relevé sur 100 bacheliers les notes obtenues en Mathématiques et en Philosophie.

P : / **M**	0-3	4-7	8-11	12-15	16-20	**Total**
0-3	3	4	2	0	0	9
4-7	6	10	8	2	0	26
8-11	1	8	20	12	3	44
12-15	0	0	8	7	3	18
16-20	0	0	1	0	2	3
Total	10	22	39	21	8	100

Le tableau des effectifs attendus sous H_0 est

P : / **M**	0-3	4-7	8-11	12-15	16-20	**Total**
0-3	0,9	1,98	3,51	1,89	0,72	9
4-7	2,6	5,72	10,14	5,46	2,08	26
8-11	4,4	9,68	17,16	9,24	3,52	44
12-15	1,8	3,96	7,02	3,78	1,44	18
16-20	0,3	0,66	1,17	0,63	0,24	3
Total	10	22	39	21	8	100

Un calcul un peu long nous permet de montrer que $\chi^2_{obs} = 51,7346$. Nous avons aussi d.d.l. $= (5-1)(5-1) = 16$. La table $\chi^2(16)$ nous donne $c_{0,05} = 26,296$, par conséquent on rejette l'hypothèse d'indépendance au niveau 5%.

Exercices

1. La distribution suivante a été dressée par Haberman (1978) à partir de données fournies par le *National Opinion Research Center* de l'Université de Chicago. Les variables sont le nombre d'années de scolarité (X) et l'attitude face à l'avortement (Y).

X : **Scolarité**	Y : **Pour**	Y : **Indifférent**	Y : **Contre**
Moins de 8 ans	31	23	56
Entre 9 et 12 ans	171	89	177
Plus de 12 ans	116	39	74

Tester l'hypothèse selon laquelle X et Y sont indépendantes, au niveau 5%.

2. On a classé 217 enfants d'après leurs performances dans des tests de langage (L) et d'équilibre physique (E). Tester au niveau 5% l'hypothèse de l'indépendance des performances de langage et d'équilibre.

	L1	**L2**	**L3**
E1	45	26	12
E2	32	50	21
E3	4	10	17

11

Le test du Khi-Deux d'homogeneite

Lors de trois sondages consécutifs, on a observé que respectivement 51%, 48% et 55% des répondants étaient en faveur d'une politique donnée sur un total de 700, 900 et 800 répondants. Entre les moments où ces sondages ont été réalisés, y a-t'il eu changement d'opinion au sujet de cette politique ? Pour répondre à cette question, nous allons procéder à un test d'homogénéité.

Nous avons 3 populations correspondant aux 3 sondages (caractère X) réparties en 2 classes : "En faveur" et "En défaveur" (caractère Y).

Y	En faveur	En défaveur	**Total**
X			
1	357	343	700
2	432	468	900
3	440	360	800
Total	1229	1171	2400

Dans le cas général, la situation se présente sous la forme du Tableau 11.1.

Dans notre exemple, on pourra convenir de $Y = 1$ si l'individu est en faveur, et $Y = 2$ s'il est en défaveur de la politique considérée. On a donc $r = 3$ et $s = 2$. Dire qu'il n'y a pas eu de changement entre les différents sondages, c'est dire que les populations $X = 1$, $X = 2$ et $X = 3$ se comportent de la même manière en ce qui concerne le caractère Y. On dit alors que ces populations sont homogènes.

Dans le cas général, soit p_{ij} la proportion de la population $X = i$ dans la classe $Y = j$, l'hypothèse nulle d'homogénéité est

$$H_0 : p_{ij} = p_j^Y, \text{ pour tous les } i = 1, \ldots, r, j = 1, \ldots, s$$

Tableau 11.1. Tableau de contingence pour un test d'homogénéité d'une variable Y

Y :	1	2	...	j	...	s	
Population							**Taille de l'échantillon**
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet s}$	n

Tableau 11.2. Effectifs attendus sous l'hypothèse H_0 d'homogénéité

Y :	1	2	...	j	...	s	
Population							**Taille de l'échantillon**
1	$\frac{n_{1\bullet}n_{\bullet 1}}{n}$	$\frac{n_{1\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{1\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{1\bullet}n_{\bullet s}}{n}$	$n_{1\bullet}$
2	$\frac{n_{2\bullet}n_{\bullet 1}}{n}$	$\frac{n_{2\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{2\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{2\bullet}n_{\bullet s}}{n}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\frac{n_{i\bullet}n_{\bullet 1}}{n}$	$\frac{n_{i\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet s}}{n}$	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$\frac{n_{r\bullet}n_{\bullet 1}}{n}$	$\frac{n_{r\bullet}n_{\bullet 2}}{n}$	...	$\frac{n_{r\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{r\bullet}n_{\bullet s}}{n}$	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet s}$	n

où r est le nombre de populations et s le nombre de modalités du caractère Y. Puisqu'on a observé $n_{i\bullet}$ individus dans la population $X = i$ et que sous H_0 une bonne estimation de p_j^Y est $p_{\bullet j}$, sous H_0, l'effectif attendu de la cellule (i, j) est

$$n_{i\bullet} \times p_{\bullet j} = \frac{n_{i\bullet} n_{\bullet j}}{n},$$

ce qui nous donne le Tableau 11.2 d'effectifs attendus sous H_0.

Dans notre exemple, ce tableau des effectifs attendus sous H_0 donne :

Y / **X**	En faveur	En défaveur	**Total**
1	$\frac{700\times1229}{2400} = 358,46$	$\frac{700\times1171}{2400} = 341,54$	700
2	$\frac{900\times1229}{2400} = 460,87$	$\frac{900\times1171}{2400} = 439,13$	900
3	$\frac{800\times1229}{2400} = 409,67$	$\frac{800\times1171}{2400} = 390,33$	800
Total	1229	1171	2400

On constate que les formules sont les mêmes que celles du test du khi-2 d'indépendance traité à la Leçon 10. En particulier, les Tableaux 11.1 et 11.2 sont identiques aux Tableaux 10.1 et 10.2.

Le Tableau 11.1 est celui des observations alors que le Tableau 11.2 est celui des effectifs théoriques (attendus sous H_0). On peut donc noter que l'observation de la cellule (i, j) est $O_{ij} = n_{ij}$ alors que sont analogue théorique est $T_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. Par analogie avec (10.1) et (10.2), la distance du khi-2 est

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n})^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} \tag{11.1}$$

que l'on note rapidement

$$\chi^2 = \sum \sum \frac{(O - T)^2}{T}. \tag{11.2}$$

Comme lors de la Leçon 10, χ^2_{obs} est positif et il ne vaut zéro que si les effectifs attendus sous H_0 et observés coïncident. Il sera d'autant plus grand que les écarts entre effectifs attendus sous H_0 et observés sont importants. Par conséquent, on aura tendance à rejeter H_0 lorsque la distance χ^2_{obs} sera grande. La régle de décision sera de la forme

$$\text{rejeter } H_0 \text{ si } \chi^2_{obs} > c_\alpha$$

où c_α est une constante à déterminer selon le niveau α désiré.

Comme en (10.3), pour le test d'homogénéité le nombre de degrés de liberté est égal à

$$\text{d.d.l.} = (r-1)(s-1), \tag{11.3}$$

ce qui signifie que l'on doit déterminer le seuil c_α à l'aide de la table de la loi du khi-2 à $(r-1)(s-1)$ degrés de liberté :

$$I\!P(\chi^2_{(r-1)(s-1)} > c_\alpha) = \alpha.$$

Remarque. La similarité des tests d'indépendance et d'homogénéité n'est pas fortuite. En fait, **un test d'homogénéité est un test d'indépendance**. En effet, se poser la question : *"Les populations $i = 1, \ldots, r$ ont-elles un comportement homogène en regard de la variable Y ?"*, c'est se poser la question de l'indépendance de la variable population : X et de la variable Y.

Dans le cas qui nous intéresse, nous avons d.d.l. $= (3-1)(2-1) = 2$ et

$$\begin{aligned}\chi^2_{obs} &= \frac{(357-358,46)^2}{358,46} + \frac{(343-341,54)^2}{341,54} + \frac{(432-460,87)^2}{460,87} \\ &\quad + \frac{(468-439,13)^2}{439,13} + \frac{(440-409,67)^2}{409,67} + \frac{(360-390,33)^2}{390,33} \\ &= 8,32.\end{aligned}$$

puisque

$$I\!P(\chi^2_2 > 5,99) = 0,05$$

et que

$$\chi^2_{obs} = 8,32 > 5,99,$$

on rejette l'hypothèse H_0 d'homogénéité au niveau 5%.

Par contre, avec un niveau égal à 1%, on accepte l'hypothèse d'homogénéité car

$$I\!P(\chi^2_2 > 9,21) = 0,01$$

et

$$8,32 \not> 9,21.$$

Exercices

1. A la sortie de deux salles de cinéma donnant le même film, on a interrogé des spectateurs quant à leur opinion sur le film. Les résultats de ce sondage d'opinion sont les suivants

	Mauvais film	Bon film	**Total**
Salle1	30	70	100
Salle 2	48	52	100
Total	78	122	200

Montrez que l'opinion est significativement liée à la salle, au niveau 5%.

2. Une enquête a été menée aux Etats-Unis pour obtenir des informations sur la consommation d'alcool en fonction du statut familial. On a selectionné au hasard 1772 adultes de plus de 18 ans et on a obtenu les résultats suivants (en nombre de verres par mois)

	Aucun	1-60	Plus de 60	**Total**
Célibataire	67	213	74	354
Marié	411	633	129	1173
Veuf	85	51	7	143
Divorcé	27	60	15	102
Total	590	957	225	1772

Peut-on conclure au vu de ces résultats que le comportement des populations "Célibataire", "Marié", "Veuf" et "Divorcé" vis-à-vis de la consommation d'alcool est globalement le même ? On fera un test de niveau 1%.

12

Le test d'ajustement de Kolmogorov-Smirnov

Ce test statistique a la même fonction que le test d'ajustement du khi-deux. Il est basé sur une autre méthode. D'une certaine manière, on peut dire que le test de Kolmogorov-Smirnov est plus général que son analogue du khi-deux : il permet, contrairement au khi-deux qui n'est valide que pour des échantillons de grande taille (supérieure à 30, en pratique), de travailler avec des petits échantillons, mais aussi avec des grands. Dans ce dernier cas, les performances des deux tests d'ajustement sont comparables.

Rappelons ce qu'est un test d'ajustement. A l'aide des données $(x_1, \ldots, x_n)$ provenant de l'observation de n variables aléatoires indépendantes de même loi inconnue $\mathcal{L}$ à déterminer, on peut donner une réponse statistique *(c'est-à-dire entachée d'une erreur possible dont on peut évaluer la probabilité, et d'autant plus fiable que le nombre n d'observations est grand)* à la question : "La loi inconnue $\mathcal{L}$ de mes observations est-elle la loi $\mathcal{L}_o$ que je me donne ?" Par exemple, mes observations proviennent-elles d'une loi uniforme sur $[0, 365]$?

Illustrons ce test à l'aide d'un exemple. Je cherche à tester la fiabilité du programme de tirage uniforme aléatoire de ma calculette. Pour cela j'observe $n = 10$ résultats de tirages. Proviennent-ils d'une loi uniforme sur $[0, 1] : \mathcal{U}(0, 1)$? J'obtiens :

$$0.62, 0.36, 0.23, 0.76, 0.65, 0.09, 0.55, 0.26, 0.38 \text{ et } 0.24.$$

Je les range par ordre croissant :

$$0.09, 0.23, 0.24, 0.26, 0.36, 0.38, 0.55, 0.62, 0.65 \text{ et } 0.76.$$

Puis je dessine l'"escalier de répartition" correspondant, dont les marches sont de hauteur $1/n = 1/10$ et se situent en chacune des valeurs observées. Si le tirage simule bien une loi $\mathcal{U}(0, 1)$, cet escalier empirique, appelé $F_n = F_{10}$ doit être proche de la fonction de répartition F_o de cette loi

$$F_o(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x \leq 1 \end{cases}$$

qui est représentée sur la figure précédente à l'aide de la droite oblique. Justifions rapidement cette proximité attendue de F_{10} et de F_o, si la loi de mes observations est bien de fonction de répartition F_o. Dire : $F_{10}(0.48) = 5/10$ signifie que 5 de nos observations sont inférieures à 0.48. Dire que $F_o(0.48) = 0.48$ signifie qu'une variable aléatoire de loi $\mathcal{U}(0,1)$ prend une valeur inférieure à 0.48 avec une probabilité égale à $0.48 = 48\%$. Intuitivement, on s'attend bien à ce que ces quantités soient proches si la loi de mes observations indépendantes est $\mathcal{U}(0,1)$, et d'autant plus que le nombre n des observations est grand. C'est une conséquence de la loi des grands nombres.

On rejettera donc l'hypothèse nulle

$$(H_0) : \text{la loi de mes observations indépendantes est } \mathcal{U}(0,1)$$

si ces deux courbes sont "trop éloignées".

Le test est basé sur l'observation du plus grand écart : d_{10}, entre l'escalier de répartition F_{10} et la fonction de répartition théorique de la loi sous (H_0) : F_o. C'est-à-dire

$$d_{10} = \sup_{x \in \mathbb{R}} |F_{10}(x) - F_o(x)|.$$

La lecture de la table de Kolmogorov-Smirnov nous indique que pour $n = 10$, au niveau $\alpha = 10\%$, si $d_{10} > 0.37$: on rejette H_0, et si $d_{10} \leq 0.37$: on ne rejette pas H_0. Dans le cas de notre expérience, nous obtenons $d_{10} = F_{10}(0.65) - F_o(0.65) = 0.25$, qui est inférieur au seuil de rejet : 0.37. Donc, on ne rejette pas H_0 au niveau 10%.

On note que pour effectuer un test du khi-deux d'ajustement, outre que $n = 10$ est trop petit, nous aurions été contraints de regrouper nos observations par classes. Par exemple en 4 classes correspondant aux tirages qui tombent dans $[0, 1/4[, [1/4, 1/2[, [1/2, 3/4[$ et $[3/4, 1]$. Le test de Kolmogorov-Smirnov est donc avatageux (par rapport au khi-deux) lorsqu'on teste l'ajustement d'un échantillon à une loi de variable aléatoire continue.

Il arrive souvent, que lors d'une approche statistique, des expérimentateurs soient tentés par *l'hypothèse gaussienne*. C'est-à-dire, que les tests statistiques mis en place soient construits sur des variables aléatoires de loi normale. En pratique, cette hypothèse de travail peut ne pas correspondre à la réalité, et sur des "petits échantillons" cela provoque des erreurs parfois énormes. La littérature des sciences humaines et médicales est malheureusement parsemée de tels abus. Il y a un moyen d'y remédier. Commencer par un test de Kolmogorov-Smirnov du caractère gaussien des variables aléatoires observées.

Exercice. On se propose de vérifier si les cinq observations suivantes proviennent d'une loi normale $\mathcal{N}(3.1, 7.85)$. On a observé : 17.6, 4.5, -2.4, 2.5, 0.7. Si X suit une loi $\mathcal{N}(3.1, 7.85)$, alors $Z = \frac{X-3.1}{\sqrt{7.85}} \simeq \frac{X-3.1}{3.80}$ suit une loi normale centrée réduite : $\mathcal{N}(0, 1)$. Or, la fonction de répartition de $\mathcal{N}(0, 1)$ est tabulée dans la Table I. On a donc accés à une F_o et à un escalier F_5 pourvu que l'on opère la même transformation : $z = \frac{x-3.1}{3.80}$ sur nos observations. Ceci nous donnent les 5 observations modifiées, ordonnées de façon croissante : -1.45, -0.63, -0.16, 0.37, 3.82. Soit :

x	-1.45	-0.63	-0.16	0.37	3.82
$F_5(x)$	0.2	0.4	0.6	0.8	1
$F_o(x)$	0.073	0.268	0.436	0.644	1

L'écart maximal entre F_o et F_5 est obtenu tout juste à gauche de $x = 3.82$ et vaut $d_5 = 1 - 0.644 = 0.356$.

On lit dans la table de Kolmogorov-Smirnov que pour $n = 5$, au niveau $\alpha = 20\%$, on rejette H_0 lorsque d_5 excède 0.45. Nous ne rejetons donc pas H_0 au niveau 20% (et a fortiori à des niveaux inférieurs).

Attention, ceci ne signifie pas que nos observations suivent effectivement la loi normale $\mathcal{N}(3.1, 7.85)$. Mais seulement, que nous ne pouvons pas affirmer le contraire.

Exercices

1. Dix observations d'une variable aléatoire nous ont donné :

$$32.4, 6.2, 11.4, 27.3, 29.2, 17.0, 30.6, 21.6, 18.7, 8.0.$$

Tester l'hypothèse nulle que X suit la loi $\mathcal{N}(20, 100)$ avec $\alpha = 20\%$.

2. Les lois exponentielles servent souvent à modéliser des temps d'attente. Leur fonction de répartition est de la forme

$$F(x) = 1 - \exp(-x/\theta),\ x \geq 0$$

avec $\theta > 0$, et $F(x) = 0$ si $x \leq 0$. L'espérance de X est $E(X) = \theta$. On observe huit temps d'attente indépendants à un guichet, ce qui nous donne (en minutes) :

$$21, 19, 44, 2, 23, 15, 11, 34.$$

Tester l'hypothèse que le temps d'attente (exprimé en minute) suit une loi exponentielle de paramètre $\theta = 15$. On fera le test aux niveaux $20, 10, 5$ et 1%.