Statistiques

M. Calciano

Dans le chapitre concernant les probabilités, les lois des processus aléatoires étaient connues, et c'est à partir de ces lois que nous avons fait les calculs.

Adoptons à présent une approche différente d'ordre pratique : on ne connaît pas la loi de X_i. Par exemple la durée de vie d'un composant électronique, le nombre de clients se présentant au guichet d'une banque un jour donné, la proportion d'un certain caractère dans une population, le résultat d'un scrutin à partir des premiers dépouillements... Il peut être raisonnable de penser qu'elle appartient à une famille de lois usuelles (ce qui définit le *modèle de loi*) qui dépend d'un paramètre (inconnu) qu'on aimerait donc estimer à partir du résultat $(x_1, x_2, \dots, x_n)$ de n expériences.

Il s'agit bien d'une *estimation*, car on ne peut pas faire une infinité d'expériences, de même qu'on ne peut pas interroger l'ensemble de la population.

I. Statistique descriptive *(laissé lâchement à la sagacité du lecteur)*

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

1) Généralités

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

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

On considère une population Ω d'individus $\omega_1, \dots, \omega_n$.

Soit $X : \Omega \to \mathbb{R}$ une statistique simple quelconque.

Notons $X(\Omega) = \{x_i,\ i \in [\![1, p]\!]\}$ (quitte à les renuméroter, on les suppose rangés dans l'ordre croissant).

a) Série statistique discrète

L'effectif de la valeur x_i correspond au cardinal, noté n_i, de $X^{-1}(\{x_i\})$.

L'effectif cumulé en x_i correspond à $\displaystyle\sum_{j=1}^{i} n_j$.

L'effectif total est : $\displaystyle\sum_{i=1}^{p} n_i$.

La fréquence de la valeur x_i correspond à $f_i = \dfrac{n_i}{\sum_{i=1}^{p} n_i}$.

La fréquence cumulée de la valeur x_i correspond à $\displaystyle\sum_{j=1}^{i} f_j$.

Remarque : *Les familles* (x_i, n_i) *ou* (x_i, f_i) *sont appelées* familles statistiques discrètes.

b) Série statistique groupée

Soit $]a, b] \subset \mathbb{R}$.

L'effectif de la classe $]a, b]$ est le cardinal de $X^{-1}(]a, b])$.

L'effectif cumulé en b est le cardinal de $X^{-1}(]-\infty, b])$.

L'effectif total est $\text{card}(\Omega)$.

La fréquence de la classe $]a, b]$ est $\dfrac{\text{card}\big(X^{-1}(]a, b])\big)}{\text{card}(\Omega)}$.

La fréquence cumulée de la classe $]a, b]$ est $\dfrac{\text{card}\big(X^{-1}(]-\infty, b])\big)}{\text{card}(\Omega)}$.

2) Représentations graphiques

a) Série statistique discrète

- Diagramme en bâtons ou en colonnes, pour les effectifs ou les fréquences, qu'ils soient cumulés ou non.
- Polygone des effectifs ou des fréquences, cumulés ou non.

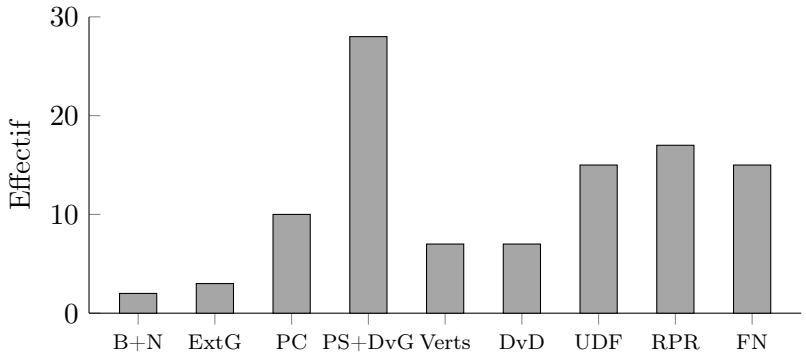

Figure 1: *

Figure 1.1. : élections législatives, diagramme en colonnes (effectifs fictifs).

b) Série statistique groupée

- Histogramme, dont les aires sont proportionnelles aux effectifs ou aux fréquences.
- Polygone des effectifs ou des fréquences, cumulés ou non.

```
> x <- c(91.6,35.7,251.3,24.3,5.4,67.3,170.9,9.5,118.4,57.1)
> abs <- c(0,26,78,130,182,234,275)
> ord <- c(0.0082,0.0077,0.0058,0.0019,0.0019,0.0019,0)
> hist(x, probability=T, breaks=seq(0,260,52), col=0, xlim=c(0,300),
      ylim=c(0,0.009))
> lines(abs, ord, lwd=5)
```

Le **mode** est le milieu de la classe correspondant au rectangle le plus haut (estimation du maximum de la densité). Ici, le mode est 26.

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

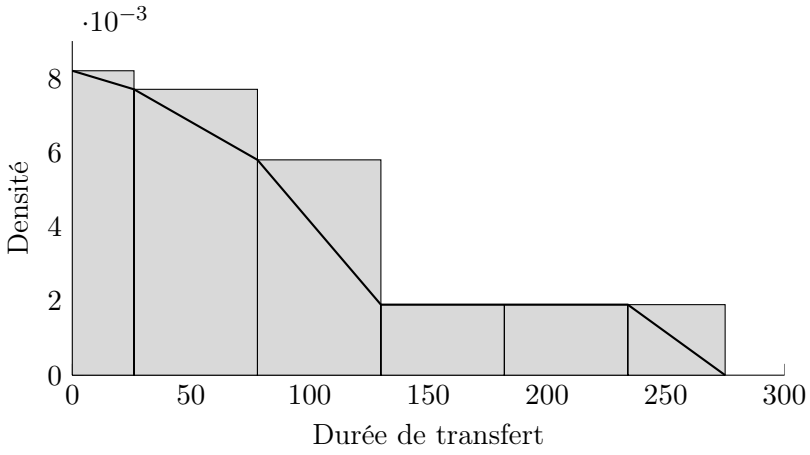

Figure 2: *

Figure 1.2. : trafic sur Internet, histogramme à classes inégales et polygone des fréquences.

3) Caractéristiques de position

a) Moyenne empirique

Si (x_i, n_i) avec $i \in [\![1, p]\!]$ est une série discrète, alors la *moyenne empirique* de l'échantillon est la moyenne arithmétique des observations :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{\sum_{i=1}^{p} n_i}$$

Remarque : $\forall (a, b) \in \mathbb{R}^2,\ E(aX + b) = aE(X) + b.$

b) Valeurs extrêmes

La plus petite valeur $x_1^* = \min_i x_i$ et la plus grande valeur $x_n^* = \max_i x_i$ d'un échantillon sont des indications intéressantes.

Problème : les deux indicateurs que l'on vient de définir sont très sensibles aux valeurs extrêmes. En particulier, il arrive parfois qu'une série statistique présente des valeurs aberrantes, c'est-à-dire des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon.

Aussi est-il important de disposer d'indicateurs qui ne soient pas trop sensibles aux valeurs aberrantes. La médiane empirique est un indicateur de localisation construit pour être insensible aux valeurs aberrantes.

c) Médiane empirique

On appelle *médiane empirique* d'une série statistique la valeur du caractère qui partage les fréquences, et donc les effectifs, en deux parties de même effectif.

Pour une série discrète :

Soit $X : \Omega \to \mathbb{R}$ une statistique simple quelconque. Notons $X(\Omega) = \{x_i,\ i \in [\![1, p]\!]\}$ (quitte à les renuméroter, on les suppose rangés dans l'ordre croissant).

- Si p est impair, la médiane de cette série est le nombre $x_{\frac{p+1}{2}}$.
- Si p est pair, la médiane de cette série est le nombre $\dfrac{x_{\frac{p}{2}} + x_{\frac{p}{2}+1}}{2}$.

Pour une série regroupée : la médiane est le réel μ correspondant à 50% des fréquences cumulées.

4) Caractéristiques de dispersion

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

Si (x_i, n_i) avec $i \in [\![1, p]\!]$ est une série discrète, on appelle *variance empirique* de la série, notée $V(X)$, la quantité :

$$V(X) = E(X^2) - E(X)^2$$

L'écart-type empirique : $\sigma(X) = \sqrt{V(X)}$.

Étude d'un exemple.

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

Table 1: *

Tableau 1.3. : températures mensuelles moyennes (°C) à New-York et à San Francisco.

La température annuelle moyenne est de 12,5 °C à New-York et de 13,7 °C à San Francisco. En se basant uniquement sur ces moyennes, on pourrait croire que les climats de ces deux villes sont similaires. Or il est clair que la différence de température entre l'hiver et l'été est beaucoup plus forte à New-York qu'à San Francisco. Ainsi, l'écart-type des températures annuelles est de 8,8 °C à New-York et de 3 °C à San Francisco, ce qui exprime bien la différence de variabilité des températures entre les deux villes.

II. Statistique descriptive à 2 variables

1) Généralités

Soient X et Y deux séries statistiques simples. On appelle *statistique double* l'application :

$$(X, Y) : \Omega \to \mathbb{R}^2, \qquad \text{notée } \big((x_i; y_i), n_{i,j}\big) \text{ avec } i \in [\![1, p]\!],\ j \in [\![1, q]\!].$$

L'espérance de XY, notée :

$$E(XY) = \frac{1}{n} \sum_{i \in [\![1,p]\!]} \sum_{j \in [\![1,q]\!]} n_{i,j}\, x_i\, y_j$$

La covariance de X, Y, notée :

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y)$$

2) Coefficient de corrélation linéaire

On a :

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sigma_X\, \sigma_Y}$$

Ainsi, si X et Y sont indépendantes, comme $E(XY) = E(X)E(Y)$, alors $\operatorname{cov}(X, Y) = 0$: le coefficient de corrélation est donc nul, et les variables X et Y sont dites *décorrélées*.

Remarque : *La réciproque est cependant fausse : prendre par exemple* $E(X) = 0$ *et* $Y = |X|$.

Étude d'un exemple.

X	Y
1	2
2	3
3	5
4	4
5	6

On obtient : 1,8 pour la covariance et 0,906 pour le coefficient de corrélation !

En résumé (en bleu, les commandes R permettant l'obtention des calculs) :

Paramètres statistiques d'un échantillon

Mesures de tendance centrale (position)

- Moyenne : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (`mean`)
- Médiane : partage les valeurs en deux parties (`median`)
- Quantiles : partagent les valeurs en k parties (`quantile`)
- Quartiles ($k = 4$) : Q_1, Q_2 (médiane), Q_3 (`quantile`)
- Mode(s) : la (les) valeur(s) de plus grande fréquence

Mesures de dispersion

- Étendue : $x_{(n)} - x_{(1)}$ (`max - min`)
- Intervalle interquartile (IQR) : $Q_3 - Q_1$ (`IQR`)
- Variance de l'échantillon (`var`) :

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

- Écart-type de l'échantillon : s (`sd`)
- Coefficient de variation : $s/\overline{x}$

3) Méthode des moindres carrés

Positionnement du problème De nombreuses séries statistiques $(x_i; y_i)_{1\leq i\leq n}$ sont reliées par des conditions du type $y = ax+b$. Ce peut être aussi le cas de grandeurs issues de la physique. En général, en raison des erreurs de mesure, les points $(x_i; y_i)_{1\leq i\leq n}$ ne sont pas alignés, mais sont « presque » sur une même droite. Il faut alors choisir a et b de sorte que la droite soit la meilleure possible.

Pour cela, il faut choisir une mesure de l'écart entre une droite $y = ax + b$ et le nuage de points expérimentaux $(x_i; y_i)_{1\leq i\leq n}$. On choisit en général le carré de la différence entre le point théorique et le point expérimental, c'est-à-dire $(y_i - (ax_i + b))^2$. L'écart total est donc :

$$J(a, b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Effectuer une *régression linéaire au sens des moindres carrés*, c'est trouver la droite qui minimise l'écart

précédent, c'est-à-dire la somme des carrés des différences : on parle de *droite des moindres carrés*.

Théorème Si la variance $\mathrm{Var}(X)$ de la série statistique $X = (x_i)$ est non nulle, il existe une unique droite qui minimise la quantité $J(a,b)$, donnée par :

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \qquad \text{et} \qquad b = \overline{Y} - a\overline{X},$$

où $\mathrm{Cov}(X,Y)$ désigne la covariance de X et de Y, $\overline{X}$ la moyenne de (x_i) et $\overline{Y}$ la moyenne de (y_i).

Démonstration :

Posons $f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$, *fonction de classe* $\mathcal{C}^\infty$ *sur* $\mathbb{R}^2$. *En développant :*

$$f(a,b) = n\Big(E(Y^2) + a^2E(X^2) + b^2 - 2aE(XY) - 2bE(Y) + 2abE(X)\Big)$$

d'où les dérivées partielles :

$$\frac{\partial f}{\partial a} = 2n\big(aE(X^2) - E(XY) + bE(X)\big), \qquad \frac{\partial f}{\partial b} = 2n\big(b - E(Y) + aE(X)\big)$$

Le point critique (a,b) *annule ce gradient ; en résolvant ce système linéaire, on obtient :*

$$a = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \qquad b = E(Y) - aE(X) = \overline{Y} - a\overline{X}.$$

On note que ce point critique existe bien car $\mathrm{Var}(X) \neq 0$ *par hypothèse.*

Les dérivées secondes de la matrice hessienne valent :

$$r = \frac{\partial^2 f}{\partial a^2} = 2nE(X^2), \qquad t = \frac{\partial^2 f}{\partial b^2} = 2n, \qquad s = \frac{\partial^2 f}{\partial a\,\partial b} = 2nE(X)$$

On a alors :

$$rt - s^2 = 4n^2E(X^2) - 4n^2E(X)^2 = 4n^2\,\mathrm{Var}(X) > 0 \qquad \textit{et} \qquad r > 0,$$

donc le point critique est un minimum, et ce minimum est unique car f *est strictement convexe (la hessienne est définie positive en tout point).* ■

Exemple La tension U aux bornes d'une batterie de force électromotrice E et de résistance interne R est $U = E - RI$, où I est l'intensité. On a procédé à différentes mesures :

Intensité mesurée (A) :	0	0,1	0,4	1
Tension mesurée (V) :	12	11	7	1

La régression linéaire donne $E = 11{,}9$ et $R = 11{,}07$.[1]

III. Estimation

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

[1] Exemple et données issus de la fiche « Régression linéaire » du site bibmath.net.

Les techniques de statistique descriptive, comme l'histogramme ou le graphe de probabilités, permettent de faire des hypothèses sur la nature de la loi de probabilité des X_i.

Des techniques statistiques plus sophistiquées, appelées *tests d'adéquation*, permettent de valider ou non ces hypothèses. On supposera ici que ces techniques ont permis d'adopter une famille de lois de probabilité bien précise (par exemple, loi normale, loi binomiale, etc.) pour la loi des X_i, mais que la valeur du ou des paramètres de cette loi est inconnue.

On notera θ le paramètre inconnu. A priori, θ peut être un paramètre à plusieurs dimensions, mais on supposera ici que θ **est un réel**. On notera $F(x;\theta)$ la fonction de répartition des X_i. Pour les variables aléatoires discrètes, on notera $P(X = x;\theta)$ les probabilités élémentaires, et pour les variables aléatoires continues, $f(x;\theta)$ la densité.

Le problème traité dans ce chapitre est celui de l'estimation du paramètre θ. Il s'agit de donner, au vu des observations $x_1, \dots, x_n$, une approximation de θ que l'on espère la plus proche possible de la vraie valeur inconnue.

On pourra proposer une unique valeur vraisemblable pour θ (**estimation ponctuelle**) ou un ensemble de valeurs vraisemblables (**intervalle de confiance**) à l'aide d'estimateurs (**sans biais**, **convergents**, etc.).

1) Généralités

a) Définition

Un *n-échantillon* de X est un n-uplet $(X_1, X_2, \dots, X_n)$ tel que les X_k ont la même loi que X et sont indépendantes.

Une *réalisation* de l'échantillon est alors un n-uplet $(x_1, x_2, \dots, x_n)$ de valeurs prises par l'échantillon.

Il faut bien distinguer $(X_1, X_2, \dots, X_n)$, qui est une variable aléatoire, de $(x_1, x_2, \dots, x_n)$, qui est un vecteur de $\mathbb{R}^n$.

Exemple On dispose d'une pièce dont on ne connaît pas la probabilité de tomber sur Pile, que l'on note p, et que l'on aimerait estimer. On considère alors une v.a. X (qui vaut 1 en cas de Pile et 0 sinon), qu'on va identifier parmi un ensemble de lois de Bernoulli, ici $\mathcal{M}_\Theta = \{B(p)\,;\ p \in [0;1]\}$.

On lance cette pièce n fois et on note X_i la v.a. qui vaut 1 si la pièce tombe sur Pile au i-ème lancer, et 0 sinon.

Le n-uplet $(X_1, X_2, \dots, X_n)$ est un n-échantillon de X. On décide de prendre $n = 10$. Le résultat des 10 lancers donne $(1, 1, 0, 0, 1, 1, 0, 0, 0, 1)$, qui représente donc une réalisation du 10-échantillon $(X_1, X_2, \dots, X_n)$.

Il va donc falloir introduire un *estimateur*, c'est-à-dire une fonction de l'échantillon qui, appliquée à l'observation, donnera une estimation du paramètre « p » cherché.

b) Définition

Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon d'une v.a. X dont la loi dépend d'un paramètre θ, θ appartenant à une partie $\Theta \subset \mathbb{R}$.

On appelle **estimateur de** θ toute variable aléatoire T_n de la forme :

$$T_n = \varphi(X_1, X_2, \dots, X_n),$$

où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, à valeurs dans Θ, éventuellement dépendante de n, et dont l'expression ne fait pas apparaître θ.

Toute réalisation d'un estimateur est appelée estimation ponctuelle de θ.

Reprenons l'exemple précédent. La variable aléatoire $\overline{X_{10}} = \dfrac{X_1 + \cdots + X_{10}}{10}$ est un estimateur de p. À

partir de l'observation $(1,1,0,0,1,1,0,0,0,1)$, $\overline{X_{10}}$ donne une estimation (ponctuelle) de $1/2$ pour p.

Remarque : *Certains estimateurs visent à estimer non pas le paramètre, mais une fonction $g(\theta)$ du paramètre. Par exemple, en reprenant l'exemple précédent, on peut vouloir estimer la variance de X, à savoir la quantité $g(p) = p(1-p)$, qui est bien une fonction du paramètre p. Les deux variables aléatoires ci-dessous sont ainsi des estimateurs de $g(p)$:*

$$\frac{1}{10}\sum_{i=1}^{10}(X_i - \overline{X_{10}})^2 \qquad et \qquad \overline{X_{10}}\,(1-\overline{X_{10}})$$

Remarque : *La notion d'estimateur paraît alors très floue (ou très générale), dans le sens « destinée à approcher la valeur de θ », puisqu'en réalité, toute fonction des données est donc un estimateur... Le choix de l'estimateur est une vraie question, d'où la définition de critères, qui vont suivre, dont le but sera d'apprécier la qualité de l'estimateur.*

Exemple Soit X une variable aléatoire dont la loi dépend d'un paramètre θ, et soit $(X_1, \dots, X_n)$ un n-échantillon de la variable X.

$X_1 + \cdots + X_n$ et $\sqrt{X_1 + \cdots + X_n}$ sont des estimateurs de θ.

Alors que $X_1 + \cdots + X_n - \theta$ n'est pas un estimateur de θ, puisqu'il s'exprime avec θ !

c) Définition

Comme évoqué précédemment, n'importe quelle fonction des observations à valeurs dans l'ensemble des valeurs possibles de θ est un estimateur de θ. Mais un estimateur d'un paramètre θ ne sera satisfaisant que si, pour n'importe quelle observation $x_1, x_2, \dots, x_n$, il est « proche », en un certain sens, de θ.

Pour cela, il faut d'abord que, si on répète plusieurs fois l'expérience, la moyenne des estimations obtenues soit très proche, et dans l'idéal égale, à θ. Cela revient à souhaiter que l'espérance de l'estimateur soit égale à θ.

Il est ainsi naturel d'évaluer l'écart entre l'espérance de l'estimateur et θ.

Le biais de l'estimateur T de θ est $E[T] - \theta$.

S'il est nul, on dit que T est un estimateur sans biais.

L'estimateur T_n est asymptotiquement sans biais si $\lim\limits_{n\to+\infty} E[T_n] = \theta$.

Remarque : *On note souvent le biais $b_\theta(T)$.*

Exercice « classique »
Soit $(X_1, \dots, X_n)$ un n-échantillon de la variable $X \in \mathcal{M}_\Theta = \{\mathcal{U}[0,\theta]\,;\ \theta \in \mathbb{R}^+\}$.

1) Montrer que $\overline{X_n}$ est un estimateur biaisé de θ, préciser $b_\theta(\overline{X_n})$.
2) On considère l'estimateur $M_n = \max(X_1, \dots, X_n)$.
   a) Déterminer la fonction de répartition F de M_n.
   b) En déduire une densité f de M_n.
   c) Montrer que $E_\theta(M_n) = \dfrac{n\theta}{n+1}$.
   d) En déduire un estimateur sans biais Z_n à partir de M_n.

d) Définition

L'estimateur T_n est dit convergent si la suite (T_n) converge en probabilité vers θ_0 :

$$\forall \varepsilon > 0, \qquad \lim_{n\to+\infty} P\big(|T_n - \theta_0| > \varepsilon\big) = 0$$

Remarque : *D'après l'inégalité de Bienaymé-Tchebychev, pour qu'un estimateur asymptotiquement sans biais soit convergent, il suffit que* $V(T_n) \to 0$.

Exercice
Soit $(X_1, \dots, X_n)$ un n-échantillon de la variable X admettant pour espérance m et pour variance σ^2.

1) On suppose que m est connu. Montrer que $T_n = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est un estimateur non biaisé de σ^2.

2) On suppose que m n'est pas connu. On note $\overline{X_n}$ la moyenne empirique de l'échantillon, et on pose $U_n = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$.

   a) Montrer que $\forall i \in [\![1, n]\!]$, $E\big((X_i - \overline{X_n})^2\big) = V(X_i - \overline{X_n})$.

   b) Montrer que $\forall i \in [\![1, n]\!]$, $V(X_i - \overline{X_n}) = \left(1 - \dfrac{1}{n}\right)^2 V(X_i) + \dfrac{1}{n^2}\sum_{k \neq i} V(X_k)$.

   c) En déduire que $V(X_i - \overline{X_n}) = \dfrac{n-1}{n}\sigma^2$.

   d) Montrer alors que U_n est un estimateur biaisé de σ^2. En déduire un estimateur sans biais de σ^2.

Remarque : *Disposer d'un estimateur sans biais semble présenter un réel intérêt, puisque son espérance est égale au paramètre cherché ; mais ce seul fait ne garantit pas que l'estimateur fournit de bonnes approximations. Pour évaluer le défaut de moyenne, et juger de la qualité d'un estimateur, on calcule la moyenne des carrés des écarts au paramètre.*

e) Définition

Soit T un estimateur de θ.

Le risque quadratique est défini par :

$$R(T, \theta) = E\big[(T - \theta)^2\big] = V(T) + b_\theta(T)^2$$

On dit que T_1 est un meilleur estimateur que T_2 si $\forall \theta \in I,\ R(T_1, \theta) \leq R(T_2, \theta)$.

De deux estimateurs sans biais, le meilleur est celui qui a la plus petite variance. C'est logique : il faut non seulement que la moyenne des estimations soit proche de θ, mais aussi que chaque estimation soit la plus proche possible de θ, donc que la variabilité de l'estimateur soit faible. Finalement, on considère que le meilleur estimateur possible de θ est un estimateur sans biais et de variance minimale (ESBVM). Un tel estimateur n'existe pas forcément.

f) Exemples classiques

Soit X une variable aléatoire telle que $E[X] = m$ et $\mathrm{Var}(X) = \sigma^2$.

$\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent de l'espérance mathématique m.

2) Estimateurs par la méthode des moments

C'est la méthode la plus naturelle. L'idée de base est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc.

Si le paramètre à estimer est l'espérance de la loi des X_i, alors on peut l'estimer par la moyenne empirique de l'échantillon. Autrement dit, si $\theta = E[X]$, alors l'estimateur de θ par la méthode des moments (EMM) est :

$$\widetilde{\theta_n} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

De la même manière, on estime la variance de la loi des X_i par la variance empirique de l'échantillon :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2$$

Exemple n°1 (loi de Bernoulli) Si $X_1, \dots, X_n$ sont indépendantes et de même loi de Bernoulli $B(p)$, alors $E[X] = p$. L'estimateur de p par la méthode des moments est $\widetilde{p_n} = \overline{X_n}$. Cet estimateur n'est autre que la proportion de 1 dans l'échantillon : on retrouve le principe d'estimation d'une probabilité par une proportion.

Exemple n°2 (loi exponentielle) Si $X_1, \dots, X_n$ sont indépendantes et de même loi exponentielle $\mathcal{E}(\lambda)$, alors $E[X] = 1/\lambda$. L'estimateur de λ par la méthode des moments est $\widetilde{\lambda_n} = \dfrac{1}{\overline{X_n}}$.

Exemple n°3 (loi normale) Si $X_1, \dots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(m, \sigma^2)$, alors $E[X] = m$ et $\mathrm{Var}(X) = \sigma^2$, donc les estimateurs de m et σ^2 par la méthode des moments sont $\widetilde{m_n} = \overline{X_n}$ et $\widetilde{\sigma_n^2} = S_n^2$.

3) Estimateurs par intervalle de confiance

Pour l'estimation ponctuelle, on considère un paramètre inconnu θ, un ensemble de valeurs observées $(x_1, \dots, x_n)$, réalisations d'un échantillon aléatoire $(X_1, \dots, X_n)$, et son estimation ponctuelle $\overline{x_n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

Les estimations ponctuelles n'apportent pas d'information sur la précision des résultats, c'est-à-dire qu'elles ne tiennent pas compte des erreurs dues aux fluctuations d'échantillonnage.

Pour évaluer la confiance que l'on peut avoir en une valeur, il est nécessaire de déterminer un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur du paramètre : c'est l'*estimation par intervalle de confiance*.

a) Définition

Soit $(X_1, \dots, X_n)$ un n-échantillon aléatoire et θ un paramètre inconnu de la loi des X_i. Soit $\alpha \in]0, 1[$. S'il existe des v.a.r. $\theta_{\min}(X_1, \dots, X_n)$ et $\theta_{\max}(X_1, \dots, X_n)$ telles que :

$$P\Big(\theta \in \big[\theta_{\min}(X_1, \dots, X_n),\, \theta_{\max}(X_1, \dots, X_n)\big]\Big) = 1 - \alpha,$$

on dit alors que $\big[\theta_{\min}(X_1, \dots, X_n),\, \theta_{\max}(X_1, \dots, X_n)\big]$ est un *intervalle de confiance* pour θ, avec un coefficient de confiance $1 - \alpha$, noté IC. Dans la pratique, on peut prendre par exemple $\alpha = 5\%$, ce qui nous donne un IC à 95%. Cela signifie qu'il y a 95% de chance que la valeur inconnue θ soit comprise entre $\theta_{\min}(x_1, \dots, x_n)$ et $\theta_{\max}(x_1, \dots, x_n)$.

b) Intervalle de confiance pour l'espérance et la variance dans un échantillon gaussien

Soit $(X_1, \dots, X_n)$ un n-échantillon de v.a.r. de loi $\mathcal{N}(\mu, \sigma^2)$.

Si la variance σ^2 est connue Pour estimer μ, on utilise la moyenne empirique $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, qui a pour loi $\mathcal{N}(\mu, \sigma^2/n)$. Ainsi :

$$\sqrt{n}\,\frac{\overline{X_n} - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

On détermine $z_{1-\alpha/2}$ tel que $P\left(-z_{1-\alpha/2} \leq \sqrt{n}\dfrac{\overline{X_n} - \mu}{\sigma} \leq z_{1-\alpha/2}\right) = 1-\alpha$, et on obtient un :

$$\text{IC} = \left[\overline{X_n} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\,;\ \overline{X_n} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Pourquoi $z_{1-\alpha/2}$ et non $z_{1-\alpha}$? :
Soit $X \sim \mathcal{N}(0,1)$. Pour $x > 0$:

$$P(-x \leq X \leq x) = P(X \leq x) - P(X \leq -x)$$

Or, par symétrie de la loi gaussienne : $P(X \leq -x) = P(X \geq x) = 1 - P(X \leq x)$. Donc :

$$P(-x \leq X \leq x) = P(X \leq x) - \big(1 - P(X \leq x)\big) = 2P(X \leq x) - 1$$

Ainsi :

$$P(-x \leq X \leq x) = 1-\alpha \iff 2P(X \leq x) - 1 = 1-\alpha \iff P(X \leq x) = 1 - \frac{\alpha}{2}$$

d'où $x = z_{1-\alpha/2}$, et non $z_{1-\alpha}$. ■

D'où l'utilisation de $\text{IC} = \left[\overline{x_n} - z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\,;\ \overline{x_n} + z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$, où $\overline{x_n}$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, \ldots, X_n)$.

Remarque : *Dans le cadre d'un intervalle à 95%, on a $z_{1-\alpha/2} \approx 1{,}96$.*

Table 2: *

Table de la fonction de répartition Φ de la loi $\mathcal{N}(0,1)$.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Si la variance σ^2 est inconnue Lorsque la variance σ^2 est inconnue, il est alors nécessaire de remplacer, dans les formules précédentes, cette quantité par la variance empirique, qui en est un estimateur convergent. Il faut donc considérer non plus la quantité $\sqrt{n}\dfrac{\overline{X_n}-\mu}{\sigma}$, mais $\sqrt{n}\dfrac{\overline{X_n}-\mu}{S_n}$, qui ne suit pas une loi normale mais une loi de Student à $n-1$ degrés de liberté. Ainsi :

$$\sqrt{n}\,\frac{\overline{X_n}-\mu}{S_n} \sim T_{n-1}$$

À l'aide des tables de la loi de Student, on détermine $t_{1-\alpha/2}$ tel que $P\left(-t_{1-\alpha/2} \leq \sqrt{n}\dfrac{\overline{X_n}-\mu}{S_n} \leq t_{1-\alpha/2}\right) = 1-\alpha$, et on obtient un :

$$\text{IC} = \left[\overline{X_n} - t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\,;\ \overline{X_n} + t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right]$$

D'où l'utilisation de $\text{IC} = \left[\overline{x_n} - t_{1-\alpha/2}\dfrac{S_n}{\sqrt{n}}\,;\ \overline{x_n} + t_{1-\alpha/2}\dfrac{S_n}{\sqrt{n}}\right]$, où $\overline{x_n}$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, \ldots, X_n)$.

Remarque : *Dans le cadre d'un intervalle à* 95% *avec* $n-1=3$ *degrés de liberté, on a* $t_{3;0,975}=3{,}182$.

Table 3: *

Fractiles de la loi de Student.

$\nu \backslash P$	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,706	31,821	63,657	318,309	636,619
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,327	31,599
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,255	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,299	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,526	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,416
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,364	2,626	3,174	3,390
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,340
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,107	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

c) Étude d'un exemple

Une entreprise chimique commercialise un polymère servant à la fabrication de microprocesseurs, et stocké dans une cuve dont la caractéristique à contrôler est la viscosité ; celle-ci doit être comprise entre 75 et 95 pour pouvoir commercialiser le polymère.

Quatre extractions ont été réalisées dans des zones différentes de la cuve et ont conduit aux valeurs de l'échantillon : $x_1 = 78$, $x_2 = 85$, $x_3 = 91$, $x_4 = 76$, réalisations des variables aléatoires X_1, X_2, X_3, X_4.

L'entreprise a besoin d'estimer la viscosité et aussi de connaître la précision de cette estimation. Ayant choisi a priori un seuil de 5%, il s'agit de fournir aux clients des intervalles de confiance à 95% pour μ.

1) Résumer les estimations ponctuelles.

2) Donner l'intervalle de confiance avec $\sigma = 5$.

3) Donner l'intervalle de confiance sans connaître la variance.

Éléments de résolution :

- *Le modèle considère que les variables X_i sont indépendantes, de loi $\mathcal{N}(\mu, \sigma^2)$; μ représente la moyenne de la viscosité dans la cuve, tandis que σ^2 prend en compte la variabilité de la viscosité au sein de la cuve et celle due à l'erreur de mesure.*
- *Les paramètres sont la moyenne μ et la variance σ^2.*
- *Les estimateurs sont $\overline{X}$ de μ et S^2 de σ^2.*
- *Les estimations ponctuelles sont $\overline{x} = 82{,}5$ et $s = 6{,}86$.*

Si l'on admet que la variabilité du processus de fabrication est constante et connue, avec $\sigma = 5$, l'estimateur de μ est gaussien, $z_{1-\alpha/2} = 1{,}96$, et les formules précédentes conduisent à l'intervalle de confiance de μ :

$$\left[82{,}5 - 1{,}96 \times \frac{5}{2}\,;\ 82{,}5 + 1{,}96 \times \frac{5}{2}\right] = [77{,}6\,;\ 87{,}4]$$

L'intervalle obtenu est bien à l'intérieur de la spécification $[75\,;\,95]$.

Si la variance n'est plus supposée constante et connue, elle doit être estimée. L'estimation de l'écart-type est $s = 6{,}86$. Celle-ci est certes plus importante que la valeur théorique précédente, mais surtout, l'estimateur de la moyenne μ suit maintenant une loi de Student à $n-1 = 3$ degrés de liberté. La table de cette loi fournit le quantile $t_{3;0{,}975} = 3{,}182$, d'où :

$$\left[82{,}5 - 3{,}182 \times \frac{6{,}86}{2}\,;\ 82{,}5 + 3{,}182 \times \frac{6{,}86}{2}\right] = [71{,}6\,;\ 93{,}4]$$

L'intervalle n'est cette fois pas contenu dans la spécification. On note l'augmentation sensible de la taille de cet intervalle, due au simple fait de devoir estimer la variance plutôt que de la supposer connue.
■

4) Estimation d'une proportion

On désire évaluer la probabilité p qu'un événement A se produise au cours d'une expérience donnée : $p = P(A)$. Pour cela, on fait n expériences identiques et indépendantes, et on compte le nombre x de fois où A s'est produit. x est la réalisation d'une variable aléatoire X qu'on sait être de loi binomiale $B(n, p)$.

Exemple Une élection oppose deux candidats A et B. Un institut de sondage interroge 800 personnes sur leurs intentions de vote : 420 déclarent voter pour A et 380 pour B. Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra le candidat A. En supposant que les réponses des 800 personnes interrogées sont indépendantes, on est bien dans le cas de figure de l'estimation d'une proportion.

a) Estimation ponctuelle

Remarquons que nous n'avons ici qu'une seule réalisation de X (ne pas confondre la variable binomiale X, qui compte les succès, avec les Bernoulli), c'est-à-dire un échantillon de taille 1.

Pour une fois, la notation n ne désigne pas la taille de l'échantillon. Il est naturel d'estimer la probabilité p que A se produise par le pourcentage $\dfrac{X}{n}$ de fois où A s'est produit au cours des n expériences.

Par la méthode des moments, on a $E(X) = np$, donc l'EMM de p est $\dfrac{X}{n}$.

b) Estimation par intervalle de confiance

On peut, dans le cas où $np \geq 5$ et $n(1-p) \geq 5$, approcher la loi binomiale $B(n,p)$ par la loi normale $\mathcal{N}\big(np,\, np(1-p)\big)$, ainsi $\dfrac{X-np}{\sqrt{np(1-p)}}$ suit approximativement la loi $\mathcal{N}(0,1)$.

On peut alors déterminer u_α tel que : $P\left(\left|\dfrac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha\right) = 1-\alpha$.

Il reste à construire l'IC à partir de : $\left|\dfrac{X-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$.

Remarque : *On peut utiliser une version approchée de l'intervalle de confiance d'une proportion, avec la formule :*

$$\text{IC} = \left[p_n - z_{1-\alpha/2}\frac{1}{2\sqrt{n}}\,;\; p_n + z_{1-\alpha/2}\frac{1}{2\sqrt{n}}\right]$$

où p_n représente la proportion observée dans l'échantillon de taille n, avec erreur à $\alpha\%$ (avec $z_{1-\alpha/2}$ tel que $P(-z_{1-\alpha/2} \leq X \leq z_{1-\alpha/2}) = 1-\alpha$ où $X \sim \mathcal{N}(0,1)$; formule basée uniquement sur la majoration de $\sqrt{p(1-p)}$ par $\dfrac{1}{2}$).

IV. Exercices

Exercice n°1 *(autocorrection)*
Soit n un entier supérieur ou égal à 2.

Soit $(x_i; y_i)$ pour $i \in [\![1,n]\!]$ une série statistique discrète telle que le moment d'ordre 2 et la variance de (x_i) soient non nuls.

La droite de régression au sens des moindres carrés est la droite d'équation $y = ax+b$ avec $(a,b) \in \mathbb{R}^2$ qui minimise $\delta_{a,b} = \sum_{i=1}^{n}(y_i - ax_i - b)^2$.

En considérant la fonction f définie sur $\mathbb{R}^2$ par $f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$, on veut montrer que $\delta_{a,b}$ admet un minimum sur $\mathbb{R}^2$ et préciser pour quelles valeurs de a et de b celui-ci est atteint.

1) Exprimer $f(a,b)$ en fonction de $E(X^2)$, $E(XY)$, $E(X)$, $E(Y)$ et $E(Y^2)$.

2) Déterminer le point critique de f.

3) Déterminer les coefficients de la matrice hessienne.

4) Conclure.

(Corrigé : voir la démonstration du théorème de la section II.3 ci-dessus, qui détaille exactement ce calcul.)

Année	CA
1	3
2	7
3	10
4	18
5	21
6	35
7	53
8	82

Exercice n°2 *(autocorrection)*
Une entreprise réalise une étude sur l'évolution de son chiffre d'affaires, en millions d'euros, lors des 8 dernières années (cf. tableau de l'exercice précédent).

On note X la variable correspondant à l'année et Y la variable correspondant au chiffre d'affaires.

1) Représenter le nuage de points correspondant à (X, Y).

2) Donner les paramètres de la droite de régression de Y en X.

3) L'ajustement précédent est-il acceptable ?

Corrigé :

$E(X) = 4{,}5$, $E(Y) = 28{,}625$, $E(X^2) = 25{,}5$, $E(Y^2) = 1460{,}125$.

$V(X) = 25{,}5 - 4{,}5^2 = 5{,}25$.

$V(Y) = 1460{,}125 - 28{,}625^2 = 640{,}734$.

$E(XY) = \dfrac{1461}{8} = 182{,}625$.

$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 53{,}813$.

Soit $a = \mathrm{cov}(X, Y)/V(X) = 10{,}25$ *et* $b = E(Y) - aE(X) = -17{,}5$, *soit* $y = 10{,}25x - 17{,}5$.

Le coefficient de corrélation : $\mathrm{cov}(X, Y)/\sqrt{V(X)V(Y)} = 0{,}928 > 0{,}75$, *l'ajustement est donc acceptable !* ■

Exercice n°3 *(autocorrection)*
Après une épreuve de philosophie à un concours, on s'intéresse au lien existant, pour une centaine de candidats pris au hasard, entre le temps consacré à la philosophie durant leur année scolaire et le résultat obtenu à cette épreuve.

En notant X leur note sur 20 et Y leur temps de travail en heures en philosophie durant l'année, on obtient les résultats suivants :

1) Déterminer la moyenne en philosophie de cet échantillon, ainsi que le temps moyen consacré par cet échantillon.

$X \setminus Y$	$[0,20[$	$[20,50[$	$[50,100[$	$[100,200[$
$[0,4[$	6	6	2	5
$[4,8[$	7	17	15	3
$[8,12[$	6	7	6	2
$[12,16[$	2	7	6	2
$[16,20[$	0	1	0	0

2) Déterminer la droite de régression de X en Y.

3) Calculer le coefficient de corrélation linéaire de (X,Y). Conclure.

Corrigé :

$E(X) = 7{,}56$ *et* $E(Y) = 55{,}15$.

$E(X^2) = \dfrac{7344}{100} = 73{,}44$, *donc* $V(X) = E(X^2) - E(X)^2 = 16{,}286$.

$E(Y^2) = \dfrac{481775}{100} = 4817{,}75$, *donc* $V(Y) = 1776{,}228$.

$E(XY) = \dfrac{41170}{100} = 411{,}7$.

$\mathrm{Cov}(X,Y) = -5{,}234$, *soit* $a = -0{,}003$ *et* $b = 7{,}723$, *c'est-à-dire* $x = -0{,}003y + 7{,}723$, *mais avec une corrélation de* $-0{,}03$ *: l'ajustement linéaire n'est donc pas pertinent ici.* ■

Exercice n°4

Un sondage portant sur 100 fumeurs choisis au hasard et avec remise au sein d'une population donnée permet de constater que leur consommation moyenne est $\mu = 9{,}37$ cigarettes par jour, avec un écart-type $\sigma = 2{,}14$.

1) Déterminer un intervalle de confiance à 95% de la consommation de cigarettes par jour au sein de cette population.

2) Même question avec un intervalle à 99%.

Exercice n°5

À la veille du second tour d'une élection présidentielle opposant deux candidats A et B d'importances comparables, un institut de sondage désire estimer la probabilité que le candidat A soit élu.

On effectue l'hypothèse que les suffrages des différents électeurs sont indépendants, et que chacun d'entre eux a une probabilité p de voter pour A à ce second tour.

1) L'institut de sondage désire obtenir une estimation à $\pm 3\%$ maximum du score du candidat A à ce second tour.

   a) Combien d'électeurs doit-il sonder au minimum pour obtenir un intervalle de confiance à 95% de son estimation ?

   b) Même question à 99%.

   c) On considère que cet institut effectue un sondage auprès de 1068 électeurs, parmi lesquels 552 se déclarent en faveur de A. Déterminer l'estimation p_{obs} de p (sachant que tous les électeurs se prononcent), puis un intervalle à 95% puis à 99% de p.

2) Soit P la variable aléatoire égale à la prévision du score de A à ce second tour, que l'on peut effectuer à l'aide du sondage précédent. On admet que P suit une loi normale.

a) Déterminer les paramètres de la loi de P.

b) Déterminer enfin la probabilité que A gagne cette élection.

Exercice n°6

Pour déterminer la teneur en potassium d'une solution, on effectue des dosages à l'aide d'une technique expérimentale donnée. On admet que le résultat d'un dosage est une variable aléatoire suivant une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ dont l'espérance μ est la valeur que l'on cherche à déterminer, et dont l'écart-type σ est de 1 mg/litre si l'on suppose que le protocole expérimental a été suivi scrupuleusement.

Les résultats de cinq dosages indépendants, réalisés en suivant rigoureusement le protocole expérimental, sont les suivants (en mg/litre) : 74,0, 71,6, 73,4, 74,3, 72,2.

1) Déterminer, à partir de ces mesures, un intervalle de confiance pour μ de niveau de confiance 95%, et calculer l'intervalle observé.

2) Quelle taille d'échantillon est nécessaire pour avoir, au même niveau de confiance, un intervalle de longueur inférieure à 0,1 mg/litre ?

Exercice n°7

Lors d'un sondage effectué en Île-de-France auprès de 550 personnes, il est apparu que 42 avaient de l'asthme. On se propose d'estimer par intervalle de confiance la probabilité p d'avoir de l'asthme en Île-de-France. On note Z_i la variable aléatoire qui vaut 1 si la i-ème personne de l'échantillon sondé est atteinte, et 0 sinon.

On admet que les variables $Z_1, \ldots, Z_n$ sont indépendantes et de même loi. On note $\overline{Z_n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Z_i$.

Comme $\overline{Z_n}$ est un estimateur consistant de p, on peut approcher la variance des Z_i par $\overline{Z_n}(1 - \overline{Z_n})$.

1) Rappeler le théorème central limite.

2) Calculer des intervalles de confiance pour p, basés sur l'approximation gaussienne, de coefficients de sécurité asymptotiques 90%, 95% puis 99%. Calculer les intervalles observés.

Exercice n°8

Soit X une VAR d'espérance m et de variance v.

On dispose d'un n-échantillon $(X_1, \ldots, X_n)$ de X. On appelle variance empirique de X la variable $W_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2$, où $\overline{X_n}$ est la moyenne empirique de X.

1. Calculer $E(\overline{X_n})$ et $V(\overline{X_n})$.

2. En déduire $E(\overline{X_n}^2)$.

3. Calculer $E(W_n)$, et en déduire un estimateur sans biais de v.

Exercice n°9

Soit X une VAR de loi uniforme sur un intervalle $[0, a]$, où a est un paramètre inconnu, et on dispose de $(X_1, \ldots, X_n)$ un n-échantillon de X. On note $\overline{X_n}$ la moyenne empirique de X.

Rappel : si X est une VAR de loi uniforme sur $[0, a]$, alors sa fonction de répartition est :

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x/a & \text{si } 0 \leq x \leq a \\ 1 & \text{si } x \geq a \end{cases}$$

1. Soit $T_n = 2\overline{X_n}$. Montrer que T_n est un estimateur sans biais de a.
2. Calculer son risque quadratique.
3. Soit $T'_n = \max(X_1, \ldots, X_n)$. Donner la fonction de répartition de T'_n.
4. En déduire une densité de T'_n, puis son biais et son risque quadratique.
5. Soit $T''_n = \dfrac{n+1}{n} T'_n$. Déterminer son biais et son risque quadratique.
   Pour de grandes valeurs de n, quel est le meilleur estimateur de a ?

Exercice n°10

Lors d'un sondage sur 100 personnes interrogées, 60 pensent voter pour A.

On modélise ce résultat par un échantillon $(X_1, X_2, \ldots, X_{100})$ de variables indépendantes de même loi de Bernoulli de paramètre p. On cherche à déterminer un intervalle de confiance pour p au niveau de confiance 99%.

1. Déterminer l'espérance et la variance de la moyenne empirique $F = \dfrac{1}{100} \displaystyle\sum_{i=1}^{100} X_i$.
2. On note F^* la variable centrée réduite associée à F. Par quelle loi peut-on approcher celle de F^* ? Déterminer t tel que $P(-t \leq F^* \leq t) \geq 0{,}99$.
3. En déduire que $P\left(F - t\dfrac{\sqrt{p(1-p)}}{10} \leq p \leq F + t\dfrac{\sqrt{p(1-p)}}{10}\right) \geq 0{,}99$.
4. Montrer que pour tout $p \in [0,1]$, $p(1-p) \leq \dfrac{1}{4}$, et en déduire un intervalle de confiance pour p au niveau de confiance 0,99, puis en donner une estimation.

Exercice n°11

Afin d'étudier la proportion p de consommateurs satisfaits par un produit, on a interrogé 100 consommateurs. 56 d'entre eux ont déclaré être satisfaits par le produit.

Donner un intervalle de confiance à 95% de p.

Exercice n°12

Une usine fabrique des câbles. On suppose que la charge maximale supportée par un câble, exprimée en tonnes, est une variable aléatoire suivant une loi normale d'espérance m et de variance 0,5.

Une étude portant sur 50 câbles a donné une moyenne des charges maximales supportées égale à 12,2 tonnes.

1. Déterminer l'intervalle de confiance à 99% de la charge maximale moyenne de tous les câbles fabriqués par l'usine.
2. Quelle doit être la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance à 99% soit inférieure ou égale à 0,2 ?

Exercice n°13

Les individus d'une population possèdent un caractère X qui suit une loi de probabilité dont la densité est donnée par $f_\theta(x) = kx^2/\theta^3$ pour $x \in [0,\theta]$, et $f_\theta(x) = 0$ sinon. On s'intéresse au paramètre $\theta > 0$.

1. Déterminer la constante k pour que f_θ soit une densité.
2. On se donne $(X_1, \ldots, X_n)$ un échantillon de même loi que X. Montrer que la moyenne $\overline{X_n}$ n'est pas un estimateur sans biais de θ.
3. En déduire un estimateur sans biais de θ.

Exercice n°14

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes ayant pour densité : $f_\theta(x) = \dfrac{1}{2}(1+\theta x)\,\mathbf{1}_{[-1,1]}(x)$.

1) Quelle contrainte doit-on imposer à θ pour que f_θ soit bien une densité de probabilité ?
2) Par la méthode des moments, donner un estimateur $\widehat{\theta}$ de θ.
3) Peut-on dire que $\widehat{\theta}$ est un bon estimateur de θ ?

Exercice n°15 *(oraux ESSEC)*

Soit θ un paramètre réel strictement positif.

Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une variable aléatoire ayant pour densité la fonction f définie par :

$$\forall t \in \mathbb{R}, \qquad f(t) = \frac{2t}{\theta^2}\,\mathbf{1}_{[0,\theta]}(t)$$

On considère $(X_i)_{i\in\mathbb{N}^*}$ une suite de variables aléatoires définies sur cet espace, indépendantes et identiquement distribuées, de même loi que X.

Pour tout $n \in \mathbb{N}^*$, on note $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

1) a) Vérifier que f est une densité de probabilité. On notera X une variable aléatoire à densité de densité f.
   b) Déterminer la fonction de répartition de la variable X.
   c) Montrer que X admet une espérance et une variance, et les calculer.
2) a) Calculer $E(\overline{X_n})$ et $V(\overline{X_n})$.
   b) Construire, à partir de $\overline{X_n}$, un estimateur sans biais convergent du paramètre θ. On le notera T_n.
3) Appliquer le théorème central limite à la variable $\overline{X_n}$. En déduire un intervalle de confiance asymptotique de θ, au niveau de confiance 95%.

   Si on note Φ la fonction de répartition d'une variable aléatoire qui suit une loi $\mathcal{N}(0,1)$, on donne, pour $t = 1{,}96$: $2\Phi(t) - 1 = 0{,}95$.
4) Soit G_n la variable aléatoire définie par : $G_n = \sup(X_1, \ldots, X_n)$.
   a) Déterminer une densité de la variable G_n.
   b) Calculer l'espérance et la variance de G_n, puis montrer que $\lim\limits_{n\to+\infty} E(G_n) = \theta$.
   c) Construire, à l'aide de G_n, un estimateur sans biais et convergent de θ, noté W_n.

Exercice n°16
Soient $X_1, \dots, X_n$ n variables aléatoires indépendantes identiquement distribuées de loi de Poisson de paramètre $\lambda > 0$.

Rappelons que pour tout entier positif k, $P(X_1 = k) = \dfrac{\lambda^k}{k!}e^{-\lambda}$, $E[X_1] = \lambda$ et $E[X_1^2] = \lambda + \lambda^2$.

Proposer un estimateur sans biais de λ et calculer son risque quadratique.

Exercice n°17
Soient $X_1, \dots, X_n$ n variables aléatoires indépendantes identiquement distribuées de densité : $p_\theta(x) = \exp(\theta - x)\,\mathbf{1}_{\{x \geq \theta\}}$, $\theta > 0$.

On cherche dans cet exercice à estimer le paramètre de translation θ.

1. Proposer un estimateur sans biais de θ par la méthode des moments : on le notera $\widehat{\theta_n}$.
2. Calculer le risque quadratique de l'estimateur $\widehat{\theta_n}$.
3. On considère à présent l'estimateur $\widetilde{\theta_n}$ défini par $\widetilde{\theta_n}(X) = \inf_{1 \leq i \leq n} X_i$. Calculer sa loi.
4. L'estimateur $\widetilde{\theta_n}$ est-il sans biais ?
5. Calculer le risque quadratique de l'estimateur $\widetilde{\theta_n}$.

Exercice n°18
Soient $X_1, \dots, X_n$ n variables aléatoires indépendantes de loi $B(\theta)$ (Bernoulli de paramètre θ).

1. Montrer que $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est un estimateur de θ efficace.
2. Trouver un estimateur sans biais de la variance des X_i, de la forme $\widehat{v} = \eta\,\overline{X}(1 - \overline{X})$.

Exercice n°19 *(oral ESSEC)*
Afin d'avoir une idée du nombre de poissons N présents dans un étang, on en pêche une certaine quantité K, que l'on marque, puis que l'on remet à l'eau. On revient un jour plus tard, et l'on pêche n poissons (avec remise, pour simplifier). Pour tout $i \in [\![1, n]\!]$, on note X_i la variable qui vaut 1 si le i-ème poisson pêché est marqué, et 0 sinon. On suppose que ces variables sont indépendantes, et on introduit la moyenne empirique $\overline{X_n}$.

(1) Déterminer la loi de X_i, puis calculer l'espérance et la variance de $\overline{X_n}$.

(2) Montrer que, si on note $p = K/N$, alors :

$$\sqrt{\frac{n}{p(1-p)}}\,(\overline{X_n} - p) \xrightarrow{\mathcal{L}} X, \qquad X \hookrightarrow \mathcal{N}(0,1).$$

(3) Déterminer alors un intervalle de confiance asymptotique de risque α pour $p = K/N$, construit sur $\overline{X_n}$. En déduire un intervalle de confiance asymptotique de même risque pour N, construit sur $\overline{X_n}$.

(4) Application numérique : $K = 50$, $n = 300$, $\overline{X_n} = 0{,}6$, $\alpha = 0{,}05$.

Exercice n°20 *(Edhec)*
Soit X une v.a.r. réelle admettant une espérance m et une variance σ^2. Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon de la v.a.r. X.

1. Montrer que $T_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur sans biais de m.
2. On pose $V_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(X_k - T_n)^2$.

a) Montrer que $V_n = \dfrac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - (T_n)^2$.

b) Montrer que $V_n = \dfrac{1}{n}\left(\sum_{k=1}^{n} (X_k - m)^2\right) - (T_n - m)^2$.

c) Montrer que $\mathbb{E}(V_n) = \dfrac{n-1}{n}\,\sigma^2$.

d) Construire, à partir de V_n, un estimateur sans biais $\widehat{V_n}$ de σ^2.

e) On dispose d'un échantillon de n observations $(x_1, x_2, \ldots, x_n)$ de la v.a.r. X. Donner une méthode pour obtenir une estimation ponctuelle de σ^2 à partir de ces observations.

Exercice n°21 *(extrait ESSEC 2007 — Maths III)*

La sécurité routière fait une enquête sur le nombre d'accidents survenus par semaine sur un tronçon d'autoroute. Soit X la v.a.r. égale au nombre d'accidents par semaine. On suppose que X suit une loi de Poisson de paramètre θ inconnu ($\theta \in]0, +\infty[$). On se propose d'évaluer le paramètre $e^{-\theta} = \mathbb{P}([X = 0])$.

On note $X_1, X_2, \ldots, X_n$ les résultats des observations faites pendant n semaines. On suppose $X_1, \ldots, X_n$ indépendantes et de même loi que X.

1. Pour tout $i \in [\![1, n]\!]$, on définit Y_i par : $Y_i = 1$ si $X_i = 0$, et $Y_i = 0$ sinon. On note aussi $\overline{Y_n} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.

   a) Pour tout $i \in [\![1, n]\!]$, donner la loi de Y_i.

   b) Montrer que $\overline{Y_n}$ est un estimateur sans biais de $e^{-\theta}$.

   c) Calculer le risque quadratique de $\overline{Y_n}$.

   d) Montrer que $\overline{Y_n}$ est un estimateur convergent de $e^{-\theta}$.

   e) Expliquer pourquoi $\overline{Y_n}$ est un estimateur « naturel » de $e^{-\theta}$. Cet estimateur ne tient toutefois pas compte du fait que X suit une loi de Poisson : on peut donc espérer trouver un meilleur estimateur sans biais convergent de $e^{-\theta}$.

2. On pose $S_n = \sum_{i=1}^{n} X_i$.

   a) Quelle est la loi de S_n ?

   b) Calculer l'espérance de $e^{-S_n/n}$ à l'aide du théorème de transfert.

   c) Montrer que $e^{-S_n/n}$ est un estimateur biaisé de $e^{-\theta}$.

   d) Montrer que $e^{-S_n/n}$ est asymptotiquement sans biais, c'est-à-dire que $\lim_{n\to+\infty} \mathbb{E}\big(e^{-S_n/n}\big) = e^{-\theta}$.