

TD2 : Échantillonnage et estimation



Version web

Ex 1

Propriétés fondamentales des estimateurs

fJtB

Soit $(X_1, X_2, \dots, X_n)$ un n -échantillon d'une variable aléatoire X d'espérance $\mu = \mathbb{E}(X)$ et de variance $\sigma^2 = \text{Var}(X)$ finies. On définit les estimateurs suivants :

$$T_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad T_2 = \frac{1}{n-1} \sum_{i=1}^n X_i \quad T_3 = X_1$$

1. Rappeler les définitions d'un estimateur sans biais et d'un estimateur convergent.



Estimateur sans biais : Un estimateur $\hat{\theta}_n$ du paramètre θ est dit sans biais si $\mathbb{E}(\hat{\theta}_n) = \theta$ pour tout n .

Estimateur convergent : Un estimateur $\hat{\theta}_n$ est dit convergent si $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta$ (convergence en probabilité vers θ).

2. Les trois estimateurs T_1 , T_2 et T_3 sont-ils des estimateurs sans biais de μ ? Justifier.



Pour T_1 : $\mathbb{E}(T_1) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n} \times n\mu = \mu$

Donc T_1 est sans biais.

Pour T_2 : $\mathbb{E}(T_2) = \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}(X_i) = \frac{n\mu}{n-1} = \mu \times \frac{n}{n-1} \neq \mu$

Donc T_2 est biaisé. Le biais est $b(T_2) = \frac{\mu}{n-1}$.

Pour T_3 : $\mathbb{E}(T_3) = \mathbb{E}(X_1) = \mu$

Donc T_3 est sans biais.

3. Calculer la variance de chacun des trois estimateurs.



Pour T_1 :

$$\text{Var}(T_1) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

Pour T_2 :

$$\text{Var}(T_2) = \frac{1}{(n-1)^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{n\sigma^2}{(n-1)^2}$$

Pour T_3 :

$$\text{Var}(T_3) = \text{Var}(X_1) = \sigma^2$$

4. Parmi les estimateurs sans biais de μ , lequel est le plus efficace? Justifier.



Les estimateurs sans biais sont T_1 et T_3 .

On a $\text{Var}(T_1) = \frac{\sigma^2}{n}$ et $\text{Var}(T_3) = \sigma^2$.

Pour tout $n \geq 2$, on a $\frac{\sigma^2}{n} < \sigma^2$, donc $\text{Var}(T_1) < \text{Var}(T_3)$.

Par conséquent, T_1 (la moyenne empirique) est l'estimateur le plus efficace parmi les estimateurs sans biais proposés.

5. On définit l'erreur quadratique moyenne (EQM) d'un estimateur $\hat{\theta}$ de θ par :

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Montrer que $\text{EQM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + b^2(\hat{\theta})$ où $b(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$ est le biais.



On pose $b = \mathbb{E}(\hat{\theta}) - \theta$. Alors :

$$\begin{aligned}\text{EQM}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] + (\mathbb{E}(\hat{\theta}) - \theta)^2\end{aligned}$$

Le terme du milieu vaut :

$$2(\mathbb{E}(\hat{\theta}) - \theta) \times \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})] = 2(\mathbb{E}(\hat{\theta}) - \theta) \times 0 = 0$$

D'où : $\text{EQM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + b^2(\hat{\theta})$

6. Calculer l'erreur quadratique moyenne de chacun des trois estimateurs. Quel est le meilleur estimateur de μ au sens de l'EQM ?



Pour T_1 : $\text{EQM}(T_1) = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}$

Pour T_2 : $\text{EQM}(T_2) = \frac{n\sigma^2}{(n-1)^2} + \left(\frac{\mu}{n-1}\right)^2 = \frac{n\sigma^2 + \mu^2}{(n-1)^2}$

Pour T_3 : $\text{EQM}(T_3) = \sigma^2 + 0^2 = \sigma^2$

Pour comparer les estimateurs, on analyse leur EQM pour $n \geq 2$ (en supposant $\sigma^2 > 0$).

Comparaison de T_1 et T_3 : On a clairement $\text{EQM}(T_1) = \frac{\sigma^2}{n} < \sigma^2 = \text{EQM}(T_3)$. Donc T_1 est meilleur que T_3 .

Comparaison de T_1 et T_2 : Pour $n \geq 2$, on a $n^2 > (n-1)^2$, ce qui implique $\frac{n}{(n-1)^2} > \frac{1}{n}$. Par conséquent, $\text{Var}(T_2) = \frac{n\sigma^2}{(n-1)^2} > \frac{\sigma^2}{n} = \text{EQM}(T_1)$. Comme $\text{EQM}(T_2) = \text{Var}(T_2) + b(T_2)^2$ et que le terme de biais au carré est positif ou nul, il s'ensuit que $\text{EQM}(T_2) > \text{EQM}(T_1)$.

Conclusion : En comparant les trois, T_1 est l'estimateur qui possède la plus faible erreur quadratique moyenne pour tout $n \geq 2$.

Ex 2

Lois statistiques et étude d'estimateurs

3kdw

On considère un échantillon (X_i) de taille $n = 5$ dans une population suivant une loi normale de paramètres μ et σ^2 . On pose

$$T_1 = \frac{1}{5} \sum_{i=1}^5 X_i \quad T_2 = \frac{1}{5}(X_1 + X_2) + \frac{1}{4}(X_3 + X_4 + X_5) \quad T_3 = \frac{1}{10}(2X_1 + 3X_2) + \frac{1}{8}(X_3 + 2X_4 + X_5)$$

$$U = \frac{1}{\sigma^2} \sum_{i=1}^5 (X_i - \mu)^2 \quad V = \frac{1}{\sigma^2} \sum_{i=1}^5 (X_i - T_1)^2$$

1. Quelle est la loi suivie par la variable $X_1 - X_2$? Justifier.



D'après le cours, $X_1 - X_2$ suit une loi normale d'espérance $\mathbb{E}(X_1 - X_2) = \mu - \mu = 0$. Par indépendance, $V(X_1 - X_2) = V(X_1) + (-1)^2 V(X_2) = 2\sigma^2$.

2. On cherche à estimer μ à l'aide des estimateurs T_1, T_2, T_3 . Étudier leur biais et comparer l'efficacité des estimateurs sans biais.



Par linéarité de l'espérance, on calcule $\mathbb{E}(T_1) = \frac{5\mu}{5} = \mu$, $\mathbb{E}(T_2) = \frac{2\mu}{5} + \frac{3\mu}{4}$, $\mathbb{E}(T_3) = \mu$. Par conséquent, $B(T_1) = B(T_3) = 0$ et $B(T_2) = \mathbb{E}(T_2) - \mu = \frac{3\mu}{20}$.

Pour comparer l'efficacité des deux estimateurs sans biais, on calcule leur EQM (ce qui revient à calculer leur variance.) Par indépendance des variables, on a :

$V(T_1) = \frac{\sigma^2}{5} < V(T_3) = \frac{147\sigma^2}{800}$. Le plus efficace est donc l'estimateur T_1 qui est la moyenne empirique.

3. Quelle est la loi suivie par la variable U ? la variable V ? justifier.



$U = \sum_{i=1}^5 \left(\frac{X_i - \mu}{\sigma} \right)^2$; or les X_i sont des variables aléatoires indépendantes et $\frac{X_i - \mu}{\sigma}$ suit une loi $\mathcal{N}(0, 1)$ donc par définition, U suit une loi de $\chi^2(5)$.

De plus, $T_1 = \bar{X}$ est l'estimateur de moyenne empirique donc d'après le théorème de Fisher, V suit une loi de $\chi^2(5 - 1) = \chi^2(4)$.

4. Déterminer $x \in \mathbb{R}$ tel que $P(U > x) = 0,05$.



On lit dans la table de loi $P(U < x) = 0,95$ pour $x = 11,07$.

5. En utilisant T_1 et U , construire une variable Y qui suive une loi de Student dont on précisera le paramètre.



On pose $Z = \frac{T_1 - \mu}{\frac{\sigma}{\sqrt{5}}}$ variable distribuée selon une loi $\mathcal{N}(0, 1)$. Soit alors $Y = \frac{Z}{\sqrt{\frac{U}{5}}}$: par définition, Y suit une loi $St(5)$. Après simplification, on peut réécrire $Y = \frac{T_1 - \mu}{\frac{\sigma}{\sqrt{5}}}$.

Ex 3

Maximum de vraisemblance pour une loi géométrique

jemd

On rappelle qu'une variable X suit une loi géométrique de paramètre $p \in]0; 1[$ si X est à valeurs dans $\mathbb{N}^* = \{1; 2; 3; \dots\}$ et si pour tout $k \in \mathbb{N}^*$,

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

On cherche à estimer ce paramètre p à partir d'un échantillon.

- On considère un échantillon $(X_1, X_2, X_3, X_4, X_5)$ ayant pour loi mère une loi géométrique de paramètre p et on suppose que la suite $(3; 4; 4; 2; 3)$ constitue une réalisation de cet échantillon.
 - Exprimer la fonction de vraisemblance associée à cet échantillon.



D'après le cours, la fonction de vraisemblance associée à cet échantillon est donnée par :

$$\begin{aligned} L(p) &= P(X_1 = 3) \times P(X_2 = 4) \times P(X_3 = 4) \times P(X_4 = 2) \times P(X_5 = 3) \\ &= p(1-p)^2 \times p(1-p)^3 \times p(1-p)^3 \times p(1-p) \times p(1-p)^2 \\ &= p^5(1-p)^{11} \end{aligned}$$

- (b) En déduire une estimation de p par la méthode du maximum de vraisemblance.



On cherche à maximiser la fonction L sur $]0; 1[$. Pour cela, on calcule la dérivée de L :

$$\begin{aligned} L'(p) &= 5p^4(1-p)^{11} - 11p^5(1-p)^{10} \\ &= p^4(1-p)^{10}(5-11p) \end{aligned}$$

La fonction L est dérivable sur $]0; 1[$ et $L'(p) = 0$ si et seulement si $p = 0$, $p = 1$ ou $p = \frac{5}{11}$.
Or, $L(0) = 0$, $L(1) = 0$ et $L(\frac{5}{11}) > 0$.

Donc, L admet un maximum en $p = \frac{5}{11}$.

En conclusion, la valeur la plus vraisemblable pour p est $\frac{5}{11}$. Il s'agit donc d'une estimation du paramètre p par la méthode du maximum de vraisemblance.

2. Afin de déterminer un estimateur de p , on considère maintenant un n -échantillon $(X_1, \dots, X_n)$ ayant pour loi mère une loi géométrique de paramètre p et n entiers non nuls $(x_1, \dots, x_n)$ constituant une réalisation de cet échantillon.

- (a) Exprimer la fonction de vraisemblance associée à cet échantillon.



$$\begin{aligned} L(x_1, \dots, x_n, p) &= P(X_1 = x_1) \times \dots \times P(X_n = x_n) \\ &= p(1-p)^{x_1-1} \times \dots \times p(1-p)^{x_n-1} \\ &= p^n(1-p)^{x_1+\dots+x_n-n} \end{aligned}$$

- (b) En utilisant la méthode du maximum de vraisemblance, déterminer un estimateur du paramètre p .



On cherche à maximiser la fonction L sur $]0; 1[$. Pour cela, on calcule la dérivée partielle de L par rapport à p :

$$\begin{aligned} \frac{\partial L}{\partial p}(x_1, \dots, x_n, p) &= np^{n-1}(1-p)^{x_1+\dots+x_n-n} - p^n(x_1+\dots+x_n-n)(1-p)^{x_1+\dots+x_n-n-1} \\ &= p^{n-1}(1-p)^{x_1+\dots+x_n-n-1}(n-(x_1+\dots+x_n-n)p) \end{aligned}$$

La fonction L est dérivable sur $]0; 1[$ et $\frac{\partial L}{\partial p}(p) = 0$ si et seulement si $p = 0$, $p = 1$ ou $p = \frac{n}{\sum_{i=1}^n x_i}$.

Or, $L(0) = 0$, $L(1) = 0$ et $L\left(\frac{n}{\sum_{i=1}^n x_i}\right) > 0$.

Donc, L admet un maximum en $p = \frac{n}{\sum_{i=1}^n x_i}$.

Ceci étant vrai pour toute réalisation $(x_1, \dots, x_n)$ de l'échantillon, on en déduit que $\hat{p} = \frac{n}{\sum_{i=1}^n X_i}$ est un estimateur du paramètre p par la méthode du maximum de vraisemblance.

Ex 4

Estimation du paramètre d'une loi

dXNn

On fixe un réel $a > 0$ et on définit une variable aléatoire X sur $\mathbb{N}$ dont la loi de probabilité est définie par

$$P(X = k) = \frac{a^k}{(1+a)^{k+1}}$$

pour tout $k \in \mathbb{N}$.

A l'aide de la méthode du maximum de vraisemblance, définir un estimateur de a et déterminer son biais.



On définit un échantillon $X_1, \dots, X_n$ et on considère la probabilité que cet échantillon réalise un ensemble de valeurs $V = \{x_1, \dots, x_n\}$. La fonction de vraisemblance s'écrit alors

$$L(x_1, \dots, x_n, a) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_{k=1}^n P(X = x_k) = \prod_{k=1}^n \frac{a^{x_k}}{(1+a)^{x_k+1}} = \frac{a^{\sum x_k}}{(1+a)^{\sum x_k+1}}$$

On dérive ce quotient par rapport à a , on factorise par $a^{\sum x_k-1}(1+a)^{n+\sum x_k-1}$ et on trouve que

$$\frac{\partial L}{\partial a}(x_1, \dots, x_n, a) = 0 \iff \sum_{k=0}^n x_k(1+a) - a(n + \sum_{k=0}^n x_k) = 0 \iff a = \frac{1}{n} \sum_{k=0}^n x_k$$

On a donc trouvé un meilleur estimateur du paramètre a selon la méthode du maximum de vraisemblance : il s'agit de l'estimateur $T = \frac{1}{n} \sum_{k=0}^n X_k$.

Reste à calculer le biais de cet estimateur, autrement dit à calculer $B(T) = \mathbb{E}(T - a)$. Or pour tout entier i , X_i suit la loi définie ci-dessus et on espère se calcule de la manière suivante :

$$\mathbb{E}(X_i) = \sum_{k=0}^n k P(X_i = k) = \sum_{k=0}^n k \left(\frac{a}{1+a} \right)^{k-1} \frac{a}{(1+a)^2} = \frac{1}{\left(1 - \frac{a}{1+a}\right)^2} \times \frac{a}{(1+a)^2} = a$$

Donc $B(T) = \frac{1}{n} \times n \times a - a = 0$: la variable T est donc un estimateur sans biais du paramètre a .

Ex 5

Estimation des paramètres d'une loi normale

PwkL

★ ★

On considère un échantillon $X_1, \dots, X_n$ suivant une loi normale $\mathcal{N}(\mu, \sigma)$. On cherche un estimateur de μ et de σ par la méthode du maximum de vraisemblance. On note $(x_1, \dots, x_n)$ une réalisation de cet échantillon. On rappelle que la densité d'une loi normale est

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

1. Exprimer la fonction de vraisemblance $L(x_1, \dots, x_n, \mu, \sigma)$, puis son logarithme.



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

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

2. Dériver $\ln L(x_1, \dots, x_n, \mu, \sigma)$ par rapport à μ .



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

3. En déduire un estimateur de μ .



$$\frac{\partial \ln L(x_1, \dots, x_n, \mu, \sigma)}{\partial \mu} = 0 \Leftrightarrow \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \sum_{i=1}^n x_i - n\mu = 0 \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

donc $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ est un estimateur de μ .

4. Déterminer un estimateur de σ avec une démarche analogue.



$$\frac{\partial \ln L(x_1, \dots, x_n, \mu, \sigma)}{\partial \sigma} = \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma}$$

$$\frac{\partial \ln L(x_1, \dots, x_n, \mu, \sigma)}{\partial \sigma} = 0 \Leftrightarrow \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0 \Leftrightarrow \sum_{i=1}^n (x_i - \mu)^2 = n\sigma^2$$

$$\Leftrightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

donc $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$ est un estimateur de σ .

Ex 6

Estimateurs et intervalles de confiance

0e0x

Un centre médical mène des études statistiques sur le taux de cholestérol de son personnel, mesuré en centigramme par litre de sang. Les observations sur 20 individus supposés sains, tirés au sort, sont les suivantes :

Taux de cholestérol en cg	120	160	200	240	280	320
Nombre de sujets observés	2	4	5	4	3	2

Le taux de cholestérol chez un individu sain (noté X) est supposé suivre une loi normale.

1. Donner une estimation $\bar{x}$ de l'espérance mathématique de X et une estimation s^2 de la variance de X . On précisera les estimateurs choisis et on en donnera les propriétés.



Avec l'estimateur usuel $\bar{X}$, on obtient à partir de la réalisation de cet échantillon de taille 20 une estimation $\bar{x} = 216$. Avec l'estimateur de variance corrigée S^2 , on obtient une estimation de la variance $s^2 \approx 3604,21$.

2. Donner l'intervalle de confiance de l'espérance de X , centré sur $\bar{x}$, au risque $\alpha = 0.01$.



Au vu de la taille de l'échantillon, la variable mère est supposée suivre une loi normale donc on utilise une loi de Student $St(19)$ pour calculer la réalisation de l'intervalle de confiance avec la formule du cours : on trouve $[177,5; 254,5]$.

Feuille de calcul

Ex 7

Intervalles de confiance

gk1A

Une association de consommateurs a effectué une enquête sur le prix d'un produit dans le supermarchés. Pour ce produit, les prix suivants (en euros) ont été relevés dans 7 supermarchés différents :

52 52 43 51 69 55 49

1. En supposant que le prix X de ce produit est distribué selon une loi normale $\mathcal{N}(\mu, \sigma)$, calculer une estimation $\bar{x}$ de l'espérance $\mathbb{E}(X)$ et une estimation s de l'écart-type $\sigma(X)$. On précisera les estimateurs choisis et on en donnera les propriétés.



Avec l'estimateur usuel $\bar{X}$, on obtient à partir de la réalisation de cet échantillon de taille 7 une estimation $\bar{x} = 53$. Avec l'estimateur de variance corrigée S^2 , on obtient une estimation de la variance $s^2 \approx 63,667$ d'où une estimation de l'écart type de 7.98.

2. Déterminer les intervalles de confiances symétriques au seuil de 90% et 99% centrés en $\mu = \mathbb{E}(X)$.



Au vu de la taille de l'échantillon, la variable mère est supposée suivre une loi normale donc on utilise une loi de Student $St(6)$ pour calculer la réalisation de l'intervalle de confiance avec la formule du cours :

Au seuil de 90%, on obtient l'intervalle $[47.1; 58.9]$;

Au seuil de 99%, on obtient l'intervalle $[41.8; 64.2]$.

Détail des calculs sur tableur

Ex 8

Intervalles de confiance

5cbY

★ ★ ★

Pour étudier le degré de pollution des eaux d'une rivière par les déchets d'une usine, on mesure la teneur en milligrammes d'un certain polluant par litre d'eau.

1. On effectue 17 mesures pour lesquelles les valeurs observées des teneurs en polluant ont une moyenne de 5.2 mg/L et un écart-type de 1.2 mg/L.
 - (a) Donner une estimation sans biais de la moyenne et de l'écart type de la teneur en polluant dans cette rivière.
 - (b) Donner un intervalle de confiance au seuil de 5% pour la teneur moyenne en polluant de cette rivière.
 - (c) A partir de ces résultats, peut-on considérer que l'usine respecte la législation en vigueur selon laquelle cette teneur moyenne en polluant ne doit pas dépasser 7 mg/L ?



- (a) La moyenne observée sur l'échantillon, $\bar{x} = 5.2$ mg/L, est une estimation ponctuelle et sans biais de la moyenne de la teneur en polluant. L'écart-type observé de 1.2 mg/L est une estimation biaisée. Pour obtenir une estimation non biaisée, on calcule d'abord la variance non biaisée (ou corrigée) $s^2 : s^2 = \frac{n}{n-1} \times (\text{écart-type obs.})^2 = \frac{17}{16} \times (1.2)^2 = 1.53$. L'estimation non biaisée de l'écart-type est donc $s = \sqrt{1.53} \approx 1.237$ mg/L.
- (b) L'écart-type de la population est inconnu et la taille de l'échantillon est faible ($n = 17 < 30$), on utilise donc une loi de Student à $n - 1 = 16$ degrés de liberté. On suppose que la distribution de la teneur en polluant est normale. Le quantile $t_{16;0.025}$ est d'environ 2.12. L'intervalle de confiance est donné par $\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}$. $I.C_{.95\%} = \left[5.2 - 2.12 \times \frac{1.237}{\sqrt{17}} ; 5.2 + 2.12 \times \frac{1.237}{\sqrt{17}} \right] \approx [5.2 - 0.636 ; 5.2 + 0.636]$. L'intervalle de confiance est donc environ **[4.56 ; 5.84] mg/L**.
- (c) La borne supérieure de l'intervalle de confiance, 5.84 mg/L, est inférieure à la limite légale de 7 mg/L. Au niveau de confiance de 95%, on peut donc considérer que l'usine respecte la législation.

2. On suppose ici l'écart type de la population connu et égal à 1.2 mg/L. Combien de mesures devrait-on faire pour estimer la teneur moyenne en polluant avec une précision de 0.1 mg/L au niveau de confiance 95% ?



L'écart-type de la population σ étant connu ($\sigma = 1.2$), on utilise la loi normale centrée réduite. La précision souhaitée $d = 0.1$ mg/L correspond à la marge d'erreur (ou demi-longueur de l'intervalle). La formule de la marge d'erreur est $d = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

On cherche n tel que $d \leq 0.1$. Pour un niveau de confiance de 95%, $z_{\alpha/2} = z_{0.025} \approx 1.96$.

On doit donc résoudre l'inéquation : $1.96 \times \frac{1.2}{\sqrt{n}} \leq 0.1$

$$\text{Soit } \sqrt{n} \geq \frac{1.96 \times 1.2}{0.1}$$

$$\sqrt{n} \geq 23.52$$

$$n \geq (23.52)^2$$

$$n \geq 553.19$$

Comme n doit être un entier, on doit effectuer au minimum **$n = 554$** mesures.

Détail des calculs sur tableur

Ex 9

Intervalle de confiance pour une proportion

iboj

★ ★ ★

On effectue un contrôle de fabrication sur des pièces dont la proportion p est défectueuse.

1. On contrôle un lot de 200 pièces et on trouve $n = 24$ pièces défectueuses. Donner les intervalles de confiance pour l'estimation de p , au risque $\alpha = 0.05$.



On commence par calculer la fréquence observée de pièces défectueuses : $f = \frac{n}{N} = \frac{24}{200} = 0.12$. Le risque étant $\alpha = 0.05$, le niveau de confiance est de $1 - \alpha = 0.95$. Pour un intervalle de confiance bilatéral, on cherche la valeur de $z_{\alpha/2}$ telle que $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$, où Z suit une loi normale centrée réduite. Pour un niveau de confiance de 95

On peut maintenant calculer les bornes de l'intervalle de confiance : La marge d'erreur est

$$E = z_{\alpha/2} \sqrt{\frac{f(1-f)}{N}} = 1.96 \sqrt{\frac{0.12(1-0.12)}{200}} \approx 1.96 \sqrt{\frac{0.1056}{200}} \approx 1.96 \times 0.02298 \approx 0.045.$$

L'intervalle de confiance est donc : $I_C = [0.12 - 0.045 ; 0.12 + 0.045] = [0.075 ; 0.165]$.

On peut donc estimer, avec un risque d'erreur de 5

2. Combien faudrait-il contrôler de pièces pour que, au risque $\alpha = 0.05$, l'intervalle de confiance centré soit d'amplitude 0.04 ?



L'amplitude A de l'intervalle de confiance est donnée par $A = 2 \times E = 2 \times z_{\alpha/2} \sqrt{\frac{f(1-f)}{n}}$. On souhaite une amplitude $A = 0.04$, avec un risque $\alpha = 0.05$, ce qui donne $z_{\alpha/2} \approx 1.96$.

On a donc l'équation : $0.04 = 2 \times 1.96 \sqrt{\frac{f(1-f)}{n}}$.

Pour la fréquence f , nous pouvons utiliser l'estimation de la question 1, soit $f = 0.12$. $0.04 = 3.92 \sqrt{\frac{0.12(1-0.12)}{n}}$ $\frac{0.04}{3.92} = \sqrt{\frac{0.1056}{n}}$ $(\frac{0.04}{3.92})^2 = \frac{0.1056}{n}$ $n = 0.1056 \times (\frac{3.92}{0.04})^2 = 0.1056 \times (98)^2 = 0.1056 \times 9604 \approx 1014.18$.

Comme le nombre de pièces doit être un entier, on arrondit à l'entier supérieur. Il faudrait donc contrôler $n = 1015$ pièces pour obtenir une amplitude d'intervalle de confiance de 0.04 en utilisant l'estimation de la proportion de 12

Alternative : Si on ne dispose d'aucune estimation pour la proportion p , on se place dans le cas le plus défavorable qui maximise l'incertitude. La variance $f(1-f)$ est maximale lorsque $f = 0.5$. $n = f(1-f) \times (\frac{2 \times z_{\alpha/2}}{A})^2$ $n = 0.5(1-0.5) \times (\frac{2 \times 1.96}{0.04})^2 = 0.25 \times (98)^2 = 0.25 \times 9604 = 2401$. Sans estimation préalable de la proportion, il faudrait contrôler 2401 pièces.

Ex 10

Étude complète d'un processus industriel

p7QA

★★★★

Une entreprise fabrique des câbles électriques dont la résistance R (en ohms) est modélisée par une variable aléatoire suivant une loi exponentielle de paramètre $\lambda > 0$. La densité de cette loi est donnée par :

$$f(x) = \lambda e^{-\lambda x} \quad \text{pour } x \geq 0$$

On rappelle que pour $X \sim \mathcal{E}(\lambda)$: $\mathbb{E}(X) = \frac{1}{\lambda}$ et $\text{Var}(X) = \frac{1}{\lambda^2}$.

Pour contrôler la qualité de production, on mesure la résistance de $n = 10$ câbles tirés au hasard. On obtient les valeurs suivantes (en ohms) :

2.3 1.8 3.1 2.7 1.5 2.9 2.1 3.4 2.5 1.9

Les normes de l'industrie exigent que la résistance moyenne des câbles soit au moins égale à 2.5 ohms, c'est-à-dire $\mathbb{E}(R) \geq 2.5$.

1. Exprimer la fonction de vraisemblance $L(r_1, \dots, r_{10}, \lambda)$ associée à cet échantillon, où $r_1, \dots, r_{10}$ désignent les valeurs observées.



La fonction de vraisemblance est le produit des densités :

$$L(r_1, \dots, r_{10}, \lambda) = \prod_{i=1}^{10} \lambda e^{-\lambda r_i} = \lambda^{10} e^{-\lambda \sum_{i=1}^{10} r_i}$$

2. En utilisant la méthode du maximum de vraisemblance, déterminer un estimateur $\hat{\Lambda}_n$ du paramètre λ à partir d'un n -échantillon $(X_1, \dots, X_n)$.



On a $\ln L = n \ln \lambda - \lambda \sum_{i=1}^n X_i$.

En dérivant : $\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n X_i = 0$

D'où l'estimateur du maximum de vraisemblance :

$$\hat{\Lambda}_n = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}_n}$$

3. L'estimateur $\hat{\Lambda}_n$ est-il sans biais ? Justifier.



L'estimateur est biaisé. Par l'inégalité de Jensen (car $x \mapsto 1/x$ est convexe sur $\mathbb{R}_+^*$), on a :

$$\mathbb{E}(\hat{\Lambda}_n) = \mathbb{E}\left(\frac{1}{\bar{X}_n}\right) \geq \frac{1}{\mathbb{E}(\bar{X}_n)} = \frac{1}{1/\lambda} = \lambda$$

Donc $\hat{\Lambda}_n$ est biaisé positivement. Cependant, il est asymptotiquement sans biais.

4. Calculer une estimation $\hat{\lambda}$ du paramètre λ à partir des données observées.



On calcule $\bar{r} = \frac{1}{10} \sum_{i=1}^{10} r_i = \frac{24.2}{10} = 2.42$ ohms.

Donc $\hat{\lambda} = \frac{1}{2.42} \approx 0.413$.

5. En déduire une estimation ponctuelle de l'espérance $\mathbb{E}(R)$ de la résistance.



L'estimation de l'espérance est $\widehat{\mathbb{E}(R)} = \frac{1}{\hat{\lambda}} = 2.42$ ohms.

6. On admet que pour un n -échantillon de loi exponentielle $\mathcal{E}(\lambda)$, la variable $2\lambda n \bar{X}_n$ suit une loi du chi-deux à $2n$ degrés de liberté. Déterminer un intervalle de confiance à 95% pour le paramètre λ .



On a $\mathbb{P}\left(\frac{\chi_{0.025, 20}^2}{2n\bar{X}_n} \leq \lambda \leq \frac{\chi_{0.975, 20}^2}{2n\bar{X}_n}\right) = 0.95$

Avec $n = 10$, $\bar{r} = 2.42$, $\chi_{0.025, 20}^2 \approx 9.59$ et $\chi_{0.975, 20}^2 \approx 34.17$, on obtient :

$$IC_{95\%}(\lambda) = \left[\frac{9.59}{20 \times 2.42}, \frac{34.17}{20 \times 2.42} \right] \approx [0.198, 0.706]$$

7. En déduire un intervalle de confiance à 95% pour l'espérance $\mathbb{E}(R)$.



Puisque $\mathbb{E}(R) = 1/\lambda$, on inverse les bornes :

$$IC_{95\%}(\mathbb{E}(R)) = \left[\frac{1}{0.706}, \frac{1}{0.198} \right] \approx [1.42, 5.05] \text{ ohms}$$

8. Au vu de ces résultats, peut-on affirmer avec 95% de confiance que l'entreprise respecte les normes de l'industrie ($\mathbb{E}(R) \geq 2.5$ ohms) ? Justifier votre réponse.



Non, on ne peut pas l'affirmer avec 95% de confiance. En effet, l'intervalle de confiance $[1.42, 5.05]$ contient des valeurs inférieures à 2.5 ohms. La borne inférieure (1.42 ohms) est même nettement inférieure à la norme exigée.

Bien que l'estimation ponctuelle $\widehat{\mathbb{E}(R)} = 2.42$ ohms soit proche de 2.5 ohms, l'incertitude statistique (reflétée par l'intervalle de confiance) ne permet pas de conclure que la norme est respectée avec un niveau de confiance de 95%.

L'entreprise devrait soit augmenter la taille de l'échantillon pour réduire l'incertitude, soit améliorer son processus de fabrication pour augmenter la résistance moyenne.