Exercices : Estimations

Exercice 1:

Soit X de loi $\mathcal{U}([0;a])$ et $(X_1, \cdots, X_n)$ un n-échantillon de X. On cherche à estimer a. On note $\overline{X_n}$ la moyenne empirique de X.

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

Exercice 2:

Soit X ayant une espérance m et une variance v, et soit $(X_1, \cdots, X_n)$ un n-échantillon 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. Soit Y ayant une espérance et une variance. Exprimer $E(Y^2)$ en fonction de $V(Y)$ et $E(Y)$.
2. Calculer $E(\overline{X_n})$ et $V(\overline{X_n})$ et en déduire $E(\overline{X_n}^2)$.
3. Calculer alors $E(W_n)$ et en déduire un estimateur sans biais de v.

Exercice 3:

Un séricicultеur a pesé 100 cocons provenant de son élevage de vers à soie. Il a obtenu les résultats suivants :

Poids en g	0,64	0,65	0,66	0,67	0,68	0,69	0,70	0,71
Effectif	3	5	2	6	6	10	12	10

Poids en g	0,72	0,73	0,74	0,75	0,76	0,77	0,78	0,79
Effectif	9	8	8	6	5	4	3	3

On appelle m la moyenne des poids de tous les cocons de l'élevage et on suppose que le poids X d'un cocon suit une loi normale de variance $0,0016g^2$.

1. Préciser les paramètres de cette loi normale.
2. On choisit au hasard un échantillon de 100 cocons. On appelle X_i le poids du i-ème cocon de cet échantillon. On suppose que $X_1, \cdots, X_{100}$ sont mutuellement indépendantes, de même loi que X. On pose $T = \dfrac{1}{100}\displaystyle\sum_{i=1}^{100} X_i$ et on admet que T suit une loi normale.
   Préciser les paramètres de cette loi normale et déterminer la variable centrée réduite T^* associée à T.
3. Déterminer un réel t tel que $P(-t \leqslant T^* \leqslant t) \geqslant 0,95$.
4. Montrer que $P(T - 0,00784 \leqslant m \leqslant T + 0,00784) \geqslant 0,95$.
5. En déduire un intervalle de confiance pour m au niveau de confiance $0,95$ et donner une estimation de cet intervalle.

Exercice 4:

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

On modélise le choix par un échantillon $(X_1, \cdots, X_{100})$ 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^* ? On suppose désormais que F^* suit la loi $\mathcal{N}(0,1)$.
3. Déterminer t tel que $P(-t \leqslant F^* \leqslant t) \geqslant 0,99$ et en déduire que
$$P\left(F - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0,99$$
4. Montrer que pour tout $p \in [0;1]$, $p(1-p) \leqslant \dfrac{1}{4}$ et en déduire un intervalle de confiance pour p au niveau de confiance $0,99$ puis en donner une estimation.

$\boxed{\textit{Correction}}$

Exercice 1:

1. T_n s'exprime en fonction des X_i donc c'est un estimateur de a. On cherche le biais de T_n donc il nous faut calculer son espérance.

$$E(T_n) = E(2\overline{X_n}) = 2 \times \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{2}{n}\sum_{i=1}^{n}\frac{a}{2} = a$$

On a donc $b(T_n) = E(T_n) - a = 0$ et ainsi T_n est un estimateur sans biais de a.
Comme T_n est un estimateur sans biais de a, son risque quadratique est égal à sa variance. Or comme les X_i sont indépendantes on a :

$$V(T_n) = V\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{4}{n^2}\sum_{i=1}^{n}\frac{a^2}{12} = \frac{a^2}{3n}$$

2. • D'après le cours la fonction de répartition de X est définie par $F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \dfrac{x}{a} & \text{si } 0 \leqslant x \leqslant a \\ 1 & \text{si } x > a \end{cases}$.

• Notons G_n la fonction de répartition de T'_n. Par définition $G_n(x) = P(T'_n \leqslant x)$. Or $[T_n \leqslant x] = [X_1 \leqslant x] \cap \cdots \cap [X_n \leqslant x]$. Donc comme les X_i sont indépendantes :

$$G_n(x) = P(X_1 \leqslant x) \times \cdots \times P(X_n \leqslant x) = (F(x))^n$$

On a donc $G_n(x) = \begin{cases} 0 & \text{si } x < 0 \\ \left(\dfrac{x}{a}\right)^n & \text{si } 0 \leqslant x \leqslant a \\ 1 & \text{si } x > a \end{cases}$.

• Une densité de T'_n s'obtient en dérivant G_n là où elle est dérivable et en complétant les points manquants :

$$g_n(t) = \begin{cases} 0 & \text{si } t < 0 \text{ ou } t > a \\ \dfrac{n}{a}\left(\dfrac{t}{a}\right)^{n-1} & \text{si } 0 \leqslant t \leqslant a \end{cases}$$

• Pour calculer le biais de T'_n il nous faut calculer son espérance. Comme la densité de T'_n est nulle en dehors de $[0;a]$ et que $t \to t g_n(t)$ est continue sur $[0;a]$, T'_n admet une espérance et :

$$E(T'_n) = \int_0^a t \times \frac{n}{a}\left(\frac{t}{a}\right)^{n-1} dt = \frac{n}{a^n}\int_0^a t^n\, dt = \frac{n}{a^n}\left[\frac{t^{n+1}}{n+1}\right]_0^a = \frac{na}{n+1}$$

Donc $b(T'_n) = \dfrac{na}{n+1} - a = \dfrac{-a}{n+1}$
• Pour les mêmes raison que pour l'espérance, T'_n admet une variance et :

$$\begin{aligned} V(T'_n) &= \int_0^a t^2 \times \frac{n}{a}\left(\frac{t}{a}\right)^{n-1} dt - E(T'_n)^2 = \frac{n}{a^n}\int_0^a t^{n+1}\, dt - \frac{n^2a^2}{(n+1)^2} \\ &= \frac{n}{a^n}\left[\frac{t^{n+2}}{n+2}\right]_0^a - \frac{n^2a^2}{(n+1)^2} = \frac{na^2}{n+2} - \frac{n^2a^2}{(n+1)^2} \\ &= \frac{n}{(n+2)(n+1)^2}a^2 \end{aligned}$$

On a donc :

$$r(T'_n) = V(T'_n) + b(T'_n)^2 = \frac{n}{(n+2)(n+1)^2}a^2 + \frac{a^2}{(n+1)^2}$$
$$= \frac{(2n+2)a^2}{(n+2)(n+1)^2} = \frac{2a^2}{(n+2)(n+1)}$$

3. On a vu que $E(T'_n) = \dfrac{na}{n+1}$ donc $E(T''_n) = a$ et ainsi T''_n est un estimateur sans biais de a.

   De plus comme $V(T'_n) = \dfrac{n}{(n+2)(n+1)^2}a^2$ on a $V(T''_n) = \dfrac{a^2}{n(n+2)}$ et donc $r(T''_n) = \dfrac{a^2}{n(n+2)}$.

4. T_n et T''_n sont meilleurs que T'_n car ils sont des estimateurs sans biais. De plus lorsque n tend vers $+\infty$ on a :

$$r(T_n) \underset{n\to+\infty}{\sim} \frac{a^2}{3} \times \frac{1}{n} \qquad \text{et} \qquad r(T''_n) \underset{n\to+\infty}{\sim} a^2 \times \frac{1}{n^2}$$

   Donc le meilleur estimateur est T''_n.

Exercice 2:

1. On sait que $V(Y) = E(Y^2) - E(Y)^2$ donc on a $E(Y^2) = V(Y) + E(Y)^2$.
2. On a :

$$E(\overline{X_n}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = m$$

   et comme les variables X_i sont indépendantes :

$$V(\overline{X_n}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} v = \frac{v}{n}$$

   Grâce à la question précédente on a donc $E(\overline{X_n}^2) = \dfrac{v}{n} + m^2$.

3. On a donc :

$$E(W_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X_n}^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(v+m^2) - \left(\frac{v}{n} + m^2\right)$$
$$= v + m^2 - \frac{v}{n} - m^2 = \frac{n-1}{n}v$$

   Si on pose $W'_n = \dfrac{n}{n-1}W_n$ alors on a bien que W'_n est un estimateur de v et de plus $E(W'_n) = \dfrac{n-1}{n}E(W_n) = v$ donc c'est bien un estimateur sans biais.

Exercice 3:

1. Les paramètres d'une loi normale sont son espérance et sa variance. $X \hookrightarrow \mathcal{N}(m; 0,0016)$
2. $E(T) = m$ et $V(T) = \dfrac{0,0016}{100}$ donc $T \hookrightarrow \mathcal{N}\left(m; \dfrac{0,0016}{100}\right)$.

   On a par définition :

$$T^* = \frac{T - E(T)}{\sqrt{V(T)}} = \frac{T-m}{\sqrt{0,0016/100}} = \frac{T-m}{0,004}$$

3. On sait donc que $T^* \hookrightarrow \mathcal{N}(0,1)$. Donc $P(-t \leqslant T^* \leqslant t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$.
   On a donc
   $$P(-t \leqslant T^* \leqslant t) \geqslant 0,95 \Leftrightarrow 2\Phi(t) - 1 \geqslant 0,95 \Leftrightarrow \Phi(t) \geqslant 0,975 \Leftrightarrow t \geqslant 1,96$$
   On peut donc choisir $t = 1,96$.
4. D'après la question précédente :
   $$P(-1,96 \leqslant T^* \leqslant 1,96) \geqslant 0,95 \Leftrightarrow P\left(-1,96 \leqslant \frac{T-m}{0,004} \leqslant 1,96\right) \geqslant 0,95$$
   $$\Leftrightarrow P\left(-0,004 \times 1,96 \leqslant T - m \leqslant 0,004 \times 1,96\right) \geqslant 0,95$$
   $$\Leftrightarrow P\left(T - 0,00784 \leqslant m \leqslant T + 0,00784\right) \geqslant 0,95$$
5. D'après la question précédente $[T - 0,00784; T + 0,00784]$ est un intervalle de confiance pour m au niveau de confiance $0,95$.
   Le tableau de valeurs nous permet de calculer une réalisation de T : $0,7132$
   Un intervalle de confiance réalisé est donc $[0,70536; 0,72104]$.

Exercice 4:

1. $E(F) = \dfrac{1}{100}\displaystyle\sum_{i=1}^{100} E(X_i) = p$ et, comme les X_i sont indépendantes,
   $V(F) = \dfrac{1}{10000}\displaystyle\sum_{i=1}^{100} V(X_i) = \dfrac{p(1-p)}{100}$.
2. D'après le théorème de la limite centrée, comme $100 > 30$, que les X_i sont indépendantes, suivent toutes la même loi et ont une variance non nulle, F^* suit approximativement la loi normale centrée réduite.
3. • On a $P(-t \leqslant F^* \leqslant t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$ donc
   $$P(-t \leqslant F^* \leqslant t) \geqslant 0,99 \Leftrightarrow 2\Phi(t) - 1 \geqslant 0,99 \Leftrightarrow \Phi(t) \geqslant 0,995 \Leftrightarrow t \geqslant 2,58$$
   On choisit $t = 2,58$.
   • De plus comme $F^* = \dfrac{F-p}{\sqrt{p(1-p)}/10}$, on a
   $$P(-t \leqslant F^* \leqslant t) = P\left(F - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\frac{\sqrt{p(1-p)}}{10}\right)$$
   On a donc bien $P\left(F - t\dfrac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\dfrac{\sqrt{p(1-p)}}{10}\right) \geqslant 0,99$
4. Il suffit ici d'étudier la fonction $f : x \to x(1-x)$ sur $[0;1]$ et de remarquer qu'elle atteint son maximum en $x = \dfrac{1}{2}$ et que ce maximum vaut $\dfrac{1}{4}$.
   On a donc $\left[F - t\dfrac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\dfrac{\sqrt{p(1-p)}}{10}\right] \subset \left[F - t\dfrac{1}{20} \leqslant p \leqslant F + t\dfrac{1}{20}\right]$ donc :
   $$P\left(F - t\frac{1}{20} \leqslant p \leqslant F + t\frac{1}{20}\right) \geqslant P\left(F - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0,99$$
   Donc $[F - 0,129 \leqslant p \leqslant F + 0,129]$ est un intervalle de confiance de p au niveau de confiance 0,99.
   De plus un réalisation de F est d'après l'énoncé : $0,6$ donc une estimation de cet intervalle est $[0,471; 0,729]$.