# Corrigé exam stats déc. 22

## 1. Exercice 1

a) Donner un exemple d'estimateur biaisé et d'un estimateur sans biais, en précisant la paramètre estimé.

$T_1 = \overline{X}$ est un estimateur sans biais de $\mu = E(X)$.

$S^2 = \dfrac{1}{n}\sum (X_i - \overline{X})^2$ est un estimateur biaisé de $\sigma^2 = Var(X)$.

b) Si $T_n$ est un estimateur du paramètre $\theta$, quelles sont les significations possibles de l'expression : $T_n$ est un estimateur convergent.

Cela signifie que $T_n$ converge vers $\theta$. Il s'agir d'une convergence d'une suite de v.a. vers la v.a. constante $= \theta$.

Il peut s'agir d'une convergence en probabilités, presque sûre ou en moyenne quadratique.

c) Si on considère un procès au tribunal comme un test d'hypothèses avec $(H_0):$ Innocence, et $(H_1):$ Culpabilité, exprimer les risques de 1ère et de 2ème espèce en terme de condamnation, d'innocent, de coupable, etc.

$\alpha =$ risque de 1ère espèce $=$ probabbilité de condamner un innocent.

$\beta =$ risque de 2ème espèce $=$ probabilité d'acquitter un coupable.

d) Comment se transforme un intervalle de confiance lorsqu'on augmente la taille de l'échantillon ? Lorsqu'on augmente le risque $\alpha$ ?

La diamètre d'un IDC est en général proportionnel à $\dfrac{1}{\sqrt{n}}$. Il sera donc plus étroit lorsque $n$ augmente.

Augmenter $\alpha$ c'est augmenter la probabilité de se trouver à l'extérieur de l'IDC. Celui-ci devient donc plus étroit.

## 2. Exercice 2

Soit $\theta$ un nombre réel vérifiant $|\theta| \leqslant \dfrac{1}{2}$, et $X$ une variable aléatoire continue de densité définie par

$$f(x) = \begin{cases} -4\theta x + 2\theta + 1 & \text{si } \quad x \in [0,1] \\ 0 & \text{sinon} \end{cases}$$

Soit $(X_1, \dots, X_n)$ un échantillon de variables indépendantes de même loi que $X$.

1. Montrer que $\mathbb{E}(X) = \dfrac{1}{2} - \dfrac{\theta}{3}$, et que $Var(X) = \dfrac{3 - 4\theta^2}{36}$.

- $\mathbb{E}(X) = \displaystyle\int_{-\infty}^{+\infty} x f(x)\,dx = \int_0^1 x(-4\theta x + 2\theta + 1)\,dx = \left[-4\theta\frac{x^3}{3} + \theta x^2 + \frac{x^2}{2}\right]_0^1$

$$\mathbb{E}(X) = \frac{1}{2} - \frac{\theta}{3}.$$

- $\mathbb{E}\left(X^2\right) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx = \left[-\theta x^4 + 2\theta\frac{x^3}{3} + \frac{x^3}{3}\right]_0^1 = \frac{1}{3} - \frac{\theta}{3}$

$$Var(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 = \frac{1}{3} - \frac{\theta}{3} - \left(\frac{1}{2} - \frac{\theta}{3}\right)^2 = \frac{3-4\theta^2}{36}.$$

2. Montrer que la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur biaisé de $\theta$.

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X) = \frac{1}{2} - \frac{\theta}{3} \implies b(\overline{X}) = \frac{1}{2} - \frac{\theta}{3} - \theta \neq 0$$

3. On se propose d'estimer le paramètre $\theta$ à l'aide de l'estimateur $T = a\,\overline{X} + b$ , où $a$ et $b$ sont des paramètres à fixer.
   Déterminer a et b pour que $T$ soit un estimateur sans biais de $\theta$.

$$\mathbb{E}(T) = \mathbb{E}(a\overline{X} + b) = a\left(\frac{1}{2} - \frac{\theta}{3}\right) + b = -\frac{a\theta}{3} + \frac{a}{2} + b$$

$$b(T) = 0 \iff \mathbb{E}(T) = \theta \iff \begin{cases} -\frac{a}{3} & = 1 \\ \frac{a}{2} + b & = 0 \end{cases} \iff \begin{cases} a & = -3 \\ b & = \frac{3}{2} \end{cases}$$

$$T = -3\overline{X} + \frac{3}{2}$$

4. Que peut-on dire de la convergence en **probabilité** de $T$.
   La loi des grands nombres assure que $\overline{X}$ converge en probabilités vers $\mathbb{E}(X)$.
   Par conséquent : $T = -3\overline{X} + \frac{3}{2}$ converge en proba. vers $\theta$.
5. Calculer le risque quadratique de $T$, et préciser sa limite lorsque $n \longrightarrow +\infty$.
   Comme $T$ est sans biais, on a :

$$R_\theta(T) = Var(T) = Var\left(-3\overline{X} + \frac{3}{2}\right) = 9Var(\overline{X}) = 9\frac{Var(X)}{n}$$

$$R_\theta(T) = \frac{3-4\theta^2}{4n}$$

6. Que peut-on en conclure ?

$$\lim_{n \to +\infty} R_\theta(T) = 0 \implies T \text{ converge en moyenne quadratique vers } \theta.$$

## 3. Exercice 3

En 1970, la Française moyenne mesurait 160,4 cm. En 2006, une enquête portant sur les mensurations des français a observé, sur un échantillon de 780 femmes, une taille moyenne de 162,5 cm avec un écart-type de 4 cm.

1. On suppose l'écart-type connu $\sigma = 4$cm. Donner un intervalle de confiance à $98\%$, c à d avec $\alpha = 2\%$, pour la taille moyenne des françaises en 2006.

   La loi utilisée est : $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$.

   La table des fractiles donne comme solution à $1-\alpha = 0.98 = P(-t \leqslant Z \leqslant t)$ le nombre $t = 2.3263$

   L'IDC est alors : $[a,b]$ avec : $a = \overline{x} - t\dfrac{\sigma}{\sqrt{n}} = 162.5 - 2.1701\dfrac{4}{\sqrt{780}} \simeq 162.17$

   et $b = \overline{x} + t\dfrac{\sigma}{\sqrt{n}} \simeq 162.83$

2. On suppose que l'échantillon ne comportait que 23 femmes et que $\sigma$ est inconnu. Quelle hypothèse faut-il alors avancer pour calculer cet intervalle de confiance ? Quelle loi sera utilisée dans ce cas ?

   On doit supposer l'échantillon gaussien.

   La loi utile est alors : $T = \frac{\overline{X}-\mu}{S^*}\sqrt{n} \sim T_{n-1}$ : loi de Student à $n-1$ d.d.l.

3. Recalculer cet IDC avec $n = 23,\ s^* = 4$ cm et $\overline{x} = 162.5$ cm.

   La table des Student donne comme solution à $1-\alpha = 0.98 = P(-t \leqslant T_{22} \leqslant t)$ le nombre $t = 2.5083$

   l'IDC est alors : $[a,b]$ avec :

   $a = \overline{x} - t\dfrac{s^*}{\sqrt{n}} = 162.5 - 2.5083\dfrac{4}{\sqrt{780}} \simeq 162.14$

   et $b = \overline{x} + t\dfrac{s^*}{\sqrt{n}} \simeq 162.86$

## 4. Exercice 4

Les moteurs des appareils électroménagers d'une marque $M$ ont une durée de vie moyenne de $3000$ heures avec un écart-type de $\sigma = 150$ heures.
A la suite d'une modification dans la fabrication des moteurs, le fabricant affirme que les nouveaux moteurs ont une durée de vie supérieure à celle des anciens.
Pour vérifier cette affirmation, nous avons testé un échantillon de $n = 50$ nouveaux moteurs et avons trouvé une durée de vie moyenne de $3075$ heures.

1. Préciser les deux hypothèses et la signification des deux risques de 1ère et de 2ème espèce.

Si on note $\mu = \mathbb{E}(X)$, les hypothèses à tester sont : $\begin{cases} (H_0): & \mu = \mu_0 = 3000 \\ (H_1): & \mu = \mu_1 > 3000 \end{cases}$

$\alpha =$ Proba. de décider que la durée moyenne est supérieure à $3000$ alors qu'elle est toujours égale à $3000$.

$\beta =$ Proba. de décider que la durée est toujours de $3000$ alors qu'elle est supérieure.

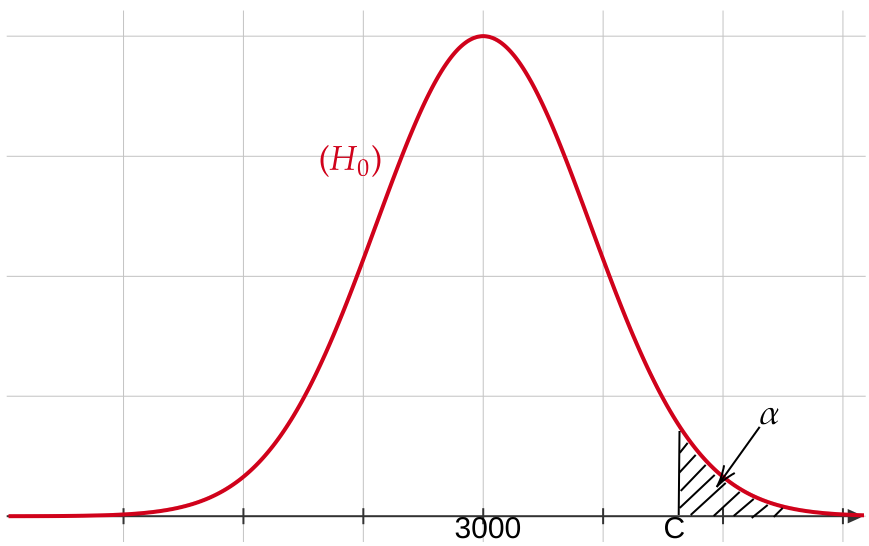
2. Quelle est la variable de décision ? Quelle loi suit-elle ? Justifier.

   La variable de décision est $\overline{X}$. Avec $n = 50 > 30$ et $\sigma$ connu, la loi utile est :

$$Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1).$$

3. Justifier la forme de la région critique et représenter la ainsi que les risques sur un graphique.

   L'hypothèses alternative se trouve du côté droit de l'hypothèse nulle, par conséquent la région critique a la forme : $W = \{\overline{X} > C\}$.




1. A partir d'un risque $\alpha = 1\%$ déterminer le seuil et préciser les règles de décision.

$$\alpha = 0.01 = P_{(H_0)}(\overline{X} > C) = P_{(H_0)}\left(\frac{\overline{X}-\mu}{\sigma}\sqrt{n} > \frac{C-\mu}{\sigma}\sqrt{n}\right)$$

$$0.01 = P\left(Z > \frac{C-\mu_0}{\sigma}\sqrt{n}\right) \iff F_Z\left(\frac{C-\mu_0}{\sigma}\sqrt{n}\right) = 0.99$$

   La table des fractiles de $\mathcal{N}(0,1)$ donne : $\frac{C-\mu_0}{\sigma}\sqrt{n} = 2.3263$.

$$C = \mu_0 + 2.3263\frac{\sigma}{\sqrt{n}} \simeq 3049$$

2. Quelle conclusion obtient-t-on avec notre échantillon ?

   $\overline{x} = 3075 > C \implies$ on valide $(H_1)$. Les nouveaux moteurs durent plus longtemps en moyenne.

3. Retrouver ce résultat en calculant la p-valeur.

$$\text{P-valeur} = P_{(H_0)}(\overline{X} > \overline{x}) = P_{(H_0)}\left(\frac{\overline{X}-\mu}{\sigma}\sqrt{n} > \frac{C-\mu}{\sigma}\sqrt{n}\right) = P\left(Z > \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right)$$
$$= 1 - F_Z\left(\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right) = 1 - F_Z\left(\frac{3075-3000}{150}\sqrt{50}\right)$$
$$= 1 - F_Z(3.54) = 1 - 0.9998 = 0.0002$$

P-valeur $< \alpha \quad \Longrightarrow$ on valide $(H_1)$.

## 5. Exercice 5

On souhaite savoir si, chez les étudiants d'une université, l'intérêt pour la politique est lié à la filière suivie ou non.

Pour cela on a demandé à $880$ étudiants d'indiquer leur filière et leur niveau d'intérêt pour la politique, à choisir entre "Peu/pas du tout", "Moyen" ou "Assez/beaucoup". Les résultats sont groupés dans le tableau suivant :

| Intérêt \ Filière | Sociologie | Droit | STAPS | Sciences | Total |
|---|---|---|---|---|---|
| Peu/pas du tout | 76 | 51 | 94 | 104 | 325 |
| Moyen | 71 | 53 | 66 | 59 | 249 |
| Assez/beaucoup | 74 | 123 | 51 | 58 | 306 |
| Total | 221 | 227 | 211 | 221 | 880 |

1. Quel est le nom du test à effectuer ? Quelles en sont les hypothèses ?

   Il s'agit d'un test d'indépendance du khi deux.

   Les hyoithèses sont : $\begin{cases} (H_0): & \text{les 2 variables sont indépendantes} \\ (H_1): & \text{les 2 variables sont liées} \end{cases}$

2. Pour effectuer ce test, nous avons besoin d'un tableau de contingence théorique comme le suivant :

| Intérêt \ Filière | Sociologie | Droit | STAPS | Sciences | Total |
|---|---|---|---|---|---|
| Peu/pas du tout | 81.6 | 83.8 | 77.9 | 81.6 | 325 |
| Moyen | **??** | 64.1 | 59.7 | 62.5 | 249 |
| Assez/beaucoup | 76.8 | 78.9 | 73.4 | **??** | 306 |

| | | | | | |
|---|---|---|---|---|---|
| Total | 221 | 227 | 211 | 221 | 880 |

Donner la signification de ce tableau et compléter les deux cases vides en expliquant la formule utilisée.

Il s'agit du tableau de contingence théorique en cas d'indépendance des 2 variables. Chaque case est obtenu en multipliant le total de la ligne avec le total de la colonne et en divisant par l'effectif total.

Moyen-Sociologie $= \dfrac{249 \times 221}{880} = 62.5$

Assez/beaucoup-Sciences $= \dfrac{306 \times 221}{880} = 76.8$

3. Définir la variable de décision $D$ en précisant sa méthode de calcul.
   La variable de décision est la distance du khi-2 calculée de la manière suivante :
$$D = \sum \frac{(n_{obs} - n_{théo})^2}{n_{théo}}$$
4. Quelle loi suit $D$ sous l'hypothèse $(H_0)$ ? Quel est le seuil correspondant pour $\alpha = 5\%$ ?
   Sous l'hypothèse $(H_0)$, $D$ suit la loi du khi-2 à $(4-1)(3-1) = 6$ d.d.l. $(D \sim \chi^2_6)$.
   Le seuil à $5\%$ donné par la table du khi-2 est $C = 12.592$
5. Le calcul de la distance $D$ donne $62.8$. Quelle conclusion en tirez-vous sur le lien entre filière et intérêt pour la politique ?
   $d = 62.8 > C \implies$ on valide $(H_1)$. Les 2 variables sont liées.
   L'intérêt pour la politique est bien lié à la filière suivie.