Corrigé exam Stats déc 21

Exo 1 :

Soit T un estimateur pour un paramètre θ.

1. Rappeler la définition du biais $b(T)$ et de l'erreur ou risque quadratique $R_\theta(T)$.

   Biais : $b(T) = E(T) - \theta$

   Risque quadratique : $R_\theta(T) = E\left((T-\theta)^2\right) = Var(T) + (b(T))^2$

2. Pourquoi entre deux estimateurs sans biais, doit-on choisir celui qui a la plus petite variance ?

   Dans ce cas, Risque quadratique=variance. On prend celui qui a le plus petit risque quadratique (erreur quadratique moyenne).

3. Rappeler ce qu'est la région critique W et donner un exemple où :

$$W = \left\{Y < C_1 \text{ ou } Y > C_2\right\}.$$

   Il faut penser à un test bilatéral, par exemple : $\begin{cases} (H_0) & \mu = \mu_0 \\ (H_1) & \mu \neq \mu_0 \end{cases}$

Exo 2 :

Soit X une variable aléatoire qui suit la loi uniforme sur un intervalle $[0, \theta]$ où θ est un paramètre positif inconnu.

On rappelle que la densité de X est données par : $f_X(x) = \begin{cases} \dfrac{1}{\theta} & \text{si } x \in [0, \theta] \\ 0 & \text{sinon} \end{cases}$

On dispose de $(X_1, \ldots, X_n)$ un n-échantillon de X. On note $\overline{X}$ la moyenne empirique de X.

1. Montrer que $E(X) = \dfrac{\theta}{2}$ et que $Var(X) = \dfrac{\theta^2}{12}$.

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)\, dx = \int_0^\theta \frac{x}{\theta} dx = \frac{\theta}{2}.$$

$$E\left(X^2\right) = \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx = \int_0^\theta \frac{x^2}{\theta} dx = \frac{\theta^2}{3}.$$

$$Var(X) = E\left(X^2\right) - (E(X))^2 = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}$$

On sait que $T_1 = 2\overline{X}$ est un estimateur sans biais et convergent de θ.

Soit $T_2 = \max(X_1, \ldots, X_n)$ un deuxième estimateur de θ.

On admet que $E(T_2) = \dfrac{n}{n+1}\theta$ et que $Var(T_2) = \dfrac{n}{(n+2)(n+1)^2}\theta^2$.

2. Calculer le biais et le risque quadratique de T_2.

$$b(T_2) = E(T_2) - \theta = \frac{n}{n+1}\theta - \theta = \frac{-\theta}{n+1}$$

$$R_\theta(T_2) = Var(T_2) + b(T_2)^2 = \frac{n}{(n+2)(n+1)^2}\theta^2 + \frac{\theta^2}{(n+1)^2} = \frac{(2n+2)\theta^2}{(n+2)(n+1)^2} = \frac{2\theta^2}{(n+2)(n+1)}$$

3. Soit $T_3 = \frac{n+1}{n}T_2$. Déterminer son biais et son risque quadratique.

$$E(T_3) = \frac{n+1}{n}E(T_2) = \theta \implies b(T_3) = 0.$$

$$R_\theta(T_3) = Var(T_3) = \left(\frac{n+1}{n}\right)^2 Var(T_2) = \frac{\theta^2}{n(n+2)}$$

4. Entre ces trois estimateurs et pour n assez grand, lequel donnera la meilleure estimation de θ et sans biais ?

T_1 et T_3 sont les seuls sans biais, et $R_\theta(T_1) = Var(T_1) = Var(2\overline{X}) = 4\dfrac{Var(X)}{n} = \dfrac{4\theta^2}{12n}$.

Alors que $R_\theta(T_3) = \dfrac{\theta^2}{n(n+2)}$.

Pour n assez grand, ce sera T_3 qui aura le plus petit risque quadratique.

C'est donc T_3 qu'il faut choisir, celui qui converge le plus vite.

Exo 3 :

On suppose que le poids, à la naissance, d'un bébé est une variable aléatoire de moyenne μ et de variance σ^2.

1. Dans un hôpital parisien, on a relevé les poids de $n_1 = 105$ bébés nés d'une mère primipare(qui accouche pour la première fois) et on a trouvé une moyenne empirique $\overline{x} = 3.41$ kg et un écart-type $s^* = 0.505$ kg.

   Donner un intervalle de confiance à 95% pour μ.

   $n_1 > 30 \overset{TCL}{\implies} Z = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$. Mais σ est inconnu remplacé par S^*.

   On doit passer la la loi de Student à 104 d.d.l. qui est confondue avec la loi normale standard. Finalement on va partir de la valeur $t = 1.96$ correspondant à $(-t < Z < t) = 0.95$.

   Ce qui nous donne l'IDC :

   $$[a, b] = \left[\overline{x} - t\frac{s^*}{\sqrt{n_1}}\ ;\ \overline{x} + t\frac{s^*}{\sqrt{n_1}}\right] = [3.31\ ;\ 3.51]$$

2. Même question pour un échantillon de $n_2 = 95$ mères multipares(qui ont déjà accouché) qui a donné $\overline{x} = 3.197$ kg et $s^* = 0.458$ kg.

   Le même raisonnement donne l'IDC suivant :

$$[a,b] = \left[\overline{x} - t\frac{s^*}{\sqrt{n_2}}\ ;\ \overline{x} + t\frac{s^*}{\sqrt{n_2}}\right] = [3.10\ ;\ 3.29]$$

Exo 4 :

On admet que la consommation d'oxygène d'une personne, exprimée en ml/kg/min, est une variable gaussienne vérifiant :

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

La valeur normale de la moyenne est $\mu = \mu_0 = 25$.

On veut tester si des patients atteints de la maladie de Parkinson voient leur consommation baisser et tomber à $\mu = \mu_1 = 20$ et s'il faut donc les oxygéner. On utilise pour cela un échantillon de taille $n = 15$.

1. Enoncer les 2 hypothèses et expliciter les risques de 1ère et de 2ème espèce.

$$\begin{cases} (H_0) & \mu = \mu_0 = 25 \\ (H_1) & \mu = \mu_1 = 20 \end{cases}$$

Risque de 1ère espèce : $\alpha =$ Proba. de valider (H_1) à tort $=$ Proba. de décider d'oxygéner les patients alors qu'ils n'en ont pas besoin.

Risque de 2ème espèce : $\beta =$ Proba. de valider (H_0) à tort $=$ Proba. de ne pas oxygéner les patients alors qu'ils en ont besoin.

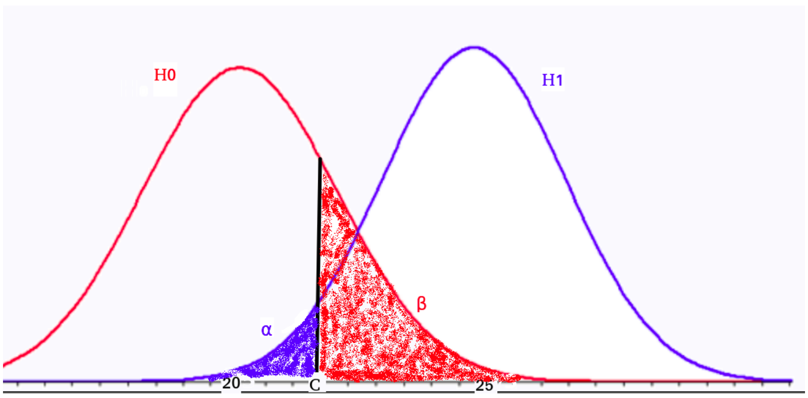
2. Quelle sera la variable de décision et quelle sera la loi utilisée si on sait que $\sigma^2 = 36$.

Variable de décision : moyenne empirique $\overline{X}$

X gaussienne $\Longrightarrow\ Z = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$

3. Faire une représentation graphique montrant la région critique et les risques de 1ère et 2ème espèce.

Région critique : $W = \{\overline{X} < C\}$



4. Déterminer le seuil de décision pour un risque de $\alpha = 5\%$.

$$\alpha = 0.05 = P_{(H_0)}(\overline{X} < C) = P_{(H_0)}\left(\frac{\overline{X} - \mu}{\sigma}\sqrt{n} < \frac{C - \mu}{\sigma}\sqrt{n}\right)$$

$$\alpha = 0.05 = P\left(Z < \frac{C-\mu_0}{\sigma}\sqrt{n}\right)$$

La table des fractiles de $\mathcal{N}(0,1)$ donne :

$$\frac{C-\mu_0}{\sigma}\sqrt{n} = -1.6449 \implies C = \mu_0 - 1.6449\frac{\sigma}{\sqrt{n}} = 25 - 1.6449\frac{6}{\sqrt{15}}$$

$$C = 22.45$$

5. Un échantillon de $n = 15$ personnes malades a donné la valeur $\overline{x} = 23.1$
   Doit-on prendre la décision de les oxygéner ?
   $\overline{x} = 23.1 > C \implies$ On valide (H_0). Il n'y a pas de raison d'oxygéner ces patients.
6. **Question bonus :**
   Retrouver la décision précédente en calculant la p-valeur.

   P-valeur $= P_{(H_0)}(\overline{X} < \overline{x}\ ) = P\left(\frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \leqslant \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right) = P\left(Z \leqslant \frac{23.1-25}{6}\sqrt{15}\right)$

   P-valeur $= F_Z(-1.22) = 1 - F_Z(1.22) = 1 - 0.88877 = 0.11123$

   P-valeur $> \alpha \implies$ On valide (H_0).

Exo 5 :

La distribution X du nombre d'enfants par famille en France peut être résumée par la loi discrète suivante :

Nombre d'enfants k	0	1	2	3	4	$\geqslant 5$
$P(X=k)$	0.15	0.2	0.3	0.2	0.1	0.05

Nous avons relevé le nombre d'enfants dans $n = 900$ familles belges et souhaitons savoir si la répartition est équivalente.

- ➢ Quel est le nom du test à effectuer, et quelles sont les hypothèses ?
  Test d'adéquation d'un échantillon à une variable discrète.

  $$\begin{cases} (H_0) & \text{L'échantillon suit la loi indiquée} \\ (H_1) & \text{L'échantillon ne suit pas la loi indiquée} \end{cases}$$

  $$\iff \begin{cases} (H_0) & \text{La distribution du nbre d'enfants en Belgique est la même qu'en France} \\ (H_1) & \text{La distribution du nbre d'enfants en Belgique est différente de celle en France} \end{cases}$$

- ➢ Compléter le tableau suivant permettant de répondre à la question précédente.
  On multiplie les probabilités des modalités apr l'effectif total $n = 900$.

  Ensuite on calcule la distance à l'aide de la formule : $\dfrac{(n_{obs}-n_{th})^2}{n_{th}}$

Nombre d'enfants k	0	1	2	3	4	$\geqslant 5$
Effectif observé n_{obs}	151	197	240	161	110	41
Effectif théorique n_{th}	135	180	270	180	90	45
Distance	1.90	1.61	3.33	2.01	4.44	0.36

- Donner la formule permettant de calculer la distance totale ainsi que la loi suivie sous (H_0).

  $$D_n = \sum_{i=1}^{6} \frac{(n_{obs,i} - n_{th,i})^2}{n_{th,i}}.$$

  Sous (H_0), D_n suit la loi du khi2 à $6 - 1 = 5$ d.d.l. (nombre de modalités retenues moins 1).

- Donner la conclusion de ce test avec un risque de $\alpha = 5\%$ puis avec $\alpha = 1\%$.

  La table du khi2 à 5 d.d.l. donne : $C_1 = 11.70$ pour $\alpha = 5\%$ et $C_2 = 15.09$ pour $\alpha = 1\%$

  La distance trouvée est de $D_n = 13.64$.

  - Au risque $\alpha = 5\%$, $D_n > C \implies$ On valide (H_1): la distribution des nombres d'enfants par famille n'est pas la même en France qu'en Belgique.
  - Au risque $\alpha = 1\%$, $D_n < C \implies$ On valide (H_0): la distribution des nombres d'enfants par famille est la même en France qu'en Belgique.