Corrigé du rattrapage Stats ING2-GI
février 2023

1. Exercice 1 :

Justifier clairement vos réponses aux questions suivantes :

1. Soit $(X_1, X_2,..., X_n)$ un échantillon d'une loi X. Et soit $\overline{X}$ la moyenne empirique, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculer $Var(\overline{X})$ en fonction de $Var(X)$ et n.
$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n Var(X) = \frac{Var(X)}{n}$$
2. Comment peut-on choisir entre deux estimateurs sans biais d'un même paramètre ?
On choisit celui qui a le plus petit risque quadratique, donc la plus petite variance.
3. Définir la p-valeur d'un test et expliquer comment elle est utilisée pour valider (H_0) ou (H_1).
P-valeur $=$ probabailité, sous (H_0), d'obtenir des valeurs encores plus extrêmes que celle de l'échantillon donné.

- P-valeur $< \alpha \implies$ on valide (H_1).
- P-valeur $> \alpha \implies$ on valide (H_0).

4. Quel type de test doit-on effectuer pour vérifier qu'un échantillon donné suit une loi discrète ? Quelle loi est suivie par la variable de décision ?
Il s'agit d'un test d'adéquation. La variable de décision suit la loi du khi-2.

2. Exercice 2 :

On note X une variable aléatoire dont la densité est donnée par :

$$f_X(x) = \begin{cases} \frac{3\theta^3}{x^4}, & \text{si } x \geqslant \theta \\ 0, & \text{sinon} \end{cases}, \text{ avec } \theta > 0.$$

1. Justifier que f_X est bien une densité de probabilité et montrer que $\mathbb{E}(X) = \frac{3\theta}{2}$.
f_X est bien positive, continue partout sauf en θ et vérifie :
$$\int_{-\infty}^{+\infty} f_X(x)\,dx = \int_{\theta}^{+\infty} \frac{3\theta^3}{x^4}\,dx = \left[-\frac{\theta^3}{x^3}\right]_{\theta}^{+\infty} = 1.$$

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

2. Montrer que $Var(X) = \frac{3\theta^2}{4}$.

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

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

3. Justifier que $T_1 = \overline{X}$ est un estimateur biaisé de θ, et déterminer le réel λ tel que $T_2 = \lambda \overline{X}$ soit un estimateur sans biais.

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X) = \frac{3\theta}{2} \neq \theta \implies T_1 = \overline{X} \text{ est un estimateur biaisé de } \theta.$$

Pour obtenir un estimateur non biaisé de la forme $\lambda \overline{X}$, il suffit de prendre $\lambda = \frac{2}{3}$.

$$T_2 = \frac{2}{3}\overline{X}$$

4. Justifier la convergence en probabilités, puis en moyenne quadratique de T_2.

D'après la loi faible des grands nombres, $\overline{X}$ converge en probabilité vers $\mathbb{E}(X) = \frac{3\theta}{2}$.

En multipliant par $\frac{2}{3}$, on obtient bien : $T_2 \xrightarrow{\mathcal{P}} \theta$.

Par ailleurs, $R_\theta(T_2) = Var(T_2) = Var\left(\frac{2}{3}\overline{X}\right) = \frac{4}{9} Var(\overline{X}) = \frac{\theta^2}{3n}$.

$$\lim_{n \to +\infty} R_\theta(T_2) = 0 \implies T_2 \text{ converge en moyenne quadratique vers } \theta.$$

3. Exercice 3 :

On s'intéresse à la proportion de femmes parmi les patients lombalgiques.
Dans une certaine clinique, sur $n = 2620$ patients atteints de lombalgie, on a relevé 1360 femmes.

1. En déduire un intervalle de confiance à 97% (au risque $\alpha = 3\%$) pour la proportion p des femmes parmi les personnes atteintes de lombalgie.

Notons F_n la fréquence empirique. On sait que :

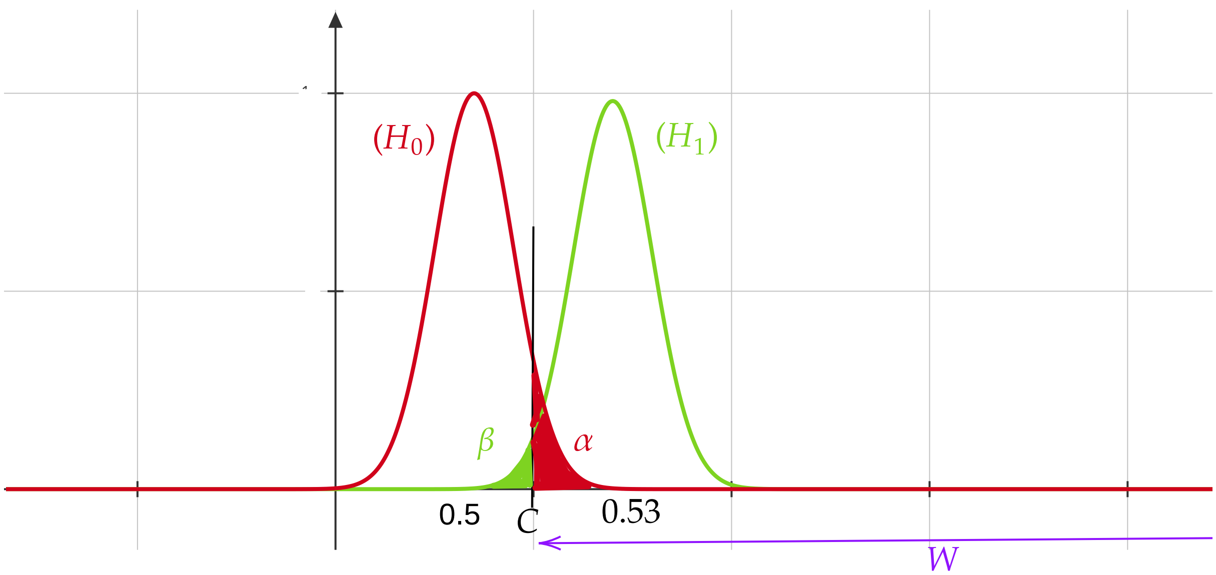
$$n = 2620 \gg 30 \implies Z = \frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n} \sim \mathcal{N}(0,1).$$

Pour construire un IDC à 97%, on a besoin du décile 0.985 de $\mathcal{N}(0,1) \implies t = 2.17$

Avec $f_n = \dfrac{1360}{2620} = 0.519$

L'IDC recherché est : $\left[f_n - t\sqrt{\dfrac{f_n(1-f_n)}{n}} \;;\; f_n - t\sqrt{\dfrac{f_n(1-f_n)}{n}} \right] = [0.498 \;;\; 0.54]$

2. Un cadre administratif affirme que la proportion de femmes est de 53% ($p_1 = 0,53$), contre l'avis de tous ses collègues qui pensent qu'il y a autant de femmes que d'hommes ($p_0 = 0,5$). On veut utiliser l'échantillon précédent pour tester si ce cadre a raison.
   Quels sont les 2 hypothèses en jeu ? Quelle est la variable de décision ?
   $$\begin{cases} (H_0) : \quad p = p_0 = 0.5 \\ (H_1) : \quad p = p_1 = 0.53 \end{cases}.$$ Variable de décision $= F_n$.
3. Représenter graphiquement la loi de cette variable selon les 2 hypothèses, ainsi que la région critique et les 2 risques.



4. Calculer le seuil à 5% et conclure à l'aide des valeurs de l'échantillon.
   $$\alpha = 0.05 = P_{(H_0)}(F_n > C) = P_{(H_0)}\left(\frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n} > \frac{C-p}{\sqrt{p(1-p)}}\sqrt{n} \right)$$
   $$\alpha = 0.05 = P\left(Z > \frac{C-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \right) \implies \frac{C-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = 1.645$$
   $$C = p_0 + 1.645\sqrt{\frac{p_0(1-p_0)}{n}} = 0.5 + 1.645\sqrt{\frac{0.5(0.5)}{2620}} = 0.516\,.$$
   Ici, $f_n = 0.519 > C \implies$ on valide (H_1). Le cadre administratif a raison.

5. Calculer le risque de 2ème espèce et la puissance du test.

$$\beta = P_{(H_1)}(F_n < C) = P_{(H_1)}\left(\frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n} < \frac{C-p}{\sqrt{p(1-p)}}\sqrt{n}\right)$$

$$\beta = P\left(Z < \frac{C-p_1}{\sqrt{p_1(1-p_1)}}\sqrt{n}\right) = F_Z\left(\frac{0.516-0.53}{\sqrt{0.53(1-0.53)}}\sqrt{2620}\right)$$

$$\beta = F_Z(-1.44) = 1 - F_Z(1.44) = 1 - 0.925 = 0.075\,.$$

La puissance du test est : $1-\beta = 0.925$

6. Retrouver la conclusion à l'aide de la p-valeur.

$$\text{P-valeur} = P_{(H_0)}(F_n > f_n) = P_{(H_0)}\left(\frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n} > \frac{f_n-p}{\sqrt{p(1-p)}}\sqrt{n}\right)$$

$$= P\left(Z > \frac{f_n-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\right) = P\left(Z > \frac{0.519-0.5}{\sqrt{0.5(0.5)}}\sqrt{2620}\right)$$

$$\text{P-valeur} = 1 - F_Z(1.94) = 1 - 0.974 = 0.026$$

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

4. Exercie 4 :

Un traitement est administré à trois doses différentes D1, D2, D3, à un groupe de sujets atteints d'une même maladie. On compte le nombre de guérisons pour chaque dose. Les résultats sont les suivants :

	Sujets Guéris	Sujets non guéris	Total
D1	30	80	
D2	42	35	
D3	58	31	
Total			

On souhaite tester, au risque de 1%, si la guérison dépend de la dose administrée.

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

   Test d'indépendance du khi-2 entre 2 variables qualitatives.

$$\begin{cases} (H_0): & \text{Les 2 vars sont indépdtes} \\ (H_1): & \text{Les 2 vars sont liées} \end{cases}$$

2. Dresser un tableau de contingence théorique en cas d'indépendance et calculer les effectifs théoriques.

   Tableau de contingence théorique (cas d'indépendance) :

	Sujets Guéris	Sujets non guéris	Tot

D1	51,8	58,2	110
D2	36,3	40,7	77
D3	41,9	47,1	89
Tot	130	146	276

Les effectifs théorique sont obtenus par la formule : $\dfrac{\text{Total ligne x Total colonne}}{\text{Total général}}$

3. Comment est calculée la variable de décision de ce test. On demande uniquement les formules pas les calculs.

   La variable de décision D_n est la distance du khi-2 calculée entre la distribution observée et la théorique par la formules suivante :

$$D_n = \sum \frac{(n_{theo} - n_{obs})^2}{n_{theo}} = \frac{(51.8 - 30)^2}{51.8} + \frac{(58.2 - 80)^2}{58.2} + ...$$

4. Quelle loi suit cette variable, sous (H_0), et quel est le seuil à 1% pour notre cas ?

   Sous l'hypothèse d'indépendance, $D_n \sim \chi^2_2$: loi du khi-2 à $(3-1)(2-1) = 2$ d.d.l.

   Le seuil à 1%, donné par la table du khi-2, est de $C = 9.21$

5. Le calcul exact de cette variable de décision sur notre échantillon donne : $D = 30.7$
   Quelle est votre conclusion quant au lien entre guérison et dose administrée ?

   $D = 30.7 > C \implies$ on valide (H_1). Les variables sont liées.

   La guérison dépend bien de la dose administrée.