Sorbonne Université Année 2020/2021

L3 Deuxième semestre

DM de statistique

Exercice 1 : Soit Y une variable aléatoire ayant pour densité f_Y définie pour tout $x \in \mathbb{R}$ par

$$f_Y(x) = \frac{1}{2(1-x)^{1/2}} \mathbf{1}_{[0,1]}(x).$$

avec $f_Y(1) = +\infty$. Soit $\theta > 0$, on considère maintenant la variable aléatoire $X = \theta Y$ et on admet que sa densité f_θ est définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{1}{2\theta(1-x/\theta)^{1/2}} \mathbf{1}_{[x,+\infty[}(\theta).$$

1. On admet que $\mathbb{E}[Y] = \frac{2}{3}$ et $\mathbb{V}[Y] = \frac{4}{45}$. Calculer $\mathbb{E}[X]$ et $\mathbb{V}[X]$.

$$\mathbb{E}[X] = \frac{2}{3}\theta \quad \text{et} \quad \mathbb{V}[X] = \frac{4}{45}\theta^2.$$

2. Par la méthode des moments, proposer un estimateur de θ.
   On a $\hat{\theta}_n = \frac{3}{2}\overline{X}_n$.

3. Est-il consistant ? Asymptotiquement normal ?
   Par la loi faible des grands nombres, $\overline{X}_n$ converge en probabilité vers $(2/3)\theta$ et on obtient donc la convergence en probabilité de $\hat{\theta}_n$. De plus, le TLC donne

$$\sqrt{n}\left(\overline{X}_n - \frac{2}{3}\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{4}{45}\theta^2\right)$$

   et on a donc

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{3}{2}\sqrt{n}\left(\overline{X}_n - \frac{2}{3}\theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{5}\theta^2\right)$$

4. Soit $\alpha \in (0,1)$, donner un intervalle de confiance asymptotique de niveau $1-\alpha$ pour θ.
   Grâce à Slutsky, on a

$$\sqrt{5n}\frac{\left(\hat{\theta}_n - \theta\right)}{\hat{\theta}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

   Soit $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite, on obtient

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta}_n \pm q_{1-\alpha/2}\frac{\hat{\theta}_n}{\sqrt{5n}}\right]$$

5. Montrer que l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{MV} = X_{(n)}$.

   On a

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} \frac{1}{2\theta(1 - X_i/\theta)^{1/2}} \mathbf{1}_{[X_i, +\infty[}(\theta) = \prod_{i=1}^{n} \frac{1}{2\sqrt{\theta}(\theta - X_i)^{1/2}} \mathbf{1}_{[X_{(n)}, +\infty[}(\theta)$$

   qui est clairement strictement décroissante sur $[X_{(n)}, +\infty[$ et on obtient donc que le maximum est unique et est atteint en $X_{(n)}$

6. Montrer que pour tout $x \in [0,1]$, $F_Y(x) = 1 - (1-x)^{1/2}$ où F_Y est la fonction de répartition de Y.

   On a

$$F_Y(x) = \int_0^x \frac{1}{2(1-t)^{1/2}} dt = \left[-(1-t)^{1/2} \right]_0^x = 1 - (1-x)^{1/2}.$$

7. En déduire les fonctions de répartition de X et X_n.

   Pour tout $x \in [0, \theta]$, on a

$$F_X(x) = \mathbb{P}\left[X \leq x\right] = \mathbb{P}\left[Y \leq \frac{x}{\theta}\right] = 1 - \left(1 - \frac{x}{\theta}\right)^{1/2}$$

   De plus, par indépendance des X_i,

$$F_{X_{(n)}}(x) = \left(1 - \left(1 - \frac{x}{\theta}\right)^{1/2}\right)^n$$

8. En déduire que $n\left(\theta - X_{(n)}\right)^{1/2}$ converge en loi vers $E \sim \mathcal{E}\left(\theta^{-1/2}\right)$

   On a pour tout $x \leq n\sqrt{\theta}$,

$$\begin{aligned}
\mathbb{P}\left[n\left(\theta - X_{(n)}\right)^{1/2} \leq x\right] &= \mathbb{P}\left[X_{(n)} \geq \theta - \frac{x^2}{n^2}\right] \\
&= 1 - \left(1 - \left(1 - \frac{\theta - x^2 n^{-2}}{\theta}\right)^{1/2}\right)^n \\
&= 1 - \left(1 - \frac{x}{n\sqrt{\theta}}\right)^n \\
&= 1 - \exp\left(n \ln\left(1 - \frac{x}{n\sqrt{\theta}}\right)\right) \\
&\sim 1 - \exp\left(-\frac{x}{\sqrt{\theta}}\right)
\end{aligned}$$

   ce qui est la fonction de répartition de la loi exponentielle de paramètre $1/\sqrt{\theta}$.

9. Soit $\alpha \in (0,1)$, trouver $c_\alpha > 1$ tel que

$$\mathbb{P}\left[\theta \leq X_{(n)} c_\alpha\right] = 1 - \alpha.$$

On a

$$\begin{aligned}1-\alpha &= \mathbb{P}\left[X_{(n)} \geq \frac{\theta}{c_\alpha}\right] \\ &= 1-\left(1-\left(1-\frac{1}{c_\alpha}\right)^{1/2}\right)^n\end{aligned}$$

On obtient donc

$$\begin{aligned}1-\alpha = 1-\left(1-\left(1-\frac{1}{c_\alpha}\right)^{1/2}\right)^n &\Leftrightarrow 1-\left(1-\frac{1}{c_\alpha}\right)^{1/2} = \alpha^{1/n} \\ &\Leftrightarrow 1-\frac{1}{c_\alpha} = 1-2\alpha^{1/n}+\alpha^{2/n} \\ &\Leftrightarrow c_\alpha = \frac{1}{2\alpha^{1/n}-\alpha^{2/n}} = \frac{1}{\alpha^{1/n}\left(2-\alpha^{1/n}\right)}.\end{aligned}$$

10. En déduire un nouvel intervalle de confiance.
    On obtient donc
    $$IC_{1-\alpha}(\theta) = \left[X_{(n)}, X_{(n)}\frac{1}{\alpha^{1/n}\left(2-\alpha^{1/n}\right)}\right].$$
11. Quel estimateur choisiriez vous ?
    L'estimateur du maximum de vraisemblance converge beaucoup plus rapidement que l'estimateur obtenu via la méthode des moments.

Exercice 2 : On relève chez 9 patients une glycémie moyenne de 1.12 g/l et une variance de 0.01. On admettra que la glycémie des patients suit une loi normale et on note m sa moyenne. Ces patients sont issus d'une population dont la glycémie moyenne vaut $m_0 = 1$ g/l. Au risque de 5%, tester si cet échantillon est représentatif du reste de la population. On pourra s'aider des résultats suivants :

$$\begin{array}{ccc}\mathbb{P}\left[T_8 \leq 2.306\right] = 0.975 & \mathbb{P}\left[T_8 \leq 1.860\right] = 0.95 & \mathbb{P}\left[T_9 \leq 2.262\right] = 0.975 \\ \mathbb{P}\left[T_9 \leq 1.833\right] = 0.95 & \mathbb{P}\left[N \leq 1.960\right] = 0.975 & \mathbb{P}\left[N \leq 1.645\right] = 0.95\end{array}$$

où T_k suit une loi de Student à k degrés de liberté et N suit une loi normale centrée réduite.

Dans ce qui suit, on note $X_1, \ldots, X_9$ les glycémies des patients, $\overline{X}_9$ l'estimateur de la moyenne, et S_9^2 l'estimateur sans biais de la variance.

On teste au risque de 5% H0 : "$m = m_0$" contre H1 : "$m \neq m_0$".
On a la statistique de test

$$Z = \sqrt{9}\frac{\overline{X}_9 - m_0}{S_9} \sim T_8 \quad \text{sous H0}$$

On a la zone de rejet

$$ZR = \left\{\sqrt{9}\frac{\left|\overline{X}_9 - m_0\right|}{S_9} > t_{8,0.975}\right\}$$

où $t_{8,0.975} = 2.306$ est le quantile d'ordre 0.975 de la loi de Student à 8 degrés de liberté. Ici , $z_{obs} = \frac{3\times 0.12}{0.1} = 3.6 > 2.306$ et on rejette donc H0.

Exercice 3 : En vue d'estimer les différences de productivité qui peuvent exister entre plusieurs types de forêts de hêtre (Fagus sylvatica L.) de l'Ardenne belge, on a mesuré, en différents endroits, la hauteur des arbres les plus gros [Dagnélie, 1956-1957]. La hauteur de ces arbres, qui est étroitement liée à la production en volume, peut en effet être considérée comme une mesure simple, mais fiable, du niveau de productivité des forêts. Nous n'envisageons ici que trois types de hêtraies, au sein desquels on a observé les hauteurs des arbres, respectivement en 13 endroits, en 14 endroits et en 10 endroits différents, choisis au hasard et indépendamment les uns des autres. En chaque endroit, les cinq arbres les plus gros situés au sein d'une parcelle circulaire d'environ 15 m de rayon (soit environ 7 ares) ont été mesurés, chacun deux fois, et pour chaque lieu, la moyenne des 10 observations a été calculée. Pour éviter la présentation de données trop nombreuses, nous ne considérons que ces moyennes.
On souhaite comparer la productivité des forêts 1 et 2 en utilisant la hauteur des arbres.
On trouvera dans le tableau ci-dessous les mesures obtenues.

Forêt 1	Forêt 2
$(x_{j,1})_{1\leq j\leq 13}$	$(x_{j,2})_{1\leq j\leq 14}$
23.4 24.4 24.6 24.9	22.5 22.9 23.7 24.0
25.0 26.2 26.3 26.8	24.4 24.5 25.3 26.0
26.8 26.9 27.0 27.6	26.2 26.4 26.7 26.9
27.7	27.4 28.5
$n_1 = 13$	$n_2 = 14$

Tab. 1: Mesures des hauteur des arbres des forêts 1 et 2.

On notera $x_{1,1}, x_{2,1}, \ldots, x_{13,1}$ les mesures obtenues pour la forêt 1 et $x_{1,2}, x_{2,2}, \ldots, x_{14,2}$ celles de la forêt 2. On supposera que les données $x_{1,1}, x_{2,1}, \ldots, x_{13,1}$ sont des réalisations indépendantes d'une variable aléatoire d'espérance μ_1 et de variance σ_1^2 et que $x_{1,2}, x_{2,2}, \ldots, x_{14,2}$ sont des réalisations indépendantes d'une variable aléatoire d'espérance μ_2 et de variance σ_2^2.

1. On trouvera dans la figure ci-dessous les boîtes à moustaches des deux séries de données ainsi que les histogrammes en fréquences des 2 séries.

   Quelles informations peut-on tirer de ces deux graphiques ?

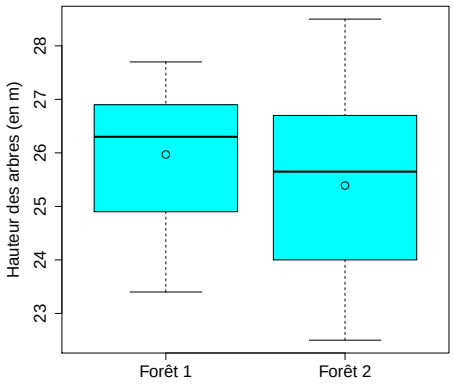

Fig. 1: Boîtes à moustaches.

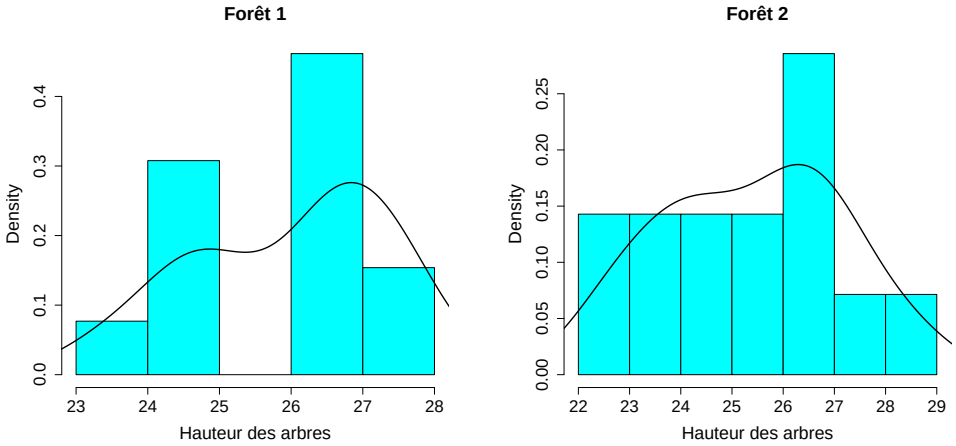

Fig. 2: Histogrammes en fréquences et densités.

Figure 1, on peut voir que les moyennes des deux échantillons sont assez semblable ainsi que les variances, même si la variance de l'échantillon 2 semble légèrement plus importante que celle de l'échantillon 1. Dans la Figure 2, on voit qu'il n'est pas "délirant" de ne pas rejeter le caractère gaussien des échantillons.

2. On effectue le test de Shapiro-Wilk sur chacune des séries de mesures et on obtient les p-values respectives 0,27 et 0,91. Que pouvez-vous en conclure ?

   Le test de Shapiro Wilk nous permet de ne pas rejeter le caractère gaussien des échantillons, i.e jusqu'à un risque de 27%, on ne rejette pas le caractère gaussien des échantillons.

3. La mise en oeuvre, avec le logiciel R et la fonction `var.test`, du test de comparaison de variances de Fisher conduit au résultat suivant :

   ```
   F test to compare two variances
   ```

```
data:  Hauteur by Foret
F = 0.58696, num df = 12, denom df = 13, p-value = 0.3647
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1861495 1.9013248
sample estimates:
ratio of variances
         0.5869621
```

Tester si σ_1^2 est égale à σ_2^2 au risque de 5% et au risque de 1%.

On test au risque de 5% H0 :"$\sigma_1^2 = \sigma_2^2$ contre H1 :"$\sigma_1^2 \neq \sigma_2^2$.

On a la stastistique de test

$$Z = \frac{S_{1,13}^2}{S_{2,14}^2} \sim \mathcal{F}_{12,13}$$

On a la zone de rejet

$$ZR = \{Z \leq f_{12,13,0.025}\} \bigcup \{Z \geq f_{12,13,0.975}\}.$$

Ici la p-value vaut $0.3647 > 0.05$, et on ne rejette donc pas $H0$.

De plus, comme la p-value est plus grande que 0.01, on ne rejette pas H0 au risque de 1% non plus.

4. On souhaite comparer la productivité des deux forêts. Quel test doit-on mettre en oeuvre pour cela. Rappeler le cadre théorique nécessaire à la mise en oeuvre de ce test.

   Compte-tenu des résultats des questions précédentes, peut-on dire que les conditions d'application du test sont satisfaites ? On justifiera la réponse.

   Afin de comparer les productions des deux forêts, on va mettre en place un test de Student. Pour cela, il faut que les données soient des réalisations de variables aléatoires indépendantes. On peut ici partir du principe que la hauteur d'un arbre n'affecte pas celle d'un autre arbre.

   Il faut que pour chaque échantillon, les données soient des réalisations de variables aléatoires suivant une loi normale, ce que l'on a vérifié grâce au test de Shapiro Wilk. Enfin, il faut que les variances de chaque échantillon soient égales, ce que l'on a vérifié à l'aide du test de Fisher.

5. Tester aux risques 5% et 1% l'hypothèse selon laquelle la productivité des deux forêts est la même. On précisera les hypothèses testées, la statistique de test et sa loi sous H_0, les 2 zones de rejet et on conclura. Pour vous aider, on fournit une partie du résultat donné par la fonction `t.test` du logiciel R :

   ```
   Two Sample t-test
   ```

```
data:  c1 and c2
t = 0.96377, df = 24.168, p-value = 0.3447
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6656157  1.8326487
sample estimates:
mean of x mean of y
 25.96923  25.38571
```

On test au risque de 5% H0 : "$\mu_1 = \mu_2$" contre H1 : "$\mu_1 \neq \mu_2$". On a la statistique de test

$$Z = \frac{\sqrt{13 \times 14}}{\sqrt{27}} \frac{\overline{X}_{1,13} - \overline{X}_{1,14}}{S} \sim T_{25} \quad \text{sous H0}$$

On a la zone de rejet

$$ZR = \{|Z| \geq t_{25,0.975}\}$$

où $t_{25,0.975}$ est le quantile d'ordre 0.975 de la loi de Student à 25 degrés de liberté. Ici, la p-value vaut $0.3447 > 0.05$ et on ne rejette donc pas H0 au risque de 5% et également au risque de 1%.