M1 Océanographie $1^{ème}$ année **Mise à niveau - Probabilités & Statistiques**

Examen de première session - Décembre 2016 - durée 2h

Sans document - Avec calculatrice

Exercice n°1.
Soit X une variable aléatoire réelle qui suit une loi exponentielle. Sa fonction de répartition est donnée par :

$$F_X(x) = \begin{cases} 1 - \exp\left(-\frac{1}{\theta}x\right) & \text{si} \quad x \geq 0 \\ 0 & \text{sinon,} \end{cases}$$

où $\theta > 0$ est un paramètre inconnu. Pour la suite, on suppose que l'on dispose d'un échantillon *i.i.d.* de taille n noté $\{X_1, \cdots, X_n\}$ de même loi mère que X. L'objectif de l'exercice est de construire un estimateur de θ.

1. Donner la densité de X et la tracer pour $\theta = 1$. Montrer que l'espérance de X est telle que $E(X) = \theta$.
2. Soit
$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
un estimateur de θ. En admettant que $V(X) = \theta^2$, calculer $E(T_n)$ et $V(T_n)$.
3. Etudier le biais et la variance de cet estimateur. Qu'en déduisez-vous?

Exercice n°2.
Vous êtes sur la paillasse et vous utilisez un instrument de mesure qui vous permet de doser des nitrates dans un échantillon d'eau de mer. Soit X la concentration en nitrates (mg/l), vous faites $n = 6$ mesures dans des conditions similaires et vous obtenez

102.94 103.75 103.89 96.17 102.94 112.08.

1. Vous vous penchez à l'arrière de votre appareil et vous lisez que la variance des mesures est de $\sigma_0^2 = 25$. Montrez qu'il n'y a pas de raison de remettre en doute la calibration de la machine. Que représente la figure 1.?
2. Donner un intervalle de confiance à 95% de la variance de l'appareil.
3. A l'endroit des prélèvements, la législation stipule que l'eau est polluée si la concentration moyenne en nitrates est supérieure à $\mu = \mu_0 = 100\,mg/l$. Construire un test pour votre échantillon :
   (a) lorsque $\sigma^2 = \sigma_0^2$
   (b) lorsque vous estimez σ^2 avec vos mesures.
4. Expliquer les différences de vos résultats en appuyant votre argumentaire sur la figure 2.
On donne $\chi^2_{5;0.025} = 0.83$, $\chi^2_{5;0.05} = 1.145$, $\chi^2_{5;0.95} = 11.07$, $\chi^2_{5;0.975} = 12.83$, $t_{5;0.05} = -2.01$, $t_{5;0.01} = -3.36$, $z_{0.05} = -1.64$, $z_{0.01} = -2.32$.

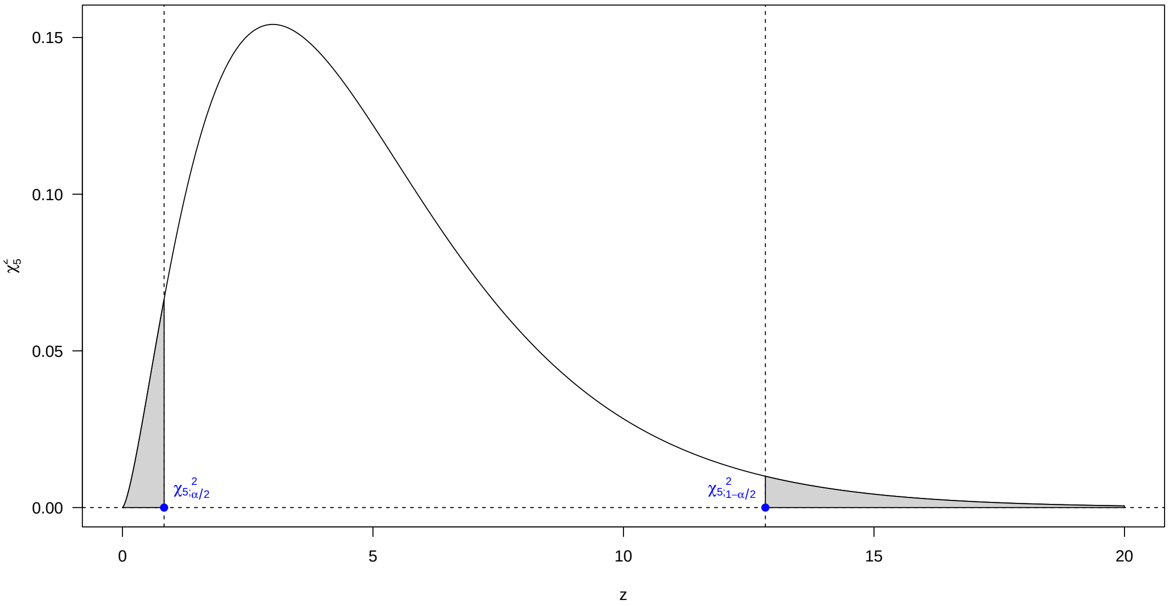

Figure 1:

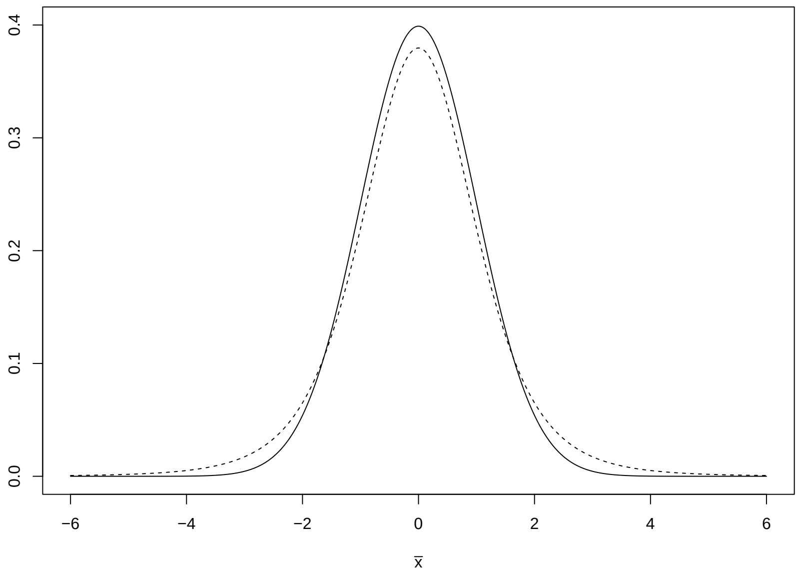

Figure 2:

Correction exercice n°1.

1. La densité est obtenues en dérivant la fonction de répatition et on obtient :

$$f_X(x) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{1}{\theta}x\right) & \text{si} \quad x \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

Son espérance se calcule avec :

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f_X(x)\,dx \\ &= 0 + \frac{1}{\theta}\int_0^{+\infty} x \times \exp\left(-\frac{1}{\theta}x\right) dx. \end{aligned}$$

En posant $u(x) = x$, $du = dx$ et $dv = \exp\left(-\frac{1}{\theta}x\right) dx$, $v(x) = -\theta\exp\left(-\frac{1}{\theta}x\right)$, on obtient en intégrant par partie :

$$\begin{aligned} E(X) &= \frac{1}{\theta}\int_{-\infty}^{+\infty} u \times dv \\ &= \frac{1}{\theta}\left[u(x) \times v(x)\right]_0^{+\infty} - \frac{1}{\theta}\int_0^{+\infty} v \times du \\ &= \frac{1}{\theta}\left[-\theta x\exp\left(-\frac{1}{\theta}x\right)\right]_0^{+\infty} + \frac{1}{\theta}\int_0^{+\infty} \theta exp\left(-\frac{1}{\theta}x\right) dx \\ &= 0 + \left[-\theta\exp\left(-\frac{1}{\theta}x\right)\right]_0^{+\infty} = \theta. \end{aligned}$$

2. Calcul de l'espérance de T_n :

$$\begin{aligned} E(T_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n} \times n\theta \\ &= \theta \end{aligned}$$

L'estimateur T_n est sans biais. Le calcul de la variance est le suivant :

$$\begin{aligned} V(T_n) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} \times nV(X) \\ &= \frac{\theta^2}{n}. \end{aligned}$$

On voit facilement que

$$\lim_{n\to+\infty} V(T_n) = 0.$$

L'estimateur T_n est sans biais et convergent. On peut l'utiliser pour estimer le paramètre θ, à partir d'une réalisation de taille n d'un échantillon.

Correction exercice n°2.

1. On suppose que $X \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$, les paramètres de la gaussienne étant inconnus. On suppose l'échantillon $\{X_1, \cdots, X_6\}$ *i.i.d* et de même loi mère que X. On utilise les estimateurs classiques : pour estimer la moyenne populationnelle μ, on utilise la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et pour estimer la variance σ^2, l'estimateur sans biais $S_{n-1}^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$, avec $n = 6$. On sait que la variable

$$Z = \frac{(n-1)\, S_{n-1}^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2.$$

On veut tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ contre l'alternative $H_1 : \sigma^2 \neq \sigma_0^2$ avec $\sigma_0^2 = 25\,\mu g/l$. Sous H_0, la variable

$$Z = \frac{5 \times S_5^2}{\sigma_0^2} \rightsquigarrow \chi_5^2,$$

c'est ce que représente la figure 1. Sous H_1, la variable S_5^2 et donc Z prendront des valeurs plus petites ou plus grandes puisque la machine serait mal calibrée. On est amené à construire un test bilatéral avec zone de rejet à gauche et à droite, matérialisées sur la figure 1. par les zones en griset. Fixons le niveau du test à $\alpha = 0.05$. Les bornes de rejet du test sont données par le quantile d'ordre $\alpha/2$ du χ_5^2, c'est à dire $\chi_{5;0.025}^2 = 0.83$ et par le quantile d'ordre $1 - \alpha/2$ soit $\chi_{5;0.975}^2 = 12.83$. La zone de non-rejet de H_0 est donc donnée par

$$\overline{RH_0} = [0.83;\ 12.83].$$

On a observé les valeurs $\overline{x}_{obs} = 103.63\,\mu g/l$, $s_{obs}^2 = \frac{128}{5}\,mg^2/l^2$ et on en déduit $\chi_{obs}^2 = \frac{128}{25} = 5.12$. On constate que $\chi_{obs}^2 \in \overline{RH_0}$: l'hypothèse H_0 n'est pas rejetée. Avec une probabilité de 0.95, il n'y a pas de raison de considérer que la machine est mal calibrée.

2. On sait que $Z \rightsquigarrow \chi_5^2$. On peut calculer la probabilité que Z soit comprise entre deux quantiles d'ordre fixé :

$$P\left(\chi_{5;\alpha/2}^2 \leq Z \leq \chi_{5;1-\alpha/2}^2\right) = 1 - \alpha.$$

Si on fixe le risque $\alpha = 0.05$, on obtient $\chi_{5;0.025}^2 = 0.83$ et $\chi_{5;0.975}^2 = 12.83$. L'$IC_{0.95}$ est donc le suivant

$$0.83 \leq \frac{5S_5^2}{\sigma^2} \leq 12.83$$
$$\frac{5S_5^2}{12.83} \leq \sigma^2 \leq \frac{5S_5^2}{0.83}.$$

Les bornes de cet intervalle sont aléatoires et dépendent de la valeur prise par S_5^2. On a observé $s_{obs}^2 = \frac{128}{5}$ et donc

$$9.98 = \frac{128}{12.83} \leq \sigma^2 \leq \frac{128}{0.83} = 154.22.$$

Il y a donc 95 % de chance d'avoir un écart-type de la population compris entre ces deux valeurs. Ces résultats sont compatibles avec ceux de la question 1.

3. On veut ici tester une hypothèse sur la moyenne de la population dans deux cas.

   (a) Considérons le premier cas : celui où la variance de la population est connue et telle que $\sigma_0^2 = 25$. On souhaite tester l'hypothèse $H_0 : \mu = \mu_0$ contre l'alternative $H_1 : \mu > \mu_0$. Si l'on suppose que les X_i sont gaussiennes d'espérance μ et de variance σ_0^2, alors la moyenne $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu, \sigma_0^2/n\right)$et sa version centrée-réduite :

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma_0}{\sqrt{n}}}. \rightsquigarrow \mathcal{N}(0,1).$$

   Sous H_0, $\overline{X} \rightsquigarrow \mathcal{N}\left(\mu = \mu_0, \sigma_0^2/n\right)$. Sous H_1, l'eau est plus polluée, les valeurs de $\overline{X}$ seront plus élevées que sous H_0, celles de Z également. Il s'agit donc d'effectuer un test gaussien unilatéral avec zone de rejet à droite. Fixons $\alpha = 0.05$. La zone de rejet de H_0 est donnée par $RH_0 =]z_{1-\alpha}; +\infty[$ où $z_{1-\alpha} = z_{0.95} = -z_{0.05} = 1.64$ par symétrie de la gaussienne. On a observé

$$z_{obs} = \frac{\overline{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}} = \frac{103.63 - 100}{\frac{5}{\sqrt{6}}} = 1.78.$$

On voit immédiatement que $z_{obs} \in RH_0$: l'échantillon proviendrait d'eaux polluées, ceci en supposant que la variance populationnelle est bien égale à $\sigma_0^2 = 25$.

(b) Voyons voir le second cas : celui où σ_0 doit être estimé avec l'échantillon. Dans ce cas, si les X_i sont gaussiennes, on sait que la moyenne $\overline{X}$ est gaussienne de moyenne μ et de variance σ_0^2 inconnues. Dans ce cas, la variable

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{S_{n-1}^2}{n}}}$$

suit une loi de Student à $n-1$ degré de liberté. L'écart-type est ici estimé avec sa version sans biais

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Rappelons que l'on a observé

$$s_{obs}^2 = \frac{128}{5}.$$

Sous H_0, $Z \rightsquigarrow t\,(5)$ avec $\mu = \mu_0$. Sous H_1, l'eau est plus polluée, les valeurs de $\overline{X}$ (donc de Z) seront plus élevées que sous H_0. Il s'agit donc d'effectuer un test de Student unilatéral avec zone de rejet à droite. Fixons $\alpha = 0.05$ comme précédemment. La zone de rejet de H_0 est donnée par $RH_0 =]t_{5;1-\alpha}; +\infty[$ où $t_{5;1-\alpha} = t_{5;0.95} = -t_{5;0.05} = 2.01$ par symétrie de la loi de Student. On a observé

$$z_{obs} = \frac{\overline{x} - \mu_0}{\sqrt{\frac{s_{obs}^2}{n}}} = \frac{103.63 - 100}{\sqrt{\frac{128}{5\times 6}}} = 1.76.$$

On voit immédiatement que $z_{obs} \notin RH_0$: l'échantillon proviendrait d'eaux non polluées, ceci en utilisant un estimateur de la variance. Ces résultats sont contradictoires avec le test précédent.

4. En fait, la figure 2. représente la distribution gaussienne et la distribution de Student sous H_0. Ce que l'on constate, c'est que la loi de Student en pointillée, est plus étalée que la gaussienne, du fait d'avoir rajouter une source d'incertitude en évaluant la variance avec un échantillon aléatoire de petit effectif ($n = 6$). Les bornes de rejet sont donc reculées par rapport à celles de la loi normale. Et comme la valeur de z_{obs} est proche de la zone de rejet, dans un cas le test permet de rejetter H_0, mais pas dans l'autre. Si l'on considère un test gaussien plus sévère, avec un risque de rejetter à tort l'hypothèse nulle en diminution ($\alpha = 0.01$), alors l'hypothèse nulle ne serait pas rejettée, comme pour le test de Student. Nous sommes ici dans un cas où la décision est incertaine : il faudrait idéalement reconduire des expériences pour augmenter l'effectif de l'échantillon.