Université de Montpellier **HLMA408**
2018/2019 **Joseph Salmon**

TD N° 2 : Estimation ponctuelle

Exercice 1. (QQ-plot)
Ci-dessous vous trouverez les quantiles associés aux probabilités $0.05, 0.10, \ldots, 0.95$ de la durée de la grossesse (en jours) pour les mères de l'étude CHDS. Tracez ces quantiles en fonction de ceux d'une loi $\mathcal{U}(0,1)$ uniforme sur le segment $[0,1]$. Décrire la forme de la distribution de la durée de la grossesse par rapport à la loi uniforme.

250, 262, 267, 270, 272, 274, 275, 277, 278, 280, 282, 283, 284, 286, 288, 290, 292, 295, 302.

Exercice 2. (Espérance et aléatoire)
On considère une population de 6 individus sur lesquels une variable x vaut

$$x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 4, x_5 = 4, x_6 = 5$$

Dans la suite on travaille sur un échantillon aléatoire simple de 2 individus.

a) Calculez la distribution exacte de la moyenne de x sur l'échantillon.

b) Utilisez cette distribution exacte pour calculer l'espérance et la variance de cet estimateur.

Correction:

a) On rappelle qu'un échantillonnage aléatoire simple consiste à piocher n individus parmi N sans remise. Puisque ici $N = 6$ et $n = 2$, on a donc $\binom{6}{2} = 15$ échantillons possibles qui sont

$$(1,2),(1,2),(1,4),(1,4),(1,5),(2,2),(2,4),(2,4),(2,5),(2,4),(2,4),(2,5),(4,4),(4,5),(4,5),$$

et dont les moyennes sont respectivement

$$1.5, 1.5, 2.5, 2.5, 3, 2, 3, 3, 3.5, 3, 3, 3.5, 4, 4.5, 4.5.$$

La distribution exacte de la moyenne est donc

$$\Pr[\bar{x}_n = k] = \begin{cases} 2/15, & \text{si } k = 1.5 \\ 1/15, & \text{si } k = 2 \\ 2/15, & \text{si } k = 2.5 \\ 1/3, & \text{si } k = 3 \\ 2/15, & \text{si } k = 3.5 \\ 1/15, & \text{si } k = 4 \\ 2/15, & \text{si } k = 4.5 \\ 0, & \text{sinon.} \end{cases}$$

On a alors

$$\mathbb{E}[\overline{X}] = 1.5 \times \frac{2}{15} + \cdots + 4.5 \times \frac{2}{15} = 3,$$

et

$$\mathrm{Var}[\overline{X}] = \mathbb{E}[\overline{X}^2] - \mathbb{E}[\overline{X}]^2 = 1.5^2 \times \frac{2}{15} + \cdots + 4.5^2 \times \frac{2}{15} - 3^2 = 9.8 - 9 = 0.8.$$

b) Pour calculer l'espérance et la variance de cet estimateur on utilise alors

$$\mathbb{E}[\bar{x}_n] = 1.5 \times \frac{2}{15} + \cdots + 4.5 \times \frac{2}{15} = 3,$$

et

$$\mathrm{Var}[\bar{x}_n] = \mathbb{E}[\bar{x}_n^2] - \mathbb{E}[\bar{x}_n]^2 = 1.5^2 \times \frac{2}{15} + \cdots + 4.5^2 \times \frac{2}{15} - 3^2 = 9.8 - 9 = 0.8.$$

Exercice 3. (Quantiles gaussiens)

Supposez que les quantiles y_p d'une loi $\mathcal{N}(\mu, \sigma^2)$ sont tracés en fonction des quantiles z_p d'une loi $\mathcal{N}(0,1)$. Montrez que la pente et l'ordonnée à l'origine de la droite des points sont σ et μ respectivement.

Correction:

Soient $z_p = \Phi^{-1}(p)$ et $y_p = \Phi^{-1}_{\mu,\sigma^2}(p)$. Ainsi $p = \Phi_{\mu,\sigma^2}(x) = \mathbb{P}(Z \leqslant x) = \mathbb{P}(\mu + \sigma X \leqslant x) = \mathbb{P}(X \leqslant \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$ avec $X \sim \mathcal{N}(0,1)$. Ainsi

$$
\begin{aligned}
& p = \Phi_{\mu,\sigma^2}(y_p) \\
\iff \quad & p = \Phi_{0,1}(\frac{y_p - \mu}{\sigma}) \\
\iff \quad & \Phi^{-1}_{0,1}(p) = \frac{y_p - \mu}{\sigma} \\
\iff \quad & y_p = \sigma \Phi^{-1}_{0,1}(p) + \mu = \sigma z_p + \mu
\end{aligned}
$$

Exercice 4. (Déterministe vs. aléatoire)

En gardant les notations du cours, dîtes quels sont les éléments aléatoires ou non et expliquez pourquoi.

$$x_1, x_{i_2}, \bar{x}_n, N, \mu, i_1, n$$

Correction:

- x_1 n'est pas aléatoire c'est le premier individu.
- x_{i_2} est aléatoire c'est la valeur prise par le premier « individu pioché ». Comme on ne peut pas savoir quel sera ce premier individu on ne peut pas connaître sa valeur.
- $\overline{X}$ est aléatoire c'est une moyenne sur n individus piochés au hasard or on ne sait pas lesquels seront piochés.
- μ est un paramètre, c'est la moyenne sur la population. Il n'est pas aléatoire.
- i_1 est aléatoire c'est le « numéro » du premier individu pioché.
- n n'est pas aléatoire c'est le nombre d'individu choisi au hasard lors de notre échantillonnage aléatoire simple.

Exercice 5. (Intervalle de confiance et théorème central limite)

Lors d'un contrôle de qualité dans une firme pharmaceutique, la quantité d'acide acétylsalicylique x dans un comprimé d'aspirine a été mesurée pour $n = 500$ comprimés prélevés dans une production de $N = 500000$ comprimés. On a ainsi obtenu après collecte des 500 mesures :

$$\bar{x}_n = \frac{1}{500} \sum_{i=1}^{500} x_i = 0.496\, mg \ .$$

On suppose connue la variance (théorique) des mesures, qui vaut $\sigma^2 = 0.004\, mg^2$. Construire un intervalle de confiance au niveau 95% de la quantité d'acide acétylsalicylique dans un comprimé (on utilisera pour cela une approximation donnée par le théorème central limite).

Correction:

to update... plus fait comme ça dans le cours, du coup prendre l'ecart-type théorique fourni. ne pas parler de "facteur correctif".

Exercice 6. (Covariance et échantillon aléatoire)

On considère un échantillon de taille $n = 2$ issu d'un échantillonnage aléatoire simple sur une population de taille $N = 100$. Ainsi on tire un couple (x_{i_1}, x_{i_2}) de manière uniforme parmi tous les couples possibles. Supposons que $x_i = 0$ ou 1 pour tout $i \in [\![1, N]\!]$ et que la proportion de 1 dans la population est p. Calculez $\mathbb{E}[x_{i_1} x_{i_2}]$ et en déduire la covariance entre x_{i_1} et x_{i_2} pour ce cas particulier.

Correction:

On a

$$\mathbb{E}[x_{i_1} x_{i_2}] = \Pr[x_{i_1} = 1, x_{i_2} = 1] = \frac{100p}{100} \times \frac{100p - 1}{99} = \frac{p(100p-1)}{99}.$$

Par définition on a

$$\mathrm{Cov}(x_{i_1}, x_{i_2}) = \mathbb{E}[x_{i_1} x_{i_2}] - \mathbb{E}[x_{i_1}]\mathbb{E}[x_{i_2}] = \frac{p(100p-1)}{99} - p^2 = \frac{p(p-1)}{99}$$

Exercice 7. (Moyenne empirique et optimisation)

Montrez que $\bar{x}_n$ (moyenne empirique des $x_1, \ldots, x_n$) est la valeur qui minimise la fonction f définie pour tout $x \in \mathbb{R}$ par

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2 \quad .$$

Aide : Montrer que pour tout réel x, la relation suivante est vraie :

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2 \quad .$$

En déduire le résultat en étudiant la formulation quadratique ainsi donnée.

Correction:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n + \bar{x}_n - x)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_n - x)^2 + \frac{2}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)(\bar{x}_n - x) \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2 + \frac{2}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)(\bar{x}_n - x) \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2 + \frac{2(\bar{x}_n - x)}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n) \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2 + \frac{2(\bar{x}_n - x)}{n} \underbrace{\left(\sum_{i=1}^{n} x_i - n\bar{x}_n \right)}_{=0} \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2
\end{aligned}$$

Pour l'exercice suivant on pourra utiliser :

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2 - (\bar{x}_n - x)^2$$

Exercice 8. (Biais de la variance empirique)

On suppose que le $x_1, \ldots, x_n$ sont *i.i.d.* et ont comme espérance μ et comme variance σ^2. En déduire la valeur du biais de la variance empirique

$$s_n^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

c'est-à-dire la valeur de $\mathbb{B}(s_n^2(\mathbf{x})) = \mathbb{E}(s_n^2(\mathbf{x})) - \sigma^2$. Proposer une modification de l'estimateur précédent pour le rendre non biaisé. Aide : utiliser l'exercice précédent.

Correction:

Rappel : il peut être bon de rappeler les propriétés simples de l'espérance et de la variance dont on se

servira dans la suite.
En appliquant la relation de l'exercice précédent pour avec le choix $x = \mu$, on obtient (notamment à la 3e ligne) :

$$\begin{aligned}
\mathbb{B}(s^2(\mathbf{x})) =& \mathbb{E}(s^2(\mathbf{x})) - \sigma^2 \\
=& \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right) - \sigma^2 \\
=& \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x}_n - \mu)^2\right) - \sigma^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(x_i) - \mathbb{E}(\bar{x}_n - \mu)^2 - \sigma^2 \\
=& \mathbb{V}\mathrm{ar}(x_1) - \mathbb{V}\mathrm{ar}(\bar{x}_n) - \sigma^2 \\
=& \sigma^2 - \frac{\sigma^2}{n} - \sigma^2 = -\frac{1}{n}\sigma^2
\end{aligned}$$

Ainsi on peut montrer que $\frac{n}{n-1}s^2(\mathbf{x})$ est un estimateur sans biais de la variance σ^2 : en effet on a montré que $\mathbb{E}(s^2(\mathbf{x})) = \frac{n-1}{n}\sigma$ au dessus, et donc $\mathbb{E}(\frac{n}{n-1}s^2(\mathbf{x})) = \sigma^2$.