

TD n° 2 : Estimation ponctuelle

EXERCICE 1. (QQ-plot)

Ci-dessous vous trouverez les quantiles associés aux probabilités 0.05, 0.10, ..., 0.95 de la durée de la grossesse (en jours) pour les mères de l'étude CHDS vue en cours. Tracez (idéalement utiliser un notebook pour cela) 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.

 Les quantiles peuvent être obtenus depuis la base de données `babies23.data` en utilisant les lignes de codes disponibles dans `StatsDescriptives.ipynb`.

Correction:

Commençons par calculer les quantiles de la loi uniforme $\mathcal{U}(0, 1)$. La fonction de répartition F de cette loi est très simple :

$$F(x) = \begin{cases} x, & \text{si } x \in [0, 1]; \\ 0, & \text{sinon.} \end{cases}$$

Si on note le quantile d'ordre p de cette loi par q_p , alors par définition $q_p = q$ est équivalent à $F(q) = p$; mais puisque par définition $q \in [0, 1]$, on a $F(q) = q$ et donc au final $q_p = p$ (ou dit autrement la fonction identité de $[0, 1]$ dans $[0, 1]$ est bijective et elle est égale à sa fonction réciproque.) Du coup, la courbe quantile contre quantile est la courbe passant par les 19 points de coordonnées suivants :

$$(0.05, 250), (0.1, 262), (0.15, 267), \dots, (0.95, 302).$$

On peut tracer ces points à la main ou en utilisant un notebook, on obtient alors le graphe de la Figure 1. En calculant les quantiles à partir des données, on ne retrouve pas exactement les valeurs de l'énoncé. On remarque sur la figure que si on oublie les points extrémaux, les points de ce nuage sont presque alignés. On en déduit que si Y est la variable aléatoire modélisant les durées de grossesse et X est la loi uniforme sur l'intervalle $[0, 1]$, on peut supposer que l'on est proche d'une relation du type $Y = aX + b$, avec a et b des constantes. Et donc la variable Y est proche d'une loi uniforme sur l'intervalle image de $[0, 1]$ par la fonction $f(x) = ax + b$, c'est à dire l'intervalle $[b, b + a]$. Pour en dire plus, il faudrait calculer les valeurs de a et b de sorte que la droite $y = ax + b$ soit la plus proche possible des points de la figure. Ceci sera l'objet d'un chapitre ultérieur de ce cours.

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 (tirage sans remise) de 2 individus :

- Donner la loi de distribution de la moyenne sur l'échantillon.
- Utilisez cette distribution pour calculer l'espérance et la variance de cet estimateur.

Correction:

- 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 espérances sont respectivement

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

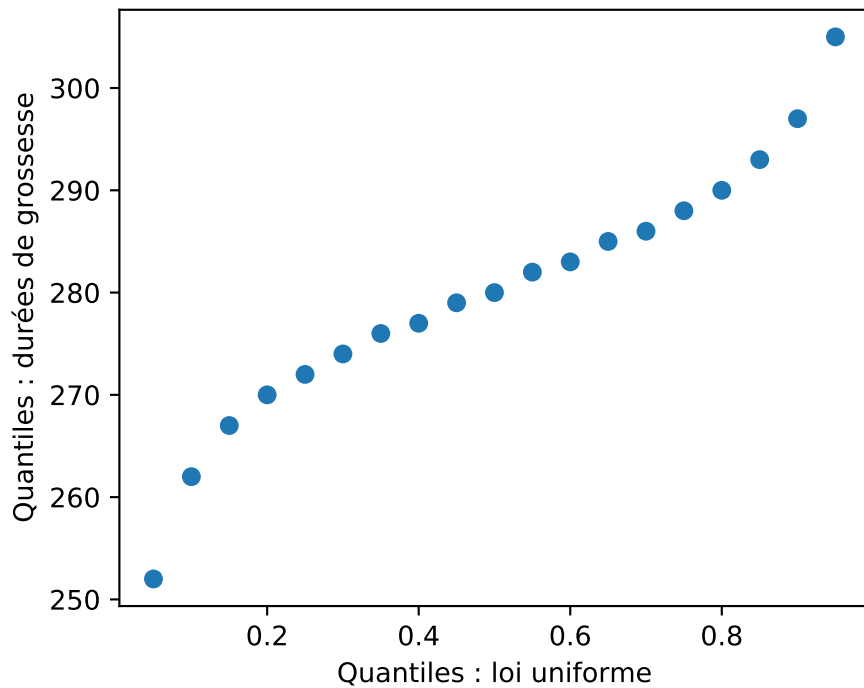


FIGURE 1 – Le nuage de points “quantiles contre quantiles”.

La distribution 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

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

et

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

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

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

et

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

EXERCICE 3. (Quantiles gaussiens)

Supposez que les quantiles y_p (pour tout $p \in \mathbb{R}$) 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 $y_p = \Phi_{\mu, \sigma^2}^{-1}(p)$. Tout d'abord on a que $Z = \mu + \sigma X$. Ainsi $p = \Phi_{\mu, \sigma^2}(x) = \mathbb{P}(Z \leq x) = \mathbb{P}(\mu + \sigma X \leq x) = \mathbb{P}(X \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$ avec $X \sim \mathcal{N}(0, 1)$. Ainsi

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

EXERCICE 4. (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 $i \in \llbracket 1, N \rrbracket$ et que la proportion de 1 dans la population soit (en espérance) p . Calculez $\mathbb{E}[x_{i_1} x_{i_2}]$ et en déduire la covariance entre x_{i_1} et x_{i_2} .

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

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

$$\begin{aligned} \mathbb{E}[x_{i_1}] &= \mathbb{P}(x_{i_1} = 1) = p, \\ \mathbb{E}[x_{i_2}] &= \mathbb{P}(x_{i_2} = 1) = \Pr[x_{i_1} = 1, x_{i_2} = 1] + \Pr[x_{i_1} = 0, x_{i_2} = 1] \\ &= p \frac{100p-1}{99} + (1-p) \frac{100p}{99} = p. \end{aligned}$$

Remarquons que par symétrie on s'attendait à $\mathbb{E}[x_{i_1}] = \mathbb{E}[x_{i_2}]$

EXERCICE 5. (Moyenne empirique et optimisation)

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

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

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

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

En déduire le résultat en étudiant la formulation quadratique ainsi donnée. Que vaut la fonction f quand elle atteint son minimum ?

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 6. (Biais de la variance empirique)

On suppose que le x_1, \dots, 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 la relation (1) de 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 3^e ligne) :

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

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 $\mathbb{E}(\frac{n}{n-1}s^2(\mathbf{x})) = \sigma^2$.

EXERCICE 7. (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 $N = 500\,000$ comprimés. On a ainsi obtenu après collecte des 500 mesures :

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

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

Correction: