Correction TD 2 : Estimation

Correction exercice 1. Manipulation de lois normales, et lecture de table

Prenez bien soin de faire un dessin, cela aide énormément pour tous ces exercices.

1. Dans cette table, on lit directement la fonction de répartition, donc il faut regarder en face de la ligne 1.2 et de la colonne 0.04 pour pouvoir lire $\mathbb{P}(X \leq 1.24)$. On trouve

$$p_1 = \mathbb{P}(X \leq 1.24) = 0.89251.$$

On peut de même chercher un quantile dans la table, c'est à dire chercher le nombre pour lequel la fonction de répartion se lit dans la table, ici, on trouve alors, en face de 0.75175,

$$q_1 = 0.68$$

2. Pour tout $t \in \mathbb{R}$, on a $\mathbb{P}(X = t) = 0$, comme pour toutes les variables aléatoires à densité. On a donc

$$\mathbb{P}(X \geq t) = \mathbb{P}(X > t) + \mathbb{P}(X = t) = 1 - \mathbb{P}(X \leq t) + 0 = 1 - \mathbb{P}(X \leq t).$$

3. On rappelle que la loi normale centrée réduite a pour densité la fonction $f_{\mathcal{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, qui est paire, on a donc

$$\mathbb{P}(X \leq -t) = \int_{-\infty}^{-t} f(x)\mathrm{d}x = \int_{t}^{+\infty} f(x)\mathrm{d}x = \mathbb{P}(X \geq t) = 1 - \mathbb{P}(X \leq t).$$

4. Pour trouver p_2 dans la table, on utilise la propriété précédente, on a donc

$$p_2 = \mathbb{P}(X \leq -1.04) = 1 - \mathbb{P}(X \leq 1.04),$$

ce dernier nombre peut se lire directement dans la table : $\mathbb{P}(X \leq 1.04) = 0.85083$, d'où

$$p_2 = 1 - 0.85083 = 0.14917.$$

De même, on cherche à se ramener à une probabilité supérieure à $1/2$, ce qu'on peut faire en utilisant

$$\mathbb{P}(X \leq -q_2) = 1 - \mathbb{P}(X \leq q_2) = 1 - 0.4522 = 0.54776$$

On peut alors chercher le quantile dans la table, et on trouve

$$-q_2 = 0.12 \Leftrightarrow q_2 = -0.12.$$

5. Il suffit d'écrire

$$\mathbb{P}(|X| < t) = 1 - \mathbb{P}(|X| > t) = 1 - \Big(\mathbb{P}(X > t) + \mathbb{P}(X < -t)\Big) = 1 - \Big(1 - \mathbb{P}(X \leq t) + 1 - \mathbb{P}(X < t)\Big) = 2\mathbb{P}(X \leq t) - 1.$$

6. On peut écrire, en lisant dans la table $\mathbb{P}(X \leq 1.96) = 0.975$,

$$p_3 = \mathbb{P}(|X| \leq 1.96) = 2\mathbb{P}(X \leq 1.96) - 1 = 2 \times 0.975 - 1 = 0.95.$$

De la même façon, on peut chercher q_3 en remarquant que

$$\begin{aligned}
\mathbb{P}(|X| \geq q_3) = 0.101 &\Leftrightarrow \mathbb{P}(|X| \leq q_3) = 1 - 0.101 \\
&\Leftrightarrow 2\mathbb{P}(X \leq q_3) - 1 = 0.899 \\
&\Leftrightarrow \mathbb{P}(X \leq q_3) = \frac{0.899 + 1}{2} = 0.9495
\end{aligned}$$

Ce qui permet de lire dans la table $q_3 = 1.68$.

7. Le théorème de stabilité des lois normales indique que $2X+1$ reste de loi normale. Il ne reste plus qu'à calculer son espérance et sa variance. On trouve

$$\mathbb{E}[2X+1] = 2\mathbb{E}[X]+1 = 1$$
$$\mathrm{var}(2X+1) = 4\,\mathrm{var}(X)+\mathrm{var}(1) = 4.$$

Donc

$$Z \sim \mathcal{N}(1,4).$$

8. On a

$$\mathbb{P}(Z \leq 3.56) = \mathbb{P}(2X+1 \leq 3.56) = \mathbb{P}(X \leq \frac{3.56-1}{2}) = \mathbb{P}(X \leq 1.28) = 0.89973,$$

où la dernière valeur peut se lire dans la table.

Attention : Ici, on ne peut pas utliser directement la table, car Z n'est pas de loi $\mathcal{N}(0,1)$. On doit donc se ramener à une variable de loi $\mathcal{N}(0,1)$. La variable $X = \frac{Z-1}{2} \sim \mathcal{N}(0,1)$ par définition de Z, donc l'objectif est de se ramener à une probabilité qui la fait intervenir. En réalité, on pourra toujours le faire pour $Y \sim \mathcal{N}(m,\sigma^2)$, en posant $Y_{\text{centrée réduite}} = \frac{Y-m}{\sigma}$ qui suit bien une $\mathcal{N}(0,1)$.

On peut raisonner de même pour trouver q_4:

$$\begin{aligned}\mathbb{P}(Z > q_4) = 0.0401 &\Leftrightarrow \mathbb{P}\Big(\frac{Z-1}{2} > \frac{q_4-1}{2}\Big) = 0.04006 \\ &\Leftrightarrow \mathbb{P}\Big(\mathcal{N}(0,1) > \frac{q_4-1}{2}\Big) = 0.04006 \\ &\Leftrightarrow \mathbb{P}\Big(\mathcal{N}(0,1) \leq \frac{q_4-1}{2}\Big) = 1-0.04006 = 0.95994\end{aligned}$$

Ce qui permet de lire dans la table $\frac{q_4-1}{2} = 1.75$, ce qui donne $q_4 = 2.1.75+1 = 4.5$.

On synthétise les formules de relation entre les aires dans le graphique ci dessous (attention, les p_i ici ne sont pas celles de l'exo !) :

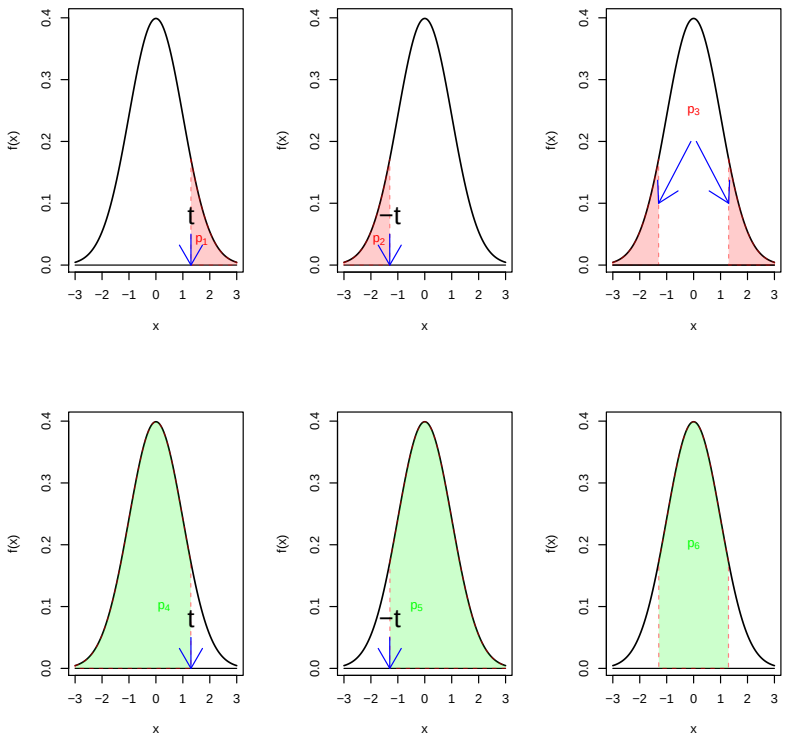

Figure 1: Cette figure permet de comprendre le lien entre les probabilités (attention, les p_i ici ne sont pas celles de l'exo !) $p_1 = \mathbb{P}(X > t), p_2 = \mathbb{P}(X < -t), p_3 = \mathbb{P}(|X| > t), p_4 = \mathbb{P}(X < t), p_5 = \mathbb{P}(X > -t), p_6 = \mathbb{P}(|X| < t)$. Par symétrie, on a $p_1 = p_2$ et $p_4 = p_5$, de plus, en utilisant que l'aire totale vaut 1, on a aussi $p_1 = 1-p_4$, $p_2 = 1-p_5$ et $p_3 = 1-p_6$. Enfin, on a $p_3 = 2p_1 = 2p_2$. En utilisant toutes ces égalités, on peut retrouver par exemple $\mathbb{P}(|X| < t) = p_6 = 1-p_3 = 1-2p_1 = 1-2(1-p_4) = 2p_4-1 = 2\mathbb{P}(X \leq t)-1$

Soit $X_1 \sim \mathcal{N}(1,9)$ et $X_2 \sim \mathcal{N}(-1,4)$, indépendantes.

1. Le théorème de stabilité des lois normale stipule que des combinaisons affines de lois normale indépendante reste de loi normale, il ne reste plus qu'à calculer les paramètres de cette loi.

   On a alors

$$\begin{aligned}
\mathbb{E}[Y] &= -2\Big(\mathbb{E}[X_1] - 2\mathbb{E}[X_2]\Big) \\
&= -2(1 - 2(-1)) \\
&= -6 \\
\operatorname{var}(Y) &= 4\operatorname{var}(X_1 - 2X_2) \\
&= 4\big(\operatorname{var}(X_1) + (-2)^2 \operatorname{var}(X_2)\big) \text{ par indépendance} \\
&= 4\big(9 + 16\big) \text{ par indépendance} \\
&= 100
\end{aligned}$$

   Donc $m = -6$ et $\sigma = 10$ ($\sigma^2 = 100$), et

$$Y \sim \mathcal{N}(-6, 100).$$

2. "Centrer et réduire" correspond à chercher une transformation affine $\tilde{Y}$ de Y qui soit de loi $\mathcal{N}(0,1)$. Il faut soustraire l'espérance (centrer) puis diviser par l'écart-type (réduire). On pose donc

$$\tilde{Y} = \frac{Y - (-6)}{10} = \frac{Y + 6}{10}.$$

   On pose donc

3. On peut alors calculer

$$\begin{aligned}
\mathbb{P}(Y > 13.6) &= \mathbb{P}\Big(\frac{Y+6}{10} > \frac{13.6+6}{10}\Big) \\
&= \mathbb{P}\big(\tilde{Y} > 1.96\big) \\
&= \mathbb{P}(\mathcal{N}(0,1) > 1.96) \\
&= 0.025 \text{ d'après la table}
\end{aligned}$$

4. De même, on peut chercher q en écrivant

$$\begin{aligned}
\mathbb{P}(|Y+6| > q) = 0.89656 &\Leftrightarrow \mathbb{P}\left(\frac{|Y+6|}{10} > \frac{q}{10}\right) = 0.89656 \\
&\Leftrightarrow \mathbb{P}\left(\left|\tilde{Y}\right| > \frac{q}{10}\right) = 0.89656 \\
&\Leftrightarrow \mathbb{P}\left(\left|\mathcal{N}(0,1)\right| > \frac{q}{10}\right) = 0.89656 \\
&\Leftrightarrow \mathbb{P}\Big(\mathcal{N}(0,1) > \frac{q}{10}\Big) = \frac{0.8966}{2} = 0.4483 \\
&\Leftrightarrow \mathbb{P}\Big(\mathcal{N}(0,1) < \frac{q}{10}\Big) = 1 - 0.44828 = 0.55172
\end{aligned}$$

   En utilisant la table, on trouve alors $\frac{q}{10} = 0.13$ et enfin

$$q = 1.3$$

Correction exercice 2. Les premiers estimateurs

Soient X_1, X_2 deux variables i.i.d., d'espérance m inconnue, et de variance σ^2 (que l'on suppose connue).

On pose $\hat{M}_1 = m, \hat{M}_2 = 1$, $\hat{M}_3 = 2X_1 + X_2, \hat{M}_4 = \frac{2X_1+X_2}{3}, \hat{M}_5 = \frac{X_1+X_2}{2}$.

1. Seule $\hat{M}_1$ n'est pas une statistique, parce qu'elle dépend de paramètres inconnus du modèle.

2. Il suffit de calculer les espérances :

$$\begin{aligned}
\mathbb{E}[\hat{M}_2] &= 1 \\
\mathbb{E}[\hat{M}_3] &= 2\mathbb{E}[X_1] + \mathbb{E}[X_2] = 2m + m = 3m \\
\mathbb{E}[\hat{M}_4] &= \frac{2\mathbb{E}[X_1] + \mathbb{E}[X_2]}{3} = m \\
\mathbb{E}[\hat{M}_5] &= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2]}{2} = m
\end{aligned}$$

   Seuls les estimateurs $\hat{M}_4$ et $\hat{M}_5$ sont donc sans biais.

3. Il suffit de calculer les variances de $\hat{M}_4$ et $\hat{M}_5$:

$$\begin{aligned}
\operatorname{var}\left(\hat{M}_4\right) &= \frac{4\operatorname{var}(X_1) + \operatorname{var}(X_2)}{9} \\
&= \frac{5}{9}\sigma^2 \\
\operatorname{var}\left(\hat{M}_5\right) &= \frac{\operatorname{var}(X_1) + \operatorname{var}(X_2)}{4} \\
&= \frac{1}{2}\sigma^2
\end{aligned}$$

   On observe donc que $\hat{M}_5$ a une plus petite variance que $\hat{M}_4$, il lui sera donc préférable, au sens du critère de la variance minimale pour des estimateurs sans biais.

4. On a $\operatorname{var}(\hat{M}_2) = 0$, on remarque donc que $\hat{M}_2$ a une variance très petite. Pour autant, on ne le préfèrera pas, car il s'agit d'un estimateur biaisé.

Correction exercice 3. Les premiers estimateurs

Soient X_1, X_2 deux variables indépendantes, d'espérance m inconnue de variances respectives σ_1^2 et σ_2^2, connues. On cherche à définir un estimateur de m de la forme $\hat{M} = aX_1 + bX_2$.

1. Pour que $\hat{M}$ soit sans biais, il faut et il suffit que $\forall m \in \mathbb{R}, \mathbb{E}[\hat{M}] = m$. Cela donne $\forall m \in \mathbb{R}, am+bm = m$, i.e. $a + b = 1$.

   Il s'agit bien d'une condition nécessaire et suffisante.

2. On peut alors directement calculer la variance de $\hat{M}$:

$$\begin{aligned}
\operatorname{var}\left(\hat{M}\right) &= a^2\operatorname{var}(X_1) + b^2\operatorname{var}(X_2) \text{ en utilisant l'indépendance.} \\
&= a^2\sigma_1^2 + b^2\sigma_2^2
\end{aligned}$$

3. On pose $b = 1 - a$, et on obtient alors la formule de la variance de $\hat{M}_a = aX_1 + (1-a)X_2$ donnée par $f(a) := \operatorname{var}\left(\hat{M}_a\right) = a^2\sigma_1^2 + (1-a)^2\sigma_2^2$. Pour minimiser cette variance, il suffit d'annuler la dérivée par rapport à a :

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}a} f(a^\star) = 0 &\Leftrightarrow 2a^\star\sigma_1^2 - 2(1-a^\star)\sigma_2^2 = 0 \\
&\Leftrightarrow a^\star = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}
\end{aligned}$$

   On peut aussi montrer qu'il s'agit bien d'un minimum, par exemple en redérivant par rapport à a:

$$\frac{\mathrm{d}^2}{\mathrm{d}a^2} f(a^\star) = 2\sigma_1^2 + 2\sigma_2^2.$$

   Cette dernière quantité étant positive, $a^\star$ correspond bien à l'argmin de f (et c'est le seul, vu que f est $\mathcal{C}_1$.)

On trouve finalement l'estimateur sans biais de variance minimale, qui est donné par

$$\hat{M^\star} = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}X_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}X_2.$$

Au passage, on peut remarque que cela revient à utiliser beaucoup plus l'information fournise par la variable ayant la plus faible variance pour estimer la moyenne. Par exemple, si $\sigma_1^2 >> \sigma_2^2$, alors $\hat{M^\star} \approx X_2$, et on n'utilise presque pas l'information fournie par X_1, en considérant que sa variance est trop grande par rapport à celle de X_2. Au passage, cette façon d'estimer peut être utilisée pour agréger simplement des estimateurs ayant des variances différentes (dans ce cas là, X_1 et X_2 seraient eux-même des estimateurs provenant de sources différentes par exemple, ce qui permet l'indépendance).

Correction exercice 4. Principe d'un intervalle de confiance et idée d'un test statistique

Un déodorant affiche une durée de protection moyenne de $6h$. Un institut prend 16 mesures (que l'on considère comme fiables) de durées de protection, et on admet que l'on peut modéliser ces durées comme des réalisations i.i.d. de lois normales d'espérance m inconnue, et d'écart-type 1 (en heures). Soit $X_1, \cdots, X_{16}$ les mesures prises.

1. On peut directement calculer l'espérance, en utilisant la linéarité :

$$\mathbb{E}[\bar{X}] = \frac{1}{16}\sum_{i=1}^{16}\mathbb{E}[X_i],$$

Comme les X_i ont même loi, elles ont en particulier même espérance m inconnue, ce qui donne

$$\mathbb{E}[\bar{X}] = \frac{1}{16}\sum_{i=1}^{16} m = \frac{1}{16}\times 16m = m.$$

2. Comme $\mathbb{E}[\bar{X}] = m$ (quelque soit m et $X_i \sim_{i.i.d} \mathcal{N}(m,1)$), et que $\bar{X}$ est bien une statistique, il s'agit d'un estimateur sans biais de m.

   $\bar{X}$ est-il un estimateur sans biais de m ?

3. Par stabilité des lois normales, $\bar{X}$ est de loi normale, et on a déjà calculé son espérance. Il reste à calculer sa variance. On a

$$\begin{aligned}
\text{var}(\bar{X}) &= \frac{1}{16^2}\,\text{var}\Big(\sum_{i=1}^{16} X_i\Big) \\
&= \frac{1}{16^2}\sum_{i=1}^{16}\text{var}\Big(X_i\Big) \text{ par indépendance} \\
&= \frac{1}{16^2}16\,\text{var}\Big(X_i\Big) \text{ par indépendance} \\
&= \frac{1}{16} \text{ car } \text{var}(X_i) = 1.
\end{aligned}$$

   On a donc

$$\bar{X} \sim \mathcal{N}(m, \frac{1}{16})$$

4. Par stabilité des lois normales, T est de loi normale, il reste à calculer son espérance et sa variance. On a

$$\begin{aligned}
\mathbb{E}[T] &= 4(\mathbb{E}[\bar{X} - m]) \\
&= 0 \\
\text{var}(T) &= 16\,\text{var}(\bar{X}) \\
&= 1
\end{aligned}$$

   Donc

$$T \sim \mathcal{N}(0,1).$$

5. On peut alors chercher a tel que

$$\begin{aligned}\mathbb{P}(|T| > a) = 0.1 &\Leftrightarrow \mathbb{P}(T > a) = 0.05 \text{ par symétrie} \\ &\Leftrightarrow \mathbb{P}(T < a) = 0.95\end{aligned}$$

En lisant la table, on trouve $a \approx 1.65$.

6. On vient de montrer que $\mathbb{P}\Big(|T| \leq 1.65\Big) = 0.95$.

On peut manipuler un peu l'événement $|T| \leq 1.65$ pour le reformuler sous la forme $m \in I$, où I sera un intervalle aléatoire :

$$\begin{aligned}|T| \leq 1.65 &\Leftrightarrow \Big|4(\bar{X} - m)\Big| \leq 1.65 \\ &\Leftrightarrow \Big|\bar{X} - m\Big| \leq \frac{1.65}{4} \\ &\Leftrightarrow -\frac{1.65}{4} \leq m - \bar{X} \leq \frac{1.65}{4} \\ &\Leftrightarrow \bar{X} - \frac{1.65}{4} \leq m \leq \bar{X} + \frac{1.65}{4}\end{aligned}$$

On peut donc poser $A_1 = A_1(X_1, ..., X_{16}) = \bar{X} - \frac{1.65}{4}$, $A_2 = A_2(X_1, ..., X_{16}) = \bar{X} + \frac{1.65}{4}$, et $I = [A_1, A_2]$ pour avoir

$$\mathbb{P}(m \in I) = 0.95.$$

L'intervalle I est aléatoire, donc la probabilité porte sur I et non sur m (qui n'est pas aléatoire). Cela signifie que s'il on refait l'expérience, l'intervalle I changera, et que sur 95% des expériences, le paramètre inconnu m sera dans l'intervalle I.

On parle d'intervalle de confiance pour m avec un degrés de 95%, et on le notera $IC_{95\%}(m)$. Attention, si on fait une application numérique, et qu'on trouve par exemple $a_1 = 5.8, a_2 = 6.7$, on peut écrire l'intervalle de confiance **ponctuel** (ponctuel faisant ici référence à une réalisation de l'aléa) $ic_{95\%}(m) = [5.8, 6.7]$, mais il ne faudra jamais écrire $\mathbb{P}(m \in [5.8, 6.7]) = 0.95$, vu que dans cette dernière expression, plus rien n'est aléatoire !!

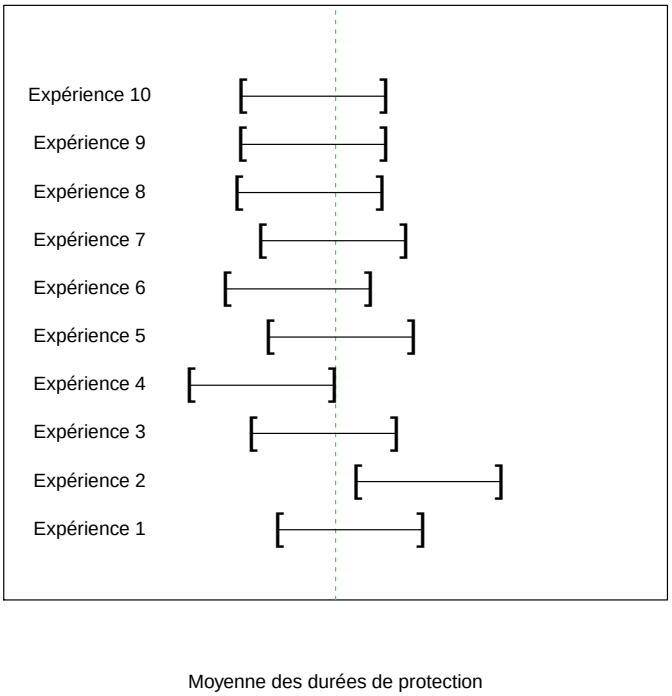

Figure 2: Cette figure illustre la variabilité des intervalles de confiance. Chaque expérience correspond à 16 mesures, puis au calcul de la moyenne empirique et de l'intervalle de confiance. En vert, la "vraie" valeur inconnue, mais commune à chaque expérience. À chaque expérience, l'intervalle de confiance change, mais on a une garantie théorique qu'avec une probabilité de 95%, il contient la "vraie valeur". Pour autant, une fois que l'on a fait l'expérience, il est impossible de savoir si l'on est ou non dans les 5% qui posent problème.

Correction exercice 5. Sondage

Du contexte pour la culture :

On suppose que l'on interroge 1000 personnes tirées au hasard, uniformément dans la population (et indépendamment, avec remise), qu'elles acceptent toutes de répondre, qu'elles ne mentent pas, et que parmi elles k disent qu'elles vont voter pour le candidat A, et $1000-k$ pour le candidat B.

1. X_1 est à valeurs dans $\{0,1\}$, donc X_1 suit la loi de Bernoulli. Le paramètre p correspond à la proportion de personnes (dans la population totale) de personnes qui pensent voter pour le candidat A.
2. S_n est une somme de variables aléatoire de Bernoulli indépendantes. Donc $S_n \sim \mathcal{B}in(n,p)$.
3. La loi d'une variable aléatoire $\mathcal{B}in(n,p)$ est donné par $\forall k \in [\![0,n]\!], \mathbb{P}\Big(S_n = k\Big) = \binom{n}{k} p^k (1-p)^{n-k}$. En appliquant pour $k = 0, 1$, on trouve

$$\mathbb{P}(S_n = 0) = (1-p)^{1000}$$
$$\mathbb{P}(S_n = 1) = 1000p(1-p)^{999}$$

4. On a $\mathbb{E}[X_1] = p$ et $\mathrm{var}(X_1) = p(1-p)$. Le théorème central limite nous indique que

$$\sqrt{n}\frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\mathrm{var}(X_1)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

donc

$$Z_n := \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

5. Comme $Z \sim \mathcal{N}(0,1)$, on peut utiliser la table de quantiles de la loi normale centrée réduite pour trouver :

$$\begin{aligned}\mathbb{P}(|Z| > a) = 0.01 &\Leftrightarrow \mathbb{P}(Z > a) = \frac{0.01}{2} = 0.005 \\ &\Leftrightarrow \mathbb{P}(Z \leq a) = 1 - 0.005 = 0.995 \\ &\Leftrightarrow \mathbb{P}(\mathcal{N}(0,1) \leq a) = 1 - 0.005 = 0.995\end{aligned}$$

ce qui donne $a = 2.58$.

6. Comme $Z_n \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1)$, on a $\mathbb{P}(|Z_n| > a) \to \mathbb{P}(|Z| > a) = 0.01$, c'est à dire

$$\mathbb{P}\left(\sqrt{n}\Big|\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}\Big| > a\right) \to 0.01.$$

D'où

$$\mathbb{P}\Big(|\bar{X}_n - p| > \frac{a\sqrt{p(1-p)}}{n}\Big) \to 0.01$$

On peut donc poser $b(n,p) = \frac{a\sqrt{p(1-p)}}{n}$.

7. Soit $p \in [0,1]$, comme $p(1-p) \leq 0.25$, on a

$$|\bar{X}_n - p| > \frac{a\sqrt{p(1-p)}}{n} \Leftarrow |\bar{X}_n - p| > \frac{a\sqrt{0.25}}{\sqrt{n}},$$

donc

$$\mathbb{P}\Big(|\bar{X}_n - p| > \frac{a\sqrt{0.25}}{\sqrt{n}}\Big) \leq \mathbb{P}\Big(|\bar{X}_n - p| > \frac{a\sqrt{p(1-p)}}{n}\Big).$$

On peut alors en déduire

$\limsup \mathbb{P}\Big(|\bar{X}_n - p| > \frac{a\sqrt{0.25}}{\sqrt{n}}\Big) \leq 0.01$

On a donc $c \approx \frac{2.58\sqrt{0.25}}{\sqrt{1000}} \approx 0.041$.

On peut donc en conclure, qu'avec un niveau de risque de 1%, le résultat du sondage (i.e. la proportion des 1000 personnes disant voter pour A) correspondra, avec une marge d'erreur de $\pm 4\%$, à la proportion, **dans la population complète** du nombre de personnes souhaitant voter pour A.

Le *niveau* de risque de 1% signifie ici, que dans 1% des expériences (i.e. pour 1% des sondages avec 1000 personnes), la marge d'erreur sera en réalité supérieure à 4%. Cette ordre de grandeur est une bonne indication des marges d'erreur dans les sondages réels. Attention, il faut toujours garder en tête la difficulté de réellement "tirer au hasard uniformément dans la population", sachant qu'il faut absolument que la personne tirée réponde pour ne pas biaiser le sondage.

Correction exercice 6. Avec la loi exponentielle

Un fabriquant d'ampoules explique que la durée de vie de ses ampoules suit une loi exponentielle.

On a fait l'expérience en mesurant la durée d'utilisation de $n = 20$ ampoules, qu'on a modélisées par des réalisations $y_1, \cdots, y_n$ de v.a. $Y_1, \cdots, Y_n$ de loi exponentielle i.i.d. de paramètre λ (inconnu).

1. On peut refaire le calcul :

$$\begin{aligned}\mathbb{E}[Y_1] &= \int_0^{+\infty} x\lambda e^{-\lambda x}\mathrm{d}x \overset{I.P.P.}{=} 0 - \int_0^{+\infty} \lambda\frac{e^{-\lambda x}}{-\lambda}\mathrm{d}x \\ &= \int_0^{+\infty} e^{-\lambda x}\mathrm{d}x \\ &= \Big[\frac{e^{-\lambda x}}{-\lambda}\Big]_0^{+\infty} \\ &= \frac{1}{\lambda}.\end{aligned}$$

La durée moyenne d'une ampoule est donc de $\frac{1}{\lambda}$.

2. On peut calculer

$$\mathbb{P}(Y_1 \geq s) = \int_s^{+\infty} \lambda e^{-\lambda x} \mathrm{d}x = \Big[\lambda \frac{e^{-\lambda x}}{-\lambda}\Big]_s^{+\infty} = e^{-\lambda s}.$$

On en déduit

$$\mathbb{P}(Y_1 \geq \frac{2}{\lambda}) = e^{-\lambda \frac{2}{\lambda}} = e^{-2} \approx 0.13$$

3. On cherche q_p tel que $\mathbb{P}(Y_1 \leq q_p) = p$, i.e. tel que $1 - e^{-\lambda q_p} = p$. On trouve alors

$$q_p = -\frac{\log(1-p)}{\lambda}.$$

On en déduit la médiane

$$q_{\frac{1}{2}} = -\frac{\log(1/2)}{\lambda} = \frac{\log(2)}{\lambda} \approx \frac{0.69}{\lambda}.$$

4. On veut calculer

$$\begin{aligned}
\mathbb{P}\Big(Y_1 \geq s+t \mid Y_1 \geq t\Big) &= \frac{\mathbb{P}\Big(Y_1 \geq s+t \cap Y_1 \geq t\Big)}{\mathbb{P}\Big(\ \mid Y_1 \geq t\Big)} \\
&= \frac{\mathbb{P}\Big(Y_1 \geq s+t\Big)}{\mathbb{P}\Big(\ \mid Y_1 \geq t\Big)} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\
&= e^{-\lambda s} \\
&= \mathbb{P}\Big(Y_1 \geq s\Big)
\end{aligned}$$

On peut interpréter ce résultat comme "l'absence de mémoire" de la loi exponentielle. Si un modèle comme cela était réellement adapté aux ampoules, cela signifie qu'il n'y a pas d'usure, puisque une ampoule qui a déjà vécu un certain temps a la même probabilité de cesser de fonctionner qu'une ampoule neuve.

5. Si le fabriquant disait la vérité sur la durée moyenne, on aurait $\lambda = \frac{1}{5000}$. Et si la loi de la durée de vie des ampoules était bien exponentielle, alors en notant $B_i = \{Y_i > 3465\}$, on aurait

$$B_i \sim \mathcal{B}\Big(\mathbb{P}(Y_i > 3465)\Big), \text{ avec } \mathbb{P}(Y_i > 3465) = e^{-\frac{3465}{5000}} = \frac{1}{2}.$$

En utilisant l'indépendance des B_i, le nombre d'ampoules N qui ont duré plus de 3465 serait donc de loi Binomiale de paramètres 20 et $\frac{1}{2}$. Donc

$$\begin{aligned}
\mathbb{P}(N \leq 4) &= \mathbb{P}(N=0) + \mathbb{P}(N=1) + \mathbb{P}(N=2) + \mathbb{P}(N=3) + \mathbb{P}(N=4) \\
&= \frac{1}{2^{20}}\left(1 + 20 + \binom{20}{2} + \binom{20}{3} + \binom{20}{4}\right) \\
&\approx 0.0059
\end{aligned}$$

Si le fabriquant disait la vérité (et que les ampoules suivaient bien une loi exponentielle), alors il y aurait moins de 6 chances sur 1000 que moins de 4 parmi les 20 durent plus de 3465 heures. De là à affirmer qu'il ment, il n'y a qu'un pas, et ce pas, c'est le formalisme d'un test statistique (que l'on verra plus tard en cours) :

$$\text{Modèle : } B_i \sim_{i.i.d.} \mathcal{B}(e^{-3465\lambda})$$
$$H_0 : \lambda = \frac{1}{5000}$$
$$H_1 : \lambda \neq \frac{1}{5000}$$
$$\text{l'expérience donne } \sum_{i=1}^{20} b_i = 4$$
$$\text{or sous } H_0 : \mathbb{P}(\sum_{i=1}^{20} B_i \leq 4) = 0.0059 < 0.05,$$
$$\text{Donc on rejette le test au niveau } \alpha = 0.05$$

Correction exercice 7. Simple Intervalle de Confiance

On cherche à contrôler la température dans un processus industriel. On prend simultanément 25 mesures de température, que l'on modélise comme des lois normales i.i.d. d'espérance m inconnue, et d'écart type $\sigma = 0.2°C$ connu. On déclenchera la régulation dès que la température de $18°C$ ne tombe pas dans l'intervalle de confiance.

1. Pour constuire l'intervalle de confiance, on cherche une expression qui dépend de m et des observations, donc on connaît la loi. Ici, on peut utiliser la stabilité des lois normales, qui après calcul de l'espérance et de la variance, nous donne

$$\bar{X} \sim \mathcal{N}\Big(m, \frac{\sigma^2}{25}\Big).$$

En réduisant cette variable, on a

$$Z := 5\frac{\bar{X} - m}{\sigma} \sim \mathcal{N}(0, 1),$$

donc, d'après la table de quantiles,

$$\mathbb{P}(|Z| \leq 2.58) = 0.99$$

Or

$$\begin{aligned} |Z| < 2.58 &\Leftrightarrow \Big|5\frac{\bar{X} - m}{\sigma}\Big| \leq 2.58 \\ &\Leftrightarrow \Big|\bar{X} - m\Big| \leq \frac{2.58}{5}\sigma \\ &\Leftrightarrow \bar{X} - \frac{2.58}{5}\sigma \leq m \leq \bar{X} + \frac{2.58}{5}\sigma \end{aligned}$$

On a donc

$$\mathbb{P}\left(m \in \Big[\bar{X} - \frac{2.58}{5}\sigma, \bar{X} + \frac{2.58}{5}\sigma\Big]\right) = \mathbb{P}\left(m \in \Big[\bar{X} \pm \frac{2.58}{5}\sigma\Big]\right) = 0.99,$$

ce qui correspond à la définition d'un intervalle de confiance. On a donc

$$IC_{99\%}(m) = [\bar{X} \pm \frac{2.58}{5}\sigma] = [\bar{X} \pm 0.1032].$$

2. L'intervalle de confiance ponctuel donne donc :

$$ic_{99\%}(m) = [\bar{x} \pm \frac{2.58}{5}\sigma] = [18.1 \pm 0.1032] = [17.997, 18.203].$$

On ne déclenche donc pas la régulation.

3. Comme l'intervalle de confiance est par essence aléatoire, si les mesures sont indépendantes, en moyenne, seul 99% des intervalles de confiances contiendront la "vraie" valeur. Donc même si la température réelle était de 18 degrés, en moyenne toutes les 100 mesures indépendantes, on déclencherait à tord la régulation (ce qui a priori dans ce sens là ne serait pas grave).