Eléments de corrections pour les exercices de Travaux Dirigés.

Feuille 1 - Exercice 1.

D'après le cours, par propriétés de l'espérance (et en particulier sans condition sur l'indépendance de X et de Y), on a

$$\mathbb{E}(5X - 4Y) = 5\,\mathbb{E}(X) - 4\,\mathbb{E}(Y) = 5 \times 2 - 4 \times 2 = 2.$$

Et par propriétés de la variance, puisque X et Y sont indépendantes, on a

$$\mathbb{V}(5X - 4Y) = 5^2\,\mathbb{V}(X) + (-4)^2\,\mathbb{E}(Y) = 25 \times 1 + 16 \times 1 = 41.$$

Feuille 1 - Exercice 2.

On rappelle que les tables de valeurs numériques concernant les lois de Poisson, binomiale et normale donnent les valeurs de la fonction de répartition, c'est-à-dire les valeurs de $\mathbb{P}(X \leqslant t)$ en fonction de t.

1. On regarde le tableau qui concerne la loi de Poisson, à la colonne $\lambda = 8$. La ligne $k = 8$ donne

   $$\mathbb{P}(X \leqslant 8) = 0{,}5925.$$

   Pour obtenir les autres valeurs, il faut effectuer des manipulations simples :

   $$\begin{aligned}
   \mathbb{P}(X = 8) &= \mathbb{P}(X \leqslant 8 \text{ mais pas } X \leqslant 7) \\
   &= \mathbb{P}(X \leqslant 8) - \mathbb{P}(X \leqslant 7) \text{ car } [X \leqslant 7] \subset [X \leqslant 8] \\
   &= 0{,}5925 - 0{,}4530 \\
   &= 0{,}1395. \\
   \mathbb{P}(X > 8) &= 1 - \mathbb{P}(X \leqslant 8) \\
   &= 1 - 0{,}5925 \\
   &= 0{,}4075. \\
   \mathbb{P}(X > 9) &= 1 - \mathbb{P}(X \leqslant 9) \\
   &= 1 - 0{,}7166 \\
   &= 0{,}2834.
   \end{aligned}$$

   On parcourt le tableau pour trouver le plus petit entier u tel que $\mathbb{P}(X \leqslant u) \geqslant 0{,}85$. On trouve $u = 11$. Pour v, on utilise que $\mathbb{P}(X > v) = 1 - \mathbb{P}(X \leqslant v)$ donc

   $$\mathbb{P}(X > v) \leqslant 10\% \Leftrightarrow \mathbb{P}(X \leqslant v) \geqslant 90\%.$$

   Ce qu'on cherche est donc le plus petit v tel que $\mathbb{P}(X \leqslant v) \geqslant 0{,}9$ et le tableau nous donne $v = 12$. Enfin, pour w on utilise que $\mathbb{P}(X \geqslant w) = 1 - \mathbb{P}(X < w)$ et que $\mathbb{P}(X < w) = \mathbb{P}(X \leqslant w - 1)$ si w est un entier. On a donc

   $$\mathbb{P}(X \geqslant w) \geqslant 70\% \Leftrightarrow \mathbb{P}(X \leqslant w - 1) \leqslant 30\%.$$

   Ce qu'on cherche est donc le plus grand w tel que $\mathbb{P}(X \leqslant w-1) \leqslant 0{,}3$ et on trouve $w-1 = 5$ donc $w = 6$.

2. On regarde le tableau qui donne la fonction de répartition de la loi binomiale de paramètres 20 et $0{,}5$: il se trouve à la première page du polycopié de tables, et on s'intéresse à la colonne $n = 20$ (attention, on trouve

plus loin un tableau concernant $n = 20$ pour différentes valeurs de p, mais $p = 0{,}5$ n'y figure pas).

$$\begin{aligned}
\mathbb{P}(2 < X \leqslant 6) &= \mathbb{P}(X \leqslant 6 \text{ mais pas } X \leqslant 2) \\
&= \mathbb{P}(X \leqslant 6) - \mathbb{P}(X \leqslant 2) \text{ car } [X \leqslant 2] \subset [X \leqslant 6] \\
&= 0{,}0577 - 0{,}0002 \\
&= 0{,}0575. \\
\mathbb{P}(5 \leqslant X \leqslant 15) &= \mathbb{P}(X \leqslant 15 \text{ mais pas } X \leqslant 4) \\
&= \mathbb{P}(X \leqslant 15) - \mathbb{P}(X \leqslant 4) \text{ car } [X \leqslant 4] \subset [X \leqslant 15] \\
&= 0{,}9941 - 0{,}0059 \\
&= 0{,}9882. \\
\mathbb{P}(X = 13) &= \mathbb{P}(X \leqslant 13 \text{ mais pas } X \leqslant 12) \\
&= \mathbb{P}(X \leqslant 13) - \mathbb{P}(X \leqslant 12) \text{ car } [X \leqslant 12] \subset [X \leqslant 13] \\
&= 0{,}9423 - 0{,}8684 \\
&= 0{,}0739.
\end{aligned}$$

Enfin, $\mathbb{P}(X \geqslant v) = 1 - \mathbb{P}(X < v)$ et on a $\mathbb{P}(X < v) = \mathbb{P}(X \leqslant v - 1)$ pour v entier, donc

$$\mathbb{P}(X \geqslant v) \geqslant 90\% \Leftrightarrow \mathbb{P}(X \leqslant v - 1) \leqslant 10\%.$$

On cherche donc le plus grand entier v tel que et $\mathbb{P}(X \leqslant v - 1) \leqslant 0{,}1$. Le tableau nous donne $v - 1 = 6$, donc $v = 7$.

3. On regarde le tableau pour la loi normale centrée réduite. On remarque que ce tableau ne nous donne les valeurs de $\mathbb{P}(X \geqslant t)$ que pour $t \geqslant 0$. On trouve directement (ligne $1{,}6$ colonne $0{,}03$) que $\mathbb{P}(X \leqslant 1{,}63) = 0{,}9484$. On utilise ensuite

$$\mathbb{P}(X \geqslant 0{,}53) = 1 - \mathbb{P}(X < 0{,}53) = 1 - \mathbb{P}(X \leqslant 0{,}53)$$

où la deuxième égalité est vraie parce que X est une variable à densité. On trouve dans le tableau (ligne $0{,}5$ colonne $0{,}03$) que $\mathbb{P}(X \leqslant 0{,}53) = 0{,}7019$ et donc $\mathbb{P}(X \geqslant 0{,}53) = 0{,}2981$. On utilise ensuite la symétrie $\mathbb{P}(X \leqslant a) = \mathbb{P}(X \geqslant -a)$ pour écrire

$$\mathbb{P}(X \leqslant -1{,}14) = \mathbb{P}(X \geqslant 1{,}14) = 1 - \mathbb{P}(X < 1{,}14) = 1 - \mathbb{P}(X \leqslant 1{,}14)$$

où, encore une fois, la deuxième égalité est vraie parce que X est une variable à densité. On trouve dans le tableau (ligne $1{,}1$ colonne $0{,}04$) que $\mathbb{P}(X \leqslant 1{,}14) = 0{,}8729$ et donc $\mathbb{P}(X \leqslant -1{,}14) = 0{,}1271$. On a

$$\begin{aligned}
\mathbb{P}(|X| \geqslant 1{,}27) &= \mathbb{P}(X \leqslant -1{,}27 \text{ ou } X \geqslant 1{,}27) \\
&= \mathbb{P}(X \leqslant -1{,}27) + \mathbb{P}(X \geqslant 1{,}27) \\
&= 2\,\mathbb{P}(X \geqslant 1{,}27) \\
&= 2\big(1 - \mathbb{P}(X < 1{,}27)\big) \\
&= 2\big(1 - \mathbb{P}(X \leqslant 1{,}27)\big)
\end{aligned}$$

et le tableau nous donne $\mathbb{P}(X \leqslant 1{,}27) = 0{,}8980$ donc $\mathbb{P}(|X| \geqslant 1{,}27) = 0{,}2040$. Enfin,

$$\begin{aligned}
\mathbb{P}(-u \leqslant X \leqslant +u) &= \mathbb{P}(X \leqslant u) - \mathbb{P}(X < -u) \\
&= \mathbb{P}(X \leqslant u) - \mathbb{P}(X < -u) \\
&= \mathbb{P}(X \leqslant u) - \mathbb{P}(X > u) \\
&= \mathbb{P}(X \leqslant u) - \big(1 - \mathbb{P}(X \leqslant u)\big) \\
&= 2\,\mathbb{P}(X \leqslant u) - 1.
\end{aligned}$$

On cherche donc u tel que $\mathbb{P}(X \leqslant u) = 0{,}935$. Le tableau nous montre que u sera compris entre $1{,}51$ et $1{,}52$.

Feuille 1 - Exercice 3.

Les calculs sont ceux de l'exercice suivant.

Feuille 1 - Exercice 4.

1. Les unités de m et σ sont en $\mu g/m^3$, ces paramètres représentent respectivement la concentration réelle d'ozone dans l'air et l'imprécision de la mesure de cette concentration.

2. (a) On cherche à calculer $\mathbb{P}(X \geqslant 180)$ lorsque $X \sim \mathcal{N}(178\ ;\ 3{,}1)$. En centrant et renormalisant X, on obtient :

$$\mathbb{P}(X \geqslant 180) = \mathbb{P}\left(\frac{X-178}{\sqrt{3{,}1}} \geqslant \frac{180-178}{\sqrt{3{,}1}}\right) = \mathbb{P}(U \geqslant 1{,}14) = 1 - F(1{,}14)$$

où $U = \dfrac{X-178}{\sqrt{3{,}1}}$ et où F est la fonction de répartition de la loi normale centrée réduite.
On lit sur la table $F(1{,}14) = 0{,}8729$ et on obtient donc $\mathbb{P}(X \geqslant 180) = 12{,}71\%$.

(b) Soit Z la variable aléatoire représentant la valeur moyenne de trois mesures supposées indépendantes. Alors $Z = \frac{1}{3}\sum_{i=1}^{3} X_i$ où X_i représente la i-ème mesure. Comme les variables X_i sont indépendantes entre elles, Z suit la loi $\mathcal{N}(m\ ;\ \dfrac{3{,}1}{3})$.
On cherche la valeur de $\mathbb{P}(Z \geqslant 180)$. En centrant et en réduisant Z, on obtient :

$$\mathbb{P}(Z \geqslant 180) = \mathbb{P}\left(V \geqslant \frac{\sqrt{3}(180-178)}{\sqrt{3{,}1}}\right) = 1 - F(\sqrt{\frac{12}{3{,}1}})$$

où $V = \dfrac{\sqrt{3}(X-178)}{\sqrt{3{,}1}}$ et où F est la fonction de répartition de la loi normale centrée réduite.
En lisant la table on obtient : $\mathbb{P}(Z \geqslant 180) = 0.0244$. La probabilité que la moyenne de trois mesures dépasse $180\mu g/m^3$ est de $2{,}44\%$.

(c) Soit Z_n la variable aléatoire représentant la valeur moyenne de n mesures supposées indépendantes. Alors $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ où X_i représente la i-ème mesure. Comme les variables X_i sont indépendantes entre elles, Z_n suit la loi $\mathcal{N}(m\ ;\ \dfrac{3{,}1}{n})$.
On cherche n tel que $\mathbb{P}(Z_n \geqslant 180) \leqslant 0{,}01$. En centrant et en réduisant Z_n, on obtient :

$$\mathbb{P}(Z_n \geqslant 180) = \mathbb{P}\left(W \geqslant \frac{\sqrt{n}(180-178)}{\sqrt{3{,}1}}\right) = 1 - F(2\sqrt{\frac{n}{3{,}1}})$$

où $W = \dfrac{\sqrt{n}(X-178)}{\sqrt{3{,}1}}$ et où F est la fonction de répartition de la loi normale centrée réduite.
L'inéquation $\mathbb{P}(Z_n \geqslant 180) \leqslant 0{,}01$ équivaut donc à l'inéquation $1 - F(2\sqrt{\frac{n}{3{,}1}}) \leqslant 0{,}01$, et

$$1 - F(2\sqrt{\frac{n}{3{,}1}}) \leqslant 0{,}01 \Leftrightarrow F(2\sqrt{\frac{n}{3{,}1}}) \geqslant 0{,}99.$$

On lit sur la table $F(2{,}32) = 98{,}98\%$ et $F(2{,}33) = 99{,}01\%$. On prend donc n tel que $2\sqrt{\frac{n}{3{,}1}} \geqslant 2{,}33$ (la fonction de répartition est croissante), c'est à dire $n \geqslant (2{,}33/2)^2 \times 3{,}1 \simeq 4{,}21$ On obtient finalement $n = 5$ (plus petit entier supérieur à 4,21).

Feuille 1 - Exercice 5.

1. La probabilité qu'un champignon soit comestible est 1/5 (ou 0,2), la probabilité qu'un champignon soit non comestible est 4/5 (ou 0,8). Le nombre de champignons non comestibles sur un panier de 20 champignons, noté X, est donc une variable aléatoire de loi $\mathcal{B}(20\ ;\ 0{,}8)$. On cherche à calculer $\mathbb{P}(X \leqslant 13)$, et cette probabilité ne se trouve pas dans la table de valeurs numériques mise à notre disposition.

   Notons Y le nombre de champignons comestibles sur un panier de 20 champignons. Cette variable suit la loi $\mathcal{B}(20\ ;\ 0{,}2)$ et on a :

   $$\mathbb{P}(X \leqslant 13) = \mathbb{P}(Y \geqslant 7) = 1 - \mathbb{P}(Y < 7) = 1 - \mathbb{P}(Y \leqslant 6) = 1 - F(6)$$

   où F est la fonction de répartition de la loi $\mathcal{B}(20\ ;\ 0{,}2)$. On lit sur la table de valeurs numériques $F(6) = 0{,}9133$, on conclut : $\mathbb{P}(X \leqslant 13) = 1 - 0{,}9133 = 8{,}67\%$.

2. Le nombre de champignons comestibles sur un panier de 100 champignons, noté X, est une variable aléatoire de loi $\mathcal{B}(100\ ;\ 0{,}2)$. On cherche à calculer $\mathbb{P}(15 \leqslant X \leqslant 20)$, et cette probabilité ne se trouve pas dans la table de valeurs numériques mise à notre disposition.

   Les conditions $100 \geqslant 50$, $100 * 0{,}2 = 20 \geqslant 15$, $100 * 0{,}8 = 80 \geqslant 15$ étant réunies, on approxime la loi de X par la loi normale $\mathcal{N}(20\ ;\ 16)$.

   Soit on calcule $\mathbb{P}(15 \leqslant X \leqslant 20)$, soit on applique une correction de continuité et on calcule $\mathbb{P}(14{,}5 \leqslant X \leqslant 20{,}5)$.

   En centrant et en réduisant, on obtient :

   $$\mathbb{P}(15 \leqslant X \leqslant 20) = \mathbb{P}\left(\frac{15-20}{\sqrt{16}} \leqslant \frac{X-20}{\sqrt{16}} \leqslant \frac{20-20}{\sqrt{16}}\right) \simeq \mathbb{P}(-1{,}25 \leqslant W \leqslant 0) = F(0) - F(-1{,}25)$$

   où $W = \dfrac{X-20}{\sqrt{16}}$ et où F est la fonction de répartition de la loi normale centrée réduite.

   Avec $F(0) = 0{,}5$ et $F(-1{,}25) = 1 - F(1{,}25) = 1 - 0{,}8944$, on obtient $\mathbb{P}(15 \leqslant X \leqslant 20) \simeq 39{,}44\%$.

   Le calcul avec correction de continuité donne :

   $$\mathbb{P}(15 \leqslant X \leqslant 20) \simeq \mathbb{P}(-1{,}375 \leqslant W \leqslant 0{,}125) \simeq F(0{,}13) - F(-1{,}38) = 46{,}79\%.$$

   Le calcul exact avec WIMS donne $\mathbb{P}(15 \leqslant X \leqslant 20) = 47{,}90\%$, la correction de continuité améliore donc beaucoup la qualité de l'approximation.

3. Parmi les champignons comestibles, un dixième sont des cèpes, la probabilité qu'un champignon soit un cèpe est donc 1/50 (ou 0,02). Notons X la variable aléatoire égale au nombre de cèpes sur un panier de 150 champignons. Cette variable suit la loi $\mathcal{B}(150\ ;\ 1/50)$ dont les valeurs ne sont pas données dans la table de valeurs numériques mise à notre disposition. Les conditions $150 \geqslant 50$, $150 * 1/50 = 3 < 15$, $1/50 < 0{,}1$ étant réunies, on approxime la loi de X par la loi de Poisson $\mathcal{P}(3)$. On cherche à calculer $\mathbb{P}(X \geqslant 5)$:

   $$\mathbb{P}(X \geqslant 5) = 1 - \mathbb{P}(X < 5) = 1 - \mathbb{P}(X \leqslant 4) \simeq 1 - F(4)$$

   où F est la fonction de répartition de la loi $\mathcal{P}(3)$. On lit sur la table de valeurs numériques $F(4) = 0{,}8153$, on conclut : $\mathbb{P}(X \leqslant 13) = 1 - 0{,}8153 = 18{,}47\%$. Le calcul exact par ordinateur donne $18{,}30\%$.

4. Notons X la variable aléatoire égale au nombre de champignons comestibles sur un panier de n champignons. La probabilité pour un champignon d'être comestible est 0,2 donc X suit la loi $\mathcal{B}(n\ ;\ 0{,}2)$, que l'on peut approximer, si $n \times 0{,}2 \geqslant 30$, c'est à dire si $n \geqslant 150$, par la loi normale $\mathcal{N}(n \times 0{,}2\ ;\ n \times 0{,}2 \times 0{,}8)$.

   On cherche n tel que $\mathbb{P}(X \geqslant 40) \geqslant 90\%$ (ou $\mathbb{P}(X \geqslant 39{,}5) \geqslant 90\%$ si on applique une correction de continuité). En centrant et en réduisant, on obtient :

$$\begin{aligned}
\mathbb{P}(X \geqslant 40) \geqslant 90\% &\iff \mathbb{P}\left(\frac{X - 0{,}2 \times n}{\sqrt{0{,}16 \times n}} \geqslant \frac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \geqslant 90\% \\
&\iff \mathbb{P}\left(W \geqslant \frac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \geqslant 90\% \\
&\iff 1 - \mathbb{P}\left(W \leqslant \frac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \geqslant 90\% \\
&\iff \mathbb{P}\left(W \leqslant \frac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \leqslant 10\%.
\end{aligned}$$

   En approximant la loi de W par la loi $\mathcal{N}(0\ ;\ 1)$ et en notant F la fonction de répartition de cette loi, on obtient :

$$F\left(\frac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \leqslant 10\%.$$

   Le réel $\dfrac{40 - 0{,}2 \times n}{\sqrt{0{,}16 \times n}}$ est négatif, on utilise alors $F(t) = 1 - F(-t)$ et on obtient :

$$1 - F\left(\frac{-40 + 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \leqslant 10\%$$

   c'est à dire

$$F\left(\frac{-40 + 0{,}2 \times n}{\sqrt{0{,}16 \times n}}\right) \geqslant 90\%.$$

   Sur la table de valeurs numériques, on lit que l'on doit avoir

$$\frac{-40 + 0{,}2 \times n}{\sqrt{0{,}16 \times n}} \geqslant 1{,}29.$$

   En posant $u = \sqrt{n}$, on est ramené à étudier l'inégalité $0{,}2\ u^2 - 1{,}29 \times 0{,}4\ u - 40 \geqslant 0$. Les deux racines de ce polynôme sont -12,9108 et 15,4908, le polynôme est positif ou nul pour $u \geqslant 15{,}4908$ soit $n \geqslant 239{,}96$ c'est à dire $n \geqslant 240$.

Feuille 1 - Exercice 6.

La loi de S est la loi binômiale $\mathcal{B}(100\ ;\ 0{,}75)$, d'espérance $100 \times 0{,}75 = 75$ et de variance $100 \times 0{,}75 \times 0{,}25 = 18{,}75$.

Les conditions $100 \geqslant 50$; $100 \times 0{,}75 \geqslant 15$; $100 \times 0{,}25 \geqslant 15$ étant satisfaites, on peut approximer la loi S par la loi normale $\mathcal{N}(75\ ;\ 18{,}75)$.

La probabilité d'observer au moins 80 mâles est $\mathbb{P}(S \geqslant 80)$, ou bien, si on a l'intention d'utiliser la correction de continuité, $\mathbb{P}(S \geqslant 79{,}5)$. On a :

$$\begin{aligned} \mathbb{P}(S \geqslant 79{,}5) &= \mathbb{P}\left(\frac{S-75}{\sqrt{18{,}75}} \geqslant \frac{79{,}5-75}{\sqrt{18{,}75}}\right) \\ &\simeq \mathbb{P}(W \geqslant 1{,}04) = 1 - F(1{,}04) \end{aligned}$$

où $W = \dfrac{S-75}{\sqrt{18{,}75}}$ et où F est la fonction de répartition de la loi normale centrée réduite. On lit sur la table de la loi normale centrée réduite $F(1{,}04) = 0{,}8508$ et on conclut que la probabilité d'observer au moins 80 mâles est environ 14,92%.

Notons que le calcul exact (sans approximation normale) effectué par un logiciel donne 14,88% et que le calcul sans correction de continuité donne 12,51%. La correction de continuité améliore donc beaucoup l'approximation.

On veut ensuite calculer $\mathbb{P}(68 \leqslant S \leqslant 82)$, que l'on écrit $\mathbb{P}(67{,}5 \leqslant S \leqslant 82{,}5)$ si on a l'intention d'utiliser la correction de continuité. On a :

$$\begin{aligned} \mathbb{P}(67{,}5 \leqslant S \leqslant 82{,}5) &= \mathbb{P}\left(\frac{67{,}5-75}{\sqrt{18{,}75}} \leqslant \frac{S-75}{\sqrt{18{,}75}} \leqslant \frac{82{,}5-75}{\sqrt{18{,}75}}\right) \\ &\simeq \mathbb{P}(-1{,}73 \leqslant W \leqslant 1{,}73) = F(1{,}73) - F(-1{,}73) \end{aligned}$$

où $W = \dfrac{S-75}{\sqrt{18{,}75}}$ et où F est la fonction de répartition de la loi normale centrée réduite. On utilise $F(-t) = 1 - F(t)$ et on obtient :

$$\mathbb{P}(67{,}5 \leqslant S \leqslant 82{,}5) \simeq 2\ F(1{,}73) - 1 = 2 \times 0{,}9582 - 1 = 91{,}64\%$$

d'après la table de la loi normale centrée réduite.

Notons que le calcul exact (sans approximation normale) effectué par un logiciel donne

$$\mathbb{P}(68 \leqslant S \leqslant 82) = F_{\mathcal{B}(100\ ;\ 0{,}75)}(82) - F_{\mathcal{B}(100\ ;\ 0{,}75)}(67) = 0{,}9624 - 0{,}0446 = 91{,}78\%$$

et que le calcul sans correction de continuité donne :

$$\mathbb{P}(68 \leqslant S \leqslant 82) \simeq F(1{,}62) - F(-1{,}62) = 2 \times 0{,}9474 - 1 = 89{,}48\%.$$

La correction de continuité améliore donc beaucoup l'approximation.

Feuille 2 - Exercice 1. *Même exercice que l'exercice 2.*

Feuille 2 - Exercice 2.

Première partie *La correction insiste sur les 7 points de la construction d'un test.*

1. Préciser le modèle.

   On considère un individu A dont le vrai taux d'hématocrite dans le sang est τ. On modélise le résultat d'une mesure de ce taux par une variable aléatoire X de loi $\mathcal{N}(\tau; 4)$: l'erreur de mesure provient de l'accumulation de nombreux aléas indépendants et est donc modélisée par une loi normale, d'espérance égale au vrai taux et de variance $\sigma^2 = 4 = 2^2$ car les instruments sont précis à deux unités de mesure près. On se pose la question : "l'individu A a-t-il triché ?" qui se traduit avec les notations adoptées : *τ est-il égal à 45% (pas de triche) ou supérieur à 45% (triche) ?*

2. Déterminer les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test.

   On définit $\mathcal{H}_0 : \tau = 45$ et $\mathcal{H}_1 : \tau > 45$.

3. Déterminer la statistique de test.

   Pour faire un test sur τ, on choisit la variable aléatoire X (taux mesuré). Sa loi est connue sous $\mathcal{H}_0 : X \sim \mathcal{N}(45; 4)$ et sa loi sous $\mathcal{H}_1$ est différente de sa loi sous $\mathcal{H}_0$: sous $\mathcal{H}_1 : X \sim \mathcal{N}(\tau > 45; 4)$.

   Notons que le choix de
   $$U_0 := \frac{X - 45}{2}$$
   comme statistique de test est également possible. La loi de U_0 est connue sous $\mathcal{H}_0$, c'est la loi $\mathcal{N}(0\ ; 1)$ et la loi de U_0 sous $\mathcal{H}_1$ est la loi $\mathcal{N}(\delta; 1)$, avec $\delta = \frac{(\tau - 45)}{2}$ ce qui implique que la densité de U_0 sous $\mathcal{H}_1$ est à droite de la densité de U_0 sous $\mathcal{H}_0$.

4. Établir la région de rejet de $\mathcal{H}_0$.

   On remarque que la loi de X sous $\mathcal{H}_1$ est à droite de la loi de X sous $\mathcal{H}_0$ (*faire un dessin pour s'en convaincre*). On rejettera donc $\mathcal{H}_0$ si on observe une valeur trop supérieure à 45%. Autrement dit, le domaine de rejet de $\mathcal{H}_0$ est de la forme $\mathcal{R} = \{X \in [45 + a, +\infty[\}$. Autrement dit, on rejettera H_0 si $X_{obs} \geqslant 45 + a$.

   Notons que le choix de U_0 comme statistique de test nous conduirait à prendre comme domaine de rejet de $\mathcal{H}_0$ le domaine $\mathcal{R} = \{U_0 \in [b, +\infty[\}$ car la statistique U_0 est une fonction croissante de X, ce que l'on pourrait noter $U_0 = f(X)$ où f est une fonction croissante (c'est une droite de coefficient directeur positif) et donc $X \geqslant 45 + a$ équivaut à $U_0 \geqslant b = f(45 + a)$.

5. Calculer le seuil de la région de rejet grâce à la loi de X sous $\mathcal{H}_0$.

   On calcule le seuil a grâce à l'équation :
   $$\mathbb{P}_{\mathcal{H}_0}(X \in [45 + a, +\infty[) = \alpha$$
   qui équivaut à l'équation :
   $$\mathbb{P}_{\mathcal{H}_0}(X \in]-\infty; 45 + a[) = 1 - \alpha$$
   On adopte $\alpha = 1\%$ (faible niveau car on ne veut pas accuser quelqu'un de triche à tort). Sous $\mathcal{H}_0 : X \sim \mathcal{N}(45{,}4)$. On centre et réduit :
   $$\begin{aligned} \mathbb{P}_{\mathcal{H}_0}(X < 45 + a) &= \mathbb{P}_{\mathcal{H}_0}(X \leqslant 45 + a) \\ &= \mathbb{P}_{\mathcal{H}_0}\left(\frac{X - 45}{2} \leqslant \frac{45 + a - 45}{2}\right) \\ &= \mathbb{P}(Z \leqslant a/2) = F(a/2) \end{aligned}$$

où Z suit la loi $\mathcal{N}(0,1)$ et où F est sa fonction de répartition. Après lecture dans les tables, $F(a/2) = 1 - 0.01 = 0.99$ donne $a/2 = 2.33$, c'est-à-dire $a = 4.66$.

Notons que le choix de U_0 comme statistique de test nous conduirait à trouver $b = 2.33$ puisque l'équation de niveau s'écrit :

$$\mathbb{P}_{\mathcal{H}_0}(U_0 \in [b, +\infty[) = \alpha$$

qui équivaut à :

$$\mathbb{P}_{\mathcal{H}_0}(U_0 \in]-\infty, b]) = 1 - \alpha$$

c'est à dire à $F(b) = 1 - 0.01 = 0.99$ qui donne immédiatement $b = 2.325$. Le calcul du seuil est donc plus simple avec la statistique U_0.

6. On calcule la valeur observée et on conclut.

Première méthode pour conclure : on compare la valeur de la statistique au seuil de la région de rejet.

La règle de décision est donc

- si $x_{obs} \geqslant 49.66$, on décide de rejeter $\mathcal{H}_0$,
- sinon, on décide de conserver $\mathcal{H}_0$.

Et si on a choisi U_0 comme statistique de test, la règle de décision est :

- si $U_0^{obs} \geqslant 2.33$, on décide de rejeter $\mathcal{H}_0$,
- sinon, on décide de conserver $\mathcal{H}_0$.

Les valeurs observées pour C.L. et B.J. sont respectivement $X_1^{obs} = 49$ et $X_2^{obs} = 50$. Ces valeurs conduisent à conserver $\mathcal{H}_0$ pour C.L. et à la rejeter pour B.J.

Et si on a choisi U_0 comme statistique de test, on trouve pour C.L. et B.J. respectivement les valeurs $U_0^{obs} = 2$ et $U_0^{obs} = 2.5$, ce qui nous conduit également à conserver $\mathcal{H}_0$ pour C.L. et à la rejeter pour B.J.

Seconde méthode pour conclure : on calcule la p-valeur (*faire un dessin*).

L'observation pour C.L. est 49. Pour rejeter $\mathcal{H}_0$ avec cette observation, le domaine de rejet de $\mathcal{H}_0$ doit être de la forme $\mathcal{R} = \{X \in [49, +\infty[\}$: c'est le domaine de rejet de $\mathcal{H}_0$ du test qui rejette $\mathcal{H}_0$ si $X_{obs} \geqslant 49$, c'est le plus petit domaine permettant de rejeter $\mathcal{H}_0$ avec l'observation $X^{obs} = 49$ (*faire un dessin*). Le niveau du test associé à cette zone de rejet de $\mathcal{H}_0$ est :

$$p_{value} = 1 - \mathbb{P}_{\mathcal{H}_0}(X < 49) = 1 - \mathbb{P}_{\mathcal{H}_0}(X \leqslant 49) = 1 - \mathbb{P}(Z \leqslant 2) = 1 - F(2)$$

où Z suit la loi $\mathcal{N}(0,1)$ et où F est sa fonction de répartition. On lit dans la table de cette loi $p_{value} = 0.0228$. Autrement dit, le risque pris en décidant $\mathcal{H}_1$ avec l'observation de C.L. est égal à 2.28%.

De façon similaire, l'observation pour B.J. est 50. Pour rejeter $\mathcal{H}_0$ avec cette observation, le domaine de rejet de $\mathcal{H}_0$ doit être de la forme $\mathcal{R} = \{X \in [50, +\infty[\}$: c'est le domaine de rejet de $\mathcal{H}_0$ du test qui rejette $\mathcal{H}_0$ si $X_{obs} \geqslant 50$, c'est le plus petit domaine permettant de rejeter $\mathcal{H}_0$ avec l'observation $X^{obs} = 50$ (*faire un dessin*). Le niveau du test associé à cette zone de rejet de $\mathcal{H}_0$ est :

$$p_{value} = 1 - \mathbb{P}_{\mathcal{H}_0}(X < 50) = 1 - \mathbb{P}_{\mathcal{H}_0}(X \leqslant 50) = 1 - \mathbb{P}(Z \leqslant 2.5) = 1 - F(2.5)$$

où Z suit la loi $\mathcal{N}(0,1)$ et où F est sa fonction de répartition. On lit dans la table de cette loi $p_{value} = 0.0062$. Autrement dit, le risque pris en décidant $\mathcal{H}_1$ avec l'observation de B.J est égal à 0.62%.

Et si on a choisi U_0 comme statistique de test, on trouve

$$p_{value} = \begin{cases} \mathbb{P}_{\mathcal{H}_0}(U_0 \geqslant 2) = 1 - \mathbb{P}_{\mathcal{H}_0}(U_0 < 2) = 1 - \mathbb{P}_{\mathcal{H}_0}(U_0 \leqslant 2) = 1 - F(2) = 2.28\% \text{ pour C.L.} \\ \\ \mathbb{P}_{\mathcal{H}_0}(U_0 \geqslant 2.5) = 1 - \mathbb{P}_{\mathcal{H}_0}(U_0 < 2.5) = 1 - \mathbb{P}_{\mathcal{H}_0}(U_0 \leqslant 2.5) = 1 - F(2.5) = 0.62\% \text{ pour B.J.} \end{cases}$$

7. Interprétation du résultat du test.

   En s'autorisant un risque de 1% on peut dire que le taux d'hématocrite de B.J. est anormalement élevé, mais on ne peut pas le dire pour C.L.

Seconde partie *On reprend la première partie. La correction insiste sur les 7 points de la construction d'un test.*

1. Préciser le modèle.

   On considère un individu A dont le vrai taux d'hématocrite dans le sang est τ. On modélise le résultat d'une mesure de ce taux par une variable aléatoire X de loi $\mathcal{N}(\tau; 4)$: l'erreur de mesure provient de l'accumulation de nombreux aléas indépendants et est donc modélisée par une loi normale, d'espérance égale au vrai taux et de variance $\sigma^2 = 4 = 2^2$ car les instruments sont précis à deux unités de mesure près. On a ici 9 mesures X_1,...,X_9 indépendantes et identiquement distribuées de loi $\mathcal{N}(\tau; 4)$. On se pose la question : "l'individu A a-t-il triché ?" qui se traduit avec les notations adoptées : *τ est-il égal à 45% (pas de triche) ou supérieur à 45% (triche) ?*

2. Déterminer les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test.

   On définit $\mathcal{H}_0 : \tau = 45$ et $\mathcal{H}_1 : \tau > 45$.

3. Déterminer la statistique de test.

   Pour faire un test sur τ, on choisit la variable aléatoire $\bar{X}$ (taux moyen mesuré). Sa loi est connue sous $\mathcal{H}_0 : \bar{X} \sim \mathcal{N}(45; 4/9)$ car les X_i sont indépendants et sa loi sous $\mathcal{H}_1$ est différente de sa loi sous $\mathcal{H}_0$: sous $\mathcal{H}_1 : \bar{X} \sim \mathcal{N}(\tau > 45; 4/9)$.

   Notons que le choix de
   $$U_0 := \frac{\sqrt{9}(\bar{X} - 45)}{2}$$
   comme statistique de test est également possible, voire plus judicieux. La loi de U_0 est connue sous $\mathcal{H}_0$, c'est la loi $\mathcal{N}(0\ ; 1)$ et la loi de U_0 sous $\mathcal{H}_1$ est la loi $\mathcal{N}(\delta\ ; 1)$, avec $\delta = \frac{\sqrt{9}(\tau - 45)}{2}$ ce qui implique que la densité de U_0 sous $\mathcal{H}_1$ est à droite de la densité de U_0 sous $\mathcal{H}_0$.

4. Établir la région de rejet de $\mathcal{H}_0$.

   On remarque que la loi de $\bar{X}$ sous $\mathcal{H}_1$ est à droite de la loi de $\bar{X}$ sous $\mathcal{H}_0$ (*faire un dessin pour s'en convaincre*). On rejettera donc $\mathcal{H}_0$ si on observe une valeur trop supérieure à 45%. Autrement dit, le domaine de rejet de $\mathcal{H}_0$ est de la forme $\mathcal{R} = \{\bar{X} \in [45 + a, +\infty[\}$. Autrement dit, on rejettera $\mathcal{H}_0$ si $\bar{X}_{obs} \geqslant 45 + a$.

   Notons que le choix de U_0 comme statistique de test nous conduisait à prendre comme domaine de rejet de $\mathcal{H}_0$ le domaine $\mathcal{R} = \{U_0 \in [b, +\infty[\}$.

5. Calculer le seuil de la région de rejet grâce à la loi de $\bar{X}$ sous $\mathcal{H}_0$.

   On calcule le seuil a grâce à l'équation de niveau :
   $$\mathbb{P}_{\mathcal{H}_0}(\bar{X} \in [45 + a, +\infty[) = \alpha$$
   qui équivaut à l'équation :
   $$\mathbb{P}_{\mathcal{H}_0}(\bar{X} \in]-\infty; 45 + a[) = 1 - \alpha$$

On adopte $\alpha = 1\%$ (faible niveau car on ne veut pas accuser quelqu'un de triche à tort). Sous $\mathcal{H}_0 : \bar{X} \sim \mathcal{N}(45{,}4/9)$. On centre et réduit :

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}(\bar{X} < 45 + a) &= \mathbb{P}_{\mathcal{H}_0}(\bar{X} \leqslant 45 + a) \\ &= \mathbb{P}_{\mathcal{H}_0}\left(\frac{\sqrt{9}(\bar{X} - 45)}{2} \leqslant \frac{\sqrt{9}(45 + a - 45)}{2}\right) \\ &= \mathbb{P}(Z \leqslant 3a/2) = F(3a/2)\end{aligned}$$

où Z suit la loi $\mathcal{N}(0{,}1)$ et où F est sa fonction de répartition. Après lecture dans les tables, $F(3a/2) = 1 - 0.01 = 0.99$ donne $3a/2 = 2.325$, c'est-à-dire $a = 1.55$.

Notons que le choix de U_0 comme statistique de test nous conduisait à trouver $b = 2.325$.

6. On calcule la valeur observée et on conclut.

   Première méthode pour conclure : on compare la valeur de la statistique au seuil de la région de rejet.

   La règle de décision est donc :

   - si $\bar{X}_{obs} \geqslant 46.55$, on décide de rejeter $\mathcal{H}_0$,
   - sinon, on décide de conserver $\mathcal{H}_0$.

   Les valeurs observées pour C.L. et B.J. sont respectivement $\bar{X}_1^{obs} = 46.44$ et $\bar{X}_2^{obs} = 47$. Ces valeurs conduisent à conserver $\mathcal{H}_0$ pour C.L. et à la rejeter pour B.J.

   Notons que le choix de U_0 comme statistique de test nous conduisait aux mêmes conclusions.

   Seconde méthode pour conclure : on calcule la p-valeur (*faire un dessin*).

   L'observation pour C.L. est 46.44. Pour rejeter $\mathcal{H}_0$ avec cette observation, le domaine de rejet de $\mathcal{H}_0$ doit être de la forme $\mathcal{R} = \{\bar{X} \in [46.44, +\infty[\}$, c'est le domaine de rejet de $\mathcal{H}_0$ du test qui rejette $\mathcal{H}_0$ si $\bar{X}_{obs} \geqslant 46.55$, c'est le plus petit domaine permettant de rejeter $\mathcal{H}_0$ avec l'observation $\bar{X}^{obs} = 46.44$ (*faire un dessin*). Le niveau du test associé à cette zone de rejet de $\mathcal{H}_0$ est :

   $$p_{value} = \mathbb{P}_{\mathcal{H}_0}(\bar{X} \geqslant 46.44) = 1 - \mathbb{P}_{\mathcal{H}_0}(\bar{X} < 46.44) = 1 - \mathbb{P}_{\mathcal{H}_0}(\bar{X} \leqslant 46.44) = 1 - \mathbb{P}(Z \leqslant 2.16) = 1 - F(2.16)$$

   où Z suit la loi $\mathcal{N}(0{,}1)$ et où F est sa fonction de répartition. On lit dans la table de cette loi $p_{value} = 0.0154$. Autrement dit, le risque pris en décidant $\mathcal{H}_1$ avec l'observation de C.L. est égal à 1.54%.

   De façon similaire, l'observation pour B.J. est 47. Pour rejeter $\mathcal{H}_0$ avec cette observation, le domaine de rejet de $\mathcal{H}_0$ doit être de la forme $\mathcal{R} = \{\bar{X} \in [47, +\infty[\}$, c'est le domaine de rejet de $\mathcal{H}_0$ du test qui rejette $\mathcal{H}_0$ si $\bar{X}_{obs} \geqslant 47$, c'est le plus petit domaine permettant de rejeter $\mathcal{H}_0$ avec l'observation $\bar{X}^{obs} = 47$ (*faire un dessin*). Le niveau du test associé à cette zone de rejet de $\mathcal{H}_0$ est :

   $$p_{value} = \mathbb{P}_{\mathcal{H}_0}(\bar{X} \geqslant 47) = 1 - \mathbb{P}_{\mathcal{H}_0}(\bar{X} < 47) = 1 - \mathbb{P}_{\mathcal{H}_0}(\bar{X} \leqslant 47) = 1 - \mathbb{P}(Z \leqslant 3) = 1 - F(3)$$

   où Z suit la loi $\mathcal{N}(0{,}1)$ et où F est sa fonction de répartition. On lit dans la table de cette loi $p_{value} = 0.0013$. Autrement dit, le risque pris en décidant $\mathcal{H}_1$ avec l'observation de C.L. est égal à 0.13%.

   Notons que le choix de U_0 comme statistique de test nous conduisait à trouver les mêmes p-valeurs.

7. Interprétation du résultat du test.

   En s'autorisant un risque de 1% on peut dire que le taux d'hématocrite de B.J. est anormalement élevé, mais on ne peut pas le dire pour C.L. La conclusion du test avec neuf mesures confirme la conclusion du test avec une mesure unique.

Feuille 2 - Exercice 3.

Encore une fois, il est impératif de rédiger le test en sept points.

1. Préciser le modèle.

   On considère une population (les Montpelliérains) dont la tension artérielle systolique moyenne est m, avec un écart-type de 2. On cherche à savoir si $m = 12$ comme pour la population française dans son ensemble. On effectue des mesures sur 28 Montpelliérains. Chaque mesure peut être modélisée par une variable aléatoire de loi normale $\mathcal{N}(m,2^2)$ et on suppose que les différentes mesures sont indépendantes entre elles. On modélise donc le résultat de ces mesures par $X_1, \ldots, X_{28}$, un 28-échantillon de loi $\mathcal{N}(m,2^2)$.

2. Déterminer les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test.

   On cherche à savoir si $m = 12$ ou $m \neq 12$. Il est naturel de supposer *a priori* que les Montpelliérains ont la même tension artérielle que le reste des Français ; de plus, il est nécessaire d'avoir une hypothèse $\mathcal{H}_0$ sous laquelle on connaît entièrement la loi de la variable de test. On définit donc $\mathcal{H}_0 : m = 12$ et $\mathcal{H}_1 : m \neq 12$.

3. Déterminer la statistique de test.

   Pour faire un test sur m, on choisit la variable aléatoire $\bar{X}_{28}$ (tension moyenne sur les 28 individus). Sa loi est connue sous $\mathcal{H}_0$ puisqu'alors $\bar{X}_{28} \sim \mathcal{N}(12; 2^2/28)$ et sa loi sous $\mathcal{H}_1$ est connue au paramètre m près, puisque ce sera $\mathcal{N}(m; 2^2/28)$. Cette variable a donc les propriétés requises d'une bonne statistique de test.

   Notons que le choix de

   $$U_0 := \frac{\sqrt{28}(\bar{X} - 12)}{\sqrt{4}} = \sqrt{7}(\bar{X} - 12)$$

   comme statistique de test est également possible, voire plus judicieux. En effet U_0 est une transformation affine de $\bar{X}$ (ce qui entraîne que si l'on connaît l'une des deux variables, on connaît l'autre) et contient donc la même information que $\bar{X}$. De plus, la loi de U_0 est connue sous $\mathcal{H}_0$, c'est la loi $\mathcal{N}(0\ ;1)$ et la densité de U_0 sous $\mathcal{H}_1$ est différente de la densité de U_0 sous $\mathcal{H}_0$. Plus précisément, la la loi de U_0 sous $\mathcal{H}_1$ est la loi $\mathcal{N}(\delta\ ;1)$ avec $\delta = \frac{\sqrt{28}(m-12)}{2}$. Ainsi, si $m > 12$, cette loi est « à droite » de la loi $\mathcal{N}(0\ ;1)$, tandis que si $m < 12$, cette loi est « à gauche » de la loi $\mathcal{N}(0\ ;1)$.

4. Établir la région de rejet de $\mathcal{H}_0$.

   On remarque que $\bar{X}_{28}$ est, sous $\mathcal{H}_1$, plutôt plus grand (si $m > 12$) ou plutôt plus petit (si $m < 12$) que sous $\mathcal{H}_0$. On rejettera donc $\mathcal{H}_0$ si on observe une valeur trop éloignée (trop supérieure ou trop inférieure) de 12. Autrement dit, le domaine de conservation I de $\mathcal{H}_0$ est de la forme $I = [12 - a,12 + a]$, c'est à dire que l'on conservera $\mathcal{H}_0$ ssi $12 - a \leqslant \bar{X} \leqslant 12 + a$ et que l'on rejetera $\mathcal{H}_0$ ssi $\bar{X} < 12 - a$ ou $\bar{X} > 12 + a$.

   Si l'on a choisi d'utiliser U_0 comme statistique de test, d'après la remarque sur la loi de U_0 du point précédent, on décidera $\mathcal{H}_1$ si U_0^{obs} est trop à gauche de 0 (on décidera alors que $m < 12$) ou trop à droite de 0 (on décidera alors que $m < 12$). Le domaine de conservation de $\mathcal{H}_0$ est de la forme $J = [-b,b]$, c'est à dire que l'on conservera $\mathcal{H}_0$ ssi $-b \leqslant U_0 \leqslant b$, ce qui peut aussi s'écrire $|U_0| \leqslant b$, et que l'on rejettera $\mathcal{H}_0$ ssi $|U_0| > b$.

5. Calculer le seuil de la région de rejet.

   Le a recherché doit avoir la propriété

   $$\mathbb{P}_{\mathcal{H}_0}(\bar{X}_{28} \notin [12 - a,12 + a]) = \alpha \tag{1}$$

où $\alpha = 3\%$ ou 5%. Or on sait que sous $\mathcal{H}_0$, $\bar{X}_{28} \sim \mathcal{N}(2{,}1/7)$. On centre et réduit :

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}(\bar{X}_{28} \in [12-a{,}12+a]) &= \mathbb{P}_{\mathcal{H}_0}(12-a \leqslant \bar{X}_{28} \leqslant 12+a) \\ &= \mathbb{P}_{\mathcal{H}_0}\Big(\frac{12-a-12}{\sqrt{1/7}} \leqslant \frac{\bar{X}_{28}-12}{\sqrt{1/7}} \leqslant \frac{12+a-12}{\sqrt{1/7}}\Big) \\ &= \mathbb{P}(-a\sqrt{7} \leqslant Z \leqslant +a\sqrt{7})\end{aligned}$$

où Z suit la loi $\mathcal{N}(0{,}1)$. Par les manipulations habituelles, ceci vaut $2P(Z \leqslant a\sqrt{7}) - 1$. Au niveau $\alpha = 3\%$, le a recherché doit être $a_{3\%}$ satisfaisant

$$2P(Z \leqslant a_{3\%}\sqrt{7}) - 1 = 0{,}97 \quad \text{donc} \quad P(Z \leqslant a_{3\%}\sqrt{7}) = 0{,}985$$

et la lecture des tables donne $a_{3\%}\sqrt{7} = 2{,}17$, donc $a_{3\%} = 0{,}82$. Au niveau $\alpha = 5\%$, le a recherché doit être $a_{5\%}$ satisfaisant

$$2P(Z \leqslant a_{5\%}\sqrt{7}) - 1 = 0{,}95 \quad \text{donc} \quad P(Z \leqslant a_{5\%}\sqrt{7}) = 0{,}975$$

et la lecture des tables donne $a_{5\%}\sqrt{7} = 1{,}96$ donc $a_{5\%} = 0{,}74$.

La règle de décision est donc :

- au niveau $\alpha = 3\%$, si $11{,}18 \leqslant \bar{x}_{28}^{obs} \leqslant 12{,}82$, on conserve $\mathcal{H}_0$, sinon, on rejette $\mathcal{H}_0$;
- au niveau $\alpha = 5\%$, si $11{,}26 \leqslant \bar{x}_{28}^{obs} \leqslant 12{,}74$, on conserve $\mathcal{H}_0$, sinon, on rejette $\mathcal{H}_0$.

On observe bien que sous $\alpha = 3\%$, le test est plus conservateur (au sens où l'on conserve plus facilement $\mathcal{H}_0$) que sous $\alpha = 5\%$.

Et si l'on a choisi d'utiliser U_0 comme statistique de test, l'équation de niveau

$$\mathbb{P}_{\mathcal{H}_0}(U_0 \notin [-b{,}b]) = \alpha \text{ (= 3\% ou 5\%)}$$

se résout très simplement en utilisant les propriétés de la loi normale centrée réduite : on sait que

$$\mathbb{P}_{\mathcal{H}_0}(U_0 \notin [-b{,}b]) = 2(1 - F_{\mathcal{H}_0}(b)).$$

Ainsi, on obtient

$$2(1 - F_{\mathcal{N}(0,1)}(b)) = \alpha \text{ (= 3\% ou 5\%)},$$

c'est à dire

$$F_{\mathcal{N}(0,1)}(b)) = \frac{1-\alpha}{2} \text{ (= 1,5\% ou 2,5\%)},$$

soit $b = 2{,}17$ si $\alpha = 3\%$ et $b = 1{,}96$ si $\alpha = 5\%$.

6. On calcule la valeur observée et on conclut.

Première méthode pour conclure : on compare la valeur de la statistique au seuil de la région de rejet.

La valeur observée pour $\bar{X}_{28}$ est $\bar{x}_{28}^{obs} \simeq 12{,}99$. On rejette donc l'hypothèse $\mathcal{H}_0$, que ce soit au niveau $\alpha = 3\%$ ou au niveau $\alpha = 5\%$.

Et si l'on a choisi d'utiliser U_0 comme statistique de test, on trouve $U_0^{obs} = \sqrt{7}(12{,}99 - 12) \simeq 2{,}62$, ce qui nous conduit à rejeter l'hypothèse $\mathcal{H}_0$, que ce soit au niveau $\alpha = 3\%$ ou au niveau $\alpha = 5\%$.

Seconde méthode pour conclure : on calcule la p-valeur (*faire un dessin*).

Le “cas limite" pour la conclusion correspond au cas où $12+a = 12{,}99$ donc $a = 0{,}99$. D'après les calculs ci-dessus, le risque qui ferait basculer la décision est la p_{valeur} vérifiant

$$p_{valeur} = \mathbb{P}_{\mathcal{H}_0}(X \notin [11{,}01\ ;\ 12{,}99] = \mathbb{P}_{\mathcal{H}_0}(\frac{\bar{X}_{28} - 12}{\sqrt{1/7}} \notin [-0{,}99\sqrt{7}\ ;\ 0{,}99\sqrt{7}]) = 2 \times \mathbb{P}(Z > 0{,}99\sqrt{7})$$

où Z suit la loi $\mathcal{N}(0,1)$. Comme $a\sqrt{7} \simeq 2{,}62$ on trouve $p_{valeur} = 2(1 - F_{\mathcal{N}(0,1)}(2{,}62)) \simeq 0{,}9\%$.

Et si l'on a choisi d'utiliser U_0 comme statistique de test, on a

$$p_{valeur} = \mathbb{P}_{\mathcal{H}_0}(U_0 \notin [-2{,}62, 2{,}62]) = 2(1 - F_{\mathcal{N}(0,1)}(2{,}62)) \simeq 0{,}9\%.$$

7. Interprétation du résultat du test.

   En s'autorisant un risque de 3% (et donc à fortiori un risque de 5%) on conclut que les Montpelliérains ont une tension artérielle systolique au repos moyenne différente, et plus précisément supérieure, à celle de la population française prise dans son ensemble.

Feuille 2 - Exercice 4.

Quasiment le même exercice que l'exercice précédent mais à variance inconnue, on utilise le test de Student à 27 degrés de liberté au lieu d'utiliser le test gaussien. Au lieu d'utiliser

$$U_0 := \frac{\sqrt{28}(\bar{X} - 12)}{\sqrt{4}}$$

qui n'est plus une statistique car non calculable à partir des données, on utilise la statistique

$$T_0 := \frac{\sqrt{28}(\bar{X} - 12)}{\sqrt{S^2}}$$

où S^2 est l'estimateur de la variance empirique :

$$S^2 = \frac{1}{27}\sum_{i=1}^{28}(X_i - \bar{X})^2 = (\frac{1}{27}\sum_{i=1}^{28} X_i^2) - \frac{28}{27}(\bar{X})^2.$$

La loi de T_0 sous H_0 est une loi de Student à 27 degrés de liberté et sa déformation sous H_1 est analogue à la déformation de U_0 sous H_1.

Application numérique :

$S^2_{obs} = 5.8058$; $\overline{X}_{obs} = 12.9882$; $U_0^{obs} = 2.6146$ (statistique gaussienne) ; $T_0^{obs} = 2.1702$ (statistique de Student)

Les seuils du test bilatéral de niveau 5% sont $\pm 2{,}052$ et la méthode 1 conduit donc à rejeter H_0 par la droite et à conclure que $m > 12$.

Pour évaluer la p-valeur, on procède à un encadrement : $2{,}052 < 2{,}1702 < 2{,}473$ donc $F_{\mathcal{T}(27)}(2{,}052) < F_{\mathcal{T}(27)}(2{,}1702) < F_{\mathcal{T}(27)}(2{,}473)$ et au final, $2\% <$ p-valeur $< 5\%$: la seconde méthode nous conduit également à rejeter H_0. Notons en passant que le seul calcul de la p-valeur ne permet pas de conclure que $m > 12$, seulement que $m \neq 12$: il faut revenir aux données pour constater que $\overline{X}_{obs}$ étant supérieur à 12, le rejet de H_0 ne peut être qu'à droite.

Feuille 3 - Exercice 1.

Il s'agit d'un test du chi-deux d'adéquation. On peut approximer la loi de la statistique par une loi du chi-deux à 5-1 =4 degrés de liberté car tous les effectifs espéré sous H_0 sont supérieurs à 5.

Classe d'âge	0-19	20-39	40-59	60-74	> 74
Effectifs espérés sous H_0	69,9	79,8	73,8	38,6	21,9

La région de rejet de H_0 au niveau 5% est $\mathcal{R} = \{Z \geqslant 9{,}488\}$. La valeur observée de la statistique est $Z_{obs} = 1{,}09$.

Feuille 3 - Exercice 2.

★ Deux modélisations sont possibles : on peut soit effectuer un test du chi-deux d'homogénéité, soit un test du chi-deux d'indépendance. Dans ce cours, seul le test du chi-deux d'indépendance a été complètement développé.

☞ Chi-deux d'homogénéité.

Classe d'âge j	0-19	20-39	40-59	60-74	> 74
Estimation des p_j sous H_0	0,2950	0,2722	0,2458	0,1343	0,0528

$Z_{obs} \simeq 16{,}42$ pour 8 degrés de liberté, la p-valeur vaut environ 3,7% : H_0 est rejetée.

☞ Chi-deux d'indépendance.

Classe d'âge	0-19	20-39	40-59	60-74	> 74	Marges horizontales
Population A	78	87	78	39	18	300
Population B	95	58	52	37	8	250
Population C	73	82	75	36	18	284
Marges verticales	246	227	205	112	44	834

Estimation des effectifs espérés sous H_0 selon la formule (marge verticale * marge horizontale / 834) :

Classe d'âge	0-19	20-39	40-59	60-74	> 74	Total
Population A	88.4892	81.6547	73.7410	40.2878	15.8273	300
Population B	73.7410	68.0456	61.4508	33.5731	13.1894	250
Population C	83.7698	77.2998	69.8082	38.1391	14.9832	284
Total	246	227	205	112	44	834

$Z_{obs} \simeq 16{,}42$ pour 8 degrés de liberté, la p-valeur vaut environ 3,7% : H_0 est rejetée.

★ Idem que l'étoile précédente, les chi-deux ont maintenant $(5-1)*(2-1) = 4$ degrés de liberté. Pour les populations A et B, voici le tableau des effectifs espérés sous H_0 :

Classe d'âge	0-19	20-39	40-59	60-74	> 74	Total
Population A	94.36	79.09	70.91	41.45	14.18	300.
Population B	78.64	65.91	59.09	34.55	11.82	250.
Total	173.	145.	130.	76.	26.	550.

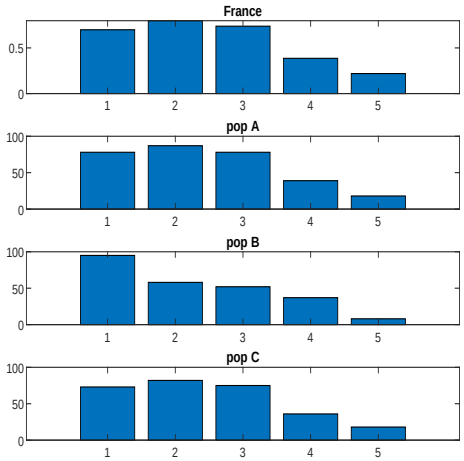
Le Z_{obs} est $12{,}124$ et le $Z_{seuil} = 9.4877$ donc on rejette.

★ Chi-deux d'adéquation comme dans le premier exercice de la feuille. Le chi-deux a 4 degrés de liberté. Pour les populations A voici le tableau des effectifs espérés sous H_0 :

Classe d'âge	0-19	20-39	40-59	60-74	> 74
Population A	73.8	84.3	78.	40.8	23.1

Le Z_{obs} est 1.5309 et le $Z_{seuil} = 9.4877$ donc on ne rejette pas.

★ Toujours représenter ses données, et parfois il est inutile de réaliser un test statistique : la visualisation suffit. On voit sur les histogrammes que les sous-populations A et C sont comparables à la population française mais que ce n'est pas le cas de la sous-population B.



Feuille 3 - Exercice 3, Partie I.

1. **Modèle.** On note par X_k la variable aléatoire représentant la base qui apparait sur la $k-$ème position dans la séquence ADN, $1 \leqslant k \leqslant 100$. Alors $X_1, \ldots, X_{100}$ est un $100-$échantillon de loi $\mathcal{L}$ donnée par

$$P(X_1 = A) = p_A,\ P(X_1 = T) = p_T,\ P(X_1 = C) = p_C,\ P(X_1 = G) = p_G.$$

La question du test est de savoir si la séquence ADN correspond à un organisme $\mathcal{O}$, c'est-à-dire,

$$p_A = 0.2,\ p_T = 0.3,\ p_C = 0.24,\ p_G = 0.26.$$

2. **Hypothèses.**
$\mathcal{H}_0 : p_A = 0{,}2$ et $p_T = 0{,}3$ et $p_C = 0{,}24$ et $p_G = 0{,}26$..
$\mathcal{H}_1 : p_A \neq 0{,}2$ ou $p_T \neq 0{,}3$ ou $p_C \neq 0{,}24$ ou $p_G \neq 0{,}26$..

3. **Statistique.** On note par N_A le nombre d'apparitions de la base A dans la séquence ADN. Idem pour N_T, N_C et N_G. Leurs espérances sont respectivement $100p_A$, $100p_T$, $100p_C$, $100p_G$. Sous $\mathcal{H}_0$, leurs espérances sont donc respectivement

$$n_A = 20,\ n_T = 30,\ n_C = 24,\ n_G = 26.$$

On choisit la statistique

$$Z = \frac{(N_A - 20)^2}{20} + \frac{(N_T - 30)^2}{30} + \frac{(N_C - 24)^2}{24} + \frac{(N_G - 26)^2}{26}.$$

Sous $\mathcal{H}_0$, la loi de Z peut être approchée par la loi $\chi^2(3)$ car n_A, n_T, n_C, n_G sont supérieurs à 5. Sous $\mathcal{H}_1$, Z a tendance de prendre des valeurs plus grandes que sous $\mathcal{H}_0$.

4. **Zone de rejet.** Puisque Z prend des valeurs plus grandes sour $\mathcal{H}_1$ que sous $\mathcal{H}_0$, on rejette $\mathcal{H}_0$ si on observe une valeur « trop grande » de Z. On choisit donc de rejeter $\mathcal{H}_0$ sous une condition de la forme $\{Z \geqslant a\}$.

5. **Calcul du seuil de la zone de rejet.** Pour un test de niveau 5%, le seuil a doit vérifier $5\% = \mathbb{P}_{\mathcal{H}_0}(Z \geqslant a)$, or sous $\mathcal{H}_0$, Z suit approximativement une loi $\chi^2(3)$ donc

$$\mathbb{P}_{\mathcal{H}_0}(Z \geqslant a) = 1 - \mathbb{P}_{\mathcal{H}_0}(Z < a) \simeq 1 - F_{\chi^2(3)}(a)$$

donc $F_{\chi^2(3)}(a) = 0{,}95$. La table de valeurs numériques donne $a \simeq 7{,}81$.

6. **Décision.** On compte $N_A^{obs} = 19$, $N_T^{obs} = 27$, $N_C^{obs} = 22$ et $N_G^{obs} = 32$. Cela implique que la valeur observée de la statistique est $Z_{obs} \simeq 1{,}90$.

Décision par comparaison de la statistique au seuil. Comme $Z_{obs} < a$, le test statistique de niveau 5% ne rejette pas H_0.

Décision par calcul de la p-valeur. On calcule

$$p_{value} = P_{\mathcal{H}_0}(Z > z_{obs}) \simeq 1 - F_{\chi^2(3)}(1{,}90).$$

Les tables papier permettent simplement de conclure que la p-valeur se situe entre 20% et 80%. En utilisant les tables de WIMS, on trouve $F_{\chi^2(3)}(1{,}90) = 0.4066$, soit $p_{value} = 59.34\%$. Comme $p_{value} > 5\%$, le test statistique de niveau 5% ne rejette pas H_0.

7. **Phrase de conclusion.** On ne peut pas conclure que l'échantillon ne provient pas de $\mathcal{O}$.

Feuille 3 - Exercice 3, Partie II.

(a)

Bases	A	T	C	G	total
Nombre de bases **observées** dans le premier échantillon	19	27	22	32	100
Nombre de bases **observées** dans le second échantillon	18	8	12	22	60

(b) On veut comparer les répartitions observées de deux variables catégorielles, on fait donc un test du χ^2 d'homogénéité, qui consiste ici à tester l'indépendance de la variable « type de base » et de la variable « échantillon d'origine ».

Modèle : On note par X_k la variable aléatoire représentant la base qui apparaît sur la $k-$ème position dans la séquence ADN, et par Y_k la variable aléatoire représentant le numéro de l'échantillon d'origine. Les variables X prennent donc les valeurs A,C,T,G et les variables Y les valeurs 1,2. On suppose que $(X_k,Y_k)_k$ forme un 160-échantillon (c'est-à-dire que chaque (X_k,Y_k) est indépendant des autres couples $(X_{k'},Y_{k'})$).

Hypothèses : on pose les hypothèses suivantes :

H_0 : la variable X est indépendante de la variable Y,

H_1 : les variables X et Y ne sont pas indépendantes.

Statistique : pour $i \in \{A,C,T,G\}$ et $j \in \{1,2\}$, on note $N_{i,j}$ la variable aléatoire donnant l'effectif, dans l'échantillon (X_1,Y_1), …, (X_{160},Y_{160}) de la valeur (i,j) (le nombre de couples (X_n,Y_n) qui prennent la valeur (i,j)). On définit aussi $N_{i,*}$ l'effectif de i dans l'échantillon $X_1,\dots,X_{160}$ et $N_{*,j}$ l'effectif de j dans l'échantillon $Y_1,\dots,Y_{160}$. Les valeurs observées pour ces variables sont donc

	A	T	C	G	total
1	19	27	22	32	100
2	18	8	12	22	60
total	37	35	34	54	160

On définit alors la variable

$$Z = \sum_{i\in\{A,C,T,G\}} \sum_{j\in\{1,2\}} \frac{\left(N_{i,j} - \frac{N_{i,*}N_{*,j}}{160}\right)^2}{\frac{N_{i,*}N_{*,j}}{160}}.$$

Pour savoir si l'on peut approximer la loi de Z sous H_0 par une loi du chi-deux, calculons les effectifs espérés sous H_0, que l'on peut noter $n_{i,j}$, par la formule :

$$n_{i,j} = \frac{N_{i,*}^{\mathrm{obs}} N_{*,j}^{\mathrm{obs}}}{160}.$$

On obtient les valeurs suivantes pour les $n_{i,j}$:

	A	T	C	G	total
1	23,125	21,875	21,25	33,75	100
2	13,875	13,125	12,75	20,25	60
total	37	35	34	54	160

Les effectifs observées vérifient tous $n_{i,j} \geqslant 5$ pour tous les couples (i,j) donc sous l'hypothèse H_0, la variable Z suit approximativement une loi $\chi^2((2-1)\times(4-1))$.

Remarque : lorsque les effectifs espérés sous H_0 sont connus, il n'est pas nécessaire de les calculer, et dans ce cas le nombre de degrés de liberté de la loi du chi-deux est (nombre de classes pour X) $\times$ (nombre de classes pour Y) -1.

Règle de rejet : Sous l'hypothèse H_1 la variable Z ne suit plus la loi $\chi^2((2-1)\times(4-1))$ et sa densité est décalée à droite de la densité de la loi $\chi^2((2-1)\times(4-1))$ (admis). Par conséquent on va rejeter H_0 si les valeurs observées pour Z sont « trop grandes » : on rejette H_0 si $Z_{obs} \geqslant z_0$.

Calcul du seuil : On a $5\% = P_{\mathcal{H}_0}(Z \geqslant a) \simeq 1 - F_{\chi^2(3)}(a)$, d'où $a = 7{,}81$.

Décision : À partir des données précédentes, on peut calculer $Z_{\text{obs}} \simeq 5{,}48$.

Décision par comparaison de la statistique au seuil. Comme $Z_{\text{obs}} < a$, le test statistique de niveau 5% ne rejette pas H_0.

Décision par calcul de la p-valeur. On calcule

$$p_{value} = P_{\mathcal{H}_0}(Z > Z_{obs}) \simeq 1 - F_{\chi^2(3)}(Z_{obs}).$$

Les tables papier permettent simplement de conclure que la p-valeur se situe entre 10% et 15%.
En utilisant la commande pchisq(q = 5.48, df = 3) du logiciel R, on trouve $F_{\chi^2(3)}(5.48) = 0.86$, soit $p_{value} = 14\%$. Comme $p_{value} > 5\%$, le test statistique de niveau 5% ne rejette pas H_0.

Phrase de conclusion : On ne peut pas conclure que les deux séquences proviennent d'un organisme différent.

Feuille 3 - Exercice 4.

Premier cas.
QUESTION I
Sous l'hypothèse d'indépendance : $P(X = J,Y = K) = P(X = J) \times P(Y = K)$ donc ...

QUESTION 2

– Modèle : Soit X_i la v.a. qui décrit la couleur du ieme grain et soit Y_i la v.a. qui décrit la forme du ieme grain. On note $N_{[RL]}$; $N_{[R\ell]}$; $N_{[rL]}$; $N_{[r\ell]}$ les effectifs observables de chaque classe et on note $p_{[RL]}$; $p_{[R\ell]}$; $p_{[rL]}$; $p_{[r\ell]}$ sont les probabilités (inconnues) de chaque classe).

– Hypothèses : H0 : $p_{[RL]} = 9/16$ et $p_{[R\ell]} = 3/16$ et $p_{[rL]} = 3/16$ et $p_{[r\ell]} = 1/16$ contre H1 : $p_{[RL]} \neq 9/16$ ou $p_{[R\ell]} \neq 3/16$ ou $p_{[rL]} \neq 3/16$ ou $p_{[r\ell]} \neq 1/16$

– Statistique de test et sa loi sous H_0 :

$$STAT = \sum_{1}^{N_{classes}} \frac{(n_{obs} - n_{theor})^2}{n_{theor}}$$

suit loi du chi-deux à $(N_{classes} - 1)$ degrés de liberté sous si les conditions d'application $n_{theor} \geqslant 5$ sont vérifiées. Dans notre cas, le nombre de classes est 4 donc le nombre de degrés de liberté est 3, les effectifs observables sont les $N_{[JK]}$ et les effectifs espérés/théoriques sous H_0 valent $284 \times p_{[JK]}$. Les effectifs observés et les effectifs théoriques sous H_0 sont résumés dans les tableaux suivants :

X/Y	L	l	somme
R	161	57	218
r	57	9	66
somme	218	66	284

X/Y	L	l	somme
R	159.75	53.25	218
r	53.25	17.75	66
somme	218	66	284

– Zone de rejet : on rejetteH_0 si la statistique observée tombe dans $[a; +\infty[$, où a est un seuil à déterminer.

– Seuil pour le niveau 5% : on lit sur la table du chi-deux à 3 degrés de liberté $a = 7.81$.

– Conclusion : La statistique vaut 4,8513 ; méthode 1 : 4,8513 < 7,81 donc le test de niveau 5% conserve H_0 ; méthode 2 : la p-valeur vaut 18,30% donc le test de niveau 5% conserve H_0.

– Conclusion biologique : au risque 5%, les effectifs observés sont compatibles avec une indépendance génétique des deux caractères.

Second cas.
On effectue un test du chi-deux d'indépendance entre les variables couleur et forme. Lse tableaux des effectifs observés et des effectifs théoriques (estimés) sous H_0 sont :

X/Y	L	l	somme
R	161	57	218
r	57	9	66
somme	218	66	284

X/Y	L	l	somme
R	167.338	50.662	218
r	50.662	15.338	66
somme	218	66	284

La statistique du chi-deux d'indépendance vaut 4,4449 ; les degrés de liberté valent 1 ; le seuil du test de niveau 5% vaut 3,8415 ; la p-valeur vaut 3,50%. On rejette H_0, on conclut à la dépendance des deux caractères.

Bonus : je ne sais pas, c'est une question ouverte ... probablement parce que la lignée n'est pas propre et que dans certaines lignées la couleur et la forme ne sont pas indépendantes, et probablement aussi que ce n'était pas le cas dans la lignée non sauvage.