Correction TD 4

Correction 1.

(1) Le statisticien choisit pour hypothèse nulle $\mathcal{H}_0$ l'hypothèse admise par la scierie, c'est à dire $\mathcal{H}_0$: "$m = m_0 := 52$ cm". La scierie espère montrer que la longueur moyenne de ses planches n'est pas statistiquement trop éloignée de m_0. Leur point de vue est symétrique au sens où les planches ne doivent pas être trop petites ni trop grandes. Ainsi, on choisira pour $\mathcal{H}_1$: "$m \neq m_0$".

(2) On note $n = 10$, le nombre d'observations, et pour tout $i = 1, \ldots, n$, on note X_i la variable aléatoire représentant la longueur de la i-ème planche. Les variables aléatoires X_i sont indépendantes et identiquement distribuées de loi $\mathcal{N}(m, \sigma^2)$, où σ^2 est la variance de la longueur d'une planche. <u>On suppose pour l'instant que $\sigma = 2$.</u>

Pour tester la valeur de la moyenne m, on utilise la statistique suivante,

$$Z_n = \sqrt{n}\frac{\bar{X}_n - m_0}{\sigma},$$

où $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

(3) Compte tenu de la réponse à la première question, on souhaite rejeter l'hypothèse nulle lorsque $\bar{X}_n$ est trop éloigné de m_0, i.e. quand Z_n est trop éloigné de 0. Donc, on utilisera une zone de rejet de la forme $]-\infty, A[\cup]B, +\infty[$.

(4) Soit $\alpha \in]0, 1[$. Au vu de la réponse à la question précédente, il faut commencer par déterminer un couple (u_α, v_α) tel que

$$\mathbb{P}_{\mathcal{H}_0}\left(Z_n < u_\alpha \text{ ou } Z_n > v_\alpha\right) \leq \alpha. \tag{1}$$

Or, sous $\mathcal{H}_0$, $Z_n \sim \mathcal{N}(0, 1)$, donc si un couple (u_α, v_α) est tel que

$$\begin{cases} \mathbb{P}\left(\mathcal{N}(0,1) < u_\alpha\right) \leq \alpha/2, \\ \mathbb{P}\left(\mathcal{N}(0,1) > v_\alpha\right) \leq \alpha/2, \end{cases} \tag{2}$$

alors il convient (i.e. il vérifie l'équation (1)). Ainsi, le test basé sur la règle de décision suivante :

"On rejette $\mathcal{H}_0$ si $Z_n < u_\alpha$ ou si $Z_n > v_\alpha$."

est un test de niveau α.

(5) **A.N.** On a $\alpha = 0,01$. On voit sur la table de la loi normale que le couple $(u_\alpha, v_\alpha) = (-2.58, 2.58)$ vérifie (2).

De plus, on a $n = 10$, la moyenne empirique des observations est $\overline{x}_n = 50.41$, donc $z_n \approx -2.51$. Ainsi, $u_\alpha \leq t_n \leq v_\alpha$ et donc on accepte l'hypothèse nulle. Avec $\alpha = 0.05$, on choisit $(u_\alpha, v_\alpha) = (-1.96, 1.96)$ et on rejette l'hypothèse nulle. Le seuil critique est 1.2%.

(6) On cherche la probabilité de rejeter $\mathcal{H}_0$ lorsque la vraie moyenne est $m = 54$ cm. Avec les notations de la question précédente, on a

$$\mathbb{P}_{m=54}\left(Z_n < u_\alpha \text{ ou } Z_n > v_\alpha\right) = 1 - \mathbb{P}_{m=54}\left(u_\alpha \leq Z_n \leq v_\alpha\right).$$

On rappelle que

$$Z_n = \sqrt{n}\frac{\bar{X}_n - m_0}{\sigma} = \sqrt{n}\frac{\bar{X}_n - m}{\sigma} + \sqrt{n}\frac{m - m_0}{\sigma}.$$

Or, $T_n := \sqrt{n}\frac{\bar{X}_n - m}{\sigma}$ suit une loi normale $\mathcal{N}(0,1)$. Donc,

$$\mathbb{P}_{m=54}\left(Z_n < u_\alpha \text{ ou } Z_n > v_\alpha\right) = 1 - \mathbb{P}\left(u_\alpha - \sqrt{n}\frac{m - m_0}{\hat{\sigma}_n} \leq T_n \leq v_\alpha - \sqrt{n}\frac{m - m_0}{\hat{\sigma}_n}\right).$$

A.N. On a $m - m_0 = 2$ cm, donc $\sqrt{n}\frac{m-m_0}{\sigma} = 3,16$. Ainsi, la puissance du test au point $m = 54$ est d'environ 0.88.

(7) On prend $\mathcal{H}_0$: "$m \geq m_0$" et $\mathcal{H}_1$: "$m < m_0$". La statistique de test reste la même mais on utilisera une zone de rejet de la forme $]-\infty, A[$. En fait, la zone de rejet est $]-\infty, t_\alpha[$ où t_α est tel que $F(t_\alpha) = \alpha$. On lit sur la table $t_{0.01} = -2.33$ et $t_{0.05} = -1.65$. L'hypothèse nulle est donc rejetée dans les deux cas. Le seuil critique est 0.6%.

(8) On utilise la statistique de Student, c'est-à-dire

$$T_n = \sqrt{n}\frac{\bar{X}_n - m_0}{\sqrt{V_n}},$$

où $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Ensuite, au lieu de la loi normale centrée réduite, on utilise le fait que sous $\mathcal{H}_0$, T_n suit une loi de Student de paramètre $n-1$.

Correction 2. On note p la proportion de garçons qui naissent dans la population de la maternité. On teste $\mathcal{H}_0$: "$p \leq p_0 = 1/2$" contre $\mathcal{H}_1$: "$p > 1/2$". Soit $(X_1, \ldots, X_n)$ un échantillon de taille $n = 650$ de loi $\mathcal{B}(p)$. Comme n est grand, le TCL nous dit que l'on peut approximer la variable

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$$

par $Z \sim \mathcal{N}(0,1)$. Ainsi, si $p = p_0$, la variable

$$Z_n = \sqrt{n}\frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}$$

peut être approximé par $Z \sim \mathcal{N}(0,1)$. On choisit Z_n comme statistique de test et une zone de rejet unilatéralle à droite. Pour un niveau de risque $\alpha = 5\%$, on considère donc la zone de rejet $]t_\alpha, +\infty[$ où $F(t_\alpha) = 1 - \alpha$, i.e. $t_\alpha = 1.65$ ici. Dans ce cas là, on a bien $\mathbb{P}(Z_n > t_\alpha) \approx \mathbb{P}(Z > t_\alpha) = \alpha$ dès que $p = p_0$.

A.N. On a $\overline{x}_n \approx 0.51$, $p_0 = 0.5$ et $n = 650$, d'où $z_n \approx 0.47$. On ne peut donc pas conclure au risque de 5% qu'il y a plus de chance d'avoir un garçon plutôt qu'une fille.

Correction 3.
On reprend ici l'enchaînement des questions de l'exercice 1. On note m le nombre moyen de bonbons que Toto mange en une journée. Attention, m est différent de $\overline{x}_n$.

(1) Le père de Toto cherche à montrer que son fils mange trop de bonbons. C'est lorsqu'un test rejette l'hypothèse nulle que son résultat est sans parti pris. On choisit pour hypothèse nulle $\mathcal{H}_0$: "$m \leq m_0 := 10$". Le père de Toto espère montrer que Toto mange (de manière significative) plus de m_0 bonbons par jour. Son point de vue n'est pas symétrique. Ainsi, on choisira pour $\mathcal{H}_1$: "$m > m_0$".

(2) On note $n = 365$, le nombre d'observations, et pour tout $i = 1, \ldots, n$, on note X_i la variable aléatoire représentant le nombre de bonbons que Toto a mangé pendant le i-ème jour. Les variables aléatoires X_i sont indépendantes et identiquement distribuées.

Grâce au TCL et au lemme de Slutsky, on sait que

$$Z_n = \sqrt{n}\frac{\bar{X}_n - m}{\sqrt{V_n}} \xrightarrow{n \to +\infty} \mathcal{N}(0,1)$$

où $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ et $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. En particulier, sous $\mathcal{H}_0$,

$$T_n := \sqrt{n}\frac{\bar{X}_n - m_0}{\sqrt{V_n}} = \sqrt{n}\frac{\bar{X}_n - m}{\sqrt{V_n}} + \sqrt{n}\frac{m - m_0}{\sqrt{V_n}} \leq Z_n.$$

On choisira la statistique T_n pour établir notre test.

(3) Compte tenu de la réponse à la première question, on souhaite rejeter l'hypothèse nulle lorsque $\bar{X}_n$ est significativement grand par rapport à m_0, i.e. quand T_n est grand par rapport à 0. Donc, on utilisera une zone de rejet de la forme $]B, +\infty[$.

(4) Soit $\alpha \in]0,1[$. Au vu de la réponse à la question précédente, il faut commencer par déterminer u_α tel que

$$\lim_{n\to\infty} \mathbb{P}(T_n > u_\alpha) \leq \alpha \text{ dès que } \mathcal{H}_0 \text{ est vraie.} \tag{3}$$

Or, sous $\mathcal{H}_0$, $T_n \leq Z_n \to \mathcal{N}(0,1)$, donc si u_α est tel que

$$\mathbb{P}(\mathcal{N}(0,1) > u_\alpha) \leq \alpha. \tag{4}$$

alors il convient (i.e. il vérifie l'équation (3)). En effet, si $\mathcal{H}_0$ est vraie, on a

$$\mathbb{P}(T_n > u_\alpha) \leq \mathbb{P}(Z_n > u_\alpha) \xrightarrow{n\to\infty} \mathbb{P}(\mathcal{N}(0,1) > u_\alpha).$$

Ainsi, le test basé sur la règle de décision suivante :

"On rejette $\mathcal{H}_0$ si $T_n > u_\alpha$."

est un test de niveau α.

(5) **A.N.** On a $\alpha = 0,05$. On voit sur la table de la loi normale que $u_\alpha = 1,65$ vérifie (4).

De plus, on a $n = 365$, $\overline{x}_n = 10,1$ et $v_n = \hat{\sigma}_n = 1$, donc $t_n = 1,91$. Ainsi, $t_n > u_\alpha$ et donc on rejette l'hypothèse nulle. Toto ne va pas être content !

Correction 4.

(1) Pamela choisit pour hypothèse nulle l'hypothèse admise par Cruella, c'est à dire $\mathcal{H}_0$: "$v^2 \geq v_0^2 := 3$ kg^2". Elle espère montrer que la variance de son poids est bien inférieure à v_0^2. Son point de vue n'est pas symétrique. Ainsi, on choisira pour $\mathcal{H}_1$: "$v^2 < v_0^2$".

(2) On note $n = 7$, le nombre d'observations, et pour tout $i = 1, \ldots, n$, on note X_i la variable aléatoire représentant le poids de Pamela lors du i-ème jour. Les variables aléatoires X_i sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu, v^2)$, où μ est le poids moyen de Pamela. Il est inconnu.

Pour tester la valeur de la variance v^2, on utilise la statistique du chi-deux,

$$C_n = (n-1)\frac{V_n}{v_0^2},$$

où $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ et $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. En particulier, sous $\mathcal{H}_0$,

$$C_n = (n-1)\frac{V_n}{v^2}\frac{v^2}{v_0^2} \geq Z_n := (n-1)\frac{V_n}{v^2} \sim \chi^2(n-1).$$

(3) Compte tenu de la réponse à la première question, on souhaite rejeter l'hypothèse nulle lorsque V_n est significativement inférieure à v_0^2, i.e. quand C_n est bien inférieure à $n-1$. Donc, on utilisera une zone de rejet de la forme $]-\infty, A[$.

(4) Soit $\alpha \in]0,1[$. Au vu de la réponse à la question précédente, il faut commencer par déterminer u_α tel que

$$\mathbb{P}\left(C_n < u_\alpha\right) \leq \alpha \text{ dès que } \mathcal{H}_0 \text{ est vraie.} \tag{5}$$

Il suffit de prendre u_α tel que

$$\mathbb{P}\left(\chi^2(n-1) < u_\alpha\right) \leq \alpha. \tag{6}$$

En effet, sous $\mathcal{H}_0$, $C_n \geq Z_n$ et donc

$$\mathbb{P}\left(C_n < u_\alpha\right) \leq \mathbb{P}\left(Z_n < u_\alpha\right) = \mathbb{P}\left(\chi^2(n-1) < u_\alpha\right).$$

Ainsi, le test basé sur la règle de décision suivante :

"On rejette $\mathcal{H}_0$ si $C_n < u_\alpha$."

est un test de niveau α.

(5) **A.N.** On a $\alpha = 0,05$. On voit sur la table de la loi du chi-deux que $u_\alpha = 1,635$ vérifie (6).

De plus, on a $n = 7$, la moyenne empirique des observations est $\overline{x}_n \approx 54,06$ et la variance empirique non biaisée est $v_n \approx 0.7$, donc $c_n \approx 1,4$. Ainsi, $c_n < u_\alpha$ et donc on rejette l'hypothèse nulle. On peut conclure que Pamela respecte la limite de variance de manière significative. Cruella ne va pas être contente !

(6) On cherche la probabilité de rejeter $\mathcal{H}_0$ lorsque la vraie variance est $v^2 = 2,3$ kg². Avec les notations de la question précédent, on veut calculer

$$\mathbb{P}_{v^2=2,3}\left(C_n < u_\alpha\right)$$

On rappelle que

$$C_n = (n-1)\frac{\hat{v}_n^2}{v_0^2} = (n-1)\frac{\hat{v}_n^2}{v^2}\frac{v^2}{v_0^2}.$$

Or, $Z_n = (n-1)\frac{\hat{v}_n^2}{v^2}$ suit une loi du chi-deux à $n-1$ degrés de liberté. Donc,

$$\mathbb{P}_{v^2=2,3}\left(C_n < u_\alpha\right) = \mathbb{P}_{v^2=2}\left(Z_n < u_\alpha \frac{v_0^2}{v^2}\right).$$

A.N. On a $v_0^2/v^2 = 3/(2,3)$, donc $u_\alpha \frac{v_0^2}{v^2} \approx 2,13$. Ainsi, la puissance du test au point $v^2 = 2,3$ kg² est inférieure à $1/10$.

Correction 5.
Pour $i = 1, \ldots, n_1$, on note X_i la variable aléatoire représentant la résistance du i-ème sac du premier contrôle. On notera m_X et σ^2 respectivement la moyenne et la variance des X_i. On note $\overline{X}_{n_1}$ la moyenne empirique des X_i.

Pour $j = 1, \ldots, n_2$, on note Y_j la variable aléatoire représentant la résistance du j-ème sac du second contrôle. On notera m_Y et τ^2 respectivement la moyenne et la variance des Y_j. On note $\overline{Y}_{n_2}$ la moyenne empirique des Y_j.

(1) On choisit pour hypothèse nulle, $\mathcal{H}_0$: "$\sigma^2 = \tau^2$" et pour alternative $\mathcal{H}_1$: "$\sigma^2 \neq \tau^2$".

Pour tester l'égalité des variances, on utilise la statistique de Fisher,

$$F = \frac{\hat{\sigma}_{n_1}^2}{\hat{\tau}_{n_2}^2}.$$

On souhaite rejeter l'hypothèse nulle lorsque $\hat{\sigma}^2_{n_1}$ et $\hat{\tau}^2_{n_2}$ sont trop différents l'un de l'autre, i.e. quand F est trop différent de 1. Donc, on utilisera une zone de rejet de la forme $]-\infty, A[\cup]B, +\infty[$.

Soit $\alpha \in]0,1[$. Il faut commencer par déterminer un couple (u_α, v_α) tel que

$$\mathbb{P}_{\mathcal{H}_0}\left(F < u_\alpha \text{ ou } F > v_\alpha\right) \leq \alpha. \tag{7}$$

Or, sous $\mathcal{H}_0$, $F \sim \mathcal{F}(n_1-1, n_2-1)$, donc si un couple (u_α, v_α) est tel que

$$\begin{cases} \mathbb{P}\left(\mathcal{F}(n_1-1, n_2-1) < u_\alpha\right) \leq \alpha/2, \\ \mathbb{P}\left(\mathcal{F}(n_1-1, n_2-1) > v_\alpha\right) \leq \alpha/2, \end{cases} \tag{8}$$

alors il convient (i.e. il vérifie l'équation (7)). Ainsi, le test basé sur la règle de décision suivante :

"On rejette $\mathcal{H}_0$ si $F < u_\alpha$ ou si $F > v_\alpha$."

est un test de niveau α.

A.N. On a $\alpha = 0,05$. On voit sur la table de la loi de Fisher que le couple $(u_\alpha, v_\alpha) = (0,36; 2,57)$ vérifie (8).

De plus, on a $f = 0,51$. Ainsi, $u_\alpha \leq f \leq v_\alpha$ et donc on accepte l'hypothèse nulle.

(2) On suppose désormais que les variances sont égales. On note σ^2 leur valeur commune. On choisit de tester $\mathcal{H}_0$: "$m_X = m_Y$" contre $\mathcal{H}_1$: "$m_X \neq m_Y$". On estime σ^2 par

$$\hat{\sigma}^2 = \hat{\sigma}^2_{n_1,n_2} := \frac{(n_1-1)\hat{\sigma}^2_{n_1} + (n_2-1)\hat{\tau}^2_{n_2}}{n_1+n_2-2}.$$

Par un résultat du cours, on a que la variable définie par

$$Z = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \frac{(\overline{X}_{n_1} - m_X) - (\overline{Y}_{n_2} - m_Y)}{\hat{\sigma}}$$

suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté. On utilisera la statistique

$$T = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\hat{\sigma}}.$$

On souhaite rejeter l'hypothèse nulle lorsque $\overline{X}_{n_1}$ et $\overline{Y}_{n_2}$ sont trop différents l'un de l'autre, i.e. quand T est trop différent de 0. Donc, on utilisera une zone de rejet de la forme $]-\infty, A[\cup]B, +\infty[$.

Soit $\alpha \in]0,1[$. Il faut commencer par déterminer un couple (u_α, v_α) tel que

$$\mathbb{P}_{\mathcal{H}_0}\left(T < u_\alpha \text{ ou } T > v_\alpha\right) \leq \alpha. \tag{9}$$

Or, sous $\mathcal{H}_0$, $T \sim \mathcal{T}(n_1+n_2-2)$, donc si un couple (u_α, v_α) est tel que

$$\begin{cases} \mathbb{P}\left(\mathcal{T}(n_1+n_2-2) < u_\alpha\right) \leq \alpha/2, \\ \mathbb{P}\left(\mathcal{T}(n_1+n_2-2) > v_\alpha\right) \leq \alpha/2, \end{cases} \tag{10}$$

alors il convient (i.e. il vérifie l'équation (9)). Ainsi, le test basé sur la règle de décision suivante :

"On rejette $\mathcal{H}_0$ si $T_n < u_\alpha$ ou si $T_n > v_\alpha$."

est un test de niveau α.

(3) **A.N.** On a $\alpha = 0,05$. On voit sur la table de la loi de Student que le couple $(u_\alpha, v_\alpha) = (-2,042; 2,042)$ vérifie (10).

De plus, on a $t = 2,03$. Ainsi, $u_\alpha \leq t_n \leq v_\alpha$ et donc on accepte l'hypothèse nulle.

Correction 6. Le principe du test du khi-deux d'adéquation est de confronter les effectifs observés n_i pour $i \in \{0, \ldots, 4\}$ avec les effectifs théoriques $n'_i = n \times \mathbb{P}(X = i)$ obtenus si $X \sim \mathcal{B}(4, 1/2)$ et où $n = 100$ est la taille de l'échantillon. On résume ceci via le tableau suivant :

i	0	1	2	3	4
n_i	7	20	41	22	10
$\mathbb{P}(X = i)$	0.0625	0.25	0.375	0.25	0.0625
n'_i	6.25	25	37.5	25	6.25

Pour mesurer la distance entre les effectifs observés et théoriques, on utilise la variable

$$\delta = \sum_{i=0}^{4} \frac{(n_i - n'_i)^2}{n'_i} \approx 4.03.$$

Les n'_i sont bien plus grands que 5 donc on peut appliquer l'approximation : comme i prend $k = 5$ valeurs, on considère que δ est la réalisation d'une variable aléatoire de loi du khi-deux à $k - 1 = 4$ degrés de liberté. On souhaite rejeter l'hypothèse d'adéquation lorsque la distance δ est significativement trop grande.

Ainsi, pour un niveau de risque de 5%, on prend t tel que $\mathbb{P}\left(\chi^2(4) > t\right) = 5\%$, c'est à dire $t = 9.488$ en lisant sur la table de la loi du khi-deux. La distance δ ne dépasse pas t, donc on peut conserver l'hypothèse selon laquelle le nombre de garçons est distribué selon la loi $\mathcal{B}(4, 1/2)$.