Correction de l'examen de décembre 2019.

Exercice 1. (*en tout, sur* 14.5 *points*).

1. *(2 points : 0.5 point pour l'explication et l'unité de chaque quantité $m, \sigma^2, \overline{Y}, S^2$).*
   Le paramètre m est le périmètre crânien moyen des nouveau-nés masculins en région parisienne, et le paramètre σ est l'écart-type des périmètres crâniens des nouveau-nés masculins en région parisienne. L'unité de m et de σ est donc le centimètre. La quantité σ^2 est la variance (carré de l'écart-type) des périmètres crâniens des nouveau-nés masculins en région parisienne, elle s'exprime en centimètres carrés.

   La variable $\overline{Y}$ est la moyenne empirique des périmètres crâniens de 1225 nouveau-nés masculins observés en région parisienne, et est donc en centimètres. La variable S^2 est la variance empirique du périmètre crânien d'un nouveau-né masculin de la région parisienne, sur 1225 observations ; elle s'obtient en prenant la moyenne (divisée par 1224 au lieu de 1225) des carrés des écarts entre le périmètre crânien Y_i d'un nouveau né, et la moyenne empirique $\overline{Y}$. L'unité de S^2 est le centimètre carré.

2. *(total sur 6 points).* La donnée de référence pour les périmètres crâniens des nouveau-nés français en général est 35 cm, et l'on souhaite comparer m, espérance de notre échantillon, à cette valeur $m_0 := 35$. On se retrouve dans le cadre d'un modèle gaussien avec moyenne à estimer et variance inconnue, donc on effectue un test de Student :

   (1) *(0.5 point).* **Modèle :** on a un 1225-échantillon $Y_1, \ldots, Y_{1225}$ de la loi normale $\mathcal{N}(m, \sigma^2)$.

   (2) *(0.5 point).* **Hypothèses :** on souhaite tester si la moyenne m des périmètres crâniens des enfants en région parisienne est proche de la moyenne nationale m_0, ou significativement différente. On va donc tester "$m = m_0$" contre "$m \neq m_0$". Ainsi,

   — H_0 : "$m = m_0 = 35$" ;

   — H_1 : "$m \neq m_0$".

   (3) *(1 point).* **Statistique :** on travaille avec un modèle gaussien dont on ne connaît ni la moyenne, ni la variance, et pour lequel on souhaite tester la moyenne. On va donc choisir comme statistique la variable aléatoire :

   $$Z := \frac{\sqrt{1225}\,(\overline{Y} - 35)}{S}$$

   où S^2 est la variance empirique donnée dans l'énoncé. C'est un choix convenable car Z prend des valeurs significativement différentes sous H_0 et sous H_1. De plus, on sait que sous H_0, Z suit une loi de Student $\mathcal{T}(1224)$, qui est approchable par une loi normale centrée réduite $\mathcal{N}(0,1)$ car on a un très grand nombre de degrés de liberté.

   (4) *(0.5 point).* **Règle de décision :** puisque Z suit sous H_0 une loi gaussienne centrée réduite, qui est en particulier symétrique et d'espérance nulle, on va rejeter H_0 si la valeur observée n'est pas dans un intervalle donné de la forme $[-a, a]$ avec a valeur seuil à calculer.

   (5) *(1 point).* **Seuil :** par défaut, on effectue un test de niveau $\alpha = 5\%$. Cela signifie que l'on cherche

à avoir :

$$\begin{aligned}\mathbb{P}_{\mathrm{H}_0}(Z \notin [-a,a]) = 0.05 &\iff \mathbb{P}_{\mathrm{H}_0}(Z \in [-a,a]) = 0,95\\ &\iff F_{\mathcal{N}(0,1)}(a) - F_{\mathcal{N}(0,1)}(-a) = 0.95\\ &\iff 2\,F_{\mathcal{N}(0,1)}(a) - 1 = 0.95\\ &\iff F_{\mathcal{N}(0,1)}(a) = 0.975,\end{aligned}$$

ce qui donne d'après les tables de valeurs numériques : $a = 1.96$. On rejette donc H_0 si $Z_{\mathrm{obs}} \notin [-1.96, 1.96]$, et on conserve H_0 si $Z_{\mathrm{obs}} \in [-1.96, 1.96]$.

(6) *(*0.5 *point pour le calcul de la statistique,* 0.5 *point pour la comparaison aux seuils,* 1 *point pour le calcul de la p-valeur).* **Décision :** sachant que $\sqrt{1225} = 35$, on a sous l'hypothèse H_0 :

$$Z_{\mathrm{obs}} := \frac{35\,(\overline{Y}_{\mathrm{obs}} - 35)}{S_{\mathrm{obs}}} = 0.9504.$$

$\rightsquigarrow$ **Décision par comparaison aux seuils :** on a $Z_{\mathrm{obs}} \in [-1.96, 1.96]$, donc on garde l'hypothèse H_0.

$\rightsquigarrow$ **Décision par calcul de la p-valeur :** le niveau α du test était donné par l'équation $\alpha = 2\,(1 - F_{\mathcal{N}(0,1)(a)})$, où a était le seuil du test. On obtient la p-valeur α_{obs} en remplaçant dans cette équation le seuil par la valeur observée Z_{obs} :

$$\alpha_{\mathrm{obs}} = 2\,(1 - F_{\mathcal{N}(0,1)}(0.9504)) \simeq 0.3422.$$

On a $\alpha_{\mathrm{obs}} > \alpha$, donc on conserve effectivement H_0.

(7) *(*0.5 *point).* **Phrase de conclusion :** la moyenne des périmètres crâniens des nouveau-nés masculins en région parisienne n'est pas significativement différente de celle des autres nouveau-nés français.

3. *(total sur* 5.5 *point).* Cette fois, on souhaite comparer σ, écart-type de notre échantillon, à la valeur $\sigma_0 := 1.35$. Par rapport à la question précédente, on est de nouveau avec un modèle gaussien à moyenne et variance inconnue, mais cette fois-ci c'est la variance (ou l'écart-type, qui en est sa racine carrée) que l'on souhaite estimer. On s'oriente donc vers un test du chi-deux sur la variance, dont voici les 7 points :

   (1) *(*0.5 *point pour le modèle et les hypothèses).* **Modèle :** on a toujours un 1225-échantillon $Y_1, \ldots, Y_{1225}$ de la loi normale $\mathcal{N}(m, \sigma^2)$.

   (2) **Hypothèses :** on souhaite tester si l'écart-type des variables observées est proche de σ_0. On va donc tester "$\sigma = \sigma_0$" contre "$\sigma \neq \sigma_0$". Ainsi :
   — H_0 : "$\sigma = \sigma_0 = 1.35$" ;
   — H_1 : "$\sigma \neq \sigma_0$".

   (3) *(*1 *point).* **Statistique :** on travaille avec un modèle gaussien dont on ne connaît ni la moyenne, ni la variance, et pour lequel on souhaite tester la variance σ^2. On va donc choisir comme statistique la variable aléatoire :

$$Z' := \frac{1224\,S^2}{\sigma_0^2} = \frac{1224\,S^2}{(1.35)^2}$$

   où S^2 est la variance empirique donnée dans l'énoncé. C'est un choix convenable car Z' prend des valeurs significativement différentes sous H_0 et sous H_1. De plus, on sait que sous H_0, Z' suit une loi du chi-deux $\chi^2(1224)$, dont les quantiles sont donnés dans l'énoncé.

(4) *(0.5 point).* **Règle de décision :** d'après la formule la définissant, Z' va prendre des valeurs plus petites (respectivement plus grandes) lorsque σ sera plus petit (respectivement plus grand) que σ_0. On va donc rejeter H_0 lorsque la valeur observée Z'_{obs} de Z' ne sera pas dans un intervalle donné $[a, b]$, où a et b sont des seuils à calculer.

(5) *(1 point).* **Seuil :** par défaut, on effectue un test de niveau $\alpha = 5\%$. Cela signifie que l'on cherche à avoir :

$$\begin{aligned}\mathbb{P}_{\mathrm{H}_0}(Z' \notin [a,b]) = 0.05 &\iff \mathbb{P}_{\mathrm{H}_0}(Z' \in [a,b]) = 0.95\\ &\iff F_{\chi^2(1224)}(b) - F_{\chi^2(1224)}(a) = 0.95.\end{aligned}$$

Comme on n'a pas d'indication précise sur le comportement de σ sous l'hypothèse alternative H_1, on est amené à rejeter symétriquement des valeurs trop grandes (sur une zone de probabilité 2.5% sous H_0) et des valeurs trop petites (également sur une zone de probabilité 2.5%). On choisit donc a et b tels que $F_{\chi^2(1224)}(a) = 0.025$ et $F_{\chi^2(1224)}(b) = 0.975$; ceci donne d'après les tables de valeurs numériques $a = 1128.9$ et $b = 1322.9$. On rejette donc H_0 si $Z'_{\mathrm{obs}} \notin [1128.9, 1322.9]$, et on conserve H_0 si $Z'_{\mathrm{obs}} \in [1128.9, 1322.9]$.

(6) *(0.5 point pour le calcul de la statistique, 0.5 point pour la comparaison aux seuils, 1 point pour le calcul de la p-valeur).* **Décision :** on calcule

$$Z'_{\mathrm{obs}} := \frac{1224\, S^2_{\mathrm{obs}}}{(1.35)^2} = 1457.38.$$

$\rightsquigarrow$ **Décision par comparaison aux seuils :** on a $Z'_{\mathrm{obs}} \notin [1128.9, 1322.9]$: on rejette donc l'hypothèse H_0, ce qui signifique que l'écart-type σ des périmètres crâniens des nouveau-nés parisiens diffère assez significativement de celui des nouveau-nés français en général. Plus précisément, et puisque S^2 est un estimateur de σ^2, on obtient que l'écart-type est plus grand à Paris qu'ailleurs en général (on a une population plus hétérogène).

$\rightsquigarrow$ **Décision par calcul du niveau observé :** comme on a un test bilatéral, on doit avoir :

$$\alpha_{\mathrm{obs}} = 2\,(1 - F_{\chi^2(1224)}(1457.38)) < 0.001.$$

En effet, $F_{\chi^2(1224)}$ est une fonction croissante, et 1457.38 est plus grand que toutes les valeurs du tableau donné dans l'énoncé.
On a donc $\alpha_{\mathrm{obs}} < \alpha$, ce qui justifie que l'on rejette effectivement H_0.

(7) *(0.5 point).***Phrase de conclusion :** les nouveau-nés masculins en région parisienne sont significativement différents des autres nouveau-nés français en ce qui concerne la variation de leur périmètre crânien. Le rejet à droite nous indique que les nouveau-nés masculins en région parisienne ont une variabilité de leur périmètre crânien plus grande que celle des autres nouveau-nés français.

4. *(1 point).* On ne peut pas dire que la moyenne des périmètres crâniens soit différente en région parisienne et en France ; par contre, la variation moyenne des périmètres crâniens est plus grande en région parisienne qu'en France. Ce dernier point peut s'expliquer par le fait que la région parisienne est une zone de forte hétérogénéité.

Exercice 2. *(en tout, sur* 7.5 *points).*

1. *(*1 *point).* Si les proportions d'enfants marchant (M), en ébauche de marche (E) et en absence de marche (A) sont les mêmes pour l'échantillon considéré (les 120 bébés prématurés) et pour la population générale, alors les trois variables aléatoires

$$\begin{aligned} N_M &= \text{nombre d'enfants marchant dans le 120-échantillon;} \\ N_E &= \text{nombre d'enfants en ébauche de marche dans le 120-échantillon;} \\ N_A &= \text{nombre d'enfants ne marchant pas dans le 120-échantillon} \end{aligned}$$

suivent respectivement des lois binomiales $\mathcal{B}(120, 0.5)$, $\mathcal{B}(120, 0.12)$, $\mathcal{B}(120, 0.38)$ (ces trois variables ne sont pas indépendantes, mais ça ne change pas le calcul des espérances). On attend donc les moyennes suivantes :

$$\mathbb{E}[N_M] = 120 * 0.5 = 60 \qquad ; \qquad \mathbb{E}[N_E] = 120 * 0.12 = 14.4 \qquad ; \qquad \mathbb{E}[N_A] = 120 * 0.38 = 45.6.$$

Les effectifs attendus pour chaque catégorie et dans l'hypothèse neutre "les bébés prématurés marchent de la même façon que les autres" sont donc $n_M = 60$, $n_E = 14.4$ et $n_A = 45.6$.

2. *(total sur* 6.5 *points).* On cherche à comparer une caractéristique d'une population donnée (la marche des bébés prématurés) à celle de la population générale. On va réaliser un test du χ^2 d'adéquation. La procédure de test est rédigée en 7 points :

   (1) *(*1 *point).* **Modèle :** on note X_i la variable aléatoire représentant le niveau de marche (M, E ou A) du i-ème bébé prématuré observé. Les variables $X_1, \ldots, X_{120}$ forment un 120-échantillon de variables aléatoires indépendantes et identiquement distribuées selon une loi discrète $\mathcal{L} = \text{Disc}(p_M, p_E, p_A)$ donnée par

$$\mathbb{P}(X_1 = M) = p_M \qquad ; \qquad \mathbb{P}(X_1 = E) = p_E \qquad ; \qquad \mathbb{P}(X_1 = A) = p_A.$$

   (2) *(*0.5 *point).* **Hypothèses :** l'objectif du test est de déterminer si les bébés prématurés marchent de la même manière que les autres bébés, c'est-à-dire si $(p_M, p_E, p_A) = (0.5, 0.12, 0.38)$. Les hypothèses testées sont donc :
   - H_0 : "$p_M = 0.5$ et $p_E = 0.12$ et $p_A = 0.38$" ;
   - H_1 : "$p_M \neq 0.5$ ou $p_E \neq 0.12$ ou $p_A \neq 0.38$".

   (3) *(*1.5 *point, dont* 0.5 *point pour la vérification des conditions de validité du test).* **Statistique de test :** comme $n_M, n_E, n_A \geq 5$, les conditions sont réunies pour réaliser un test du chi-deux. La statistique de test choisie est

$$Z = \frac{(N_M - n_M)^2}{n_M} + \frac{(N_E - n_E)^2}{n_E} + \frac{(N_A - n_A)^2}{n_A} = \frac{(N_M - 60)^2}{60} + \frac{(N_E - 14.4)^2}{14.4} + \frac{(N_A - 45.6)^2}{45.6}.$$

   Sous H_0, Z suit approximativement une loi $\chi^2(3 - 1 = 2)$ tandis que, sous H_1, Z a tendance à prendre des valeurs plus grandes.

   (4) *(*0.5 *point).* **Région de rejet :** la zone région de rejet pour un test d'adéquation est toujours de la forme $R_t = [t, +\infty[$. On rejettera donc H_0 si $Z_{\text{obs}} \geq t$.

   (5) *(*0.5 *point).* **Calcul du seuil :** pour un seuil $\alpha = 5\%$, on a

$$\alpha = 0.05 = P_{\mathrm{H}_0}(Z \geq t) = 1 - P_{\mathrm{H}_0}(Z < t) \simeq 1 - F_{\chi^2(2)}(t).$$

   Sur les tables on trouve $t = 5.9915$, donc $R_t = [5.9915, +\infty[$.

(6) *(0.5 point pour le calcul de la statistique, 0.5 point pour la comparaison aux seuils, 1 point pour le calcul de la p-valeur).* **Décision :** $N_{M,\text{obs}} = 54$, $N_{E,\text{obs}} = 8$ et $N_{A,\text{obs}} = 58$, donc la valeur de la statistique de test est $Z_{\text{obs}} = 6.8164$.

$\rightsquigarrow$ **Décision par comparaison au seuil :** $Z_{\text{obs}} = 6.8164 > 5.9915$, on rejette donc H_0 au niveau 5%.

$\rightsquigarrow$ **Décision par calcul du niveau observé :** on calcule

$$\alpha_{\text{obs}} = P_{\text{H}_0}(Z > Z_{\text{obs}}) = 1 - F_{\chi^2(2)}(6.8164).$$

Comme $5.9915 < 6.8164 < 7.3778$, les tables permettent d'obtenir l'encadrement

$$2.5\% < \alpha_{\text{obs}} < 5\%.$$

On a donc moins de 5% de chances de se tromper en rejetant H_0.

(7) *(0.5 point).* **Phrase de conclusion :** les bébés prématurés ont donc un apprentissage de la marche signifactivement différent de celui des bébés en général. Plus précisément, on observe des valeurs N_M et N_E inférieures aux valeurs attendues n_M et n_E, et à l'inverse, une valeur N_A nettement supérieure à n_A ($N_A = 58 > 45.6 = n_A$). On a donc mis en évidence par le test une tendance à un apprentissage plus tardif de la marche chez les bébés prématurés.