Deux tests de comparaison de moyennes : exemples de rédaction

Exercice. Un groupe d'ornithologues étudie les canards du fleuve jaune. Chaque saison, ils capturent des canards, mesurent un certain nombre de paramètres, puis les mangent. Le tableau suivant indique les poids (en kilogrammes) des canards qui ont été capturés deux années consécutives, appelées 1 et 2.

Année 1	2.22	2.16	2.49	2.11	2.17	2.13	2.06	2.21	2.20	2.47	2.20	2.08
Année 2	2.52	2.31	2.11	2.73	2.32	2.38	2.30	2.41	2.40	2.15		

Un autre groupe d'ornithologues étudie quant à lui les oies sauvages du Pô. Chaque saison, ils capturent des oies, mesurent un certain nombre de paramètres, puis les relâchent après les avoir baguées. D'une saison sur l'autre, il leur arrive de recapturer des volatiles qu'ils avaient pu observer l'année précédente. Le tableau suivant indique les poids (en kilogrammes) des oies qui ont pu être observées deux années consécutives, appelées I et II.

Année I	4.78	4.85	4.09	5.00	4.90	4.86	5.37	4.68	5.28	5.11
Année II	4.99	5.05	4.53	4.97	5.48	5.03	5.01	4.78	4.99	5.07

- (a) Expliquer brièvement pourquoi les modèles statistiques de ces deux études sont profondément différents.
- (b) **Étude des canards du fleuve jaune :**
- 1) Vérifier que la variabilité du poids des canards ne dépend pas de l'année (test d'égalité des variances).
 - 2) Le poids moyen des canards du fleuve jaune a-t-il varié entre l'année 1 et l'année 2 ?
- (c) **Étude des oies du Pô :** Le poids moyen des oies du Pô a-t-il varié entre l'année I et l'année II ?

Correction.

- (a) L'équipe du fleuve Jaune étudie d'une année sur l'autre des oiseaux différents alors que l'équipe du Pô étudie les mêmes oiseaux. Dans le premier cas, les échantillons pour année 1 et année 2 sont indépendants, alors que dans le second cas, les échantillons pour année I et année II sont appariés.
- (b) **Étude des canards du fleuve jaune :**
- 1) Il s'agit de faire un test d'égalité des variances.

Modèle : On dispose de deux échantillons indépendants, celui des poids des canards capturés en année 1 et celui des poids de canards capturés en année 2. Soit X la variable aléatoire représentant le poids d'un canard capturé en année 1. On a un 12-échantillon ($n = 12$) indépendant $X_1, \dots, X_{12}$. On suppose que $X_i \sim N(m_X, \sigma_X^2)$. Soit Y la variable aléatoire représentant le poids d'un canard capturé en année 2. On a un 10-échantillon ($n_Y = 10$) indépendant $Y_1, \dots, Y_{10}$. On suppose que $Y_j \sim N(m_Y, \sigma_Y^2)$. On veut savoir si $\sigma_X = \sigma_Y$.

Hypothèses : On peut tester H_0 : la variabilité du poids des canards ne dépend pas de l'année contre H_1 : la variabilité du poids des canards dépend de l'année. Ces hypothèses se traduisent mathématiquement par : $H_0 : \sigma_X = \sigma_Y$ contre $H_1 : \sigma_X \neq \sigma_Y$.

Statistique de test : La statistique est

$$Z = \frac{S_X^2}{S_Y^2},$$

où $S_X^2 = \frac{1}{12-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ et $S_Y^2 = \frac{1}{10-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2$.
Sous H_0 , on a

$$Z = \frac{S_X^2}{S_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathcal{F}(10, 12).$$

Règle de décision : On remarque que sous H_1 , Z a tendance à prendre de grandes valeurs (si $\sigma_X > \sigma_Y$) ou bien de petites valeurs (si $\sigma_X < \sigma_Y$). On rejettera donc H_0 si $Z_{obs} > b$ ou $Z_{obs} < a$. Autrement dit, $R_{rejet} = [0, a[\cup]b, \infty[$ et $R_{accept} = [a, b]$.

Calcul des valeurs seuils : On effectue un test de niveau $\alpha = 5\%$. On cherche à déterminer a et b tels que :

$$P_{H_0}(Z \geq b) = \alpha/2$$

$$1 - P_{H_0}(Z < b) = 0.025$$

$$P_{H_0}(Z \leq b) = 0.975$$

$$F_{\mathcal{F}(11,9)}(b) = 0.975$$

$$b = F_{\mathcal{F}(11,9)}^{-1}(0.975) = \text{qf}(0.975, 11, 9) = 3.912074,$$

et de même,

$$P_{H_0}(Z \leq a) = \alpha/2$$

$$F_{\mathcal{F}(11,9)}(a) = 0.025$$

$$a = F_{\mathcal{F}(12,10)}^{-1}(0.025) = \text{qf}(0.025, 11, 9) = 0.2787147.$$

Décision : On calcule

$$S_{X,obs}^2 = \frac{1}{12-1} \sum_{i=1}^{12} (X_{i,obs} - \bar{X}_{obs})^2 = 0.0187$$

$$S_{Y,obs}^2 = \frac{1}{10-1} \sum_{j=1}^{10} (Y_{j,obs} - \bar{Y}_{obs})^2 = 0.0312.$$

$Z_{obs} = 0.5998429 \in [0.2787147, 3.912074] = R_{accept}$, donc on ne rejette pas H_0 : au niveau 5%, on peut affirmer que la variabilité du poids des canards ne dépend pas de l'année.

p -valeur : On cherche à calculer la p -valeur α_{obs} . Celle-ci vérifie

$$\begin{aligned} P_{H_0}(Z \leq Z_{obs}) &= \alpha_{obs}/2 \\ F_{\mathcal{F}(11,9)}(0.5998429) &= \alpha_{obs}/2 \\ \alpha_{obs} &= 2F_{\mathcal{F}(11,9)}(0.5998429) = 2 * \text{pf}(0.5998429, 11, 9) = 0.4190429. \end{aligned}$$

2) Il s'agit d'un test de Student de comparaison des deux moyennes.

Modèle : En se basant sur 1), on peut désormais supposer que $\sigma_X = \sigma_Y = \sigma$.
On a donc le même modèle qu'avant, avec de plus $X_i \sim N(m_X, \sigma^2)$ et $Y_j \sim N(m_Y, \sigma^2)$.

Hypothèses : On veut tester H_0 : le poids moyen des canards du fleuve jaune n'a pas varié contre H_1 : le poids moyen des canards du fleuve jaune a varié. Ces hypothèses se traduisent mathématiquement par : $H_0 : m_X = m_Y$ contre $H_1 : m_X \neq m_Y$.

Statistique de test : La statistique est

$$Z = \frac{\bar{X}_{12} - \bar{Y}_{10}}{S \sqrt{\frac{1}{12} + \frac{1}{10}}},$$

où

$$S^2 = \frac{1}{12 + 10 - 2} \left(\sum_{i=1}^{12} (X_i - \bar{X}_{12})^2 + \sum_{j=1}^{10} (Y_j - \bar{Y}_{10})^2 \right).$$

Sous H_0 , $Z \sim \mathcal{T}(12 + 10 - 2 = 20)$.

Règle de décision : On remarque que sous H_1 , $|Z| \xrightarrow[n \rightarrow \infty]{} \infty$, donc Z a tendance à prendre de grandes valeurs. On rejettera donc H_0 si $|Z_{obs}| \geq a$. Autrement dit, $R_{rejet} =] - \infty, -a[\cup] a, \infty[$ et $R_{accept} = [-a, a]$.

Calcul des valeurs seuil : On effectue un test de niveau $\alpha = 5\%$. On cherche donc à déterminer a tel que :

$$\begin{aligned} \alpha &= P_{H_0}(|Z| \geq a) \\ 1 - \alpha/2 &= P_{H_0}(Z \leq a) \\ P_{H_0}(Z \leq a) &= 0.975 \\ F_{\mathcal{T}(20)}(a) &= 0.975 \\ a &= F_{\mathcal{T}(20)}^{-1}(0.975) = \text{qt}(0.975, 20) = 2.085963. \end{aligned}$$

Décision : On calcule $Z_{obs} = -2.314 \in R_{rejet}$, donc on rejette H_0 : au niveau 5%, on peut affirmer que le poids moyen des canards du fleuve jaune a varié entre l'année 1 et l'année 2.

p-valeur : D'après le calcul précédent,

$$\begin{aligned}\alpha_{obs} &= P_{H_0}(|Z| \geq |Z_{obs}|) \\ &= 2(1 - P_{H_0}(Z \leq |Z_{obs}|)) \\ &= 2(1 - F_{\mathcal{T}(20)}(2.314)) \\ &= 2 * (1 - \text{pt}(2.314, 20)) = 0.03142125.\end{aligned}$$

(c) **Étude des oies du Pô :**

Modèle : Comme les échantillons sont appariés, on introduit un échantillon $X_1, \dots, X_{10}$ obtenu en soustrayant les données de l'année II à celles de l'année I. On modélise par $X_i \sim N(m, \sigma^2)$, avec m et σ inconnus. Les valeurs observées $X_{1,obs}, \dots, X_{10,obs}$ sont les suivantes :

-0,21	-0.2	-0,44	0,03	-0,58	-0,17	0,36	-0,1	0,29	0,04
-------	------	-------	------	-------	-------	------	------	------	------

Hypothèses : On teste H_0 : le poids moyen des oies du Pô n'a pas varié contre H_1 : le poids moyen des oies du Pô a varié, c'est-à-dire que l'on teste $H_0 : m = 0$ contre $H_1 : m \neq 0$.

Statistique de test : La statistique est

$$Z = \frac{\sqrt{10}\bar{X}_{10}}{S},$$

avec $S^2 = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X}_{10})^2$. Sous H_0 , $Z \sim \mathcal{T}(10 - 1)$.

Règle de décision : On prend $R_{rejet} =] - \infty, -a[\cup] a, \infty[$ et $R_{accept} = [-a, a]$.

Calcul des valeurs seuil : On effectue un test de niveau $\alpha = 5\%$. On cherche donc à déterminer a tel que :

$$\begin{aligned}\alpha &= P_{H_0}(|Z| \geq a) \\ 1 - \alpha &= 2P_{H_0}(Z \leq a) - 1 \\ P_{H_0}(Z \leq a) &= 0.975 \\ F_{\mathcal{T}(9)}(a) &= 0.975 \\ a &= F_{\mathcal{T}(9)}^{-1}(0.975) = \text{qt}(0.975, 9) = 2.262.\end{aligned}$$

Décision : On calcule $S_{obs} = 0.293$ puis $Z_{obs} = -1.057$. On a $Z_{obs} \in R_{accept}$. Au niveau 5%, le poids moyen des oies du Pô n'a pas varié entre l'année I et l'année II.

p-valeur : On a

$$\begin{aligned}\alpha_{obs} &= 2(1 - P_{H_0}(Z \leq 1.057)) \\ &= 2 * (1 - \text{pt}(1.057, 9)) = 0.318,\end{aligned}$$

ce qui confirme (bien sûr !) que l'on ne rejette pas H_0 au niveau 5%.