Comparaison d'échantillons et tests multiples

Solenne Gaucher

Test simple

Les tests de comparaison d'échantillons sont utilisés pour détecter l'impact d'un traitement sur un ensemble de grandeurs d'intérêt. Par exemple, des scientifiques étudient en laboratoire l'influence de la consommation de maïs génétiquement modifié Mon863 sur le poids de rats. Dans ce cadre, ils nourrissent un groupe de rats avec un régime à base de maïs Mon863, et un groupe témoin de rats avec un régime à base de maïs non modifié génétiquement. Ils cherchent ensuite à tester l'hypothèse "les rats des deux groupes ont le même poids". Dans cette partie, on travaillera avec le jeu de données "Ratweight.csv".

1 - Variance connue

On observe des variables aléatoires $X_1^{(1)}, ..., X_{n_1}^{(1)}$ et $X_1^{(2)}, ..., X_{n_2}^{(2)}$ indépendantes, de loi respective $\mathcal{N}\left(\mu_1, \sigma^2\right)$ et $\mathcal{N}\left(\mu_2, \sigma^2\right)$. Dans un premier temps, on suppose la variance σ^2 connue. On souhaite tester l'hypothèse

$$\begin{aligned} \mathcal{H}_0 &: \mu_1 = \mu_2 \\ \text{contre } \mathcal{H}_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

1 - a) On note $\overline{X^{(1)}}$ la moyenne de l'échantillon $X^{(1)} \triangleq \left(X_1^{(1)}, ..., X_{n_1}^{(1)}\right)$ et $\overline{X^{(2)}}$ la moyenne de l'échantillon $X^{(2)} \triangleq \left(X_1^{(2)}, ..., X_{n_2}^{(2)}\right)$. Sous l'hypothèse $\mathcal{H}_0$, quelle est la loi de $\overline{X^{(1)}} - \overline{X^{(2)}}$?

1 - b) Déterminer un test de niveau α pour $\mathcal{H}_0$.

2 - Variance inconnue

Dans la pratique, la variance des grandeurs mesurées est inconnue et il faut l'estimer. On note $\widehat{\sigma_1^2}$ l'estimateur sans biais de la variance σ^2 obtenu à partir de l'échantillon $X^{(1)}$ et $\widehat{\sigma_2^2}$ l'estimateur sans biais de σ^2 obtenu à partir de l'échantillon $X^{(2)}$.

2 - a) Rappeler l'expression de $\widehat{\sigma_1^2}$ et $\widehat{\sigma_2^2}$. Déterminer $\widehat{\sigma^2}$ le meilleur estimateur de la variance de l'échantillon $X \triangleq \left(X_1^{(1)}, ..., X_{n_1}^{(1)}, X_1^{(2)}, ..., X_{n_2}^{(2)}\right)$ parmi les estimateurs de la forme $\lambda\widehat{\sigma_1^2} + (1-\lambda)\widehat{\sigma_2^2}$ où $\lambda \in [0,1]$, et donner son expression en fonction de $X^{(1)}, X^{(2)}$. Montrer qu'il est sans biais.

2 - b) On rappelle le Théorème de Cochran :

Théorème (Cochran)**.** *Soit X un vecteur aléatoire de $\mathbb{R}^n$ de loi $\mathcal{N}(0_n, Id_n)$, F un sous espace de $\mathbb{R}^n$, F^T son orthogonal et P_F, P_{F^T} les matrices des projections orthogonales sur F et F^T. Alors $P_F X$, $P_{F^T} X$ sont indépendants et $\|P_F X\|^2$ suit une loi du χ^2 à dim(F) degrés de liberté.*

On note u le vecteur de $\mathbb{R}^{n_1+n_2}$ tel que $u_i = \frac{1}{\sqrt{n_1}} 1_{1 \leq i \leq n_1}$, v le vecteur tel que $v_i = \frac{1}{\sqrt{n_2}} 1_{n_1+1 \leq i \leq n_1+n_2}$ et $F = \text{Vect}(u, v)$. Montrer que le vecteur $P_F X$ a pour coordonnées $(\sqrt{n_1}\overline{X^{(1)}}, \sqrt{n_2}\overline{X^{(2)}})$ dans (u, v) une base orthonormée de F. Exprimer les coordonnées de $P_F X$ et $P_{F^T} X$ dans la base canonique. En déduire que $(\overline{X^{(1)}}, \overline{X^{(2)}})$ est indépendant de $\widehat{\sigma^2}$. Quelle est la loi de $\widehat{\sigma^2}$?

2 - c) Sous l'hypothèse $\mathcal{H}_0$, quelle est la loi de $\frac{\overline{X^{(1)}}-\overline{X^{(2)}}}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$? En déduire un test de niveau α pour $\mathcal{H}_0$.

2 - d) Charger le jeu de données "Ratweight.csv". En extraire un vecteur X correspondant aux poids à 14 semaines des rats nourris au Mon863, et un vecteur Y correspondant aux poids à 14 semaines des rats du groupe témoin. Implémenter le test décrit en question 2 - c) pour tester l'égalité des moyennes des échantillons X et Y. Conclure.

3 - Test non paramétrique

L'hypothèse $X \sim \mathcal{N}(\mu, \sigma)$ est justifiée dans la limite des grands échantillons par le théorème central limite. Cependant, dans le cadre de petits échantillons, cette hypothèse peut être problématique. On recourt donc à des tests non paramétriques, parmi lesquels se classe le test de Wilcoxon-Mann-Whitney. La logique de ce test est la suivante : dans un premier temps, on ordonne l'échantillon X. Si les espérances des deux populations sont différentes, la plupart des observations d'une population aura un faible rang, tandis que celle de l'autre aura un rang important. On considère comme statistique la somme des rangs des observations de l'échantillon $X^{(1)}$, notée R_1.

3 - a) À l'aide de méthodes de Monte-Carlo, estimer les quantiles à $0,025$ et à $0,0975$ de la statistique R_1 sous $\mathcal{H}_0$. Que peut-on en conclure pour nos données ?

3 - b) À l'aide de méthodes de Monte-Carlo, comparer les puissances du test du χ^2 et du test de Wilcoxon-Mann-Whitney dans le cas où les observations sont distribuées suivant une loi normale. Commenter.

Tests multiples

Dans le cadre d'études sur la leucémie, des chercheurs observent le niveau d'expression de différents gènes pour des patients atteints de leucémie de deux types. Ils tentent ensuite de déterminer les gènes liés au développement de ces maladies. Ils observent le jeu de données "leucemie.RData", c'est à dire une matrice X de taille 3051×38. Chaque ligne de cette matrice correspond au niveau d'expression d'un des $m = 3051$ gènes contrôlés, tandis que chaque colonne correspond à un patient. Les 27 premières colonnes correspondent aux patients atteints de leucémie de type "ALL", tandis que les 11 colonnes restantes correspondent aux patients atteints de leucémie de type "AML". Pour chaque gène i, on teste l'hypothèse $\mathcal{H}_0^i$: "l'espérance du niveau d'expression du gène i est la même dans les deux populations" contre l'hypothèse $\mathcal{H}_1^i$: "l'espérance du niveau d'expression du gène i est différentes selon les populations".

4 - Une approche naïve des tests multiples

4 - a) Pour un test d'hypothèse simple, on considère une statistique S et on note $T(s) = \mathbb{P}_{\mathcal{H}_0}(S \geq s)$. On rappelle que la p-valeur est définie comme $p = T(S)$. Montrer que sous $\mathcal{H}_0$, si la statistique T admet une densité, $p \sim \mathcal{U}([0,1])$. En conclure que le test consistant à rejeter $\mathcal{H}_0$ dès que $p < \alpha$ est de niveau α.

4 - b) Importer les données, et les séparer en deux matrices correspondant aux patients "ALL" et "AML". À l'aide de la fonction "t.test", calculer les p-valeurs p_i associées à ces hypothèses". Ordonner ces p-valeurs et les afficher. Commenter.

4 - c) Une approche naïve pour détecter les gènes liés au développement d'un cancer particulier consiste à calculer les p-valeurs liés aux différents niveaux d'expression des gènes, et à rejeter l'hypothèse "le niveau d'expression du gène n'a pas la même distribution dans les deux populations" dès que la p-valeur associée est inférieure à 0.05. En supposant que les niveaux d'expression des gènes sont indépendants, quel est le nombre moyen de faux positifs induit par cette procédure lorsque que pour chaque gène i, $\mathcal{H}_0^i$ est vraie ? Conclure.

5 - Contrôle de la "Family-Wise Error Rate"

La "Family-Wise Error Rate" (abrégée FWER) est définie comme la probabilité, lorsque pour chaque gène i $\mathcal{H}_0^i$ est vraie, de rejeter l'hypothèse $\mathcal{H}_0^i$ pour au moins un gène i. Pour $\beta \in [0,1]$ et pour chaque gène i, on rejette l'hypothèse $\mathcal{H}_0^i$ si $p_i \leq \beta$, où p_i est la p-valeur associée à l'hypothèse $\mathcal{H}_0^i$.

5 - a) On suppose que les niveaux d'expression des différents gènes sont indépendants. Montrer que

$$FWER \leq 1 - (1-\beta)^m.$$

En déduire un choix de β tel que la FWER soit plus petite que 0.05. Appliquer cette procédure au jeu de données "leucemie.RData".

5 - b) On ne suppose plus l'indépendance des niveaux d'expression des différents gènes. Montrer que

$$FWER \leq m\beta.$$

En déduire un choix de β tel que la FWER soit plus petite que 0.05. Appliquer cette procédure au jeu de données "leucemie.RData".

6 - Contrôle du taux de faux positifs

Les méthodes proposées ci-dessus pour contrôler la FWER ont pour défaut de considérablement réduire la puissance des tests mis en place. Pour palier ce défaut, on peut préférer contrôler le taux de faux positifs $TFP = \mathbb{E}\left[\frac{FP}{VP+FP}\right]$, où FP et VP dénotent respectivement le nombre de faux positifs et de vrais positifs. Pour minimiser le nombre de faux positifs tout en maximisant le nombre de vrais positifs, on étudie une procédure rejetant $\mathcal{H}_0^i$ dès que la p-valeur associée p_i est suffisamment petite.

Procédure de Benjamini-Hochberg : Rejeter l'hypothèse $\mathcal{H}_0^i$ pour les gènes i tels que $p_i \leq \frac{\alpha\widehat{k}}{m}$, où $\widehat{k} \triangleq \max\{k : p_k \leq \frac{\alpha k}{m}\}$.

On peut démontrer la majoration suivante pour le TFP pour la procédure de Benjamini-Hochberg

$$TFP = \mathbb{E}\left[\frac{\text{card}\{i \in I_0 : p_i \leq \alpha\widehat{k}/m\}}{\widehat{k}} 1_{\widehat{k}\geq 1}\right] \leq \alpha.$$

Implémenter la procédure de Benjamini-Hochberg pour le jeu de données pour un taux de faux positifs de 0.05. Commenter.