Projet noté : évaluation du risque par une méthode de Monte Carlo et tests par permutations

Une étoile * marque les questions difficiles. Le compte-rendu est à rendre sur Ametice.

L'objectif de ce TP est de regarder le test de comparaison de deux moyennes à partir de deux échantillons indépendants. Les données forment un vecteur $(x_1, \dots, x_m, y_1, \dots, y_n)$ de longueur $m+n$. Elles concernent l'effet d'un traitement médical sur $m+n$ patients. Les $m = 13$ premiers patients ont reçu le médicament A, qui est le traitement de référence. Les $n = 16$ autres patients ont reçu le médicament B, qui est le nouveau traitement. L'objectif de l'étude est de savoir si ce médicament B est plus efficace que le traitement de référence.

Les données sont :
$x_{1:m}^{\text{obs}} = (-3.06, -0.71, 11.99, 1.42, 1.84, 13.1, 4.19, -8.06, -3.96, -2.24, 9.61, 3.47, 3.77)$ et
$y_{1:n}^{\text{obs}} = (12.57, 7.44, 2.97, 10.35, 10.24, 9.89, 9.07, 8.23, 4.42, 2.9, 2.44, 0.49, 3.51, -3.05, 18.25, 12.29)$.

On modélise ces données $(x_1, \dots, x_m, y_1, \dots, y_n)$ par le vecteur de variables aléatoires $(X_1, \dots, X_m, Y_1, \dots, Y_n)$. On suppose que ces $m+n$ variables sont indépendantes, que les X_i ont tous même loi, d'espérance μ_X et que les Y_j ont tous même loi, d'espérance $\mu_X + \Delta$. On s'intéresse au test

$$H_0 : \Delta \leq 0 \quad vs \quad H_1 : \Delta > 0.$$

Partie 1 : Test de Student sous hypothèse gaussienne

Dans cette partie, on suppose que

$$\text{Loi}(X_1) = \dots = \text{Loi}(X_m) = \mathcal{N}\Big(\mu_X, \sigma^2\Big), \quad \text{et} \tag{1}$$

$$\text{Loi}(Y_1) = \dots = \text{Loi}(Y_n) = \mathcal{N}\Big(\mu_X + \Delta, \sigma^2\Big). \tag{2}$$

Toutes les variables aléatoires du modèle suivent donc une loi gaussienne. En outre, on a supposé que $\text{Var}(X_1) = \dots = \text{Var}(X_m) = \text{Var}(Y_1) = \dots = \text{Var}(Y_n)$.

On introduit

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{S_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{3}$$

où

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j, \quad S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_i - \bar{Y})^2$$

$$\text{et} \quad S_p^2 = \frac{1}{m+n-2}\Big((m-1)S_X^2 + (n-1)S_y^2\Big).$$

On admettra les résultats suivant :

- $\bar{X}$ varie autour de μ_X,
- $\bar{Y}$ varie autour de μ_Y,
- S_p^2 varie autour de σ^2,
- et, sous l'hypothèse supplémentaire que (le vrai) $\Delta = 0$, T suit la loi de Student à $m+n-2$ degrés de liberté.

L'objectif est d'abord de suivre la section 3 du plan de cours sur les tests statistiques pour mettre en place le test de Student ici.

1.1. Justifier du choix de $\Delta \leq 0$ comme hypothèse nulle.

1.2. Justifier du choix de la statistique de test T.

1.3. Justifier que la zone de rejet soit de la forme $[c; +\infty[$, où c est une constante à déterminer.

1.4. On note Φ_{m+n-2} la fonction de répartition de la loi de Student à $m+n-2$ degrés de liberté. Montrer que, si l'on fixe la taille du test à α, il faut choisir $c = \Phi^{-1}_{m+n-2}(1-\alpha)$.

1.5. Écrire une fonction (sans boucle `for` explicite), nommée `calculeT` qui calcule la valeur observée de T, à partir de deux entrées : le vecteur des x_i et le vecteur des y_i. Que vaut la valeur observée t^{obs} ici ?

1.6. En utilisant les fonctions de `scipy.stats`, calculer c dans notre cas si $\alpha = 0.05$. Quelle décision prend le test de Student ?

1.7. Montrer que la p-value est donnée ici par $p(X,Y) = 1 - \Phi_{m+n-2}(T)$. Quelle est la valeur observée de la p-value ?

Partie 2 : Étude de la puissance du test de Student

Dans cette partie, on conserve l'hypothèse de loi gaussienne des équations (1) et (2) et on note $\theta = (\mu_X, \Delta, \sigma^2)$. L'objectif est d'étudier la fonction puissance

$$\theta \mapsto \beta(\theta) = \mathbb{P}_\theta\Big(T \geq \Phi^{-1}_{m+n-2}(1-\alpha)\Big).$$

On s'intéresse aux valeurs de cette puissance lorsque

$$\Delta \in [0; 15], \quad \text{et} \quad \sigma^2 \in [20; 40]. \tag{4}$$

2.1. * Montrer que la loi de T ne dépend pas de μ_X, mais uniquement de Δ et σ^2. Dans toute la suite, on fixera $\mu_X = 0.92$ si on a besoin d'une valeur numérique.

2.2. Écrire une fonction nommée `simuleT` qui prend en entrée les valeurs de Δ et σ^2 et qui fait les choses suivantes :
- elle simule un vecteur $(X_1, \ldots, X_m)$ et un vecteur $(Y_1, \ldots, Y_n)$, de longueurs respectives $m = 13$ et $n = 17$;
- elle calcule et renvoie la valeur de T de l'équation (3), en utilisant la fonction `calculeT`.

2.3. Écrire une fonction `puissT` qui prend en entrée les valeurs de Δ, σ^2, α et N et qui fait les choses suivantes :
- elle simule N valeurs de T indépendantes ;
- elle calcule et renvoie le nombre de fois où $T \geq \Phi^{-1}_{m+n-2}(1-\alpha)$

2.4. En utilisant cette fonction pour $\sigma^2 = 20$ et $N = 10^3$, approcher les valeurs de $\beta(\theta)$ pour les 16 valeurs entières de Δ entre 0 et 15. Représenter graphiquement ces approximations en fonction de Δ.

2.5. Même question pour $\sigma^2 = 30$.

2.6. Même question pour $\sigma^2 = 40$.

2.7. Que peut-on faire pour améliorer ces approximations ? Le faire, et constater le résultat.

Partie 3 : calcul d'une p-value par une méthode de permutation Monte-Carlo

On suppose maintenant que

$$\text{Loi}(X_1) = \ldots = \text{Loi}(X_m) \text{ de fonction de répartition } F(\cdot), \quad \text{et} \tag{5}$$
$$\text{Loi}(Y_1) = \ldots = \text{Loi}(Y_n) \text{ de fonction de répartition } F(\cdot - \Delta). \tag{6}$$

Désormais, le paramètre inconnu est $\theta = (\Delta, F)$.

On note $(X_1^*, \ldots, X_m^*, Y_1^*, \ldots, Y_n^*)$ une permutation aléatoire du vecteur $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. Autrement dit, X_1^* est tiré uniformément au hasard par le vecteur $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ de longueur $m+n$, puis X_2^* est tiré uniformément au hasard parmi les $m+n-1$ coordonnées restantes de $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$,... À la fin, Y_n^* est la seule coordonnée de $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ qui n'a pas encore été tirée. On pourra utiliser la fonction `resample` dans `sklearn.utils` pour faire cette permutation aléatoire.

3.1. Écrire une fonction `approxP`, qui prend en entrée le vecteur des données $(x_1, \ldots, x_m, y_1, \ldots, y_n)$, t^{obs} et K, et qui fait les choses suivantes :
- elle calcule K valeurs de T sur K échantillons permutés $(X_1^*, \ldots, X_m^*, Y_1^*, \ldots, Y_n^*)$ des données observées ;
- elle compte, parmis ces K valeurs de T, le nombre de fois où $T > t^{\text{obs}}$;
- elle renvoie la fréquence de $T > t^{\text{obs}}$ parmi ces K simulations.

3.2. * En quelques lignes, expliquer pourquoi `approxP` permet d'approcher la p-value du test.

3.3. Rappeler comment on prend une décision en fonction de la valeur de la p-value et de la taille α du test.

3.4. Approcher la p-value du test sur les données qui nous intéresse par cette méthode avec $K = 10^4$. Quelle est la décision que l'on prend ici.

Partie 4 : Étude la puissance du test mis en place dans la partie 3

L'objectif de cette partie est de comparer le test mis en place dans la partie 1, et celui mis en place dans la partie 3, lorsque les hypothèses de la partie 1 sont satisfaites, c'est-à-dire lorsque (1) et (2) sont satisfaites. Comme dans la partie 2, on s'intéresse à des valeurs de Δ et σ^2 qui vérifient (4) et pour $\mu_X = 0.92$. On note $\theta \mapsto \beta_3(\theta)$ la fonction puissance du test de la partie 3.

4.1. En s'inspirant de la partie 2, écrire une fonction `puiss3` qui prend en entrée les valeurs de Δ, σ^2, α, K et N et qui fait les choses suivantes :
- elle simule N jeux de données suivant (1) et (2) ;
- elle calcule, sur chacun de N jeux de données, une approximation de la p-value à l'aide de `approxP` avec K permutations indépendantes ;
- elle compte le nombre de fois où ces N p-value sont inférieures à α ;
- elle renvoie la fréquence où $p < \alpha$ parmi ces N valeurs de p.

4.2. En utilisant cette fonction pour $\sigma^2 = 20$, $N = 10^3$ et $K = 400$, approcher les valeurs de $\beta_3(\theta)$ pour les 16 valeurs entières de Δ entre 0 et 15. Représenter graphiquement ces approximations en fonction de Δ, ainsi que $\beta(\theta)$ tel que calculé dans la partie 2. On représentera les deux courbes dans deux couleurs différentes.

4.3. Même question pour $\sigma^2 = 30$.

4.4. Même question pour $\sigma^2 = 40$.

4.5. Quel est le test le plus puissant sous l'hypothèse gaussienne ?

4.6. Peut-on raisonnablement faire quelque chose ici pour améliorer l'approximation de $\beta_3(\theta)$?

Partie 5 : amélioration de l'approximation

L'objectif de cette partie est d'étudier l'article de Boos et Zhang (2000, *Journal of the American Statistical Association*).

5.1. Décrire en quelques lignes quel est l'objectif de l'article en question.

5.2. Quelle est l'idée de l'algorithme ?

5.3. * Implémenter l'algorithme.

5.4. * L'utiliser et conclure.