Contrôle : Statistique pour les big data

Patrick Tardivel, Université de Bourgogne

31 mars 2023

Durée 2 heures, calculatrice autorisée, notes de cours interdite.

Exercice 1 (Question de cours) *On rappelle que la proportion de faux positifs (FDP) est le quotient du nombre de Faux Positifs* $|FP(X)|$ *sur le nombre d'hypothèses nulles Rejetées* $|R(X)|$ *:*

$$FDP(X) = \frac{|FP(X)|}{\max\{|R(X)|, 1\}}, \text{ où } R(X) = FP(X) \cup VP(X).$$

Le taux de faux positifs (FDR) est l'espérance de la proportion de faux positifs : $FDR = \mathbb{E}(FDP(X))$*. On rappelle également que la probabilité d'avoir au moins un faux positif (FWER) est* $\mathbb{P}(|FP(X)| \geq 1)$*. Montrer les résultats suivants :*

1. *Le FDR est plus petit que le FWER.*
2. *Lorsque les hypothèses nulles sont toutes vraies alors* $FDR = FWER$*.*

Exercice 2 *On considère le problème de test multiple suivant*

$$\begin{cases} X : (\Omega, \mathcal{F}, Pr) \to (\mathbb{X}, \mathcal{X}) \\ \mathbb{P}^X \in \mathcal{P} \text{ où } \mathcal{P} \text{ est une famille de loi sur } (\mathbb{X}, \mathcal{X}) \\ \text{Pour tout } j \in \{1, \dots, n\}, \mathcal{H}^{0,j} : \mathbb{P}^X \in \mathcal{P}^{0,j} \text{ où } \mathcal{P}^{0,j} \text{ est un sous-ensemble de } \mathcal{P} \end{cases}.$$

Pour $j \in \{1, \dots, n\}$, $p_j(X)$ *est une p-valeur, c'est-à-dire une variable aléatoire vérifiant l'inégalité suivante dès que* $\mathbb{P}^X \in \mathcal{P}^{0,j}$ *:*

$$\forall t \in [0,1] \quad Pr(p_j(X) \leq t) \leq t.$$

1. *Rappeler la procédure de Benjamini-Hochberg. Que peut-on dire du contrôle du FDR lorsque les p-valeurs sont indépendantes ? Même question lorsqu'aucune hypothèse d'indépendance n'est faîte sur les p-valeurs.*
2. *Soit* $p_1(X), \dots, p_5(X)$ *des p-valeurs associées aux hypothèses nulles* $\mathcal{H}_1^0, \dots, \mathcal{H}_5^0$*. Les p-valeurs expérimentales sont :* $p_1(X^{exp}) = 0,015$, $p_2(X^{exp}) = 0,002$, $p_3(X^{exp}) = 0,090$, $p_4(X^{exp}) = 0,036$ *et* $p_5(X^{exp}) = 0,032$*. Pour un contrôle du FDR au niveau* $0,05$*, donner les hypothèses nulles rejetées par la procédure de Benjamini-Hochberg lorsque*
   (a) *Les p-valeurs sont indépendantes.*
   (b) *Aucune hypothèse n'est faîtes sur les p-valeurs.*

   Vous justifierez vos réponses.

Exercice 3 *Soit* $X = (X_1, \dots, X_n)$ *un échantillon de loi normale* $\mathcal{N}(\mu, \sigma^2)$ *où* $\mu \in \mathbb{R}$ *et* $\sigma^2 > 0$ *sont inconnus. On souhaite tester l'hypothèse nulle* $\mathcal{H}^0 : \sigma^2 \leq \sigma_0^2$ *(où* σ_0^2 *est une valeur donnée) contre l'hypothèse alternative* $\mathcal{H}^1 : \sigma^2 > \sigma_0^2$*.*

1. *Réécrire formellement le problème de test.*

Soit $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$*, on pose* $Y(X) = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma_0^2$*. Avec un risque de première espèce* $\alpha \in]0,1[$ *on rejette* $\mathcal{H}^0$ *lorsque* $Y(X) > F_0^{-1}(1-\alpha)$ *où* F_0 *est la fonction de répartition d'une loi du khi deux à* $n-1$ *degrés de liberté.*

2. *Déterminer la p-valeur* $p(X)$ *de cette procédure de test.*
3. *Montrer que la p-valeur suit une loi uniforme lorsque* $\sigma^2 = \sigma_0^2$*.*
4. *On rappelle que la statistique* $\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$ *suit une loi khi deux à* $n-1$ *degrés de liberté. Montrer que lorsque* $\sigma^2 < \sigma_0^2$*, pour tout* $x \in]0,1[$ *on a* $Pr(p(X) \leq x) < x$*. Comment qualifie-t-on une p-valeur satisfaisant cette inégalité ?*

Exercice 4 *Soit* $X = (X_1, \ldots, X_{n_1})$ *un échantillon de taille* n_1 *de loi* $\mathcal{N}(\mu_X, \sigma_X^2)$ *et* $Y = (Y_1, \ldots, Y_{n_2})$ *un échantillon de taille* n_2 *indépendant du premier échantillon et de loi* $\mathcal{N}(\mu_Y, \sigma_Y^2)$*. On souhaite tester l'hypothèse nulle* $\mathcal{H}^0 : \sigma_X^2 = \sigma_Y^2$ *contre l'hypothèse alternative* $\mathcal{H}^1 : \sigma_X^2 \neq \sigma_Y^2$*. On rappelle que sous l'hypothèse nulle, lorsque* $\sigma_X^2 = \sigma_Y^2$*, la statistique*

$$F(X,Y) = \frac{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2}{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}$$

suit une loi de Fisher à $n_1 - 1, n_2 - 1$ *degrés de liberté.*

1. *Proposer une procédure de test pour* $\mathcal{H}^0$ *contre* $\mathcal{H}^1$ *contrôlant le risque de première espèce au niveau* $\alpha \in]0,1[$*.*
2. *Déterminer la p-valeur* $p(X,Y)$ *de cette procédure de test.*
3. *Montrer que la p-valeur suit une loi uniforme lorsque* $\sigma_X^2 = \sigma_Y^2$*.*