Travaux pratiques : tests multiples

Patrick Tardivel,
Université de Bourgogne, Dijon

Ces travaux pratiques sont notés. Durée : 3h, tous les documents autorisés, usage d'internet autorisé (sauf pour communiquer avec un étudiant ou une autre personne). A la fin de la séance veuillez m'envoyer le code sur R avec pour titre : « TP_Big_Data_NOM_Prénom » par courriel à l'adresse Patrick.Tardivel@u-bourgogne.fr

Exercice 1

1. *On considère les p-valeurs expérimentales suivantes :* $p_1(X^{\exp}) = 0.042$, $p_2(X^{\exp}) = 0.006$, $p_3(X^{\exp}) = 0.014$, $p_4(X^{\exp}) = 0.015$, $p_5(X^{\exp}) = 0.048$.
   (a) *Déterminer les hypothèses nulles rejetées par la procédure de Holm lorsque la probabilité d'avoir un ou plusieurs faux positifs (FWER) est contrôlée au niveau* 0.05.
   (b) *Déterminer les hypothèses nulles rejetées par la procédure de Benjamini-Hochberg lorsque le taux de faux positifs (FDR) est contrôlé au niveau* 0.05.
2. *Écrire une fonction Holm qui prend en entrée un vecteur* $p \in [0,1]^n$ *(représentant des p-valeurs) et* $\alpha \in [0,1]$ *et qui renvois en sortie les hypothèses nulles rejetées par la procédure de Holm lorsque le FWER est contrôlé au niveau* α.
3. *Écrire une fonction BH qui prend en entrée un vecteur* $p \in [0,1]^n$ *(représentant des p-valeurs) et* $\alpha \in [0,1]$ *et qui renvois en sortie les hypothèses nulles rejetées par la procédure de Benjamini-Hochberg lorsque le FDR est contrôlé au niveau* α.

Exercice 2 *Le tableau suivant est tiré d'une étude sur le risque qu'un travailleur ait des problèmes de genou en fonction de la durée de la carrière (Duration) et de la pénibilité au travail (Intensity/Frequency).*

Duration	Intensity/Frequency	OR	IC 95%	p-value
low	[1,2[	1.59	[1.35 , 1.87]	0.00000003
low	[2,3[	1.71	[1.37 , 2.14]	0.00000228
low	[3,4[	1.45	[0.95 , 2.19]	0.08556811
low	{4}	2.63	[1.83 , 3.79]	0.00000019
medium	[0,1[	0.87	[0.75 , 1.02]	0.08772095
medium	[1,2[	1.33	[1.10 , 1.59]	0.00293856
medium	[2,3[	1.64	[1.28 , 2.10]	0.00009054
medium	[3,4[	1.79	[1.19, 2.70]	0.00523358
medium	{4}	3.30	[2.04, 5.34]	0.00000115
high	[0,1[	0.93	[0.77, 1.12]	0.43871305
high	[1,2[	1.31	[1.07, 1.62]	0.00933741
high	[2,3[	1.93	[1.51, 2.47]	0.00000016
high	[3,4[	1.36	[0.85, 2.19]	0.19814336
high	{4}	1.63	[0.89, 2.98]	0.11365255

TABLE 1 – Results of adjusted logistic regression of severe knee pain with lifting.

L'Odds Ratio (OR) compare les probabilités d'avoir des problèmes de genou pour une catégorie par rapport à la catégorie de référence (durée faible et pénibilité faible). Dans le cadre de la régression logistique l'odds ratio est estimé par $\exp(\widehat{\beta})$ *où* $\widehat{\beta}$ *est un estimateur du coefficient de régression* β *d'une catégorie de plus l'intervalle*

de confiance pour l'odds ratio de niveau asymptotique $0,95$ *est donné par la formule*

$$\left[\exp\left(\widehat{\beta} - 1,96\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})}\right), \exp\left(\widehat{\beta} + 1,96\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})}\right)\right].$$

Enfin, l'estimateur $\widehat{\beta}$ *est asymptotiquement normal :*

$$\frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\mathrm{var}}(\widehat{\beta})}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

1. *On considère le problème de test* $\mathcal{H}_0 : \beta = 0$ *vs* $\mathcal{H}_1 : \beta \neq 0$. *En utilisant la normalité asymptotique, donner une procédure de test et calculer sa p-valeur. Expliquer comment à partir de la colonne « IC* 95% *» on retrouve les p-valeurs données à la colonne « p-value » .*
2. *Quelles sont les hypothèses nulles rejetées par la procédure de Holm lorsque la probabilité d'avoir au moins un faux positif (FWER) est contrôlée au niveau* $0,05$ *?*
3. *Quelles sont les hypothèses nulles rejetées pour la procédure de Benjamini-Hochberg lorsque le taux de faux positifs (FDR) est contrôlé au niveau* $0,05$ *?*

Pour les expériences numériques vous pouvez utiliser le code R suivant

```
borne_inf=c(1.35,1.37,0.95,1.83,0.75,1.10,1.28,1.19,2.04,0.77,1.07,1.51,0.85,0.89)
borne_sup=c(1.87,2.14,2.19,3.79,1.02,1.59,2.10,2.70,5.34,1.12,1.62,2.47,2.19,2.98)
p_val=c(0.00000003,0.00000228,0.08556811,0.00000019,0.08772095,0.00293856,0.00009054,
0.00523358,0.00000115,0.43871305,0.00933741,0.00000016,0.19814336,0.11365255)
```