Master 1 Mathématique — Année 2010-2011
Université d'Angers

TP 3 : TESTS

Test de comparaison de deux moyennes

Nous allons comparer les poids de poulpes mâles et femelles au stade adulte. Pour cela, nous disposons du fichier "poulpe.csv" qui contient la pesée de 28 poulpes. Nous souhaitons tester l'égalité des moyennes théoriques inconnues des poids des poulpes femelles et mâles.

1. Récupérer le fichier "poulpe.csv" et importer les données sous R à l'aide de la commande

   ```
   >poulpe<-read.table("poulpe.csv",header=T,sep=";")
   ```

2. Résumer le jeu de donnée à l'aide de la commande summary.
3. Visualiser une boîte à moustaches des données à l'aide de la commande :

   ```
   >boxplot(Poids~Sexe,ylab="Poids",xlab="Sexe",data=poulpe),
   ```

   et estimer les statistiques de base des deux sous-population.
4. Pour construire un test de comparaison des moyennes, nous faisons l'hypothèse que l'estimateur de la moyenne suit une loi normale dans chaque sous-population. Cela est vrai si la taille de l'échantillon est suffisamment grande (par le TLC), ou bien si la loi de nos observations est elle-même gaussienne. Ici les effectifs sont restreints (< 30) et l'utilisation du TLC n'est pas raisonnable. Nous allons donc tester la normalité des données pour chaque sous-population avec le test de Shapiro-Wilk. La commande

   ```
   >shapiro.test(x),
   ```

   où x est un vecteur permet de tester la normalité de l'échantillon x. Tester la normalité de chaque sous-population et commenter les résultats obtenus.
5. Pour tester l'égalité des moyennes, on connait le test de Fisher-Student vu en cours. Pour cela, il faut d'abord vérifier l'égalité des variances. Tester l'égalité des variances à l'aide de la commande var.test. Commenter la sortie R.
6. Les variances étant différentes, on utilise le test de Welch pour tester l'égalité des moyennes. Pour cela, on utilisera la fonction t.test avec l'argument var.equal=FALSE. réaliser le test au niveau 5% de l'hypothèse H_0 : "Les moyennes sont égales" contre l'alternative H_1 : "Les moyennes sont différentes". Conclure.

Test d'indépendance

L'objectif est ici de tester l'indépendance entre deux variables qualitatives. Pour cela, on utilise le test du χ^2 d'indépendance qui teste l'hypothèse H_0 : "Les deux variables sont indépendantes" contre H_1 : "Les deux variables ne sont pas indépendantes". Ce test est basé sur la statistique suivante :

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - T_{ij})^2}{T_{ij}},$$

où n_{ij} est l'effectif observé pour la modalité i de la première variable et j de la deuxième, et T_{ij} correspond à l'effectif sous l'hypothèse d'indépendance. Ainsi $T_{ij} = n\hat{p}_{i.}\hat{p}_{.j}$ avec $\hat{p}_{i.} = \frac{1}{n}\sum_j n_{ij}$ et $\hat{p}_{.j} = \frac{1}{n}\sum_i n_{ij}$. Sous l'hypothèse H_0, cette statistique converge vers une loi du χ^2 à $(I-1)(J-1)$ degré de liberté.

On étudie l'influence du sexe sur la couleur des cheveux d'élèves d'un district écossais. Nous souhaitons savoir si la couleur des cheveux est indépendante du sexe. Pour cela, on dispose du tableau de données suivant :

	Blond	Roux	Châtain	Brun	Noir de jais
Garçon	592	119	849	504	36
Fille	544	97	677	451	14

1. Saisissez le jeu de données manuellement dans une matrice :
   >tab<-matrix(c(592,544,119,97,849,677,504,451,36,14),ncol=5)
   >rownames(tab)<-c("Garçon","Fille")
   >colnames(tab)<-c("Blond","Roux","Chatain","Brun","Noir de jais")
2. Représenter les données à l'aide de diagrammes en barres, où on représente les données par sexe sur un même graphique :
   >par(mfrow=c(2,1))
   >barplot(tab[1,],main="Garçons")
   >barplot(tab2,],main="Filles")
3. Réaliser le test à l'aide de la commande :
   >resultat<-chisq.test(tab)
4. Précisez les couleurs qui contribuent le plus au Khi2. Ces contributions sont dans l'objet residuals. En divisant chaque valeur par la statistique de test (contenu dans l'objet stat), on obtient les pourcentages.

Test de Kolmogorov Smirnov

On dispose d'un échnatillon $(X_1, \ldots, X_n)$ i.i.d. de loi μ inconnue. Alors, d'après le théorème de Glivenko-Cantelli, on sait que si F_μ est la fonction de répartition de μ, la statistique

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t)$$

converge presque-sûrement uniformément vers F_μ, c'est-à-dire :

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \longrightarrow 0 \, p.s.$$

Ce théorème a été obtenu en utilisant la loi forte des grands nombres. Une version plus élaborée utilisant alors le théorème limite central permet d'aboutir au test de Kolmogorov Smirnov.

Theorem 1 *Soit $(X_1, \ldots X_n)$ un échantillon de loi μ et de fonction de répartition F_μ continue. Alors on a :*

$$K_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{\mathcal{L}} \mu_{KS},$$

où μ_{KS} est une loi de probabilité universelle indépendante de F de fonction de répartition F_{KS} donnée par, $\forall t > 0$:

$$F_{KS}(t) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k \exp(-2k^2 t^2).$$

1. Simuler N échantillons $(X_1, \ldots X_n)$ de loi normale $N(0,1)$ et représenter l'histogramme des N réalisations $K_n^i, i = 1, \ldots N$ des variables aléatoires K_n.
2. Répéter le procédé pour une loi exponentielle $\mathcal{E}(1)$ et une loi uniforme $\mathcal{U}[0,1]$.
3. Illustrer le Théorème de Kolmogorov-Smirnov en comparant les histogrammes obtenus.

A partir d'un échantillon $(X_1, \ldots, X_n)$ de loi μ inconnue, on veut utiliser le Théorème de Kolmogorov-Smirnov pour construire un test asymptotique d'hypothèse H_0 : "$\mu = \nu$" contre H_1 : "$\mu \neq \nu$".

Corollaire 1 *Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de loi μ de fonction de répartition F_μ continue. Alors pour tester l'hypothèse H_0 : "$\mu = \nu$" contre l'alternative H_1 : "$\mu \neq \nu$", le test défini par la zone de rejet suivante :*

$$R_{KS} = \{\sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F_\nu(t)| > k_{1-\alpha}^{KS}\},$$

où $k_{1-\alpha}^{KS}$ est le quantile d'ordre $1-\alpha$ de la loi μ_{KS}, est un test de niveau asymptotique α.

4. Démontrer le Corollaire.
5. Montrer que la statistique de test K_n s'écrit en réalité :

$$K_n = \sqrt{n} \max_{i=1,\ldots n} \left(F_\nu(X_{(i)}) - \frac{i}{n}, F_\nu(X_{(i)}) - \frac{i-1}{n} \right).$$

6. On peut à présent construire le test de Kolmogorov Smirnov. Ecrire une fonction ksnorm ayant un vecteur d'observations en argument et qui retourne la p-valeur du test de Kolmogorov Smirnov d'adéquation à une loi normale.
7. On va tester le générateur aléatoire rnorm de **R**. En utilisant la fonction ksnorm, tracer un graphique qui représente l'évolution des p-valeurs du test de Kolmogorov en fonction de la taille de l'échantillon $(X_1, \ldots X_n)$ généré, pour $n = 10$ à $n = 20000$ par pas de 10.