

Devoir 1

Pascal Bessonneau et Christophe Pallier

November 12, 2009

Ce devoir ne sera pas noté. Considérez le comme un entraînement. Il permettra également permettre aux enseignants d'évaluer les notions qui ont été acquises. Essayer de faire le maximum mais ne vous découragez pas si vous n'arrivez pas à répondre à certaines questions.

Vous pouvez renvoyer soit un document accompagné du script R soigneusement commenté, soit directement un script R avec les réponses aux questions en commentaires. Ceux qui sont familiers de \LaTeX , pourront utiliser Sweave.

Nous vous invitons à nous le renvoyer le 21 novembre au plus tard à pbessonneau@gmail.com et Christophe Pallier christophe@pallier.org.

1 Exploration de données

Question 1

1. Charger la table sauvée sur le site du cours dans `support/cours3/Population.txt` (attention: le séparateur de colonnes est ";")
2. Quelles sont les valeurs minimale, médiane et maximale de *Value*?
3. Afficher l'histogramme de *Value*, puis de $\log(\text{Value})$.
4. Combien de pays différents sont listés dans la colonne *Country.or.Area*
5. quelle est la valeur moyenne de *Value* pour le Togo?
6. Calculer les moyennes de $\log(\text{Value})$ par pays, puis lister les dix pays ayant les plus forte valeurs.
7. Afficher les distributions de $\log(\text{Value})$ pour chacun des niveaux de la variable *Sex*.
8. Comparer par un t.test $\log(\text{Value})$ pour les groupes Male et Female (on ignorera l'appariement par pays)

2 Pile ou face (encore!)

Question 2

1. Simuler 100 tirages pile-ou-face avec une pièce non biaisée.
2. Calculer la moyenne du nombre de 'pile'.
3. répéter l'opération 1000 fois et afficher la distribution de la moyenne.
4. Faire la même chose avec une pièce biaisée pour tomber 80% du temps sur 'pile'
5. Quelle est la probabilité qu'une pièce biaisée à 80% tombe exactement 5 fois sur pile lors de 10 tirages?
6. Simuler 30 tirage d'une pile non biaisée, et 30 tirage d'une pile biaisée à 80%; effectuer un test statistique pour comparer les proportions observées dans chacune des simulations.

3 Echantillonnage

Question 3

1. Créer une variable contenant 100 valeurs aléatoires distribuées suivant une loi normale de moyenne théorique 0 et d'écart-type theorique 1.
2. Afficher ces données sur une boîte à moustache (boxplot).
3. Afficher côte à côte 20 boxplots d'échantillons de 100 éléments tirés selon la même loi ($\text{Normal}(\text{mean}=0, \text{sd}=1)$). (remarque: vous pouvez tirer partie du fait que la fonction boxplot sur un objet "matrix" affiche un boxplot pour chaque colonne de la matrice).
4. Faire la même chose pour des échantillons de taille 10, puis pour des échantillons de taille 1000.
5. Calculer l'écart-type des moyennes de 1000 échantillons normaux ($\text{mean}=0, \text{sd}=1$) de taille 10. Refaire la même pour des taille variant de 10 à 100 par pas de 10. Afficher ces écart-types.

4 Simulations de t.test

Question 4

1. Générer un vecteur de 10 valeurs aléatoires distribuées normalement puis Utiliser 'scale' pour le transformer en vecteur ayant exactement une moyenne de 10 et un écart type de 5.
2. De la même façon, créer un vecteur de 10 éléments ayant exactement une moyenne de 11 et un écart-type de 5.
3. Comparer les deux vecteur à l'aide d'un t.test.
4. Refaire la même chose avec des vecteurs de moyennes 10 et 12, puis 10 et 13, puis 10 et 14, puis 10 et 15.

5 Analyse de données

Le but ici est de préparer une analyse de données plus fine des données du fichier patient.csv. Ce fichier a été décrit dans le TP n°2 (Cours n°4). Un fichier décrivant le contenu des colonnes est présent dans le dossier sous le nom: patient.txt

Pour rappel il s'agit de données provenant d'une enquête sur dossier de la prise en charge de la douleur chez l'enfant dans deux hopitaux différents.

Le but de ces petits exercices est de préparer une analyse plus fine.

Les quatres variables étudiées sont l'âge, le nombre de jours d'hospitalisation post-opératoires (postopj), le nombre de traitement contre la douleur (nbttt), la somme des valeurs obtenues à toutes les évaluations de la douleur (totalechelle) et le nombre d'évaluation de la douleur réalisées (nbechelle).

Description

1. Faire une description graphique des variables
2. Faire une description numérique des variables

Relation entre les variables

1. Regarder la corrélation des variables entre elles par des représentations graphiques. Qu'en concluez-vous ?
2. Regarder les variations des variables en fonction des variables CIM2 (pathologies) et ACP (prescription d'une pompe à morphine).

Perspectives

1. Est-ce que selon vous il faudrait utiliser des variables synthétiques: de nouvelles variables calculées avec les données disponibles ? Si oui pourquoi ?
2. Faire une étude de ces variables comme pour les questions précédentes.