Faculté des Sciences
Aix*Marseille Université

Année universitaire 2017/2018

Site : ☒ Luminy ☐ St-Charles ☐ St-Jérôme ☐ Cht-Gombert ☐ Aix-Montperrin ☐ Aubagne-SATIS
Sujet de : ☒ 1er semestre ☐ 2ème semestre ☐ Session 2 Durée de l'épreuve : 2h
Examen de : M1 Nom du diplôme : MASTER 1 BBSG
Code du module : Libellé du module : Analyse statistique des données biologiques
Calculatrices autorisées : NON Documents autorisés : NON

NB : Les deux parties sont indépendantes. Les réponses attendues sont des scripts R avec éventuellement un commentaire succint. Merci de bien indiquer sur la copie le numéro des questions.

1 Partie 1 Analyse de données

L'analyse proposée porte sur le jeu de données bosson.csv (fournies par le Professeur Jean-Luc Bosson). Cette base contient les informations de 209 patients venant de France ou du Vietnam :
country : Vietnam ou France
gender : F ou M
aneurysm : taille de l'anévrisme en mm
bmi : indice de masse corporelle
risk : nombre de facteurs de risque entre 0 et 5

1. Charger le fichier bosson.csv placé dans amétice, onglet EXAMEN, on le nommera BOSSON (utiliser plutôt la fonction read.csv2())
2. Nommer les colonnes du jeu de données Pays, Genre, Anévrisme, IMC, Risque.
3. Afficher les 6 premières lignes de la base. Afficher le nom des colonnes de BOSSON.
4. Quel est le type des différentes variables ?
5. Séparer dans deux data.frames les données concernant les patients vietnamiens et français.
6. Quel est le nombre de facteurs de risque moyen des hommes, des femmes ?
7. Calculer les effectifs et les proportions d'hommes et femmes.
8. Quelle est la proportion de vietnamiens dans le data frame ?
9. Tracer la répartition du genre selon le pays.
10. Tracer une représentation du nombre de facteurs de risque selon le pays.
11. Tracer un boxplot de la colonne anévrisme en fonction du genre.
12. Est-ce que les moyennes de la taille des anévrismes des hommes et des femmes diffèrent de façon significative ? On effectuera un test de Student, dont on extraira en particulier la p value.
13. De même on regardera s'il y a une différence entre les moyennes de la taille des anévrismes des français (hommes et femmes) et des vietnamiens(hommes et femmes).
14. Tracer deux histogrammes de fréquence de la taille des anévrismes des femmes en vert, et de celle des hommes en bleu.
15. Donner une représentation graphique permettant de comparer la variabilité de la variable « taille des anévrismes » chez les hommes et les femmes. Peut-on au vu de ce graphique considérer les variances égales ?
16. Donner une représentation de la variable risque selon le pays d'origine.
17. Y-a -t-il une différence significative entre l'IMC des individus français et vietnamiens ? On répondra par une méthode laissée au choix.

2 Partie 2 Simulations

1. Générer n=50 vecteurs de taille 100 suivant une loi continue uniforme sur l'intervalle [0,1]. Ces valeurs seront placées par construction sous forme de matrice M à 50 lignes.
2. Effectuer à l'aide d'une boucle 50 tests de Student de comparaison de la moyenne observée pour chaque vecteur (ligne de M), à la moyenne théorique mu=0.5. On pourra utiliser la fonction t.test avec l'argument mu=0.5.

   On conservera uniquement la p-value de ces tests, dans un vecteur numérique ValeurP.
3. Compter le nombre de tests pour lesquels on peut conclure que la moyenne observée est significativement différente de mu=0.5.
4. Refaire le même travail (à l'aide d'une fonction si possible), mais en prenant pour chaque vecteur seulement la moyenne de ses 10 premières valeurs. On placera les résultats obtenus dans la variable MO10. De même pour MO50 (moyenne des 50 premières valeurs de chaque vecteur).
5. Tracer sur le même graphique les histogramme de fréquences de M, M010, M050.
6. Superposer la courbe de la loi normale de moyenne 0.5 et variance 1/12.