STA104 : Contrôle continu

03/10/2022

Durée = 2 heures, conseil : 1 heure pour chaque logiciel

Documents et Internet autorisés

Pour chaque question, la réponse attendue est un commentaire (plus un résultat ou graphique si nécessaire) suivie d'une suite de commandes qui permet de répondre à la question. Vous pouvez rajouter des commentaires dans vos programmes. Même si plusieurs possibilités existent pour répondre à la question choisissez en une seule.

Pour la partie SAS, les réponses doivent être rédigées sur un seul document, que vous exporterez au format PDF (nom du fichier : Nom_Prenom.pdf), et que vous enverrez par mail à pierre.joly@u-bordeaux.fr.

Pour la partie R, un fichier .R (nom du fichier : Nom_Prenom.R) contenant les commandes ainsi que les codes **ou** un fichier HTML (ou PDF) compilé avec RMarkdown en laissant les codes apparents sera à envoyer par mail à benjamin.hivert@u-bordeaux.fr

© - 2022 – Université de Bordeaux, ISPED

Les données sur lesquelles vous devez travailler proviennent d'une étude pour comprendre la prévalence de l'obésité, du diabète et d'autres facteurs de risque cardiovasculaire dans le centre de la Virginie pour les Afro-Américains. Elles se composent de 13 variables qui sont :

- **id** : L'identifiant du sujet

- **chol** : Mesure de cholestérol total

- **stab.glu** : Mesure de glucose stabilisé

- **glyhb** : Mesure d'hémoglobine glyquée

- **location** : Origine de l'individu

- **age** : Âge de l'individu

- **gender** : Genre de l'individu

- **height** : Taille de l'individu (en pouces)

- **weight** : Poids de l'individu (en livres sterling)

- **bp.1s** : Pression artérielle systolique

- **bp.1d** : Pression artérielle diastolique

- **waist** : Tour de taille (en pouces)

- **hip** : Tour de hanche (en pouces)

© - 2022 – Université de Bordeaux, ISPED

Partie R (10 points)

Consignes : Restituer un fichier .R (nom du fichier : Nom_Prenom.R) contenant les commandes ainsi que les commentaires. Le rendu peut également se faire via RMarkdown (fichier HMTL ou PDF) en laissant les codes apparents.

L'évaluation portera sur l'ensemble des questions mais également sur la qualité du script et des différents graphiques.

Exercice 1 : Création de fonction

Créer une fonction R qui prend en argument une taille (en pouces) et un poids (en livres sterling) et qui renvoie l'Indice de Masse Corporelle associé (IMC).

- **Pour le calcul :** $\text{IMC} = \frac{\text{Poids (kg)}}{(\text{Taille (m)})^2}$

- **Conversion :**
  – 1 livre sterling = 0, 453592 kg

  – 1 pouce = 0,0254m

Exemple : Pour une taille de 62 pouces et un poids de 121 livres, la fonction doit renvoyer : 22.1309443

Exercice 2 : Préparation des données

1. Importer les données *diabetes.csv.*

2. Corriger le type des variables si nécessaire.

3. Recoder les modalités de la variable **gender** pour qu'elle prenne comme valeurs "Homme" et "Femme" au lieu de "male" et "female". En particulier, combien y-a-t-il d'hommes et de femmes dans l'étude ?

4. D'après les auteurs, une hémoglobine glyquée (variable **glyhb**) supérieure à 7 est généralement considérée comme un diagnostic positif de diabète. En se basant sur ce critère, créer une nouvelle variable **diabetic** (de type factor) qui vaut 1 si l'individu est diabétique, et 0 sinon. En particulier, quelle est la proportion de personnes diabétiques dans l'étude ?

5. Créer une nouvelle variable **WHR** (*Waist-Hip Ratio*) correspondant au ratio entre le tour de taille et le tour de hanche de l'individu.

6. Selon l'OMS, on considère qu'une personne souffre d'hypertension lorsque l'on constate une tension artérielle systolique supérieure ou égale à 140 mmHg ou une tension artérielle diastolique supérieure ou égale à 90 mmHg. A partir de ces

© - 2022 – Université de Bordeaux, ISPED

informations, créer une variable **hypertension** (de type factor) qui vaut 1 si l'individu souffre d'hypertension et 0 sinon. En particulier, combien de personnes souffrent d'hypertension artérielle dans l'étude ?

7. Appliquer la fonction faite à la Partie 1 pour calculer l'IMC des individus de l'étude.

Exercice 3 : Analyse de données

Les représentations graphiques doivent être réalisées, dans la mesure du possible, à l'aide du package ggplot2.

8. Décrire les variables **chol**, **stab.glu** et **location** sur l'ensemble des individus, puis suivant les modalités de la variable **hypertension**.

9. Faire une représentation graphique permettant de comparer la distribution du ratio tour de taille-tour de hanche en fonction du statut diabétique des individus.
   Faire de même pour comparer la distribution de l'IMC en fonction du statut d'hypertension des individus.

10. Créer une variable **NIVAGE** telle que :

    - **NIVAGE** prend 1 si le sujet a moins de 35ans ;

    - **NIVAGE** prend 2 si le sujet a entre 35 et 45ans ;

    - **NIVAGE** prend 3 si le sujet a entre 45 et 65ans ;

    - **NIVAGE** prend 4 si le sujet a entre 65 et 75ans

    - **NIVAGE** prend 5 si le sujet est plus âgé que 75 ans.

Quelle classe d'âge est la plus représentée dans les données ? Décrire à l'aide d'un graphique, la répartition de l'hypertension en fonction des classes d'âge construites uniquement sur les individus de la ville de Buckingham.

© - 2022 – Université de Bordeaux, ISPED

Partie SAS (10 points)

Exercice 1 :

a/ Importez les données du fichier *diabetes.csv*, et gardez le dans une table qu'on peut appeler *diab*.

b/ Corriger, si nécessaire, le type des variables et le codage des données manquantes.

c/ Recoder la variable **gender** en une variable **sexe** codée 2 si femme et 1 si homme.

d/ Créez la variable **imc.** Sachant que l'IMC = $\frac{\text{Poids (kg)}}{(\text{Taille (m)})^2}$ et que 1 livre sterling = 0, 453592 kg et 1 pouce = 0,0254 m.

e/ Créez la variable **hta** qui sera codée 1 si la tension artérielle systolique est supérieure ou égale à 140 mmHg ou si la tension artérielle diastolique supérieure ou égale à 90 mmHg et 0 sinon.

Exercice 2 :

a/ Décrivez les variables : **sexe**, **hta**, **imc**, **chol**, **age** et **location**.

b/ Comparez en fonction du sexe l'âge des sujets. Faites une représentation graphique adaptée pour illustrer ce test.

c/ Comparez en fonction du sexe la distribution de la variable **hta**. Faites une représentation graphique adaptée pour illustrer ce test.

d/ Faite une représentation graphique pour illustrer la distribution du poids en fonction de la taille. Différencier les sujets en fonction du sexe.

Exercice 3 :

Faites un (nouvel) identifiant unique pour chaque sujet de manière à ce qu'avec ce seul identifiant on puisse retrouver, le numéro d'identification du sujet (variable id), le sexe et la localisation.

© - 2022 – Université de Bordeaux, ISPED