Atelier d'introduction à R : Devoir

Comité des initiatives étudiantes du réseau de recherche en santé cardiométabolique, diabète et obésité

Le devoir comprend deux applications : la première avec un jeu de données intégré à R et la deuxième avec vos propres données. Les questions sont de difficulté variable. Vous pouvez toutes les résoudre, ou encore, sélectionner celles-qui semblent les plus appropriées pour vous selon votre niveau de familiarité avec R/RStudio.

Puisque nous n'avons pas vu l'ensemble des tests possibles durant l'atelier, vous pouvez chercher les commandes lorsque nécessaire dans l'onglet Help, ou avec une recherche Google. Vous trouverez aussi à la fin du document une liste de sites web pertinents pour trouver les commandes dans R/RStudio.

Application 1 : jeu de données intégré à R

Ouvrez un nouveau script dans RStudio et appelez le « jeu de données ToothGrowth » avec la commande *data(ToothGrowth)*. Lisez le descriptif de l'étude ToothGrowth en tapant la commande *?ToothGrowth* ou en tapant le nom dans l'onglet Help. À partir de ce jeu de données, effectuez les exercices suivants.

1. Vérifiez la structure du jeu de données.

a) Combien de sujets sont inclus dans ce jeu de données ?

b) Combien y a-t-il de variables et quel est le type de chacune ?

2. Créez deux groupes de sujets : ceux qui ont reçu le supplément de vitamine C (VC) et ceux qui ont reçu le jus d'orange (OJ). Pour le moment, il n'est pas nécessaire de créer de sous-groupes par dose.

a) Rapportez les statistiques descriptives de longueur des dents dans chaque groupe : moyenne, écart-type, médiane, 1er et 3^{e} quartile, minimum et maximum à l'aide de la commande summary().

b) Évaluez la distribution de la variable de longueur des dents dans chaque groupe avec un histogramme.

3. Comparez la moyenne de longueur des dents entre les sujets qui ont reçu la vitamine C et ceux qui ont reçu le jus d'orange. Faites un test d'hypothèse avec le test de comparaison de moyennes approprié selon 2.b. Est-ce que le type de supplément a un effet différent sur la longueur de dents ?

4. Réalisez un modèle de régression linéaire entre la dose (0.5, 1, 2) et la longueur des dents. Sur la base du coefficient bêta estimé, il y a-t-il une association entre ces deux variables ?

5. À l'aide de graphiques à boîte à moustache (box plots), évaluez la distribution de longueur des dents par sous-groupe de l'étude. Est-ce que la distribution des données suggère un effet selon le type de supplément et selon la dose ?

a) Faites tout d'abord des graphiques box plot par groupe de type de supplément (vitamine C et jus d'orange).

b) Ensuite, réalisez des graphiques box plot par sous-groupe de combinaison dose (0.5, 1, 2) et type de supplément (vitamine C et jus d'orange).

Voir URL #1 à la fin du document au besoin

6. Comparez les moyennes de chaque sous-groupe de combinaison dose et type de supplément à l'aide d'un test statistique de comparaison de moyennes pour 3 groupes et plus.

Voir URL #2 à la fin du document au besoin

a) Est-ce qu'une des moyennes diffère significativement des autres ?

b) Si oui, réalisez des tests de Wilcoxon de comparaisons de moyennes avec correction pour la multiplicité des comparaisons.

Application 2 : votre propre jeu de données

Voici l'opportunité de vous pratiquer avec R avec vos propres données. Les questions qui suivent sont un guide pour se familiariser avec des commandes plus courantes. Cela dit, n'hésitez-pas à adapter le choix des modèles statistiques selon vos besoins. Par exemple, vous pouvez réaliser une régression logistique ou un modèle de Cox pour risques proportionnels au lieu de la régression linéaire à la question 5.

1. Importez dans RStudio le jeu de données de votre choix qui comporte au moins une variable dépendante d'intérêt de forme continue.

2. Vérifiez la structure du jeu de données.

3. Rapportez les statistiques descriptives d'une variable dépendante et d'une variable indépendante de votre choix : moyenne, écart-type, médiane, 1er et 3^{e} quartile, minimum et maximum.

4. Évaluez la distribution de la variable dépendante avec un histogramme.

5. Faites un modèle de régression linéaire entre une variable indépendante de votre choix et la variable dépendante. Si votre variable dépendante n'est pas normalement distribuée, faites la transformation de votre choix (ex. échelle log).

6. Vérifiez la distribution des résidus du modèle de régression linéaire, avec un nuage de point et un graphique Q-Q, au minimum avec la commande plot().

Voir URL #3 à la fin du document au besoin

7. Facultatif (avancé) : vérifiez la pertinence d'un terme quadratique par le test d'hypothèse et par l'adéquation du modèle sur les données observées.

Sites web pertinents pour la recherche de commandes en R

En général :

http://www.sthda.com/

https://rstudio.com/resources/cheatsheets/

Pour ce devoir :

#1 Pour les graphiques de l'application 1 : https://rstudio-pubs-static.s3.amazonaws.com/46904_72122070b1d047639f36ef98974a33eb.html

#2 Pour le test Kruskal-Wallis et le test Wilcoxon avec correction pour multiplicité des comparaisons de l'application 1: http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r#multiple-pairwise-comparison-between-groups

#3 Pour l'examen des résidus application 2 : http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/