TP2: graphes, statistiques descriptives et statistiques inférentielles
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Commandes utiles pour le TP2

Commande	Description de la commande
`getwd()`	donne le dossier dans lequel on est en train de travailler
`setwd()`	permet de changer de dossier dans lequel on est en train de travailler
`readRDS()`	lit un fichier RDS placé dans le bon dossier
`str()`	donne la structure d'un jeu de données
`dim()`	donne les dimensions des données (nombre de lignes, puis nombre de colonnes)
`levels()`	affiche les valeurs (catégories) d'une variable qualitative
`attach()`	permet d'utiliser les noms de variables d'un jeu de données directement
`hist()`	crée un histogramme d'une variable quantitative
`summary()`	donne le résumé d'une variable
`sd()`	donne l'écart-type d'une variable quantitative
`mean()`	donne la moyenne d'une variable quantitative
`sqrt()`	calcule la racine carré d'un nombre positif
`qt()`	calcule un quantile de la distribution student t
`qnorm()`	calcule un quantile de la distribution normale
`t.test()`	permet de calculer un intervalle de confiance pour la moyenne (en utilisant les quantiles d'une distribution student t)
`plot()`	crée un graphique (linéaire ou en nuage de points)
`table()`	crée un table de contingence pour deux variables qualitatives

N'oubliez pas que pour obtenir de l'aide sur n'importe quelle fonction, vous pouvez taper `?nomdelafonction` dans la console même.

Données "cdc" : identification des facteurs de risque aux États-Unis

Récupérez le fichier `cdc.RDS`, disponible sur webcampus, et enregistrez-le sur votre ordinateur dans un dossier "TP2". Créez un nouveau script `R` et enregistrez-le dans le même dossier. Ouvrez le fichier sur Rstudio en utilisant la commande suivante :

```
data <- readRDS("cdc.RDS")
```

Attention, votre fichier doit se trouver dans le dossier correspondant au projet actif. Si vous avez un doute concernant la localisation de ce dossier, vous pouvez retrouver le chemin en tapant la commande `getwd()` pour voir où RStudio va chercher les fichiers.

Le jeu de données fait référence aux données collectées par les Centres de contrôle et de prévention des maladies (CDC). Il comprend un échantillon aléatoire de 20 000 personnes de l'enquête BRFSS menée en 2000. Cet enquête téléphonique annuelle est conçu pour identifier les facteurs de risque chez la population adulte aux États-Unis et pour signaler les tendances émergentes en matière de santé. Par exemple, les répondants sont interrogés sur leur régime alimentaire et leur activité physique hebdomadaire, leur statut VIH / SIDA, leur consommation de tabac et même leur niveau de couverture des soins de santé. Cet ensemble de données contient plus de 200 variables, mais nous travaillerons avec le sous-ensemble des variables suivant :

— `genhlth` : l'état de santé général des répondants
— `exerany` : si le répondant a fait du sport au cours du dernier mois ou non
— `hlthplan` : si le répondant avait une assurance santé ou non
— `smoke100` : si le répondant a fumé au moins 100 cigarettes au cours de sa vie ou non
— `height` : la taille en centimètres
— `weight` : le poids en kilogrammes
— `wtdesire` : le poids désiré en kilogrammes
— `age` : l'âge du répondant
— `gender` : le sexe du répondant

Exercices

1. Examinez les données. Quelles sont les dimensions de ce jeu de données et que représentent ces dimensions ?
2. Affichez uniquement la colonne `genhlth` de deux manières différentes. Ensuite, n'affichez que les 10 premiers répondants pour cette variable. Identifiez les valeurs que cette variable peut prendre.
3. Pour chacune des variables, identifiez son type (quantitative ou qualitative). *NB : les analyses et graphiques réalisés sur les variables quantitatives sont différents de ceux réalisés sur les variables qualitatives !*

4. Utilisez `attach()` afin de pouvoir utiliser les noms des variables directement. Créez une nouvelle variable IMC (l'indice de masse corporelle) selon la formule IMC = poids / taille^2, avec la taille en mètres.
5. Créez deux histogrammes de la variable IMC, le premier sans donner d'argument supplémentaire et le second en précisant `breaks = 50`, qu'observez-vous ?
6. Quel est l'IMC moyen et quel est son écart-type ? Trouvez également à partir de quel IMC un répondant appartient au top 25 % des répondants ayant les plus grands IMCs de cet échantillon.

7. Calculez ensuite un intervalle de confiance pour l'IMC en utilisant la formule vue au cours (Slide 20, Chapitre 2). Utilisez un niveau de $\alpha = 0.001$ et interprétez votre résultat. Est-ce qu'il est important dans cet exemple d'utiliser une distribution student t plutôt qu'une distribution normale ?
8. Vérifiez la réponse obtenue à la question précédente avec la fonction `t.test()`. Pour obtenir le bon niveau α, vous devez spécifier l'argument `conf.level`, qui est égal à $1 - \alpha$.
9. Un IMC d'un adulte en bonne santé ne devrait pas dépasser la valeur de 25. Or on suspecte que l'IMC des répondants est plus élevé que cette valeur. Faites un test d'hypothèse avec niveau $\alpha = 0.0001$ pour répondre à cette question : écrivez bien l'hypothèse nulle et l'hypothèse alternative, calculez ensuite la région de rejet associée à ce test, et concluez.

10. La fonction `summary()` s'applique également aux variables qualitatives. Trouvez combien de répondants de cette base de données ont répondu `excellent` pour leur état de santé général. Quelle est la proportion de l'échantillon indiquant être en mauvaise santé ?
11. Trouvez les effectifs des valeurs pour la variable `smoke100`. Peut-on également faire un histogramme de `smoke100` ? Si oui, faites-le ; si non, créez un graphique en barres pour les effectifs de cette variable.
12. Grâce aux arguments de la fonction `table()`, donnez le nombre de participants qui ont fumé au moins 100 cigarettes pour chaque sexe. *NB : ce type de tableau à double entrée (deux variables) s'appelle également une table de contingence.*
13. Le commande `plot()` s'adapte au type de variable qu'on lui passe. Réaliser les graphiques suivants et interprétez-les :

```
plot(smoke100, gender, xlab = "smoke100", ylab = "gender")
plot(genhlth, smoke100, ylab = "smoke100")
plot(genhlth, IMC, xlab = "genhlth", ylab = "weight")
plot(weight, wtdesire, xlim = c(0,200), ylim = c(0,200))
```