Tutoriels de référence :

Ceux cités dans le document principal.

Voir également sous Google avec les mots clés « statistique descriptive r ».

Questions :

On souhaite traiter le fichier « Census.xlsx ».

« Classe » joue un rôle particulier, la variable indique les personnes qui ont un revenu annuel supérieur (more) ou inférieur (less) à une valeur donnée.

Des indications sur les commandes à utiliser sont données. Après, si vous avez des solutions qui semblent plus appropriées, vous pouvez les utiliser également.

1. Si ce n'est pas déjà fait, installer le package « xlsx ». Le charger par la suite (library).
2. Charger le fichier « **census.xlsx** » sous R (read.xlsx)

Que constatez-vous au chargement des données ? Pour remédier à cet écueil, interrompre l'opération. Puis ouvrir le fichier « census.xlsx » dans Excel et l'exporter le au format texte avec séparateur tabulation (census.txt)

3. Charger le fichier « census.txt » sous R (read.table, attention aux options)
4. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ? (ncol, nrow)
5. Afficher le résumé des données (summary)
6. Essayer de répondre aux différentes questions suivantes : quelle est la *proportion* des hommes (sex = male) ? celle des « classe = more » ? (table ou filtrage + calcul)
7. Quelle est le nombre de personnes travaillant pour le gouvernement (workclass contenant le terme ''gov'') ? (grep + table)
8. Construire le diagramme à bandes pour les variables « marital_status » et « relationship » (http://www.statmethods.net/graphs/bar.html).
9. Pour les mêmes variables, construire les diagrammes à secteurs (http://www.statmethods.net/graphs/pie.html).
10. « Education » correspond en réalité à un niveau d'éducation atteint. C'est donc une variable qualitative ordinale avec les modalités suivantes {Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Bachelors, Masters, Prof-school, Doctorate}. Quelle est la proportion des personnes qui ont uniquement le niveau « Preschool » ? (table)
11. Quelle est la proportion de personnes qui ont au moins le niveau « Bachelors » ? (table + astuce)
12. Croiser les variables « classe » et « sex ». Quelle est la proportion des « more » dans l'échantillon global ? Parmi les hommes ? Parmi les femmes ? Est-ce que ce résultat nous permet de conclure que le niveau de revenu est différent selon que l'on est un homme ou une femme ? (table

permet de croiser deux variables, on a une matrice – type matrix - qu'on peut indicer de différentes manières, on peut aussi effectuer des calculs récapitulatifs)

13. Calculer le KHI-2 du tableau croisé entre « classe » et « sex » (chisq.test, pas de correction de continuité). Puis en déduire le v de Cramer (ex. http://www.r-bloggers.com/example-8-39-calculating-cramers-v/ ; l'objet généré par chisq.test possède la propriété statistic que l'on peut exploiter).
14. Confronter le résultat obtenu avec ce que fourni le package « lsr » (fonction cramersV) (https://cran.r-project.org/web/packages/lsr/lsr.pdf). Vos résultats concordent-ils ?
15. Croiser maintenant « relationship » et « marital status ». Pour chaque valeur de « relationship », quelle est la modalité de « marital status » qui lui est le plus associé ? Et inversement ? Est-ce que la relation est symétrique ? (*il faut appliquer un which.max pour chaque modalité ligne (et colonne pour la 2ème partie de la question)* ; *avant de se lancer dans une boucle, voir du côté de apply* – (https://eric.univ-lyon2.fr/~ricco/cours/slides/tableaux%20et%20matrices%20avec%20r.pdf *pages 12 et 13*).
16. Penchons-nous maintenant sur la variable « age ». Calculer sa moyenne et son écart-type (mean, sd).
17. Centrer et réduire « age » (scale). Recalculer la moyenne et l'écart-type sur les données transformées. Que constate-t-on ?
18. Calculer la médiane et les quartiles d'ordre 1 et 3 des variables (median, quantile).
19. Construire le graphique BOXPLOT (boîte de Tukey) pour la variable « âge » (boxplot). Que remarque-t-on ?
20. Produire l'histogramme de la variable âge (http://www.statmethods.net/graphs/density.html)
21. Calculer la corrélation entre « age » et « hours per week » (cor). Peut-on dire que ces deux variables sont liées ? Réaliser le graphique nuage de points entre ces deux variables pour affiner votre réponse. Que conclure ?
22. Construire le boxplot de « âge » selon « relationship ». Il y a des choses à remarquer dans ce graphique ? (http://www.statmethods.net/graphs/boxplot.html)
23. Calculer la moyenne de l'âge pour chaque valeur de « relationship » (tapply). Le calcul confirme l'impression laissée par le graphique précédent ?
24. Calculer le rapport de corrélation η^2 entre « âge » et « relationship » (voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf ; section 4.6) (var permet de calculer la variance d'une variable)