TD 3

Correlation (suite), identifier les outliers, test d’hypothese

M2 – semestre automne 2021

3pts total

Reponds aux questions suivantes et envois moi une copie electronique de tes reponses par email avant Lundi 6/12 a 23h00. Les reponses peuvent etre courtes et peuvent etre sous la forme de bullet points sauf si indication contraire. Je vous encourage a faire le travail en groupe et a discuter – cela dit assures toi que tu comprends les reponses que tu ecris. Tu peux inclure des screenshots des outputs R ou bien tu peux aussi m’envoyer ton script R avec ce word doc (ou juste le script R si tu as repondu aux questions ecrites dessus en commentaire).

Probleme 1 – Le quartet d’Anscombe. (1 point)

Le quartet d’Anscombe est compose de quatre sets de donnees avec des proprieties particulieres. Pour ce premier probleme nous allons explorer ce set de donnees. Pour cette exercice, tu auras besoin du fichier anscombe.csv (disponible sur discord). Commence par importer ce fichier dans ton espace de travail sur RStudio (soit avec read.csv, soit avec le bouton Import Dataset)

1. Quelles sont les variables contenues dans ces sets de donnees, et de quel type sont elles? **(0.10 pts)**
2. Complete le tableau suivant (montre moi quel functions tu as utilise dans R pour le faire). Que peut on dire des quatres set de donnees en se basant sur les informations de ce tableau? **(0.10)**

|  |  |  |
| --- | --- | --- |
| Dataset | Moyenne | Ecart type |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |

1. Fais un graph de ces quatres set de donnees en utilisant ggplot. **(0.25 – 0.5 total)**
   1. Change le theme du graph **(+0.05)**
   2. Ajoute en titre “Le quartet d’Anscombe” et change la taille du texte a 12pt **(+0.05)**
   3. Change la couleur des points **(+0.05)**
   4. Ajoute un ligne de regression sur les graphs **(+0.10)**
2. Comment peut on caracteriser la relation entre X et Y dans le dataset 1? **(0.10)**
3. Quel aspect important de l’analyse de donnees nous enseigne le quartet d’Anscombe? **(0.20)**

Probleme 2 – Dataset de Ponari et al., (2018), trouve sur osf.io (<https://osf.io/teyqn/>) et disponible sur le discord du cours. L’etude a pour titre “Learning abstract words and concepts: insights from developmental language disorder”. Les auteurs ont compare les competences en decision lexicale pour des mots abstraits vs. concrets pour des enfants avec un trouble du development du language (note DLD dans le dataset) vs typically developing children (TD). Comme pour le probleme 1, commence par lire le fichier csv dans ton espace de travail (read.csv ou Import Dataset).

1. Combien y-a-t-il de variables contenu dans ce dataset? Combien de participants? **(0.05)**
2. Pour chaque participant, les auteurs ont mesure:
   1. Leur capacite en vocabulaire avec le British Picture Vocabulary Scale (BPVS)
   2. Leur capacite de raisonnement avec les matrice de Wechsler (reasoningSkills)
   3. Leur age
   4. Pour les participants avec un trouble du development, leur capacite en langue avec le Clinical Evaluation of Language Fundamentals test (CELF

On va s’interesser au vocabulaire (BPVS), raisonnement (reasoningSkills) et age.

1. On veut deja s’assurer que, comme les auteurs l’affirment, les participants sont tous dans une tranche d’age similaire.
   1. Calcule la moyenne d’age de chaque groupe. Que peut on en deduire? **(0.10)**
   2. Fais un graph avec ggplot qui a pour ordonnee l’identifiant des participants (Subject) et pour abscisse leur age (age). Decris le graph brievement**. (0.10)**
      1. Maintenant donne une couleur au point en function du groupe auquel les participants appartiennent. Hint: il faut ajouter color=group quelque part. En regardant ce graph, que penses-tu de la distribution de l’age des participants dans chaque groupe – est elle equivalente selon toi (pas de bonne ou mauvaise reponse)? **(0.15).**
2. Maintenant on veut voir les relations entre raisonnement, vocabulaire et age des participants.
   1. Cree un subset de la base de donnees qui ne contient que les colonnes “reasoningSkills”, “BVPS”, et “age”. **(0.20)**
   2. Fais un graph de BVPS en function des reasoningSkills avec ggplot **(0.10)**
      1. Change la couleur des point en function de l’age **(+0.10)**
      2. Change les labels des axes X et Y en “Capacite de raisonnement” et “Capacite de Vocabulaire” (+0.05).
   3. Cree un matrice de correlation avec R et complete le tableau (ou poste un screenshot):

|  |  |  |  |
| --- | --- | --- | --- |
|  | reasoningSkills | BPVS | age |
| reasoningSkills |  |  |  |
| BPVS |  |  |  |
| age |  |  |  |

Avec tes propres mots, explique ce que represente le coefficient de correlation entre age et BPVS (i.e., que nous indique-t-il?) **(0.25)**

Probleme 3. Questions ecrites sur le cours (pas de R)

1. Au debut de la recherche sur la lecture, des etudes ont trouve une correlation entre les mouvements des yeux des lecteurs et leur capacite en lecture. Plus particulierement, les mauvais lecteurs avaient des mouvements occulaires plus erratiques (par exemple, plus de mouvements de droite a gauche, plus de fixations sur une ligne,etc…) compare aux bons lecteurs. Les chercheur.euses ont fait l’hypothese que les problemes occulomoteurs etaient la cause des problemes de lecture. De nombreux programmes d’entrainement des mouvements des yeux furent developpes. Est ce que les resultats des etudes justifiaient cette hypothese et le developpement des ces programmes? Pourquoi? (0.25)
2. Imagine que tu prepares une experience sur le lien entre les deficits attentionels (e.g., ADHD) et la vitesse d’acquisition de la lecture en comparant un groupe d’enfants qui ont ces deficits a un groupe d’enfants qui ne les ont pas. Tu fais l’hypothese H1 que les enfants avec des deficits d’attention seront plus lent a acquirir la lecture que leur pairs qui n’ont pas ces deficits.
   1. Quelle est l’hypothese nulle dans ce cas? **(0.25)**
   2. Ton experience faite, les resultats ne te permettent pas de rejeter l’hypothese nulle (i.e., tu a trouve une p-value superieure a 0.05). Mais la realite est qu’il existe vraiment un effet des deficits de l’attention sur la lecture. Tu as fais une erreur de quelle type? **(0.25)**
3. Qu’est ce que la p-value? (0.25)