TD5

Test d’hypotheses, ANOVA

M2 – semestre automne 2021

5pts total

Reponds aux questions suivantes et envois moi une copie electronique de tes reponses par email avant le 10 Janvier 2022. 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).

Questions ecrites sur le cours (pas de R). (1 point)

1. Un groupe de chercheur.euses cherchent a evaluer les differences de production des consonnes voisees chez les francophones monolingues et les francophones bilingues avec l’anglais. Pour cela ils enregistrent 20 monolingues et 20 bilingues prononcant le son /ba/. Iels mesurent ensuite, pour chacun des deux groupes, la moyenne des voiced onset time (VOT), une caracteristique acoustique importante pour le voisement. Quel test statistique doivent iels utiliser pour comparer les VOT moyens des deux groupes? (0.25)
2. Dans quelle situation doit on utiliser un “paired” t-test? (0.25)
3. Une etude evalue les capacites de reconnaissance vocale de Siri, Alexa et Ok Google. Chaque assistant vocale est teste avec une suite de phrases plus ou moins compliquees et recoit une note entre 1 et 10 pour chaque. Siri recoit une moyenne de 9.5, Alexa de 8.7 et Ok Google de 8.5. L’etude reporte un test ANOVA significatif (p<0.05). Les auteur.ices en concluent que Siri est l’assistant le plus performant en reconnaissance vocale. Est il possible de tirer cette conclusion du test statistique utilise? (0.25)
4. Lorsqu’on obitent une p-value inferieure a 0.05, que peut on dire de la probabilite d’avoir obtenue nos donnees si l’hypothese nulle etait vraie? (0.25)

Question en R (4 points).

Cet exercice est a faire avec la base de donnees orthoProject. Le but de cette etude etait d’explorer les effets de l’orthographe sur la perception de la parole. En effet, on sait que certaines formes orales sont ‘preferees’ par les gens meme quand elles sont tres rares. Par exemple, un participant francais va reagir plus vite sur une tache de decision lexicale a la pronunciation *bracElet* (avec un schwa au milieu, forme canonique) malgre le fait que *bra/s/let* (forme reduite) est beaucoup plus commune dans la langue. Une hypothese pour expliquer cette preference pour la forme canonique est qu’elle correspond a la forme orthographique. En d’autres termes, notre connaissance de la forme orthographique des mots influe sur notre perception de ceux ci, meme a l’orale. Dans cette experience, les participants ont appris des nouveaux mots (e.g., *byddo*, *senno*, *troddy*…) associes a des orthographes (soit canonique: “bytto”, soit reduite: “senno”. Soit aucune: none). Ensuite on a mesure si l’orthographe apprise influe sur la perception de la pronunciation canonique des mots (i.e., *bytto*, alors que le mot etait generalement prononce *byddo*). Les variables importantes pour l’exercice sont:

* ID: un identifiant unique pour chaque participant
* Spelling.learned: l’orthographe que les participants ont apprise pour chacun des mots nouveaux (e.g., “bytto” pour *byddo*).
* Pronunciation: les differentes prononciations des mots avec lesquelles les participants furent teste (pour les nouveaux mots/new words et les mots familiers/familiar words mais aussi les nonmots/nonwords)

1. Importe le fichier orthoProject.csv dans R. (0.5)
   1. Liste les variables incluent dans ce fichier et leur type.
   2. Transforme les variables ID, group, pronunciation et spelling.learned en facteur.
   3. Combien de niveau(x) ont ces facteurs? Combien y a t’il de participant dans cette experience?
2. Exploration des donnees, calcul de moyenne et representation graphique (2 pt)
   1. On veut savoir si certains participants ont mal fait la tache (i.e., outliers). On sait que la performance sur les mot connus (Familiar words) et les non mots (Nonwords) doit etre assez haute (plus que 50% correct). Calcule la moyenne de la variable “correct” en function des variables **pronunciation** et **ID**. Attention, on ne s’interesse QUE aux niveaux Nonwords et Familiar words de pronunciation (i.e., pense a faire un subset).
   2. Quels sont les participants qui ont une performance inferieure a 50% (i.e., correct < 0.5). Explique comment tu as fais pour trouver (plusieurs facon de faire).
   3. Enleve ces participants du dataset (soit en creant un subset, soit en les enlevant directement du dataset original).
   4. Maintenant on s’interesse a la performance des participants en fonction de l’orthographe apprise (i.e., spelling.learned):
      1. On ne s’interesse qu’aux niveaux canonical, reduced et none de la variable spelling.learned. Cree un subset qui ne contient que ces trois.
      2. Calcule la moyenne de correct par spelling.learned, store dans un objet appele meanSpelling
      3. Calcule la moyenne de correct par spelling.learned ET ID (i.e., par participant); store dans un objet appele meanSpelling.id
      4. Cree un bar plot (geom\_bar) qui represente les moyennes meanSpelling et avec des points representant les moyennes pour chaque participant dans meanSpelling.id.
3. Maintenant on veut analyser les resultats pour savoir si il existe une difference significative de performance en function de l’orthographe apprise. (1.5 pt)
   1. Sachant qu’on a une variable independante qualitative (spelling.learned) a trois niveaux (canonical, reduced, none) et une variable dependante continue (percent correct), quel test statistique est approprie?
   2. Conduit ce test en R. Fais un imprime ecran du resultat et interprete le.
   3. Ce test permet il de savoir lesquels des trois groupes different entre eux? Si non, utilise un test post-hoc pour le savoir en prenant en compte les questions d’ajustement de la p-value.
   4. Qu’en conclut on sur l’influence de l’orthographe sur la perception de la prononciation canonique des mots?