Département STID - IUT de Paris
Deuxième année - FA - VCOD 2022-2023

Tests paramétriques

TP : Tests de comparaison à deux échantillons et test du chi-deux
Date limite de rendu du compte-rendu : dimanche 26 février 2023

Instructions : Un compte-rendu du TP rédigé sous la forme d'un fichier pdf est à rendre par étudiant. Vous devrez déposer votre compte rendu sous moodle. Chaque étudiant remettra un document pdf ayant pour nom *nometudiant_tptest.pdf*.

Exercice 1

Une étude sur l'insuffisance pondérale des nouveaux nés à été menée en 1986 aux États-Unis (Baystate Medical Center, Springfield, Massachusetts) auprès de 189 mères. Les données sont disponibles sur le commun et sur moodle sous le nom de `birthwt.dat`. On qualifie dans la suite un bébé d'hypotrophe si son poids de naissance est inférieur à 2 500 grammes.

Les variables de la base sont les suivantes :

- low : indicateur d'insuffisance pondérale (1 si poids du bébé inférieur à 2 500 g, 0 sinon)
- age : âge de la mère (en années)
- smoke : tabacologie de la mère durant la grossesse (1 si fumeur, 0 si non fumeur)
- lwt : poids de la mère au début de la grossesse (en livre)
- ht : hypertension pendant la grossesse (1 si hypertension, 0 sinon)
- ui : infection urinaire au cours de la grossesse (1 si infection, 0 sinon)
- bwt : poids du bébé à la naissance (en grammes).

1. Faire une étude **univariée** descriptive de la base de données (univariée signifie que l'on regarde chaque variable séparément). On présentera les indicateurs statistiques et les graphiques pertinents pour chaque variable. Conclure avec une brève description des variables qui constituent cette base de données.
2. On s'intéresse ici au poids du bébé à la naissance (bwt).
   (a) Étudier de manière descriptive le lien entre la variable poids du bébé à la naissance et toutes les autres variables de la base de données (citées ci-dessus). On présentera les indicateurs statistiques et les graphiques pertinents pour chaque couple de variables. Commenter les différents liens observés avec la variable poids du bébé.
   (b) Le but maintenant est de tester le lien entre la variable poids du bébé à la naissance et toutes les autres variables de la base de données. Pour cela, présenter la méthodologie statistique pour chaque test envisagé : donner les conditions d'applications des tests que vous devez vérifier, le type de test à utiliser selon le type de situation etc.

      Remarque : pour chaque couple de variables, on proposera tous les tests possibles, parmi ceux qui ont été vus en cours.
   (c) Appliquez tous les tests envisagés dans la question précédente. Vous pouvez résumer vos résultats dans un tableau contenant les p-valeurs des tests.

(d) Concluez : d'après vos différentes analyses, quelles sont les variables qui influent sur le poids du bébé à la naissance ?

3. On s'intéresse à présent à la variable indicateur d'insuffisance pondérale (low). Refaire toutes les questions précédentes 2. (a), (b), (c) et (d) pour cette nouvelle variable.
4. Une des limitations de la méthodologie proposée dans cet exercice est que les comparaisons sont toutes faites deux à deux. Cela peut poser des problèmes si certains variables sont liées entre elles. Par exemple, on peut observer sur le jeu de données que les variables hypertension (ht) et poids de la mère (lwt) semblent liées. Il est possible alors d'observer un impact du poids de la mère sur le poids du bébé à la naissance simplement parce que les poids les plus élevés (de la mère) correspondent aux hypertensions les plus élevées sans que le poids de la mère ne soit réellement responsable du poids du bébé à la naissance. C'est ce que l'on appelle un **facteur de confusion**. Montrez sur les données l'existence de ce facteur de confusion.

   Existe-t-il une méthode statistique multivariée qui permette de modéliser le poids du bébé à la naissance en fonction de toutes les autres variables en même temps ? Et donc de prendre en compte des éventuels facteurs de confusion. On ne demande pas d'implémenter cette méthode.

Exercice 2

On s'intéresse à la base de données `employes.csv` disponible sous moodle. On veut étudier l'influence du sexe sur la catégorie d'employé dans cette base de données.

1. Quelle méthodologie permet d'étudier le lien entre sexe et catégorie d'employés ?
2. Donner le tableau de contingence de la catégorie et du sexe.
3. Quelle est la proportion d'employés qui sont des hommes parmi les cadres ? Parmi les secrétaires ?
4. Quelle est la proportion d'employés qui sont des secrétaires parmi les hommes ? Parmi les femmes ?
5. Proposer un graphique permettant de visualiser le lien entre le sexe et la catégorie d'employés. Mettre en oeuvre le test permettant d'étudier le lien entre ces 2 variables. Quelle est la statistique de ce test, sa loi sous l'hypothèse (H_0), sa valeur produite par R ?
6. La catégorie des employés est-elle liée à leur sexe ? Justifier votre réponse à partir de vos analyses.