Université Grenoble Alpes
L2/STA351

TP7 : Tests de comparaison de deux échantillons

Objectifs : savoir utiliser les tests de comparaison de deux échantillons appariés ou indépendants.

Exercice 1
On utilise le jeu de données `titanic.csv`.

1. Charger le jeu de données. Appeler les variables `pclass`, `survived`, `gender` et `age` respectivement `P`, `S`, `G` et `A`.
2. Calculer les indicateurs statistiques de `A` selon les valeurs de `G` :
   `by(A,G,summary)`
3. Tracer les boites à moustaches de `A` pour chaque groupe défini par les valeurs de `G` :
   `boxplot(A~G)`
4. Extraire les ages des femmes et les ages des hommes et appeler `Af` et `Am` ces deux échantillons.
5. On veut tester si l'age moyen des hommes est différent de celui des femmes. Définir les hypothèses nulle et alternative, notées $\mathcal{H}_0$ et $\mathcal{H}_1$ et mettre en oeuvre le test en calculant la statistique de test et la p-valeur. Comparer les résultats obtenus avec ceux de la fonction de test de student de R `t.test`. Quelle est la conclusion du test ?
6. On veut savoir si l'âge moyen des femmes est inférieur à celui des hommes. Utiliser la fonction `t.test` avec l'option `alternative="less"`. Conclure.
7. On veut à présent effectuer le test de comparaison des moyennes en utilisant des échantillons de petite taille.
   (a) Extraire les 20 premiers termes de `Af` et les 15 premiers termes de `Am` et les affecter respectivement à `Afr` et `Amr`.
   (b) Quelles conditions doit-on poser sur les variables âge des femmes et âge des hommes pour pouvoir mettre en oeuvre le test de comparaison des moyennes ?
   (c) Sous ces conditions quelle est la statistique de test utilisée et sa loi sous $\mathcal{H}_0$?
   (d) Calculer la statistique puis la p-valeur du test.
   (e) Retrouver les résultats précédents en appliquant la fonction `t.test` aux deux échantillons `Afr` et `Afm` avec les choix adaptés des options `var.equal` et `paired`.
   (f) Vérifier à présent que la condition posée sur les variances des deux variables considérées est raisonnable à l'aide du test de comparaison des variances. Préciser les deux hypothèses du test réalisé ainsi que les conditions requises sur les variables pour l'appliquer. Calculer ensuite la valeur de la statistique de test puis la p-valeur. Retrouver ces résultats avec la fonction `var.test`.
8. Recommencer en remplaçant le genre `G` par le fait de survivre `S` et en utilisant les échantillons complets. Est ce que l'âge des survivants est le même que l'âge des non survivants ?
9. On veut savoir si l'âge moyen est ou non le même dans les trois classes. Tracer les boites à moustaches de l'âge selon les classes.
10. Extraire les âges des 3 classes.
11. Est ce que les passagers de première classe sont plus âgés que ceux de deuxième classe ? que ceux de troisième classe ? Est ce que ceux de deuxième classe sont plus âgés que ceux de troisième classe ?

Exercice 2

On utilise le jeu de données `her.txt`.

1. Charger le jeu de données. On étudie la variable `BMI` (indice de masse corporelle) et on veut comparer l'indice entre les hommes et les femmes. Extraire les deux échantillons des hommes et des femmes et les appeler `BH` et `BF`.
2. Calculer les principaux indicateurs statistiques de ces deux variables, et proposer une représentation graphique permettant de les comparer.
3. On cherche à tester si la différence moyenne entre le BMI des hommes et celui des femmes est nulle ou non. Définir l'hypothèse nulle et alternative, notées $\mathcal{H}_0$ et $\mathcal{H}_1$.

   Mettre en place le test à l'aide de la fonction `t.test`. Interpreter.
4. Recommencer en testant la différence de pression systolique entre patients traités et non traités.
5. On veut à présent savoir si en en moyenne la pression systolique diffère de la pression diastolique. Faire le test permettant de répondre à ce problème. Préciser les conditions sur les variables permettant d'appliquer ce test, définir les deux hypothèses testées $\mathcal{H}_0$ et $\mathcal{H}_1$, calculer la statistique de test puis la p-valeur. Retrouver vos résultats à l'aide de la fonction `t.test` et conclure.
6. Quel test unilatéral faudrait-il faire pour préciser la conclusion précédente ?

Exercice 3

On utilise le jeu de données `cardiaque.csv`.

1. Charger le jeu de données. On étudie la variable `systolique` (pression sanguine systolique) et on veut comparer l'indice entre les personnes hyperactives et les autres. Extraire les deux échantillons des hyperactifs ou non et les appeler `SHA` et `SA` (les individus hyperactifs sont ceux pour lesquels la variable `ActivitéBinaire` prend la modalité 1).
2. Calculer les principaux indicateurs statistiques de ces deux variables, et proposer une représentation graphique permettant de les comparer.
3. On cherche à tester si la moyenne entre la pression systolique des hyperactifs est supérieure à celle des non hyperactifs. Définir les hypothèses nulle et alternative, notées $\mathcal{H}_0$ et $\mathcal{H}_1$.

   Mettre en place le test à l'aide de la fonction `t.test`. Interpréter.
4. Recommencer en testant la différence de pression systolique entre patients traités et non traités (les patients traités sont ceux pour lesquels la variable `cardiaque` prend la modalité 1).