Université de Montpellier
2018/2019

HLMA408: Traitement de données
Joseph Salmon, Emmanuel Bonnet

TP n° 3 : Inférence sur échantillons gaussiens

Objectifs du TP : savoir réaliser des tests simples en `Python` sur des populations suivant des lois gaussiennes.

1 Inférence sur la moyenne d'un échantillon

1.1 Test de Student

On se place ici dans le cadre où le statisticien dispose d'un seul échantillon à analyser. On note X la variable aléatoire du modèle. L'espérance $\mathbb{E}(X)$ est inconnue. Soit μ_0 un réel fixée. On désire tester

$$\mathcal{H}_0 \ : \ \mathbb{E}(X) \ = \ \mu_0 \quad \text{contre} \quad \mathcal{H}_1 \ : \ \mathbb{E}(X) \neq \mu_0 \ ,$$

à l'aide du test de Student. Les tests de Student s'effectuent à l'aide de la commande `t.test`. Taper `help(t.test)` pour obtenir de l'aide sur cette commande.

En 1879, le physicien américain Michelson a fait plusieurs expériences pour vérifier la valeur de la vitesse de la lumière c proposée par le physicien français Cornu en 1876. La valeur proposée par Cornu était $c = 299\,990$ km/s. Michelson a obtenu 20 mesures pour la vitesse de la lumière (les valeurs données sont les valeurs mesurées par Michelson auxquelles on a soustrait 299 000 afin de ne pas avoir à manipuler des chiffres trop grands) :

Ces 20 observations peuvent être considérées comme les valeurs observées de 20 variables aléatoires ayant une espérance commune mais inconnue $\mathbb{E}(X)$. Si les conditions expérimentales pour mesurer la vitesse de la lumière sont sans biais, il est alors raisonnable de supposer que $\mathbb{E}(X)$ est la vraie vitesse de la lumière.

1) Télécharger les données[1] du fichier `"michelson.txt"` et les enregistrer dans le répertoire `~/HLMA408/TP2`
2) Importer les données à l'aide de la commande `read.table` et nommez la base de données `michelson`
3) Représenter les données sous forme d'histogramme et commenter en particulier l'hypothèse de données gaussiennes.
4) En supposant que la variance théorique est connue et vaut $\sigma = 105$, tester si les mesures de Michelson confirme la valeur de la vitesse de la lumière proposée par Cornu. On pourra regarder la p-value obtenue par $2 - 2\Phi\left(\frac{|\mu-\mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)$.
5) Sans faire l'hypothèse que la variance est connue utiliser la commande `t.test` pour tester si les mesures de Michelson confirme la valeur de la vitesse de la lumière proposée par Cornu. Quelle est la conclusion du test ?
6) Donner un intervalle de confiance pour la moyenne, aux niveaux 0.95 et 0.90.
7) Refaire le test pour confirmer la vitesse calculée par Cornu en créant une variable `michelson$true_speed` contenant la vraie valeur des mesures. Cela change-t-il la conclusion ?

1. Données disponibles ici : http://josephsalmon.eu/enseignement/datasets/michelson.txt

2 Comparaison de deux échantillons indépendants

On se place dans le cadre où le statisticien dispose de deux échantillons indépendants à analyser. On note μ_1 l'espérance de la variable dans le modèle de la première population et μ_2 l'espérance de la variable dans le modèle de la seconde. Bien-sûr, ces deux paramètres sont inconnus. On souhaite tester

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 \ : \ \mu_1 \neq \mu_2$$

avec un test de Student. On pose $\mu_{\text{diff}} = \mu_1 - \mu_2$.

Chercher dans l'aide de la commande `t.test` comment comparer la moyenne de deux échantillons indépendants avec R.

Dans une étude[2] sur les mécanismes de détoxication, on dispose de la concentration du DDT et de ses dérivés, DDD et DDE, (en mg/g) contenus dans des brochets du Nord (*Esox lucius*), capturés dans la rivière Richelieu (province de Québec). Les données en question sont relatives aux brochets de 2 et 3 ans.

1) Ouvrir le fichier de données et regarder comment il est organisé.
2) Importer le jeu de données et nommez le `brochet`
3) Re-coder la variable age à l'aide de la fonction `replace` pour avoir 2 et 3 dans la variable `age` à la place de `"deux_ans"` et `"trois_ans"`
4) Calculer et commenter les statistiques résumées obtenues avec les commandes ci-dessous :

```
summary(brochet$conc[brochet$age == "2"])
summary(brochet$conc[brochet$age == "3"])
```

5) Afficher un graphique en "violon" de la concentration pour chacun des âges.
6) Effectuer un test de Student pour deux échantillons indépendants avec la commande `t.test` et interpréter les résultats.

2.1 Test d'égalité des variances : les brochets

Lorsque l'on a deux échantillons indépendants, pour déterminer le choix correct pour l'option `var.equal` dans l'utilisation de la commande `t.test`, il est préférable de tester au préalable l'égalité de variance par un test de Fisher.

Pour σ_1^2 la vraie variance de la première population et σ_2^2 la vraie variance de seconde population, la statistique de ce test sous l'hypothèse nulle :

$$\mathcal{H}_0 \ : \ \sigma_1^2 \ = \ \sigma_2^2 \ ,$$

est la suivante :

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1) \ ,$$

où $\mathcal{F}$ est la loi de Fisher[3]. Pour effectuer ce test on utilise la commande `var.test(ech1, ech2)` où `ech1` (resp. `ech2`) est le vecteur des données du premier (resp. deuxième) échantillon.

Pour appliquer correctement ce test, il faut vérifier que les deux échantillons proviennent de distributions normales.

Par défaut, l'hypothèse alternative est $\mathcal{H}_1 \ : \sigma_1^2 \neq \sigma_2^2$ (cas bilatéral).[4]

On reprend l'exemple des brochets vu précédemment.

2. Données disponibles ici : http://josephsalmon.eu/enseignement/datasets/brochet2.dat
3. La loi de Fisher est décrite ici : https://fr.wikipedia.org/wiki/Loi_de_Fisher
4. On peut changer d'hypothèse alternative pour des tests unilatéraux : pour avoir $H_1 \ : \sigma_1^2 > \sigma_2^2$ (resp. $H_1 \ : \sigma_1^2 < \sigma_2^2$)

7) Pour l'égalité des variances à partir des échantillons de l'exemple sur les brochets dans la deuxième partie (avec le jeu de données **brochet2.dat**), créer les deux échantillons avec

```
ech1 <- brochet$conc[brochet$age == "2"]
ech2 <- brochet$conc[brochet$age == "3"]
```

8) Lancer le test d'égalité des variances avec `var.test(ech1, ech2)` et conclure.

Remarque. Rappelons que si l'hypothèse d'égalité des variances est validée, on peut utiliser des tests de Student reposant sur l'hypothèse d'égalité des variance' pour accepter ou rejeter l'égalité des moyennes. Pour faire un tel test avec R, il faut utiliser l'option `var.equal=TRUE` lorsque l'on lance la commande `t.test`.

9) Reprendre le test de Student d'égalité des moyennes avec l'option `var.equal=TRUE` ou `var.equal=FALSE` en fonction de la conclusion sur l'égalité des variances.

```
t.test(brochet$conc[brochet$age == "2"],
       brochet$conc[brochet$age == "3"], var.equal= ???)
```

2.2 Comparaison de la pollution sur Toulouse et sur Montpellier

Nous allons maintenant utiliser des données de pollution recueillies sur diverses communes de l'Occitanie [5] entre

10) Charger les données dans R dans un dataframe que vous nommerez `pol_occ`. (attention aux séparateurs !)

11) Observer ce que donne les commandes suivantes :

```
unique(pol_occ$polluant)
unique(pol_occ$nom_com)
```

12) Compléter le script suivant pour visualiser les violons des divers polluants sur Montpellier et Toulouse sur la période étudiée :

```
library(vioplot)
pol <- c("O3", "PM10", "NO", "NO2", "PM2.5")
pol_toulouse_montpellier <- subset(pol_occ, nom_com == c("TOULOUSE",
 "MONTPELLIER"))
par(mfrow=c(1,length(pol))
for (i in 1:length(pol)){
  pol_i <- cbind(as.character(pol_toulouse_montpellier$nom_com),
                 pol_toulouse_montpellier$valeur_originale
                 )[pol_toulouse_montpellier$polluant==pol[i],]
  vioplot(???????,main=pol[i])
}
```

13) En utilisant une boucle `for`, testez l'égalité des niveaux des pollutions pour les cinq polluants vus précédemment.

5. Données disponibles ici : https://moodle.umontpellier.fr/pluginfile.php/665853/mod_folder/content/0/Mesure_journaliere_Region_Occitanie_Polluants_Principaux.csv?forcedownload=1

3 Comparaison d'échantillon appariés

On se place dans le cas où le statisticien dispose de deux échantillons appariés à analyser. On désire faire le test (de Student) de l'hypothèse nulle

$$H_0 \ : \ \mu_{\text{diff}} \ = \ 0$$

avec $\mu_{\text{diff}} = \mu_1 - \mu_2$ où μ_1 est la vraie moyenne (inconnue) de la première variable dans la population et μ_2 est la vraie moyenne (inconnue) de la seconde variable dans la population.
Manipulation avec R. Il faut utiliser la commande `t.test` avec `paired=TRUE`. Afficher la page d'aide de cette commande en tapant `help(t.test)`. On peut préciser la valeur de μ_{diff} que l'on désire tester dans l'hypothèse nulle. Par défaut, on teste l'égalité des moyennes, c'est à dire avec $\mu_{\text{diff}} = 0$.

3.1 Traitement des eaux usées : différences entre deux filtres

Dans une étude sur le traitement des eaux usées, l'efficacité de deux filtres, l'un en fibre de verre (variable **verre**) et l'autre en papier filtre Whatman numéro 40 (variable **papier**) a été testée. Sur des prélèvements de 200 millilitres d'eau provenant d'usine de pâte à papier, la quantité de solides en suspension retenus par les deux filtres a été mesurée. Les résultats de ces analyses sont contenus dans la fichier **filtre.dat**.

14) Téléchargez les données[6] du fichier `filtre.dat`. Ouvrez le jeu de données et regardez la manière dont il a été saisi. L'importer dans R et mettre en forme les données

15) Créez une variable **delta** égale à la différence entre les deux mesures. Pour cela, tapez dans la fenêtre de commande :

```
filtre$delta <- filtre$verre - filtre$papier.
```

16) Représentez la distribution de cette variable `filtre$delta` à l'aide d'un histogramme.

17) Faire un test de Shapiro-Wilk[7] sur cette dernière variable à l'aide de la fonction `shapiro.test` et conclure.

18) Effectuez le test de Student d'égalité des moyennes pour ces deux échantillons appariés. Commentez.

3.2 Comparaison de mesures de hauteur d'un arbre : avec ou sans abattage

L'objectif de l'étude[8] décrite ci-dessous est de valider ou non une nouvelle technique de mesure de la taille des arbres sur pied (non abattus) en prenant un risque de 5%.

Dans une forêt, on choisit des arbres au hasard que l'on mesure sur pied (debout). Ensuite, on les abat puis on les mesure à nouveau. Chaque arbre a donc été mesuré deux fois. On veut tester l'égalité des moyennes de ces deux séries pour comparer les deux méthodes de mesure.

19) Importer ces données dans R.

20) Créez une variable **delta** égale à la différence entre les 2 mesures et représentez la densité de cette variable. Commentez.

21) Peut-on dire au risque $\alpha = 5\%$ que la nouvelle technique de mesure est (en moyenne) valide ou biaisée ?

6. Données disponibles ici : http://josephsalmon.eu/enseignement/datasets/filtre.dat
7. https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk
8. Données disponibles ici : http://josephsalmon.eu/enseignement/datasets/tailles_arbres.csv