

TD de statistique : comparaisons de moyenne et corrélation linéaire

Jean-Baptiste Lamy

6 octobre 2008

1 Comparaison de moyenne

1.1 Cas 1 : comparer deux moyennes observées dans des échantillons différents

Pour comparer 2 moyennes observées sur des échantillons (ou sous-échantillons) différents, on utilise le test de Welch Student :

```
> t.test(variable1_numerique, variable2_numerique)
...
t = ..., df = ..., p-value = ...
...
```

Dans ce cas, on distingue en général deux groupes d'individus : un groupe "témoin" et un groupe "expérimental", et l'on souhaite comparer la valeur moyenne d'une variable entre ces deux groupes.

1.2 Cas 2 : comparer deux moyennes observées sur un même échantillon

Pour comparer de deux moyennes observées sur un même échantillon, les valeurs étant appariées 2 à 2, on utilise le test T de student apparié :

```
> t.test(variable1_numerique, variable2_numerique, paired = TRUE)
```

Dans ce cas, les deux variables doivent avoir le même nombre de valeurs, chaque paire (1^{ère} valeur de la variable 1, 1^{ère} valeur de la variable 2), (2^{ème} valeur de la variable 1, 2^{ème} valeur de la variable 2),... correspondant à un seul individu. C'est notamment le cas des études du type "avant - après".

NB En général, ce type d'étude (sur des valeurs appariées) conduit à une meilleure sensibilité que le cas précédent (sur des valeurs non appariées), cependant il est parfois plus difficile, voire impossible, à mettre en oeuvre!

1.3 Cas 3 : comparer une moyenne observée dans un échantillon à une moyenne théorique

Pour comparer une moyenne observée dans un échantillon à une moyenne théorique, on utilise le test T de student :

```
> t.test(variable_numerique, mu = moyenne_theorique)
```

2 Corrélation et régression linéaire

La fonction `cor()` permet de calculer le coefficient de corrélation linéaire :

```
> cor(variable1_numérique, variable2_numérique)
```

Ce coefficient indique si les deux variables sont liées de manière linéaire. Une valeur de 0 indique une absence de liaison, une valeur de 1 ou de -1 indique une linéarité parfaite.

En cas de relation linéaire, la fonction `lm()` permet d'effectuer une régression linéaire :

```
> lm(variable1_numérique ~ variable2_numérique)
Call :
lm(formula = variable1_numérique ~ variable2_numérique)
Coefficients :
(Intercept)          variable2_numérique
           Y0                pente
# pente est la pente de la droite
# Y0 est son ordonnée à l'origine
```

La variable passée en premier à la fonction `lm()` doit être la variable dont on cherche à expliquer les variations, à partir de la seconde variable.

3 Exercice 1

Nous allons reprendre l'étude sur les yeux des lapins (cf TD 3). Un laboratoire cosmétique souhaite vérifier l'absence de propriétés irritantes sur trois nouveaux produits, afin de pouvoir les inclure dans des shampoings. Pour cela, des gouttes de produit sont placés dans l'oeil de lapins, et on mesure dix minutes après la rougeur de l'oeil du lapin (unité arbitraire). L'expérience est réalisée sur quarante lapins, et chaque lapin reçoit successivement les trois produits ainsi qu'un témoin.

1. Le tableau de données est enregistré dans le fichier `lapin.csv`. Charger ce fichier.

Réponse :

```
> t = read.table("lapin.csv", sep=",", header=TRUE)
> attach(t)
```

2. Tracer un graphique pour représenter ces données.

Réponse :

```
> boxplot(temoin, produit1, produit2, produit3)
```

3. Nous souhaitons comparer la rougeur moyenne sur le témoin avec celle du produit 1. Quelle est l'hypothèse H_0 ? l'hypothèse H_1 ? Dans quel cas de comparaison de moyenne sommes-nous?

Réponse : H_0 : pas de différence de rougeur entre le témoin et le produit 1.

H_1 : il y a une différence de rougeur entre le témoin et le produit 1.

Il s'agit d'une comparaison de deux moyennes issus d'un même échantillon (appariement).

```
> t.test(temoin, produit1,paired=TRUE)
      Welch Two Sample t-test
data : temoin and produit1
t = 0.1924, df = 61.753, p-value = 0.848
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
 -0.1891363  0.2294228
sample estimates :
mean of x mean of y
 5.005509  4.985366
```

4. Même question pour les produits 2 et 3.

Réponse :

```
> t.test(temoin, produit2,paired=TRUE)
t = -26.5028, df = 59.744, p-value < 2.2e-16
> t.test(temoin, produit3,paired=TRUE)
t = -2.4422, df = 65.915, p-value = 0.01729
```

4 Exercice 2

Nous allons reprendre l'étude sur les OGM (cf TD 3). Afin de tester la toxicité d'une variété de maïs OGM, 3 groupes de 10 rats a été nourri avec ce maïs. Le maïs OGM représentait 11% de la ration alimentaire dans le premier groupe, 22% dans le second, et 33% dans le troisième. Un quatrième groupe témoin de 60 rats a été nourri avec du maïs non-OGM. Après 90 jours, on mesure le poids du foie de chaque rat.

1. Le tableau de données est enregistré dans le fichier `ogm.csv`. Charger ce fichier.

Réponse :

```
> t = read.table("ogm.csv", sep=",", header=TRUE)
> attach(t)
```

2. Représenter graphiquement la taille du foie en fonction de la quantité d'OGM présent dans l'alimentation des rats.

Réponse :

```
> plot(foie, ogm)
> boxplot(foie ~ ogm)
```

3. Créer les variables `g0`, `g11`, `g22` et `g33` contenant les lignes du tableau correspondant aux rats nourris avec 0%, 11%, 22% et 33% d'OGM, respectivement.

Réponse :

```
> g0 = t[t["ogm"] == 0,]
> g11 = t[t["ogm"] == 11,]
> g22 = t[t["ogm"] == 22,]
> g33 = t[t["ogm"] == 33,]
```

4. Comparer le poids moyen du foie des rats du groupe sans OGM, avec celui des rats du groupe à 11% d'OGM, puis à 22% et à 33%.

Réponse :

```

> t.test(g0["foie"], g11["foie"])
t = -0.2052, df = 11.572, p-value = 0.841
> t.test(g0["foie"], g22["foie"])
t = -0.5421, df = 10.123, p-value = 0.5995
> t.test(g0["foie"], g33["foie"])
t = -3.1829, df = 13.433, p-value = 0.006946

```

5. Pourquoi l'étude n'a-t-elle pas utilisé un protocole avec appariement des rats, type "avant consommation d'OGM" versus "après consommation d'OGM" (comme cela avait été fait pour les lapins, cf exercice précédent)?

Réponse : Parce que, pour mesurer le poids du foie d'un rat, il faut le sacrifier!

6. Calculer le coefficient de corrélation linéaire entre la taille du foie des rats et la quantité d'OGM absorbée. Y a-t-il un effet dose dans la toxicité de cet OGM? Que peut-on en déduire sur le mécanisme d'action de cette toxicité?

Réponse :

```

> cor(foie, ogm)
[1] 0.2529605

```

Il y a un effet dose, mais assez limité. Les mécanismes de type allergique ne sont pas dose-dépendants.

7. Effectuer une régression linéaire de la quantité d'OGM consommée sur le poids du foie. Qu'en déduit-on?

Réponse :

```

> lm(foie ~ ogm)
(Intercept)      ogm
 16.09224      0.01653

```

Le poids du foie normal est de 16,09 g. Si l'on s'en tient au modèle linéaire, chaque pourcent d'OGM dans la consommation augmente le poids du foie de 0,017 g.

5 Exercice 3

Un nutritionniste souhaite étudier un nouveau régime amaigrissant à forte teneur en protéine. Pour cela, il a testé ce régime sur 50 patients pendant une durée de trois mois. Le poids des patients a été mesuré avant et après le régime.

1. Le tableau de données est enregistré dans le fichier `regime.csv`. Charger ce fichier.

Réponse :

```

> t = read.table("regime.csv", sep=";", header=TRUE)
> attach(t)

```

2. Représenter graphiquement le poids après le régime en fonction du poids avant le régime.

Réponse :

```

> plot(poids_avant, poids_apres)

```

3. Le régime a-t-il provoqué une perte de poids significative, en moyenne?

Réponse :

```

> t.test(poids_avant, poids_apres, paired=TRUE)
t = 24.1307, df = 49, p-value < 2.2e-16

```

4. Étudier la corrélation entre le poids avant régime et le poids après. Effectuer une régression linéaire si possible. Que peut-on en déduire?

Réponse :

```

> cor(poids_avant, poids_apres)
[1] 0.8861524
> lm(poids_apres ~ poids_avant)
Call :
lm(formula = poids_apres ~ poids_avant)
Coefficients :
(Intercept)  poids_avant
  -13.885      1.050

```

La perte de poids est quasiment indépendante du poids initial du patient (la pente est presque égale à 1).

5. Ajouter au tableau de donnée une colonne "kilo_perdus" indiquant pour chaque patient le nombre de kilos perdus lors du régime.

Réponse :

```

> t["kilo_perdus"] = poids_avant - poids_apres
> attach(t)

```

6. Nous souhaitons comparer ce nouveau régime à un autre régime qui repose sur un faible apport calorique, et qui permet de faire perdre en moyenne 8kg. Le nouveau régime est-il significativement plus efficace que l'ancien?

Réponse :

```
> t.test(kilo_perdus,mu = 8)
t = 4.6842, df = 49, p-value = 2.263e-05
+ regarder le sens de la différence!
```

7. Le régime est-il significativement plus efficace chez les hommes ou chez les femmes ?

Réponse :

```
> hommes = t[t["sexe"] == "homme",]
> femmes = t[t["sexe"] == "femme",]
> t.test(hommes["kilo_perdus"], femmes["kilo_perdus"])
t = -4.0523, df = 47.98, p-value = 0.0001848
mean of x mean of y
8.734664 11.573624
```

6 Exercice 4

Une étude a été réalisé sur 100 patients d'un service hospitalier afin de vérifier la relation entre le tabac et les problèmes pulmonaires. Pour cela, nous avons demandé à chaque personne son âge, son sexe, sa situation (célibataire, mariée,...), sa consommation de tabac (nombre de cigarettes par jour), la présence de tabagisme passif, et la présence de problème pulmonaire (cancer du poumon, BPCO,...) chez cette personne.

1. Le tableau de données est enregistré dans le fichier tabac.csv. Charger ce fichier.

Réponse :

```
> t = read.table("tabac.csv", sep=";", header=TRUE)
> attach(t)
```

2. Ajouter une colonne "fumeur" de type booléenne.

Réponse :

```
> t["fumeur"] = tabac > 0
> attach(t)
```

3. Dans ce service hospitalier, l'âge moyen est de 40 ans. Est-ce que notre échantillon est conforme à ce chiffre ?

Réponse :

```
> t.test(age, mu=40)
One Sample t-test
data : age
t = 1.0505, df = 99, p-value = 0.2961
alternative hypothesis : true mean is not equal to 40
95 percent confidence interval :
38.77333 43.98667
sample estimates :
mean of x
41.38
```

$p > 0.05 \Rightarrow$ pas de différence entre la moyenne d'âge théorique et la moyenne observée.

4. Dans cet échantillon, les femmes fument-elles plus ou moins que les hommes ?

Réponse : Les femmes fument moins (3,6 cigarettes par jour contre 4,3). Pas de test statistique ici car on reste dans l'échantillon ! La question ne demande pas d'extrapoler à une population plus grande.

5. D'après cet échantillon, peut-on dire que les femmes fument-elles plus ou moins que les hommes au sein de ce service ?

Réponse :

```
> femmes = t[t["sexe"] == "femme",]
> hommes = t[t["sexe"] == "homme",]
> t.test(femmes["tabac"], hommes["tabac"])
t = -0.7001, df = 97.93, p-value = 0.4855
```

6. Les personnes ayant des problèmes pulmonaires ont-elles une consommation de tabac significativement supérieure ?

Réponse :

```
> malades = t[t["probleme_pulmonaire"] == TRUE ,]
> sains = t[t["probleme_pulmonaire"] == FALSE,]
> t.test(malades["tabac"], sains["tabac"])
t = 18.1068, df = 48.2, p-value < 2.2e-16
```