

Travail Pratique 5

Comparaison de deux variances

Test de t

Condition de normalité

STT1039
BIOLOGIE QUANTITATIVE

Réalisé par
Marc Pépino



Département des sciences de l'environnement

1. Objectifs du TP5

Le travail pratique 5 (TP 5) a pour objectif général d'initier les étudiants aux tests statistiques réalisés à l'aide du logiciel R. À la fin du TP 5, l'étudiant doit savoir : comparer les variances de deux échantillons, utiliser le test de t de Student pour comparer les moyennes de deux échantillons, et vérifier la condition de normalité.

Note : Tout ce qui est mis en rouge et en gras dans les exemples (encadrés) sont des noms donnés par l'utilisateur. Ces noms sont à modifier en conséquence si les exemples sont utilisés avec d'autres bases de données.

Les tests statistiques présentés dans la plupart des exemples seront réalisés à partir de la base de données *TP5_Exemples.csv* disponible sur le portail du cours :

```
> #Importation de la base de données
> tab <- read.csv("D:/bureau/TP5_Exemples.csv",sep=";",dec=".",h=T)
>
> #Visualisation de la structure du tableau de données
> str(tab)
'data.frame': 100 obs. of 4 variables:
 $ x : num 211 208 192 192 212 ...
 $ y : num 729 745 650 580 664 ...
 $ f : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 ...
 $ f2: Factor w/ 2 levels "F","M": 1 2 2 2 1 2 1 2 2 1 ...
```

OPTIONNEL :

```
#Une base de données similaire peut être obtenue à partir du code suivant:
#Construction des variables
x <- c(rnorm(60,200,10),rnorm(40,210,7))
f <- c(rep("A",60),rep("B",40))
f2 <- c(rep("M",50),rep("F",50))[sample(100)]
y <- rep(NA,100)
y[1:60] <- 55 + 3*x[1:60] + rnorm(60,0,50)
y[61:100] <- 60 + 3*x[61:100] + rnorm(40,0,60)

#Construction du tableau de données
tab <- as.data.frame(x)
tab$y <- y
tab$f <- as.factor(f)
tab$f2 <- as.factor(f2)
```

Pour plus de détails sur les tests statistiques avec R, vous pouvez consulter, entre autres, **Crawley (2015)**, **Millot (2018)**, ou **Bertrand et Maumy-Bertrand (2010)**.

2. Comparaisons de deux variances : test de F

La fonction `var.test()` permet de faire directement la comparaison des variances de deux échantillons indépendant (test de F). Si n_1 et n_2 sont respectivement les effectifs des échantillons A et B, alors le rapport des variances obéit à une loi de Fisher-Snedecor à n_1-1 et n_2-1 degré de liberté.

```
> #Exemple 1
> #Comparer les variances de x entre les catégories M et F de la variable f2
> F <- tab[tab$f2=="F",]$x
> M <- tab[tab$f2=="M",]$x
> var.test(F,M)
```

F test to compare two variances

```
data: F and M
F = 0.8155, num df = 49, denom df = 49, p-value = 0.478
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.4627924    1.4371131
sample estimates:
ratio of variances
    0.8155274
>
> #Alternative: test de F à partir du tableau de données
> var.test(tab$x~tab$f2)
```

F test to compare two variances

```
data: tab$x by tab$f2
F = 0.8155, num df = 49, denom df = 49, p-value = 0.478
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.4627924    1.4371131
sample estimates:
ratio of variances
    0.8155274
```

Note : Le symbole « ~ » signifie *en fonction de*. Il est souvent utilisé pour écrire des modèles statistiques.

Interprétation : Le ratio des variances (F) est égal à 0.82, les degrés de liberté sont de 49 au numérateur et de 49 au dénominateur, et la valeur de p est de 0.48. Par défaut, un test bilatéral est effectué (c'est-à-dire que l'on cherche à savoir si le ratio des variances est différent de 1. Ici (exemple 1), la valeur de p et l'intervalle de confiance à 95% nous indiquent que les deux variances sont égales au seuil de signification de 5%.

Important : l'argument « **alternative** » peut être ajouté à l'intérieur de la fonction pour indiquer si le test doit être unilatéral ou bilatéral. Il y a trois options possibles :

Tableau 1. Trois options possibles pour l'argument « alternative » de la fonction *var.test()*

alternative =	Test de F	Définition
"two.sided"	bilatéral	Le rapport des variances est différent de 1
"less"	unilatéral	Le rapport des variances est plus petit que 1
"greater"	unilatéral	Le rapport des variances est plus grand que 1

```
> #Exemple 2
> #Comparer les variances de x entre les catégories A et B de la variable f
> #Hypothèse: la variance de A est supérieure à la variance de B
> var.test(tab$x~tab$f,alternative="greater")
```

F test to compare two variances

```
data: tab$x by tab$f
F = 1.3217, num df = 59, denom df = 39, p-value = 0.1787
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
0.8024435      Inf
sample estimates:
ratio of variances
1.321684
```

Note : Les fonctions *df()*, *pf()* et *qf()* sont respectivement les fonctions de **densité de probabilité**, de **répartition**, et de **quantile** de la loi de Fisher. Il est possible de calculer le ratio des variances et d'utiliser ses fonctions pour effectuer le test de F sans passer par la fonction *var.test()*. Un exemple sera donné dans les exercices. Pour plus de détails sur ces fonctions, consulter l'aide en ligne.

3. Test de t

La fonction ***t.test()*** permet de faire la comparaison de deux moyennes et devrait être utilisée dans le cas de petits échantillons. Les arguments de la fonction permettent de spécifier s'il s'agit d'un test pour des échantillons indépendants de variances égales (test de Student), pour des échantillons indépendants de variances inégales (test de Welch), ou d'un test de Student pour des données appariées.

Tableau 2. Les arguments de la fonction *t.test()*

Arguments	Définition
alternative	Les deux moyennes sont inégales ("two.sided" ; défaut), la première moyenne est inférieure à la deuxième ("less"), la première moyenne est supérieure à la deuxième ("greater")
paired	Test de t pour des données appariées (TRUE) ou indépendantes (FALSE ; défaut)
var.equal	Test de t pour des variances égales (TRUE) ou inégales (FALSE ; défaut)

```
> #Exemple 3
> #Comparer les moyennes de x entre les catégories M et F de la variable f2
> #Condition: l'exemple 1 nous a montré que les variances sont égales
> t.test(tab$x~tab$f2,var.equal=TRUE)
```

Two Sample t-test

```
data: tab$x by tab$f2
t = -0.9498, df = 98, p-value = 0.3445
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.269954    2.210783
sample estimates:
mean in group F mean in group M
202.0148    204.0444
```

La valeur de t est de -0.95, le nombre de degré de liberté est de 98 (nombre de données - 2), et la valeur de p est de 0.34 (test non significatif au seuil de 5%). Les moyennes pour les groupes « F » (202.0) et « M » (204.0) sont également indiquées avec l'intervalle de confiance à 95% de la différence entre ces deux moyennes (IC_{95%} : -6.27 – 2.21).

Note : Les fonctions ***dt()***, ***pt()*** et ***qt()*** sont respectivement les fonctions de **densité de probabilité**, de **répartition**, et de **quantile** de la loi de Student. Pour plus de détails sur ces fonctions, consulter l'aide en ligne ou le cours sur les lois de probabilité.

4. Condition d'application : la normalité

Une des conditions d'applications du test de F ou du test de t est la normalité de chacun des deux échantillons (toutefois, le test de t est relativement robuste à la violation de cette condition). La normalité peut être vérifiée graphiquement à l'aide d'un **histogramme** (voir TP 2), du diagramme quantile-quantile (« **normal QQ-plot** » ; voir TP 4) ou à l'aide d'un test statistique tel que celui de Shapiro-Wilk. La fonction pour effectuer ce test est ***shapiro.test()***.

```
> #Exemple 4  
> #Test de normalité de x avec le test de Shapiro-Wilk  
> shapiro.test(tab$x)
```

Shapiro-Wilk normality test

```
data: tab$x  
W = 0.9866, p-value = 0.4136
```

Dans l'exemple 4, les données de x ne diffèrent pas significativement d'une distribution normale au seuil de signification de 5% ($p > 0.05$).

Important : la condition de normalité doit se vérifier sur chacun des groupes.

```
> #Exemple 5  
> #Test de normalité de x pour le groupe F  
> shapiro.test(tab$x[tab$f2=="F"])
```

Shapiro-Wilk normality test

```
data: tab$x[tab$f2 == "F"]  
W = 0.96885, p-value = 0.2077
```

```
> #Test de normalité de x pour le groupe M  
> shapiro.test(tab$x[tab$f2=="M"])
```

Shapiro-Wilk normality test

```
data: tab$x[tab$f2 == "M"]  
W = 0.98093, p-value = 0.5917
```

5. Exercices

Répondez aux exercices dans un fichier texte « Nom_Prenom_TP5.txt ». Au début du fichier, identifiez-vous par votre nom et prénom, ajouter la date et le numéro du TP, le tout en utilisant des commentaires. Utilisez des commentaires pour structurer vos réponses.

Exercice 1. Soit les variables indépendantes x_1 et x_2 suivantes :

```
x1 <- rnorm(27,mean=220,sd=3)
x2 <- rnorm(23,mean=210,sd=4)
```

Les variances de x_1 et x_2 sont-elles différentes au seuil de signification de 5% (on assume que les données sont normales) ? Retrouver la valeur de F donnée dans la sortie du test en calculant le ratio des variances. **Optionnel :** retrouver la valeur de p du test à l'aide de la fonction de distribution de F , soit la fonction $pf()$. Quelle est la valeur de la borne supérieure de F qui donnerait une valeur de p de 5% avec ces deux échantillons dans le cas d'un test bilatéral. Indice : utiliser la fonction $qf()$. Interprétation (aucune analyse supplémentaire n'est demandée) : comparer cette valeur au ratio des variances et valider le résultat du test de F .

Exercice 2. Soit les variables indépendantes y_1 et y_2 suivantes :

```
y1 <- rnorm(22,mean=170,sd=8)
y2 <- rnorm(28,mean=175,sd=8)
```

Effectuer le test approprié pour comparer les moyennes de y_1 et y_2 (on assume que les données sont normales). Les moyennes de y_1 et de y_2 sont-elles différentes au seuil de signification de 5% ? Est-ce qu'il aurait été possible de faire un test de t pour des données appariées ? Justifier brièvement votre réponse.

Exercice 3. Vous voulez déterminer la perte de poids de 25 poissons suite à cinq jours de jeune. Les variables p_1 et p_2 représentent respectivement le poids des 25 poissons avant et après les cinq jours de jeune.

```
p1 <- rnorm(25,mean=150,sd=8)
p2 <- p1 - rnorm(25,mean=1,sd=0.5)
```

Le jeune a-t-il entraîné une perte de poids significative au seuil de 5% ? Quelle est la perte de poids moyenne de ces 25 poissons ? Comparer le résultat à un test de t pour des variables indépendantes. Visualiser la distribution de p_1 et de p_2 sous forme de diagramme à moustache (« boxplot »). **Optionnel :** Pouvez-vous interpréter la différence éventuelle entre ces deux résultats ?

Bibliographie

- Bertrand, F. et Maumy-Bertrand, M. 2010. Initiation à la statistique avec R. Dunod, Paris, France.
- Crawley, M. J. 2013. The R book (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Crawley, M. J. 2015. Statistics: an introduction using R (2^{ème} éd.). John Wiley and Sons, Chichester, England.
- Legendre, P. 2009. Introduction au langage statistique R.
http://biol09.biol.umontreal.ca/BIO6077/Introduction_R.pdf.
- Millot, G. 2018. Comprendre et réaliser les tests statistiques à l'aide de R : manuel pour les débutants (4^{ème} éd.). De Boeck, Bruxelles, Belgique.
- Murrell, P. 2011. R graphics (2^{ème} éd.). Chapman & Hall/CRC The R series, Boca Raton, Floride.
- Paradis, E. 2005. R pour les débutants.
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf.
- Short, T. 2004. R Reference Card. <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>.

Liens utiles :

- Site officiel de R.
<http://www.r-project.org>
- Ateliers R du CSBQ.
<https://qcbs.ca/wiki/r>
- Quick R.
<http://www.statmethods.net/>
- Apprentissage en ligne du logiciel R.
<https://www.datacamp.com/>
- Beginners statistics introduction.
http://influentialpoints.com/Training/basic_statistics_introduction.htm
- R Graphics Cookbook.
<http://www.cookbook-r.com/Graphs/>
- Liste des bases de données disponibles sous R.
<https://vincentarelbundock.github.io/Rdatasets/datasets.html>