TD R - 2.2
Tests statistiques

Vincent Payet

« *Dans une situation tendue, quand tu parles fermement avec un calibre en pogne, personne ne conteste. Y'a des statistiques là-dessus.* »
— Jean Gabin, Mélodie en sous-sol (1963), dialogues de Michel Audiard

Objectifs : Rappels sur quelques tests statistiques communs, réalisation de ces tests avec R, les calculs et figures associés à ces tests.

Table des matières

1 Principe

Dans un contexte expérimental, il est classique de comparer les résultats de mesures réalisées sur deux sous-groupes (on parle d'échantillons). Il s'agit généralement d'appliquer un traitement différent à chaque échantillon et de suivre sur chaque individu de l'echantillon une variable considérée comme une réponse. Même en cas de non action du traitement, il n'est pas raisonnable d'attendre des valeurs identiques, ni pour les individus ni pour les moyennes d'échantillon. On peut identifier plusieurs sources de variation : la variabilité inhérente au système considéré (notamment pour des expériences sur du vivant), les facteurs non contrôlés, les imprécisions de la mesures... On parle globalement de **fluctuation d'échantillonnage**.

Dans cette situation comment établir qu'une différence observée résulte de l'effet du traitement et non de ces autres sources de variation ? Les tests statistiques permettent d'aborder ces situations.

Les tests statistiques nécessitent de définir un couple d'hypothèses (de référence et alternative, nommées respectivement H_0 et H_1), un niveau de risque acceptable si on décide de rejeter l'hypothèse H_0 (le risque de première espèce α) et une variable de test dont on connait le comportement si H_0 est vraie. Il est nécessaire de conclure statistiquement (conservation ou rejet de H_0) puis concrètement vis-à-vis du problème posé.

2 Comparaisons de moyennes

Contexte

Pour comparer deux échantillons en terme de moyenne, en particulier pour tester l'égalité des moyennes μ_1 et μ_2 on considère les hypothèses $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (ou $\mu_1 > \mu_2$ ou $\mu_1 < \mu_2$ en versions unilatérales).

La variable du test est :

$$t = \frac{\mu_1 - \mu_2}{\sigma_{\Delta m}} \quad \text{avec} \quad \sigma_{\Delta m} \quad \text{l'écart-type de la variable : } \mu_1 - \mu_2$$

Le paramètre $\sigma_{\Delta m}$ dépend des informations sur les échantillons (variances connues ou non, effectifs...).

Ce test est appelé couramment **test-t** ou **test de student**. Il s'agit d'un test paramétrique, c'est-à-dire que l'on fait des suppositions sur la distribution des données pour construire le test. En l'occurrence on suppose que les données de chaque échantillon suivent une loi normale et que les variances des deux groupes sont comparables.

La démarche de la comparaison de moyennes est la suivante :

1. Description des données par des paramètres et des graphiques.
2. Vérification de la normalité.
3. Vérification de l'homogénéité des variances.
4. Test des moyennes.

Un premier exemple, application des calculs

On suit la production laitière de deux lots de vaches : un lot témoin élevé dans des conditions classiques et un lot essai élevée dans des conditions devant modifier leur production (valeurs en litre/jour).

```
> essai <- c(22,37,31,17,31,16,29,31,30,33,24,26,34,30,25)
> temoin <- c(19,28,25,24,31,16,21,19,20,26,22,22,27,21,26)
> mean(essai);sd(essai)
```

```
[1] 27,73333

[1] 6,017435
```

```
> mean(temoin);sd(temoin)

[1] 23,13333

[1] 4,033196
```

On veut tester si les deux moyennes sont comparables afin de savoir si les conditions du lot essai ont permis de modifier la production laitière. Commençons par visualiser les données :

```
> par(mfrow=c(1,2))
> boxplot(temoin, essai,names=c("Témoin","Essai"),
+         ylab="Production laitière")
> stripchart(c(temoin,essai)~gl(2,15), method="jitter",
+            pch=16, group.names=c("Témoin", "Essai"),
+            ylab="", vertical = TRUE)
> points(2,mean(essai),col="red",pch=9)
> points(1,mean(temoin),col="red",pch=9)
```

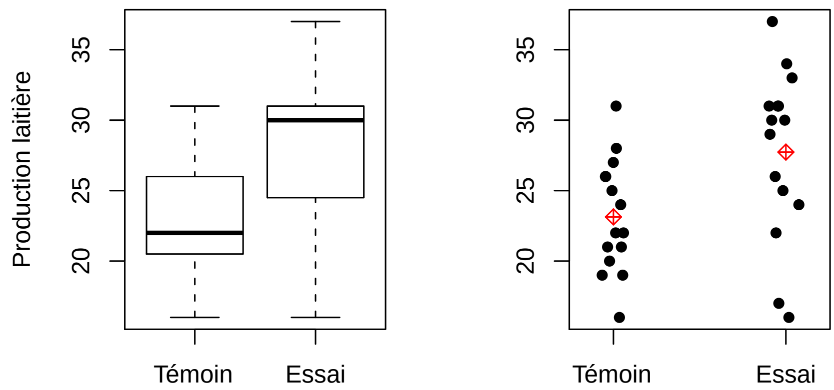

FIGURE 1 – Production de lait pour les lots Témoin et Essai. Représentation avec des boxplots (à gauche) ou avec tous les points et la moyenne en rouge (à droite).

La production de lait du lot essai semble supérieure mais les deux distributions sont recoupantes et nous ne disposons pas de beaucoup de données. Nous ne pouvons répondre à la question sérieusement sans tests statistiques.

La normalité

Pour la normalité nous nous contenterons dans un premier temps d'une analyse graphique. Il faut bien regarder la normalité des deux lots séparément. Dans le cas où H_1 est fortement vraie la distribution de l'ensemble des données sera non normale et bimodale !

```
> par(mfrow=c(1,2))
> hist(essai, main="distribution, lot essai")
> hist(temoin, main="distribution, lot témoin")
```

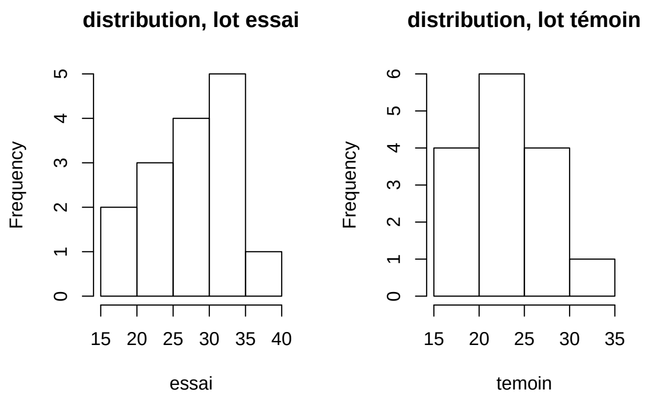

FIGURE 2 – Production de lait pour les lots Témoin et Essai. Représentation avec des histogrammes.

Les histogrammes sont peu adaptés à cette situation avec de petits effectifs. Notons que les distributions sont globalement unimodales et en cloche. Ce test est robuste à un certain écart à la normalité. Nous pouvons continuer.

Les variances

Le test F fait le rapport des variances estimées et compare le résultat à une variable de Fisher à $(n_1-1; n_2-1)$ ddl. La plus grande variance doit être au numérateur. L'hypothèse nulle suppose l'égalité des variances (L'espérance du rapport des variances vaut 1 si H_0 est vraie).

```
> var(essai);var(temoin)

[1] 36,20952

[1] 16,26667

> var(essai)/var(temoin)

[1] 2,225995

> qf(0.95,14,14)

[1] 2,483726
```

`qf()` renvoie un quantile de la loi de Fisher, c'est à dire la valeur seuil du test qu'on peut lire sur les tables statistiques. Le F calculé est inférieur au F théorique, on conserve l'hypothèse d'égalité des variances.

La comparaison de moyenne

On est dans le cas de variables normales (on a accepté cette hypothèse), de variances inconnues et avec de petits effectifs. Il faut estimer la variance commune $\widehat{\sigma^2}$:

$$\widehat{\sigma^2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

On a aussi besoin de la variance calculée d'un échantillon. La fonction `var`() calcule la variance estimée en $SCE/(n-1)$. On crée donc une fonction `varech`() pour calculer la variance d'un échantillon. On réalise ensuite le calcul comme à la calculatrice :

```
> varech <- function(x){sum((x-mean(x))^2)/length(x)}
> n1 <- length(essai)
> n2 <- length(temoin)
> var1 <- varech(essai)
> var2 <- varech(temoin)
> mu1 <- mean(essai)
> mu2 <- mean(temoin)
> varCom <- (n1*var1+n2*var2)/(n1+n2-2)
> tcalc <- (mu1-mu2)/sqrt(varCom/n1+varCom/n2) ; tcalc

[1] 2,459361

> qt(0.975,n1+n2-2) # le test est bilatéral

[1] 2,048407
```

Le t calculé est supérieur au t théorique : on peut rejeter H_0 avec un risque de se tromper d'au plus 5%. Quelle est la probabilité d'observer une différence de moyennes au moins aussi grande ? On cherche $P(\mu_1 - \mu_2 > 2,45936)$, c'est la p-value du test.

```
> 1-pt(tcalc,n1+n2-2)

[1] 0,01017898

> # Le risque est bilatéral :
> (1-pt(tcalc,n1+n2-2))*2

[1] 0,02035795
```

`pt`() retourne la fonction de répartition d'une loi de Student.Avec un bon niveau de confiance, nous pouvons affirmer que les conditions expérimentales téstées améliorent la production laitière.

Le test avec R

La fonction `t.test`() permet de réaliser le test. Retrouvons nous les valeurs calculées "à la main" (notamment *t*, *ddl* et *p-value*) ? `df` signifie *degree of freedom*, c'est-à-dire degré de liberté.

```
> t.test(essai, temoin)
```

```
        Welch Two Sample t-test

data:  essai and temoin
t = 2,4594, df = 24,466, p-value = 0,02135
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0,7435654 8,4564346
sample estimates:
mean of x mean of y
 27,73333  23,13333
```

Dans l'aide de la fonction `?t.test`, on note plusieurs paramètres intéressants. On peut notament présiser si les variances sont consiérées commes égales :

```
> t.test(essai, temoin, var.equal=TRUE)

        Two Sample t-test

data:  essai and temoin
t = 2,4594, df = 28, p-value = 0,02036
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0,7686502 8,4313498
sample estimates:
mean of x mean of y
 27,73333  23,13333
```

C'est mieux, n'est-ce pas ? Le test par défaut, avec `var.equal=FALSE` est le test de Welch qui permet de comparer deux groupes de variances différentes. On peut aussi préciser une valeur `mu` pour comparer une moyenne à une valeur ou préciser s'il s'agit de données appariées (avec `paired`).

3 Comparaisons de variances

Contexte

H_0 : les variances des populations sont égales, $F = \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}} = 1$.

Le test

La fonction `var.test()` permet de réaliser le test. Dans l'aide de la fonction on remarque un paramètre `alternative` permettant de spécifier si le test est uni- ou bilatéral.

Exemple

Avec l'exemple de la production laitière, le test s'écrit ainsi :

```
> var.test(essai, temoin, alternative="greater")
```

```
        F test to compare two variances

data:  essai and temoin
F = 2,226, num df = 14, denom df = 14, p-value = 0,07325
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0,8962323       Inf
sample estimates:
ratio of variances
          2,225995
```

Retrouvez les valeurs calculées précédemment. À quoi correspond la *p-value* ? Retrouvez cette valeur par le calcul sachant que `pf()` retourne la fonction de répartition de la loi de Fisher.

4 Comparaisons de proportions

Le test

La fonction `prop.test()` permet de réaliser un test d'égalité de proportion. Pour les proportions, pensez aussi au test du χ^2 (Chi deux). Une fiche R est disponible sur le cours de première année pour le χ^2.

5 Quelques autres tests

Pour tester la normalité : `shapiro.test()`

Pour des données non normales, test non paramètrique de comparaison de médiane : `wilcox.test()`

On peut connaître toutes les fonctions installées qui contiennent le mot "test" :

```
> apropos("test")
```

6 Exercices

Pour ces exercices, vous devez importer les fichiers de données à récupérer sur e-campus.

Exercice 1 On veut comparer le rendement en blé de deux parcelles travaillées avec ou sans labour (resp. L ou NL). Données : `Tillage.txt`

```
> till <- read.table("Tillage.txt",h=T)
> dim(till)

[1] 148   2

> names(till)

[1] "culture"   "rendement"
```

```
> culture <- till$culture
> rdt <- till$rendement
> # calculons les moyennes
> tapply(rdt,culture, mean)

       L       NL
33,13514 29,59459
```

Les données sont ici fournies de façon “classique” en statistique : un tableau individus-variables. Combien de variables ici ? Combien d'individus ?

La fonction `tapply`() permet d'appliquer une fonction (`mean`(), ici) à une variable (`rdt`) scindée en blocs par une variable qualitative (`culture`).

Plusieurs fonctions (`boxplot`(), les tests aussi) acceptent une écriture de ce type :

`rdt~culture`

Cela signifie `rdt` en fonction de `culture`. Testez `boxplot(rdt~culture)` pour voir.

Réalisez les graphiques et les tests nécessaires puis proposez une conclusion pour ces deux lots.

Exercice 2 On dispose de quatre groupes de patients d'effectifs variables et pour chaque groupe on a compté les fumeurs. On se demande si la proportion de fumeurs est identique dans les 4 populations dont proviennent les patients [1].

```
> # ?prop.test
> fumeurs  <- c( 83, 90, 129, 70 )
> patients <- c( 86, 93, 136, 82 )
> tab <- rbind(fumeurs, patients)
> colnames(tab) <- paste("pop",1:4,sep="")
> tab

         pop1 pop2 pop3 pop4
fumeurs    83   90  129   70
patients   86   93  136   82
```

- Calculez les proportions de fumeurs par population, représentez les données graphiquement et réalisez un test pour savoir si la proportion de fumeurs est la même dans les 4 groupes.
- Ces données peuvent également être traitées avec un test de χ^2 après quelques transformation. En utilisant la fiche χ^2, proposez une interprétation de ces données.

Exercice 3 Si vous avez du temps, mettez en forme dans un fichier de traitement de texte vos deux exercices précédents et conservez des scripts fonctionnels pour chaque tests.

1. Données de Fleiss J.L., 1981, Statistical methods for rates and proportions. New York, Wiley