Tuto et TP: tests de comparaison de moyenne

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2017-10-02

Contents

But de ce TP

Au cours de ce TP nous effectuerons des tests de comparaison de moyennes sur deux types de données.

1. Données artificielles générées selon des distributions normales, soit sous hypothèse nulle (H_0) soit sous hypothèse alternative (H_1). Ceci nous permettra de réaliser des tests dans des situations où nous connaissons les paramètres des populations à comparer (μ_1, μ_2, σ_1, σ_2), en connaissant donc la réponse correcte du test. Le but de cet exercice sera de

   - nous familiariser avec les tests de comparaison de moyenne: choix d'un test en fonction des caractéristiques des données, choix des paramètres du test, interprétation des résultats;
   - évaluer l'adéquation des tests en fonction des types de données;
   - mesurer empiriquement les taux d'erreurs de types I et II, et vérifier s'ils correspondent aux attentes théoriques.

2. Données de transcriptome obtenues au moyen de puces à ADN (DNA microarrays) chez 190 patients souffrant de leucémies lymphoblastiques aigues (LLA), classifiées en différents groupes selon les perturbations chromosomiques (hyperploidie) ou génétiques (mutations d'un gène particulier) supposées être à l'origine du cancer.

Le but de cette analyse sera de détecter les gènes exprimés différentiellement (***differentially expressed genes***, **DEG**) entre deux sous-types particuliers de cancer.

Jeux de données artificiels

Pour les jeux de données artificiels, nous poserons arbitrairement un seuil $\alpha = 0.05$ sur le risque d'erreur de première espèce.

Sous hypothèse nulle

Dans un premier temps, nous allons délibérément générer des données sous hypothèse nulle (H_0) c'est-à-dire en tirant des échantillons dans deux populations de taille égale. Les données seront générées dans les conditions d'applicabilité du test de Student: populations normales de variance égale. Le but de l'exercice sera de mesurer le **taux de faux posiitfs**, c'est-à-dire la proportion des tests déclarés (à tort) positifs, alors qu'il n'existe pas de différence entre les moyennes des deux populations.

1. Génération des données

Au moyen de la fonction `rnorm()`, tirez deux échantillons de taille $n_1 = n_2 = 10$ à partir de populations normales de moyennes $\mu_1 = \mu_2 = 7$ et d'écart-types $\sigma_1 = \sigma_2 = 2$.

2. Application du test avec la fonction `t.test()`

Effectuez un test de comparaison de moyenne en utilisant la fonction `t.test()`. Lisez attentivement l'aide de cette fonction, et choisissez les paramètres en tenant compte des caractéristiques de vos données. Nous effectuerons ici un test bilatéral (*two-tailed*).

Interprétez le résultat de la fonction `t.test()` (paramètres, décision, interprétation de la p-valeur).

```
n1 <- 10
n2 <- 10
mu1 <- 7
mu2 <- 7
sigma1 <- 2
sigma2 <- 2

x1 <- rnorm(n = n1, mean = mu1, sd = sigma1)
x2 <- rnorm(n = n2, mean = mu2, sd = sigma2)

m1 <- mean(x1)
m2 <- mean(x2)

alpha <- 0.05

t.test(x1, x2,
       var.equal = TRUE,
       alternative = "two.sided",
       conf.level = 1 - alpha)
```

```
    Two Sample t-test

data:  x1 and x2
t = -2.7684, df = 18, p-value = 0.01267
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -3.9329955 -0.5391139
sample estimates:
mean of x mean of y
 5.127889  7.363944
```

3. Calcul manuel des statistiques de Student

Au moyen de la fonction R `sum()`, calculez les paramètres de vos échantillons nécessaires au test de Student ($\bar{x}_1, \bar{x}_2, s_1, s_2$).

```
n1 <- length(x1)
m1 <- sum(x1) / n1
s1 <- sqrt(sum((x1 - m1)^2)/n1)
n2 <- length(x2)
m2 <- sum(x2) / n2
s2 <- sqrt(sum((x2 - m2)^2)/n2)
print(paste("n1 =", n1, "; m1 =", round(digits=2, m1), "; s1 = ", round(digits=2, s1)))
```

```
[1] "n1 = 10 ; m1 = 5.13 ; s1 =  1.79"
```

```
print(paste("n2 =", n2, "; m2 =", round(digits=2, m2), "; s2 = ", round(digits=2, s2)))
```

```
[1] "n2 = 10 ; m2 = 7.36 ; s2 =  1.63"
```

Au moyen de cette même fonction `sum()`, calculez les estimateurs de ces paramètres pour les populations ($\mu x_1, \mu x_2, \sigma_1, \sigma_2$).

```
mu1.est <- m1
mu2.est <- m2
sigma1.est <- s1 * sqrt(n1/(n1-1))
sigma2.est <- s2 * sqrt(n2/(n2-1))

print(paste("mu1.est = ", round(digits=2, mu1.est), "sigma1.est", round(digits=2, sigma1.est)))
```

```
[1] "mu1.est =  5.13 sigma1.est 1.89"
```

```
print(paste("mu2.est = ", round(digits=2, mu2.est), "sigma2.est", round(digits=2, sigma2.est)))
```

```
[1] "mu2.est =  7.36 sigma2.est 1.72"
```

Recalculez ces paramètres au moyen des fonctions R `mean()` et `sd()`.

```
mean1 <- mean(x1)
mean2 <- mean(x2)
sd1 <- sd(x1)
sd2 <- sd(x2)

print(paste("mean1 = ", round(digits=2, mean1), "sd1", round(digits=2, sd1)))
```

```
[1] "mean1 =  5.13 sd1 1.89"
```

```
print(paste("mean2 = ", round(digits=2, mean2), "sd2", round(digits=2, sd2)))
```

```
[1] "mean2 =  7.36 sd2 1.72"
```

Attention, les fonctions R `var()` et `sd()` ne calculent pas les paramètres d'échantillons (s^2, s) mais les estimateurs des paramètres de population correspondants ($\hat{\sigma}^2$, $\hat{\sigma}$).

Sur base de ces paramètres, calculez la statistique t_{obs} de Student. Justifiez le choix des paramètres que vous choisirez pour cette fonction.

$$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

```
## Difference between sample means
diff <- m1 - m2

## Estimation of the standard error on the difference between sample means
Student.diff.err <- sqrt((1/n1 + 1/n2) * (n1 * s1^2 + n2 * s2^2) / (n1 + n2 - 2))

## Student statistics
Student.t <- diff / Student.diff.err


## Print Student statistics with 5 significant digits
print(paste("Student statistics  t =", signif(digits=5, Student.t)))
```

```
[1] "Student statistics  t = -2.7684"
```

Avec la fonction $pt()$, calculez la p-valeur de cette valeur t.

Interprétez les résultats (décision, interprétation de la p-valeur).

4. Mesure empirique du taux de faux-positifs

Dans cet exercice, nous allons réaliser un grand nombre de tests de Student en nous plaçant sous hypothèse nulle, et compter le nombre de tests retournant une réponse positive.

a. Avant de commencer l'expérience, indiquez le nombre de faux positifs attendus *a prior* si l'one ffectue $R = 10^4$ tests sous hypothèse nulle, avec un seuil critique de $\alpha = 0.05$.

b. Répétez 10^{4} fois le test de Student au moyen de la fonction `t.test()` et récupérez dans deux vecteurs séparés les valeurs rapportées pour la statistique t et pour la p valeur.

c. Dessinez l'histogramme des valeurs t obtenues empiriquement.

d. Calculez la proportion de faux-positifs. Cette proportion correspond-elle à vos attentes ?

5. Test sous hypothèse alternative

Effectuez $R = 10^4$ tests de comparaison de moyenne sur des échantillons aléatoires tirés dans des populations de moyennes respectives $\mu_1 = 6$ et $\mu_2 8$, ayant toutes deux un écart-type $\sigma = 1$.

a. Choisissez le test et ses paramètres en fonction des caractéristiqus de vos données.

b. Comptez le nombre de résultats déclarés positifs et négatifs avec un seuil $\alpha = 0.05$.

c. Interprétez ce résultat.

6. Test avec variances inégales

Effectuez les mêmes tests (10^{4} répliques, d'abord sous hypothèse nulle puis sous hypothèse alternative) avec des données tirées de population de variances inégales: $\sigma_1^2 = 4$, $\sigma_2^2 = 25$.

a. Justiifez le choix du test et des paramètres.

b. Comptez le nombre de résultats déclarés positifs ou négatifs.

c. Interprétez le résultat.

7. Test avec données non-normales

Au moyen de la fonction R `rpois()`, générez des données selon une loi de Poisson dont l'espérance vaut $\lambda_1 = 4$ pour la première population et $\lambda_2 = 6$ pour la seconde.

a. Choisissez le test approprié.

b. Mesurez les proportions de tests respectivement déclarés positifs et négatifs.

c. Sur ces mêmes données, effectuez un test paramétrique de comparaison de moyennes (Student ou Welch, à vous de choisi en le justifiant).

d. Interprétez le résultat.

Données transcriptomiques obtenues par puces à ADN

Si vous arrivez ici, suivez ce tutoriel:

http://pedagogix-tagc.univ-mrs.fr/ASG/practicals/microarrays_student_test/DenBoer_Student_test.html

---

Formules mathématiques

- Greek symbols (μ, σ) denote population-wide statistics, and roman symbols ($\bar{x}$, s) sample-based statistics.
- The "hat" ($\hat{}$) symbol is used to denote sample-based estimates of population parameters.

Symbol	Description
μ_1, μ_2	Mean of expression values for the gene i in the whole populations 1 and 2, respectively (in our case, the popultations correspond to all the blood samples that could possibly be taken from patients suffering from cancer of type 1 and 2).
σ_1, σ_2	Standard deviation of expression values for the gene i in the whole populations 1 and 2, respectively.
n_1, n_2	"Sample sizes" in the statistical sense, i.e. number of observations for the groups 1 and 2, respectively.
$\bar{x}_1, \bar{x}_2$	Mean of expression values for the gene i in samples of the groups 1 and 2, respectively.
s_1^2, s_2^2	Variance of expression values for the gene i in samples of the groups 1 and 2, respectively.
s_1, s_2	Standard deviations of expression values for the gene i in samples of the groups 1 and 2, respectively.

Symbol	Description
$\hat{\sigma}_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$	Pooled standard deviation, used as estimator for the standard deviation of two groups altogether, when their variances are assumed equal.
$d = \hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$	d = Effect size (difference between sample means), used as estimator of the difference between population means δ.
$\hat{\sigma}_\delta = \hat{\sigma}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	Standard error about the difference between means of two groups whose variances are assumed equal (Student).
$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Student t statistics
$t_W = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Welch t statistics