Bachelor - Tests Statistique

S. Jaubert

05 janvier 2022

Les exemples sont tirés du fascicule : D:\OneDrive - CFAI Centre\Bachelor\Test statistique\App_test-stat-R.pdf

Z-test

On considère un échantillon de taille n, $X_1, X_2, ..., X_n$ avec $X_i \hookrightarrow \mathcal{N}(\mu; \sigma)$ et un risque α

- si l'on teste $H_0 : \mu = m_0$ (Test bilatéral)

La statistique de test sous l'hypothèse nulle est :

$$z = \sqrt{n}\frac{\bar{X_n} - m_0}{\sigma}$$

qui suit une loi normale $\mathcal{N}(0; 1)$

Si $|z|$, la réalisation de la statistique de test, est supérieur au quantile d'ordre $1 - \frac{\alpha}{2}$ alors on rejette l'hypothèse nulle.

- Si l'on teste $H_0 : m \leqslant m_0$

Si z est supérieur au quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$ alors on rejette l'hypothèse nulle.

- Si l'on teste $H_0 : m \geqslant m_0$ Si z est inférieur au quantile d'ordre α de la loi $\mathcal{N}(0, 1)$ alors on rejette l'hypothèse nulle.

Remarque : si l'on note υ_α le quantile d'ordre α de la loi $\mathcal{N}(0, 1)$, alors on a l'égalité $\upsilon_\alpha = -\upsilon_{1-\alpha}$

```
x<-c(6.47,7.02,7.15,7.22,7.44,6.99,7.47,7.61,
     7.32,7.22,7.52,6.92,7.28,6.69,7.24,7.19,
     6.97,7.52,6.22,7.13,7.32,7.67,7.24,6.21)
n<-length(x)
```

Exemple 1 page 13 D'après l'énoncé nous avons : $X \hookrightarrow \mathcal{N}(\mu; \sigma = 0.38)$

et nous savons que $\bar{X} \hookrightarrow \mathcal{N}(\mu; \frac{\sigma}{\sqrt{n}})$

Nous posons :

$H_0 : \mu = 7.3$

$H_1 : \mu \neq 7.3$

Moyenne observée :

```r
(mo<-mean(x))
```

```
## [1] 7.12625
```

Nous connaissant l'écart-type $\sigma = 0.38$, nous allons faire un test Z

Sous H_0 on cherche h tel que : $Pr(7.3 - h < \bar{X} < 7.3 + h) = 1 - \alpha \Leftrightarrow Pr(\bar{X} < 7.3 + h) = 1 - \frac{\alpha}{2}$

Pour $\alpha = 0.05$, on a : $\prod \left(\frac{h}{\frac{\sigma}{\sqrt{n}}} \right) = \prod (t_{0.975})$

Soit :

```r
(h<-qnorm(0.975)*0.38/sqrt(n))
```

```
## [1] 0.1520289
```

Nous devrions avoir dans 95% des cas (si on est bien sous H_0) :

$\overline{X} \in [7.3 - h; 7.3 + h] = [7.14; 7.45]$

Or nous observons que $mo = 7.126 \neq [7.14; 7.45]$

La probabilité d'observer (sous H_0) une valeur aussi lointaine est (d'un seul côté) :

```r
pnorm(mo,7.3,0.38/sqrt(n))
```

```
## [1] 0.01254566
```

Soit une p-value de (on multiplie par 2 car le test est bilatéral) :

```r
2*pnorm(mo,7.3,0.38/sqrt(n))
```

```
## [1] 0.02509132
```

Retrouvons ces résultats directement avec R :

```r
library(TeachingDemos) #librairie pour effectuer un test Z
```

```
## Warning: le package 'TeachingDemos' a été compilé avec la version R 4.1.2
```

```r
z.test(x,mu = 7.3,stdev = 0.38)
```

```
##
##  One Sample z-test
##
## data:  x
## z = -2.24, n = 24.000000, Std. Dev. = 0.380000, Std. Dev. of the sample
## mean = 0.077567, p-value = 0.02509
## alternative hypothesis: true mean is not equal to 7.3
## 95 percent confidence interval:
##  6.974221 7.278279
## sample estimates:
## mean of x
##   7.12625
```

Nous retrouvons bien la p-value, la valeur z est, en nombre d'écart-type, la distance qui sépare μ de la valeur observée

```
(mo-7.3)/(0.38/sqrt(n))
```

```
## [1] -2.239994
```

Ainsi, on peut affirmer que le fournisseur ne respecte pas ses engagements avec un risque de se tromper de 2.6 chances sur 100

t-test

σ population inconnu

on remplace sa variance σ^2 par son estimateur sans biais

$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

La statistique de test sous l'hypothèse nulle est :

$$z = \sqrt{n}\frac{\bar{X}_n - m_0}{S_{n-1}}$$

qui suit alors une loi de Student à n − 1 degrés de liberté sous l'hypothèse nulle (c'est le théorème de Cochran).

```
x2<-c(10.1,9.8,10.2,10.3,10.4,9.8,9.9,10.4,10.2,9.5,10.4,9.6)
```

Exemple 2 page 14 La statistique t_{obs} est :

```
(t_obs<-(mean(x2)-10)*sqrt(length(x2))/sd(x2))
```

```
## [1] 0.540403
```

La probabilité d'observer une telle valeur est :

```
pt(t_obs,df = 11,lower.tail = F)
```

```
## [1] 0.299845
```

Nous pouvons retrouver ce résultat très simplement par :

```
t.test(x2,mu = 10,alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  x2
## t = 0.5404, df = 11, p-value = 0.2998
## alternative hypothesis: true mean is greater than 10
## 95 percent confidence interval:
##  9.883838      Inf
## sample estimates:
## mean of x
##     10.05
```

Nous n'avons donc aucune raison de rejeter l'hypothèse nulle

Exemple 3 page 15 Nous observons comme moyenne :

```
x<-c(232,277,235,245,250,268,256,245)
(mean(x)->mo)
```

```
## [1] 251
```

L'hypothèse nulle $H_0 : mu = 276$

La statistique de test est t_0 :

```
(to<-(mo-276)*sqrt(8)/sd(x))
```

```
## [1] -4.564355
```

Pour obtenir la p-value :

```
pt(to,df = 7)*2
```

```
## [1] 0.00259146
```

Avec la fonction t-test nous retrouverons ces valeurs :

```
t.test(x = x,mu = 276)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -4.5644, df = 7, p-value = 0.002591
## alternative hypothesis: true mean is not equal to 276
## 95 percent confidence interval:
##  238.0484 263.9516
## sample estimates:
## mean of x
##       251
```

Observer une telle valeur mo sous H_0 est donc très peu probable, nous rejetterons cette hypothèse.

Exemple 4 page 16 Soit p la proportion inconnue de haricots fins.

On pose $H_0 : p = 0.25$ contre $H_1 : p \neq 0.25$

On a pour fréquence observée $fo = 118/400 = 0.295$

La p-value est de :

```
pnorm(0.295,0.25,sd = sqrt(0.25*0.75/400),lower.tail = F)*2
```

```
## [1] 0.03766692
```

Si on utilise le test 1-prop-Z-Test de R, on retrouve la même chose :

```
prop.test(118,400,0.25,correct = F,conf.level = 0.95)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  118 out of 400, null probability 0.25
## X-squared = 4.32, df = 1, p-value = 0.03767
## alternative hypothesis: true p is not equal to 0.25
## 95 percent confidence interval:
##  0.252429 0.341471
## sample estimates:
##     p
## 0.295
```

Le X^2 se retrouve par :

```
(0.295-0.25)/sqrt(0.25*0.75/400)->X
X^2
```

```
## [1] 4.32
```

Comme p-value < 0.05, on peut affirmer, au risque 5%, que le producteur a tort.

Tests d'homogénéité

```
X1<-c(106.7,107.02,107.13,107.22,107.41,106.39,107.47,107.61,107.38,107.22)
X2<-c(107.68,106.69,107.24,107.69,106.97,107.52,106.22,107.23,107.32)
```

Exemple 1 page 20 On a : $X_1 \hookrightarrow \mathcal{N}(\mu_1; \sigma_1 = 1.3)$ et $X_2 \hookrightarrow \mathcal{N}(\mu_2; \sigma_2 = 0.9)$

Soit l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ au risque $\alpha = 0.05$

Alors sous H_0, $D = \bar{X}_1 - \bar{X}_2 \hookrightarrow \mathcal{N}(0; \sqrt{\frac{1.3^2}{10} + \frac{0.9^2}{9}})$

```
(h<-qnorm(0.975)*sqrt(1.3^2/10+0.9^2/9))
```

```
## [1] 0.9974657
```

Et $D \in [-0.997; 0.997]$ dans 95% des cas.

```
(D_obs<-mean(X1)-mean(X2))
```

```
## [1] -0.01833333
```

D_obs est bien dans l'intervalle nous ne pouvons rejeter l'hypothèse nulle, de plus la probabilité d'observer une telle différence est :

```
(p_valeur<-pnorm(-0.01833333,mean = 0,sd = sqrt(1.3^2/10+0.9^2/9))*2)
```

```
## [1] 0.9712633
```

Avec la fonction mean_test2 de R, on retrouvera directement la même chose :

```
library(OneTwoSamples)
```

```
## Warning: le package 'OneTwoSamples' a été compilé avec la version R 4.1.1
```

```
mean_test2(X1,X2,sigma = c(1.3,0.9))
```

```
##          mean df           Z   p_value
## 1 -0.01833333 19 -0.03602397 0.9712632
```

Exemple 2 page 21 On considère deux lots de tasses et on souhaite comparer la solidité de ceux-ci. Pour chacun des deux lots, on dispose d'un échantillon de 10 tasses et on mesure la résistance de chacune d'entre eux. Les résultats sont :

- pour le premier échantillon :

```r
X1<-c(31.70,31.98,32.24,32.35,31.18,32.19,32.63,31.19,31.54,31.89)
```

- pour le deuxième échantillon :

```r
X2<-c(31.61,31.10,31.20,31.11,32.66,31.15,31.71,31.22,31.16,31.21)
```

Peut-on affirmer que ces deux échantillons ne proviennent pas de la même production ?

Tests d'indépendance

Voir cours page 31 ou ici : https://sjaubert.github.io/SPCR/Test_du_Khi2.html

```r
A<-matrix(c(50,47,56,5,14,8),nrow = 2,byrow = T)
rownames(A)<-c("Brillants","Médiocres")
colnames(A)<-c("A","B","C")
addmargins(A)
```

Exemple page 33

```
##            A  B  C Sum
## Brillants 50 47 56 153
## Médiocres  5 14  8  27
## Sum       55 61 64 180
```

Détails du calcul de la matrice théorique :

```r
V_th<-c()
for (i in 1:2){
  for (j in 1:3){
    V_th<-c(V_th,sum(A[i,])*sum(A[,j])/sum(A))
  }

}
(A_th<-matrix(V_th,nrow = 2,byrow = T))
```

```
##       [,1]  [,2] [,3]
## [1,] 46.75 51.85 54.4
## [2,]  8.25  9.15  9.6
```

Que l'on retrouve avec :

```r
chisq.test(A)$expected
```

```
##               A     B    C
## Brillants 46.75 51.85 54.4
## Médiocres  8.25  9.15  9.6
```

La statistique du χ^2 se calcule :

```
(chi2_obs <-(50-46.75)^2/46.75+(47-51.85)^2/51.85+(56-54.4)^2/54.4+(5-8.25)^2/8.25+(14-9.15)^2/9.15+(8-9
```

```
## [1] 4.844394
```

Que l'on retrouve facilement avec :

```
chisq.test(A)$statistic
```

```
## X-squared
##  4.844394
```

Et la p-value (probabilité d'observer une valeur aussi extrême) se détermine par :

```
pchisq(q = 4.844394 ,df = 2,lower.tail = F)
```

```
## [1] 0.08872647
```

Enfin nous pouvons avoir tous ces résultats directement par :

```
chisq.test(A)
```

```
##
##  Pearson's Chi-squared test
##
## data:  A
## X-squared = 4.8444, df = 2, p-value = 0.08873
```

Tests d'indépendance cas avec des caractères quantitatifs

Soient X et Y deux caractères quantitatifs.

On souhaite affirmer à partir des données observées que X et Y ne sont pas indépendants.

On considère alors l'hypothèse :

- H_0 : "X et Y sont indépendants"

On fait l'hypothèse que si dépendance il y a, alors elle est linéaire, donc : H_0 : "X et Y sont indépendants <=> $\rho = 0$

On démontre que sous H_0, la statistique :

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

suit une loi de Student à $(n-2)$ degrés de liberté.

La p-value associée au test de nullité du coefficient de corrélation est :

$$\mathbb{P}\left(|T| \geq |t_{obs}|\right)$$

Exemple page 36 Sur 14 familles composées d'un père et d'un fils, on examine le QI du père et le QI du fils. Les résultats sont :

```
Qp<-c(121,142,108,111,97,139,131,90,115,107,124,103,115,151)
Qf<-c(102,138,126,133,95,146,115,100,142,105,130,120,109,123)
```

Peut-on affirmer qu'il y a une liaison significative entre le QI du père et le QI du fils ?

$t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

```
n<-length(Qp)
r<-cor(Qp,Qf)
(t_obs<-r/sqrt((1-r^2)/(n-2)))
```

```
## [1] 2.290343
```

Soit une p-value :

```
(p_value<-2*pt(t_obs,df = n-2,lower.tail = F))
```

```
## [1] 0.04090612
```

Résultats que l'on a directement avec :

```
cor.test(Qp,Qf)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Qp and Qf
## t = 2.2903, df = 12, p-value = 0.04091
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02960175 0.83713281
## sample estimates:
##       cor
## 0.5515191
```

On peut donc affirmer qu'il y a une liaison significative entre le QI du père et celui du fils.

```
chequiers<-read.table(file = "https://sjaubert.github.io/SPCR/chequiers.txt",header = T)
colnames(chequiers)<-c("Interdit","age")
```

```
(table(chequiers)->tab)
```

Exercice 10 page 42

```
##         age
## Interdit ai25 ai35 ai45 ai55 ai75
##        0   84  136  196  165  171
##        1    6   20   16    9    7
```

Essayez de comprendre les résultats suivants :

```
(chisq.test(tab)->res)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 11.423, df = 4, p-value = 0.0222
```

```
res$statistic
```

```
## X-squared
##   11.4228
```

```
res$expected
```

```
##         age
## Interdit      ai25      ai35      ai45      ai55      ai75
##        0 83.555556 144.82963 196.81975 161.54074 165.25432
##        1  6.444444  11.17037  15.18025  12.45926  12.74568
```