Ejercicios Tema 4 - Contraste hipótesis. Taller 2

Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir

Curso completo de estadística inferencial con R y Python

Contenidos

1 Contraste hipótesis taller 2.

1.1 Librerías y datos necesarios

Para este taller necesitaremos los siguientes paquetes: `faraway, nortest, car` si no los tenéis instalados podéis ejecutar lo siguiente:

```
install.packages("faraway")
install.packages("nortest")
install.packages("car")
```

Para utilizarlos, deberéis cargarlos ejecutando las siguientes instrucciones:

```
library("faraway")
library("nortest")
library("car")
```

También necesitáis el fichero "zinc.txt".

1.2 Ejercicio 1

El iris data set es una tabla clásica de datos que recopiló Ronald Fisher publicó en 1936. En este data set hay 150 flores de tres especies las que se mide la longitud y anchura de sus pétalos y sépalos.

La medias globales de toda la población son

```
library(tidyverse)
resumen1=iris %>% summarise(Media_muestral_Sepal.Length=mean(Sepal.Length),
                            Desviacion_muestral_Sepal.Length=sd(Sepal.Length))
resumen1
```

```
##   Media_muestral_Sepal.Length Desviacion_muestral_Sepal.Length
## 1                    5.843333                         0.8280661
```

Consideremos una muestra de tamaño $n = 50$ de la longitud del sépalo del data set iris que generamos con el siguiente código

```r
set.seed(333)# por reproducibilidad, para fijar la muestra
muestra_50=sample(iris$Sepal.Length,size=50,replace = TRUE)
```

1. Contrastar si podemos aceptar que la media de la muestra es igual a la media poblacional es igual a 5.5 contra que es distinta, resolver utilizando el p-valor.
2. Calcular un intervalo de confianza del tipo $(-\infty, x_0)$ para la media poblacional de la muestra al nivel de confianza del 95%

1.2.1 Solución

Para la primera cuestión y bajo estas condiciones, $n = 50$ muestra grande varianza desconocida podemos utilizar un t-test

```r
t.test(muestra_50,mu=5.5,alternative = "two.sided",conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  muestra_50
## t = 3.3027, df = 49, p-value = 0.001793
## alternative hypothesis: true mean is not equal to 5.5
## 95 percent confidence interval:
##  5.654262 6.133738
## sample estimates:
## mean of x
##     5.894
```

El p-valor del contraste es c(t = 3.3026648158547), c(df = 49), 0.00179334930855166, c(5.65426248875515, 6.13373751124485), c(mean of x = 5.894), c(mean = 5.5), 0.11929760419785, two.sided, One Sample t-test, muestra_50 muy pequeño así que no podemos rechazar que la media sea 5.5 (fijémosnos que media real es 5.843333)

Para la segunda cuestión podemos utilizar la función t.test

```r
t.test(muestra_50,alternative="less",conf.level=0.95)$conf.int
```

```
## [1]     -Inf 6.094009
## attr(,"conf.level")
## [1] 0.95
```

1.3 Ejercicio 2

Si consideramos el data set iris la población la proporción poblacional p de flores que tienen la longitud del sépalo mayor que 5 es

```r
Sepalo_mayor_5=prop.table(table(iris$Sepal.Length>5))["TRUE"]
Sepalo_mayor_5
```

```
##      TRUE
## 0.7866667
```

Tomamos una muestra de tamaño $n = 30$ de la población de iris y calculamos en ella la proporción de flores con sépalo mayor que 5.

```r
set.seed(44)
muestra_30=sample(iris$Sepal.Length,size=30,replace = TRUE)
x=table(muestra_30>5)["TRUE"]
x
```

```
## TRUE
##   25
```

```
phat=as.numeric(prop.table(table(muestra_30>5))["TRUE"])
phat
```

```
## [1] 0.8333333
```

1. Queremos contrastar si esta muestra confirma la proporción de flores con sépalo mayor que 5 es 0.75 contra que es mayor de 0.75 con el test exacto y el test aproximado.
2. Extraer de los dos test los intervalos de confiaza asociados al contraste y decir qué fórmula se utiliza para cada intervalo.
3. Extraer de los dos test el valor del estadístico de conytraste y el p valor.

1.3.1 Solución

El contraste que nos piden es

$$\begin{cases} H_0: & p \leq 0.75 \\ H_1: & p > 0.75 \end{cases}$$

los datos son

```
p0=0.75

n=length(muestra_30)
n
```

```
## [1] 30
```

```
x=as.numeric(table(muestra_30>5)["TRUE"])
x
```

```
## [1] 25
```

Para resolver el test exacto hacemos

```
binom.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  25 and 30
## number of successes = 25, number of trials = 30, p-value = 0.2026
## alternative hypothesis: true probability of success is greater than 0.75
## 95 percent confidence interval:
##  0.6810288 1.0000000
## sample estimates:
## probability of success
##              0.8333333
```

```
prop.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  25 out of 30, null probability 0.75
## X-squared = 0.71111, df = 1, p-value = 0.1995
## alternative hypothesis: true p is greater than 0.75
```

```
## 95 percent confidence interval:
##  0.6761382 1.0000000
## sample estimates:
##         p
## 0.8333333
```

En ambos casos el p-valor no es pequeño así que no podemos rechazar la hipótesis nula. Fijémonos que estamos jugando con una dirfencia pequeña sobre el valor real.

Los apartados 2. y 3. se contestan con el siguiente código

```
binom.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)$conf.int
```

```
## [1] 0.6810288 1.0000000
## attr(,"conf.level")
## [1] 0.95
```

```
binom.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)$p.value
```

```
## [1] 0.2025981
```

```
binom.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)$statistic
```

```
## number of successes
##                  25
```

```
prop.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95,correct=FALSE)$conf.int
```

```
## [1] 0.6950795 1.0000000
## attr(,"conf.level")
## [1] 0.95
```

```
prop.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)$p.value
```

```
## [1] 0.1995376
```

```
prop.test(x=25,n=30,p=0.75,alternative = "greater",conf.level = 0.95)$statistic
```

```
## X-squared
## 0.7111111
```

El intervalo de confianza del test exacto`binom.test` es el intervalo de Clopper-Pearson y del test aproximado `prop.test` es el intervalo asintótico para muestras grandes de la proporción con correccion de continuidad (Yates)

1.4 Ejercicio 3

Concentración de zinc

El rastro de metales en al agua potable afecta el sabor y una concentración inusualmente alta puede representar un riesgo para la salud. El fichero *zinc.txt* contiene la concentración de zinc en el fondo y en la superficie de botellas de agua.

Se cree que la concentración media de zinc del agua en el fondo de la botella es mayor que la de la superficie. Suponiendo que los datos siguen una ley normal, ¿hay evidencia suficiente para asegurarlo con un nivel de confianza del 95%?

Plantea un contraste de hipótesis para estudiar si existe dicha evidencia. En particular, especifica la hipótesis nula H_0, la hipótesis alternativa H_1, la expresión del estadístico que se calculará a partir de los datos observados y la distribución de dicho estadístico.

1.4.1 Solución

Denotesmos por μ_F y μ_S las medias poblacionales del al concentración de zinc en el fondo y en la supercficie, respectivamente. No temos que la muestras son emparajadas pues correcponden a la misma botella.

Con esta notación de finimos el contraste:

$$\begin{cases} H_0 : \mu_F \leq \mu_S \\ H_1 : \mu_F > \mu_S, \end{cases}$$

o equivalentemente,

$$\begin{cases} H_0 : \mu_F - \mu_S \leq 0 \\ H_1 : \mu_F - \mu_S > 0 \end{cases}$$

Sabemos que las mediciones siguen una ley normal y la muestra es de tamaño pequeño. Por tanto, se utiliza el test t de muestras dependientes o emparejadas. Sea d la variable diferencia, el estadístico es:

$$t = \frac{\bar{d}}{\frac{\tilde{S}_d}{\sqrt{n}}}$$

Y cuya distribución (de ser cierto H_0) es la distribución T de Student de $n-1$grados de libertad, es decir, de 10-1=9 grados de libertad.

1.5 Ejercicio 4

Continuación ejercicio concentración de zinc

Carga el fichero en la variable `conc.zinc`. Utiliza la función de `R` que calcule el test de hipótesis definido. Interpreta los resultados.

1.5.1 Solución

Con `R` calculamos el test de comparación de medias emparejadas mediante la siguiente instrucción:

```
conc.zinc = read.table("zinc.txt", header=TRUE)
test = t.test(conc.zinc$bottom, conc.zinc$surface,
              alternative="greater",
              paired=TRUE,
              conf.level=0.95)
test
```

```
##
##  Paired t-test
##
## data:  conc.zinc$bottom and conc.zinc$surface
## t = 4.8638, df = 9, p-value = 0.0004456
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0500982       Inf
## sample estimates:
## mean of the differences
##                  0.0804
```

El valor del estadístico es $t = 4.8638127$. Observemos que el p-valor es de 0.0004456, menor que el nivel de significación $\alpha = 0.05$, por lo que se rechaza la hipótesis nula. Hay suficiente evidencia para sugerir que la concentración media de zinc en el agua en el fondo de la botella es mayor que la de la superficie. También se

podría llegar a la misma conclusión observando que el parámetro poblacional, $\mu_F - \mu_S = 0$, está fuera del intervalo de confianza.

1.6 Ejercicio 5

Continuación ejercicio concentración de zinc.

Encuentra la región crítica (es decir el intervalo en el cual se rechaza la hipótesis nula) y la región de aceptación (es decir el intervalo de no rechazo de la hipótesis nula)

1.6.1 Solución

Este es un test unilateral, por lo que el único valor crítico entre las regiones es el cuantil $1-\alpha$ de la distribución t de Student. En este caso es:

```
n=dim(conc.zinc)[1] # tamaño de la muestra: número de filas del data.frame
n
```

```
## [1] 10
```

```
alpha=0.05
nu = n-1 # grados de libertad de la t de Student
nu
```

```
## [1] 9
```

```
cuantil_t=qt(1-alpha,nu)# cuantil de la t-student
cuantil_t
```

```
## [1] 1.833113
```

Dado que la hipótesis alternativa es $H_1 : \mu_F - \mu_S > 0$, la región crítica es $]1.8331129, \infty[$, y la de aceptación $]-\infty, 1.8331129[$