Inferencia Estadística

Profesor(es): Jarnishs Beltran
Ayudante: Pablo Rivera

Pauta ayudantía Nº11
Primavera 2020

Contrastes de hipótesis en R

Prueba t para una muestra, dos muestras independientes y dos muestras dependientes con **t.test**

i) Se obtuvieron durante 132 días las concentraciones máximas de ozono (en partes por 10^9) en una determinada zona de Nueva York. Estados Unidos fija como requirimiento un nivel máximo de 120 de ozono. De los 132 días, 2 días presentaron niveles de ozono por encima de 120.

Contrasta si la proporción de días con nivel de ozono mayor que el permitido es menor o igual que 0,05 y calcula un intervalo de confianza al 95 %.

a) Definir nuestras hipótesis

Solution: Queremos probar (nuestra **hipótesis alternativa**) si la proporción de días con nivel de ozono mayor que el permitido es menor o igual que 0.05. Por lo tanto, tenemos que:

Hipótesis nula es $H_0 : p > 0{,}05$

Hipótesis alternativa es $H_1 : p \leq 0{,}05$

b) Realizar el contraste

Solution: El contraste *binom.test* lleva a cabo un contraste exacto sobre el valor de la probabilidad de éxito en un experimento de Bernoulli.

```r
binom.test( x = 2, # los 2 días con niveles ozono superiores
           n = 132, # el total de días, los 132
           p = 0.05,
           alternative = "less", # en relación a la H.alternativa
           conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  2 and 132
## number of successes = 2, number of trials = 132, p-value = 0.03658
## alternative hypothesis: true probability of success is less than 0.05
## 95 percent confidence interval:
##  0.00000000 0.04692521
## sample estimates:
## probability of success
##             0.01515152
```

c) Interpretar los resultados

Solution:

Con un p-value = 0,03658 menor de 0,05 se rechaza la hipótesis nula H_0. Nos quedamos con la hipótesis alternativa H_1. Por lo tanto, podemos concluir que la proporción de días con nivel de ozono mayor que el permitido es menor o igual que 0,05

ii) A unos pacientes se les ha administrado unos medicamentos para ver si son efectivos en la disminución de ciertas moléculas en sangre. Se han tomado medidas al inicio del estudio, a los 3 meses y a los 6 meses. Los valores obtenidos están recogidos en el fichero medicamentos.csv.

Nuestros datos

```
dfmedicamentos <- read.table(file = "medicamentos.csv",
                             header = TRUE,
                             sep = ";",
                             dec = ".",
                             encoding = "UTF-8",
                             stringsAsFactors = FALSE) # cargamos los datos


head( dfmedicamentos ) # comprobamos que se han leido bien
```

```
##   ID Sex Group    Month0    Month3   Month6
## 1  1   F     P 12.741917 10.302912 8.302369
## 2  2   F     P  8.870604  8.831782 7.822960
## 3  3   F     P 10.726257 10.737613 9.031419
## 4  4   F     P 11.265725 10.589309 9.327378
## 5  5   F     P 10.808537  9.441481 9.693284
## 6  6   F     P  9.787751  7.327527 9.513506
```

Comprobamos nuestros datos antes de empezar con los análisis. Para examinar la estructura de los datos se usa la función $str()$.

```
str ( dfmedicamentos ) # detalle de las variables
```

```
## 'data.frame':    96 obs. of  6 variables:
##  $ ID    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sex   : chr  "F" "F" "F" "F" ...
##  $ Group : chr  "P" "P" "P" "P" ...
##  $ Month0: num  12.74 8.87 10.73 11.27 10.81 ...
##  $ Month3: num  10.3 8.83 10.74 10.59 9.44 ...
##  $ Month6: num  8.3 7.82 9.03 9.33 9.69 ...
```

Codificar adecuadamente las variables categóricas o cualitativas. Las variables Sex y $Group$ podemos considerarlas categóricas. Ahora están en chr, por lo que hay que codificarlas. Las transformamos a $factor$, de forma que R reconozca sus valores como niveles de una variable categórica. Con la función $as.factor()$:

```
dfmedicamentos$Sex  <- as.factor( dfmedicamentos$Sex ) # transformación de los datos a factor

dfmedicamentos$Group <- as.factor( dfmedicamentos$Group ) # transformación de los datos a factor
```

```
str( dfmedicamentos ) # comprobar que ahora son factor
```

```
## 'data.frame':    96 obs. of  6 variables:
##  $ ID    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sex   : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Group : Factor w/ 2 levels "M1","P": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Month0: num  12.74 8.87 10.73 11.27 10.81 ...
##  $ Month3: num  10.3 8.83 10.74 10.59 9.44 ...
##  $ Month6: num  8.3 7.82 9.03 9.33 9.69 ...
```

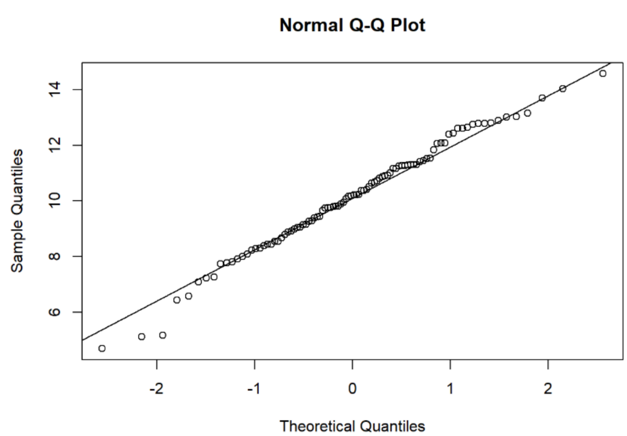
a) Contrastar, con nivel de significación $\alpha = 0{,}05$, si la media de los valores en el mes inicial $Month0$ es 10.

Estamos ante un **contraste de una muestra** (mes inicial). Para una muestra se debe comprobar el supuesto de normalidad. Despúes se realiza el contraste sobre lo que queremos probar, en nuestro caso si la media de los valores en el mes inicial es 10.

SUPUESTO DE NORMALIDAD: Con el **gráfico Q-Q** se hace una primera **aproximación visual** de si hay o no normalidad. Hay que tener en cuenta que este gráfico es meramente descriptivo.

```
Interpretación: La nube de puntos se sitúa sobre la recta. En un principio, visualmente se aprecia que nuestros datos cumple
n el supuesto de normalidad.
```

```
# Gráfico Q-Q
qqnorm( dfmedicamentos$Month0 ) # la nube de puntos
qqline( dfmedicamentos$Month0 ) # la recta
```



Realizar el **contraste para normalidad**. En este contraste *la hipótesis nula es la hipótesis de normalidad*, esto es, no hay diferencias entre nuestra distribución y una distribución normal con esa media y esa desviación típica. Para contrastar la normalidad usamos el test de Shapiro-Wilk, con la función *shapiro.test*().

```
shapiro.test ( dfmedicamentos$Month0 )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dfmedicamentos$Month0
## W = 0.98767, p-value = 0.5142
```

```
Interpretación: Con un p-value = 0.5142 mayor de 0.05 no podemos rechazar la hipótesis nula (hipótesis de normalidad). Por l
o tanto, podemos concluir que nuestros datos cumplen el supuesto de normalidad.
```

CONTRASTE DE HIPÓTESIS: Se supone normalidad en nuestros datos, podemos realizar el contraste.
Definimos nuestras hipótesis. Queremos probar si la media de los valores en el mes inicial es 10. Por lo tanto, tenemos que:
Hipótesis nula es $H_0 : \mu = 10$
Hipótesis alternativa es $H_1 : \mu \neq 10$
Realizamos el contraste. La prueba t para una muestra se utiliza cuando tenemos una variable de medida y un valor esperado para la media, y se supone normalidad de los datos (o muestra grande). Para este contraste sobre una media utilizamos el *t.test*:

```
t.test( dfmedicamentos$Month0,
        mu = 10,
        alternative = "two.sided" ) # contraste bilateral
```

```
##
##  One Sample t-test
##
## data:  dfmedicamentos$Month0
## t = 0.63124, df = 95, p-value = 0.5294
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.724152 10.533047
## sample estimates:
## mean of x
##   10.1286
```

Interpretamos los resultados. Con un p-value = 0,5294 mayor de 0,05 no podemos rechazar la hipótesis nula H_0. Podemos concluir que la media de los valores en el mes inicial Month0 es 10. El intervalo de confianza incluye el 10 (9,724152 − 10,533047).

b) ¿Debemos aceptar o rechazar la diferencia de la media del mes inicial Month0 según el sexo Sex, para $\alpha = 0{,}05$?

Estamos ante un contraste para **dos muestras independientes** (hombres y mujeres). Para dos muestras independientes se debe comprobar el supuesto de normalidad y el supuesto de homocedasticidad. Despúes se realiza el contraste sobre lo que queremos probar, en nuestro caso si la media de los hombres es distinta de la media de las mujeres para el mes inicial.

PREPARAMOS NUESTROS DATOS. Creamos $HombresIni$ solo con los datos del mes inicial ($Month0$) de hombres ($Sex ==$ "M"), y creamos MujeresIni solo con los datos del mes inicial ($Month0$) de mujeres ($Sex ==$ "F"). Serán nuestras dos muestras independientes.

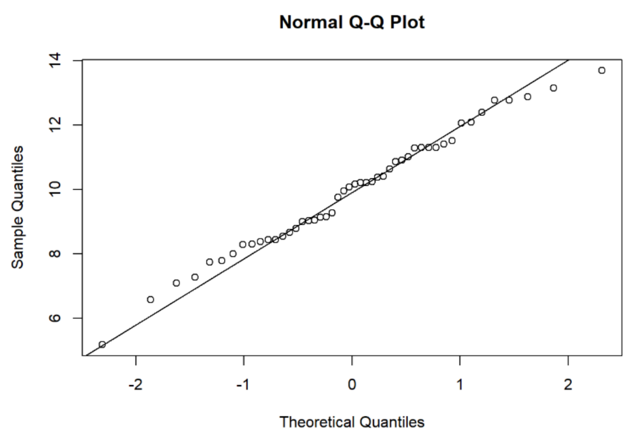
```r
HombresIni <- dfmedicamentos$Month0[dfmedicamentos$Sex == "M"]

MujeresIni <- dfmedicamentos$Month0[dfmedicamentos$Sex == "F"]
```

<u>SUPUESTO DE NORMALIDAD</u>: Con el **gráfico Q-Q** se hace una primera aproximación visual, y con el test de Shapiro-Wilk se realiza el **contraste para normalidad**. La normalidad se comprueba para cada una de las muestras (Hombres y Mujeres).

Supuesto de normalidad para los Hombres

```r
# Gráfico Q-Q
qqnorm( HombresIni ) # la nube de puntos
qqline( HombresIni ) # la recta
```



```
shapiro.test ( HombresIni ) # contraste de normalidad
```

```
##
##  Shapiro-Wilk normality test
##
## data:  HombresIni
## W = 0.98733, p-value = 0.8789
```

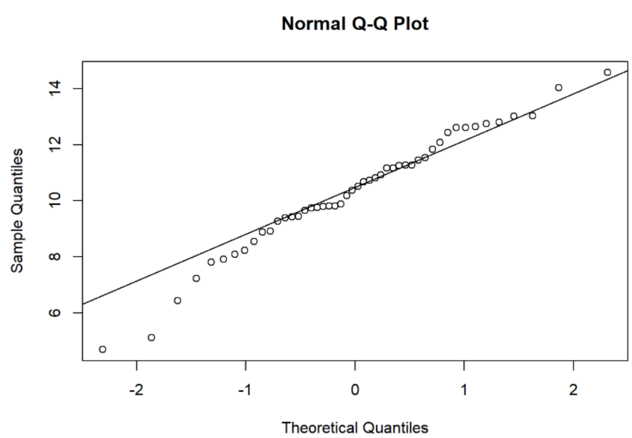
```
Interpretación Hombres:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuest
o de normalidad.

Con un p-value = 0.8789, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que nuestros d
atos cumplen el supuesto de normalidad.
```

Supuesto de normalidad para las Mujeres:

```
# Gráfico Q-Q
qqnorm( MujeresIni ) # la nube de puntos
qqline( MujeresIni ) # la recta
```



```
shapiro.test ( MujeresIni ) # contraste de normalidad
```

```
##
##  Shapiro-Wilk normality test
##
## data:  MujeresIni
## W = 0.97567, p-value = 0.4135
```

```
Interpretación Mujeres:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuest
o de normalidad.

Con un p-value = 0.4135, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que nuestros d
atos cumplen el supuesto de normalidad.
```

SUPUESTO DE HOMOCEDASTICIDAD (homogeneidad de varianzas). En el contraste de homogeneidad de varianzas la **hipótesis nula** es la **varianza es constante (no varía)** en los diferentes grupos. Para contrastarla podemos utilizar el test F de Snedecor con $var.test()$, que se aplica cuando solo hay dos grupos.

```
var.test( HombresIni, MujeresIni ) # contraste de homogeneidad de varianzas
```

```
##
##  F test to compare two variances
##
## data:  HombresIni and MujeresIni
## F = 0.78694, num df = 47, denom df = 47, p-value = 0.4145
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4411467 1.4037792
## sample estimates:
## ratio of variances
##          0.7869387
```

```
Interpretación:
Con un p-value = 0.4145, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto suponemos homogeneidad de varian
zas.
```

CONTRASTE DE HIPÓTESIS: Se supone normalidad y homocedasticidad u homogeneidad de varianzas, podemos realizar nuestro contraste.
Definimos nuestras hipótesis. Queremos probar si la media de los hombres es distinta de la media de las mujeres para el mes inicial. Por lo tanto, tenemos que:
Hipótesis nula es $H_0 : \mu_H = \mu_M$

Hipótesis alternativa es $H_1 : \mu_H \neq \mu_M$

Realizamos el contraste. Para la prueba t para dos muestras independientes usamos la función $t.test()$, sobre muestras independientes paired $= FALSE$, en un contraste bilateral (de dos colas) alternative $=$ "$two.sided$".

```
t.test( HombresIni, MujeresIni, # dos muestras
        alternative = "two.sided", # contraste bilateral
        paired = FALSE, # muestras independientes
        var.equal = TRUE ) # se supone homocedasticidad
```

```
##
##  Two Sample t-test
##
## data:  HombresIni and MujeresIni
## t = -0.952, df = 94, p-value = 0.3435
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1974887  0.4213205
## sample estimates:
## mean of x mean of y
##  9.934557 10.322642
```

```
# Otra forma de escribirlo:
# t.test( Month0 ~ Sex,
#         dfmedicamentos,
#         alternative = "t",
#         paired = FALSE,
#         var.equal = TRUE)
```

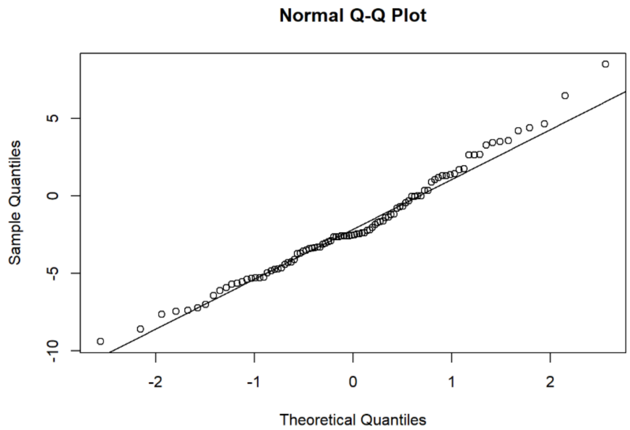
Interpretamos los resultados. Con un p-value = 0,3435 mayor de 0,05 no podemos rechazar la hipótesis nula H_0 de igualdad de medias. Esto es, no hay diferencias significativas entre las medias. Podemos concluir que la media de los hombres y la media de las mujeres no son distintas para el mes inicial.

c) Los investigadores afirman que hay diferencia entre los valores tomados en el mes inicial Month0 y en el tercer mes $Month3$. ¿Tienen razón?.

Solution:
Estamos ante un contraste para **dos muestras dependientes** (mes inicial y tercer mes, sobre los mismos individuos). Para dos muestras dependientes se debe comprobar el supuesto de normalidad. Despúes se realiza el contraste sobre lo que queremos probar, en nuestro caso si la media en el mes inicial $Month0$ es distinta de la media en el tercer mes $Month3$.

SUPUESTO DE NORMALIDAD : Con el **gráfico Q-Q** se hace una primera aproximación visual, y con el test de Shapiro-Wilk se realiza el **contraste para normalidad**. La normalidad se comprueba de forma conjunta para las dos muestras (mes inicial y tercer mes).

```
# Gráfico Q-Q
qqnorm( dfmedicamentos$Month3 - dfmedicamentos$Month0 ) # la nube de puntos
qqline( dfmedicamentos$Month3 - dfmedicamentos$Month0 ) # la recta
```




```
# contraste de normalidad
shapiro.test ( dfmedicamentos$Month3 - dfmedicamentos$Month0 )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dfmedicamentos$Month3 - dfmedicamentos$Month0
## W = 0.98201, p-value = 0.2116
```

```
Interpretación:

Los puntos se agrupan en torno a la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuesto de
normalidad.

Con un p-value = 0.2116, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que nuestros d
atos cumplen el supuesto de normalidad.
```

CONTRASTE DE HIPÓTESIS: Se supone normalidad en nuestros datos, podemos realizar el contraste.

Definimos nuestras hipótesis. Queremos probar si la media de los valores en el mes inicial $Month0$ es distinta de la media en el tercer mes $Month3$.

Hipótesis nula es $H_0 : \mu_{m0} = \mu_{m3}$

Hipótesis alternativa es $H_1 : \mu_{m0} \neq \mu_{m3}$

Realizamos el contraste. La prueba t para muestras dependientes se utiliza cuando tenemos dos variables dependientes (p.e. sobre los mismos individuos). Es equivalente al de una muestra si tomamos la variable diferencia. Se supone normalidad de las diferencias (o muestra grande). Usamos la función $t.test()$, sobre muestras dependientes paired $= TRUE$, en un contraste bilateral (de dos colas) alternative $=$ "$two.sided$".

```
t.test( dfmedicamentos$Month3,
        dfmedicamentos$Month0,
        alternative = "two.sided",
        paired = TRUE ) # contraste muestras dependientes
```

```
##
##  Paired t-test
##
## data:  dfmedicamentos$Month3 and dfmedicamentos$Month0
## t = -5.7578, df = 95, p-value = 1.043e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.698015 -1.314517
## sample estimates:
## mean of the differences
##               -2.006266
```

Interpretamos los resultados. Con un p-value $= 1{,}043e-07$ menor de 0,05 podemos rechazar la hipótesis nula H_0 de igualdad de medias. Podemos concluir que existen diferencias entre la media de los valores en el mes inicial $Month0$ y la de los valores en el tercer mes $Month3$. Por lo tanto, los investigadores tienen razón.