Inferencia Estadística

Profesor(es): Jarnishs Beltran
Ayudante: Pablo Rivera

Pauta ayudantía Nº8
Primavera 2020

Ejercicio en R

0.1. Prueba de hipótesis para el cociente de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$

i) Se realiza un estudio para comparar dos tratamientos que se aplicarán a frijoles crudos con el objetivo de reducir el tiempo de cocción. El tratamiento T1 es a base de bicarbonato de sodio, el T2 es a base de cloruro de sodio o sal común. La variable respuesta es el tiempo de cocción en minutos. Los datos se muestran abajo. ¿Son las varianzas de los tiempos iguales o diferentes? Usar $\alpha = 0{,}05$.

T1: 76, 85, 74, 78, 82, 75, 82.

T2: 57, 67, 55, 64, 61, 63, 63.

Solution: En este problema interesa probar si las varianzas poblacionales son iguales o no, por esta razón el cociente de $\frac{\sigma_{T_1}^2}{\sigma_{T_2}^2}$ se iguala al valor de 1 que será el valor de referencia de la prueba.

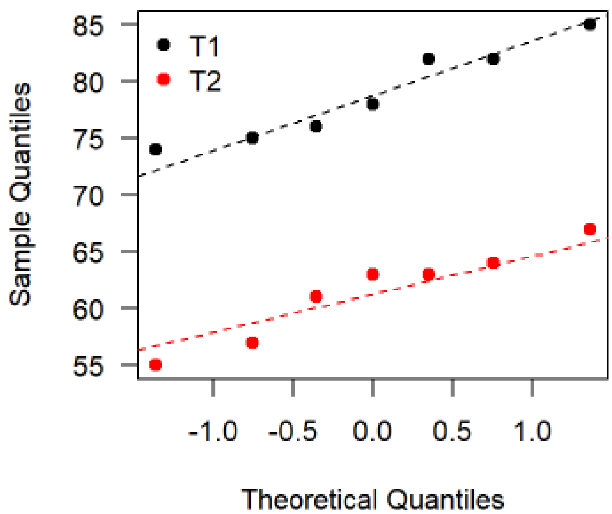
$$H_0 : \frac{\sigma_{T_1}^2}{\theta_{T_2}^2} = 1$$
$$H_1 : \frac{\sigma_{T_1}^2}{\theta_{T_2}^2} \neq 1$$

Para ingresar los datos se hace lo siguiente:

```
T1 <- c(76, 85, 74,78, 82, 75, 82)
T2 <- c(57, 67, 55, 64, 61, 63, 63)
```

Primero se debe explorar si las muestras provienen de una población normal y para esto se construyen los QQplot que se muestran en la Figura

```r
q1 <- qqnorm(T1, plot.it=FALSE)
q2 <- qqnorm(T2, plot.it=FALSE)
plot(range(q1$x, q2$x), range(q1$y, q2$y), type="n", las=1,
     xlab='Theoretical Quantiles', ylab='Sample Quantiles')
points(q1, pch=19)
points(q2, col="red", pch=19)
qqline(T1, lty='dashed')
qqline(T2, col="red", lty="dashed")
legend('topleft', legend=c('T1', 'T2'), bty='n',
       col=c('black', 'red'), pch=19)
```



De la Figura se observa que los puntos están bastante alineados lo cual nos lleva a pensar que las muestras si provienen de una población normal, para estar más seguros se aplicará una prueba formal para estudiar la normalidad.

A continuación el código para aplicar la prueba de normalidad Kolmogorov-Smirnov a cada una de las muestras.

```
require(nortest)  # Se debe tener instalado
lillie.test(T1)$p.value
```

```
## [1] 0.520505
```

```
lillie.test(T2)$p.value
```

```
## [1] 0.3952748
```

Del QQplot mostrado en la Figura y las pruebas de normalidad se observa que se puede asumir que las poblaciones son normales.

La función *var.test* se puede usar para probar H_0 a continuación el código para realizar la prueba.

```
var.test(T1, T2, null.value=1, alternative="two.sided",
         conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  T1  and  T2
## F = 1.011, num df = 6, denom df = 6, p-value = 0.9897
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1737219 5.8838861
## sample estimates:
## ratio of variances
##           1.011019
```

Como el valor-P es 0,9897 (reportado como 1 en la salida anterior), muy superior al nivel α de significancia 5 %, se puede concluir que las varianzas son similares.

ii) El arsénico en agua potable es un posible riesgo para la salud. Un artículo reciente reportó concentraciones de arsénico en agua potable en partes por billón (ppb) para diez comunidades urbanas y diez comunidades rurales. Los datos son los siguientes:

Urbana: 3, 7, 25, 10, 15, 6, 12, 25, 15, 7

Rural: 48, 44, 40, 38, 33, 21, 20, 12, 1, 18

Solution:

¿Son las varianzas de las concentraciones iguales o diferentes? Usar α.

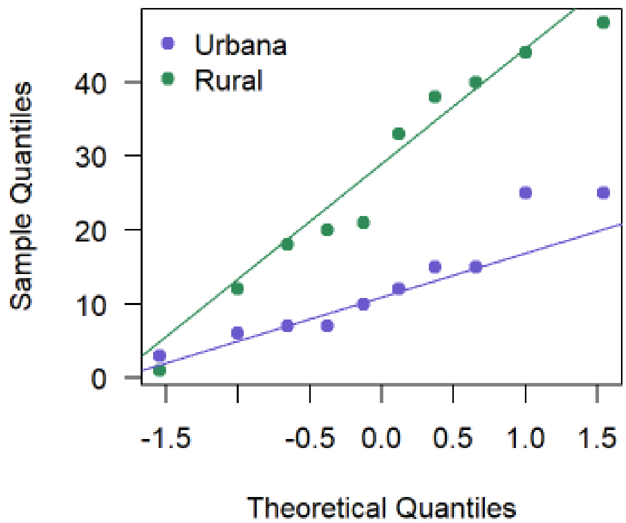
En este problema interesa probar:

$$H_0 : \frac{\sigma^2_{Urb}}{\theta^2_{Rur}} = 1$$
$$H_1 : \frac{\sigma^2_{Urb}}{\theta^2_{Rur}} \neq 1$$

Para ingresar los datos se hace lo siguiente:

```
urb <- c(3, 7, 25, 10, 15, 6, 12, 25, 15, 7)
rur <- c(48, 44, 40, 38, 33, 21, 20, 12, 1, 18)
```

Primero se debe explorar si las muestras provienen de una población normal, para esto se construyen los QQplot mostrados en la Figura anterior.



A continuación el código para aplicar la prueba de normalidad Kolmogorov-Smirnov, a continuación el código usado.

```
require(nortest)  # Se debe tener instalado
lillie.test(urb)$p.value
```

```
## [1] 0.5522105
```

```
lillie.test(rur)$p.value
```

```
## [1] 0.6249628
```

Del QQplot mostrado en la Figura y las pruebas de normalidad se observa que se pueden asumir poblaciones normales.

La función *var.test* se puede usar para probar H_0, a continuación el código para realizar la prueba.

```
var.test(urb, rur, null.value=1, alternative="two.sided",
         conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  urb  and  rur
## F = 0.24735, num df = 9, denom df = 9, p-value = 0.04936
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.06143758 0.99581888
## sample estimates:
## ratio of variances
##          0.2473473
```

Como el valor-P es 0,0493604 (reportado como 0,05 en la salida anterior) y es menor que el nivel de significancia $\alpha = 0{,}05$, se puede concluir que las varianzas no son iguales.

¿Notó que las funciones `var.test` y `var.test` son diferentes?

`var.test` sirve para prueba de hipótesis sobre σ^2.

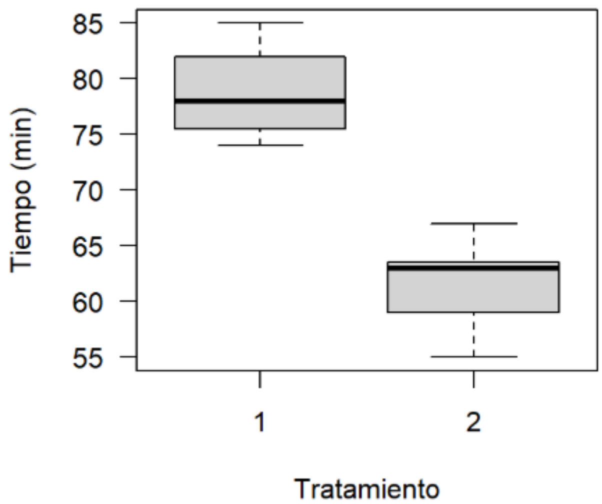
`var.test` sirve para prueba de hipótesis sobre σ_1^2/σ_2^2.

0.2. Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas iguales

iii) Retomando el ejemplo de los fríjoles (item iv), ¿existen diferencias entre los tiempos de cocción de los fríjoles con T1 y T2? Usar un nivel de significancia del 5 %.

Solution: Primero se construirá un boxplot comparativo para los tiempos de cocción diferenciando por el tratamiento que recibieron. Abajo el código para obtener en este caso el boxplot. En la Figura se muestra el boxplot, de esta figura se observa que las cajas de los boxplot no se traslapan, esto es un indicio de que las medias poblacionales, μ_1 y μ_2 , son diferentes, se observa también que el boxplot para el tratamiento T1 está por encima del T2.

```
datos <- data.frame(tiempo=c(T1, T2), trat=rep(1:2, each=7))
boxplot(tiempo ~ trat, data=datos, las=1,
        xlab='Tratamiento', ylab='Tiempo (min)')
```



En este problema interesa estudiar el siguiente conjunto de hipótesis.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

El código para realizar la prueba es el siguiente:

```
t.test(x=T1, y=T2, alternative="two.sided", mu=0,
       paired=FALSE, var.equal=TRUE, conf.level=0.97)
```

```
##
##  Two Sample t-test
##
## data:  T1 and T2
## t = 7.8209, df = 12, p-value = 4.737e-06
## alternative hypothesis: true difference in means is not equal to 0
## 97 percent confidence interval:
##  11.94503 22.91212
## sample estimates:
## mean of x mean of y
##  78.85714  61.42857
```

De la prueba se obtiene un valor-P muy pequeño, por lo tanto, podemos concluir que si hay diferencias significativas entre los tiempos promedios de cocción con T1 y T2, resultado que ya se sospechaba al observar la Figura boxplot.

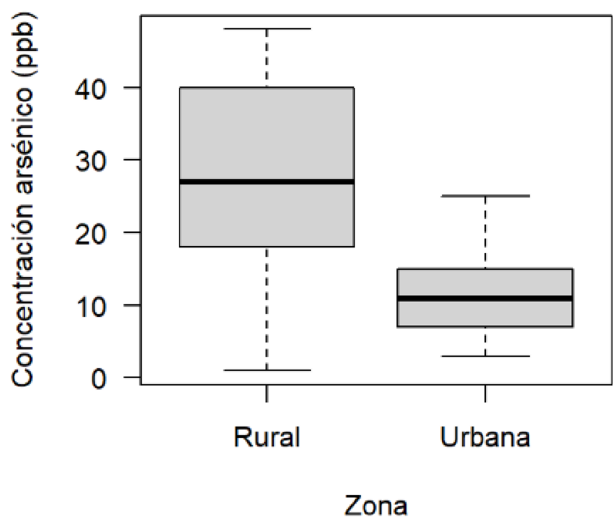
Si el objetivo fuese elegir el tratamiento que minimice los tiempos de cocción se recomendaría el tratamiento T2, remojo de fríjoles en agua con sal.

0.3. Prueba de hipótesis para la diferencia de medias $\mu_1 - \mu_2$ con varianzas diferentes

iv) Retomando el ejemplo de la concentración de arsénico en el agua (item v), ¿existen diferencias entre las concentraciones de arsénico de la zona urbana y rural? Usar un nivel de significancia del 5 %.

Solution: Primero se construirá un boxplot comparativo para las concentraciones de arsénico diferenciando por la zona donde se tomaron las muestras. Abajo el código para obtener en este caso el boxplot. En la figura se muestra el boxplot, de esta figura se observa que las cajas de los boxplot no se traslapan, esto es un indicio de que las medias poblacionales, , son diferentes, se observa también que el boxplot para la zona rural está por encima del de la zona urbana.

```
datos <- data.frame(Concentracion=c(urb, rur),
                    Zona=rep(c('Urbana', 'Rural'), each=10))
boxplot(Concentracion ~ Zona, data=datos, las=1,
        xlab='Zona', ylab='Concentración arsénico (ppb)')
```



En este problema interesa estudiar el siguiente conjunto de hipótesis.

$$H_o : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

El código para realizar la prueba es el siguiente:

```
t.test(x=urb, y=rur, alternative="two.sided", mu=0,
       paired=FALSE, var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  urb and rur
## t = -2.7669, df = 13.196, p-value = 0.01583
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.694067  -3.305933
## sample estimates:
## mean of x mean of y
##      12.5      27.5
```

De la prueba se obtiene un valor-P pequeño, por lo tanto, podemos concluir que si hay diferencias significativas entre las concentraciones de arsénico del agua entre las dos zonas, resultado que ya se sospechaba al observar la Figura. La zona que presenta mayor concentración media de arsénico en el agua es la rural.

Para todas las pruebas se incluyó un intervalo de confianza, revise si la conclusión obtenida con el IC coincide con la obtenida con PH.

0.4. Prueba de hipótesis para la diferencia de proporciones $p_1 - p_2$

v) Se quiere determinar si un cambio en el método de fabricación de una piezas ha sido efectivo o no. Para esta comparación se tomaron 2 muestras, una antes y otra después del cambio en el proceso y los resultados obtenidos son los siguientes.

Num piezas	Antes	Después
Defectuosas	75	80
Analizadas	1500	2000

Realizar una prueba de hipótesis con un nivel de significancia del 10 %.

Solution: En este problema interesa estudiar el siguiente conjunto de hipótesis.

$$H_0 : p_{antes} - p_{despues} = 0$$
$$H_1 : p_{antes} - p_{despues} > 0$$

Para realizar la prueba se usa la función *prop.test* como se muestra a continuación.

```r
prop.test(x=c(75, 80), n=c(1500, 2000),
         alternative='greater', conf.level=0.90)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(75, 80) out of c(1500, 2000)
## X-squared = 1.7958, df = 1, p-value = 0.09011
## alternative hypothesis: greater
## 90 percent confidence interval:
##  0.0002765293 1.0000000000
## sample estimates:
## prop 1 prop 2
##   0.05   0.04
```

Del reporte anterior se observa que el Valor-P es 9 %, por lo tanto no hay evidencias suficientes para pensar que el porcentaje de defectuosos después del cambio ha disminuído.

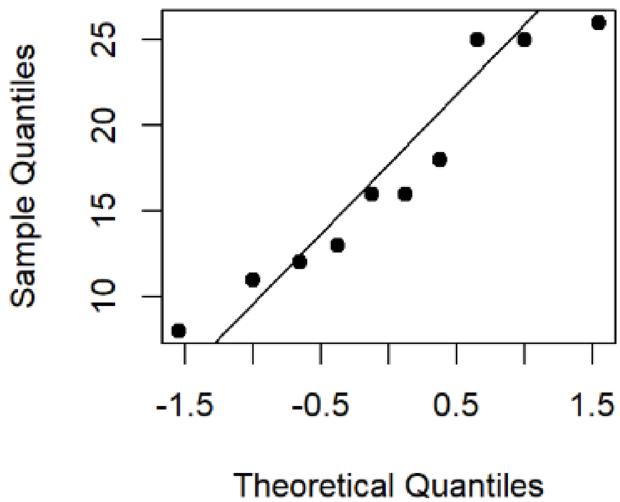
0.5. Prueba de hipótesis para la diferencia de medias pareadas

vi) Diez individuos participaron de programa para perder peso corporal por medio de una dieta. Los voluntarios fueron pesados antes y después de haber participado del programa y los datos en libras aparecen abajo. ¿Hay evidencia que soporte la afirmación de la dieta disminuye el peso medio de los participantes? Usar nivel de significancia del 5 %.

Sujeto	001 002 003 004 005 006 007 008 009 010
Antes	195 213 247 201 187 210 215 246 294 310
Después	187 195 221 190 175 197 199 221 278 285

Solution: Primero se debe explorar si las diferencias de peso (antes-después) provienen de una población normal, para esto se construye el QQplot mostrado en la Figura siguiente. De la figura no se observa un alejamiento serio de la recta de referencia, por lo tanto se puede asumir que las diferencias se distribuyen en forma aproximadamente normal.

```
antes <- c(195, 213, 247, 201, 187, 210, 215, 246, 294, 310)
despu <- c(187, 195, 221, 190, 175, 197, 199, 221, 278, 285)
dif <- antes - despu
qqnorm(dif, pch=19, main='')
qqline(dif)
```



Se puede aplicar la prueba de normalidad Kolmogorov-Smirnov para estudiar si las diferencias dif provienen de una población normal, esto se puede realizar por medio del siguiente código.

```
require(nortest)
lillie.test(dif)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dif
## D = 0.19393, p-value = 0.3552
```

De la salida anterior se observa que el valor-P de la prueba es grande por lo tanto se puede asumir que las diferencias se distribuyen en forma aproximadamente normal.

En este problema interesa estudiar el siguiente conjunto de hipótesis.

$$H_0 : \mu_{antes} - \mu_{despues} = 0$$
$$H_1 : \mu_{antes} - \mu_{despues} > 0$$

El código para realizar la prueba es el siguiente:

```r
t.test(x=antes, y=despu, alternative="greater", mu=0,
       paired=TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  antes and despu
## t = 8.3843, df = 9, p-value = 7.593e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.2832     Inf
## sample estimates:
## mean of the differences
##                      17
```

De la prueba se obtiene un valor-P pequeño, por lo tanto, podemos concluir que el peso μ_{antes} es mayor que $\mu_{despues}$, en otras palabras, la dieta si ayudó a disminuir el peso corporal.