Entorno de programación RStudio: inferencia (1/2)

Román Salmerón Gómez

Abril de 2017

*R* es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos que está disponible como software libre en <http://www.r-project.org/>. Por otro lado, *RStudio* es un entorno de desarrollo integrado (IDE) para *R* que Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

El objetivo del presente documento es el de crear un breve informe con **RStudio** mediante el cual se muestre al alumno como realizar inferencia en un conjunto de datos.

## Funciones a usar

Para comparar medias y varianzas se disponen de los comandos *t.test* y *var.test*, respectivamente:

**t.test(var1, var2=NULL, alternative="two.sided", mu = 0, paired=FALSE, var.equal=FALSE, conf.level=0.95)**

**var.test(var1, var2, ratio=1, alternative = "two.sided", conf.level=0.95)**

Las opciones indicadas son las ofrecidas por defecto:

* Si para *t.test* no se especifica el segundo conjunto de datos se trabajará con una única muestra.
* El argumento *alternative* indica el tipo de contraste: bilateral si se especifica *two.sided*, unilateral con hipótesis alternativa del tipo mayor para *greater* y menor para *less*.
* En el argumento *mu* se especifica el valor de la hipótesis nula en el caso de comparación de medias, mientras que para la comparación de varianzas se hace con el argumento *ratio*.
* En el argumento *paired* se especifica si los datos están relacionados (pareados), *True*, o no, *False*.
* En el argumento *var.equal* se especifica, en caso de dos muestras, si las varianzas son iguales (*True*) o no (*False*).
* Con el argumento *conf.level* se indica el nivel de confianza a usar.

## Conjunto de datos *cars*

Veamos algunos ejemplos que ilustren las órdenes anteriores usando los datos de la base *cars* (para más información *help(cars)*):

# cars # quitar primera almohadilla para ver el conjunto de datos  
 attach(cars)

Con la segunda orden puedo usar directamente el nombre de las variables: *speed* (velocidad en millas por hora) y *dist* (distancia de frenado en pies).

### p-valor

Se define el p-valor como el mínimo valor de significación al partir del cual se rechaza la hipótesi nula. Por tanto:

* Si *p-valor < 0.1*, se rechaza la hipótesis nula al 10% de significación.
* Si *p-valor < 0.05*, se rechaza la hipótesis nula al 5% de significación.
* Si *p-valor < 0.01*, se rechaza la hipótesis nula al 1% de significación.

Como se verá a lo largo de esta práctica, el p-valor resultará muy útil para interpretar los resultados obtenidos.

### Inferencia para una población

Por ejemplo, para contrastar si la velocidad media es igual, mayor o menor a 17 millas por hora haríamos:

t.test(speed, mu=17) # por defecto es a dos colas

##   
## One Sample t-test  
##   
## data: speed  
## t = -2.1397, df = 49, p-value = 0.03739  
## alternative hypothesis: true mean is not equal to 17  
## 95 percent confidence interval:  
## 13.89727 16.90273  
## sample estimates:  
## mean of x   
## 15.4

t.test(speed, mu=17, alternative="greater") # como alternativa la suposición que nos interesa

##   
## One Sample t-test  
##   
## data: speed  
## t = -2.1397, df = 49, p-value = 0.9813  
## alternative hypothesis: true mean is greater than 17  
## 95 percent confidence interval:  
## 14.1463 Inf  
## sample estimates:  
## mean of x   
## 15.4

t.test(speed, mu=17, alternative="less")

##   
## One Sample t-test  
##   
## data: speed  
## t = -2.1397, df = 49, p-value = 0.01869  
## alternative hypothesis: true mean is less than 17  
## 95 percent confidence interval:  
## -Inf 16.6537  
## sample estimates:  
## mean of x   
## 15.4

Atendiendo al p-valor, rechazo que la media sea igual a 17 frente a la alternativa de que sea distinta (p-valor=0.03739<0.05), no rechazo que la media sea 17 frente a la alternativa de que sea mayor (p-valor=0.9813>0.05) y rechazo la hipótesis de que sea igual a 17 frente a la alternativa de que sea menor (p-valor=0.01869<0.05).

Adviértase que en todos los casos se nos proporciona el correspondiente intervalo de confianza, obteniéndose los mismos resultados, y el estimador puntual de la media (media muestral del conjunto de datos).

### Inferencia para dos poblaciones

Empecemos viendo si *speed* y *dist* tienen la misma varianza. Observando las cuasivarianzas muestrales se debería verificar que la primera es menor que la segunda, por tal motivo nos interesamos por esta hipótesis:

var(speed)

## [1] 27.95918

var(dist)

## [1] 664.0608

var.test(speed, dist)

##   
## F test to compare two variances  
##   
## data: speed and dist  
## F = 0.042103, num df = 49, denom df = 49, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.02389265 0.07419404  
## sample estimates:  
## ratio of variances   
## 0.04210335

var.test(speed, dist, alternative="l") # con la inicial es suficiente

##   
## F test to compare two variances  
##   
## data: speed and dist  
## F = 0.042103, num df = 49, denom df = 49, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is less than 1  
## 95 percent confidence interval:  
## 0.00000000 0.06767227  
## sample estimates:  
## ratio of variances   
## 0.04210335

Atendiendo a los resultados, en primer lugar se rechaza la hipótesis nula de que sean iguales frente a la alternativa de que sean distintas (p-valor=2.2·10^(-16)<0.05) y en segundo lugar también se rechaza frente a la alternativa de que sea menor (p-valor=2.2·10^(-16)<0.05). Adviértase que con el primer contraste sólo se podría afirmar (al nivel de significación usado) que las varianzas son distintas, sin embargo, con el segundo se podría decir que la primera varianza es menor que la segunda.

Al igual que antes también se obtienen los correspondientes intervalos de confianza, a partir de los cuales se obtendrían las mismas conclusiones (adviértase que el segundo es unilateral).

Para estudiar si *speed* y *dist* tienen la misma media habría que especificar que las varianzas son distintas:

t.test(speed, dist, var.equal=F)

##   
## Welch Two Sample t-test  
##   
## data: speed and dist  
## t = -7.4134, df = 53.119, p-value = 9.628e-10  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -35.04152 -20.11848  
## sample estimates:  
## mean of x mean of y   
## 15.40 42.98

Observando las medias muestrales (se obtienen como salida del comando anterior), como era esperable, rechazo la hipótesis de igualdad (p-valor=0.000000009628). Además, atendiendo al intervalo de confianza obtenido (siempre es negativo: (-35.04152, -20.11848)) se podría decir que la media de *speed* es menor que la de *dist*. Esta conclusión se obtiene también especificando el siguiente contraste:

t.test(speed, dist, var.equal=F, alternative="l")

##   
## Welch Two Sample t-test  
##   
## data: speed and dist  
## t = -7.4134, df = 53.119, p-value = 4.814e-10  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -21.35209  
## sample estimates:  
## mean of x mean of y   
## 15.40 42.98

Como el p-valor=0.0000000004814 es menor que 0.05 se rechaza la hipótesis nula en favor de la laternativa de menor.

### Inferencia para dos poblaciones (bis)

Consideremos que se dispone información para coches europeos y americanos, la cual es codificada como sigue:

europeo = c(0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0)   
 # europeo igual a 1 si el coche es europeo y 0 en caso contrario

Usando la variable anterior vamos dividir en dos la muestra correspondiente a *speed*. Sus principales características se pueden calcular como:

factores = factor(europeo) # codifico 0 y 1 como factores  
 tapply(speed, factores, mean) # help(tapply)

## 0 1   
## 15.05882 16.12500

tapply(speed, factores, var) # en lugar de mean y var se puede usar cualquier función

## 0 1   
## 32.23886 19.58333

Para ver si *speed* tiene la misma media en los coches europeos y americanos debemos ejecutar:

var.test(speed~europeo)

##   
## F test to compare two variances  
##   
## data: speed by europeo  
## F = 1.6462, num df = 33, denom df = 15, p-value = 0.3049  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.6277125 3.7222238  
## sample estimates:  
## ratio of variances   
## 1.64624

# t.test(speed~, var.equal=F)  
 t.test(speed~europeo, var.equal=T)

##   
## Two Sample t-test  
##   
## data: speed by europeo  
## t = -0.66126, df = 48, p-value = 0.5116  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.307999 2.175646  
## sample estimates:  
## mean in group 0 mean in group 1   
## 15.05882 16.12500

En primer lugar se tiene que las varianzas pueden considerarse iguales ya que no se rechaza la hipótesis nula (p-valor=0.3049>0.05). Considerando que el intervalo calculado, (0.6277125, 3.7222238), se llegaría a la misma conclusión ya que contiene al 1, por lo que el cociente puede considerarse igual a 1 y, por tanto, las varianzas serían iguales.

Atendiendo al resultado anterior, tendríamos en cuenta el último contraste (varianzas iguales). De nuevo no se rechaza la hipótesis de que las medias sean iguales (p-valor=0.5116>0.05), es decir, la velocidad media es la misma en los coches europeos y americanos al 5% de significación. Considerando el intervalo de confianza asociado, (-4.307999, 2.175646), se llega a la misma conclusión ya que el intervalo contiene al 0.

De esta forma se puede contrastar si dos medias son iguales dependiendo de cualquier factor con dos alternativas (poseer o no una característica de interés).

### Generalización a más de dos poblaciones

Si el factor tiene más de dos alternativas, entonces se ha de realizar una ANOVA de un factor:

terna = rbinom(50,2,0.5) # genero aleatoriamente un factor 0, 1, 2  
 table(terna)

## terna  
## 0 1 2   
## 15 21 14

factores.terna=factor(terna)  
 tapply(speed, factores.terna, summary)

## $`0`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.00 12.50 15.00 15.67 19.00 25.00   
##   
## $`1`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.0 11.0 14.0 14.1 19.0 24.0   
##   
## $`2`  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.00 14.25 17.50 17.07 20.00 24.00

tapply(speed, factores.terna, var)

## 0 1 2   
## 23.52381 32.49048 24.22527

anova.speed=aov(speed~factores.terna) # hago el análisis de la varianza  
 require(car) # necesario para el test de Levene de igualdad de varianzas

## Loading required package: car

leveneTest(anova.speed, center=mean) # homogeneidad de varianzas

## Levene's Test for Homogeneity of Variance (center = mean)  
## Df F value Pr(>F)  
## group 2 0.3483 0.7077  
## 47

summary(anova.speed) # resumen del ANOVA

## Df Sum Sq Mean Sq F value Pr(>F)  
## factores.terna 2 75.9 37.96 1.379 0.262  
## Residuals 47 1294.1 27.53

En este caso no rechazo que las varianzas sean iguales (p-valor=0.4853>0.05) y que las tres medias son iguales (p.valor=0.884>0.05). En caso de rechazarla, sería interesante ver donde falla la igualdad. Con la siguiente orden se hacen todos los posibles pares:

TukeyHSD(anova.speed)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = speed ~ factores.terna)  
##   
## $factores.terna  
## diff lwr upr p adj  
## 1-0 -1.571429 -5.864453 2.721596 0.6518033  
## 2-0 1.404762 -3.314312 6.123835 0.7526960  
## 2-1 2.976190 -1.405359 7.357740 0.2376441

## Ejercicios propuestos

* Estudiar si la distancia de frenado es igual a 43 pies.
* Estudiar si la distancia de frenado es mayor que 45 pies.
* Estudiar si la distancia de frenado es menor que 40 pies.
* Calcular la media y desviación típica de *dist* en función de si el coche es europeo o americano.
* Estudia si la varianza de *dist* coincide para valores de la variable que distingue entre coches europeos y americanos.
* Estudia si la media de *dist* coincide para valores de la variable que distingue entre coches europeos y americanos.
* Generar una variable que tenga tres valores de forma aleatoria y ver si *dist* tiene la misma media en los grupos formados por dicha variable aleatoria.