Jugando con pruebas estadísticas

00R Team

Abril 2016

Índice

1. Nueve de cada diez dentistas

Ya sabemos que la prueba t de Student sirve para comparar medias de poblaciones y nos permite calcular intervalos de confianza para la media. Pero, ¿y si nuestras variables no son siquiera continuas? En ese caso, hemos de utilizar otro contraste diferente. En la siguiente tabla se recogen la principales pruebas, paramétricas y no paramétricas, que podemos aplicar según el tipo de datos que tengamos y el contraste deseado:

Comparar \ Tipo de datos	Paramétricos	Ordinales	Categóricos
Dos grupos indep	t independiente	Mann-Whitney	Exacto de Fisher
Dos grupos dep	t dependiente	Wilcoxon	McNemar
Dos o más grupos indep	ANOVA de una vía	Kruskal-Wallis	Chi-cuadrado
Dos o más grupos dep	ANOVA medidas repetidas	Friedman	Q de Cochran

Supongamos que queremos comprobar si es equivalente que 9 de cada 10 dentistas avalen un producto a que lo hagan 90 de cada 100. Utiliza la función `binom.test()` para calcular los intervalos de confianza para el parámetro "proporción de dentistas que avalan el producto" en cada caso.

Ojo: Lee la ayuda de R con la expresión `?binom.test()` y fíjate bien en la información que te facilita el test: `number of successes` (número de éxitos) y `number of trials` (número de ensayos).

```
# Caso 9 de cada 10
# prop.test(9, 10)
binom.test(9, 10)
# binom.test(c(9,10)) # Mal
```

```
# Caso 90 de cada 100
# prop.test(90, 100)
binom.test(90, 100)
```

A modo de reflexión (2)

1. En el primer caso (9 de cada 10 dentistas) si nos fijamos en el intervalo de confianza ¿podemos afirmar que la proporción de dentistas que avalan el producto es distinto de 0.6? ¿Podemos afirmarlo para el segundo caso (90 de cada 100)?
2. A la vista de los resultados ¿es equivalente afirmar que 9 de cada 10 dentistas avalan un producto a que lo hagan 90 de cada 100? ¿Cuál es la diferencia? ¿Qué ocurre con 900 de cada 1000?

2. Un caso típico de contraste con dos poblaciones

El conjunto de datos `turtles2.csv` contiene medidas sobre 54 tortugas tomadas de dos grupos: SREL (*Savannah River Ecology Lab*) y CMNH (*Carnegie Museum of Natural History*), estos datos también están disponibles en `http://ares.inf.um.es/00Rteam/datos`. El sexo de cada tortuga es descrito como "F" o "M", también se especifica la longitud y el ancho del caparazón y, por último, la altura del caparazón y el plastrón (el equivalente al caparazón inferior).

Lee la base de datos `turtles2.csv` y, siguiendo el material de trabajo de clase (comprobación de supuestos y especificación de los argumentos en la función), realiza un contraste de hipótesis para comprobar si la altura (`height`) es diferente entre grupos (`SREL` o `CMNH`).

```r
par( cex = 0.75)
# Conjunto de datos con muchos detalles a tener en cuenta
# turtles <- read.csv("http://people.sc.fsu.edu/~jburkardt/datasets/stats/turtles.csv",
#                 header = FALSE, sep = ",", skip = 1, strip.white = TRUE,
#                 col.names = c("index", "collection", "sex", "length", "width", "height"))

turtles <- read.table( "turtles2.csv", header = TRUE, sep = "," )
head( turtles )
dim( turtles )

shapiro.test( turtles[ turtles$collection == "SREL", "height" ] )
shapiro.test( turtles[ turtles$collection == "CMNH", "height" ] )

var.test( turtles$height ~ turtles$collection )

t.test( turtles$height ~ turtles$collection, paired = FALSE, var.equal = TRUE )

boxplot( turtles$height ~ turtles$collection )
```

A modo de reflexión (3)

1. ¿Los tamaños de las muestras según los grupos están balanceados?
2. ¿Aceptamos la condición de normalidad?
3. Por defecto, si no especificamos la homocedasticidad en la función `t.test()` con el argumento `var.equal`, ¿se realiza la prueba t con corrección de Welch?
4. ¿Tenemos criterio para afirmar la diferencia entre los dos grupos? ¿Podemos afirmar que no hay diferencia de altura entre los grupos?
5. Realizando un boxplot de la altura (`height`) según el grupo (`collection`) ¿te podrías haber hecho una idea de lo que te iba a salir el contraste de homocedasticidad?

3. Alternativas al contraste con dos poblaciones

Las distintas pruebas estadísticas pueden aplicarse para responder a la misma pregunta por vías alternativas. Vamos a analizar las diferencias y semejanzas en los resultados en la comparación de dos poblaciones independientes con la ayuda de: t–test, ANOVA, regresión lineal y test de aleatorización.

Para ello, estudiaremos como afecta el sexo a la estatura de estudiantes de la Licenciatura de Biología matriculados en el curso 200-03 en la asignatura "Ecología metodológica y cuantitativa". Utilizaremos los datos recogidos en el fichero `biom2003.dat`, que asignaremos al objeto `biom`; estos datos están disponibles en `http://ares.inf.um.es/00Rteam/datos`.

```
biom <- read.table( "biom2003.dat" )
head( biom )
summary( biom )
```

Gráficamente podemos ver las diferencias mediante un diagrama de cajas y bigotes:

```
boxplot( biom$Altura ~ biom$Sexo, cex.axis = 0.75 )
```

o mediante esta alternativa:

```
par( cex = 0.75  )
orden <- order( biom$Sexo )
dotchart( biom$Altura[ orden ], group = factor( biom$Sexo[ orden ] ) )
```

Verificamos la normalidad y homogeneidad de varianzas:

```
shapiro.test( biom$Altura[ biom$Sexo == 1 ] )
shapiro.test( biom$Altura[ biom$Sexo == 2 ] )
var.test( biom$Altura ~ biom$Sexo )
```

3.1. t–test

Comprueba en las ayudas de la función `t.test` qué significan los distintos argumentos empleados en el siguiente código; evalúa las expresiones y analiza los resultados; ten cuidado, pues algunas expresiones pueden ser erróneas.

```
t.test( biom$Altura, biom$Sexo )
t.test( biom$Altura, biom$Sexo, var.equal = TRUE )
t.test( biom$Altura, factor( biom$Sexo ), var.equal = TRUE  )
t.test( biom$Altura ~ biom$Sexo, var.equal = TRUE  )
t.test( biom$Altura ~ factor( biom$Sexo ), var.equal = TRUE  )
```

A modo de reflexión (4)

1. ¿Se obtienen los mismos resultados con todas las expresiones?
2. Si alguna de la expresiones no funciona, intenta explicar el motivo. Pista: contrasta lo indicado en la expresión y lo indicado en la ayuda.
3. ¿Es necesario usar el argumento `paired = FALSE`?
4. ¿Qué expresión te parece más clara? ¿Por qué?

3.2. ANOVA

Podemos considerar que la prueba t es un caso particular de ANOVA cuando tenemos sólo dos niveles en el factor. Para realizar una prueba ANOVA recurrimos a la función `aov`:

```
summary( aov( biom$Altura ~ biom$Sexo ) )
```

A modo de reflexión (5)

1. ¿Podemos decir que los resultados obtenidos con esta prueba son equivalentes a los obtenidos con t–test?
2. ¿Se utilizan los mismos estadísticos en las dos pruebas? ¿Se obtiene el mismo `p-value`?

3.3. Regresión lineal

Comparemos, ahora, los resultados considerando que existe una recta que pasa por las medias de los dos grupos (hombre y mujeres).

```
by( biom$Altura, factor( biom$Sexo ), mean )
diff( by( biom$Altura, factor( biom$Sexo ), mean ) )
summary( lm( biom$Altura ~ biom$Sexo ) )
summary( lm( biom$Altura ~ factor( biom$Sexo ) ) )
```

A modo de reflexión (6)

1. ¿Podemos decir que los resultados obtenidos con esta prueba son equivalentes a los obtenidos con t–test? ¿Se utilizan lo mismo estadísticos en las dos pruebas? ¿Se obtiene el mismo `p-value`?
2. Comparando los resultados de aplicar `lm()` considerando, o no, la variable sexo como un factor. ¿Qué valor se obtiene para la intersección con el origen de ordenadas (`Intercept`)? ¿Puedes relacionar este valor con alguna de las medias de los grupos?
   ¿Qué pasa con el coeficiente asociado a sexo?
3. En caso de variables cualitativas codificadas con valores numéricos ¿Crees que es necesario trabajar con estas consideradas como factores?

3.4. Aleatorización

Una alternativa a los test convencionales es recurrir a la aleatorización, calculando, en este caso las diferencias muestrales asignando de forma aleatoria el sexo a las observaciones y comparando estos valores con el valor muestral.

Para simplificar el código creamos en primer lugar una función que devuelva el valor absoluto de las diferencias por niveles, asignando los niveles aleatoriamente o dejando el valor muestral.

```
diferenciaMediasAle <- function( x, y, alea = TRUE ){
    y <- factor( y )
    if( alea ) y <- sample( y )
    abs( diff( by( x, y, mean ) ) )
}
```

Para nuestro estudio aleatorizamos obteniendo 999 medias donde *no debería influir* la variable sexo y la correspondiente al valor muestral. Si el valor muestral es mayor que cualquier aleatorización, o la mayoría de ellas, podremos afirmar que hay diferencias.

```
difMuestral <- diferenciaMediasAle( biom$Altura, biom$Sexo, alea = FALSE )
mediasAelatorias <- sapply( 1:999, function( x ) diferenciaMediasAle( biom$Altura, biom$Sexo )  )
medias <- c( mediasAelatorias, difMuestral )
```

Para ver el valor muestral en relación con los aleatorizados utilizamos un método gráfico o contamos directamente los valores.

```
hist( medias )
points( difMuestral, 0, pch = 18, col = 2, cex = 1.5 )
sum( medias < difMuestral )
```

A modo de reflexión (7)

1. El resultado obtenido nos lleva a la misma conclusión que las pruebas anteriores.
2. ¿Qué habría ocurrido de estar en una situación donde no hubiera diferencias?
   Simula una situación para dos variables independientes, una con distribución $\mathrm{N}(150, 25)$ y la otra con una $\mathrm{B}(1, 0{,}5)$ con un tamaño muestral de 50.