Taller Estadística en R: Sesión 3

Abril 2022

Prueba Binomial y Pruebas Chi Cuadrado

Prueba Binomial

La prueba binomial, también conocida como la distribución de Bernoulli, compara las frecuencias observadas de las dos categorías de una variable dicotómica con las frecuencias esperadas en una distribución binomial con un parámetro de probabilidad especifica. El test binomial exacto se emplea para estudiar si la proporción de eventos verdaderos de una variable de tipo binomial se diferencia significativamente de la frecuencia teórica con la que se esperaría que apareciesen.

Supuestos de la prueba binomial

La prueba binomial tiene 3 supuestos:

1. La cantidad de ensayos (n) es fija
2. Los ensayos son independientes
3. La probabilidad de éxito es igual para cada ensayo

Y puede usarse para:

1. Calcular la probabilidad de cierto escenario
2. Probar la hipótesis de si las frecuencias de las categorías son iguales

Cómo calcular la probabilidad de cierto escenario

Para determinar la probabilidad de éxitos se usa la siguiente fórmula:

$$Pr[Exito] = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

donde p es la probabilidad de éxitos, n es el número de ensayos y x es el número de éxitos

$$p = \frac{x}{n}$$

En R, se puede usar utilizando el comando **dbinom()**, el cual tiene la siguiente estructura:

dbinom(x,n,p), en donde x es el número de éxitos, n es el número de ensayos y p es la probabilidad de éxito

Por ejemplo: Digamos que tenemos una población con 27 individuos. La probabilidad de tener el pistilo hacia la derecha es de 0.25. ¿Cuál es la probabilidad de escoger 6 flores con el pistilo a la derecha?

```
dbinom(6,27,0.25)
```

```
## [1] 0.171883
```

```
#La probabilidad de tener 6 éxitos, en 27 ensayos cuando la probabilidad de éxito es de 0.25
```

Esto significa que tenemos un 17% de probabilidades de escoger 6 flores con el pistilo hacia la derecha

Cómo probar si las frecuencias observadas para ambas categorías son diferentes

Para esto podemos usar el comando **binom.test()**, el cual tiene la siguiente estructura:

binom.test(x,n,p,alternative) en donde x es el número de éxitos, n es el número de ensayos, p es la probabilidad de éxito bajo la hipótesis nula, alternative es si la prueba es de una cola o dos colas, y conf.level establece el intervalo de confianza

Por ejemplo digamos que en un quiz de estadística hay 10 preguntas, con 5 probables respuestas cada una. Un estudiante respondió correctamente 4 de 10 preguntas. ¿Sabe el estudiante de estadística?

H_0 = El estudiante contestó al azar. $p_0 = 0.2$

H_A = El estudiante contestó mejor que azar. $p_0 > 0.2$

```
binom.test(4,10,0.2,alternative = "greater",conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  4 and 10
## number of successes = 4, number of trials = 10, p-value = 0.1209
## alternative hypothesis: true probability of success is greater than 0.2
## 95 percent confidence interval:
##  0.1500282 1.0000000
## sample estimates:
## probability of success
##                    0.4
```

```
#como queremos saber si respondió mejor que el azar, ponemos alternative="greater"
```

Como el valor p es mayor a 0.05, podemos decir que fallamos en rechazar la hipótesis nula y concluimos que el estudiante no contestó mejor que al azar

Digamos que queremos saber si las abejas pueden distinguir la presencia de la araña depredadora *Thomisus spectabilis* en las flores de las cuales se alimentan si tenemos 33 flores seleccionadas por las abejas, en donde 24 tenían arañas.

H_0 = Las abejas seleccionan las flores al azar. $p = 0.5$

H_A = Las abejas no seleccionan las flores al azar. $p \neq 0.5$

```
binom.test(24,33,0.5,alternative = "two.sided",conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  24 and 33
## number of successes = 24, number of trials = 33, p-value = 0.01353
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5447616 0.8670036
## sample estimates:
## probability of success
##              0.7272727
```

Pruebas Chi Cuadrado

Las pruebas Chi Cuadrado son pruebas de bondad de ajuste, es decir, que nos permiten saber qué tan bien se ajustan nuestras observaciones a una distribución establecida o pruebas de contigencia, que nos permiten analizar asociación

Supuestos de las pruebas Chi Cuadrado

- Muestreo al azar
- Frecuencia esperada nunca menor de 1
- Menos del 20% de las categorías pueden tener esperados menores de 5

Si no se cumplen estos supuestos se pueden juntar categorías para que se cumplan los supuestos

Hay 3 pruebas chi cuadrado importantes:

- Chi Cuadrado por modelo proporcional (También conocida como Goodness-of-fit)
- Chi cuadrado de Poisson
- Chi cuadrado de contigencia

Chi Cuadrado por modelo proporcional Es una extensión aproximada de la prueba binomial que se puede usar cuando hay más de dos categorías. La hipótesis nula de la prueba es que las observaciones siguen una distribución acorde a la distribución del modelo de referencia. El modelo proporcional es la hipótesis nula, la frecuencia de eventos es proporcional al número de oportunidades. Solo se usa con valores positivos.

$\chi^2 = \sum \frac{(f_{obs} - f_{esp})^2}{f_{esp}}$

Si el χ^2 resultante es alto significa que hay mucha diferencia entre lo esperado y lo observado

¿Cómo sé si rechazo o no la hipótesis nula?

A partir del *valor crítico* se establece un límite o "bondad" en la cola de distribución

El valor crítico se establece a partir de los *grados de libertad*, los grados de libertad están dados por el número de valores que pueden ser asignados de forma arbitraria, antes de que el resto de las variables tomen un valor automáticamente, producto de establecerse las que son libres; esto, con el fin de compensar e igualar un resultado el cual se ha conocido previamente.

En este caso, los grados de libertad son el número de categorías (k) - el número de parámetros. Al tener los grados de libertad podemos mirar en la tabla cual es el valor crítico para esos grados de libertad y ese α específico

Si el valor de χ^2 calculado es **mayor** que el valor de χ^2 de la tabla **se rechaza la hipótesis nula**

Si el valor de χ^2 calculado es **menor** que el valor de χ^2 de la tabla **se falla en rechazar la hipótesis nula**

En R se puede hacer esta prueba con el comando **chisq.test()**

Digamos que queremos saber si la riqueza aves se da al azar dentro de 7 relictos de bosques.

H_0 : La probabilidad de registrar diferentes especies de aves es igual en todos los relictos de bosques H_A : La probabilidad de registrar diferentes especies de aves no igual en todos los relictos de bosques

```
riqueza_aves<-c(33,42,63,64,47,56,47)
probabilidad_nula<-rep((1/7),7)
chisq.test(riqueza_aves,p=probabilidad_nula,correct = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  riqueza_aves
## X-squared = 15.341, df = 6, p-value = 0.01776
```

Como el valor p<0.05, concluimos que rechazamos la hipotésis nula e inferimos que no en todos los relictos de bosque detecto la misma cantidad de especies de aves

```r
qchisq(0.95,6) #Valor crítico para un alfa de 0.05 y 6 grados de libertad
```

```
## [1] 12.59159
```

Como el valor crítico nos dio menor que el valor de χ^2 que hallamos, concluimos de nuevo que se rechaza la hipótesis nula.

Chi cuadrado de Poisson Se aplica cuando el interés está en el tiempo o en el espacio.

La hipótesis nula es que los eventos ocurren al azar en el tiempo o en el espacio. La fórmula en este caso es:

$$Pr[X éxitos] = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

$$\lambda = \bar{Y}$$

Digamos que por ejemplo queremos saber si la distribución de murciélagos en ciertos árboles frutales se da al azar o no.

Y tenemos que en 120 árboles no habían murciélagos, en 82 árboles había 1 murciélago, en 49 árboles habían 2 murciélagos, en 74 árboles habían 3, en 21 árboles habían 4, en 15 árboles habían 5 y en 7 árboles habían 6 murcis.

La hipótesis nula es que los murciélagos se distribuyen aleatoriamente entre los árboles

```r
murcis<-rep(c(0:6),c(120,82,49,74,21,15,7))
murcis
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [223] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3
## [260] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [297] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4
## [334] 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6
```

```r
chisq.test(0:6,p=dpois(0:6,mean(murcis)),rescale.p = TRUE,simulate.p.value = TRUE)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  0:6
## X-squared = 385.85, df = NA, p-value = 0.0004998
```

```r
#rescale.p hace que si los valores de probabilidad sumen 1
#simulate.p hace que se simulen 2000 veces los valores de valor P
```

Como el valor P>0.05, concluimos que la distribución de murciélagos en los árboles no es aleatoria.

Chi cuadrado por contigencia Es una prueba para estudiar relaciones de dependencia entre variables categóricas

La hipótesis nula es que hay independencia entre las dos variables.

En esta prueba los grados de libertad son: (número de filas-1)(número de columnas-1)

Digamos que queremos saber si hay relación entre el sexo de los estudiantes y aprueban o no un examen.

Nuestra hipótesis nula es que no hay asociación entre ambas variables

```
datos_estudiantes<-matrix(c(15,30,40,26),nrow = 2,ncol = 2,byrow = T,
dimnames = list("sexo"=c("Hombres","Mujeres"),"nota"
                =c("Aprobado","Reprobado")))
datos_estudiantes
```

```
##          nota
## sexo      Aprobado Reprobado
##   Hombres       15        30
##   Mujeres       40        26
```

```
chisq.test(datos_estudiantes,correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  datos_estudiantes
## X-squared = 7.9613, df = 1, p-value = 0.004779
```

```
#La corrección de Yares sobrestima el valor p, por eso lo quitamos
```

Como nuestro valor P<0.05, concluimos que rechazamos la hipótesis nula y si hay asociación entre el sexo y si aprueban o no un examen