Ejercicios 3

Estadística No Paramétrica

Joaquin Cavieres G.

Introducción

Dentro de la estadística no paramétrica uno de los test estadísticos que podemos encontrar es el test (o prueba) X^2. Este test permite (en los casos que vamos a evaluar hoy en clases) considerar una muestra aleatoria de una variable y realizar una bondad de ajuste. Generalmente los datos se observan en frecuencias y basados en la función de distribución X^2. En muchas ocasiones nosotros estamos interesados en saber si una muestra aleatoria n proviene de una función distribución asociada a la población. Para esto, y mediante este test, podemos generar categorías o subconjuntos finitos $O_1, O_2, ..., O_k$ para luego compararlas con las frecuencias esperadas a cada una de ellas.

Test de bondad de ajuste X^2

Asumamos que tenemos un número k de clases en las cuales tenemos n observaciones asociadas a cada una de ellas. Estas clases las llamaremos como *frecuencias observadas* y las denotaremos como $O_1, O_2, ..., O_k$. La sumatoria de todas las frecuencias asociadas a estas clases es igual a n.

La idea principal de este test es comparar las frecuencias observadas con las *frecuencias esperadas* (esperanza matemática) quienes comúnmente son denotadas como $E_1, E_2, ..., E_k$, en donde la sumatoria de estas frecuencias esperadas también es igual a n.

La siguiente tabla muestra una idea general de lo anterior:

	$Clase_1$	$Clase_2$	....	$Clase_k$	Total
Freq Obs	O_1	O_2	....	O_k	n
Freq Esperada	E_1	E_2	....	E_k	n

de acuerdo a la tabla anterior entonces vamos a determinar si las frecuencias esperadas estan en concordancia con las frecuencias observadas. Para determinar esto primero determinar el estadístico de prueba:

$$T = \sum_{i=1}^{k} \frac{(O_i - \mathbb{E}_i)^2}{\mathbb{E}_i}, \tag{1}$$

para cada O_i y E_i correspondiente a cada clase k y el estadístico T corresponde a la suma de k números positivos. Cuanto menor sea el valor del estadístico T entonces más concordancia entre las frecuencias observadas y las frecuencias esperadas. En este tipo de prueba se rechaza H_0 cuando el estadístico T es mayor a un determinado valor númerico α.

Dado lo anterior es necesario señalar que:

- El valor de T se podrá aproximar a una distribución X^2 cuando el tamaño muestral sea grande ($n \geq 30$) y todas las frecuencias esperadas $\geq$ a 5.
- Los datos observados deben ser obtenidos desde una población con clases o categorías bien definidas.

Test de bondad de ajuste X^2 multinomial

Un test de bondad de ajuste asumiendo una distribución multinomial para las frecuencias observadas es una generalización del test de bondad de ajuste binomial. En este caso debemos considerar.

- El ensayo contiene n pruebas idénticas e independientes.
- En cada test hay un número k de resultados posibles.
- Cada uno de los k posibles resultados tiene una probabilidad de ocurrencia p_i con $(p_1 + p_2 + ... + p_k = 1)$, la cual permanece constante den todo el proceso.
- Se obtienen del ensayo/experimento frecuencias observadas $O_1, O_2, ..., O_k$ para cada resultado con la sumatoria de todas ellas igual a n.

El estadístico de prueba es el mismo que en (1) con $k-1$ grados de libertad. La frecuencia esperada se calcula como:

$$\mathtt{E}_i = n * p_i, \qquad i = 1, ..., k \tag{2}$$

Ejercicios

Un profesor le comunica a otro que en su curso el 15 % de los estudiantes de álgebra terminan el semestre con un 6, el 20 % termina con 5, un 25 % con un 4, un 10 % con un 3 y un 30 % con un 2. Dentro del curso había 1000 y los resultados finales fueron: 160 con un 6, 190 con un 5, 240 con un 4, 115 con un 3 y 295 con un 2.

Determine si las calificaciones observadas difieren significativamente de la distribución descrita por el profesor con un α del 10 %.

Desarrollo manual

```
n = 1000
O_i = c(160,190,240,115,295)
E_i = c(.15,.20,.25,.10,.30)
E = n * E_i
E
```

```
## [1] 150 200 250 100 300
```

```
# Calculo manual del estadístico T
T_test = sum((O_i - E)^2/ E)
T_test
```

```
## [1] 3.9
```

```
# Calculo manual del p-value del test con 5 - 1 grados de libertad (k -1)
pchisq(c(3.9), df=4, lower.tail=FALSE)
```

```
## [1] 0.4197085
```

Ya que el p-value $>\alpha$ no rechazamos la hipótesis nula (H_0), por lo tanto, la distribución asumida es la correcta.

Desarrollo en R

En R podemos hacer este mismo calculo de bondad de ajuste a través de la función `chisq.test`. Esta función tiene como argumentos a `(x, correct, p)` donde `x` es el vector de datos de las frecuencias observadas, `correct` indica si se desea utilizar la correción de Yates (esto es generalmente `TRUE` cuando el vector `x` contiene varias entradas con números pequeños) y `p` es un fector de frecuencias relativas esperadas especificada por H_0.

```
O_i = c(160,190,240,115,295)
E_i = c(.15,.20,.25,.10,.30)
O_i
```

```
## [1] 160 190 240 115 295
```

```r
E_i
```

```
## [1] 0.15 0.20 0.25 0.10 0.30
```

```r
chisq.test(O_i, correct = FALSE, p = E_i)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  O_i
## X-squared = 3.9, df = 4, p-value = 0.4197
```

Ejercicio 2

Existen 8 secciones (de la A a la H respectivamente) ofreciendo ciertos cursos para un determinado semestre y se supone que existen el mismo número de alumnos matriculados para cada una de las 8 secciones. Dado los antecedentes luego de las primeras semanas las matrículas son las siguientes:

Sección A	Sección B	Sección C	Sección D	Sección E	Sección F	Sección G	Sección H
32	29	30	26	33	11	24	27

Determine si las matrículas observadas difieren significativamente de la distribución descrita con un α del 10 %.

```r
O_i = c(32,29,30,26,33,11,24, 27)
E_i = c(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)
chisq.test(O_i, correct=FALSE, p=E_i)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  O_i
## X-squared = 12.755, df = 7, p-value = 0.07832
```

Dado que el p-value $<\alpha$, rechazamos H_0 (con un nivel de significancia del 10 %). Esto significa que los números de matrículas no estan distribuidas igualmente en todas las secciones

Ejercicio 3

Según un estudio preliminar en Valparaíso la proporción de mujeres nacidas es mayor a la de los hombres durante los últimos 5 años. Para determinar si esto es cierto se toma una muestra aleatoria durante los últimos 2 años constatando el género de cada uno de ellos. La pregunta que queremos responder es, ¿existen diferencias significativas en las proporciones de nacidos?

Desarrollo en R

Hipótesis nula: H_0 = No existen diferencias significativas en la proporción de ambos sexos ($p = 0{,}5$)
Hipótesis alternativa: H_1 = Existen diferencias significativas en la proporción de ambos sexos ($p \neq 0{,}5$)

```
data = c("masculino", "masculino", "masculino", "masculino", "femenino",
           "masculino", "masculino", "masculino", "femenino", "femenino",
           "femenino", "femenino", "femenino", "femenino", "masculino",
           "masculino", "femenino", "masculino", "femenino", "masculino",
           "femenino", "masculino", "masculino", "masculino", "femenino",
           "masculino", "masculino", "femenino", "masculino", "femenino")
tabla = table(data)
tabla
```

```
## data
##  femenino masculino
##        13        17
```

Test X^2 bondad de ajuste:

```
chisq.test(x = c(13, 17), p = c(0.5, 0.5), correct = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(13, 17)
## X-squared = 0.53333, df = 1, p-value = 0.4652
```

```
# O tambien podemos declarar el test incluyendo la tabla:
chisq.test(tabla, p = c(0.5, 0.5), correct = FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tabla
## X-squared = 0.53333, df = 1, p-value = 0.4652
```

Se determinó mediante el test X^2 que no existen diferencias significativas entre las frecuencias observadas y las frecuencias esperadas si la verdadera probabilidad de nacimiento es del 50 % para ambos sexos.