Estadística No Paramétrica

Clase 3

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Test X^2

Generalmente cuando hacemos inferencia sobre los parámetros de una población, debemos identificar si los datos observados se ajustan a alguna distribución de probabilidad conocida (por ejemplo, Normal, Gamma, Poisson, etc). Este procedimiento se le conoce como Bondad de ajuste y en el cual queremos contrastar:

$$H_0 : F_X(x) = F_X^0(x)$$
$$H_1 : F_X(x) \neq F_X^0(x)$$

Donde F_X^0 puede ser una función que no esta completamente especificada y F_X es distribución de los datos observados.

Test X^2

En los test relacionados a la distribución Binomial analizamos variables aleatorias con sólo dos posibles resultados para la variable respuesta. Ahora, mediante el test X^2, podemos extender esa idea a variablea aleatorias discretas cuyo rango consiste un número general de categorías.

Test X^2

En los test relacionados a la distribución Binomial analizamos variables aleatorias con sólo dos posibles resultados para la variable respuesta. Ahora, mediante el test X^2, podemos extender esa idea a variablea aleatorias discretas cuyo rango consiste un número general de categorías.

Definición

El estadístico de la prueba X^2 es esencialmente la suma de las categorías de las diferencias al cuadrado y estandarizadas entre las frecuencias observadas y esperadas, donde las frecuencias esperadas se formulan bajo el supuesto de que la hipótesis nula (H_0) es verdadera.

Test X^2

En general, bajo H_0, este estadístico de prueba tiene una distribución X^2 asintótica con grados de libertad iguales al número de categorías menos el número de parámetros para formar las frecuencias esperadas.

Test X^2

En general, bajo H_0, este estadístico de prueba tiene una distribución X^2 asintótica con grados de libertad iguales al número de categorías menos el número de parámetros para formar las frecuencias esperadas.

Observación

Puede utilizarse la distribución nula exacta o la distribución asintótica para evaluar este test.

Ejemplo

Se tiene una muestra aleatoria $X_1, ..., X_n$ desde F_X. De lo anterior se plantea la siguiente hipótesis:

$$H_0 : F_X(x) = F_X^0(x)$$
$$H_1 : F_X(x) \neq F_X^0(x)$$

Desarrollo

Dividimos el rango de observaciones en k clases y construimos una tabla de contingencia en donde se cuenta el número de observaciones en casa clade k:

Clase 1	Clase 2	Clase 3	$\cdots$	Clase k
O_1	O_2	O_3	$\cdots$	O_k

Test X^2

F_X es la verdadera pero desconocida distribución y F_X^0 es una función completamente especificada conocida (si es así, entonces podemos ejecutar la prueba estadística)

Estadístico de prueba

Sea p_i la probabilidad de que una observación se encuentre en alguna clase i bajo H_0, es decir, $F_X(x) = F_X^0(x)$. Por lo tanto, podemos definir a la esperanza matemática $\mathbb{E}_i$ como el valor esperado en cada clase como:

$$\mathbb{E}_i = \mathbb{E}(O_i) = p_i n, \quad i = 1, ..., k$$

donde $\mathbb{E}_i$ es la esperanza de las observaciones en la clase i bajo H_0.

Test X^2

Ahora, considere:

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

Entonces, bajo H_0, se esperaría que T tenga valores pequeños ya que se espera que O_i sea muy cercano a $\mathbb{E}_i$.

Test X^2

Ahora, considere:

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

Entonces, bajo H_0, se esperaría que T tenga valores pequeños ya que se espera que O_i sea muy cercano a $\mathbb{E}_i$.

Observaciones

- En algunos libros se puede encontrar que T (el cual hemos definido previamente) sea expresado como X^2.
- Las clases O_i pueden llamarse como las *frecuencias observadas* de las categorías de X.
- Las frecuencias observadas O_i estan restringias a $\sum_{i=1}^{k} O_i = n$, así que hay $k-1$ categorías libres.

Test X^2

El teorema de Pearson garantiza que bajo H_0:

$$T \sim X^2_{(k-1)}$$

por lo tanto, rechazamos H_0 si T tiene un valor alto.

Ejemplo 1

Suponga que lanzamos un dado 350 veces ($n = 350$) y observamos las siguientes frecuencias (55, 50, 60, 65, 66, 60). Nosotros estamos interesados en testear que los lanzamientos estan efectivamente equilibrados, entonces la hipótesis nula es $H_0 : p = p_i = 1/6$.

Ver codigo en `R`

Test X^2

Desarrollo en R

```
x = c(55, 50, 60, 65, 66, 60)
Tfit = chisq.test(x)
Tfit
round(Tfit$expected,digits=4)
```

`round((Tfit$residuals)`^{2}`,digits=4)` (aka. residuales de Pearson)

Test X^2

Desarrollo en R

```
x = c(55, 50, 60, 65, 66, 60)
Tfit = chisq.test(x)
Tfit
round(Tfit$expected,digits=4)
```

`round((Tfit$residuals)`^{2}`,digits=4)` (aka. residuales de Pearson)

Por lo tanto, no hay evidencia que respalde que el dado no este equilibrado!!

Test X^2

Ejemplo 2

Sólo se conoce la forma de la densidad de probabilidad y los parámetros deben ser estimados, entonces, los valores esperados son las estimaciones de $\mathbb{E}_i$ basadas en densidad de probabilidad. La siguiente tabla muestra el número de varones en los primeros siete hijos de ministros suecos para un $n = 1334$ (John Kloke, J and McKean, J.W., 2014 (Nonparametric Statistical Methods Using R)).

Num de varones	0	1	2	3	4	5	6	7
Num de ministros	6	57	207	362	365	256	69	13

Test X^2

Hipótesis nula: El número de hijos tiene una distribución Binomial con probabilidad de éxito p.

Desarrollo

El valor de p puede encontrarse mediante máxima verosimilitud como:

$$\hat{p} = \frac{\sum_{i=0}^{7} i * O_i}{7 * 1334} = 0.5140$$

y la esperanza es cálculada mediante:

$$\mathbb{E}_i = n \binom{7}{i} \hat{p}^i (1 - \hat{p})^{7-i}$$

Test X^2

Los valores de la densidad de probabilidad pueden ser calculadas en R junto al test X^2.

Desarrollo en R

Ver codigo (Ejemplo 2)

Test X^2

Regla de decisión

Rechazar H_0 si :

$$T > X_{k-1}^{2(1-\alpha)}$$

Consideraciones importantes

- Si algunos $\mathbb{E}_i$ tienen valores pequeños entonces la aproximación hacia X^2 no es buena.
- El número de clases k es arbitrario.
- Clases con $\mathbb{E}_i$ pequeños deben combinarse con otras clases con el fin de que sólo 20% de las $\mathbb{E}_i$ sean menores a 5 y ninguna menor a 1.

Test X^2

Distribución exacta de T

El estadístico de prueba T que vimos previamente sigue una distribución aproximada a X^2, entonces, ¿cual es la distribución exacta?

La obtención explítica de la distribución de T es complicada debido a los calculos que hay detrás, pero gracias a las simulaciones computacionales, podemos aproximarla adecuadamente junto a sus caracteristicas propias.

Test X^2

Considere la siguiente situación:

F_X^0 esta completamente especificado, las clases k estan definidas, por lo que se pueden determinar las p_i asociadas a cada una de ellas. Dado lo anterior es que el vector de frecuencias observadas $O_1, ..., O_k$ sigue una distribución multinomial de parámetros $(n, p_1, ..., p_k)$ con densidad:

$$\mathbb{P}(O_1 = o_1, O_2 = o_2, ..., O_k = o_k) = \frac{n!}{o_1!o_2!, ..., o_k!} p_1^{o_1}, ..., p_k^{o_k}$$

Test X^2

Vamos a simular el vector $(O_1, ..., O_k)$ para obtener una gran cantidad de estadísticos T (simulados). Además vamos fijar un n grande de simulaciones para estimar los cuantiles de la distribución exacta de T.

Desarrollo en R

Ver codigo (Ejemplo 4)