Clase 23 de Bioestadística

Tablas de contingencia

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Prueba χ^2 de independencia, cuadro de contingencia

En R

Prueba χ^2 de independencia, cuadro de contingencia

- Queremos es determinar si existe o no relación de dependencia entre ciertas características de los datos.
- Dicho de otra forma, queremos determinar si dos variables aleatorias son independientes
- Más concretamente, tenemos dos muestras apareadas $(X_1, Y_1), \ldots (X_n, Y_n)$
- Las variables X_i toman una cantidad finita de valores (c_i)
- Las variables Y_j toman otra cantidad finita (d_j).
- Queremos determinar si las características c son independientes o no de las d.

- Planteamos entonces la hipótesis nula
- H_0 : las X son idependientes de las Y
- H_1 : las X **no** son idependientes de las Y

Ejemplo:

- c_1 indica que la persona es hombre
- c_2 indica que la persona es mujer
- d_1 es si fuma actualmente,
- d_2 es si fumaba y ya no lo hace,
- d_3 si nunca fumó.

Ejemplo numérico

- Tenemos 402 individuos
- 193 son hombres y 209 son mujeres, primero se construye la tabla siguiente, que se denomina tabla de contingencia:

$Y \mid X$	Hombre	Mujer	Total
Nunca fumó	149	148	297
Fue fumador y no fuma	13	24	37
Fuma actualmente	31	37	68
Total	193	209	**402**

Esta tabla nos dice, por ejemplo:

- 13 de los 193 hombres fueron fumadores pero no fuman actualmente,
- 37 de las 209 mujeres fuma actualmente,
- 149 hombres nunca fumaron.
- Bajo la hipótesis de independencia entre los hábitos como fumador o no fumador y el género debería verificarse que, por ejemplo:
- 149 que es la cantidad observada de personas que nunca fumó *y* es hombre, debería ser, en promedio $297 \times 193/402$ que es la cantidad esperada si fueran independientes.
- Y eso para cualquier otra entrada de la matriz 3×2 anterior.

- ¿Cómo se construye el estadístico?
- Calculamos las diferencias entre lo observado y lo esperado.
- Para eso calculamos la tabla con valores esperados:

$Y \vert X$	Hombre	Mujer	Total	%
NF	$193 \times 0{,}739$	$209 \times 0{,}739$	297	73.9
F y NF	$193 \times 0{,}092$	$209 \times 0{,}092$	37	9,2
F	$193 \times 0{,}169$	$209 \times 0{,}169$	68	16.9
Total	193	209	**402**	**100**

- El estadístico que se plantea tiene en cuenta estas diferencias de la siguiente forma.
- Se calculan 6 sumandos uno por cada entrada de la matriz:
- Cada sumando incluye la diferencia de lo esperado y lo observado al cuadrado.
- Para escalar todos los sumandos de la misma manera, dividimos por el valor esperado

El resultado es:

$$\frac{\left(149-\frac{193\times 297}{402}\right)^2}{\frac{193\times 297}{402}}+\frac{\left(148-\frac{209\times 297}{402}\right)^2}{\frac{209\times 297}{402}}$$
$$+\frac{\left(13-\frac{37\times 193}{402}\right)^2}{\frac{37\times 193}{402}}+\frac{\left(24-\frac{37\times 209}{402}\right)^2}{\frac{37\times 209}{402}}$$
$$+\frac{\left(31-\frac{68\times 193}{402}\right)^2}{\frac{68\times 193}{402}}+\frac{\left(37-\frac{68\times 209}{402}\right)^2}{\frac{68\times 209}{402}}=\mathbf{3.17}$$

¿Este valor es grande o pequeño?

Escribamos en términos teóricos

- Llamemos o_{ij} a cada entrada de la matriz anterior, el valor observado con $i = 1, 2, 3$ filas y $j = 1, 2$ columnas.
- Calculamos los valores esperados en la hipótesis nula (independencia):

$$e_{ij} = \frac{(\text{ total de la fila } i) \times (\text{total de la columna } j)}{(\text{total de datos})},$$

- elevamos al cuadrado cada una de estas 2×3 diferencias:

$$(o_{ij} - e_{ij})^2$$

- Para que todos tengan el mismo peso, dividimos entre e_{ij} cada diferencia
- Por último sumamos

El resultado es el estadístico:

$$X_n^2 = \sum_{i=1}^{\text{nro filas}} \sum_{j=1}^{\text{nro col}} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Teorema Cuando $n \to \infty$ se tiene

$$X_n^2 \to \chi_{df}^2$$

donde χ_{df}^2 es una v.a. con distribución Chi cuadrado, con df grados de libertad, donde

$$df = (\text{filas} - 1)(\text{columnas} - 1)$$

son los grados de libertad

- La variable aleatoria χ^2_ℓ (se lee Chi cuadrado con ℓ grados de libertad) es la suma de ℓ variables aleatorias normales estándar, es decir

$$\chi^2_\ell = (Z_1)^2 + \cdots + (Z_\ell)^2$$

donde las Z_i son normales estándar.

- Es como que cada sumando al cuadrado converge a una normal al cuadrado, pero algunas de ellas desaparecen.

Conclusión

- Rechazamos la hipótesis nula de que son independientes si X_n^2 supera un valor crítico t_n.
- La región crítica es de la forma

$$RC = \{X_n^2 > t_n\}$$

- En nuestro caso $df = (3-1)(2-1) = 2$
- Calculamos

$$\texttt{qchisq(0.95,2)=5.99} = t_n$$

- ¿Rechazamos la hipótesis nula?: No

En R

- Primero ingresamos los datos, en formato de vector:
  `datosv<-c(149,13,31,148,24,37)`
- OJO: la tabla se ingresa por columnas
- Ahora ordenamos los datos en una matriz de 2 columnas
  `datosm<-matrix(datosv,ncol=2)`
- Ahora lo convertimos al formato de tabla, que es el que utiliza el test:
  `datost<-as.table(datosm)`
- El comando `chisq.test(datost)` calcula el test anterior

Resultado en R

```
> datosv<-c(149,13,31,148,24,37)
> datosm<-matrix(datosv,ncol=2)
> datost<-as.table(datosm)
> chisq.test(datost)

        Pearson's Chi-squared test

data:  datost
X-squared = 3.1713, df = 2, p-value = 0.2048
```

Conclusión: no rechazo H_0.
No hay evidencia para suponer que no son independientes.

Otro ejemplo

Nos preguntamos la independencia zurdo derecho en relación al género[1]:

	Diestro	Zurdo	TOTAL
Hombre	43	9	52
Mujer	44	4	48
TOTAL	87	13	100

Los datos en vector son.

$$43, \quad 44, \quad 9, \quad 4$$

[1] Ejemplo tomado de Wikipedia

El resultado en R

```
> datos2<-as.table(matrix(c(43,44,9,4),ncol=2))
> chisq.test(datos2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  datos2
X-squared = 1.0725, df = 1, p-value = 0.3004
```

La corrección de Yates se aplica cuando hay pocos datos. Se trata de restar 0.5 a las diferencias:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0{,}5)^2}{E_i}$$

Así se logra una mejor aproximación a la Chi-cuadrado.