Colegio de Postgraduados

Analisis de datos categoricos

Humberto Vaquera Huerta
2024-06-19

Introduccion

Los datos categóricos aparecen cuando una variable se mide en una escala nominal u ordinal. Por ejemplo, una encuesta donde se recoge informacion sobre variables como el genero, estado civil o afiliacion politica. Ademas de distinguir una variable como categorica (cualitativa) o continua (cuantitativa), las variables tambien se puede clasificar como independiente o dependientes. El termino independiente se refiere a una variable que se puede manipular experimentalmente (e.j. el tipo de tratamiento que se le asigna a cada persona), pero también se aplica a menudo a una variable que se utiliza para predecir otra variable (e.j. nivel socioeconomico).

Tablas de Contingencia

Una tabla de contingencia es una forma de resumir datos categoricos. En general, el interes se centra en estudiar si existe alguna asociacion entre una variable denominada fila y otra variable denominada columna y se calcula la intensidad de dicha asociacion.De manera formal, se consideran X e Y dos variables categoricas con I y J categorıas respectivamente. Una observacion puede venir clasificada en una de las posibles $I \times J$ categorıas que existen. Cuando las casillas de la tabla contienen las frecuencias observadas, la tabla se denomina **tabla de contingencia**.

Ejemplo de tabla de Contingencia

A partir de los datos obtenidos a través de los cuestionarios, se analiza la estructura de la opinión de los encuestados sobre el tipo de agricultura (extensiva o intensiva) que se necesita para aumentar la producción y la rentabilidad de las granjas en una Region, dependiendo del tipo de relieve que coincida con las propiedades de los encuestados.

- X -¿Qué tipo de práctica agrícola es apropiada para el condado? (extensivo o intensivo) (I=2)
- Y - Forma de relieve, que coincide con las propiedades de los encuestados (llanuras, colinas, montañas).(J=3)

Especificacion	Extensivo	Intensivo
Planicie	11	40
Colina	17	21
Montaña	10	1

Prueba de independencia chi-cuadrada

La prueba de independencia de chi-cuadrado se utiliza para analizar una tabla de frecuencias formada por dos variables categóricas. La prueba de chi-cuadrada evalúa si existe una asociación significativa entre las categorías de las dos variables.

- **Hipótesis nula (H_0)**: las variables de fila y columna de la tabla de contingencia son independientes.
- **Hipótesis alternativa (H_1)**: las variables de fila y columna dependen

Para cada celda de la tabla, tenemos que calcular el valor esperado bajo una hipótesis nula.

Para una celda dada, el valor esperado se calcula de la siguiente manera:

$$e = \frac{renglon.suma * col.suma}{suma.total}$$

Prueba de independencia chi-cuadrada

La estadística Chi-cuadrada se calcula de la siguiente manera:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

- o is la frecuencia observada
- e is la frecuencia esperada

Esta estadística de Chi-cuadrada calculada se compara con el valor crítico (obtenido de tablas estadísticas) con $df = (r-1)(c-1)$ grados de libertad y $\alpha = 0.05$.

- r es el número de filas en la tabla de contingencia
- c es el número de columnas en la tabla de contingencia

Si el estadístico Chi-cuadrado calculado es mayor que el valor crítico, entonces debemos concluir que las variables de fila y columna no son independientes entre sí. Esto implica que están significativamente asociadas.

Ejemplo

```
datos <- matrix(c(11,17,10,40,21,1),nrow=3,ncol=2)
datos
```

```
##      [,1] [,2]
## [1,]   11   40
## [2,]   17   21
## [3,]   10    1
```

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

```
## Warning in chisq.test(datos, correct = FALSE): Chi-squared approximation ma
## incorrect
```

```
##
##     Pearson's Chi-squared test
##
## data:  datos
## X-squared = 19.647, df = 2, p-value = 5.418e-05
```

Ejemplo 2

En este ejemplo, queremos saber si los rayos X blandos y duros difieren en su efecto sobre la larva de saltamontes. La pregunta es si las larvas alcanzan o no un cierto ciclo de vida dependiendo de si están expuestas a rayos X blandos, rayos X duros, luz o rayos beta. Los datos para este ejemplo se proporcionan a continuación.

	Mitosis not reached	**Mitosis reached**
$X - raysoft$	21	14
$X - rayhard$	18	13
$Beta - rays$	24	12
$Light$	13	30

Datos

```
tabla<- matrix(c(21,14,18,13,24,12,13,30), byrow=T,nrow=4)
colnames(tabla) <- c("reached", "notreached")
rownames(tabla) <- c("rsoft", "rhard", "beta", "light")
tabla
```

```
##       reached notreached
## rsoft      21          14
## rhard      18          13
## beta       24          12
## light      13          30
```

Correlación de Cramer

La correlación V de Cramer mide la asociación entre dos atributos y su valor varía de 0 (que indica que no hay relación entre los atributos) a 1 (que indica una asociación completa entre variables). Alcanza el valor 1 sólo cuando un atributo está completamente determinado por el otro atributo.

Correlación V de Cramer:

$$V = \sqrt{\frac{\chi_c^2}{n(m-1)}}$$

Donde:

- χ^2 es el estadístico chi cuadrado calculado a partir de la tabla de contingencia.
- n es el tamaño total de la muestra (la suma de todas las celdas en la tabla de contingencia).
- m es el menor entre el número de filas y el número de columnas en la tabla de contingencia.

Ejemplo V cramer en r

```
datos2 <- matrix(c(11,17,10,40,21,1),nrow=3,ncol=2)
datos2
```

```
##      [,1] [,2]
## [1,]   11   40
## [2,]   17   21
## [3,]   10    1
```

```
library(rcompanion)
```

```
## Warning: package 'rcompanion' was built under R version 4.2.3
```

```
cramerV(datos2,ci=TRUE)
```

```
##   Cramer.V lower.ci upper.ci
## 1   0.4432   0.2958   0.6028
```

- **Este valor de 0.44 indica una asociación moderada**

Prueba de homogeneidad Chi-Cuadrada de Pearson

La prueba de homogeneidad Chi-cuadrada nos permite evaluar si dos muestras se distribuyen por igual en varios niveles / categorías. El valor p-value indica el nivel de significación estadística de la diferencia entre las distribuciones observadas y esperadas.

Prueba de homogeneidad: ejemplo

Se quiere evaluar si la germinación o no de semillas está asociada a la condición de haber sido tratadas con un fungicida. Para ello dos lotes de tamaño similar, fueron tratadas con fungicida o dejadas como control no tratadas. Luego las semillas se hicieron germinar y se registró el número de germinadas y no germinadas en cada uno de los grupos: control y tratadas con fungicida. El resultado de este conteo se presenta a continuación:

Condición	no germinó	germinó	Total
Control	245	1190	1435
Fungicida	123	1358	1481
Total	368	2548	2916

Ejemplo

Un grupo de ecólogos estudia la distribución de frecuencias de edad en tres especies de roedores asociados a un curso de agua. Se muestrearon 83 ejemplares de cada especie, obteniéndose la siguiente tabla de frecuencias:

$Edad(meses)$	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Especie 1	1	5	30	17	22	8	0
Especie 2	2	5	19	28	23	6	0
Especie 3	3	9	18	24	18	9	2

Probar la hipótesis de que las especies poseen una distribución de edades semejante.

Estudio longitudinal de Framingham sobre enfermedades coronarias.

Muestra 1329 pacientes clasificados de forma cruzada por el nivel de colesterol sérico (por debajo o por encima 260) y la presencia o ausencia de enfermedad cardíaca.

Serum Cholesterol and Heart Disease

Serum Cholesterol	Heart Disease		Total
	Present	Absent	
< 260	51	992	1043
260+	41	245	286
Total	92	1237	1329