Práctica dirigida 8

Contents

FACULTAD DE CIENCIAS SOCIALES - PUCP

Curso: POL 278 - Estadística para el análisis político 1 | Semestre 2023 - 2

Tablas de contingencia y prueba Chi2

Tablas de contingencia

- Son tablas de doble entrada, en las cuales se cruzan las categorías de dos variables de interés.
- En las casillas de la tabla se ubica la frecuencia o el número de casos de cada cruce.
- Conceptos importantes: Frecuencias observadas y frecuencias esperadas.

		VARIABLE 1		
		Valor 1	**Valor 2**	**Valor 3**
VARIABLE 2	**Valor 1**	Frecuencia	Frecuencia	Frecuencia
	Valor 2	Frecuencia	Frecuencia	Frecuencia
	Valor 3	Frecuencia	Frecuencia	Frecuencia

Ejemplo

	Fumadores	No fumadores	Totales
Hombres	120	60	180
Mujeres	50	70	120
Totales	170	130	300

Frecuencias observadas y esperadas

- Frecuencia esperada: Estas son las frecuencisa que deberían darse si las variables fueran independientes.
- Frecuencia observada: Estas son las frecuencias reales que se observa en nuestra data.

Ejemplo:

```
> tabla2 #valores observado

                       Baja Media baja Media alta alta
  Menos que secundaria   50         38         30   15
  Secundaria completa   121        112         95   83
  Mas que secundaria    195        280        261  172
> chisq.test(tabla2)$expected #valores esperados

                          Baja Media baja Media alta      alta
  Menos que secundaria  33.52479   39.38705   35.35675  24.73140
  Secundaria completa  103.59917  121.71488  109.26033  76.42562
  Mas que secundaria   228.87603  268.89807  241.38292 168.84298
```

Prueba Chi2

Chi2 es una prueba para estimar el grado de asociación entre variables categóricas: “Nominal - Nominal”, “Nominal - Ordinal” y “Ordinal - Ordinal”. Esto significa que una parte de la variabilidad de una variable puede ser explicada por otra variable.

Supuestos:

Para analizar asociación se requiere que el número de observaciones esperadas en cada celda de la tabla de contingencia debe ser suficientemente grande.

Para fines de este curso, al menos cada celda de la TC de frecuencias esperadas debe ser de 5.

Ten en cuenta que si estas condiciones no se cumplen, entonces la prueba podría no funcionar adecuadamente y los resultados de la prueba podrían no ser válidos. Si es que encuentran que no se cumple este supuesto: Repórtalo!

Hipótesis:

- Hipótesis nula (H0): Las variables son estadísticamente independientes (No hay asociación).
- Hipótesis alternativa (H1): Las variables son estadísticamente dependientes (Sí hay asociación).

Ejercicios

Utilizaremos data sobre elecciones presidenciales en Estados Unidos, de hace 20 años. La base de datos continene información sobre las preferencias electorales antes y después de las elecciones, así como informacion de las preferencias políticas, situación económica, religión, y participación política de los encuestados.

Cargamos la data:

```
library(rio)
eda=import("eda.sav")
```

Ejercicio 1. Relación entre sexo y situación económica respecto del año pasado**

PASO 0: Revisamos la estructura de las variables que nos interesan:

Variable sexo: nominal

```
str(eda$sexo)
```

Les damos el formato adecuado:

```
eda$sexo = factor(eda$sexo, labels = c("Hombre","Mujer"))
table(eda$sexo)
```

```
##
## Hombre  Mujer
##    790   1017
```

Situación económica: ordinal

```
str(eda$su_ecopas)
```

```
##  num [1:1807] 1 1 3 NA NA 2 NA 1 NA NA ...
##  - attr(*, "label")= chr "Su situacion economica el año pasado?"
##  - attr(*, "format.spss")= chr "F1.0"
##  - attr(*, "labels")= Named num [1:3] 1 2 3
##   ..- attr(*, "names")= chr [1:3] "1. Mejor" "2. Igual" "3. Peor"
```

```r
eda$su_ecopas= factor(eda$su_ecopas,
                      levels = c(1:3),
                      labels = c("Mejor","Igual","Peor"),
                      ordered = T)
table(eda$su_ecopas)
```

```
##
## Mejor Igual  Peor
##   311   482   130
```

```r
prop.table(table(eda$su_ecopas))*100
```

```
##
##    Mejor    Igual     Peor
## 33.69447 52.22102 14.08451
```

PASO 1: Tabla de contigencias

Los valores observados son los valores de nuestra tabla tal como la tenemos en nuestra base

```r
tabla1.1 = table(eda$su_ecopas, eda$sexo)
tabla1.1 #tabla simple
```

```
##
##         Hombre Mujer
##   Mejor    164   147
##   Igual    196   286
##   Peor      48    82
```

Creamos porcentajes por columna:

```r
library(tidyverse)
tabla1.2 = tabla1.1 %>%
           prop.table(2) %>%  # porcentaje por columna
           round(3)
tabla1.2
```

```
##
##         Hombre Mujer
##   Mejor  0.402 0.285
##   Igual  0.480 0.555
##   Peor   0.118 0.159
```

> Existe diferencia con lo que vemos a nivel de cada subgrupo (hombre y mujer) respecto a lo que habíamos visto a nivel de toda la muestra?

PASO 2: Diagrama de barras apiladas

Preparamos la data para graficar:

```r
toPlot1 = as.data.frame(tabla1.2)
names(toPlot1) = c("Categoria", "Sexo", "Porcentaje")
toPlot1
```

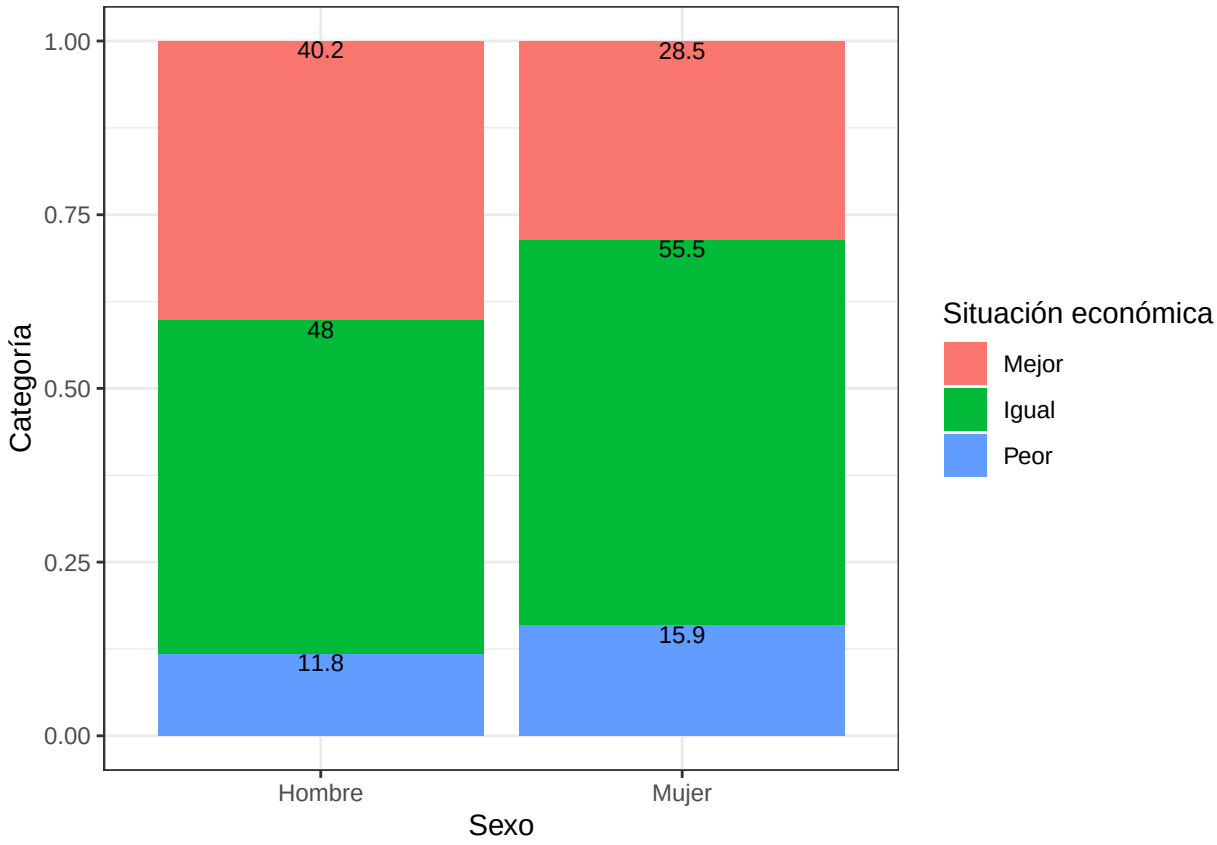
```
##   Categoria   Sexo Porcentaje
## 1     Mejor Hombre      0.402
## 2     Igual Hombre      0.480
## 3      Peor Hombre      0.118
## 4     Mejor  Mujer      0.285
## 5     Igual  Mujer      0.555
## 6      Peor  Mujer      0.159
```

Generamos el gráfico y lo solicitamos:

```r
library(ggplot2)

barras_apiladas<-toPlot1 |>
  ggplot()+
  aes(x=Sexo, y=Porcentaje, fill=Categoria) +
  geom_bar(position="stack", stat="identity")+
  geom_text(aes(label=Porcentaje*100),
            position = position_stack(),
            vjust=1, size = 3)+
  labs(x="Sexo", y="Categoría", fill="Situación económica")+
  theme_bw()

barras_apiladas
```



De forma preliminar, ves diferencias entre la forma cómo se distribuye la variable "Situación Económica" en cada subgrupo (hombre y mujer)?

PASO 3: Prueba Chi cuadrado

- H0: El sexo es estadísticamente independiente de la situación económica respecto del año pasado
- HA: El sexo es estadísticamente dependiente de la situación económica respecto del año pasado

```
chisq.test(tabla1.1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla1.1
## X-squared = 14.416, df = 2, p-value = 0.0007406
```

De acuerdo al p–value obtenido en la prueba de hipótesis de Chi2, al ser menor de 0.05, podemos rechazar la hipótesis nula (Las variables son independientes).

Por lo tanto, concluimos existe dependencia entre las variables escogidas: sexo y situación económica actual.

SUPUESTO

Ten en cuenta que si te piden verificar el supuesto sólo tienes que solicitar la tabla de frecuencias esperadas y ver que efectivamente todas las celdas tienen un número igual o mayor a 5.

```
chisq.test(tabla1.1)$expected
```

```
##
##            Hombre     Mujer
##   Mejor 137.47346 173.52654
##   Igual 213.06176 268.93824
##   Peor   57.46479  72.53521
```

En este caso sí cumple el supuesto!

Ejercicio 2: Relación entre nivel educativo (ordinal) y nivel de confianza en la política (ordinal)

PASO 0: Revisamos la estructura de las variables que nos interesan:

Variable nivel educativo: ordinal

```
str(eda$educ)
```

```
eda$educ = factor(eda$educ,
                  levels = c(1:3),
                  labels = c("Menos que secundaria","Secundaria completa","Mas que secundaria"),
                  ordered = T)
table(eda$educ)
```

```
##
## Menos que secundaria  Secundaria completa   Mas que secundaria
##                  180                  519                 1101
```

Confianza en la política: ordinal

```
eda$confipolR = factor(eda$confipolR,
                       levels = c(1:3),
                       labels = c("Baja","Media", "Alta"),
                       ordered = T)
table(eda$confipolR)
```

```
##
##  Baja Media  Alta
##   367   817   270
```

PASO 1: Tabla de contigencia

Los valores observados son los valores de nuestra tabla tal como la tenemos en nuestra base

```
tabla2.1 = table(eda$confipolR, eda$educ)
tabla2.1
```

```
##
##         Menos que secundaria Secundaria completa Mas que secundaria
##   Baja                    50                 121                195
##   Media                   68                 207                541
##   Alta                    15                  83                172
```

Creamos porcentajes por columna:

```
library(tidyverse)
tabla2.2 = tabla2.1 %>%
  prop.table(2) %>%  # porcentaje por columna
  round(3)
tabla2.2
```

```
##
##         Menos que secundaria Secundaria completa Mas que secundaria
##   Baja                 0.376                0.294              0.215
##   Media                0.511                0.504              0.596
##   Alta                 0.113                0.202              0.189
```

Creamos porcentajes por fila:

PASO 2: Diagrama de barras apiladas

Preparamos la data para graficar:

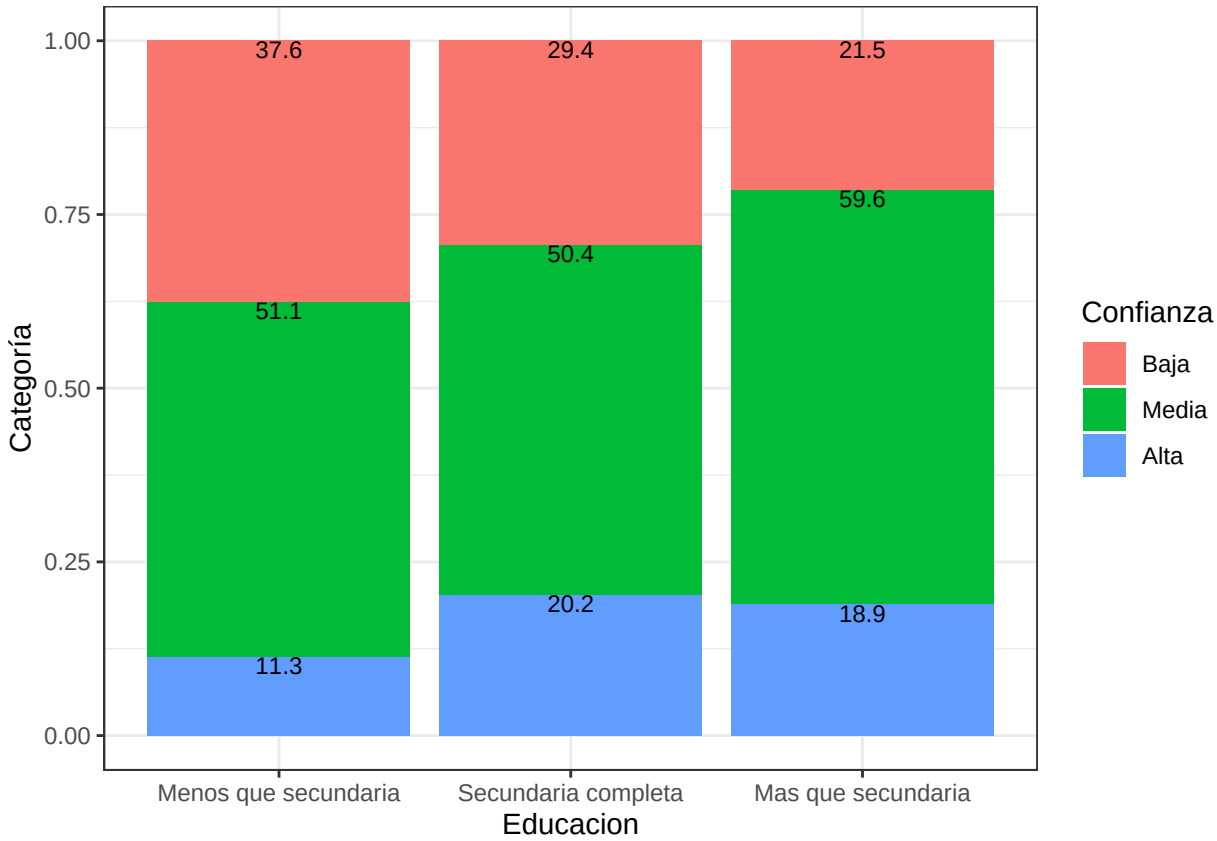
```
toPlot2 = as.data.frame(tabla2.2)
names(toPlot2) = c("Categoria", "Educacion", "Porcentaje")
toPlot2
```

```
##   Categoria            Educacion Porcentaje
## 1      Baja Menos que secundaria      0.376
## 2     Media Menos que secundaria      0.511
## 3      Alta Menos que secundaria      0.113
## 4      Baja  Secundaria completa      0.294
## 5     Media  Secundaria completa      0.504
## 6      Alta  Secundaria completa      0.202
## 7      Baja   Mas que secundaria      0.215
## 8     Media   Mas que secundaria      0.596
## 9      Alta   Mas que secundaria      0.189
```

Generamos el gráfico y lo solicitamos:

```
barras_apiladas2<-toPlot2 |>
  ggplot()+
  aes(x=Educacion, y=Porcentaje, fill=Categoria) +
  geom_bar(position="stack", stat="identity")+
  geom_text(aes(label=Porcentaje*100),
            position = position_stack(),
            vjust=1, size = 3)+
  labs(x="Educacion", y="Categoría", fill="Confianza")+
  theme_bw()

barras_apiladas2
```



PASO 3: Prueba Chi cuadrado

- H0: El nivel educativo es estadísticamente independiente de la confianza en la política
- HA: El nivel educativo es estadísticamente dependiente de la confianza en la política

```
chisq.test(tabla2.1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla2.1
## X-squared = 25.433, df = 4, p-value = 4.116e-05
```

De acuerdo al p–value obtenido en la prueba de hipótesis de Chi2, al ser menor de 0.05, podemos rechazar la hipótesis nula (Las variables son independientes).

Por lo tanto, concluimos existe dependencia entre las variables escogidas: confianza en la política y nivel educativo.

SUPUESTO

Ten en cuenta que si te piden verificar el supuesto sólo tienes que solicitar la tabla de frecuencias esperadas y ver que efectivamente todas las celdas tienen un número igual o mayor a 5.

```
chisq.test(tabla2.1)$expected
```

```
##
##         Menos que secundaria Secundaria completa Mas que secundaria
##   Baja              33.52479           103.59917            228.876
##   Media             74.74380           230.97521            510.281
##   Alta              24.73140            76.42562            168.843
```

En este caso también cumple el supuesto!