Práctica 12. Comparación de proporciones (II)

Jesús Martín Fernández

Contenidos

1. Introducción

Las comparaciones muestrales de proporciones son herramientas estadísticas esenciales para evaluar las diferencias significativas en la frecuencia de ocurrencia de eventos entre grupos. Estas comparaciones pueden realizarse en contextos de datos individuales y datos emparejados. En el caso de los datos individuales, donde se analizan dos o más grupos independientes, se emplean pruebas como el Chi-cuadrado de Pearson. Para llevar a cabo estos contrastes, es fundamental construir una tabla que resuma las frecuencias observadas y esperadas, la cual puede elaborarse manualmente o generarse a partir de las variables de un dataframe. En contraste, cuando se trabaja con datos emparejados, se examinan las proporciones de eventos en situaciones donde los sujetos están relacionados, como en estudios que analizan resultados antes y después de una intervención. En este contexto, se utiliza el Test de McNemar para identificar cambios en la proporción de respuestas entre los pares de sujetos. Sin embargo, cuando se dispone de una variable con más de dos categorías en datos emparejados, el Test de Cochran se convierte en una herramienta adecuada, permitiendo evaluar la simetría y las diferencias en las proporciones a lo largo de múltiples niveles. Adicionalmente, se puede aplicar el Chi-cuadrado de tendencia lineal para explorar si la proporción de un carácter presenta una tendencia lineal en función de una variable ordinal, lo que es crucial para obtener conclusiones significativas en estudios que involucran categorías complejas

1.1. Preparando la base de datos

En primer lugar, vamos a seleccionar, como siempre, nuestro directorio de trabajo y a obtener la dase de datos de trabajo, en este caso df_iam3, que puedes obtener de la Carpeta de la Práctica 12 en el Aula Virtual:

```
#setwd()
#getwd ()

df_iam3<-read.csv ("df_iam3.csv")
```

Recuerda que , en esta base se recogen una serie de características de 984 sujetos. De todos se incluyeron características sociodemográficas (`fech_nac`, `sex` y `clas_soc`), clínicas (`hta` , `DM`, `colesterol` , `salud`) y el hábito tabáquico (`fum`). En un momento en el tiempo se recogió qué sujetos habían tenido un evento tipo infarto de miocardio `iam`). Posteriormente se siguió a los sujetos hasta otro punto en el tiempo. Tras ese punto t, sólo se recogió si el sujeto seguía fumando tras el primer infarto (`fum_p`) , la cifra de colesterol (`colesterol_p`), la percepción de salud posterior al infarto (`salud_p`) y la ocurrencia de un nuevo reinfarto (`iam2`).

Vamos a obtener la variable edad (en años cumplidos), suponiendo que la fecha final de seguimiento es el 31/12/2023 a partir de la variable `fech_nac`

```
#fech_nac viene definida como character, cambiamos a formato fecha
df_iam3$fech_nac <- as.Date(df_iam3$fech_nac)
fecha_fin <- as.Date("2023-12-31")
df_iam3$edad <- (fecha_fin-df_iam3$fech_nac)/365.25
df_iam3$edad <- as.numeric (round (df_iam3$edad,0))
```

Seguidamente vamos a convertir en factor las variables `sex` (etiquetas: 0=“mujer; 1=”Varón”), `hta` , `DM` , `fum` , `fum_p` , `iam1` e `iam2` , con las etiquetas (0=“No”, 1=“Sí”) y la variable `clas_soc` con las etiquetas (0= “baja”; 1=” Media”, 2= Alta”), y `salud` y `salud_p` con las etiquetas (0=“Muy mala”; 1=“Mala”; 2= “Regular”; 3= “Buena”; 4= “Muy buena”)

```
df_iam3$sex <- factor(df_iam3$sex, levels = c(0, 1),
                      labels = c("Mujer", "Varón"))
df_iam3$hta <- factor(df_iam3$hta, levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$DM <- factor(df_iam3$DM, levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$fum <- factor(df_iam3$fum, levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$fum_p <- factor(df_iam3$fum_p,
                        levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$salud <- factor(df_iam3$salud,
                        levels = c(0,1, 2,3,4),
```

```
                        labels = c("Muy mala", "Mala",
                                   "Regular", "Buena", "Muy buena"))
df_iam3$salud_p <- factor(df_iam3$salud_p,
                        levels = c(0,1, 2,3,4),
                        labels = c("Muy mala", "Mala",
                                   "Regular", "Buena", "Muy buena"))
df_iam3$iam1 <- factor(df_iam3$iam1, levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$iam2 <- factor(df_iam3$iam2, levels = c(0, 1), labels = c("No", "Sí"))
df_iam3$clas_soc <- factor(df_iam3$clas_soc,
                           levels = c(0, 1, 2),
                           labels = c("Baja", "Media", "Alta"))
```

Comprueba que ha funcionado todo bien

```
head (df_iam3)
```

```
    fech_nac   sex   alt peso      imc hta fum DM colesterol   salud clas_soc
1 1982-08-06 Mujer 159.9 56.9 22.25437  No  No Sí        255   Buena    Media
2 1982-09-23 Varón 165.7 72.7 26.47826  Sí  No No        192    Mala    Media
3 1939-03-04 Mujer 156.3 51.3 20.99904  No  No No        188   Buena     Baja
4 1936-01-15 Varón 176.6 91.6 29.37068  Sí  No No        174 Regular     Baja
5 1940-03-23 Varón 169.2 89.2 31.15761  No  No No        140   Buena     Baja
6 1962-11-10 Mujer 151.7 45.0 19.55426  Sí  No No        140   Buena    Media
  iam1 fum_p colesterol_p salud_p iam2 edad
1   No    No          255   Buena   No   41
2   No    No          195    Mala   No   41
3   No    No          189   Buena   No   85
4   No    No          182 Regular   No   88
5   No    No          140   Buena   No   84
6   No    No          144   Buena   No   61
```

2. Comparación de proporciones para datos agrupados

Vamos a comparar la proporción de fumadores inicial (`fum`) con la proporción de fumadores tras el primer infarto (`fum_p`).

La prueba de McNemar es un test no paramétrico utilizado para comparar proporciones en datos emparejados o repetidos, como antes y después de un tratamiento, o en dos condiciones diferentes para los mismos sujetos. Esta prueba se aplica a tablas de contingencia 2x2, donde los cambios en las respuestas de una categoría a otra son de interés.

```r
# Crear la tabla de contingencia entre fum (antes) y fum_p (después)

tabla_fum <- table(df_iam3$fum, df_iam3$fum_p)


tabla_fum
```

```
     No  Sí
  No 742   0
  Sí  65 177
```

```r
mcnemar_test <- mcnemar.test(tabla_fum)

mcnemar_test
```

```
    McNemar's Chi-squared test with continuity correction

data:  tabla_fum
McNemar's chi-squared = 63.015, df = 1, p-value = 2.051e-15
```

3. Volviendo sobre el estudio de las diferencias de proporciones

Recordamos aquí que la comparación de proporciones y el intervalo de confianza (IC) para la diferencia de proporciones son enfoques complementarios en la estadística inferencial.

La comparación de proporciones, utilizando pruebas como Chi-cuadrado o Fisher, determina si hay una diferencia significativa entre las proporciones de diferentes grupos, asumiendo inicialmente que no hay diferencias. El valor p de estas pruebas indica la probabilidad de que la diferencia observada se deba al azar, pero no muestra cuán grande o relevante es esa diferencia. En cambio, el intervalo de confianza (IC) para la diferencia de proporciones proporciona un rango probable para la verdadera diferencia con un nivel de confianza determinado, como el 95%. El IC no solo indica si hay una diferencia significativa (si no incluye el 0), sino que también ayuda a valorar la importancia práctica de esa diferencia para la toma de decisiones.

Vamos a comparar las proporciones de infartos `iam1`, en personas con/sin `DM`

```r
tabla_DM_iam <- table(df_iam3$DM, df_iam3$iam1)

# Ver la tabla
print(tabla_DM_iam)
```

```
     No  Sí
  No 733  99
  Sí 126  26
```

```r
#Ver proporciones

proporciones<- prop.table (tabla_DM_iam,1)
proporciones
```

```
           No        Sí
  No 0.8810096 0.1189904
  Sí 0.8289474 0.1710526
```

```r
# Extraer los valores de la tabla de contingencia
DM_con_iam <- tabla_DM_iam["Sí", "Sí"]  # DM con iam
DM_sin_iam <- tabla_DM_iam["Sí", "No"]  # DM sin iam

total_con_iam <- sum(tabla_DM_iam[, "Sí"])  # Total de pacientes con iam
total_sin_iam <- sum(tabla_DM_iam[, "No"])  # Total de pacientes sin iam

# Aplicar el test de proporciones
resultado <- prop.test(x = c(DM_con_iam, DM_sin_iam),
n = c(total_con_iam, total_sin_iam),
                       correct = FALSE)
# Se puede añadir correct=TRUE para corrección de continuidad

# Ver el resultado
print(resultado)
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(DM_con_iam, DM_sin_iam) out of c(total_con_iam, total_sin_iam)
X-squared = 3.1413, df = 1, p-value = 0.07633
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01366458  0.13630020
sample estimates:
   prop 1    prop 2
0.2080000 0.1466822
```

El valor de p 0.07633 nos lleva, bajo la convención habitual, a no rechazar la hipótesis nula de la deiferencia de medias. Pero la aproximación del IC es mucho más informativa. Sabemos que la diferencia de medias vale 0.208 - 0.1466822= 0.0613178, con un IC del 95% de -0.01366458 a 0.13630020. Aunque no se puede descartar que la diferencia de medias sea nula, el investigador puede pensar en una asociación entre DM e IAM.

4. Comparación de proporciones, tablas de n x m

4.1 Datos independientes

Puede darse el caso de que una de las dos variables que estudiamos tenga más de 2 categorías.

Vamos a ver si las proporciones de hipertensos son diferentes en las diferentes categorías de la variable `clas_soc`

```
# Crear la tabla de contingencia
tabla_htaclass <- table(df_iam3$hta, df_iam3$clas_soc)


tabla_htaclass
```

```
    Baja Media Alta
  No  230   264   50
  Sí  180   216   44
```

```
# Realizar la prueba Chi-Cuadrado
chi_test <- chisq.test(tabla_htaclass)

# Mostrar resultados de la prueba
print(chi_test)
```

```
    Pearson's Chi-squared test

data:  tabla_htaclass
X-squared = 0.29193, df = 2, p-value = 0.8642
```

```r
print(chi_test$expected)
```

```
         Baja    Media     Alta
  No 226.6667 265.3659 51.96748
  Sí 183.3333 214.6341 42.03252
```

No hay diferencias en las proporciones de HT por clase social y no hay ninguna frecuencia esperada menos que 5, podríamos hacer el test de Chi cuadrado sin la corrección de Yates

```r
chi_test <- chisq.test(tabla_htaclass, correct = FALSE)
chi_test
```

```
    Pearson's Chi-squared test

data:  tabla_htaclass
X-squared = 0.29193, df = 2, p-value = 0.8642
```

El resultado da exactamente igual, con la n grande la corrección de Yates apenas tiene impacto.

4.2 Datos apareados

Vamos a utilizar los datos de percepción del estado de salud antes y después del primer infarto para ver si hay diferencias (`salud` y `salud_p`)

Se tratan de datos apareados con 5 niveles en cada variable (ordenados)

Utilizaremos la prueba de Friedman, el equivalente no paramétrico de un diseño de medidas repetidas de una muestra o un análisis de varianza de dos factores con una observación por casilla. Friedman contrasta la hipótesis nula de que las variables relacionadas procedan de la misma población. Para cada caso, a las variables se les asignan los rangos 1 a k. El estadístico de contraste se basa en estos rangos.

```r
tabla_friedman <- table(df_iam3$salud, df_iam3$salud_p)
tabla_friedman
```

```
           Muy mala Mala Regular Buena Muy buena
  Muy mala       75    0       0     0         0
  Mala           16  147       0     0         0
  Regular         0   32     239     0         0
  Buena           0    0      24   336         0
  Muy buena       0    0       0     9       106
```

```r
#Friendman necesita un formato largo

# Crear una matriz de salud antes y después
matriz_salud <- as.matrix(cbind(df_iam3$salud, df_iam3$salud_p))

# Aplicar la Prueba de Friedman
friedman_result <- friedman.test(matriz_salud)
friedman_result
```

```
    Friedman rank sum test

data:  matriz_salud
Friedman chi-squared = 81, df = 1, p-value < 2.2e-16
```

Hay diferencias entre los grupos no achacables al azar.

5. Test de tendencia lineal

Los tests de tendencia lineal son herramientas estadísticas utilizadas para evaluar si existe una relación sistemática y creciente o decreciente entre una variable ordinal y una variable categórica, permitiendo identificar patrones de cambio en las proporciones a medida que varía la categoría de la variable independiente.

Podríamos preguntarnos si hay menos proporción de fumadores a medida que aumenta el nivel social. Vamos a intentar resolver esta cuestión con un test de tendencia lineal.

Primero generamos la tabla

```r
tabla_contingencia <- table(df_iam3$fum, df_iam3$clas_soc)

# Mostrar la tabla de contingencia
print(tabla_contingencia)
```

```
    Baja Media Alta
  No  307   356   79
  Sí  103   124   15
```

```r
# Calcular las proporciones de la tabla de contingencia
proporciones <- prop.table(tabla_contingencia, margin = 2)
proporciones
```

```
         Baja     Media      Alta
  No 0.7487805 0.7416667 0.8404255
  Sí 0.2512195 0.2583333 0.1595745
```

Y ahora hay que disponer los datos para poder usar la función `prop.trend.test`

```r
# Definir el número de fumadores (x)) y el tamaño de cada grupo (n)
x <- c(103, 124, 15)
n <- c(307 + 103, 356 + 124, 79 + 15)

# Realizar el test de tendencia
tendencia_resultado <- prop.trend.test(x, n, score = seq_along(x))


print(tendencia_resultado)
```

```
    Chi-squared Test for Trend in Proportions

data:  x out of n ,
 using scores: 1 2 3
X-squared = 1.4169, df = 1, p-value = 0.2339
```

Aunque parece haber cierta tendencia decreciente, no se puede rechazar la hipótesis nula , que no reconoce ninguna tendencia lineal.