Práctica 11.Comparación de proporciones (I)

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 y el Test de Fisher. Para llevar a cabo estos contrastes, es fundamental construir una tabla que resuma las frecuencias observadas y esperadas. Esta tabla puede elaborarse manualmente, utilizando una matriz, o generarse a partir de las variables de un dataframe mediante las funciones aprendidas en la sección de descriptiva.

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. 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.

1.1. Preparando nuestra 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 10 en el Aula Virtual:

```
#setwd()
getwd ()
```

```
[1] "~/Práctica11"
```

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

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`). Posteriormenete 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. Contraste frente a un valor de referencia.

En esta base de datos compararemos si las proporciones de hipertensos (variable `hta`) hacen pensar que la muestra es representativa de la población en la cuál la población de hipertensión es del 42%.

```r
# Primero obtenemos proporción de hipertensos en la muestra
tabla_hta <- table(df_iam3$hta)

# Calcular el número de hipertensos y el total
n <- length(df_iam3$hta)  # Tamaño total de la muestra
n_hta <- sum(df_iam3$hta == "Sí", na.rm = TRUE)
# Número de hipertensos en la muestra

# Proporción esperada en la población
p_c <- 0.42

# Realizar la prueba de proporción
test <- prop.test(n_hta, n, p = p_c, conf.level = 0.95)

# Mostrar el resultado
test
```

```
    1-sample proportions test with continuity correction

data:  n_hta out of n, null probability p_c
X-squared = 2.8681, df = 1, p-value = 0.09035
alternative hypothesis: true p is not equal to 0.42
95 percent confidence interval:
 0.4158513 0.4788754
sample estimates:
        p
0.4471545
```

No podemos descartar que la muestra provenga de la población con un porcentaje de hipertensos del 42%

3. Comparación de proporciones muestrales, tablas de 2x2

3.1 Test de Chi cuadrado

En esta base de datos compararemos las proporciones de hipertensos (variable `hta`) en aquellos que sufrieron el primer infarto (variable `iam1`) y los que no con una Chi cuadrado. Este test compara las frecuencias observadas con las esperadas, que son los valores que esperaríamos ver en cada celda de una tabla de contingencia si no hubiera asociación entre las variables (es

decir, si las variables fueran independientes). Se calculan multiplicando el total de la fila por el total de la columna , dividiéndole por el total general.

```
# Tabla cruzada de iam1 y hta
tabla_hta_iam1 <- table(df_iam3$hta, df_iam3$iam1)

# Ver la tabla de frecuencias
tabla_hta_iam1
```

```
     No  Sí
  No 468  76
  Sí 391  49
```

```
# Calcular las proporciones de hta dentro de cada grupo de iam1
prop_hta_iam1 <- prop.table(tabla_hta_iam1, margin = 1)

# Ver las proporciones
prop_hta_iam1
```

```
           No        Sí
  No 0.8602941 0.1397059
  Sí 0.8886364 0.1113636
```

```
# Test chi-cuadrado de independencia
chi_test_hta_iam1 <- chisq.test(tabla_hta_iam1)

# Ver el resultado del test
chi_test_hta_iam1
```

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

data:  tabla_hta_iam1
X-squared = 1.5157, df = 1, p-value = 0.2183
```

Como se ve, nos ofrece el resultado de la Chi cuadrado con la correción de Yates. Dicha corrección es una modificación del test Chi cuadrado aplicada a tablas 2x2 para ajustar la sobrestimación de significancia cuando las frecuencias esperadas son bajas, típicamente inferiores

a 5. Su función es reducir la diferencia entre las frecuencias observadas y esperadas, mejorando la precisión del test en muestras pequeñas, aunque puede resultar demasiado conservadora. Para desactivarla en R, se puede usar el argumento `correct = FALSE` en la función `chisq.test()`, lo cual es útil cuando las frecuencias no son tan pequeñas y la corrección se considera innecesaria.

Primero vamos a aprender a obtener las frecuencias esperadas

```
#Obtener frecuencias esperadas

chi_test_hta_iam1$expected
```

```
           No       Sí
  No 474.8943 69.10569
  Sí 384.1057 55.89431
```

```
chi_test_hta_iam1_sin_yates <- chisq.test(tabla_hta_iam1, correct = FALSE)

chi_test_hta_iam1_sin_yates
```

```
    Pearson's Chi-squared test

data:  tabla_hta_iam1
X-squared = 1.762, df = 1, p-value = 0.1844
```

Vemos que las frecuencias esperadas son grandes, por lo que no haría falta la corrección de Yates. El p-value (0,1844) nos lleva a la misma conclusión que el obtenido con la corrección de Yates (0,2183). El significado de ese valor, es que no se puede rechazar la hipótesis nula de igualdad de proporciones.

Debes saber que, teniendo los datos de la tabla 2x2 se pueden calcular directamente el p-value para el valor de Chi cuadrado resultante

```
# Crear la matriz con los valores de la tabla de contingencia
m_hta_iam1 <- matrix(c(468, 76, 391, 49), nrow = 2, byrow = TRUE,
          dimnames = list(c("hta = No", "hta = Sí"),
                          c("iam1 = No", "iam1 = Sí")))


m_hta_iam1
```

```
          iam1 = No iam1 = Sí
hta = No        468        76
hta = Sí        391        49
```

```r
# Aplicar la prueba chi-cuadrado a la matriz
chi_test_hta_iam1 <- chisq.test(m_hta_iam1)

# Ver el resultado del test chi-cuadrado
chi_test_hta_iam1
```

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

data:  m_hta_iam1
X-squared = 1.5157, df = 1, p-value = 0.2183
```

El resultado es el mismo y su interpretación, al ser un p-value alto, es que no se puede rechazar la hipótesis nula de que las proporciones sean iguales.

3.2 Test de Fisher

El Test exacto de Fisher es preferible cuando hay pequeñas frecuencias en las celdas de una tabla de contingencia (frecuencias esperadas menores de 5) o cuando se desea una mayor precisión en tablas 2x2. Aunque en este caso no sería imprescindible el test de Fisher, vamos a repasara las frecuencias esperadas y a utilizarlo en el ejemplo anterior.

```r
#Las frecuencias esperadas solo se pueden calcular
directamente con el test Chi cuadrado o hacerse manualmente
# Calcular frecuencias esperadas manualmente
n <- sum(m_hta_iam1)  # Total general
a <- m_hta_iam1[1, 1]  # hta = No, iam1 = No
b <- m_hta_iam1[1, 2]  # hta = No, iam1 = Sí
c <- m_hta_iam1[2, 1]  # hta = Sí, iam1 = No
d <- m_hta_iam1[2, 2]  # hta1= Sí, iam1 = Sí

# Calcular frecuencias esperadas
E_11 <- (a + b) * (a + c) / n  # Frecuencia esperada para hta = No, iam1 = No
E_12 <- (a + b) * (b + d) / n  # Frecuencia esperada para hta = No, iam1 = Sí
E_21 <- (c + d) * (a + c) / n  # Frecuencia esperada para hta = Sí, iam1 = No
E_22 <- (c + d) * (b + d) / n  # Frecuencia esperada para hta = Sí, iam1 = Sí
```

```r
# Mostrar frecuencias esperadas
frecuencias_esperadas <- matrix(c(E_11, E_12, E_21, E_22), nrow = 2,
                      byrow = TRUE,
                      dimnames = list(c("hta = No", "hta = Sí"),
                      c("iam1 = No", "iam1 = Sí")))
print("Matriz de Frecuencias Esperadas:")
```

```
[1] "Matriz de Frecuencias Esperadas:"
```

```r
print(frecuencias_esperadas)
```

```
          iam1 = No iam1 = Sí
hta = No   474.8943  69.10569
hta = Sí   384.1057  55.89431
```

```r
#Hacemos el test de Fisher
fisher_test_hta_iam1 <- fisher.test(tabla_hta_iam1)
fisher_test_hta_iam1
```

```
    Fisher's Exact Test for Count Data

data:  tabla_hta_iam1
p-value = 0.2107
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5143338 1.1502070
sample estimates:
odds ratio
 0.7719081
```

La hipótesis nula en el test de Fisher es que la Odds Ratio es 1 y no puede descartarse en este caso, como era de esperar.

3.3 Datos apareados, test de McNemar

Ahora 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.

```
# 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
```

```
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
```

El p-value tan pequeño hace poco probable que las dos proporciones sean iguales y nos habilita para tratarlas como diferentes.

Como hay un valor de una casilla =0 , puede ser es apropiada la corrección de Yates pero para estar seguros deberíamos calcular los valores esperados

```
# Frecuencias observadas
cat("Frecuencias observadas:\n")
```

```
Frecuencias observadas:
```

```
print(tabla_fum)
```

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

```r
# Calcular frecuencias esperadas manualmente
n <- sum(tabla_fum)  # Total general
a <- tabla_fum[1, 1]  # Fumador, Fumador
b <- tabla_fum[1, 2]  # Fumador, No Fumador
c <- tabla_fum[2, 1]  # No Fumador, Fumador
d <- tabla_fum[2, 2]  # No Fumador, No Fumador

# Calcular frecuencias esperadas
E_11 <- (a + b) * (a + c) / n
E_12 <- (a + b) * (b + d) / n
E_21 <- (c + d) * (a + c) / n
E_22 <- (c + d) * (b + d) / n

# Mostrar frecuencias esperadas
cat("\nFrecuencias esperadas:\n")
```

```
Frecuencias esperadas:
```

```r
cat("Frecuencia esperada (Fumador, Fumador):", E_11, "\n")
```

```
Frecuencia esperada (Fumador, Fumador): 608.5305
```

```r
cat("Frecuencia esperada (Fumador, No Fumador):", E_12, "\n")
```

```
Frecuencia esperada (Fumador, No Fumador): 133.4695
```

```r
cat("Frecuencia esperada (No Fumador, Fumador):", E_21, "\n")
```

```
Frecuencia esperada (No Fumador, Fumador): 198.4695
```

```r
cat("Frecuencia esperada (No Fumador, No Fumador):", E_22, "\n")
```

```
Frecuencia esperada (No Fumador, No Fumador): 43.53049
```

Pues no, no hay ninguna frecuencia esperada menor que 5, así que podíamos quitar la correción de yates

```
mcnemar_test_sin_Yates <- mcnemar.test(tabla_fum, correct= FALSE)
mcnemar_test_sin_Yates
```

```
    McNemar's Chi-squared test

data:  tabla_fum
McNemar's chi-squared = 65, df = 1, p-value = 7.49e-16
```

El resultado no es muy diferente del anterior

4. Intervalos de confianza para la diferencia de proporciones

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, mediante pruebas como el test de Chi-cuadrado o el test de Fisher, evalúa si existe una diferencia significativa entre las proporciones de dos o más grupos, basándose en la hipótesis nula de que no hay diferencia entre ellas. Sin embargo, el valor p obtenido en estas pruebas no refleja la magnitud o relevancia de la diferencia, sino lo improbable que sería observar esa diferencia debido al azar si la hipótesis nula fuera verdadera. Por otro lado, el IC para la diferencia de proporciones ofrece un rango de valores donde probablemente se encuentra la verdadera diferencia, con un cierto nivel de confianza (como el 95%). El IC no solo indica si la diferencia es estadísticamente significativa (si no incluye el valor 0), sino que aporta un significado intrínseco, permitiendo evaluar la importancia clínica o práctica de la diferencia, proporcionando una idea más clara sobre su relevancia en la toma de decisiones.

El IC de la diferencia de proporciones se calcula así

$IC = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \cdot SE(\hat{p}_1 - \hat{p}_2)$

donde

$\hat{p}_1$: Proporción observada en el primer grupo.
$\hat{p}_2$: Proporción observada en el segundo grupo.
n_1 : Tamaño de la muestra del primer grupo.
n_2 : Tamaño de la muestra del segundo grupo.
$Z_{\alpha/2}$: Valor crítico de la distribución normal estándar.
$SE(\hat{p}_1 - \hat{p}_2)$: Error estándar de la diferencia de proporciones.

$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Vamos a calcular las proporciones de fumadores (`fum`) en pacientes con y sin infarto (`iam1`)y luego la diferencia de proporciones , interpretando su significado:

```
tabla_fum_iam <- table(df_iam3$fum, df_iam3$iam1)

# Ver la tabla
print(tabla_fum_iam)
```

```
     No  Sí
  No 674  68
  Sí 185  57
```

```
# Extraer los valores de la tabla de contingencia
fum_con_iam <- tabla_fum_iam["Sí", "Sí"]  # Fumadores con iam
fum_sin_iam <- tabla_fum_iam["Sí", "No"]  # Fumadores sin iam

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

# Aplicar el test de proporciones
resultado <- prop.test(x = c(fum_con_iam, fum_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(fum_con_iam, fum_sin_iam) out of c(total_con_iam, total_sin_iam)
X-squared = 34.071, df = 1, p-value = 5.313e-09
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1490958 0.3321708
sample estimates:
   prop 1    prop 2
0.4560000 0.2153667
```

El valor de p es muy pequeño luego es poco probable que la diferencia se deba al azar. Pero ¿Cuál es esa diferencia?

```
0.4560000 - 0.2153667= 0.2402333
```

¿? Y su IC del 95%

```
0.1490958 a 0.3321708
```

Esto significa que la diferencia de proporciones, con una confianza del 95% será de, al menos 14,91%.

Si esta misma diferencia la encontrásemos en unas muestras mucho más pequeñas , podría ocurrir que la p no fuese significativa. Mira el ejemplo de más abajo

```r
# Crear la matriz original
m_original <- matrix(c(674, 68, 185, 57), nrow = 2, byrow = TRUE,
                  dimnames = list(c("No fumador", "Fumador"),
                  c("iam1 = No", "iam1 = Sí")))

# Dividir la matriz por 10 y redondear sin decimales
m_dividida_redondeada <- round(m_original / 10)

# Ver la matriz redondeada
print(m_dividida_redondeada)
```

```
           iam1 = No iam1 = Sí
No fumador        67         7
Fumador           18         6
```

```r
fum_con_iam <- m_dividida_redondeada["Fumador", "iam1 = Sí"]
# No fumadores con iam
fum_sin_iam <- m_dividida_redondeada["Fumador", "iam1 = No"]
# No fumadores sin iam

total_con_iam <- sum(m_dividida_redondeada[, "iam1 = Sí"])
# Total de pacientes con iam
total_sin_iam <- sum(m_dividida_redondeada[, "iam1 = No"])
# Total de pacientes sin iam

# Aplicar el test de proporciones
```

```r
resultado <- prop.test(x = c(fum_con_iam, fum_sin_iam),
                       n = c(total_con_iam, total_sin_iam),
                       correct = FALSE)
```

```
Warning in prop.test(x = c(fum_con_iam, fum_sin_iam), n = c(total_con_iam, :
Chi-squared approximation may be incorrect
```

```r
# sin corrección de continuidad

# Ver el resultado
print(resultado)
```

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

data:  c(fum_con_iam, fum_sin_iam) out of c(total_con_iam, total_sin_iam)
X-squared = 3.804, df = 1, p-value = 0.05113
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03479754  0.53434505
sample estimates:
   prop 1    prop 2
0.4615385 0.2117647
```

En este caso el p-value, aunque pequeño, está por encima del umbral de 0,05 (muy poco por encima, seguiríamos desconfiando de que la diferencia se deba al azar aunque formalmente no rechazemos la hipótesis nula), pero la diferencias de proporciones sería igual a la de antes (casi igual, varían los decimales)

```
0.4615385 - 0.2117647= 0.2497738
```

Sin embargo el nuevo IC de la diferencia, sí incluye el 0

```
-0.03479754 a 0.53434505
```

En definitiva el IC de la diferencia de proporciones, sí nos da una idea de la importancia de la diferencia, que no nos ofrece el valor del p-value