Lectura 8

Inferencia no paramétrica con proporciones

Si eres una persona observadora, habrás notado que el título de este capítulo lleva la frase **no paramétrica** para referirse a inferencias con proporciones, pero ¿qué significa esto?

En el capítulo 5 conocimos las pruebas Z y t de Student. Ambas formulan hipótesis relativas al parámetro μ de una distribución normal (o la diferencia $\mu_1 - \mu_2$ de dos distribuciones normales). Así estas pruebas (y otras que se verán más adelante) hacen una fuerte suposición acerca de la distribución que subyace a las poblaciones estudiadas, lo que permite inferir sobre los parámetros de esas distribuciones. Lo mismo ocurre con las pruebas de Wald y Wilson estudiadas en el capítulo 7, las cuales contrastan hipótesis en torno a un cierto valor para el parámetro p de una población que sigue una distribución binomial (o la diferencia de los parámetros $p_1 - p_2$ de dos de estas poblaciones).

En este capítulo conoceremos algunas pruebas para inferir acerca de proporciones cuyas hipótesis nula y alternativa **no mencionan parámetro alguno**. Es más, **ninguna de ellas hace alguna suposición sobre la distribución de la población** desde donde proviene la muestra analizada. Es por esta razón que a estas pruebas (y a otras que se abordan en capítulos posteriores) se les denomina **no paramétricas** o **libres de distribución**.

Las pruebas no paramétricas nos ofrecen una ventaja evidente: **son menos restrictivas** que las pruebas paramétricas, porque imponen menos supuestos a las poblaciones para poder trabajar con ellas. Asegurar que una población sigue una distribución normal o binomial, por ejemplo, puede ser una tarea difícil y, en la práctica, no es infrecuente encontrarse con conjuntos de datos que no parecen seguir alguna de estas distribuciones. Pero... si las pruebas no paramétricas parecen tan ventajosas, ¿por qué no usarlas siempre? Por dos grandes razones:

- Las pruebas no paramétricas **nos entregan menos información**. Como veremos en este capítulo para el caso de las proporciones, estas pruebas se limitan a trabajar con hipótesis del tipo "las poblaciones muestran las mismas proporciones" versus "las poblaciones muestran proporciones distintas", pero **ninguna indica cuáles serían esas proporciones** en realidad, ni siquiera si es mayor en una o en la otra.
- Cuando sí se cumplen las condiciones para aplicar una prueba paramétrica, las versiones no paramétricas presentan **menor poder estadístico** y, en consecuencia, suelen necesitar muestras de mayor tamaño para detectar diferencias significativas que pudieran existir entre las poblaciones comparadas.

Prueba chi-cuadrado de Pearson

->Esta sirve para inferir con proporciones cuando disponemos de dos variables categóricas y una de ellas es dicotómica.

Condiciones a verificar

1) Las observaciones deben ser independientes entre sí
2) Debe haber a lo menos 5 observaciones esperadas en cada grupo

Conocida también como **Prueba χ^2 de Asociación**, la **prueba chi-cuadrado de Pearson** sirve para inferir con proporciones cuando disponemos de dos variables categóricas y una de ellas es dicotómica (es decir, tiene solo dos niveles). En este caso, podemos registrar las frecuencias observadas para las posibles combinaciones de ambas variables mediante una tabla de contingencia o matriz de confusión, como ya estudiamos en el capítulo 2. En adelante, nos referiremos a cada una de estas combinaciones como un grupo.

Debemos verificar algunas condiciones antes de poder usar la prueba chi-cuadrado:

1. Las observaciones deben ser independientes entre sí.
2. Debe haber a lo menos 5 observaciones esperadas en cada grupo.

La primera de estas condiciones ya la hemos encontrado antes, mientras que explicaremos la segunda a medida que avancemos en el estudio de la prueba chi-cuadrado.

Si bien en esta sección estamos hablando de una una única prueba, que sigue siempre el mismo procedimiento, es común encontrarla como tres pruebas diferentes:

- Prueba χ^2 de homogeneidad.
- Prueba χ^2 de bondad de ajuste
- Prueba χ^2 de independencia.

La diferencia entre ellas es **conceptual** (no matemática) y tiene relación con cómo se miren las variables y las poblaciones involucradas en el problema.

Prueba chi-cuadrado de Homogeneidad

Esta prueba resulta adecuada si queremos determinar si **dos poblaciones** (la variable dicotómica) presentan **las mismas proporciones en los diferentes niveles de una variable categórica**.

Por ejemplo, supongamos que la Sociedad Científica de Computación (SCC) ha realizado una encuesta a 300 programadores con más de 3 años de experiencia de todo el país, escogidos al azar, y les ha preguntado cuál es su lenguaje de programación favorito. La tabla 8.1 muestra las preferencias para cada lenguaje, separadas en programadores (varones) y programadoras (mujeres). ¿Son similares las preferencias de lenguaje de programación entre hombres y mujeres?

Lenguaje	**C**	**Java**	**Python**	**Ruby**	**Otro**	**Total**
Programadores	42	56	51	27	24	200
Programadoras	25	24	27	15	9	100
Total	67	80	78	42	33	300

Tabla 8.1: tabla de frecuencias para el lenguaje de programación favorito de la muestra.

Si fuera cierto que ambas poblaciones tienen las mismas preferencias, esperaríamos encontrar proporciones similares en las muestras, pese a la variabilidad. En consecuencia, necesitamos determinar si las diferencias entre las cantidades observadas y las esperadas son lo suficientemente grandes como para proporcionar evidencia convincente de que las preferencias son disímiles. La tabla 8.2 muestra las frecuencias esperadas para cada lenguaje de programación bajo este supuesto, calculadas mediante la ecuación 8.1, donde:

- n_i: total de observaciones en la fila i.
- n_j: total de observaciones en la columna j.
- n: tamaño de la muestra.

$$E_{i,j} = \frac{n_i \cdot n_j}{n} \tag{8.1}$$

Ahora que ya sabemos cómo determinar la cantidad de observaciones esperadas en cada grupo, podemos verificar que, para cada caso, este valor es mayor que 5. Adicionalmente, es razonable suponer la muestra representa menos del 10 % de los programadores del país y sabemos que fue seleccionada de manera aleatoria, por lo que podemos proceder con la prueba χ^2 de homogeneidad.

Lenguaje	C	Java	Python	Ruby	Otro	Total
Programadores	44,7	53,3	52,0	28,0	22,0	200
Programadoras	22,3	26,7	26,0	14,0	11,0	100
Total	67	80	78	42	33	300

Tabla 8.2: frecuencias esperadas si hombres y mujeres tienen las mismas preferencias.

Las hipótesis a contrastar son:

H_0: programadores hombres y mujeres tienen las mismas preferencias en lenguaje de programación favorito (ambas poblaciones muestras las mismas proporciones para cada lenguaje estudiado).

H_A: programadores hombres y mujeres tienen preferencias distintas en lenguajes de programación favorito.

Recordemos que la primera aproximación para construir un estadístico de prueba adecuado está dada por la ecuación 4.5, que reproducimos aquí:

$$Z = \frac{\text{estimador puntual} - \text{valor nulo}}{SE_{\text{estimador puntual}}}$$

Podríamos usar esta fórmula de la diferencia estandarizada para cada uno de los grupos, donde:

- El estimador puntual corresponde a la frecuencia observada para el grupo.
- El valor nulo corresponde a la frecuencia esperada para el grupo.
- El error estándar del estimador puntual es la raíz cuadrada del valor nulo.

Así, para los programadores (varones) en C se tiene:

$$Z_{H\_C} = \frac{42 - 44,7}{\sqrt{44,7}} = -0,404$$

Al repetir el procedimiento para cada grupo, se obtienen los valores Z presentados en la tabla 8.3.

Lenguaje	C	Java	Python	Ruby	Otro
Programadores	-0,404	0,370	-0,139	-0,189	0,426
Programadoras	0,572	-0,523	0,196	0,267	-0,603

Tabla 8.3: valor Z para cada grupo.

Pero necesitamos transformar estos estadísticos por cada grupo en un único estadístico de prueba. Para ello, se considera la suma de sus cuadrados, pues así todos los valores son positivos y las diferencias significativas se incrementan aún más (como en el caso de la varianza). Así, se define el estadístico de prueba χ^2, definido en la ecuación 8.2, donde m y n son, respectivamente, la cantidad de filas y la cantidad de columnas de la tabla de frecuencias, sin considerar los totales (puede ser útil en este punto repasar lo que aprendimos en el capítulo 3 sobre la distribución χ^2).

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\text{cantidad observada} - \text{cantidad esperada})^2}{\text{cantidad esperada}} \tag{8.2}$$

Para el ejemplo tenemos entonces:

$$\begin{aligned}\chi^2 = (-0.404)^2 + (0.370)^2 + (-0.139)^2 + (-0.189)^2 + (0.426)^2 + (0.572)^2 + \\ +(-0.523)^2 + (0.196)^2 + (0.267)^2 + (-0.603)^2 = 1,611\end{aligned}$$

Como estamos sumando $m \cdot n$ valores Z al cuadrado, **el estadístico χ^2 sigue una distribución chi-cuadrado, con $\nu = (m-1)\cdot(n-1)$ grados de libertad**. En el ejemplo, $\nu = (2-1)\cdot(5-1) = 4$.

El valor p para la prueba chi-cuadrado está dado por el área bajo la curva de la distribución chi-cuadrado con valores mayores al obtenido para el estadístico de prueba. En este caso, gracias a la llamada en R `pchisq(1.611, df = 4, lower.tail = FALSE)`, obtenemos que $p = 0,807$. Suponiendo un nivel de significación $\alpha = 0,05$, $p > \alpha$, por lo que se falla al rechazar la hipótesis nula. Es decir, no hay evidencia suficientemente fuerte que sugiera, con 95 % de confianza, que programadores hombres y mujeres prefieran lenguajes de programación distintos.

En R, podemos realizar la prueba chi-cuadrado de homogeneidad como muestra el script 8.1, usando para ello la función `chisq.test(x)`, donde `x` corresponde a la matriz de confusión.

Al ejecutar el script, debemos tener en cuenta que el valor p obtenido usando R es ligeramente diferente debido a los redondeos aplicados en la tabla 8.2 y al resolver la ecuación 8.2.

Script 8.1: prueba chi-cuadrado de homogeneidad.

```r
# Crear tabla de contingencia.
programadores <- c(42, 56, 51, 27, 24)
programadoras <- c(25, 24, 27, 15, 9)

tabla <- as.table(rbind(programadores, programadoras))

dimnames(tabla) <- list(sexo = c("programadores", "programadoras"),
                        lenguajes = c("C", "Java", "Python", "Ruby", "Otro"))

print(tabla)

# Hacer prueba chi-cuadrado de homogeneidad.
prueba <- chisq.test(tabla)
print(prueba)
```

Prueba chi-cuadrado de bondad de ajuste

Esta prueba **permite comprobar si una distribución de frecuencias observada se asemeja a una distribución esperada**. Usualmente se emplea para comprobar si una muestra es representativa de la población (NIST/SEMATECH, 2013, p. 1.3.5.15).

Para entender mejor esta idea, supongamos ahora que una gran empresa de desarrollo de software cuenta con una nómina de 660 programadores, especialistas en diferentes lenguajes de programación. El gerente ha seleccionado un subconjunto de 55 programadores, supuestamente de forma aleatoria, para enviarlos a cursos de perfeccionamiento en sus respectivos lenguajes, pero el sindicato lo ha acusado de "seleccionar estas personas a conveniencia de los intereses mezquinos de la gerencia, impidiendo que el grupo sea representativo a fin de asegurar una mejora en la productividad de toda la empresa". Ante el inminente riesgo de movilizaciones, el gerente necesita demostrar que el grupo seleccionado es una muestra representativa de sus programadores.

La tabla 8.4 muestra la cantidad de especialistas en cada lenguaje, tanto para la nómina de la empresa como para la muestra seleccionada.

Como ya es habitual, comencemos por verificar las condiciones. Puesto que la muestra representa menos del 10 % de la población y fue elegida de manera aleatoria, las observaciones son independientes entre sí.

La segunda condición resulta algo más compleja. Comencemos por calcular la proporción de programadores de la nómina (población) especialista en cada lenguaje. Para el caso de C, tenemos:

Lenguaje	C	Java	Python	Ruby	Otro
Nómina	236	78	204	76	66
Muestra	17	9	14	10	5

Tabla 8.4: frecuencias por lenguaje de programación para la toda la nómina y para la muestra.

$$P_C = \frac{n_C}{n} = \frac{236}{660} = 0,358$$

En consecuencia, esperaríamos la misma proporción de especialistas en C en la muestra, es decir:

$$E_C = P_C \cdot n = 0,358 \cdot 55 = 19,690$$

Repitiendo este proceso para los lenguajes restantes, obtenemos las proporciones para la población y valores esperados para la muestra que se presentan en la tabla 8.5. En ella podemos ver que para cada grupo se esperan más de 5 observaciones, por lo que se verifica la segunda condición.

Lenguaje	C	Java	Python	Ruby	Otro
Proporciones nómina	0,358	0,118	0,309	0,115	0,100
Valores esperados muestra	19,690	6,490	16,995	6,325	5,500

Tabla 8.5: proporciones de la población y valores esperados de la muestra.

En este ejemplo, las hipótesis a contrastar son:

H_0: las proporciones de especialistas en cada lenguaje son las mismas para la nómina y la muestra.
H_A: las proporciones de especialistas en cada lenguaje son diferentes en la nómina que en la muestra.

En este caso, se puede proceder de igual manera que para la prueba de bondad de ajuste, como muestra el script 8.2. Para este ejemplo, el valor p resultante es $p = 0,461$, por lo que se falla al rechazar la hipótesis nula con un nivel de significación $\alpha = 0,05$. En consecuencia, podemos concluir con 95 % de confianza que la muestra seleccionada es, en efecto, representativa de la nómina de programadores de la empresa, por lo que la acusación del sindicato no tiene fundamentos.

Script 8.2: prueba chi-cuadrado de bondad de ajuste.

```r
# Crear tabla de contingencia.
nomina <- c(236, 78, 204, 76, 66)
muestra <- c(17, 9, 14, 10, 5)

tabla <- as.table(rbind(nomina, muestra))

dimnames(tabla) <- list(grupo = c("Nómina", "Muestra"),
                        lenguajes = c("C", "Java", "Python", "Ruby", "Otro"))

print(tabla)

# Verificar si se esperan más de 5 observaciones por cada grupo.
n_nomina <- sum(nomina)
n_muestra <- 55
proporciones <- round(nomina/n_nomina, 3)
esperados <- round(proporciones * n_muestra, 3)
print(esperados)

# Hacer prueba chi-cuadrado de homogeneidad.
prueba <- chisq.test(tabla, correct = FALSE)
print(prueba)
```

Prueba chi-cuadrado de independencia

Al ejecutar la prueba en R (script 8.3) obtenemos que el valor para el estadístico de prueba es $\chi^2 = 485,64$, con $\nu = 4$ grados de libertad y un valor p $< 2,2 \cdot 10^{-16}$. Aún para un nivel de significación muy exigente, como $\alpha = 0,01$, el valor p obtenido nos permite rechazar la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, concluimos con 99 % de confianza que las variables clase y forma del sombrero están relacionadas (son dependientes).

Script 8.3: prueba chi-cuadrado de independencia.

```
# Crear tabla de contingencia.
comestible <- c(404, 1948, 32, 228, 1596)
venenoso <- c(48, 1708, 0, 600, 1556)

tabla <- as.table(rbind(comestible, venenoso))

dimnames(tabla) <- list(tipo = c("comestible", "venenoso"),
                        sombrero = c("campana", "convexo", "hundido",
                                     "nudoso", "plano"))


print(tabla)

# Hacer prueba chi-cuadrado de independencia.
prueba <- chisq.test(tabla)
cat("\nLa prueba internamente calcula los valores esperados:\n")
esperados <- round(prueba[["expected"]], 3)
print(esperados)

cat("\nResultado de la prueba:\n")
print(prueba)
```

Esta prueba permite **determinar si dos variables categóricas, de una misma población, son estadísticamente independientes** o si, por el contrario, están relacionadas.

Tomemos en este caso como ejemplo que un micólogo desea determinar si existe relación entre la forma del sombrero de los hongos y si estos son o no comestibles. Para ello, tras recolectar una muestra de 8.120 hongos, obtiene la tabla de contingencia que se muestra en la tabla 8.6[1].

		Forma del sombrero					
		Campana	**Convexo**	**Hundido**	**Nudoso**	**Plano**	**Total**
Clase	**Comestible**	404	1.948	32	228	1.596	4.208
	Venenoso	48	1.708	0	600	1.556	3.912
	Total	452	3.656	32	828	3.152	8.120

Tabla 8.6: tabla de contingencia para las características de los hongos.

Una vez más, comencemos por verificar las condiciones. Podemos suponer que la muestra fue obtenida de manera aleatoria, ya que se trata de un estudio publicado en una revista científica, y, desde luego, representa menos del 10 % de la población mundial de hongos. En consecuencia, se verifica la condición de independencia de las observaciones en las muestras.

Ahora debemos determinar cuántas observaciones esperaríamos tener en cada grupo si las variables fueran independientes. En este caso, la frecuencia esperada para cada celda está dado por la ecuación 8.3, donde:

- n_i: total de observaciones en la fila i.
- n_j: total de observaciones en la columna j.
- n: tamaño de la muestra.

$$E_{i,j} = \frac{n_i \cdot n_j}{n} \tag{8.3}$$

De acuerdo a esto, la cantidad de hongos comestibles con sombrero en forma de campana que esperaríamos encontrar es:

$$E_{1,1} = \frac{4.208 \cdot 452}{8.120} = 234,238$$

Si replicamos este cálculo para cada celda de nuestra matriz de confusión, se obtienen los valores esperados que se presentan en la tabla 8.7. Podemos ver que todos los valores esperados superan las 5 observaciones, por lo que podemos proceder con la prueba χ^2 de independencia.

		Forma del sombrero				
		Campana	**Convexo**	**Hundido**	**Nudoso**	**Plano**
Clase	**Comestible**	234,238	1.894,636	16,583	429,092	1.633,450
	Venenoso	217,762	1.761,364	15,417	398,908	1.518,550

Tabla 8.7: frecuencias esperadas para los hongos.

En este caso, las hipótesis a docimar son:

H_0: las variables clase y forma del sombrero son independientes.
H_A: las variables clase y forma del sombrero están relacionadas.

[1]Datos obtenidos desde el conjunto de datos Mushroom, disponible en https://archive.ics.uci.edu/ml/datasets/mushroom (última visita: 26-04-2021).

Prueba exacta de Fisher

La **prueba exacta de Fisher** es una alternativa a la prueba χ^2 de independencia en el caso de que **ambas variables sean dicotómicas**. Así, las hipótesis a contrastar son:

H_0: las variables son independientes.
H_A: las variables están relacionadas.

En este escenario, las frecuencias de la muestra pueden resumirse en una tabla de contingencia de 2×2, como muestra la tabla 8.8.

		Variable 1		
		Presente	**Ausente**	**Total**
Variable 2	**Presente**	a	b	a+b
	Ausente	c	d	c+d
	Total	a+c	b+d	n

Tabla 8.8: tabla de contingencia para dos variables categóricas con dos niveles cada una.

correspondiente a la función de distribución hipergeométrica.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \tag{8.4}$$

La prueba lleva en su nombre la palabra **exacta** porque internamente construye todas las tablas posibles con los mismos totales marginales que recibe como entrada y, para cada una de ellas, determina la probabilidad exacta de observarla. El valor p corresponde en este caso a la suma de las probabilidades de todas las tablas con probabilidad menor o igual que la tabla dada.

Para entender mejor esta prueba, supongamos que un controvertido estudio desea determinar si dos vacunas, Argh y Grrr, son igualmente efectivas para inmunizar a la población ante una mordida de vampiro. Para ello, los investigadores reclutaron a 17 voluntarios de todo el mundo, de los cuales 6 recibieron la vacuna Argh y los 11 restantes, la Grrr. Al cabo de tres meses, sometieron a cada uno de los participantes a una mordida de vampiro y observaron que ninguno de los voluntarios que recibieron la vacuna Argh resultó afectado, mientras que 5 de los que recibieron la vacuna Grrr se convirtieron en vampiros, como resume la tabla 8.9.

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	0	5	5
	Humano	6	6	12
	Total	6	11	17

Tabla 8.9: tabla de contingencia con los contagios producidos en el experimento.

La probabilidad de observar un conjunto de frecuencias con los mismos totales por fila y por columna, si las variables son realmente independientes está dada por:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} = \frac{5!12!6!11!}{17!0!5!6!6!} = 0,075$$

Son cinco las posibles tablas (además de la obtenida) con iguales valores marginales, como podemos ver en la tabla 8.10.

Calculando las probabilidades para cada una de ellas de acuerdo a la ecuación 8.4, se tiene que:

- Probabilidad para la tabla 8.10a: 0,001.
- Probabilidad para la tabla 8.10b: 0,320.
- Probabilidad para la tabla 8.10c: 0,027.
- Probabilidad para la tabla 8.10d: 0,400.
- Probabilidad para la tabla 8.10e: 0,178.

Así, el valor p está dado por la suma de las probabilidades de las tablas con probabilidad menor o igual a la de los datos observados:

$$p = 0,075 + 0,001 + 0,027 = 0,103$$

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Infectado**	5	0	5
	Sano	1	11	12
	Total	6	11	17

(a)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Infectado**	1	4	5
	Sano	5	7	12
	Total	6	11	17

(b)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Infectado**	4	1	5
	Sano	2	10	12
	Total	6	11	17

(c)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Infectado**	2	3	5
	Sano	4	8	12
	Total	6	11	17

(d)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Infectado**	3	2	5
	Sano	3	9	12
	Total	6	11	17

(e)

Tabla 8.10: tablas con los mismos valores marginales que los obtenidos.

Considerando un nivel de significación $\alpha = 0,05$, se falla al rechazar la hipótesis nula. En consecuencia, se concluye con 95 % de confianza que no hay una asociación estadísticamente significativa entre la cantidad de nuevos vampiros y la vacuna recibida.

En R, podemos llevar a cabo esta prueba mediante la función `fisher.test(x, conf.level)`, donde `x` corresponde a la tabla de contingencia y `conf.level`, al nivel de confianza. El script 8.4 muestra el su uso para el ejemplo (con una pequeña diferencia en el valor p obtenido debido a los redondeos efectuados en el cálculo anterior).

Script 8.4: prueba exacta de Fisher.

```
# Construir la tabla de contingencia.
vacuna <- c(rep("Argh", 6), rep("Grrr", 11))
resultado <- c(rep("Humano", 12), rep("Vampiro", 5))
datos <- data.frame(resultado, vacuna)
tabla <- xtabs(~., datos)
print(tabla)

# Aplicar prueba exacta de Fisher.
alfa <- 0.05
prueba <- fisher.test(tabla, 1-alfa)


print(prueba)
```

Prueba de mcNemar

Esta prueba resulta apropiada cuando una misma característica, con respuesta dicotómica, se mide en dos ocasiones diferentes para los mismos sujetos (muestras pareadas) y queremos determinar si se produce o no un cambio significativo entre ambas mediciones. Una vez más, podemos registrar las frecuencias en una matriz de confusión como la que vimos en 8.8. En ella, podemos ver que las celdas a y d corresponde a instancias en que no hay cambios. La celda b en dicha tabla representa a las instancias que cambian de `Presente` a `Ausente` y la celda c, a instancias que cambian de `Ausente` a `Presente`.

Las hipótesis asociadas a la prueba de mcNemar son:

H_0: **no** hay cambios significativos en las respuestas.
H_A: **sí** hay cambios significativos en las respuestas.

Puesto que nos interesa medir los cambios, solo nos sirven las celdas b y c de la tabla de contingencia. La cantidad de instancias en que se producen cambios es $b+c$ y, de acuerdo a la hipótesis nula, se esperaría que $(b+c)/2$ cambien en un sentido y que las $(b+c)/2$ restantes lo hicieran en sentido contrario. Así, b y c cuentan respectivamente los éxitos y los fracasos de una distribución binomial de $b+c$ intentos con probabilidad de éxito igual a $1/2$. Cuando $(b+c) > 10$, esta distribución binomial se asemeja a una distribución normal con la misma media $((b+c)/2)$ y desviación estándar $\sqrt{(b+c)/4}$, a partir de la cual se puede obtener un estadístico z. Sin embargo, la mayoría de los paquetes de software para estadística (incluido R) reportan el cuadrado de dicho estadístico (e ignoran completamente los casos en que hay 10 o menos cambios entre las mediciones), el cual sigue una distribución χ^2 con un grado de libertad y se calcula como muestra la ecuación 8.5 (Agresti, 2019).

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{8.5}$$

Puesto que los datos siguen una distribución binomial (discreta), pero se está usando como aproximación la distribución chi-cuadrado (continua), suele emplearse un **factor de corrección de continuidad** propuesta por Frank Yates en 1934. El estadístico de prueba con la corrección de Yates se calcula en realidad como muestra la ecuación 8.6.

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{8.6}$$

Para ilustrar el funcionamiento de la prueba de mcNemar, suponga que un cientista de datos ha construido dos modelos para predecir, a partir de las notas obtenidas en cursos previos, si sus estudiantes aprobarán o no la asignatura de aprendizaje automático. Al probar sus modelos con los 25 estudiantes del semestre anterior, observó que predijeron el resultado final de cada estudiante como muestra la tabla 8.11 y se resume en la matriz de confusión de la tabla 8.12.

El cientista de datos desea saber si existe diferencia entre el desempeño de ambos algoritmos, por lo que decide emplear la prueba de mcNemar. Al calcular el estadístico de prueba (con el factor de corrección), obtiene:

$$\chi^2 = \frac{(|5-7|-1)^2}{5+7} = \chi^2 = \frac{(5-7)^2}{5+7} = 0,083$$

Alumno	Modelo 1	Modelo 2
1	Correcto	Correcto
2	Correcto	Correcto
3	Correcto	Correcto
4	Correcto	Correcto
5	Correcto	Correcto
6	Correcto	Correcto
7	Correcto	Correcto
8	Correcto	Correcto
9	Correcto	Correcto
10	Correcto	Incorrecto
11	Correcto	Incorrecto
12	Correcto	Incorrecto
13	Correcto	Incorrecto
14	Correcto	Incorrecto
15	Correcto	Incorrecto
16	Correcto	Incorrecto
17	Incorrecto	Incorrecto
18	Incorrecto	Incorrecto
19	Incorrecto	Incorrecto
20	Incorrecto	Incorrecto
21	Incorrecto	Correcto
22	Incorrecto	Correcto
23	Incorrecto	Correcto
24	Incorrecto	Correcto
25	Incorrecto	Correcto

Tabla 8.11: resultados de la predicción para cada estudiante con ambos modelos.

		Modelo 1		
		Correcto	**Incorrecto**	**Total**
Modelo 2	**Correcto**	9	5	14
	Incorrecto	7	4	11
	Total	16	9	25

Tabla 8.12: tabla de contingencia con las predicciones de los resultados finales de los estudiantes.

El valor p está dado por el área bajo la cola superior de la distribución chi-cuadrado, que en R puede calcularse como `pchisq(0.083, 1, lower.tail = FALSE)`, obteniéndose que $p = 0,773$. En consecuencia, se falla al rechazar la hipótesis nula (para un nivel de significación $\alpha = 0,05$) y se concluye que no hay diferencia en el desempeño de ambos clasificadores.

La función de R para esta prueba, que por defecto aplica el factor de corrección, es `mcNemar.test(x)`, donde `x` corresponde a la tabla de contingencia. El script 8.5 muestra su aplicación para el ejemplo dado.

Script 8.5: prueba de mcNemar.

```
# Construir la tabla de contingencia.
alumno <- seq(1:25)
modelo_1 <- c(rep("Correcto", 16), rep("Incorrecto", 9))
modelo_2 <- c(rep("Correcto", 9), rep("Incorrecto", 11), rep("Correcto", 5))
datos <- data.frame(alumno, modelo_2, modelo_1)
tabla <- table(modelo_2, modelo_1)
print(tabla)

# Aplicar prueba de McNemar.

prueba <- mcnemar.test(tabla)
print(prueba)
```

Prueba de Cochran

La **prueba Q de Cochran** es una extensión de la prueba de mcNemar, adecuada cuando la variable de respuesta es dicotómica y la variable independiente tiene más de dos observaciones pareadas (cuando ambas variables son dicotómicas, esta prueba es equivalente a la de mcNemar). Como tal, debería estar incluida en la sección precedente, pero le dedicaremos una sección aparte pues la explicación requiere de algunos conceptos importantes que no hemos estudiado aún.

Veamos esta prueba por medio de un ejemplo. Elsa Capunta, estudiante de un curso de algoritmos, tiene como tarea determinar si existe una diferencia significativa en el desempeño de tres metaheurísticas que buscan resolver el el problema del vendedor viajero. Para ello, el profesor le ha proporcionado los datos presentados en la tabla 8.13, donde la columna `instancia` identifica cada instancia del problema empleada para evaluar las metaheurísticas y las restantes columnas indican si la metaheurística en cuestión encontró (1) o no (0) la solución óptima para dicha instancia.

Instancia	Simulated Annealing	Colonia de hormigas	Algoritmo genético
1	0	0	1
2	1	0	0
3	0	1	1
4	0	0	1
5	0	0	1
6	0	1	1
7	0	0	0
8	1	0	1
9	0	0	0
10	0	1	1
11	0	0	1
12	0	0	0
13	1	0	0
14	0	0	1
15	0	1	1

Tabla 8.13: resultados de las metaheurísticas para cada instancia con ambos modelos.

Las hipótesis contrastadas por la prueba Q de Cochran son:

H_0: la proporción de "éxitos" es la misma para todos los grupos.
H_A: la proporción de "éxitos" es distinta para al menos un grupo.

Como ya debemos suponer, esta prueba también requiere que se cumplan algunas condiciones:

1. La variable de respuesta es dicotómica.
2. La variable independiente es categórica.
3. Las observaciones son independientes entre sí.
4. El tamaño de la muestra es lo suficientemente grande. Glen (2016a) sugiere que $n \cdot k \geq 24$, donde n es el tamaño de la muestra (la cantidad de instancias, para el ejemplo) y k, la cantidad de niveles en la variable independiente.

El estadístico de prueba se calcula como muestra la ecuación 8.7, donde:

- b: cantidad de bloques.
- k: cantidad de bloques (niveles de la variable independiente).
- x_j: total de éxitos en la columna j.
- x_i: total de éxitos en la fila i.
- N: número total de éxitos.

$$Q = k(k-1)\frac{\sum_{j=1}^{k}\left(x_j - \frac{N}{k}\right)^2}{\sum_{i=1}^{b} x_i(k - x_i)} \tag{8.7}$$

Podemos ver que los cálculos que se llevan a cabo para esta prueba son complejos, por lo que suele hacerse mediante software. En R, esta prueba está implementada en la función `cochran.qtest(formula, data, alpha = 0.05)` del paquete `RVAideMemoire`, donde:

- `formula`: fórmula de la forma `respuesta ~ independiente | bloques`.
- `data`: matriz de datos en formato largo.
- `alpha`: nivel de significación.

Al ejecutar el script 8.6, obtenemos el resultado que se muestra en la figura 8.1. Tenemos que el valor p es $p = 0,028$, menor que el nivel de significación $\alpha = 0,05$, por lo que rechazamos la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, Elsa concluye con 95 % de confianza que al menos una de las metaheurísticas tiene un desempeño diferente a las demás.

```
Cochran's Q test

data:  resultado by metaheuristica, block = instancia
Q = 7.1667, df = 2, p-value = 0.02778
alternative hypothesis: true difference in probabilities is not equal to 0
sample estimates:
proba in group annealing  proba in group genetico  proba in group hormigas
               0.2000000                0.6666667                0.2666667
```

```
        Pairwise comparisons using Wilcoxon sign test

          annealing genetico
genetico   0.09814        -
hormigas   1.00000  0.09375

P value adjustment method: fdr
```

Figura 8.1: resultado de la prueba Q de Cochran.

Todo buen estudiante sabe que Elsa debe entregar en su tarea una respuesta más detallada que la que hemos obtenido hasta ahora, pues el profesor esperaría un análisis de las diferencias.

En este punto, debemos mencionar que la hipótesis nula de la prueba Q de Cochran no es específica, sino que comprueba la igualdad de todas las proporciones. Esta clase de hipótesis nula suele llamarse **ómnibus** (en ocasiones también colectiva o global). Así, se dice que la prueba Q de Cochran es una prueba ómnibus porque tiene esta clase de hipótesis nula, con la dificultad de que solo detecta si existe al menos bloque con una proporción de "éxito" diferente. Sin embargo, de ser afirmativa la respuesta, no nos dice qué grupos presentan diferencias (Lane, s.f.). Desde luego, existen métodos para responder a esta última pregunta, llamados **pruebas *post-hoc***, o también ***a posteriori***. Reciben este nombre porque se realizan una vez que se ha concluido gracias a la prueba ómnibus que existen diferencias significativas.

Algo importante que debemos recordar: **solo haremos un procedimiento post-hoc si la prueba ómnibus rechaza la hipótesis nula** en favor de la hipótesis alternativa. Además, el procedimiento post-hoc

realizado debe considerar el mismo nivel de significación que la prueba ómnibus.

En el caso de la prueba Q de Cochran, el procedimiento post-hoc consiste en efectuar pruebas de mcNemar entre cada par de bloques. R nos permite hacer esto mediante la función `pairwiseMcnemar(formula, data, method)` del paquete `rcompanion`, donde `formula` y `data` son las mismas que para la prueba Q de Cochran y `method` nos permite determinar el método para ajustar los valores p de las comparaciones. Pero... ¿por qué querríamos ajustar los valores p?

Como explican Goeman y Solari (2014), cuando contrastamos hipótesis acotamos la probabilidad de cometer errores tipo I por medio del nivel de significación α. Sin embargo, cuando hacemos múltiples contrastes de hipótesis simultáneamente, cada uno de ellos tendrá una probabilidad α de cometer un error de tipo I. Esto se traduce en un **incremento de la probabilidad de cometer este tipo de errores** a medida que aumenta la cantidad de hipótesis contrastadas y, en consecuencia, en una reducción del poder estadístico.

Muchos factores de corrección tienen por objeto distribuir el nivel de significación empleado para la prueba ómnibus en cada prueba de pares de bloques. El método más sencillo para ajustar los valores p es la **corrección de Bonferroni**. Como explica la ayuda de R, esta corrección simplemente multiplica el valor p obtenido en cada prueba por la cantidad de pruebas realizadas. En general, no se recomienda el uso del método de Bonferroni, especialmente si el número de grupos es alto, pues es considerado muy **conservador**, lo que significa que mantiene la probabilidad de cometer un error tipo I más baja que el nivel de significación establecido (y es, por ende, más propensa a cometer errores tipo II).

Otra alterativa es la **corrección de Holm** (Glen, 2016b), con mayor poder estadístico que la de Bonferroni. Esta corrección comienza por efectuar las pruebas entre pares de bloques y luego ordena los valores p en forma creciente. A continuación, se calcula el factor de Holm, HB, para cada par de bloques, dado por la ecuación 8.8, donde:

- α: nivel de significación.
- N: cantidad de comparaciones efectuadas.
- i: importancia de la comparación (posición en la lista de valores p ordenados).

$$HB_i = \frac{\alpha}{N - i + 1} \tag{8.8}$$

Luego, se compara el valor p con su respectivo factor de Holm y, si el valor p es menor, se considera que existe una diferencia significativa. R implementa esta corrección de manera ligeramente diferente, de modo que el valor p ajustado pueda ser comparado con el nivel de significación original.

Si has estado leyendo de manera atenta, habrás notado que en el resultado entregado por `cochran.qtest()` para el ejemplo (figura 8.1) aparece otro procedimiento post-hoc adecuado para la prueba Q de Cochran, aunque no lo presentaremos aquí pues se basa en una prueba que estudiaremos en capítulos posteriores.

El script 8.6 incluye también los procedimientos *post-hoc* mediante pruebas de mcNemar usando las correcciones de Holm y Bonferroni, obteniéndose los resultados que se muestran en la figura 8.2.

Podemos ver en la figura 8.2 que, aún cuando la prueba Q de Cochran indica que existen diferencias significativas entre las metaheurísticas, ninguno de los procedimientos post-hoc ha detectado diferencias significativas entre pares de bloques. En consecuencia, la respuesta que Elsa debe dar a su profesor es que la evidencia no es lo suficientemente fuerte para poder afirmar que existen diferencias entre las metaheurísticas, pero que podría ser adecuado hacer un estudio con una muestra mayor puesto que los resultados de la prueba Q de Cochran y de los procedimientos post-hoc son contradictorios.

Script 8.6: prueba Q de Cochran.

```
library(tidyverse)
library(RVAideMemoire)
library(rcompanion)

# Crear matriz de datos.
instancia <- 1:15
```

```
Cochran's Q test

Procedimiento post-hoc con corrección de Bonferroni
$Test.method
   Test
1 exact

$Adustment.method
      Method
1 bonferroni

$Pairwise
                Comparison Successes Trials p.value p.adjust
1 annealing - genetico = 0         2     11  0.0654   0.1960
2 annealing - hormigas = 0         3      7       1   1.0000
3  genetico - hormigas = 0         6      6  0.0313   0.0939


Procedimiento post-hoc con corrección de Holm
$Test.method
   Test
1 exact

$Adustment.method
  Method
1   holm

$Pairwise
                Comparison Successes Trials p.value p.adjust
1 annealing - genetico = 0         2     11  0.0654   0.1310
2 annealing - hormigas = 0         3      7       1   1.0000
3  genetico - hormigas = 0         6      6  0.0313   0.0939
```

Figura 8.2: resultados de los procedimientos post-hoc.

```r
annealing <- c(0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0)
hormigas <- c(0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1)
genetico <- c(1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1)
datos <- data.frame(instancia, annealing, hormigas, genetico)

# Llevar matriz de datos a formato largo.
datos <- datos %>% pivot_longer(c("annealing", "hormigas", "genetico"),
                                names_to = "metaheuristica",
                                values_to = "resultado")

datos[["instancia"]] <- factor(datos[["instancia"]])
datos[["metaheuristica"]] <- factor(datos[["metaheuristica"]])

# Hacer prueba Q de Cochran.
prueba <- cochran.qtest(resultado ~ metaheuristica | instancia,
                        data = datos, alpha = 0.05)

print(prueba)

# Procedimiento post-hoc con corrección de Bonferroni.

post_hoc_1 <- pairwiseMcnemar(resultado ~ metaheuristica | instancia,
                              data = datos, method = "bonferroni")

cat("\nProcedimiento post-hoc con corrección de Bonferroni\n")
print(post_hoc_1)

# Procedimiento post-hoc con corrección de Holm.
post_hoc_2 <- pairwiseMcnemar(resultado ~ metaheuristica | instancia,
                              data = datos, method = "holm")

cat("\nProcedimiento post-hoc con corrección de Holm\n")
print(post_hoc_2)
```