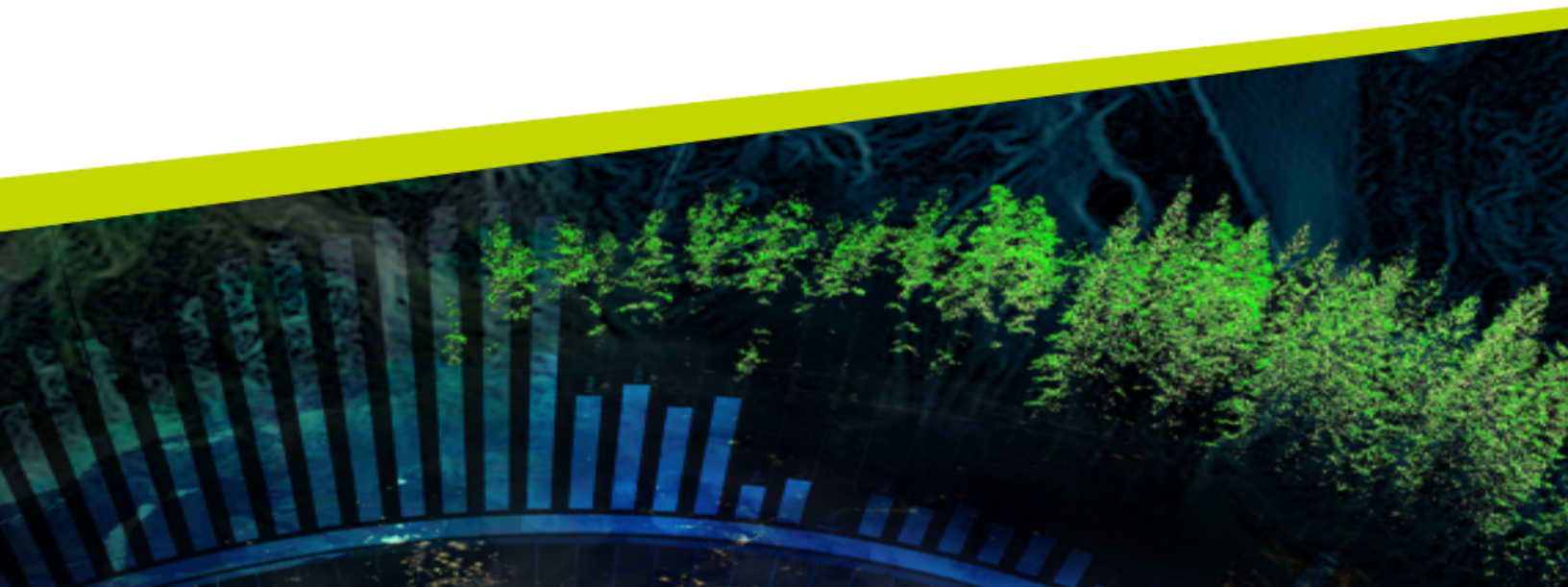
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 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 6, 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? La respuesta a esta pregunta es que existen 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 indicación de **qué valores** tendrían tales proporciones, 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.

Como ya hemos dicho, en este capítulo conoceremos algunas pruebas no paramétricas para estudiar la relación entre dos variables categóricas, con base en Diez et al. (2017, pp. 286-302), Pértega y Pita (2004), Glen (2016a), Goeman y Solari (2014) y Mangiafico (2016).

8.1 PRUEBA EXACTA DE FISHER

Cuenta la leyenda que en la década de 1920 una colega del estadístico Ronald Fisher[1] aseguraba que podía distinguir si en su taza de té inglés se le había puesto la leche antes o después del verter el té. Fisher planificó un experimento con algunas tazas de té para comprobar la habilidad de su colega. Del estudio de este experimento de dos variables binarias (¿se puso la leche después del té? y ¿pudo la colega reconocer este orden?) y el número reducido de observaciones, nació la **prueba exacta de Fisher**.

Entonces, la versión original de esta prueba de hipótesis permite analizar la **asociación entre dos variables dicotómicas** basándose en los conteos observados y registrados en una tabla de contingencia (o matriz de confusión) de 2×2, como muestra la tabla 8.1 para dos variables binarias con niveles "Presente" y

[1] Considerado una de las figuras centrales en la fundación de la estadística inferencial clásica, junto a Jerzy Neyman y Egon Pearson.

"Ausente". Así, la celda (a) contiene el número de observaciones que presentan la combinación (Presente, Presente), la celda (b) la frecuencia observada de la combinación (Ausente, Presente), la celda (c) el conteo de la combinación (Presente, Ausente), y la celda (d) las veces que la combinación (Ausente, Ausente) fue observada. Es común referirse a cada una de estas combinaciones (cada celda de la matriz de confusión) como "un grupo". Por último, notemos que los valores marginales son los subtotales de cada fila y columna, las que coinciden en el número total de observaciones registradas (n).

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

Tabla 8.1: tabla de contingencia para dos variables categóricas con dos niveles cada una.

Si se asume independencia entre ambas variables, la **probabilidad exacta** de observar el conjunto de frecuencias de la tabla 8.1 está dada por la ecuación 8.1, correspondiente a la función de distribución hipergeométrica:

$$P(\text{Tabla 2.1}) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} \tag{8.1}$$

De esta forma, si encontramos todas las posibles tablas de contingencia de 2×2 que tienen las mismas sumas de filas y columnas que la tabla observada, y calculamos la probabilidad de cada una de ellas bajo el supuesto de que no hay asociación entre las variables (usando la ecuación anterior), podemos determinar la **probabilidad exacta** de obtener la tabla de contingencia observada o una más extrema simplemente sumando las probabilidades de aquellas cuya probabilidad sea menor o igual a la probabilidad de la tabla observada. Es decir, se puede tener el **valor p exacto** del contraste de las siguientes hipótesis:

H_0: las variables son independientes (no hay asociación).
H_A: las variables están relacionadas (existe asociación).

Esta es la razón por lo que esta prueba lleva en su nombre la palabra **exacta**, pues la gran mayoría de las otras pruebas de hipótesis entregan valores p basados en **aproximaciones asintóticas**.

Para entender mejor el procedimiento, 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, el grupo de investigación reclutaron a 17 personas de todo el mundo, de los cuales 6 recibieron la vacuna Argh y los 11 restantes, la Grrr. La asignación de la vacuna recibida fue realizada de manera aleatoria utilizando una ruleta. Al cabo de tres meses, cada participante fue sometida/o a una mordida de vampiro. El grupo de investigación reportó 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 8.2.

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

Tabla 8.2: tabla de contingencia con los contagios producidos en el experimento.

Veamos primero si se cumplen las condiciones para aplicar la prueba exacta de Fisher:

1. El estudio plantea el cruce de dos variables binarias: vacuna recibida (Argh o Grrr) y resultado de la mordida (vampiro o humano), separando las observaciones en cuatro grupos cuyas frecuencias observadas se resumen en la tabla 8.2.
2. El enunciado no es muy claro sobre cómo fueron seleccionadas las personas en la muestra, pero podríamos asumir que se lograron conseguir solo 17 voluntarias/os por lo invasivo y peligroso del experimento. Al parecer, estas personas vendrían de países distintos, por lo que podríamos suponer que no se conocen

entre sí y que no se influenciaron mutuamente para ofrecerse para el estudio (aunque esto es cada día más difícil para grupos muy específicos, como gente dispuesta a ser mordida por un vampiro, por el auge de las redes sociales). Obviamente 17 personas es mucho menor al 10 % de la población mundial. Como la vacuna fue asignada de manera aleatoria, la vacuna recibida por una voluntaria o un voluntario no influye en qué vacuna recibe otra persona de la muestra. Así, las observaciones en cada grupo parecen ser **independientes entre sí**.

3. Si bien **no es una condición** para la confiabilidad de la prueba, podemos verificar que la muestra es pequeña y que las celdas en la tabla 8.2 contienen alrededor de cinco observaciones e incluso hay una con cero. Esto nos sugiere que el número de todas las posibles tablas de contingencia de 2×2 con las mismas sumas marginales no debería ser muy grande, haciendo factible realizar la prueba en tiempo razonable.

También es importante que hagamos explícitas las hipótesis que estaríamos contrastando:

H_0: el resultado de una mordedura de vampiro no está asociada a la vacuna recibida por la persona mordida.
H_A: el resultado de una mordedura de vampiro depende de la vacuna recibida por la persona mordida.

Realicemos el procedimiento para contrastar estas hipótesis. La probabilidad de observar este conjunto de frecuencias si las variables son realmente independientes está dada por:

$$P(\text{Tabla 2.2}) = \frac{5! \cdot 12! \cdot 6! \cdot 11!}{17! \cdot 0! \cdot 5! \cdot 6! \cdot 6!} \approx 0{,}075$$

Notemos que esta probabilidad es un valor aproximado solo porque lo hemos redondeado a tres decimales, no porque se esté usando una aproximación a una distribución probabilística.

Usando un algoritmo estándar para generar permutaciones es posible encontrar que hay cinco tablas (además de la obtenida desde la muestra) con valores marginales iguales a los observados, las que se presentan en la figura 8.1.

(a)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	5	0	5
	Humano	1	11	12
	Total	6	11	17

(b)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	1	4	5
	Humano	5	7	12
	Total	6	11	17

(c)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	4	1	5
	Humano	2	10	12
	Total	6	11	17

(d)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	2	3	5
	Humano	4	8	12
	Total	6	11	17

(e)

		Vacuna		
		Argh	**Grrr**	**Total**
Resultado	**Vampiro**	3	2	5
	Humano	3	9	12
	Total	6	11	17

Figura 8.1: tablas con los mismos valores marginales que los observados.

Calculando las probabilidades para cada una de ellas de acuerdo a la ecuación 8.1, se tiene que:

- $P(\text{Tabla (a) de la figura 8.1}) \approx 0{,}001$.
- $P(\text{Tabla (b) de la figura 8.1}) \approx 0{,}320$.
- $P(\text{Tabla (c) de la figura 8.1}) \approx 0{,}027$.
- $P(\text{Tabla (d) de la figura 8.1}) \approx 0{,}400$.
- $P(\text{Tabla (e) de la figura 8.1}) \approx 0{,}178$.

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 (incluyéndola):

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

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 existe evidencia de que exista una asociación entre el resultado de una mordedura de vampiro y la vacuna recibida por la persona mordida.

En R, la función `fisher.test(x, y = NULL, alternative = "two.sided", conf.level = 0.95)` nos permite llevar a cabo una prueba exacta de Fisher. Cuando `y = NULL` (o se omite en la llamada), `x` debe corresponder a la tabla de contingencia. Sino, `x` e `y` han de corresponder a las muestras con las observaciones de la variable dicotómica en cada grupo. El argumento `conf.level` corresponde, obviamente, al nivel de confianza definido para la prueba, que por omisión toma el valor tradicional 0,95 (es decir, un nivel de significación $\alpha = 0,05$).

Si bien la prueba original fue pensada para tablas de contingencia de 2×2, la función `fisher.test()` es capaz de manejar tablas de otras dimensiones. Sin embargo, siempre debe recordarse que internamente se construyen todas las tablas posibles con los mismos totales marginales que la original, número que crece rápidamente con el número de dimensiones o el tamaño de la muestra. Así, para no encontrarse con problemas de rendimiento al ejecutar esta función, la prueba exacta debería ser usada solamente para muestras de **tamaños pequeños**, cuando el número esperado de observaciones en una o más celdas de la tabla de contingencia sea menor de 5. De hecho la función provee otros argumentos que liberan la necesidad de obtener un valor p exacto, y le permiten usar simulaciones para conseguir uno aproximado, cuando la tabla de contingencia no es de 2×2.

La función `fisher.test()` también permite realizar pruebas unilaterales (`"greater"` o `"less"`). En este caso, las tablas de contingencia "más extremas" que la observada son las tablas que muestran una **asociación más fuerte** entre las variables en **la dirección** de la hipótesis alternativa que se está considerando. Por ejemplo, si el objetivo del estudio de las vacunas contra mordidas de vampiros fuera mostrar la mayor eficacia de la vacuna Grrr (en vez de determinar si las dos opciones son igualmente efectivas), las tablas más extremas serían aquellas que tienen un número aún mayor de personas sin infectarse en el grupo que recibió dicha vacuna (aumentando la frecuencia en la celda d) mientras se mantienen los totales marginales.

La figura 8.2 presenta el resultado entregado al ejecutar el script 8.1 para el ejemplo de la tabla 8.2 y la figura 8.1. La (pequeña) diferencia en el valor p obtenido, en relación al cálculo manual expuesto anteriormente, se debe a los redondeos aplicados en el cálculo de las probabilidades de las tablas relevantes.

Script 8.1: prueba exacta de Fisher.

```r
# Construir los datos y la tabla de contingencia
Vacuna <- c(rep("Argh", 6), rep("Grrr", 11))
Resultado <- c(rep("Humano", 12), rep("Vampiro", 5))
Resultado <- factor(Resultado, levels = c("Vampiro", "Humano"))
datos <- data.frame(Resultado, Vacuna)
tabla <- xtabs(~., datos)
print(tabla)

# Aplicar la prueba exacta de Fisher a la tabla de contingencia
prueba_1 <- fisher.test(tabla)
cat("\n")
cat("Prueba exacta de Fisher usando la tabla de contingencia\n")
cat("-------------------------------------------------------\n")
print(prueba_1)

# Aplicar la prueba exacta de Fisher directamente a las muestras
prueba_2 <- fisher.test(Vacuna, Resultado)
cat("Prueba exacta de Fisher usando las muestras:\n")
cat("--------------------------------------------\n")
print(prueba_2)
```

```
         Vacuna
Resultado Argh Grrr
  Vampiro    0    5
  Humano     6    6

Prueba exacta de Fisher usando la tabla de contingencia
-------------------------------------------------------

 Fisher's Exact Test for Count Data

data:  tabla
p-value = 0.1023
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.000000 1.726613
sample estimates:
odds ratio
         0

Prueba exacta de Fisher usando las muestras:
--------------------------------------------

 Fisher's Exact Test for Count Data

data:  Vacuna and Resultado
p-value = 0.1023
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.000000 1.726613
sample estimates:
odds ratio
         0
```

Figura 8.2: salida del script 8.1 que ejemplifica del uso de la función `fisher.test()`.

8.2 PRUEBA CHI-CUADRADO DE PEARSON

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). Como vimos en la sección anterior, podemos registrar las frecuencias observadas para las posibles combinaciones de ambas variables mediante una tabla de contingencia.

Para usar la prueba chi-cuadrado de Pearson de forma confiable (es decir, que el teorema del límite central opera y podemos usar el modelo normal) se deben cumplir las siguientes condiciones:

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 es común a las pruebas de inferencia que estamos estudiando. Recordemos que esto significa que la elección de **un caso** para la muestra no tiene influencia en la selección o no selección de **otro caso**. También recordemos que, asumiendo que la muestra se construye seleccionando casos de forma aleatoria sin reposición, el tamaño de la muestra debe ser inferior al 10 % del tamaño de la población para garantizar esta independencia.

Como en la sección anterior, cada observación pertenece a un solo grupo, es decir, cada caso es "contado" en una sola celda de la matriz de confusión. La segunda condición la iremos revisando a medida que avancemos en el estudio de la prueba, pero debería quedar en claro que la prueba chi-cuadrado de Pearson es la alternativa (aproximada) a la prueba exacta de Fisher cuando la muestra es grande y se hace ineficiente, o derechamente inviable en la práctica, obtener un valor p exacto.

Si bien en esta sección estamos hablando de 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.

8.2.1 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 de una base de datos con 400.000 profesionales, y les ha preguntado cuál es su lenguaje de programación favorito. La tabla 8.3 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 8.3: 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 8.4 muestra las frecuencias esperadas para cada lenguaje de programación bajo este supuesto, calculadas mediante la ecuación 8.2:

$$E_{(i,j)} = \frac{n_i \, n_j}{n}, \tag{8.2}$$

donde:

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

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

Tabla 8.4: frecuencias esperadas si hombres y mujeres tienen las mismas preferencias.

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, la muestra representa menos del 10 % de los programadores en la base de datos y sabemos que fue seleccionada de manera aleatoria, cumpliendo el requisito de independencia. De esta forma, podemos proceder con la prueba χ^2 de homogeneidad.

Las hipótesis a contrastar son:

H_0: las programadoras y los programadores tienen las mismas preferencias en lenguaje de programación favorito (ambas poblaciones muestras las mismas proporciones para cada lenguaje estudiado).
H_A: las programadoras y los programadores tienen preferencias distintas en lenguajes de programación favorito.

Recordemos que la primera aproximación para construir un estadístico de prueba basándose en el modelo normal está dada por la ecuación 4.5, que reproducimos aquí:

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

Podemos usar esta fórmula de la diferencia estandarizada para cada uno de los grupos: el estimador puntual corresponde a la frecuencia observada para el grupo, el valor nulo es la frecuencia esperada para el grupo y el error estándar del estimador puntual es la raíz cuadrada del valor nulo. Así, para los programadores (varones) en C se tiene:

$$Z_{\text{C}}^{\text{Hombre}} = \frac{42 - 44{,}7}{\sqrt{44{,}7}} \approx -0{,}404$$

Al repetir el procedimiento para cada grupo, se obtienen los valores Z presentados en la tabla 8.5.

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 8.5: 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 8.3, 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).

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\text{cantidad observada} - \text{cantidad esperada})^2}{\text{cantidad esperada}} \tag{8.3}$$

Para el ejemplo tenemos entonces:

$$\begin{aligned}\chi^2 \approx &(-0{,}404)^2 + \quad (0{,}370)^2 + (-0{,}139)^2 + (-0{,}189)^2 + \quad (0{,}426)^2 + \\ &(0{,}572)^2 + (-0{,}523)^2 + \quad (0{,}196)^2 + \quad (0{,}267)^2 + (-0{,}603)^2 \approx 1{,}611\end{aligned}$$

Como estamos sumando m n valores Z al cuadrado, el estadístico χ^2 **sigue una distribución chi-cuadrado**, con $\nu = (m-1)\ (n-1)$ 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 8.2, usando para ello la función `chisq.test(x)`, donde `x` corresponde a la matriz de confusión.

El resultado de ejecutar este script puede verse en la figura 8.3. Debemos tener en cuenta que el valor p obtenido usando R es ligeramente diferente debido a los redondeos aplicados en la tabla 8.4 y al aplicar la

ecuación 8.3. También es importante notar que el script calcula las frecuencias esperadas (líneas 21–27) y reporta si existen grupos con menos observaciones que las requeridas por la prueba (líneas 36–37).

Script 8.2: prueba chi-cuadrado de homogeneidad.

```
# Crear la 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"))

# Definir condiciones y el nivel de significación
minima_frec_esperada <- 5
alfa <- 0.05

# Mostrar la tabla
cat("Tabla de contingencia:\n")
cat("----------------------\n")
print(tabla)

cat("\n\nPrueba global:\n")
cat("==============\n\n")

# Obtener las frecuencias esperadas
sumas_filas <- apply(tabla, 1, sum)
sumas_columnas <- apply(tabla, 2, sum)
suma_total <- sum(tabla)
esperadas <- outer(sumas_filas, sumas_columnas, "*") / suma_total
esperadas <- round(esperadas, 1)
dimnames(esperadas) <- list(sexo = c("programadores", "programadoras"),
                            lenguajes = c("C", "Java", "Python", "Ruby", "Otro"))

# Realizar prueba chi-cuadrado de homogeneidad
prueba <- chisq.test(tabla)

# Mostrar las frecuencias esperadas y si hay grupos que no cumplen con el mínimo
cat("Frecuencias esperadas:\n")
cat("----------------------\n")
print(esperadas)
cat("Frecuencias esperadas bajo", minima_frec_esperada)
cat(":", sum(esperadas < minima_frec_esperada), "\n")

# Mostrar el resultado de la prueba
cat("\nResultado de la prueba chi-cuadrado:\n")
cat("-----------------------------------\n")
print(prueba)
```

8.2.2 Prueba chi-cuadrado de bondad de ajuste

Esta prueba permite comprobar si una **distribución de frecuencias observada** se asemeja a una **distribución de referencia**. 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 y programadoras, especialistas en diferentes lenguajes de programación. La gerencia ha seleccionado un subconjunto de 55 de personas desde esta nómina, supuestamente de forma aleatoria, para enviarlos a cursos de perfeccionamiento en sus respectivos lenguajes. Pero el sindicato ha acusado de "seleccionar estas personas a conveniencia de los intereses mezquinos de la gerencia, impidiendo

```
Tabla de contingencia:
----------------------
                lenguajes
sexo              C Java Python Ruby Otro
  programadores 42   56     51   27   24
  programadoras 25   24     27   15    9


Prueba global:
==============

Frecuencias esperadas:
----------------------
                lenguajes
sexo               C Java Python Ruby Otro
  programadores 44.7 53.3     52   28   22
  programadoras 22.3 26.7     26   14   11
Frecuencias esperadas bajo 5: 0

Resultado de la prueba chi-cuadrado:
------------------------------------

 Pearson's Chi-squared test

data:  tabla
X-squared = 1.5879, df = 4, p-value = 0.811
```

Figura 8.3: salida del script 8.2 que ejemplifica del uso de la función `chisq.test()` para realizar una prueba chi-cuadrado de homogeneidad.

que el grupo sea representativo a fin de asegurar una mejora en la productividad de toda la empresa". Ante el inminente riesgo de movilizaciones, la gerencia necesita demostrar que el grupo seleccionado es una muestra representativa de sus programadores y programadoras. La tabla 8.6 muestra la cantidad de especialistas en cada lenguaje, tanto para la nómina de la empresa como para la muestra seleccionada.

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

Tabla 8.6: frecuencias por lenguaje de programación para la toda la nómina y para la muestra.

Como ya es habitual, comencemos por verificar las condiciones. Puesto que la muestra representa menos del 10 % de la población (la nómina de programadores y programadoras con contrato) 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 personas de la nómina especialista en cada lenguaje. Para el caso del lenguaje C, tenemos:

$$P_{\mathrm{C}} = \frac{n_{\mathrm{C}}}{n} = \frac{236}{660} \approx 0{,}358$$

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

$$E_{\mathrm{C}} = P_{\mathrm{C}}\, n \approx 0{,}358 \cdot 55 = 19{,}690$$

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 8.7. En ella podemos ver que para cada grupo se esperan más de 5 observaciones, por lo que se verifica la segunda condición.
En este ejemplo, las hipótesis a contrastar son:

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 8.7: proporciones de la población y valores esperados de la muestra.

H_0: las proporciones de especialistas en en la muestra son las mismas que para la nómina completa para todos los lenguajes de programación.

H_A: las proporciones de especialistas son diferentes en la nómina que en la muestra para al menos uno de los lenguajes de programación.

En este caso se puede proceder de igual manera que para la prueba de homogeneidad, como muestra el script 8.3, cuyo resultado puede verse en la figura 8.4. Notemos que este script también muestra las frecuencias esperadas (líneas 24–27) y reporta la cantidad de grupos que no cumplen con el número mínimo de observaciones requerido para aplicar la prueba (líneas 28–29). Pero a diferencia del script anterior, las frecuencias esperadas no se obtienen haciendo cálculos manuales, sino que desde los resultados entregados por la función `chisq.test()`.

Para este ejemplo, el estadístico resultante es $\chi^2(4) = 3{,}613$ que conlleva a un valor $p = 0{,}461$, 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 no hay evidencia de que la muestra seleccionada no sea representativa de la nómina de programadores y programadoras de la empresa, por lo que la acusación del sindicato no tiene fundamentos.

Script 8.3: prueba chi-cuadrado de bondad de ajuste.

```
# Crear la 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"))

# Definir condiciones y el nivel de significación
minima_frec_esperada <- 5
alfa <- 0.05

# Mostrar la tabla
cat("Tabla de contingencia:\n")
cat("----------------------\n")
print(tabla)

cat("\n\nPrueba global:\n")
cat("==============\n\n")

# Realizar prueba chi-cuadrado de bondad de ajuste
prueba <- chisq.test(tabla)

# Mostrar las frecuencias esperadas
cat("Frecuencias esperadas:\n")
cat("----------------------\n")
esperadas <- round(prueba[["expected"]], 1)
print(esperadas)
cat("Frecuencias esperadas bajo", minima_frec_esperada)
cat(":", sum(esperadas < minima_frec_esperada), "\n")

# Mostrar el resultado de la prueba
cat("\nResultado de la prueba chi-cuadrado:\n")
cat("-----------------------------------\n")
print(prueba)
```

```
Tabla de contingencia:
---------------------
         lenguajes
grupo       C Java Python Ruby Otro
  Nómina  236   78    204   76   66
  Muestra  17    9     14   10    5


Prueba global:
==============

Frecuencias esperadas:
---------------------
         lenguajes
grupo         C Java Python Ruby Otro
  Nómina  233.5 80.3  201.2 79.4 65.5
  Muestra  19.5  6.7   16.8  6.6  5.5
Frecuencias esperadas bajo 5: 0

Resultado de la prueba chi-cuadrado:
-----------------------------------

 Pearson's Chi-squared test

data:  tabla
X-squared = 3.613, df = 4, p-value = 0.4609
```

Figura 8.4: salida del script 8.3 que ejemplifica del uso de la función `chisq.test()` para realizar una prueba chi-cuadrado de bondad de ajuste.

8.2.3 Prueba chi-cuadrado de independencia

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 8.8[2].

		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 8.8: 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. Como vimos, estas se obtiene aplicando la ecuación 8.2 en cada celda de la matriz de confusión observada en los datos. De acuerdo a esto, se obtienen los valores esperados que se presentan en la tabla 8.9.

[2]Valores del conjunto de datos Mushroom, disponible en https://archive.ics.uci.edu/ml/datasets/mushroom.

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 8.9: frecuencias esperadas para los hongos.

En este caso, las hipótesis a docimar son:

H_0: si un hongo es comestible o venenoso es independiente de la forma de su sombrero.
H_A: la forma del sombrero de un hongo está relacionada con si este es comestible o venenoso.

Al ejecutar la prueba en R, utilizando el script 8.4, obtenemos la salida presentada en la figura 8.5. El valor para el estadístico de prueba con $\nu = 4$ grados de libertad es $\chi^2(4) \approx 485{,}64$, que conlleva a un valor $p < 0{,}001$. 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 hay evidencia de que las variables clase y forma del sombrero están relacionadas (son dependientes).

Script 8.4: prueba chi-cuadrado de independencia.

```r
# Crear la 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"))

# Definir condiciones y el nivel de significación
minima_frec_esperada <- 5
alfa <- 0.05

# Mostrar la tabla
cat("Tabla de contingencia:\n")
cat("----------------------\n")
print(tabla)

cat("\n\nPrueba global:\n")
cat("==============\n\n")

# Realizar prueba chi-cuadrado de independencia
prueba <- chisq.test(tabla)

# Mostrar las frecuencias esperadas
cat("Frecuencias esperadas:\n")
cat("----------------------\n")
esperadas <- round(prueba[["expected"]], 3)
print(esperadas)
cat("Frecuencias esperadas bajo", minima_frec_esperada)
cat(":", sum(esperadas < minima_frec_esperada), "\n")

# Mostrar el resultado de la prueba
cat("\nResultado de la prueba chi-cuadrado:\n")
cat("-----------------------------------\n")
print(prueba)
```

```
Tabla de contingencia:
---------------------
            sombrero
tipo         campana convexo hundido nudoso plano
  comestible     404    1948      32    228  1596
  venenoso        48    1708       0    600  1556


Prueba ómnibus:
==============

Frecuencias esperadas:
---------------------
            sombrero
tipo         campana  convexo hundido  nudoso   plano
  comestible 234.238 1894.636  16.583 429.092 1633.45
  venenoso   217.762 1761.364  15.417 398.908 1518.55
Frecuencias esperadas bajo 5: 0

Resultado de la prueba ómnibus:
------------------------------

 Pearson's Chi-squared test

data:  tabla
X-squared = 485.64, df = 4, p-value < 2.2e-16
```

Figura 8.5: salida del script 8.4 que ejemplifica del uso de la función `chisq.test()` para realizar una prueba chi-cuadrado de independencia.

8.3 PROCEDIMIENTOS POST-HOC

Al estudiar las diferentes formas de la prueba chi-cuadrado de Pearson deberíamos haber notado que la hipótesis nula se refiere a la **igualdad de todas las proporciones**, mientras que la hipótesis alternativa a su negación: **no todas las proporciones son iguales**. Sin embargo, en caso de que se rechace la hipótesis nula, la prueba **no identifica cuáles son** las proporciones distintas.

A esta clase de hipótesis nula se les llama **ómnibus** o, en lenguaje menos técnico, "colectiva" o "global". Así, se dice que la prueba chi-cuadrado de Pearson es una prueba ómnibus porque tiene esta clase de hipótesis nula, que solo detecta si existe al menos un "tratamiento" (un grupo) con una proporción de "éxito" diferente a otro[3]. Sin embargo, de ser afirmativa la respuesta, no nos dice qué tratamientos presentan diferencias (Lane, s.f.).

Esto usualmente es un problema. Consideremos el estudio realizado por la SCC; si existiera heterogeneidad en las preferencias de lenguajes de programación entre hombres y mujeres, seguramente les gustaría conocer en qué lenguajes ocurre esto para estudiar a qué se debe la diferencia. En el ejemplo de la empresa de software, si algunas proporciones no son representativas de la nómina de expertos, de seguro la gerencia querría conocer dónde están las inconsistencias para corregirlas lo antes posible y evitar la movilización del sindicato. Análogamente, en el ejemplo del estudio micológico sería necesario conocer qué formas de sombrero están más asociados a hongos venenosos, de manera de poder advertir del peligro a la comunidad desde donde provienen los especímenes estudiados.

En general, el procedimiento post-hoc más directo corresponde a realizar **todas las comparaciones entre pares de tratamientos**. Si en una prueba ómnibus se tienen k tratamientos, la cantidad de comparaciones (N) que deberíamos efectuar está dada por la ecuación 8.4.

$$N = \binom{k}{2} = \frac{k(k-1)}{2} \tag{8.4}$$

[3]Hemos aprovechado de introducir algunos términos usuales en estadística inferencial para que no se confundan cuando revisen recursos externos a este apunte.

En el ejemplo del estudio micológico se consideraron cinco niveles: campana, convexo, hundido, nudoso y plano. Es decir, se necesitaría comparar 10 pares de estos tratamientos para determinar si proporciones distintas de hongos venenosos se observan al comparar los que tienen forma de campana con los que tienen forma convexa, o con los que muestran sombreros hundidos, o con los que tienen sombreros nudosos o con los parecen planos, y así con el resto de las combinaciones (convexo-hundido, convexo-nudoso, convexo-plano, hundido-nudoso, hundido-plano, nudoso-plano). Suponiendo que usamos un nivel de significación $\alpha = 0{,}05$ en cada una de estas pruebas, entonces, si estas pruebas fueran independientes, la probabilidad de no cometer un error de tipo I en ninguna prueba ($P(\text{no error})$) sería:

$$\begin{aligned} P(\text{no error}) &= (1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha)\,(1-\alpha) \\ &= (1-0{,}05)^{10} \\ &= 0{,}95^{10} \\ &\approx 0{,}599 \end{aligned}$$

Luego, la probabilidad de cometer al menos un error de tipo I es el complemento de esta probabilidad: $P(\text{error}) = 1 - P(\text{no error}) \approx 0{,}401$. O sea, si realizamos estas 10 pruebas de pares de tratamientos con las condiciones descritas, la probabilidad de obtener al menos un resultado significativo por azar es de aproximadamente ¡40,13 %!, muy lejos de la tasa nominal deseada de 5 %.

Evidentemente las pruebas de pares de tratamientos ¡no son independientes!. En el ejemplo anterior, la prueba del par campana-convexo usa parcialmente las mismas proporciones que los pares campana-hundido, campana-nudoso, campana-plano, convexo-hundido, convexo-nudoso y convexo-plano; por lo que todas ellas están correlacionadas. Sin embargo, la **inflación de la tasa de errores de tipo I** sigue ocurriendo, solo que la magnitud exacta de esta inflación es más difícil de calcular y depende de si las pruebas están correlacionadas positivamente (diferencias significativas en una prueba tiende a ir acompañada de diferencias significativas en otras) o negativamente (diferencias significativas en una prueba suelen estar acompañadas de diferencias no significativas en otras).

Es exactamente por esta razón que **no usamos pruebas múltiples** desde el principio, sino que una prueba ómnibus que está diseñada para determinar diferencias significativas en los tratamientos de **forma global**, manteniendo la tasa de error de tipo I cerca del nivel de significación α definido.

Desde luego, si una prueba ómnibus resulta significativa, existen métodos para identificar con más precisión entre qué tratamientos se dan estas diferencias, llamados **procedimientos post-hoc**, o **pruebas post-hoc** o también **pruebas a posteriori**. Reciben este nombre para remarcar que se realizan después de que se ha llegado a la conclusión que existen diferencias significativas por medio de una prueba ómnibus. Es decir, **solo haremos un procedimiento post-hoc si la prueba ómnibus rechaza la hipótesis nula** en favor de la hipótesis alternativa.

El diseño del procedimiento post-hoc depende de los datos que se están analizando y las preguntas que se quieren responder. Para el diseño más simple, realizando las N comparaciones de a pares de tratamientos, cada prueba ómnibus tiene asociada una o más pruebas post-hoc. Al aplicarlas, es importante **verificar el cumplimiento** de las condiciones de confiabilidad de estas pruebas y mantener el **mismo nivel de significación** que la prueba ómnibus durante todo el procedimiento.

Sin embargo, también es necesario **ajustar** los valores p obtenidos en estas pruebas para **corregir** la inflación de la tasa de error de tipo I y conseguir que en **su conjunto** tengan el nivel requerido. A esta probabilidad de cometer al menos un error de Tipo I en alguna de las pruebas se le conoce como la "tasa de error por familia de hipótesis" (FWER, del inglés *family-wise error rate*).

El método más sencillo para ajustar m comparaciones múltiples es la **corrección de Bonferroni**, que controla la FWER a un nivel α rechazando solamente las hipótesis nulas cuyo valor p original es inferior a α/m. Alternativamente, se pueden obtener **valores p ajustados** (p_i^{adj}, del inglés *adjusted p values*), que se comparan directamente con el nivel de significación de la forma usual, multiplicando el valor p obtenido en cada prueba por la cantidad de pruebas realizadas:

$$p_i^{\text{adj}} = m\, p_i \tag{8.5}$$

Si bien la corrección de Bonferroni es muy popular, probablemente por su simplicidad, su uso ya no es recomendado pues es considerada muy conservadora, especialmente si el número de comparaciones múltiples es alto, la mayoría hipótesis nulas son falsas o existen fuertes correlaciones positivas entre los valores p. Aquí, **muy conservadora** quiere decir que mantiene tasa de error familiar por debajo del nivel de significación requerido y, por ende, es más propensa a cometer errores tipo II (podría no detectar todas las diferencias significativas en el conjunto de pruebas múltiples).

Otra alterativa es la **corrección de Holm** (Glen, 2016b), que es una variante secuencial del método de Bonferroni que logra disminuir algo de su conservadurismo y conseguir mayor poder estadístico. Por esta razón, el método de Holm **siempre debería preferirse** a la corrección de Bonferroni (a menos que estemos realizando el ajuste ¡de forma manual! y necesitemos aprovechar su simplicidad).

El procedimiento iterativo es el siguiente: en el primer paso, se rechazan todas las hipótesis nulas con valores p menores o iguales a α/m, donde m es el número total de comparaciones. Supongamos que este paso deja m_1 hipótesis nulas sin rechazar. En el siguiente paso, se rechazan todas las hipótesis nulas con valores p menores o iguales a α/m_1, dejando m_2 hipótesis nulas sin rechazar, las cuales se prueban posteriormente al nivel α/m_2. El proceso se repite hasta que se rechazan todas las hipótesis nulas o hasta que un paso se complete sin rechazar alguna.

Una forma alternativa (más computacional) de describir el método de Holm es a través de los valores p ordenados: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Comparando cada valor $p_{(i)}$ con su valor crítico $\alpha/(m-i+1)$, el método de Holm encuentra la menor posición (k) tal que $p_{(k)}$ excede su valor crítico. Luego, se rechazan todas las hipótesis nulas cuyos valores p estén en posiciones menores a (k) (todas las que cumplen $p \leq \alpha/(m-k)$). Si no se puede encontrar tal posición (k), se rechazan todas las hipótesis.

Como para la prueba de Bonferroni, el algoritmo se puede invertir para obtener valores p ajustados que puedan ser comparados directamente con el nivel de significación α. Tal algoritmo puede encontrarse en Goeman y Solari (2014), pero en resumen los valores p ajustados se obtienen con la ecuación 8.6:

$$p_{(i)}^{\text{adj}} = \text{máx}\left\{(m-i+1)p_{(i)}, \quad p_{(i-1)}^{\text{adj}}\right\}, \text{ con } p_{(0)}^{\text{adj}} = 0 \tag{8.6}$$

Si bien los métodos de Bonferroni y Holm son útiles y generales, pues no hacen suposiciones acerca de la correlación que existe entre los valores p de la familia de hipótesis, su rigurosidad para evitar errores de tipo I a toda costa los hacen inadecuados para aplicaciones donde ocurren grandes cantidades de pruebas simultáneamente, como al estudiar datos genéticos o imágenes médicas, donde producen un gran número de errores de tipo II. Esto ha incentivado el desarrollo de métodos que no intentan controlar la FWER, sino que la tasa de falsos descubrimientos (FDR, del inglés *false discovery rate*), que corresponde a la **proporción esperada** de errores de tipo I entre todas las hipótesis nulas que han sido rechazadas, es decir la proporción de pruebas declaradas significativas, "descubrimientos", de forma errónea (también llamados **falsos positivos**).

Una de las correcciones basadas en FDR más conocidas es el procedimiento de Benjamini y Hochberg (1995), que presenta mayor o igual potencia estadística que los métodos basados en FWER, especialmente cuando hay numerosas hipótesis nulas falsas. Sin embargo, este método asume que los valores p de la familia son independientes o tienen una "dependencia de regresión positiva". Si bien una definición detallada de esta condición está fuera de los alcances de este libro, diremos que, en palabras muy simplistas, significa que si el valor p de una de las hipótesis nulas verdaderas fuera más alto de lo esperado, esto no debería hacer que los valores p de otras hipótesis nulas verdaderas tiendan a ser más bajos, al menos en términos de cómo afectan a funciones estrictamente crecientes que se les apliquen (como la suma acumulativa o el valor máximo). El procedimiento consiste en ordenar los m valores p de menor a mayor: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ y rechazar todas las hipótesis nulas cuyos valores p están en posiciones menores o iguales a (k) tal que:

$$k = \max_{i \in \{1,\ldots,m\}} \left\{p_{(i)} \leq \frac{i}{m}\alpha\right\} \tag{8.7}$$

Si no existe tal posición (k), no se rechaza ninguna de las m hipótesis nula.

Basado en esta idea, Benjamini y Yekutieli (2001) propusieron un cambio en la ecuación 8.7:

$$k = \max_{i \in \{1,\ldots,m\}} \left\{ p_{(i)} \leq \frac{i}{m} \frac{\alpha}{\sum_{i=1}^{m} \frac{1}{i}} \right\} \tag{8.8}$$

Con este cambio, el procedimiento permite controlar la FDR al nivel $(m_0/m)\,\alpha$, donde m_0 es el (desconocido) número de hipótesis nulas verdaderas, sin hacer ninguna suposición sobre la dependencia de la familia de valores p, aunque evidentemente es más conservador que el procedimiento original porque los valor críticos se reducen por un factor $\sum_{i=1}^{m}(1/i)$ (Goeman & Solari, 2014).

En la comunidad de R existen implementaciones de estos y otros métodos para ajustar pruebas múltiples. Por ejemplo se podrían mencionar `multtest`, `mutoss`, `pairwiseComparisons`, `fdrtool`, `qvalue`, `MCPAN`, `sgof`, `multcomp`, `emmeans`, entre otros. Sin embargo, por ahora nos quedaremos con la implementación en el paquete base `stats` que siempre es cargado en el entorno. Este paquete define la constante `p.adjust.methods` que corresponde a un vector de strings con los nombres de los métodos disponibles. Los métodos descritos anteriormente se asocian, respectivamente, a los nombres `"bonferroni"`, `"holm"`, `"BH"` o su alias `"fdr"` y `"BY"`.

Estos nombres se usan en el argumento `method` de la función `p.adjust(p, method)`. El argumento `p` corresponde al vector de valores p no ajustados. La función devuelve un vector del mismo largo con los p valores ajustados con el método especificado. Por omisión, se aplica el método de Holm.

8.3.1 Procedimiento post-hoc para la prueba chi-cuadrado

Para la prueba chi-cuadrado de Pearson, el procedimiento post-hoc más simple consiste en efectuar **múltiples pruebas chi-cuadrado** entre cada par de tratamientos, si se verifican las condiciones, o pruebas exactas de Fisher, si las frecuencias esperadas se reducen mucho. El script 8.5 muestra una implementación de esta idea, cuyo resultado se muestra en la figura 8.6.

```
Procedimiento post-hoc:
======================

   Forma 1 Forma 2        Prueba Estadístico  Valor p P.adj_Holm   P.adj_BY
1  campana convexo  Chi cuadrado      214.18 1.68e-48   1.34e-47 * 1.64e-47 *
2  campana hundido Ex. de Fisher        0.00 6.09e-02   6.09e-02   1.78e-01
3  campana  nudoso  Chi cuadrado      447.39 2.67e-99   2.67e-98 * 7.81e-98 *
4  campana   plano  Chi cuadrado      240.29 3.41e-54   3.07e-53 * 4.99e-53 *
5  convexo hundido  Chi cuadrado       27.85 1.31e-07   3.94e-07 * 4.81e-07 *
6  convexo  nudoso  Chi cuadrado      179.15 7.41e-41   5.19e-40 * 5.43e-40 *
7  convexo   plano  Chi cuadrado        4.75 2.92e-02   5.84e-02   9.51e-02
8  hundido  nudoso  Chi cuadrado       76.70 1.99e-18   9.95e-18 * 9.71e-18 *
9  hundido   plano  Chi cuadrado       30.90 2.72e-08   1.09e-07 * 1.14e-07 *
10  nudoso   plano  Chi cuadrado      140.92 1.67e-32   1.00e-31 * 9.79e-32 *
Warning message:
In chisq.test(tabla[, pares[i, ]], correct = F) :
  Chi-squared approximation may be incorrect
```

Figura 8.6: salida del script 8.5 que aplica un procedimiento post-hoc para una prueba chi-cuadrado de independencia.

Podemos ver que al comparar las proporciones de todos los pares de formas de sombreros de los hongos estudiados, ajustando con el método de Benjamini y Yekutieli, se encuentran diferencias significativas entre casi todos los casos, con la excepción de los pares campana-hundido ($OR = 0{,}00;\ p = 0{,}180$) y convexo-plano ($\chi^2(1) = 4{,}75;\ p = 0{,}095$).

Debemos notar la advertencia que nos entrega el entorno R al ejecutar el script 8.5 (que aparece en color rojo en la figura 8.6). Esto se debe a que en una de las pruebas post-hoc, específicamente entre las formas de sombrero campana y hundido, la cantidad de observaciones esperadas es menor a 5, por lo que no debemos confiar en los resultados de la prueba. Es por esta razón que el script busca estos casos y los reemplaza por pruebas exactas de Fisher (líneas 53–59).

Script 8.5: (continuación del script 8.4) procedimiento post-hoc para la prueba χ^2 de independencia.

```
# Realizar procedimiento post-hoc si corresponde
if(prueba[["p.value"]] < alfa)
{
  cat("Procedimiento post-hoc:\n")
  cat("=====================\n\n")

  # Obtener los pares de formas de sombrero
  pares <- t(combn(colnames(tabla), 2))
  N <- nrow(pares)

  # Obtener las pruebas post-hoc
  pruebas_ph <- sapply(1:N, function(i) chisq.test(tabla[, pares[i, ]],
                                                   correct = FALSE),
                       simplify = FALSE)

  # Identificar las pruebas chi-cuadrado post-hoc que no cumplen condiciones
  i_no <- which(sapply(1:N, function(i) min(pruebas_ph[[i]][["expected"]]) <
                                        minima_frec_esperada))

  # Cambiar la prueba post-hoc cuando no se cumplen las condiciones
  if(length(i_no) > 0)
    pruebas_ph[i_no] <- sapply(i_no, function(i) fisher.test(tabla[, pares[i,]]),
                               simplify = FALSE)

  # Preparar tabla con el resumen de resultados
  nombre_ph <- rep("Chi cuadrado", N)
  nombre_ph[i_no] <- "Ex. de Fisher"
  estadistico <- sapply(1:N, function(i) pruebas_ph[[i]][["statistic"]])
  estadistico[i_no] <- sapply(i_no, function(i) pruebas_ph[[i]][["estimate"]])
  estadistico <- round(unlist(estadistico), 2)
  p_val <- sapply(1:N, function(i) pruebas_ph[[i]][["p.value"]])
  p_adj_h <- p.adjust(p_val, method = "holm")
  p_adj_by <- p.adjust(p_val, method = "BY")
  sig_h <- ifelse(p_adj_h < alfa, "*", " ")
  sig_by <- ifelse(p_adj_by < alfa, "*", " ")

  resultados <- data.frame(pares[, 1], pares[, 2], nombre_ph, estadistico)
  resultados <- cbind(resultados, p_val,  p_adj_h, sig_h, p_adj_by, sig_by)
  colnames(resultados) <- c("Forma 1", "Forma 2", "Prueba", "Estadístico",
                            "Valor p", "P.adj_Holm", "", "P.adj_BY", "")

  # Mostrar resumen de resultados
  print(resultados, digits = 3)
}
```

Como es usual, existen paquetes de R que implementan procedimientos post-hoc específicos para la prueba chi-cuadrado de Pearson. Por ejemplo, la función `chisq.posthoc.test()` del paquete homónimo, o la función `pairwiseNominalIndependence()` del paquete `rcompanion`, o la función `pairwise.table()` del paquete `RVAideMemoire`, etc.

Más aún, si se dan las condiciones, también podríamos usar **múltiples pruebas de proporciones** para el análisis post-hoc. Si bien no es común realizar pruebas paramétricas en un procedimiento post-hoc para una

prueba ómnibus no paramétrica, esto tendría la ventaja de entregar estimaciones e intervalos de confianza para las diferencias entre las proporciones observadas para los tratamientos. El paquete base `stats` también proporciona la función `pairwise.prop.test(x, n, p.adjust.method)` que realiza la comparación de todos los pares de proporciones, pero solo reporta los valores p ajustados. Para una opción un poco más moderna, el paquete `rstatix` proporciona funciones al estilo `tidyverse` que permiten crear *pipelines* para aplicar tanto pruebas ómnibus como pruebas post-hoc por pares de tratamientos usando múltiples chi-cuadrado, exactas de Fisher o diferencia de proporciones, todo mediante simulaciones que permiten estimaciones más confiables. Sin embargo, estas funciones tampoco entregan detalles sino que los valores p originales y los ajustados. Así, si queremos intervalos de confianza debemos aplicar una versión manual como presentada en el script 8.5.

8.4 PRUEBA DE MCNEMAR

Las dos pruebas anteriores requieren **muestras independientes** para comparar las poblaciones subyacentes. En esta sección se considera el análisis de **frecuencias apareadas**, es decir cuando una misma característica con respuesta dicotómica se mide en dos ocasiones (o situaciones diferentes) para **el mismo grupo de casos**.

En estas condiciones, se recurre a la **prueba de McNemar** que permite determinar si se produce o no un **cambio significativo en las proporciones observadas** entre ambas mediciones.

Una vez más, podemos registrar las frecuencias en una matriz de confusión como la que vimos en la tabla 8.1. En ella, bajo este contexto, podemos reconocer que las celdas (a) y (d) corresponde a instancias en que no hay cambios, mientras que la celda (b) 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 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 ($\mu = (b+c)/2$) y desviación estándar $\sigma = \sqrt{(b+c)/4}$, 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 la condición que existan 10 o más cambios entre las mediciones), el cual sigue una distribución χ^2 con un grado de libertad y se calcula como muestra la ecuación 8.9 (Agresti, 2019):

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{8.9}$$

Puesto que los datos siguen una distribución binomial que es discreta, pero se está usando como aproximación la distribución chi-cuadrado que es continua, suele emplearse un **factor de corrección de continuidad** como el propuesto por Allen L. Edwards en 1948 (Pembury Smith & Ruxton, 2020). El estadístico de prueba con esta corrección se calcula como muestra la ecuación 8.10:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{8.10}$$

Es más, se ha sugerido que si b o c es muy pequeño o si $b+c < 25$, la potencia estadística de la prueba puede ser baja, incluso si el tamaño de la muestra es grande. En estos casos, podría ser mejor no utilizar la prueba asintótica de McNemar, sino que una alternativa **exacta** (Hazra & Gogtay, 2016). Sin embargo, recientes simulaciones que consideraron casi 10.000 escenarios distintos, mostraron que la versión original (ecuación 8.9) se comporta sorprendentemente bien, sin nunca sobrepasar una tasa de error de tipo I de 5,37 % para un nivel de significación $\alpha = 0{,}05$ (Fagerland et al., 2013). Más aún, este mismo estudio sugiere

nunca utilizar la versión con corrección de continuidad ni la versión exacta (condicional) de la prueba, sino que utilizar el **valor mid-p** que se calcula como el valor p exacto menos la mitad de la probabilidad puntual del estadístico de prueba observado. Sin embargo, dejaremos esta alternativa para otra versión del libro porque aún no hay muchas implementaciones directas en software estadísticos (en R hay al menos una en el paquete `contingencytables`).

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 8.10 y se resume en la matriz de confusión de la tabla 8.11.

Estudiante	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 8.10: 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 8.11: tabla de contingencia con las predicciones de los resultados finales de los estudiantes.

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 (sin corrección) obtiene:

$$\chi^2 = \frac{(5-7)^2}{5+7} \approx 0{,}333$$

El valor p está dado por el área bajo la cola superior de la distribución chi-cuadrado con un grado de libertad, que en R puede calcularse como `pchisq(0.333, 1, lower.tail = FALSE)`, obteniéndose que $p = 0{,}564$. En consecuencia, se falla al rechazar la hipótesis nula (para un nivel de significación $\alpha = 0,05$) y se concluye que no hay evidencia suficiente para creer que existe una diferencia en el desempeño de ambos clasificadores.

La función de R para esta prueba es `mcnemar.test(x, y = NULL, correct = FALSE)`. Cuando `y` se omite o tiene valor `NULL`, `x` debe indicar la tabla de contingencia usada en la prueba. En caso contrario, `x` e `y` especifican las muestras con los pares de observaciones de la variable dicotómica de interés. El script 8.6 muestra su aplicación para el ejemplo dado. Notemos que, por omisión, la función aplica el factor de corrección de continuidad, por lo que se debe explicitar `correct = FALSE` para conseguir la versión de la ecuación 8.9.

La figura 8.7 presenta la salida entregada por el script 8.6. Podemos ver que el resultado de la función coincide con el cálculo manual hecho más arriba.

Script 8.6: prueba de McNemar.

```r
# Construir la tabla de contingencia
estudiante <- 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(estudiante, modelo_2, modelo_1)
tabla <- table(modelo_2, modelo_1)

# Mostrar la tabla
cat("Tabla de contingencia:\n")
cat("----------------------\n")
print(tabla)

# Aplicar la prueba de McNemar a la tabla y mostrarla
prueba_1 <- mcnemar.test(tabla, correct = FALSE)
cat("\nPrueba de McNemar sobre la tabla de contingencia:\n")
cat("-------------------------------------------------\n")
print(prueba_1)

# Pero también se puede aplicar directamente a las muestras
prueba_2 <- mcnemar.test(modelo_2, modelo_1, correct = FALSE)
cat("Prueba de McNemar sobre las muestras:\n")
cat("-----------------------------------\n")
print(prueba_2)
```

```
            modelo_1
Tabla de contingencia:
----------------------
            modelo_1
modelo_2     Correcto Incorrecto
  Correcto          9          5
  Incorrecto        7          4

Prueba de McNemar sobre la tabla de contingencia:
-------------------------------------------------

 McNemar's Chi-squared test

data:  tabla
McNemar's chi-squared = 0.33333, df = 1, p-value =
0.5637

Prueba de McNemar sobre las muestras:
-----------------------------------

 McNemar's Chi-squared test

data:  modelo_2 and modelo_1
McNemar's chi-squared = 0.33333, df = 1, p-value =
0.5637
```

Figura 8.7: salida del script 8.6 que ejemplifica del uso de la función `mcnemar.test()`.

8.5 PRUEBA Q 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 apareadas** (cuando

ambas variables son dicotómicas, esta prueba es equivalente a la de McNemar).

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 8.12, donde la primera columna identifica cada una de las 15 instancias del problema empleadas para evaluar las metaheurísticas, mientras que las columnas restantes 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 8.12: resultados de las metaheurísticas para cada una de las instancias usadas en su evaluación.

Las hipótesis contrastadas por la prueba Q de Cochran son que la proporción de "éxitos" es la misma (o no) en todas las mediciones. Para el ejemplo de Elsa:

H_0: la proporción de instancias en que se encuentra la solución óptima es la misma para todas las metaheurísticas.

H_A: la proporción de instancias en que se encuentra la solución óptima es distinta para al menos una de las metaheurísticas.

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 $n_\text{b}\ n_\text{t} \geq 24$, donde n_b es el número de "bloques" en que se organizan las observaciones y n_t es la cantidad de tratamientos estudiados.

Ya sabemos lo que son los tratamientos, pero debemos introducir otro término muy usado en estadística: un **bloque** corresponde a una agrupación de unidades experimentales, es decir **casos**, que son **similares** en términos de una o más características, que pueden ser consideradas que presentan el **mismo resultado** en un estudio. En su forma más simple, un bloque corresponde a un caso, que es sometido a las distintas mediciones. Esto es lo que ocurre en el ejemplo: una misma instancia del problema del vendedor viajero es resuelta por las diferentes metaheurísticas en estudio.

Evaluemos si se cumplen las condiciones para aplicar la prueba Q de Cochran al ejemplo. La variable de respuesta es si la metaheurística consigue (1) o no (0) la solución óptima para la instancia, por lo que es dicotómica. La variable independiente (tratamientos) corresponde a las metaheurísticas utilizadas, que es categórica con tres niveles (simulated annealing, colonia de hormigas y algoritmo genético). Si una metaheurística consigue o no consigue encontrar la solución óptima para una determinada instancia no debería influir en qué consigue en otra instancia ni en cómo les va en la búsqueda a las otras metaheurísticas, por lo que las observaciones son independientes entre sí (considerando que probablemente hay infinitas instancias

y solo se han escogido 15, mucho menos del 10 % de ellas). Por último, la tabla muestra 15 bloques y 3 tratamientos, con lo que la muestra posee $15 \cdot 3 = 45 \geq 24$ observaciones. Así, el estudio del ejemplo cumple con las condiciones para confiar en los resultados de la prueba.

El estadístico de prueba se calcula como muestra la ecuación 8.11:

$$Q = \frac{\sum_{j=1}^{n_\text{t}} \left(x_{\bullet j} - \frac{N}{n_\text{t}} \right)^2}{\sum_{i=1}^{n_\text{b}} x_{i\bullet}(n_\text{t} - x_{i\bullet})} \, k \, (k-1) \tag{8.11}$$

donde:

- n_b: cantidad de bloques.
- n_t: cantidad de tratamientos.
- $x_{\bullet j}$: total de éxitos en la columna del j-ésimo tratamiento.
- $x_{i\bullet}$: total de éxitos en la fila del i-ésimo bloque.
- N: número total de éxitos.

Podemos ver que los cálculos necesarios para esta prueba son tediosos, por lo que suele hacerse mediante software. En R, esta prueba está implementada en la función `cochran.qtest(formula, data)` del paquete `RVAideMemoire`, donde:

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

El script 8.7 presenta el uso de esta función con el ejemplo de Elsa Capunta. Al ejecutar el script, se obtiene el resultado que muestra la figura 8.8. Vemos que la prueba resulta significativa ($Q(2) = 7{,}167$; $p = 0{,}028$) al nivel de significación establecido, por lo que se ha de rechazar 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.

Debemos notar que la prueba Q de Cochran, como la prueba chi-cuadrado, considera una **hipótesis nula no paramétrica ómnibus**. Por lo tanto, en caso de tener un resultado significativo, es necesario aplicar un procedimiento post-hoc para poder determinar dónde se encuentran las diferencias.

Para la prueba Q de Cochran, la prueba post-hoc más utilizada es la prueba de McNemar. En R la función `pairwiseMcnemar(formula, data, method)` del paquete `rcompanion` permite aplicar esta prueba a cada par de tratamientos, donde `formula` y `data` son las mismas que para la prueba Q de Cochran y `method` permite indicar el método para ajustar los valores p múltiples (definidos en la mencionada constante `p.adjust.methods`).

No obstante, como muestra la figura 8.8, la función `cochran.qtest()` usada en el script 8.7, parece optar por otra alternativa: aplicar a cada par de tratamientos la prueba de rangos con signo de Wilcoxon ("Wilcoxon sign test"). Estudiaremos esta prueba más adelante, pero diremos aquí que es *extraño* que se aplique en este contexto puesto que se usa para comparar variables numéricas, no binarias.

Sin embargo, revisando el código de la función (una ventaja de usar entornos de código abierto como R) ¡no es eso lo que hace! La función en realidad está realizando múltiples **pruebas exactas condicionales de McNemar** por medio de la función más general `binom.test(b, b + c, 0.5)` para hipótesis nulas sobre la probabilidad de éxito en ensayos de Bernoulli. Es decir, el mensaje es incorrecto, pero el procedimiento aplicado no lo es. Aunque debe considerarse que el uso de esta prueba exacta actualmente está siendo desaconsejada (Fagerland et al., 2013), como se mencionó en la sección 8.4, por lo que tampoco podríamos considerarla ideal.

Dicho esto, mirando el reporte en la figura 8.8 podemos ver que al comparar los pares de metaheurísticas, ajustando con el método de Benjamini y Hochberg (`fdr`), no se encuentran diferencias significativas al nivel $\alpha = 0{,}05$ ($p > 0{,}093$), lo que **¡no es consistente** con el resultado de la prueba ómnibus!

```
Prueba global:
==============

Tamaño de la muestra: 45  > 24

Resultado de la prueba Q de Cochran:
-----------------------------------

 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 8.8: resultado del script 8.7 que ejemplifica el uso de la función `cochran.qtest()`.

Script 8.7: prueba Q de Cochran.

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

# Crear la matriz de datos
instancia <- 1:15
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_anchos <- data.frame(instancia, annealing, hormigas, genetico)

# Llevar la matriz de datos a formato largo
datos_largos <- datos_anchos |>
                pivot_longer(c("annealing", "hormigas", "genetico"),
                             names_to = "metaheuristica", values_to = "resultado")
datos_largos[["instancia"]] <- factor(datos_largos[["instancia"]])
datos_largos[["metaheuristica"]] <- factor(datos_largos[["metaheuristica"]])

# Definir condiciones y el nivel de significación
minimo_tamaño_muestra <- 24
alfa <- 0.05

cat("Prueba global:\n")
cat("==============\n\n")

# Mostrar tamaño de la muestra
N <- nrow(datos_anchos[, -1]) * ncol(datos_anchos[, -1])
cat("Tamaño de la muestra:", N, " ")
cat(ifelse(N > minimo_tamaño_muestra, "> ", "<= "))
cat(minimo_tamaño_muestra, "\n")

```

```r
# Realizar la prueba Q de Cochran
prueba <- cochran.qtest(resultado ~ metaheuristica | instancia,
                        data = datos_largos, alpha = 0.05)

# Mostrar el resultado de la prueba
cat("\nResultado de la prueba Q de Cochran:\n")
cat("-----------------------------------\n")
print(prueba)
```

Siguiendo la recomendación de Fagerland et al. (2013), el script 8.8 presenta procedimientos post-hoc mediante pruebas de asintóticas de McNemar sin corrección de continuidad aplicandolas correcciones de Holm y de Benjamini y Yekutieli, obteniéndose la salida de la figura 8.9.

```
Procedimiento post-hoc:
======================

Procedimiento post-hoc con ajuste de Holm
-----------------------------------------
$Test.method
     Test
1 mcnemar

$Adustment.method
  Method
1   holm

$Pairwise
                Comparison chi.sq df p.value p.adjust
1 annealing - genetico = 0   4.45  1  0.0348   0.0696
2 annealing - hormigas = 0  0.143  1   0.705   0.7050
3  genetico - hormigas = 0      6  1  0.0143   0.0429


Procedimiento post-hoc con ajuste Benjamini y Yekutieli
-------------------------------------------------------
$Test.method
     Test
1 mcnemar

$Adustment.method
  Method
1     BY

$Pairwise
                Comparison chi.sq df p.value p.adjust
1 annealing - genetico = 0   4.45  1  0.0348   0.0957
2 annealing - hormigas = 0  0.143  1   0.705   1.0000
3  genetico - hormigas = 0      6  1  0.0143   0.0786
```

Figura 8.9: resultados de los procedimientos post-hoc para la prueba Q de Cochran de la figura 8.8.

Podemos ver que solo utilizando el ajuste de Holm se logra detectar una posible diferencia significativa entre las proporciones de las metaheurísticas basadas en colonias de hormigas y algoritmos genéticos ($\chi^2(1) = 6{,}0;\ p = 0{,}043$).

En consecuencia, la respuesta que Elsa debe dar a su profesor es que la evidencia no es lo suficientemente clara para poder afirmar que existen diferencias entre las metaheurísticas, y podría ser adecuado hacer un estudio con una muestra de instancias mayor, puesto que los resultados de la prueba Q de Cochran y de los procedimientos post-hoc son mayoritariamente contradictorios.

Script 8.8: (continuación del script 8.7) procedimiento post-hoc para la prueba Q de Cochran.

```
# Realizar procedimiento post-hoc si corresponde
if(prueba[["p.value"]] < alfa)
{
  cat("\n\nProcedimiento post-hoc:\n")
  cat("======================\n")

  # Procedimiento post-hoc con corrección de Holm
  post_hoc_1 <- pairwiseMcnemar(resultado ~ metaheuristica | instancia,
                                data = datos_largos,
                                test = "mcnemar", correct = FALSE, method = "holm")

  cat("\nProcedimiento post-hoc con ajuste de Holm\n")
  cat("----------------------------------------\n")
  print(post_hoc_1)

  # Procedimiento post-hoc con corrección de Benjamini y Yekutieli
  post_hoc_2 <- pairwiseMcnemar(resultado ~ metaheuristica | instancia,
                                data = datos_largos,
                                test = "mcnemar", correct = FALSE, method = "BY")

  cat("\nProcedimiento post-hoc con ajuste Benjamini y Yekutieli\n")
  cat("------------------------------------------------------\n")
  print(post_hoc_2)
}
```

Cerremos diciendo que, como ha sido la tónica, existen numerosas implementaciones de la prueba Q de Cochran y procedimientos post-hoc para ella en R, como por ejemplo en el ya mencionado paquete `rstatix`.

8.6 EJERCICIOS PROPUESTOS

8.1 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba exacta de Fisher. Explica bien qué variables están involucradas y enuncia las hipótesis a docimar.

8.2 Para la situación anterior, extiende la pregunta de investigación de forma que requiera usar una prueba χ^2 de independencia.

8.3 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera una prueba de McNemrar. Explica bien qué variables están involucradas y enuncia las hipótesis a docimar.

8.4 Para la situación anterior, extiende la pregunta de investigación de forma que requiera usar una prueba Q de Cochran. Explica cómo se verían los datos recogidos en este caso.

8.5 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera una prueba χ^2 de homogeneidad. Explica bien las variable involucradas y enuncia las hipótesis a docimar.

8.6 Plantea la pregunta de investigación de tu ejemplo anterior para una prueba χ^2 de bondad de ajuste. ¿Cuál versión parece más natural?

8.7 Un estudio clínico reclutó a 32 pacientes con fatiga crónica para determinar si un tratamiento basado en inyecciones intramusculares de magnesio es efectivo para esta condición. De los 15 pacientes que recibieron estas inyecciones, seleccionados de manera aleatoria, 12 reportaron sentirse mejor (80 %), mientras que solo 3 pacientes de los 17 que recibieron inyecciones placebo (18 %) reportaron mejorías.

(a) ¿Se cumplen las condiciones para aplicar una prueba exacta de Fisher al problema enunciado?

(b) ¿Cuáles serían las hipótesis nula y alternativa para esta prueba?

(c) Independientemente de la respuesta anterior, aplica la prueba usando R.

(d) ¿A qué conclusión lleva este procedimiento?

8.8 Para la situación anterior, aplica la prueba exacta de Fisher de forma manual (ayuda: hay 16 tablas que mantienen los totales marginales del enunciado).

8.9 Antes del debate de candidatos presidenciales por televisión abierta, una encuesta consultó a 1.000 personas si apoyaban o no la legalización del aborto libre, encontrando 705 personas a favor y 295 en contra. Luego de que estas personas escucharon el debate, 663 se manifestaron a favor y 337 en contra de la propuesta legal. 73 encuestados cambiaron de opinión de en contra, a en apoyo de la ley; mientras que 115 cambiaron su opinión a favor, para estar ahora en contra.

(a) ¿Se cumplen las condiciones para aplicar una prueba de McNemar al problema enunciado?

(b) ¿Cuáles serían las hipótesis nula y alternativa si usamos esta prueba?

(c) Independientemente de la respuesta anterior, aplica la prueba usando R.

(d) ¿A qué conclusión lleva este procedimiento?

8.10 Con palabras propias, explica qué es una prueba ómnibus, qué es una prueba post-hoc y cuándo se aplican. Da ejemplos que clarifiquen tus definiciones.

8.11 Con palabras propias, explica por qué cuando se estudian más de dos alternativas es problemático hacer múltiples pruebas entre pares de ellas. Da ejemplos que clarifiquen tu explicación.

8.12 Las autoridades de la universidad desean conocer si las semanas de receso (sin actividades docentes) ayuda o no al descanso del estudiantado. Para eso seleccionaron 20 estudiantes de forma aleatoria y les consultaron si se sentían "cansada/o" o "descansada/o" en tres ocasiones: el lunes, miércoles y viernes de la primera semana de receso del semestre. Los resultados se muestran en la siguiente tabla, donde 0 representa cansancio y 1 descanso.

Estudiante	Lunes	Miércoles	Viernes	Estudiante	Lunes	Miércoles	Viernes
1	1	1	1	11	1	1	0
2	0	1	1	12	1	1	1
3	0	0	1	13	0	0	0
4	0	1	0	14	1	0	1
5	1	0	0	15	0	1	1
6	0	1	1	16	0	1	0
7	0	1	1	17	0	0	1
8	0	0	1	18	0	1	1
9	0	1	1	19	1	0	1
10	0	1	0	20	0	1	1

(a) ¿Hay diferencias entre los tres periodos de tiempo sin actividades? No olvide enunciar las hipótesis, seleccionar una prueba para docimarlas y verificar si se cumplen las condiciones necesarias para realizar la prueba seleccionada.

(b) Si hay diferencias, ¿entre qué periodos se encuentran? No olvide justificar su respuesta.

8.7 BIBLIOGRAFÍA DEL CAPÍTULO

Agresti, A. (2019). *An introduction to categorical data analysis* (3.ª ed.). John Wiley & Sons, Inc.

Benjamini, Y., & Hochberg, Y. (1995).
Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001).
The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Diez, D., Barr, C. D., & Çetinkaya-Rundel, M. (2017). *OpenIntro Statistics* (3.ª ed.).
https://www.openintro.org/book/os/.

Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, *13*, 1-8.

Glen, S. (2016a). *Cochran's Q Test*.
Consultado el 9 de octubre de 2021, desde https://www.statisticshowto.com/cochrans-q-test/

Glen, S. (2016b). *Holm-Bonferroni Method: Step by Step*. Consultado el 7 de mayo de 2021, desde https://www.statisticshowto.com/holm-bonferroni-method/

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946-1978.

Hazra, A., & Gogtay, N. (2016). Biostatistics series module 4: comparing groups–categorical variables. *Indian journal of dermatology*, *61*(4), 385-392.

Lane, D. (s.f.). *Online Statistics Education: A Multimedia Course of Study*.
Consultado el 4 de mayo de 2021, desde https://onlinestatbook.com/

Mangiafico, S. S. (2016). *Cochran's Q Test for Paired Nominal Data*.
Consultado el 9 de octubre de 2021, desde https://rcompanion.org/handbook/H_07.html

NIST/SEMATECH. (2013). *e-Handbook of Statistical Methods*.
Consultado el 29 de abril de 2021, desde http://www.itl.nist.gov/div898/handbook/

Pembury Smith, M. Q., & Ruxton, G. D. (2020). Effective use of the McNemar test.
Behavioral Ecology and Sociobiology, *74*, 1-9.

Pértega, S., & Pita, S. (2004).
Asociación de variables cualitativas: El test exacto de Fisher y el test de Mcnemar. Consultado el 29 de abril de 2021, desde https://www.fisterra.com/mbe/investiga/fisher/fisher.asp#Mcnemar