		Conclusión de la prueba	
		No rechazar H_0	Rechazar H_0 en favor de H_A
Verdad	H_0 verdadera	Decisión correcta	Error tipo I
	H_A verdadera	Error tipo II	Decisión correcta

Tabla 4.1: posibles escenarios para una prueba de hipótesis.

Intervalos de confianza. El valor p. Pruebas unilaterales / bilaterales.

Un **estadístico de prueba** es un estadístico de resumen que resulta especialmente útil para evaluar hipótesis o calcular el valor p. El valor z se usa cuando el estimador puntual se acerca a la normalidad, aunque existen otros estadísticos de prueba adecuados para otros escenarios.

4.5.2 El efecto del nivel de significación

Hemos visto que el nivel de significación (α) representa la proporción de veces en que se cometería un error de tipo I (es decir, rechazar H_0 en favor de H_A, cuando H_0 es en realidad verdadera). Si resulta costoso o peligroso cometer un error de este tipo, debemos requerir evidencia más fuerte para rechazar la hipótesis nula (es decir, reducir la probabilidad de que esto ocurra), lo que podemos lograr usando un valor más pequeño para el nivel de significación, por ejemplo, $\alpha = 0,01$. Sin embargo, esto necesariamente **aumentará** la probabilidad de cometer un error de tipo II.

Si, por el contrario, el costo o el peligro de cometer un error de tipo II (no rechazar H_0 cuando en realidad H_A es verdadera) es mayor, debemos escoger un nivel de significación más elevado (por ejemplo, $\alpha = 0,10$).

Así, **el nivel de significación seleccionado para una prueba siempre debe reflejar las consecuencias de cometer errores de tipo I o de tipo II**.

El método de prueba de hipótesis usando valores p puede generalizarse para otros estimadores puntuales con distribución cercana a la normal. Para ello, Diez y col. (2017, p. 199) señalan que se debemos considerar los siguientes pasos:

Prueba de hipótesis usando el modelo normal:

1. Formular las hipótesis nula (H_0) y alternativa (H_A) en lenguaje llano y luego en notación matemática.
2. Identificar un estimador puntual (estadístico) adecuado e insesgado para el parámetro de interés.
3. Verificar las condiciones para garantizar que la estimación del error estándar sea razonable y que la distribución muestral del estimador puntual siga aproximadamente una distribución normal.
4. Calcular el error estándar. Luego, graficar la distribución muestral del estadístico bajo el supuesto de que H_0 es verdadera y sombrear las áreas que representan el valor p.
5. Usando el gráfico y el modelo normal, calcular el valor p para evaluar las hipótesis y escribir la conclusión en lenguaje llano.

4.6.2 Estimadores con otras distribuciones

Existen métodos de construcción de intervalos de confianza y prueba de hipótesis adecuados para aquellos casos en que el estimador puntual o el estadístico de prueba no son cercanos a la normal (por ejemplo, si la muestra es pequeña, se tiene una mala estimación del error estándar o el estimador puntual tiene una distribución distinta a la normal). No obstante, la selección de métodos alternativos debe hacerse siempre teniendo en cuenta la distribución muestral del estimador puntual o del estadístico de prueba.

Una consideración importante es que **siempre debemos verificar el cumplimiento de las condiciones requeridas por una herramienta estadística**, pues de lo contrario las conclusiones pueden ser erradas y carecerán de validez.

Capítulo 5: Inferencia con medias muestrales.

5.1 PRUEBA Z

Como ya adelantamos, la prueba Z es adecuada para inferir acerca de las medias con una o dos muestras, aunque aquí solo veremos el primer caso. Para poder usarla, debemos **verificar el cumplimiento** de algunas condiciones, muchas de las cuales están asociadas al modelo normal que conocimos en el capítulo anterior:

- La muestra debe tener al menos 30 observaciones. Si la muestra tiene menos de 30 observaciones, se debe conocer la varianza de la población.
- Las observaciones deben ser independientes, es decir que la elección de una observación para la muestra no influye en la selección de las otras.
- La población de donde se obtuvo la muestra sigue aproximadamente una distribución normal.

Esta prueba resulta adecuada si queremos **asegurar** o **descartar** que la media de la población tiene un cierto **valor hipotético**. Supongamos que queremos saber si, en promedio, las utilidades mensuales de una

H_0: la media de las utilidades mensuales de la empresa (μ) es de 20 millones de pesos, es decir: $\mu = 20$ [M$].

H_A: las utilidades mensuales de la empresa son, en promedio, distintas de 20 millones de pesos, es decir: $\mu \neq 20$ [M$].

5.2 PRUEBA T DE STUDENT

En la práctica, rara vez podemos conocer la desviación estándar de la población y a menudo nos encontraremos con muestras pequeñas, por lo que la prueba Z no es muy utilizada.

La prueba t de Student, basada en la distribución t, es en consecuencia la alternativa más ampliamente empleada para inferir acerca de una o dos medias muestrales.

5.2.1 Prueba t para una muestra

Aunque la prueba t no opera bajo el supuesto de normalidad, aún así requiere verificar algunas condiciones para poder usarla:

1. Las observaciones son independientes entre sí.
2. Las observaciones provienen de una distribución cercana a la normal.

El primer paso es formular las hipótesis:

H_0: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es igual a 500 milisegundos.
H_A: el tiempo promedio que tarda el algoritmo en resolver una instancia del problema es inferior a 500 milisegundos.

Recordemos que μ_0 es el valor nulo, por lo que en este caso $\mu_0 = 500$ [ms]. Matemáticamente, las hipótesis anteriores pueden formularse como:

Denotando como μ al tiempo medio que tarda la implementación del algoritmo en resolver una instancia cualquiera del problema:
H_0: $\mu = \mu_0$, esto es $\mu = 500$
H_A: $\mu < \mu_0$, es decir $\mu < 500$

5.2.2 Prueba t para dos muestras pareadas

Para este ejemplo, tenemos dos tiempos de ejecución diferentes para cada instancia del problema: uno con cada algoritmo. En consecuencia, los datos están **pareados**. Es decir, cada observación de un conjunto tiene una correspondencia o conexión especial con exactamente una observación del otro. Una forma de uso común para examinar datos pareados es usar la diferencia entre cada par de observaciones, para lo cual podemos usar la técnica de la distribución t (también llamada prueba t de Student) vista en la sección anterior.

La media de las diferencias es $\overline{x}_{dif} = -12,08591$ y la desviación estándar es $s_{dif} = 36,08183$.

Una vez más, comenzamos por formular las hipótesis:

H_0: la media de las diferencias en los tiempos de ejecución es igual a 0.
H_A: la media de las diferencias en los tiempos de ejecución es distinta de 0.

Que matemáticamente se expresan como:

Denotando la media de las diferencias en los tiempos de ejecución necesitados por ambos algoritmos para cualquier instancia del problema como μ_{dif}:
H_0: $\mu_{dif} = 0$
H_A: $\mu_{dif} \neq 0$

5.2.3 Prueba t para dos muestras independientes

En este caso, la prueba t se usa para comparar las medias de dos poblaciones en que las observaciones con que se cuenta no tienen relación con ninguna de las otras observaciones, ni influyen en su selección, ni en la misma ni en la otra muestra. En este caso la inferencia se hace sobre la diferencia de las medias: $\mu_1 - \mu_2 = d_0$, donde d_0 es un valor hipotético fijo para la diferencia. Usualmente se usa $d_0 = 0$, en cuyo caso las muestras podrían provenir de dos poblaciones distintas con igual media, o desde la misma población. Para ello, la prueba usa como estimador puntual la diferencia de las medias muestrales $(\overline{x}_1 - \overline{x}_2)$. Así, el estadístico T en este caso toma la forma de la ecuación 5.3.

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{SE_{(\overline{x}_1 - \overline{x}_2)}} \tag{5.3}$$

Al usar la distribución t de Student para la diferencia de medias, se deben cumplir los siguientes requisitos:

1. Cada muestra cumple las condiciones para usar la distribución t.
2. Las muestras son independientes entre sí.

Las hipótesis a formular en este caso son:

H_0: no hay diferencia entre la efectividad promedio de ambas vacunas.
H_A: la vacuna A es, en promedio, más efectiva que la B.

Si μ_A y μ_B son la concentraciones medias de anticuerpos presentes en personas luego de un mes de recibir la vacuna A y B, respectivamente, entonces:
H_0: $\mu_A = \mu_B$
H_A: $\mu_A > \mu_B$

CAPÍTULO 6. PODER ESTADÍSTICO

Así como el nivel de significación α corresponde a la probabilidad de cometer errores de tipo I, definimos ahora β como la probabilidad de cometer errores de tipo II. α y β están relacionados: **para un tamaño fijo de la muestra: al reducir β, α aumenta, y viceversa**. Este fenómeno se evidencia con mayor fuerza mientras más pequeña sea la muestra. No obstante, en la práctica resulta más interesante conocer la probabilidad de **no** cometer errores de tipo II. Esto nos lleva a un nuevo concepto: el **poder estadístico** de una prueba de hipótesis, dado por $1-\beta$, que se define como **la probabilidad de correctamente rechazar H_0 cuando es falsa**.

Otra forma de entender la noción de poder de una prueba es qué tan propensa es esta para distinguir un efecto real de una simple casualidad, lo que nos lleva a la noción de **tamaño del efecto**, que corresponde a una cuantificación de la diferencia entre dos grupos, o del valor observado con respecto al valor nulo.

6.1 PODER, NIVEL DE SIGNIFICACIÓN Y TAMAÑO DE LA MUESTRA

En la introducción de este capítulo vimos que el poder corresponde a la probabilidad de **no** cometer un error de tipo II, y que está muy relacionado con el tamaño de la muestra. También mencionamos que existe una relación entre el poder y el nivel de significación, la cual exploraremos en esta sección.

- El poder de la prueba aumenta mientras mayor es el tamaño del efecto (en este caso, la distancia entre el valor nulo y la media de la muestra).
- A medida que el tamaño del efecto disminuye (es decir, el estimador se acerca al valor nulo), el poder se aproxima al nivel de significación.
- Usar un valor de α más exigente (menor), manteniendo constante el tamaño de la muestra, hace que la curva de poder sea más baja para cualquier tamaño del efecto (lo que verifica la relación entre α y β).
- Usar una muestra más grande aumenta el poder de la prueba para cualquier tamaño del efecto distinto de 0.

6.2 TAMAÑO DEL EFECTO

El problema que podríamos tener al considerar el tamaño del efecto en la misma escala de la variable estudiada, como hemos hecho hasta ahora, es que esta escala varía de variable en variable. Para poder hacer comparaciones con mayor libertad, existen diferentes **medidas estandarizadas de efecto** que podemos

6.3 PODER, TAMAÑO DEL EFECTO Y TAMAÑO DE LA MUESTRA

Mencionamos en páginas anteriores que el poder puede también entenderse como qué tan propensa es una prueba estadística para distinguir un efecto real de una simple casualidad, y que podemos cuantificar este efecto.

6.4 CÁLCULO TEÓRICO DEL PODER

Como ya hemos mencionado a lo largo de este capítulo, el poder es la probabilidad de correctamente rechazar H_0 cuando es falsa, lo que equivale a la probabilidad de distinguir un efecto real de una mera casualidad. Ahora veremos algunos ejemplos de cómo podemos usar el poder.

H_0: $\mu_{(A_i - B_i)} = 0$, es decir que la media de las diferencias en el tiempo de ejecución necesitado por los algoritmos A y B, para cada posible instancia i, es cero

H_A: $\mu_{(A_i - B_i)} \neq 0$

CAPÍTULO 7. INFERENCIA CON PROPORCIONES MUESTRALES

7.1 MÉTODO DE WALD

En el capítulo 3 vimos que, cuando queremos responder preguntas del tipo "¿qué proporción de la ciudadanía apoya al gobierno actual?", estamos hablando de una variable aleatoria que sigue una distribución binomial. En general, no conocemos la **probabilidad de éxito** p de la población, por lo que tenemos que usar el estimador puntual (correspondiente a la proporción de éxito de la muestra), denotado por $\hat{p}$. Este estimador se distribuye de manera cercana a la normal cuando se cumplen las siguientes condiciones:

1. Las observaciones de la muestra son independientes.
2. Se cumple la **condición de éxito-fracaso**, que establece que se espera observar al menos 10 observaciones correspondientes a éxito y al menos 10, correspondientes a fracasos. Matemáticamente, $np \geq 10$ y $n(1-p) \geq 10$.

7.1.1 Método de Wald para una proporción

El **método de Wald** permite construir intervalos de confianza y contrastar hipótesis bajo el supuesto de normalidad para una proporción. Consideremos el siguiente ejemplo: Aquiles Baeza, ingeniero en informática, desea conocer qué proporción de las ejecuciones de un algoritmo de ordenamiento para instancias con 100.000 elementos (bajo iguales condiciones de hardware y sistema) tardan menos de 25 segundos. Para ello, registró los tiempos de ejecución para 150 instancias generadas de manera aleatoria, encontrando que 64 % de dichas instancias fueron resueltas en un tiempo menor al señalado.

H_0: el 70 % de las instancias se ejecutan en menos de 25 segundos.
H_A: más del 70 % de las instancias se ejecutan en menos de 25 segundos.

De acuerdo a las hipótesis formuladas por el jefe de Baeza, el valor nulo es $p_0 = 0,7$, con lo que estas pueden formularse matemáticamente como:

Denotando como p a la proporción de todas las instancias de tamaño 100.000 que se ejecutan en menos de 25 segundos y considerando el valor hipotético $p_0 = 0,7$ *para este parámetro:*
H_0: p = p_0
H_A: p >p_0

7.1.2 Método de Wald para dos proporciones

También podemos usar el método de Wald para estudiar la **diferencia entre las proporciones** de dos poblaciones, considerando para ello como estimador puntual la diferencia $\hat{p}_1 - \hat{p}_2$.

De manera similar a lo que ya vimos para una única proporción, también en este caso debemos verificar ciertas condiciones antes de poder aplicar el modelo normal:

1. Cada proporción, por separado, sigue el modelo normal.
2. Las dos muestras son independientes una de la otra.

Desde luego, también podemos realizar pruebas de hipótesis en este escenario. Para el ejemplo tenemos que:

H_0: no hay diferencia en la tasa de reprobación de hombres y mujeres.
H_A: las tasas de reprobación son diferentes para hombres y mujeres.

Matemáticamente:

Denotando como p_1 y p_2 a las proporciones de hombres y mujeres, respectivamente, que reprueban la asignatura de programación la primera vez que la cursan:
H_0: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$

Ya verificamos las condiciones para operar bajo el supuesto de normalidad cuando construimos el intervalo de confianza. Sin embargo, **cuando la hipótesis nula supone que no hay diferencia entre las proporciones**, la verificación de la condición de éxito-fracaso y la estimación del error estándar se realizan usando para ello la **proporción agrupada**, dada por la ecuación 7.5, donde $\hat{p}_1 n_1$ y $\hat{p}_2 n_2$ representan la cantidad de éxitos en la primera y segunda muestra, respectivamente.

Cuando contrastamos hipótesis para la **diferencia entre dos proporciones con un valor nulo distinto de 0**, el procedimiento es ligeramente diferente. En este caso, la comprobación de la condición de éxito-fracaso se realiza de manera independiente para ambas muestras y el error estándar se calcula, como ya se estudió para los intervalos de confianza, mediante la ecuación 7.4.

Supongamos ahora que la Facultad de Ingeniería de la Universidad anterior ha decidido replicar el estudio realizado para el curso de programación, esta vez para una asignatura de física. No obstante, las autoridades están convencidas de que la tasa de reprobación es 10 % mayor para los hombres y que, incluso, la diferencia podría ser mayor. Desean comprobar con un nivel de confianza de 95 % y para ello, seleccionaron aleatoriamente a 89 de los 1.023 hombres y a 61 de las 620 mujeres de la cohorte correspondiente al primer semestre de 2019. En las muestras se encuentran, respectivamente, 45 y 21 reprobaciones.

Las hipótesis son, en este caso:

H_0: la tasa de reprobación de los hombres es exactamente 10 % más alta que la de las mujeres.
H_A: la tasa de reprobación de los hombres es más de 10 % más alta que la de las mujeres.

Matemáticamente:

Denotando como p_1 y p_2 a las proporciones de hombres y mujeres, respectivamente, que reprueban la asignatura de física estudiada la primera vez que la cursan:
H_0: $p_1 - p_2 = 0,1$
H_A: $p_1 - p_2 > 0,1$

Al igual que en los ejemplos previos, las observaciones de cada muestra son independientes entre sí pues corresponden a menos del 10 % de la población y fueron escogidos aleatoriamente. A su vez, los datos proporcionados indican que se cumple la condición de éxito-fracaso para cada muestra. Como ambas muestras pertenecen a grupos diferentes de estudiantes, son independientes entre sí. En consecuencia, se cumplen las condiciones para operar bajo el modelo normal.

7.2 MÉTODO DE WILSON

El método de Wald, tratado en la sección anterior, es el método que tradicionalmente se ha usado y el que aparece en la mayoría de los libros clásicos de inferencia estadística. Sin embargo, el método está siendo muy criticado hoy en día debido a que hace importantes simplificaciones matemáticas en su procedimiento y ya hay evidencia empírica que ha demostrado sus limitaciones (Agresti & Coull, 1998).

Gracias al aumento del poder de cómputo y la disponibilidad de software estadístico, han surgido diversas alternativas, entre las cuales destaca el **método de Wilson** (junto con algunas variaciones), considerado el más robusto por diversos autores (Agresti & Coull, 1998; Brown y col., 2001; Devore, 2008; Wallis, 2013). Este método opera del mismo modo que el de Wald, aunque las fórmulas empleadas para estimar la proporción en la muestra y el error estándar son diferentes.

7.3 PODER Y PRUEBAS DE PROPORCIONES

En el capítulo anterior conocimos el poder estadístico y vimos que está relacionado con el nivel de significación, el tamaño de la muestra y el tamaño del efecto que queremos detectar.

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 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.

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

8.1 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). 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.

8.1.1 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**.

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.

8.1.2 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).

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.

8.1.3 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.

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.

8.2 PRUEBAS PARA MUESTRAS PEQUEÑAS

Hemos visto que la prueba χ^2 nos pide que las observaciones esperadas para cada grupo sean a lo menos 5. Sin embargo, hay escenarios donde esta condición no se cumple, por lo que debemos recurrir a alguna alternativa.

8.2.1 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.

8.2.2 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.

8.3 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 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.

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.

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.

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: