# Capítulo I

## Aplicaciones de la Prueba Chi-Cuadrado

*Cuando el Señor creó el mundo y las personas para vivir en él – obra que de acuerdo con la ciencia moderna, llevó mucho tiempo – podría muy bien imaginarme que razonó para sí de la siguiente manera: "Si hago todo predecible, estos seres humanos, a los que he dotado de cerebros bastante buenos, indudablemente aprenderán a predecirlo todo, y por lo tanto no tendrán aliciente para hacer nada, porque reconocerán que el futuro está totalmente determinado y en él no puede influir ninguna acción humana. Por otra parte, si todo lo hago impredecible, gradualmente descubrirán que no hay base racional para ninguna decisión y por tanto, como en el primer caso, no tendrán motivos para hacer nada. Ninguno de estos dos proyectos tiene sentido. Crearé, por lo tanto, una mezcla de los dos. Que unas cosas sean predecibles y otras impredecibles. Tendrán entonces, entre muchas otras cosas, la importante tarea de saber cuál es cuál."*

E. F. Schumacher.

## 1. Introducción

Una de las mayores utilidades de la distribución Chi-Cuadrado consiste en que permite comparar frecuencias observadas (frecuencias obtenidas en un experimento o muestreo) con frecuencias esperadas según un modelo supuesto (hipótesis nula). Esta característica de la distribución Chi-Cuadrado permite efectuar las siguientes pruebas:

1. Pruebas de bondad de ajuste a una distribución de probabilidades.
2. Prueba de homogeneidad de subpoblaciones.
3. Prueba de independencia.

La metodología a utilizar en cada uno de los tres casos será muy similar. La diferencia principal está en la forma en que se calculan las frecuencias esperadas ya que estas dependerán de la hipótesis nula en cuestión.

## 2. Pruebas Chi-Cuadrado de Bondad de Ajuste

Las pruebas de bondad de ajuste permiten evaluar cuán bien (o mejor dicho cuán mal) una variable aleatoria se ajusta a una distribución de probabilidades teórica. Otras pruebas de bondad de ajuste son la de Anderson-Darling y la de Kolmogorov-Smirnov. Mientras que la prueba Chi-Cuadrado se basa en la comparación de las frecuencias observadas con las frecuencias esperadas bajo el supuesto de que la hipótesis nula es verdadera, las pruebas de Anderson-Darling y de Kolmogorov-Smirnov se basan en la comparación de la distribución

de probabilidades acumuladas empírica (resultado de la muestra) con la distribución de probabilidades acumuladas teórica (según $H_0$).

## 2.1. Prueba de Bondad de Ajuste a una Distribución de Frecuencias

Esta prueba permite analizar si las frecuencias observadas de una variable aleatoria en *k* clases o categorías se ajustan o no a ciertas frecuencias teóricas o esperadas. Esta prueba se aplica principalmente con variables cualitativas como por ejemplo:

- Tipo de cáncer en los enfermos con cáncer en una población (1, 2, 3, 4 = otros tipos)
- Alguna característica genética heredada como por ejemplo el color de los ojos.

Hipótesis:

La hipótesis nula se define de acuerdo con las proporciones esperadas para cada una de las *k* categorías.

$H_0$: $\pi_i = \pi_{i0}$ para $i = 1, 2, \ldots k$.

$H_1$: $\pi_i \neq \pi_{i0}$ para al menos un $i$

Estadístico de Prueba:

El estadístico de prueba tiene una distribución Chi-Cuadrado con *k*-1 grados de libertad y se define de la siguiente manera:

$$\chi^2_c = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(k-1)}$$

donde $o_i$ son las frecuencias observadas y $e_i$ las frecuencias esperadas. Las frecuencias esperadas se calculan multiplicando el tamaño de muestra *n* por cada una de las proporciones supuestas en $H_0$:

$$e_i = n\pi_i$$

Regla de decisión:

La hipótesis nula se rechaza con un nivel de significación $\alpha$ sí el $\chi^2_c$ resulta mayor que el valor de tabla $\chi^2_{[1-\alpha,\, k-1]}$.

**Ejemplo 1**: Suponga que en una población de enfermos con cáncer, históricamente los 3 tipos más frecuentes siguen las proporciones 35%, 24% y 18%, y que entonces un estudio es desarrollado para evaluar si estas proporciones han cambiado (debido a la nueva tecnología médica, nuevos hábitos de vida, etc.). En este caso la hipótesis nula sería:

$H_0$: Las proporciones poblacionales no han cambiado

$H_0$: $\pi_1 = 0.35$ $\pi_2 = 0.24$ $\pi_3 = 0.18$ $\pi_4 = 0.23$ (otros tipos de cáncer)

y la hipótesis alterna:

$H_1$: Las proporciones poblacionales sí han cambiado

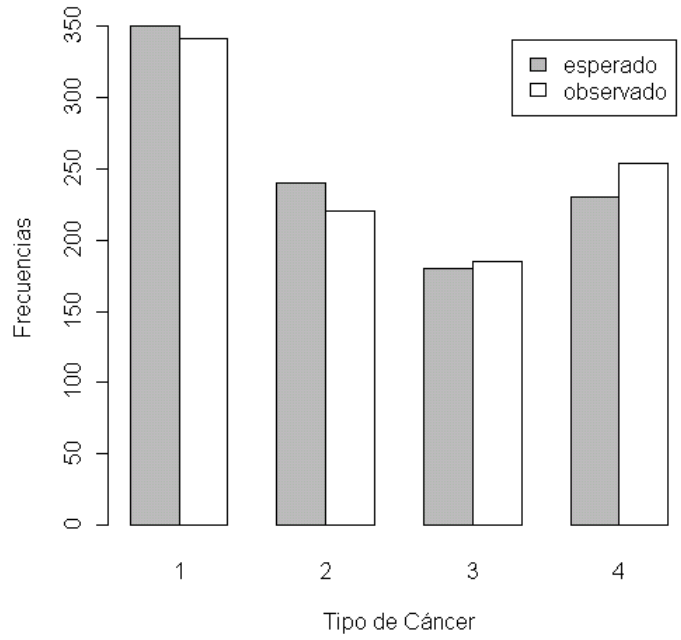
$H_1$: Al menos uno de los $\pi_i$ es diferente.

Suponga que en el estudio se obtuvieron los siguientes resultados con una muestra aleatoria de 1000 enfermos de cáncer:

| Tipo de cáncer | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Frecuencia observada | 341 | 220 | 185 | 254 |

Las frecuencias esperadas, si se supone que la hipótesis nula es verdadera (es decir que las proporciones no han cambiado), serán:

| Tipo de cáncer | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Frecuencia esperada | 350 | 240 | 180 | 230 |




Con estos datos, el estadístico de prueba resulta:

$$\chi_c^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = 4.54$$

El valor de tabla es $\chi^2_{(0.95,\ 3gl)} = 7.815$. Como el valor calculado es menor al valor de tabla, la información muestral no es suficiente para rechazar $H_0$, y se concluye que no existe suficiente evidencia estadística para aceptar que las proporciones de enfermos de cáncer han cambiado.

Cuando las frecuencias esperadas son pequeñas, la aproximación Chi-Cuadrado para la distribución del estadístico de prueba puede ser no muy buena. Para solucionar este problema

es aconsejable juntar categorías de modo que se eliminen las celdas con frecuencias esperadas muy pequeñas. Dos o más categorías pueden juntarse siempre y cuando estas sean combinables y el sentido de la hipótesis nula no se vea afectado por esta agrupación. Por otro lado, hay que tener presente, que por cada dos categorías que se junten se pierde un grado de libertad y que el poder de la prueba puede disminuir. Algunos autores recomiendan tener cuidado cuando hay muchas frecuencias esperadas menores a 5, o no permitir frecuencias esperadas menores a 1. La mayoría de los paquetes estadísticos muestran mensajes de advertencia cuando se tienen frecuencias esperadas menores a 5 ó 1.

## 2.2. Prueba de Bondad de Ajuste a una Distribución de Probabilidades

Esta prueba permite analizar si la distribución de probabilidades de una variable aleatoria se ajusta o no a una distribución de probabilidades teórica dada. En esta sección se presentarán los casos de bondad de ajuste a la distribución Binomial y a la Poisson. Sin embargo, el estudiante podrá aplicar esta metodología a cualquier otra distribución sin mucha dificultad. Antes de continuar, recuerde algunas características de las distribuciones Binomial y Poisson:

### Distribución Binomial

Una variable aleatoria $X$ tendrá distribución Binomial con parámetros $n$ y $\pi$ si cumple con las siguientes características:

- $X$ es el número de éxitos en $n$ ensayos independientes de un experimento, o el número de éxitos en una muestra de tamaño $n$. Para que los resultados de la muestra sean independientes la población debe ser infinita. Si la población es finita el muestreo debe ser con reemplazo.
- $\pi$ es la probabilidad de éxito para cada uno de los $n$ ensayos. Esta probabilidad debe ser constante para los $n$ ensayos.

Las siguientes variables podrían tener una distribución Binomial:

- Número de artículos defectuosos por lote.
- Número de personas que responden favorablemente a un tratamiento.
- Número de penales que falla un jugador en una ronda de 12.
- Número de entrevistados que sí estarían dispuestos a comprar un nuevo producto.

Decir que los $n$ ensayos son independientes implica que el resultado obtenido en un ensayo en particular no depende de los otros resultados. En el caso del número de penales fallados por un jugador, esto podría no ser cierto si se asume la existencia de un factor psicológico de modo que la confianza del jugador se vea mermada o incrementada según haya fallado o anotado en los lanzamientos anteriores. La falta de independencia entre los resultados podría ocurrir también en variables en las que todos los resultados estén afectados por algún factor común de modo que exista cierta posibilidad de que todos los elementos corran con la misma suerte; este podría ser por ejemplo el caso de la variable número de animales enfermos por corral (si es que la enfermedad es contagiosa).

**Distribución de Poisson**

Una variable aleatoria *X* tendrá distribución de Poisson con parámetro $\mu = \lambda t$ si cumple con las siguientes características:

- *X* es el número de eventos u ocurrencias aleatoriamente distribuidos por intervalo (de tiempo, longitud, volumen, etc.).
- $\lambda$ es el número medio de eventos por intervalo unitario.
- *t* es el tamaño del intervalo.
- $\mu$ es el número medio de eventos por intervalo de tamaño *t*.

A la distribución de Poisson se le conoce también como la distribución de los eventos raros (poco probables). La distribución de Poisson fue desarrollada por el matemático francés Poisson en 1837 y su primera aplicación fue la descripción del número de muertes por patada de mula en la armada prusiana.

Las siguientes variables podrían tener una distribución de Poisson:

- Número de bacterias por ml.
- Número de accidentes por semana en una intersección.
- Número de animales encontrados por $Km^2$.
- Número de emergencias atendidas en un hospital por día.

El procedimiento para la prueba será muy similar al presentado en la sección anterior. La única diferencia está en la forma de calcular las frecuencias esperadas, que en este caso se calcularán bajo el supuesto de que la variable tiene una distribución de probabilidades dada.

Hipótesis:

$H_0$: La variable *X* tiene una distribución de probabilidades dada.
$H_1$: La variable *X* no tiene una distribución de probabilidades dada.

Estadístico de prueba:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{(k\text{-}1\text{-}m)}$$

Las frecuencias esperadas se calculan de la siguiente manera:

$$e_i = np_i$$

donde $p_i$ son las probabilidades correspondientes a cada valor de *X* según la distribución de probabilidades establecida en la hipótesis nula. Si la distribución es Binomial las probabilidades se calcularán con la siguiente fórmula:

$$f(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Si la distribución es de Poisson las probabilidades se calcularán con:

$$f(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Los grados de libertad para el estadístico de prueba son ($k - 1 - m$) donde $k$ es el número de categorías y $m$ el número de parámetros estimados. En el caso de la distribución Binomial podría ser necesario estimar $\pi$ y en el caso de la distribución de Poisson podría ser necesario estimar $\mu$.

Regla de Decisión:

La hipótesis nula se rechaza con un nivel de significación $\alpha$ sí el $\chi_c^2$ resulta mayor que el valor de tabla $\chi^{2}_{[1-\alpha,\, k\,-\,1\,-\,m]}$.

**Ejemplo 2**: Hay 1000 bolsas de naranjas, cada una de las cuales contiene 10 naranjas. Algunas de las naranjas están podridas. ¿Es la distribución de probabilidades del número de naranjas podridas por bolsa una Binomial(10, $\pi$)? Los resultados obtenidos tras analizar las 1000 bolsas son los siguientes:

| Número de naranjas podridas | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Frecuencia observada (bolsas) | 334 | 369 | 191 | 63 | 22 | 12 | 9 |

$H_0$: El número de naranjas podridas por bolsa sigue una distribución Binomial (10, $\pi$) para algún $\pi$.

$H_1$: El número de naranjas podridas por bolsa no sigue una distribución Binomial (10, $\pi$)

Dado que no se conoce la proporción de naranjas podridas $\pi$, este valor será estimado con la proporción muestral $p$:

$$\hat{\pi} = p = \frac{\#\text{ de naranjas podridas}}{\#\text{ de naranjas}} = \frac{1142}{10000} = 0.1142$$

Ahora, se calculan las probabilidades binomiales para $X$ = 0, 1, 2, 3, 4, 5 y 6 ó más, y a partir de estas probabilidades se calculan las frecuencias esperadas:

| Núm. de naranjas podridas ($X$) | 0 | 1 | 2 | 3 | 4 | 5 | 6 ó + |
|---|---|---|---|---|---|---|---|
| Frecuencias observadas | 334 | 369 | 191 | 63 | 22 | 12 | 9 |
| $p(X)$ | 0.2974 | 0.3834 | 0.2224 | 0.0765 | 0.0173 | 0.0027 | 0.0003 |
| Frecuencias esperadas | 297.4 | 383.4 | 222.4 | 76.5 | 17.3 | 2.7 | 0.3 |

Note que las dos últimas frecuencias esperadas son menores a 5, por lo que será necesario agrupar las tres últimas categorías:

| Número de naranjas podridas ($X$) | 0 | 1 | 2 | 3 | 4 ó + |
|---|---|---|---|---|---|
| Frecuencias observadas | 334 | 369 | 191 | 63 | 43 |
| $p(X)$ | 0.2974 | 0.3834 | 0.2224 | 0.0765 | 0.0203 |
| Frecuencias esperadas | 297.4 | 383.4 | 222.4 | 76.5 | 20.3 |

Con estos datos el estadístico de prueba es:

$$\chi_c^2 = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = 37.24$$

Los grados de libertad para el estadístico de prueba serán 3 (5 categorías – 1 – 1 parámetro estimado). El valor de tabla para un nivel de significación del 5% es $\chi^2_{(0.95,\ 3gl)} = 7.815$. Como el valor calculado es mayor que el valor de tabla se rechaza $H_0$. En conclusión existe suficiente evidencia estadística para aceptar que el número de naranjas podridas por bolsa no sigue una distribución Binomial.

**Ejemplo 3**: Un entomólogo está analizando la distribución de una especie de insecto en una zona de cultivo. Para dicho estudio seleccionó 40 parcelas de 2m x 2m y contabilizó el número de insectos de dicha especie en cada una. Los resultados son los siguientes:

| Número de insectos | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Número de parcelas | 4 | 16 | 12 | 6 | 2 |

Pruebe con $\alpha$ =0.05 si los datos se ajustan a una distribución de Poisson.

$H_0$: El número de insectos por parcela sigue una distribución de Poisson ($\mu$) para algún $\mu$.

$H_1$: El número de insectos por parcela no sigue una distribución de Poisson ($\mu$).

Dado que no se conoce el parámetro $\mu$, este valor será estimado con la media muestral:

$$\hat{\mu} = \overline{X} = \frac{\#\text{ de insectos}}{\#\text{ de parcelas}} = \frac{66}{40} = 1.65$$

Ahora, se calculan las probabilidades de la distribución de Poisson para $X$ = 0, 1, 2, 3 y 4 ó más, y a partir de estas probabilidades se calculan las frecuencias esperadas:

| Número de insectos ($X$) | 0 | 1 | 2 | 3 | 4 ó más |
|---|---|---|---|---|---|
| Frecuencias observadas | 4 | 16 | 12 | 6 | 2 |
| $p(X)$ | 0.1920 | 0.3169 | 0.2614 | 0.1438 | 0.0859 |
| Frecuencias esperadas | 7.68 | 12.68 | 10.46 | 5.75 | 3.43 |

Agrupando las dos últimas categorías se tiene:

| Número de insectos ($X$) | 0 | 1 | 2 | 3 ó más |
|---|---|---|---|---|
| Frecuencias observadas | 4 | 16 | 12 | 8 |
| $p(X)$ | 0.1920 | 0.3169 | 0.2614 | 0.2296 |
| Frecuencias esperadas | 7.68 | 12.68 | 10.46 | 9.19 |

Con estos datos el estadístico de prueba es:

$$\chi_c^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = 3.0175$$

Los grados de libertad para el estadístico de prueba serán 2 (4 categorías – 1 – 1 parámetro estimado). El valor de tabla para un nivel de significación del 5% es $\chi^2_{(0.95,\ 2gl)} = 5.991$.

Como el valor calculado es menor al valor de tabla no se rechaza $H_0$. En conclusión no existe suficiente evidencia estadística para rechazar que el número de insectos por parcela siga una distribución de Poisson.

Una característica importante de la distribución de Poisson es que los eventos están distribuidos en forma aleatoria en el intervalo; por lo tanto, la prueba de bondad de ajuste a la distribución de Poisson puede ser utilizada para probar la aleatoriedad en la distribución de los eventos.

## 3. Pruebas Chi-Cuadrado para Tablas de Contingencia de dos Entradas

En esta sección se verán las pruebas de homogeneidad de subpoblaciones y de independencia. Si bien ambas pruebas presentan el mismo procedimiento de cálculo, las hipótesis a probar son diferentes y por lo tanto las conclusiones obtenidas también.

### 3.1. Prueba de Homogeneidad de Subpoblaciones

Esta prueba permite analizar si la distribución de probabilidades de una variable es la misma en *r* poblaciones.

Datos: Existen *r* poblaciones y una muestra aleatoria es extraída desde cada población. Sea $n_{i\bullet}$ el tamaño de la muestra extraída de la *i*-ésima población. Cada observación de cada muestra puede ser clasificada en una de *c* categorías diferentes. Los datos son arreglados en la siguiente tabla de contingencia *r*x*c*:

| | Categoría 1 | Categoría 2 | . . . | Categoría *c* | Total |
|---|---|---|---|---|---|
| Población 1 | $o_{11}$ | $o_{12}$ | . . . | $o_{1c}$ | $n_{1\bullet}$ |
| Población 2 | $o_{21}$ | $o_{22}$ | . . . | $o_{2c}$ | $n_{2\bullet}$ |
| . | . | . | | . | . |
| . | . | . | | . | . |
| . | . | . | | . | . |
| Población *r* | $o_{r2}$ | $o_{r2}$ | . . . | $o_{rc}$ | $n_{r\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | . . . | $n_{\bullet c}$ | $n_{\bullet\bullet}$ |

En la tabla, $o_{ij}$ es el número de observaciones de la muestra *i* clasificadas en la categoría *j*; $n_{\bullet j}$ es el número total de observaciones en la categoría *j* extraídas desde las *r* poblaciones y $n_{\bullet\bullet}$ es el total de observaciones extraídas desde las *r* poblaciones.

Hipótesis:

Sea $\pi_{ij}$ la probabilidad de que una observación seleccionada de la población *i* sea clasificada en la categoría *j*. Entonces las hipótesis son:

$H_0$: $\pi_{1j} = \pi_{2j} = \ . \ . \ . = \pi_{rj}$ para todo $j = 1, 2, \ldots c$.
$H_1$: Al menos una igualdad no se cumple.

Las hipótesis pueden expresarse equivalentemente de la siguiente manera:

$H_0$: La variable aleatoria tiene la misma distribución de probabilidades en las *r* poblaciones.

$H_1$: La variable aleatoria tiene una distribución de probabilidades diferente en al menos una de las poblaciones.

Estadístico de prueba:

$$\chi_c^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(r\text{-}1)(c\text{-}1)} \qquad \text{donde } e_{ij} = n_{i\bullet}\times\frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

Regla de decisión:

La hipótesis nula se rechaza con un nivel de significación $\alpha$ si el $\chi_c^2$ resulta mayor que el valor de tabla $\chi^2_{[1-\alpha,\,(r-1)(c-1)]}$.

**Ejemplo 4**: PTC es un compuesto que es amargo al sabor para algunos individuos e insípido para otros. Si uno puede o no saborear el PTC es una característica heredada. En la siguiente tabla se presentan las frecuencias de los individuos que pueden y no pueden saborear el PTC para muestras de cuatro países:

| | Irlanda | Portugal | Noruega | Italia | Total |
|---|---|---|---|---|---|
| Perciben el sabor | 558 | 345 | 185 | 402 | 1490 |
| No perciben el sabor | 225 | 109 | 81 | 134 | 549 |
| Total | 783 | 454 | 266 | 536 | 2039 |

¿Existen evidencias que indiquen que la proporción de personas que perciben el sabor amargo del PTC es diferente entre los 4 países?

En este caso las hipótesis a contrastar son las siguientes:

$H_0$: La proporción de personas que perciben el sabor del PTC es igual en los cuatro países.

$H_1$: La proporción de personas que perciben el sabor del PTC es diferente en al menos uno de los cuatro países.

Las frecuencias observadas y esperadas (frecuencias esperadas entre paréntesis) se presentan en la siguiente tabla:

| | Irlanda | Portugal | Noruega | Italia | Total |
|---|---|---|---|---|---|
| Perciben el sabor | 558 (572) | 345 (332) | 185 (194) | 402 (392) | 1490 |
| No perciben el sabor | 225 (211) | 109 (122) | 81 (72) | 134 (144) | 549 |
| Total | 783 | 454 | 266 | 536 | 2039 |

Con estos datos el estadístico de prueba es:

$$\chi_c^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 5.957 \sim \chi^2_{(2\text{-}1)(4\text{-}1)}$$

Los grados de libertad para el estadístico de prueba son (4-1)(2-1) = 3. El valor de tabla para un nivel de significación del 5% es $\chi^2_{(0.95,\ 3gl)} = 7.815$. Como el valor calculado es menor que el valor de tabla no se rechaza $H_0$ y se concluye que no existe suficiente evidencia estadística para aceptar que la proporción de personas que perciben el sabor amargo del PTC sea diferente entre los 4 países.

### 3.2. Prueba de Independencia

Esta prueba permite analizar si dos variables aleatorias son o no independientes.

Datos: Una muestra aleatoria de tamaño $n_{\bullet\bullet}$ es extraída, y cada observación de la muestra es clasificada de acuerdo a dos criterios (variables *X* y *Y*). Usando el primer criterio cada observación es clasificada en una de *r* filas y usando el segundo criterio en una de *c* columnas. Los datos son arreglados en la siguiente tabla de contingencia *r*x*c*:

| | Columna 1 | Columna 2 | . . . | Columna *c* | Total |
|---|---|---|---|---|---|
| Fila 1 | $o_{11}$ | $o_{12}$ | . . . | $o_{1c}$ | $n_{1\bullet}$ |
| Fila 2 | $o_{21}$ | $o_{22}$ | . . . | $o_{2c}$ | $n_{2\bullet}$ |
| . | . | . | | . | . |
| . | . | . | | . | . |
| . | . | . | | . | . |
| Fila *r* | $o_{r2}$ | $o_{r2}$ | . . . | $o_{rc}$ | $n_{r\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | . . . | $n_{\bullet c}$ | $n_{\bullet\bullet}$ |

En la tabla, $o_{ij}$ es el número de observaciones clasificadas en la fila *i* columna *j*, $n_{i\bullet}$ es el número total de observaciones en la fila *i* y $n_{\bullet j}$ es el número total de observaciones en la columna *j*.

Hipótesis:

Sea $\pi_{ij}$ la probabilidad de que una observación sea clasificada en la fila *i* columna *j*, $\pi_{i\bullet}$ la probabilidad de que una observación sea clasificada en la fila *i* y $\pi_{\bullet j}$ la probabilidad de que una observación sea clasificada en la columna *j*. Entonces las hipótesis son:

$H_0$: $\pi_{ij} = \pi_{i\bullet}\,\pi_{\bullet j}$ para todo $i = 1, \ldots r,\ j = 1, \ldots c$.
$H_1$: Al menos una igualdad no se cumple.

Las hipótesis pueden expresarse, en forma equivalente de la siguiente manera:

$H_0$: Las variables *X* y *Y* son independientes.
$H_1$: Las variables *X* y *Y* no son independientes.

Estadístico de prueba:

$$\chi_c^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(r\text{-}1)(c\text{-}1)} \qquad \text{donde } e_{ij} = n_{\bullet\bullet}\frac{n_{i\bullet}}{n_{\bullet\bullet}}\frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$$

Regla de decisión:

La hipótesis nula se rechaza con un nivel de significación $\alpha$ si el $\chi_c^2$ resulta mayor que el valor de tabla $\chi^2_{[1-\alpha,\,(r-1)(c-1)]}$.

**Ejemplo 5**: En un estudio sobre enfermedades al corazón en hombres, 356 voluntarios fueron clasificados de acuerdo con su nivel socioeconómico y sus hábitos de fumar. Los datos se presentan en la siguiente tabla:

| Hábito de fumar | Nivel Socioeconómico | | | |
|---|---|---|---|---|
| | Alto | Medio | Bajo | Total Filas |
| Actualmente | 51 | 22 | 43 | 116 |
| En el pasado | 92 | 21 | 28 | 141 |
| Nunca | 68 | 9 | 22 | 99 |
| Total Columnas | 211 | 52 | 93 | 356 |

¿Es el hábito de fumar independiente del nivel socioeconómico?

Las hipótesis a contrastar serán las siguientes:

$H_0$: El hábito de fumar es independiente del nivel socioeconómico.
$H_1$: El hábito de fumar no es independiente del nivel socioeconómico.

Las frecuencias observadas y esperadas (frecuencias esperadas entre paréntesis) se presentan en la siguiente tabla:

| Hábito de fumar | Nivel Socioeconómico | | | |
|---|---|---|---|---|
| | Alto | Medio | Bajo | Total Filas |
| Actualmente | 51 (68.75) | 22 (16.94) | 43 (30.30) | 116 |
| En el pasado | 92 (83.57) | 21 (20.60) | 28 (36.83) | 141 |
| Nunca | 68 (58.68) | 9 (14.46) | 22 (25.86) | 99 |
| Total Columnas | 211 | 52 | 93 | 356 |

Con estos datos el estadístico de prueba es:

$$\chi_c^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 18.510 \sim \chi^2_{(3\text{-}1)(3\text{-}1)}$$

Los grados de libertad para el estadístico de prueba son (3-1)(3-1) = 4. El valor de tabla para un nivel de significación del 5% es $\chi^2_{(0.95,\ 4\text{gl})} = 9.488$. Como el valor calculado es mayor que el valor de tabla se rechaza $H_0$ y se concluye que existe suficiente evidencia estadística para aceptar que el hábito de fumar y el nivel socioeconómico están relacionados (o no son independientes).

Esta prueba de independencia es útil principalmente cuando al menos una de las dos variables es cualitativa. Si bien es posible aplicar esta prueba con variables cuantitativas, en estos casos es posible realizar análisis más completos, los cuales pueden incluir el cálculo de un coeficiente de correlación, como por ejemplo el coeficiente de correlación de Pearson (que se verá en el capítulo 6), o los coeficientes de correlación basados en rangos como el de Spearman y el de Kendall, y el análisis del tipo de relación existente entre ambas variables (si es lineal, cuadrática, exponencial o logarítmica, etc.).

Al igual que en las pruebas de bondad de ajuste, hay que tener cuidado cuando se tengan frecuencias esperadas pequeñas, y es recomendable agrupar filas o columnas para evitar este problema.

### 3.3. Análisis de tablas 2x2

Ya sea que se esté tratando el caso de homogeneidad de subpoblaciones o el caso de independencia, si solo se tienen 2 filas y 2 columnas en la tabla de contingencias, esta se reduce a:

| | Columna 1 | Columna 2 | Total |
|---|---|---|---|
| Fila 1 | $o_{11}$ | $o_{12}$ | $n_{1\bullet}$ |
| Fila 2 | $o_{21}$ | $o_{22}$ | $n_{2\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet\bullet}$ |

y el estadístico de prueba puede simplificarse a la siguiente expresión:

$$\chi_c^{\,2} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{n_{\bullet\bullet}\left(o_{11}o_{22}-o_{12}o_{21}\right)^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}$$

Para mejorar el ajuste del estadístico de prueba a la distribución chi-cuadrado, Yates (1934) propuso utilizar una corrección de continuidad cuando el estadístico de prueba tiene solo un grado de libertad, para compensar la falta de exactitud producida por el uso de una distribución continua (la chi-cuadrado) para aproximar la distribución del estadístico de prueba que es discreta (ya que se basa en frecuencias y por lo tanto el número de posibles valores del $\chi_c^2$ es finito). Aplicando esta corrección el estadístico de prueba resulta:

$$\chi_c^{\,2} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(\left|o_{ij}-e_{ij}\right|-\frac{1}{2}\right)^2}{e_{ij}} = \frac{n_{\bullet\bullet}\left(\left|o_{11}o_{22}-o_{12}o_{21}\right|-\frac{n_{\bullet\bullet}}{2}\right)^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}$$

Sin embargo, hay que tener en cuenta que esta corrección disminuye el valor del $\chi_c^2$, y algunos autores consideran que el valor $\chi_c^2$ corregido resulta demasiado conservador.

**Ejemplo 6**: Un investigador realizó un experimento para comparar dos tratamientos con ratones enfermos. Cada tratamiento fue aplicado a una muestra de 30 ratones enfermos. La siguiente tabla muestra el número de ratones vivos luego de una semana:

| | Vivos | Muertos |
|---|---|---|
| Tratamiento 1 | 10 | 20 |
| Tratamiento 2 | 13 | 17 |

¿Hay evidencias suficientes para aceptar que alguno de los dos tratamientos sea más efectivo?

En este caso se tiene una prueba de homogeneidad de subpoblaciones y las hipótesis a contrastar son las siguientes:

$H_0$: Los dos tratamientos son iguales.

$H_1$: Alguno de los dos tratamientos es más efectivo que el otro.

Con los datos obtenidos, el estadístico de prueba sin corregir es:

$$\chi_c^2 = \frac{n_{\bullet\bullet}\left(o_{11}o_{22} - o_{12}o_{21}\right)^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{60\left((10)(17) - (20)(13)\right)^2}{(30)(30)(23)(37)} = 0.635$$

y el estadístico de prueba con la corrección de Yates:

$$\chi_c^2 = \frac{n_{\bullet\bullet}\left(\left|o_{11}o_{22} - o_{12}o_{21}\right| - \frac{n_{\bullet\bullet}}{2}\right)^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}} = \frac{60\left(\left|(10)(17) - (20)(13)\right| - \frac{60}{2}\right)^2}{(30)(30)(23)(37)} = 0.282$$

El valor de tabla para un nivel de significación del 5% es $\chi^2_{(0.95,\ 1gl)} = 3.842$. Como el valor calculado es menor que el valor de tabla no se rechaza $H_0$ y se concluye que no existe suficiente evidencia estadística para aceptar que alguno de los tratamientos sea más efectivo que el otro.

En tablas 2x2 es posible también evaluar hipótesis unilaterales. Los datos en este caso deben corresponder a una prueba de homogeneidad de subpoblaciones, esto es, se deben tener dos muestras aleatorias desde sus respectivas poblaciones y para cada muestra cada elemento debe ser clasificado en una de dos categorías, a las que se les designará como éxito y fracaso.

| | Éxito | Fracaso | Total |
|---|---|---|---|
| Población 1 | $o_{11}$ | $o_{12}$ | $n_{1\bullet}$ |
| Población 2 | $o_{21}$ | $o_{22}$ | $n_{2\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet\bullet}$ |

Las hipótesis unilaterales a contrastar serían las siguientes:

Caso A: Prueba de cola izquierda

$H_0$: $\pi_1 = \pi_2$

$H_1$: $\pi_1 < \pi_2$

Caso B: Prueba de cola derecha

$H_0$: $\pi_1 = \pi_2$
$H_1$: $\pi_1 > \pi_2$

donde $\pi_1$ y $\pi_2$ son las probabilidades de éxito en las poblaciones 1 y 2 respectivamente.

En este caso, el estadístico de prueba está dado por la raíz cuadrada de $\chi_c^2$ y su distribución se aproxima a una normal estándar:

$$Z_c = \frac{\sqrt{n_{\bullet\bullet}}\left(o_{11}o_{22} - o_{12}o_{21}\right)}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

La hipótesis nula será rechazada con un nivel de significación $\alpha$ si $Z_c$ es menor que $Z_{(\alpha)}$ en el caso de una prueba de cola izquierda y si $Z_c$ es mayor que $Z_{(1\,-\,\alpha)}$ en el caso de una prueba de cola derecha.

**Ejemplo 7**: Continuando con el ejemplo anterior, ahora se desea comparar el tratamiento 2 con un control. El objetivo del investigador es encontrar evidencias significativas de que el tratamiento es mejor que el control. Los datos se dan en la siguiente tabla:

| | Vivos | Muertos |
|---|---|---|
| Testigo | 7 | 23 |
| Tratamiento 2 | 13 | 17 |

En este caso, las hipótesis serán las siguientes:

$H_0$: El tratamiento no es efectivo (no es mejor que el testigo)
$H_1$: El tratamiento sí es efectivo (es mejor que el testigo)

En términos de la probabilidad de supervivencia, las hipótesis serían:

$H_0$: $\pi_1 = \pi_2$
$H_1$: $\pi_1 < \pi_2$

donde $\pi_1$ y $\pi_2$ son las probabilidades de que un ratón sobreviva en el grupo testigo y tratamiento respectivamente. Con los datos obtenidos, el estadístico de prueba es:

$$Z_c = \frac{\sqrt{n_{\bullet\bullet}}\left(o_{11}o_{22} - o_{12}o_{21}\right)}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}} = \frac{\sqrt{60}\left((7)(17) - (23)(13)\right)}{\sqrt{(30)(30)(20)(40)}} = -1.643$$

El valor de tabla para un nivel de significación del 5% es $Z_{(0.05)} = -1.645$. Como el valor calculado es mayor que el valor de tabla no se rechaza $H_0$ y se concluye que no existe suficiente evidencia estadística para aceptar que el tratamiento sea efectivo.

## Ejercicios

1. Los genetistas dicen que el color de los zapallos italianos debe seguir la razón 12:3:1. Un grupo de investigadores colecta la siguiente información: Blancas 155, amarillas 40 y verdes 10. ¿Son estos datos consistentes con la hipótesis de los genetistas?

2. Usted ha notado que los pinos crecen bien en algunas partes del bosque, pero no en otras. Usted especula que la distribución de los pinos está relacionada con el drenaje del terreno por lo que decide dividir el terreno en 100 parcelas igualmente espaciadas del bosque dos días después de una lluvia. Usted descubre que hay tres categorías de suelo: seco, margoso y húmedo. Como resultado de su análisis encuentra que 50 parcelas estaban secas, 30 margosas y 20 húmedas. Además, 50 parcelas tenían árboles de pino, 31 de las cuales estaban secas, 17 margosas y 2 húmedas. ¿Existe suficiente evidencia estadística para aceptar que los árboles de pino se desarrollan mejor en alguno de los tipos de suelo?

3. Se desea investigar si la distribución de buitres en un ecosistema es o no aleatoria. Con este objetivo, se colecta información sobre el número de nidos en áreas de 4 km cuadrados y se registra el número de nidos en cada área. Los resultados obtenidos fueron los siguientes:

| Número de nidos | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Número de áreas | 4 | 22 | 15 | 10 | 7 | 2 |

   Verifique el supuesto de que los nidos se distribuyen en forma aleatoria en el terreno.

4. Una compañía de seguros basa sus primas de seguros para cosechas en el número de incendios fuera de control en áreas de matorrales por año. ¿A que distribución de probabilidad podría ajustarse la variable número de incendios por año? A continuación se presenta información sobre el número de incendios en los últimos 60 años:

| Número de Incendios | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frecuencia | 8 | 10 | 16 | 14 | 12 |

   ¿Aporta esta información suficiente evidencia para rechazar su supuesto inicial?

5. El gerente de una empresa afirma que la probabilidad de producir un artículo defectuoso es 0.25 y que, dado que la condición de un artículo es independiente de la de los otros, el número de artículos defectuosos por caja debe ser una variable aleatoria con distribución Binomial. El departamento de Control de Calidad selecciona al azar 100 cajas de 4 artículos cada una obteniendo los siguientes resultados:

| N° de artículos no defectuosos | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frecuencias Observadas | 13 | 16 | 30 | 31 | 10 |

¿Presentan los datos suficiente evidencia al 5% de significación para rechazar la afirmación del gerente?

6. En un estudio ecológico se localizan 100 puntos sobre un mapa de un área forestal donde se buscará por nidos de aves. En cada locación se ubicarán los cuatro nidos más cercanos al punto y se registrará el número de nidos correspondientes a la especie I'wi (especie nativa hawaiana). Estudios anteriores dicen que la proporción de nidos de I'wi en el campo es 0.6 y que la ubicación de un nido es independiente de la de los otros. A continuación se presentan los resultados obtenidos:

| Número de nidos de I'wi en cada locación. | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Número de locaciones. | 20 | 41 | 10 | 22 | 7 |

¿Presentan los datos suficiente evidencia estadística para rechazar los supuestos antes mencionados?

7. Una muestra aleatoria de estudiantes es seleccionada aleatoriamente de escuelas privadas y otra de escuelas públicas. A los estudiantes se les aplica una prueba cuyos resultados se presentan a continuación:

| | Puntajes obtenidos | | | |
|---|---|---|---|---|
| | 0 - 25 | 26 - 50 | 51 - 75 | 76 – 100 |
| Escuelas privadas | 6 | 14 | 17 | 9 |
| Escuelas públicas | 30 | 32 | 17 | 3 |

¿Presenta esta información evidencia de que la preparación de los estudiantes es diferente en ambos tipos de escuela?

8. A continuación se presentan datos de un estudio sobre los tipos de sangre y su relación con el grupo étnico. Los datos fueron tomados del banco de sangre de Hawai.

| Tipo de Sangre | Grupo Étnico | | | |
|---|---|---|---|---|
| | Hawaiano | Hawaiano Blanco | Hawaiano Chino | Blanco |
| O | 1903 | 4469 | 2206 | 53759 |
| A | 2490 | 4671 | 2368 | 50008 |
| B | 178 | 606 | 568 | 16252 |
| AB | 99 | 236 | 243 | 5001 |

Evalúe si el tipo de sangre y el grupo étnico son variables independientes o no.

9. Los árboles frutales están sujetos a una enfermedad causada por bacterias comúnmente llamada plaga de fuego, debido a que las ramas muertas lucen como si hubiesen sido quemadas. Los siguientes tratamientos son propuestos para esta enfermedad: Tratamiento A: no acción (grupo control), tratamiento B: cuidadosa remoción de ramas afectadas y tratamiento C: frecuente rocío del follaje con un antibiótico en adición a la remoción de las ramas afectadas. Un grupo de 48 árboles es dividido aleatoriamente en tres grupos de

16 y cada grupo es asignado aleatoriamente a un tratamiento. Al cabo de un año se observa la condición del árbol y se registran tres posibles resultados: Resultado 1: el árbol ha muerto, resultado 2: el árbol no ha muerto pero sigue enfermo y resultado 3: el árbol ha sanado. Los resultados del experimento se presentan en la siguiente tabla:

| Resultado | Tratamiento *A* | *B* | *C* | Total Filas |
|---|---|---|---|---|
| 1 | 10 | 6 | 4 | 20 |
| 2 | 6 | 6 | 6 | 18 |
| 3 | 0 | 4 | 6 | 10 |
| Total Columnas | 16 | 16 | 16 | 48 |

¿Cuáles serían las hipótesis a contrastar en este caso? Plantee las hipótesis y efectué la prueba correspondiente.

10. Se realizó un estudio para determinar si el color de pelo y color de ojos guardan relación o actúan separadamente, obteniéndose los siguientes resultados:

| Color de Ojos | Color de pelo | | Total |
|---|---|---|---|
| | Rubio | Castaño | |
| Azules | 32 | 12 | 44 |
| Castaños | 14 | 22 | 36 |
| Otros | 6 | 9 | 15 |
| Total | 52 | 43 | 95 |

Efectúe la prueba correspondiente.

11. Se realizó un estudio para comparar la terapia de radiación con la cirugía en el tratamiento del cáncer. Se supone que la cirugía es más efectiva que la radiación. Para verificar estas sospechas se conduce un experimento con una muestra de 41 pacientes de los cuales 18 recibieron radiación y 23 cirugía. Los resultados se dan a continuación:

| | Cáncer controlado | Cáncer no controlado |
|---|---|---|
| Cirugía | 21 | 2 |
| Radiación | 15 | 3 |

¿Apoyan estos datos la práctica de la cirugía sobre la terapia de radiación?