

# Aplicaciones de la ji-cuadrado

## Bondad de ajuste (para una multinomial)

Esta es una prueba para comparar las probabilidades  $\pi_i$  de una distribución multinomial (lo esperado), con las obtenidas en una muestra (lo observado) para determinar si son iguales o no.

$H_0$ :  $\pi_1=p_1, \pi_2=p_2, \dots, \pi_k=p_k$

$H_1$ : las proporciones en la población no son  $\pi_1=p_1, \pi_2=p_2, \dots, \pi_k=p_k$

### Distibución Multinomial

La distribución Multinomial es una extensión de la distribución Binomial. En vez de haber solo dos posibles resultados (éxitos y fracasos) tenemos k posibles resultados.

Al igual que en la Binomial:

1. los experimentos son independientes
2. hay un número fijo de experimentos
3. la probabilidad de que ocurra cada uno de los resultados en un experimento,  $\pi_1, \pi_2, \dots, \pi_k$ , es constante.

### Ejemplo

Supongamos que en una clase de estadística un profesor ha determinado a través de los años que 10% de los estudiantes obtienen A, 20% B, 35% C, 5% D, 10% F y 20% se da de baja (W).

- Los diferentes resultados son: sacar A, B, C, D, F o W
- Los estudiantes son independientes.
- La probabilidad que tiene un estudiante de sacar A, B, C, D, F o W es:
- $\pi_A=.10 \quad \pi_B=.20 \quad \pi_C=.35 \quad \pi_D=.05 \quad \pi_F=.10 \quad \pi_W=.20$

El semestre pasado el profesor tuvo 120 ( $n=120$ ) estudiantes: de éstos 11 obtuvieron A, 10 B, 53 C, 9 D, 15 F y 22 W.

De acuerdo con esta muestra, ¿podrías decir que las proporciones de A, B, etc. no cambiaron? (usando un nivel de significancia de .05)

### Estadístico de prueba

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$O_i$  = valor observado en la muestra

$E_i$  = valor esperado cuando la hipótesis nula es cierta =  $n\pi_i$

Si el tamaño de muestra es grande y los valores esperados son mayores de 5 entonces  $\chi^2$  tiene una distribución aproximadamente ji-cuadrado con  $k-1$  grados de libertad.

### Decisión

A un nivel de significancia de  $\alpha$  se rechaza  $H_0$  si:  $p\text{-value} < \alpha$

### Ejemplo

Siguiendo con el ejemplo anterior,

#### Hipótesis:

$H_0: \pi_A=.10 \quad \pi_B=.20 \quad \pi_C=.35 \quad \pi_D=.05 \quad \pi_F=.10 \quad \pi_W=.20$

$H_1$ : las proporciones en la población no son estas

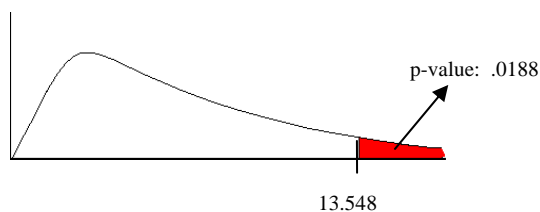
Grados de libertad: 5

Estadístico de prueba:

calculamos los valores esperados:

$n=120$

$\pi$	$E_i = n\pi_i$	$O_i$	$\frac{(O_i - E_i)^2}{E_i}$
.10	12	11	0.083
.20	24	10	8.167
.35	42	53	2.881
.05	6	9	1.500
.10	12	15	0.750
.20	24	22	0.167
Total:			$\chi^2 = 13.548$



#### Decisión:

Como el p-value (.0188) es menor que el nivel de significancia (.05), se rechaza  $H_0$ . De acuerdo con la evidencia las proporciones cambiaron.

## Tablas de contingencia y pruebas de independencia

Tablas de contingencia son tablas de doble entrada (variables cualitativas) que contienen las frecuencias con que ocurren las diferentes combinaciones de los valores de las variables.

### Ejemplo:

Una compañía de cervezas tiene dos tipos de cerveza: clara y negra. Antes de lanzar su nueva propaganda desea saber si existen diferencias en la preferencia del tipo de cervezas entre los hombres y las mujeres. Para determinar esto a un nivel de significancia de .05 hizo una encuesta a 150 personas y obtuvo la siguiente información:

		tipo de cerveza		total
		clara	negra	
género	F	55	15	70
	M	45	35	80
total		100	50	150

### Hipótesis y Estadístico de prueba

$H_0$ : las variables son independientes

$H_1$ : las variables no son independientes

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$O_{ij}$  = frecuencia observada en la muestra (de la fila  $i$ , columna  $j$  de la tabla) =  $n_{ij}$

$E_{ij}$  = valor esperado cuando la  $H_0$  es cierta =  $\frac{(\text{total de la fila } i)(\text{total de la columna } j)}{n}$

Si el tamaño de muestra es grande y los valores esperados son mayores de 5 entonces  $\chi^2$  tiene una distribución aproximadamente ji-cuadrada con  $(r-1)(c-1)$  grados de libertad. ( $r$  = número de filas,  $c$  = número de columnas)

		Variable 1		total
		A <sub>1</sub>	A <sub>2</sub>	
Variable 2	B <sub>1</sub>	$n_{11}$	$n_{12}$	$n_{1\cdot}$
	B <sub>2</sub>	$n_{21}$	$n_{22}$	$n_{2\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$n$

$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$

Total de la fila 1 (pointing to  $n_{1\cdot}$ )  
 Total de la columna 1 (pointing to  $n_{\cdot 1}$ )  
 Total de la muestra (pointing to  $n$ )

**Ejemplo** (siguiendo el ejemplo anterior)Hipótesis:

$H_0$ : las preferencias por el tipo de cerveza y el género son independientes

$H_1$ : las preferencias por el tipo de cerveza y el género no son independientes

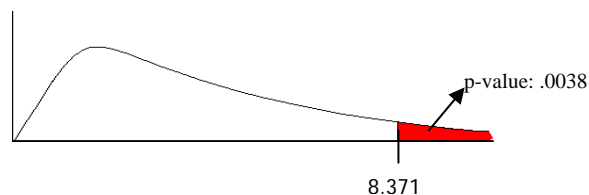
Grados de libertad:  $(2-1)(2-1)=1$  (r=2 filas y c=2 columnas)

Estadístico de prueba:

Obsevados:		tipo de cerveza		
		clara	negra	total
género	F	55	15	70
	M	45	35	80
Total		100	50	150

Esperados:		tipo de cerveza	
		clara	negra
género	F	$\frac{(70)(100)}{150} = 46.67$	$\frac{(70)(50)}{150} = 23.33$
	M	$\frac{(80)(100)}{150} = 53.33$	$\frac{(80)(50)}{150} = 26.67$

$$\chi^2 = \frac{(55 - 46.67)^2}{46.67} + \dots + \frac{(35 - 26.67)^2}{26.67} = 8.371$$

Decisión:

Como el p-value (.0038) es menor que el nivel de significancia (.05), se rechaza  $H_0$ . De acuerdo con la evidencia la preferencia por el tipo de cerveza depende del género.

Esta prueba no nos dice cuál es la dependencia.

## Prueba de hipótesis para proporciones (3 o más poblaciones independientes)

Cada población tiene una distribución Binomial  $n_i, \pi_i$  y se mide la misma característica en cada población.

La diferencia con la primera prueba es que ahora tenemos diferentes poblaciones independientes, y antes era una sola población.

### Hipótesis:

$$H_0: \pi_1 = \pi_2 = \dots = \pi_k$$

$H_1$ : por lo menos una es diferente

Se analiza de la misma forma que una tabla de contingencia de  $k$  columnas (las muestras) y 2 filas (tiene la característica y no tiene la característica)

**Estadístico de prueba:** (el mismo que para tablas de contingencia-prueba de independencia)

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$O_{ij} = n_{ij}$$

$$E_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

Si el tamaño de muestra es grande y los valores esperados son mayores de 5 entonces  $\chi^2$  tiene una distribución aproximadamente ji-cuadrada con  $k-1$  grados de libertad.

$\hat{p}_i = \frac{n_{ij}}{n_{\cdot j}}$	muestras				total	
	1	2	...	k		
característica	sí	$n_{11}$	$n_{12}$	...	$n_{1k}$	$n_{1\cdot}$
	no	$n_{21}$	$n_{22}$	...	$n_{2k}$	$n_{2\cdot}$
total		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	$n$

### Ejemplo:

Una compañía de seguros de carros desea investigar si hay diferencias entre el porcentaje de accidente que tienen las personas: menores de 25 años; entre 25 y 50 años; más de 50 años. (Usando un nivel de significancia de .01.)

A continuación se presentan los resultados:

	muestras			total
	Menores de 25	25-50	Mayores de 50	
tiene accidente	50	90	20	160
no tiene accidente	90	300	30	420
total	140	390	50	580

Hipótesis:

H0:  $\pi_1 = \pi_2 = \pi_3$

H1: por lo menos una es diferente

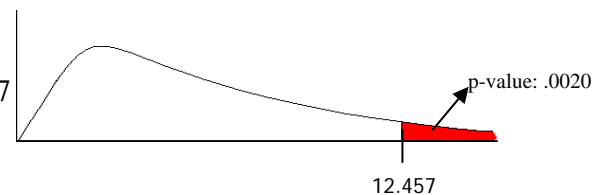
Grados de libertad:  $3 - 1 = 2$

Estadístico de prueba:

Observados:	muestras			total
	Menores de 25	25-50	Mayores de 50	
tiene accidente	50	90	20	160
no tiene accidente	90	300	30	420
total	140	390	50	580

Esperados:	muestras		
	Menores de 25	25-50	Mayores de 50
tiene accidente	38.620	107.586	13.793
no tiene accidente	101.379	282.414	36.207

$$\chi^2 = \frac{(50 - 38.620)^2}{38.620} + \dots + \frac{(30 - 36.207)^2}{36.207} = 12.457$$



Decisión:

Como el p-value (.0020) es menor que el nivel de significancia (.01), se rechaza H0. De acuerdo con la evidencia existe diferencia entre el por ciento de accidentes de estos tres grupos de edades.