Universidad Autónoma de Madrid

Gestión Aeronáutica: Estadística Teórica
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

APLICACIONES CHI-CUADRADRO

PRINCIPALES APLICACIONES DE LA CHI-CUADRADO

Al analizar en una población un carácter cualitativo o cuantitativo el estudio resulta muy tedioso por el gran número de elementos del que consta la población.

Generalmente, se examina una muestra tomada de la población, lo que lleva a tener una serie de datos, y ver hasta qué punto la muestra se pude considerar perteneciente a una distribución teórica conocida.

Siempre existirán desviaciones entre la distribución empírica u observada y la distribución teórica. Se plantea la cuestión de saber si estas desviaciones son debidas al azar o al haber tomado una distribución teórica inadecuada.

CONTRASTE DE BONDAD DEL AJUSTE

El objetivo del contraste de bondad del ajuste es saber si una muestra procede de una población teórica con determinada distribución de probabilidad.

Sea una población, donde se analiza un carácter X con $(x_1, x_2, \cdots, x_k)$ modalidades excluyentes, denotando por n_i es el número de elementos que presenta la modalidad x_i (frecuencia observada de x_i), $\sum_{i=1}^{k} n_i = n$

Por otra parte, sea $e_i = n \cdot p_i$ la frecuencia esperada o teórica de cada modalidad x_i

Se origina la **TABLA DE CONTINGENCIA**:

X	x_1	x_2	.....	x_i	.....	x_k
Frecuencia observada	n_1	n_2	.....	n_i	.....	n_k
Frecuencia esperada	(e_1)	(e_2)	.....	(e_i)	.....	(e_k)

Se plantea la hipótesis nula H_0 : La distribución teórica representa a la distribución empírica u observada

Para un nivel de significación (o riesgo) α

Se acepta H_0 : $\overbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}^{\text{estadístico observado}} < \overbrace{\chi^2_{\alpha,(k-1)}}^{\text{estadístico teórico}}$ Se rechaza H_0 : $\overbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}^{\text{estadístico observado}} \geq \overbrace{\chi^2_{\alpha,(k-1)}}^{\text{estadístico teórico}}$

El estadístico $\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n$ (útil en el cálculo)

OBSERVACIONES DE LA APLICACIÓN

a) El test de la χ^2 se puede aplicar en situaciones donde se desea decidir si una serie de datos (observaciones) se ajusta o no a una función teórica previamente determinada (Binomial, Poisson, Normal, etc.)

b) Es necesario que las frecuencias esperadas de las distintas modalidades no sea inferior a cinco. Si alguna modalidad tiene una frecuencia esperada menor que cinco se agrupan dos o más modalidades contiguas en una sola hasta conseguir que la frecuencia esperada sea mayor que cinco.

c) Los grados de libertad de la χ^2 dependen del número de parámetros que se necesitan hallar para obtener las frecuencias esperadas. En este sentido, si se requieren hallar p parámetros, los grados de libertad son $(k-p)$ si las modalidades son independientes y $(k-p-1)$ cuando las modalidades son excluyentes.

TABLAS CONTIGENCIA: CONTRASTE DE DEPENDENCIA O INDEPENDENCIA

Cuando se desea comparar dos caracteres (X, Y) en una misma población que admiten las modalidades: $X(x_1, x_2, \cdots, x_i, \cdots, x_k)$ $Y(y_1, y_2, \cdots, y_j, \cdots, y_m)$, se toma una muestra de tamaño n, representando por n_{ij} el número de elementos de la población que presentan la modalidad x_i de X e y_j de Y.

X \ Y	y_1	y_2	…	y_j	…	y_m	$\sum_{i=1}^{k} n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
$\sum_{j=1}^{m} n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	n

Se plantea la hipótesis nula H_0 : {No existe diferencia entre las distribuciones empíricas de X e Y

Bajo la hipótesis nula, cada frecuencia observada n_{ij} $(i=1,\cdots,k\ ;\ j=1,\cdots,m)$ de la tabla de contingencia $(k \times m)$ hay una frecuencia esperada (e_{ij}) que se obtiene mediante la expresión:

$$\boxed{e_{ij} = p_{ij} \cdot n = \frac{n_{i\bullet} \times n_{\bullet j}}{n}} \text{ , donde } p_{ij} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n}$$

Agrupando frecuencias observadas y esperadas en la tabla de contingencia (k x m):

X \ Y	y_1	y_2	…	y_j	…	y_m	$\sum_{i=1}^{k} n_{i\bullet}$
x_1	n_{11} (e_{11})	n_{12} (e_{12})	…	n_{1j} (e_{1j})	…	n_{1m} (e_{1m})	$n_{1\bullet}$
x_2	n_{21} (e_{21})	n_{22} (e_{22})	…	n_{2j} (e_{2j})	…	n_{2m} (e_{2m})	$n_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1} (e_{i1})	n_{i2} (e_{i2})	…	n_{ij} (e_{ij})	…	n_{im} (e_{im})	$n_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	n_{k1} (e_{k1})	n_{k2} (e_{k2})	…	n_{kj} (e_{kj})	…	n_{km} (e_{km})	$n_{k\bullet}$
$\sum_{j=1}^{m} n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	n

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exige que al menos el 80% de los valores esperados de las celdas sean mayores que 5. Cuando esto no ocurre hay que agrupar modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.
En una tabla de contingencia de 2×2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

El estadístico de contraste observado: $\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \chi^2_{(k-1)\,.\,(m-1)}$ que sigue aproximadamente una Chi-cuadrado con $(k-1) \times (m-1)$ grados de libertad.

Para un nivel de significación α se puede contrastar la diferencia significativa entre las dos distribuciones empíricas o la independencia de las distribuciones empíricas.

- **CONTRASTE DE HOMOGENEIDAD**

Se acepta H_o si : $\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} < \overbrace{\chi^2_{\alpha\,,\,(k-1)\,.\,(m-1)}}^{\text{estadístico teórico}}$

Se rechaza H_o si : $\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} \geq \overbrace{\chi^2_{\alpha\,,\,(k-1)\,.\,(m-1)}}^{\text{estadístico teórico}}$

- **CONTRASTE DE INDEPENDENCIA**

Hipótesis nula H_0 : Las distribuciones empíricas X e Y son independientes

Se acepta H_o si : $\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} < \overbrace{\chi^2_{\alpha\,,\,(k-1)\,.\,(m-1)}}^{\text{estadístico teórico}}$

Se rechaza H_o si : $\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} \geq \overbrace{\chi^2_{\alpha\,,\,(k-1)\,.\,(m-1)}}^{\text{estadístico teórico}}$

TABLAS CONTIGENCIA 2 x 2 y 2 x 3

Para las tablas de contingencia 2x2 y 2x3 se obtienen fórmulas sencillas de la χ^2 utilizando únicamente las frecuencias observadas

X \ Y	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	n

$$\chi^2_1 = \frac{n(n_{11}\,.\,n_{22} - n_{12}\,.\,n_{21})^2}{n_{1\bullet}\,.\,n_{2\bullet}\,.\,n_{\bullet 1}\,.\,n_{\bullet 2}}$$

Se acepta H_0 : $\chi^2_1 < \chi^2_{\alpha,1}$ Se rechaza H_0 : $\chi^2_1 \geq \chi^2_{\alpha,1}$

X \ Y	y_1	y_2	y_3	
x_1	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	n

$$\chi^2_2 = \frac{n}{n_{1\bullet}}\left[\frac{n^2_{11}}{n_{\bullet 1}} + \frac{n^2_{12}}{n_{\bullet 2}} + \frac{n^2_{13}}{n_{\bullet 3}}\right] + \frac{n}{n_{2\bullet}}\left[\frac{n^2_{21}}{n_{\bullet 1}} + \frac{n^2_{22}}{n_{\bullet 2}} + \frac{n^2_{23}}{n_{\bullet 3}}\right] - n$$

Se acepta H_0 : $\chi^2_2 < \chi^2_{\alpha,2}$ Se rechaza H_0 : $\chi^2_2 \geq \chi^2_{\alpha,2}$

Coeficiente de CONTINGENCIA

Es una medida del grado de relación o dependencia entre dos caracteres en la tabla de contingencia, se define:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad 0 \le C \le 1$$

Mayor valor de C indica un grado de dependencia mayor entre X e Y

FACTOR de corrección de YATES

Adviértase que como la muestra $n < 40$ se hace aconsejable el uso de la Chi-cuadrado con el factor de corrección de continuidad de Yates:

$$\text{Factor corrección} \begin{cases} n_{ij} < e_{ij} & \mapsto \quad n_{ij} + 0{,}5 \\ n_{ij} > e_{ij} & \mapsto \quad n_{ij} - 0{,}5 \end{cases}$$

Para una tabla de contingencia de 2×2 la corrección de Yates:

$$\chi_1^2 = \frac{n\left(\left|n_{11}.n_{22} - n_{12}.n_{21}\right| - \frac{n}{2}\right)^2}{n_{1\bullet}.n_{2\bullet}.n_{\bullet 1}.n_{\bullet 2}} \quad \text{la corrección no es válida cuando } \left|n_{11}.n_{22} - n_{12}.n_{21}\right| \le \frac{n}{2}$$

La corrección de Yates se hace cuando el número de grados de libertad es 1.

Test G de la razón de verosimilitud

El test de contraste de independencias por la razón de verosimilitudes (test G) es una prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el de Pearson. Se distribuye asintóticamente con una variable aleatoria χ^2 con $(k-1).(m-1)$ grados de libertad.

Se define el estadístico $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$

Se acepta la hipótesis nula H_0 si $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) < \chi^2_{\alpha,(k-1).(m-1)}$

Test de McNemar

El *test de McNemar* se utiliza para decidir si se puede aceptar o no que determinado tratamiento induce un cambio en la respuesta de los elementos sometidos al mismo, y es aplicable a los diseños del tipo **antes-después** en los que cada elemento actúa como su propio control.

Consisten en n observaciones de una variable aleatoria bidimensional (X,Y)

La escala de medición para X e Y es nominal con dos categorías, tales como positivo o negativo, hembra o macho, presencia o ausencia, que se pueden denominar 0 y 1.

X	Y		Total
	+	–	
+	**a**	**b**	$a+b$
–	**c**	**d**	$c+d$
Total	$a+c$	$b+d$	**n**

Los casos que muestran cambios entre la primera y segunda respuesta aparecen en las celdillas **b** y **c**.

Un individuo es clasificado en la celdilla **b** si cambia de + a –, en la celdilla **a** cuando la respuesta es + antes y después, en la celdilla **d** cuando la respuesta es – antes y después.

Hipótesis nula H_0 : El *tratamiento* no induce cambios significativos en las respuestas

En el *test de McNemar* para la significación de cambios solamente interesa conocer las celdas **b** y **c** que presentan cambios. Puesto que **b + c** es el número de individuos que cambiaron, bajo el supuesto de la hipótesis nula, se espera que **(b + c) / 2** casos cambien en una dirección y **(b + c) / 2** casos cambien en otra dirección.
Estadístico de contraste si **b + c < 20**:

$$\chi^2_{McNemar} = b \quad \text{se acepta } H_0 \text{ si } \chi^2_{McNemar} = b < \chi^2_{\alpha/2,1}$$

Estadístico de contraste si **b + c ≥ 20**:

$$\chi^2_{McNemar} = \chi^2_1 = \frac{(b-c)^2}{b+c} \quad \text{se acepta } H_0 \text{ si } \chi^2_{McNemar} = \chi^2_1 = \frac{(b-c)^2}{b+c} < \chi^2_{\alpha/2,1}$$

La aproximación muestral a la distribución Chi-cuadrado es más precisa si se realiza la corrección de continuidad de Yates (ya que se utiliza una distribución continua para aproximar una distribución discreta). El estadístico corregido:

$$\chi^2_{McNemar} = \chi^2_1 = \frac{(|b-c|-1)^2}{b+c} \quad \text{se acepta } H_0 \text{ si } \chi^2_{McNemar} = \chi^2_1 = \frac{(|b-c|-1)^2}{b+c} < \chi^2_{\alpha/2,1}$$

Coeficientes en distribuciones dicotómicas

Los coeficientes más utilizados en variables dicotómicas son los de correlación phi ϕ y Q de Yule.

Estos coeficientes tienen algunas propiedades comunes de interés:

a) Están normalizados, las magnitudes no dependen del tamaño de la tabla.

b) Son muy sensibles a la distribución empírica observada, traduciendo concentraciones de casos en algunas celdas en magnitudes.

c) Tienen un recorrido teórico entre [-1, 1] indicando situaciones de asociación perfecta y de independencia estadística.

Los coeficientes ϕ y Q de Yule se diferencian en la sensibilidad rinconal:

a) El coeficiente ϕ alcanza su máximo valor sólo cuando una de las dos diagonales se ha vaciado.

b) El coeficiente Q es muy sensible a la existencia de una celda que en términos relativos se está vaciando. Su valor máximo se alcanza cuando en una celda no hay ningún caso, esto es lo que se conoce como *sensibilidad rinconal*.

X	Y		Total
	y_1	y_2	
x_1	**a**	**b**	**(a + b)**
x_2	**c**	**d**	**(c + d)**
Total	**(a + c)**	**(b + d)**	**(n)**

Coeficiente Phi: $\phi = \dfrac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad 0 \le \phi \le 1$

Coeficiente Q de Yule: $Q = \dfrac{ad - bc}{ad - bc} \qquad 0 \le Q \le 1$

Test exacto de FISHER

Si las dos variables que se están analizando son dicotómicas, y la frecuencia esperada es menor que 5 en más de una celda, no resulta adecuado aplicar el test de la χ^2, aunque sí el test exacto de Fisher.

El test exacto de Fisher permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no cumple las condiciones necesarias para que la aplicación del test de la Chi-cuadrado sea idónea.

X	Y		Total
	y_1	y_2	
x_1	**a**	**b**	**(a + b)**
x_2	**c**	**d**	**(c + d)**
Total	**(a + c)**	**(b + d)**	**(n)**

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exige que al menos el 80% de los valores esperados de las celdas sean mayores que 5. De este modo, en una tabla de contingencia de 2×2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas 2×2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla observada. Cada uno de estas probabilidades se obtiene bajo la hipótesis de independencia de las dos variables que se están analizando.

La probabilidad asociada a los datos que han sido observados viene dada por:

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

La fórmula general de la probabilidad descrita deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales de filas y columnas de la tabla observada.

El valor de la p asociado al test exacto de Fisher puede calcularse sumando las probabilidades de las tablas que resulten menores o iguales a la probabilidad de la tabla que ha sido observada.

El planteamiento es bilateral, es decir, cuando la hipótesis alternativa asume la dependencia entre las variables dicotómicas, pero sin especificar de antemano en qué sentido se producen dichas diferencias, el valor de la p obtenido se multiplica por 2.

INTERPRETACIÓN DE DATOS

Se ha realizado un estudio sobre la situación laboral de las mujeres y su estado civil, los datos obtenidos fueron:

Trabajo remunerado	Estado civil		Total
	Casada	Soltera	
Si			
No			
Total	45	35	80

Los resultados obtenidos en el análisis de la tabla de contingencia fueron:

Estadísticos	Valor	p-valor
Chi-cuadrado Pearson	5,634361	0,0175
Chi-cuadrado de Yates	4,154897	0,0357
Test G	5,789645	0,0189
Chi-cuadrado NcNemar	2,94	0,0978
Correlación Phi	-0,685643	0,0178
Q de Yule	-0,812345	

Con un nivel de significación $\alpha = 0,05$, se pide:

a) ¿Se encuentra asociada la situación laboral de la mujer a su estado civil?

b) ¿Generalmente, las mujeres que realizan un trabajo remunerado con solteras?

Solución:

a) Para analizar la dependencia o no de la situación laboral de la mujer con su estado civil (asociación entre variables categóricas en una tabla de 2×2) se utiliza el test de la χ^2 de Pearson, con o sin corrección de Yates, el test G de razón de verosimilitudes. El test de McNemar mo se puede utilizar en este caso por no tratarse de muestras pareadas (antes-después).

Estableciendo las hipótesis:

H_0 : La situación laboral de la mujer es independiente de su estado civil.
H_1 : La situación laboral de la mujer depende de su estado civil.

Los tres estadísticos primeros, basados en la χ^2 , presentan un p-valor $< \alpha = 0,05$, con lo que se rechaza la hipótesis nula H_0 , concluyendo que la situación laboral de la mujer está asociada a su estado civil.

b) Partiendo de que la situación laboral de la mujer se encuentra asociada a su estado civil, falta por determinar la dirección de dicha asociación, para lo que se recurre al coeficiente de correlación Phi y la Q de Yule.

Ambos estadísticos son negativos, con un p-valor < $\alpha = 0{,}05$, pudiendo afirmar que la correlación entre la situación laboral y el estado civil de las mujeres es inversa y significativa al 5%.

Se puede concluir que la situación laboral de la mujer (sí esta trabajando) esta asociada a las solteras, con un nivel de significación del 5%.

CONTRASTE NO PARAMÉTRICO DE BONDAD DE AJUSTE

1. Para comprobar si los operarios encontraban dificultades con una prensa manual de imprimir, se hizo una prueba a cuatro operarios anotando el número de atascos sufridos al introducir el mismo número de hojas, dando lugar a la siguiente tabla:

Operario	A	B	C	D	Total
Obstrucciones	6	7	9	18	40

Con un nivel de significación del 5%, ¿existe diferencia entre los operarios?

Solución:

Estableciendo la hipótesis nula H_0 : 'No existe diferencia entre los operarios'

La probabilidad de que se atascase una hoja sería $1/4$ para todos los operarios. De este modo, el número de atascos esperados para cada uno de ellos sería $(e_i = 10)_{i=1,\cdots,4}$.

Tenemos, la tabla de contingencia 1 x 4:

Operario	A	B	C	D	Total
Obstrucciones	6 (10)	7 (10)	9 (10)	18 (10)	40 (40)

Se acepta la hipótesis nula, a un nivel de significación α si

$$\chi^2_{k-1} = \underbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n \; < \; \underbrace{\chi^2_{\alpha\,;\,k-1}}_{\text{estadístico teórico}} \qquad k \equiv \text{número intervalos}$$

o bien, la región de rechazo de la hipótesis nula: $R = \left\{ \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{\alpha\,;\,k-1} \right\}$

con lo cual, $\chi^2_3 = \sum_{i=1}^{4} \frac{n_i^2}{e_i} - n = \frac{6^2}{10} + \frac{7^2}{10} + \frac{9^2}{10} + \frac{18^2}{10} - 40 = 9$

Con el nivel de significación ($\alpha = 0{,}05$), el estadístico teórico: $\chi^2_{0{,}05\,;\,3} = 7{,}815$ siendo $\chi^2_3 = 9 > 7{,}815 = \chi^2_{0{,}05\,;\,3}$ se verifica la región de rechazo.

En consecuencia, se rechaza la hipótesis nula, concluyendo que existe diferencia significativa entre los operarios respecto al número de atascos en la prensa de imprimir.

CONTRASTE NO PARAMÉTRICO DE BONDAD DE AJUSTE A UNA POISSON CON PARÁMETRO DESCONOCIDO.

2. En un laboratorio se observó el número de partículas α que llegan a una determinada zona procedentes de una sustancia radiactiva en un corto espacio de tiempo siempre igual, obteniéndose los siguientes resultados:

Número partículas	0	1	2	3	4	5
Número períodos de tiempo	120	200	140	20	10	2

¿Se pueden ajustar los datos obtenidos a una distribución de Poisson, con un nivel de significación del 5%?

Solución:

Se establece la hipótesis nula H_0 : 'La distribución empírica se ajusta a la Poisson'

La hipótesis nula se acepta, a un nivel de significación α si

$$\chi^2_{k-p-1} = \underbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n \; < \; \underbrace{\chi^2_{\alpha\,;\,k-p-1}}_{\text{estadístico teórico}}$$

donde $k \equiv$ número intervalos

$p \equiv$ número parámetros a estimar

o bien, la región de rechazo de la hipótesis nula: $R = \left\{ \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{\alpha\,;\,k-p-1} \right\}$

La distribución de Poisson se caracteriza porque sólo depende del parámetro λ que coincide con la media.

Sea la variable aleatoria X = 'número de partículas' y n_i = 'número de períodos de tiempo'

x_i	n_i	$x_i n_i$	$P(x_i = k) = p_i$
0	**120**	**0**	**0,3012**
1	**200**	**200**	**0,3614**
2	**140**	**280**	**0,2169**
3	**20**	**60**	**0,0867**
4	**10**	**40**	**0,0260**
5	**2**	**10**	**0,0062**
	n = 492	**590**	

$$\bar{x} = \lambda = \frac{\sum x_i n_i}{n} = \frac{590}{492} = 1{,}2$$

$\lambda = 1{,}2$

en consecuencia,

$$P(x_i = k) = \frac{1{,}2^k}{k!} e^{-1{,}2} \quad k = 0, \cdots, 5$$

Las probabilidades con que llegan las partículas $k = 0, 1, \cdots, 5$ se obtienen sustituyendo los valores de k en $P(x_i = k) = \frac{1,2^k}{k!} e^{-1.2}$, o bien en las tablas con $\lambda = 1,2$

Para verificar si el ajuste de los datos a una distribución de Poisson se acepta o no, mediante una χ^2, hay que calcular las frecuencias esperadas $(e_i = n \,.\, p_i)$

x_i	0	1	2	3	4	5
Frecuencias	**120** ($e_1 = 148,2$)	**200** ($e_2 = 177,8$)	**140** ($e_3 = 106,7$)	**20** ($e_4 = 42,7$)	**10** ($e_5 = 12,8$)	**2** ($e_6 = 3,05$)

$e_1 = 492.0,3012 = 148,2$ **$e_2 = 492.0,3614 = 177,8$** **$e_3 = 492.0,2169 = 106,7$**

$e_4 = 492.0,0867 = 42,7$ **$e_5 = 492.0,0260 = 12,8$** **$e_6 = 492.0,0062 = 3,05$**

dando lugar a una tabla de contingencia 1 x 6, en donde hay que agrupar las dos últimas columnas por tener la última columna frecuencias esperadas menores que cinco.

Por tanto, se tiene la tabla de contingencia 1 x 5:

x_i	0	1	2	3	4 y 5
Frecuencias	**120** ($e_1 = 148,2$)	**200** ($e_2 = 177,8$)	**140** ($e_3 = 106,7$)	**20** ($e_4 = 42,7$)	**12** ($e_5 = 15,8$)

Así, los grados de libertad son tres $(k - p - 1 = 5 - 1 - 1 = 3)$

- El estadístico de contraste:

$$\chi^2_3 = \sum_{i=1}^{5} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{5} \frac{n_i^2}{e_i} - n = \frac{120^2}{148,2} + \frac{200^2}{177,8} + \frac{140^2}{106,27} + \frac{20^2}{42,7} + \frac{12^2}{15,8} - 492 = 32,31$$

- El estadístico teórico: $\chi^2_{0,05\,;\,3} = 7,815$

El estadístico de contraste (bondad de ajuste) es mayor que el estadístico teórico (7,815), rechazándose la hipótesis nula, es decir, la distribución NO se puede ajustar a una distribución de Poisson a un nivel de significación del 5%.

Se verifica la región de rechazo: $R = \left\{ \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{\alpha\,;\,k-p-1} \right\} \equiv \left\{ 32,31 > 7,815 \right\}$

CONTRASTE NO PARAMÉTRICO DE BONDAD DE AJUSTE A UNA NORMAL CON PARÁMETROS DESCONOCIDOS.

3. Para una muestra aleatoria simple de 350 días, el número de urgencias tratadas diariamente en un hospital A queda reflejado en la siguiente tabla:

Nº urgencias	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 - 30	Total días
Nº días	20	65	100	95	60	10	350

Contrastar, con un nivel de significación del 5%, si la distribución del número de urgencias tratadas diariamente en el hospital A se ajusta a una distribución normal.

Solución:

Para decidir si los datos se distribuyen normalmente es necesario calcular la media y desviación típica.

Se establece la hipótesis nula H_0 : 'La distribución empírica se ajusta a la normal'

Se acepta la hipótesis nula, a un nivel de significación α si

$$\chi^2_{k-p-1} = \underbrace{\sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n \; < \; \underbrace{\chi^2_{\alpha\,;\,k-p-1}}_{\text{estadístico teórico}} \quad \text{donde} \begin{cases} \mathbf{k} \equiv \textbf{número intervalos} \\ \mathbf{p} \equiv \textbf{número parámetros a estimar} \end{cases}$$

- Se obtiene la media y la desviación típica:

Intervalos	x_i	n_i	$x_i . n_i$	$x_i^2 . n_i$
0 - 5	2,5	20	50	125
5 - 10	7,5	65	487,5	3656,25
10 - 15	12,5	100	1250	15625
15 - 20	17,5	95	1662,5	29093,75
20 - 25	22,5	60	1350	30375
25 - 30	27,5	10	275	7562,5
		$n = \sum_{i=1}^{6} n_i = 350$	$\sum_{i=1}^{6} x_i n_i = 5075$	$\sum_{i=1}^{6} x_i^2 . n_i = 86437,5$

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i n_i}{350} = 14,5 \qquad \sigma_x^2 = \frac{\sum_{i=1}^{6} (x_i - \bar{x})^2 n_i}{350} = \frac{\sum_{i=1}^{6} x_i^2 . n_i}{250} - (\bar{x})^2 = 36,71 \qquad \sigma_x = 6,06$$

- Se procede al ajuste de una distribución normal $N(14,5\ ;\ 6,06)$, hallando las probabilidades de cada uno de los intervalos:

Intervalos	n_i	p_i	$e_i = p_i \cdot n$	$(n_i - e_i)^2$	$(n_i - e_i)^2 / e_i$
0 - 5	20	0,0498	17,43	6,6	0,38
5 - 10	65	0,1714	59,99	25,1	0,42
10 - 15	100	0,3023	105,81	33,76	0,32
15 - 20	95	0,2867	100,35	28,62	0,29
20 - 25	60	0,1396	48,86	124,1	2,54
25 - 30	10	0,0366	12,81	7,9	0,62
	n = 350				$\sum_{i=1}^{6} (n_i - e_i)^2 / e_i = 4,57$

$$P(0 < x < 5) = P\left[\frac{0 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{5 - 14,5}{6,06}\right] = P(-2,39 < z < -1,57) =$$

$$= P(1,57 < z < 2,39) = P(z > 1,57) - P(z > 2,39) = 0,0582 - 0,00842 = 0,04978$$

$$P(5 < x < 10) = P\left[\frac{5 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{10 - 14,5}{6,06}\right] = P(-1,57 < z < -0,74) =$$

$$= P(0,74 < z < 1,57) = P(z > 0,74) - P(z > 1,57) = 0,2296 - 0,0582 = 0,1714$$

$$P(10 < x < 15) = P\left[\frac{10 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{15 - 14,5}{6,06}\right] = P(-0,74 < z < 0,08) =$$

$$= P(0,08 < z < 0,74) = 1 - P(z > 0,74) - P(z > 0,08) = 1 - 0,4681 - 0,2296 = 0,3023$$

$$P(15 < x < 20) = P\left[\frac{15 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{20 - 14,5}{6,06}\right] = P(0,08 < z < 0,91) =$$

$$= P(z > 0,08) - P(z > 0,91) = 0,4681 - 0,1814 = 0,2867$$

$$P(20 < x < 25) = P\left[\frac{20 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{25 - 14,5}{6,06}\right] = P(0,91 < z < 1,73) =$$

$$= P(z > 0,91) - P(z > 1,73) = 0,1814 - 0,0418 = 0,1396$$

$$P(25 < x < 30) = P\left[\frac{25 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{30 - 14,5}{6,06}\right] = P(1,73 < z < 2,56) =$$

$$= P(z > 1,73) - P(z > 2,56) = 0,0418 - 0,0052 = 0,0366$$

- Se calculan las frecuencias esperadas, multiplicando las probabilidades por el número total de datos $e_i = p_i \cdot n$

- Se calcula el estadístico de contraste χ^2, donde el número de grados de libertad es $k - p - 1 = (\text{nº intervalos}) - (\text{nº parámetros a estimar}) - 1 = 6 - 2 - 1 = 3$, con lo cual,

$$\chi^2_3 = \sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = 4,57$$

Por otra parte, el estadístico teórico $\chi^2_{0,05\,;\,3} = 7{,}815$

Como $\chi^2_3 = 4{,}57 \;<\; \chi^2_{0,05\,;\,3} = 7{,}815$, se acepta la hipótesis nula a un nivel de significación del 5%. En consecuencia, la variable aleatoria número de urgencias en el hospital *A* sigue una distribución $N(14{,}5\,;\,6{,}06)$.

CONTRASTE DE HOMOGENEIDAD.

4. Para conocer la opinión de los ciudadanos sobre la actuación del alcalde de una determinada ciudad, se realiza una encuesta a 404 personas, cuyos resultados se recogen en la siguiente tabla:

	Desacuerdo	De acuerdo	No contestan
Mujeres	**84**	**78**	**37**
Varones	**118**	**62**	**25**

Contrastar, con un nivel de significación del 5%, que no existen diferencias de opinión entre hombres y mujeres ante la actuación del alcalde.

Solución:

Se trata de un contraste de homogeneidad en el que se desea comprobar si las muestras proceden de poblaciones distintas.

Se tienen dos muestras clasificadas en tres niveles, donde se desea conocer si los hombres y mujeres proceden de la misma población, es decir, si se comportan de manera semejante respecto a la opinión de la actuación del alcalde.

La hipótesis nula: H_0 **:'No existe diferencia entre hombres y mujeres respecto a la opinión'**

Región de rechazo de la hipótesis nula: $R_{rechazo} = \left\{ \chi^2_{(k-1)\,.\,(m-1)} \geq \chi^2_{\alpha\,;\;(k-1).(m-1)} \right\}$

Se forma una tabla de contingencia 2 x 3: En cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\,;\,j=1,\cdots,m}$ en la tabla de contingencia se tiene una frecuencia teórica o esperada e_{ij} que se calcula mediante la expresión: $e_{ij} = p_{ij}\,.\,n = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$, donde p_{ij} son las probabilidades de que un elemento tomado de la muestra presente las modalidades x_i de X e y_j de Y.

	Desacuerdo	De acuerdo	No contestan	$n_{i\bullet}$
Mujeres	**84** $(e_{11} = 99,5)$	**78** $(e_{12} = 68,96)$	**37** $(e_{13} = 30,53)$	**199**
Varones	**118** $(e_{21} = 102,5)$	**62** $(e_{22} = 71,03)$	**25** $(e_{23} = 31,46)$	**205**
$n_{\bullet j}$	**202**	**140**	**62**	**n = 404**

$$e_{11} = \frac{199\,.\,202}{404} = 99,5 \qquad e_{12} = \frac{199\,.\,140}{404} = 68,96 \qquad e_{13} = \frac{199\,.\,62}{404} = 30,53$$

$$e_{21} = \frac{205\,.\,202}{404} = 102,5 \qquad e_{22} = \frac{205\,.\,140}{404} = 71,03 \qquad e_{23} = \frac{205\,.\,62}{404} = 31,46$$

El estadístico de contraste: $\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(2-1)\,.\,(3-1)} = \chi^2_2$, con lo que,

$$\chi^2_2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \frac{(84-99,5)^2}{99,5} + \frac{(78-68,96)^2}{68,96} + \frac{(37-30,53)^2}{30,53} + \frac{(118-102,5)^2}{102,5} + \frac{(62-71,03)^2}{71,03} + \frac{(25-31,46)^2}{31,46} = 9,76$$

sigue una χ^2 con dos grados de libertad si es cierta la hipótesis nula con $e_{ij} > 5$ $\forall$ i, j; en caso contrario sería necesario agrupar filas o columnas contiguas.

- El estadístico de contraste: $\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(k-1)\,.\,(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - n$

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{n_{ij}^2}{e_{ij}} - n = \frac{84^2}{99,5} + \frac{78^2}{68,96} + \frac{37^2}{30,53} + \frac{118^2}{102,5} + \frac{62^2}{71,03} + \frac{25^2}{31,46} - 404 = 9,76$$

El estadístico teórico $\chi^2_{0,05\,;\,2} = 5,991$

Como $\chi^2_2 = 9,76 > \chi^2_{0,05\,;\,2} = 5,991$ se cumple la región de rechazo, concluyendo que las muestras no son homogéneas, es decir, no proceden de la misma población, hombres y mujeres no opinan lo mismo.

CONTRASTE DE INDEPENDENCIA.

5. Novecientos cincuenta escolares se clasificaron de acuerdo a sus hábitos alimenticios y a su coeficiente intelectual:

	Coeficiente Intelectual				Total
	< 80	80 - 90	90 - 99	≥ 100	
Nutrición buena	245	228	177	219	869
Nutrición pobre	31	27	13	10	81
Total	276	255	190	229	950

A un nivel de significación del 10%, ¿hay relación entre las dos variables tabuladas?

Solución:

Se trata de un contraste de independencia entre el coeficiente intelectual y los hábitos alimenticios.

Se establecen las hipótesis: $\begin{cases} H_0 : \text{'Las dos variables estudiadas son independientes'} \\ H_1 : \text{'Existe dependencia entre las dos variables'} \end{cases}$

El estadístico de contraste: $$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(k-1).(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - n$$

Siendo la región de rechazo de la hipótesis nula: $R_{rechazo} = \left\{ \chi^2_{(k-1)\,.\,(m-1)} \geq \chi^2_{\alpha\,;\,(k-1).(m-1)} \right\}$

En la tabla de contingencia 2 x 4 para cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\,;\,j=1,\cdots,m}$ se tiene una frecuencia teórica o esperada e_{ij} que se calcula mediante la expresión:

$$e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

	Coeficiente Intelectual				$n_{i\bullet}$
	< 80	80 - 90	90 - 99	≥ 100	
Nutrición buena	245 ($e_{11} = 252,46$)	228 ($e_{12} = 233,25$)	177 ($e_{13} = 173,8$)	219 ($e_{14} = 209,47$)	869
Nutrición pobre	31 ($e_{21} = 23,53$)	27 ($e_{22} = 21,74$)	13 ($e_{23} = 16,2$)	10 ($e_{24} = 19,52$)	81
$n_{\bullet j}$	276	255	190	229	950

$$e_{11} = \frac{869\,.\,276}{950} = 252,46 \qquad e_{12} = \frac{869\,.\,255}{950} = 233,25 \qquad e_{13} = \frac{869\,.\,190}{950} = 173,8 \qquad e_{14} = \frac{869\,.\,229}{950} = 209,47$$

$$e_{21} = \frac{81\,.\,276}{950} = 23,53 \qquad e_{22} = \frac{81\,.\,255}{950} = 21,74 \qquad e_{23} = \frac{81\,.\,190}{950} = 16,2 \qquad e_{24} = \frac{81\,.\,229}{950} = 19,52$$

El estadístico de contraste:

$$\chi_3^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{n_{ij}^2}{e_{ij}} - n = \frac{245^2}{252,46} + \frac{228^2}{233,25} + \frac{177^2}{173,8} + \frac{219^2}{209,47} + \frac{31^2}{23,53} + \frac{27^2}{21,74} + \frac{13^2}{16,2} + \frac{10^2}{19,52} - 950 = 9,75$$

ó bien,

$$\chi_3^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(245 - 252,46)^2}{252,46} + \frac{(228 - 233,25)^2}{233,25} + \frac{(177 - 173,8)^2}{173,8} + \frac{(219 - 209,47)^2}{209,47} +$$
$$+ \frac{(31 - 23,53)^2}{23,53} + \frac{(27 - 21,74)^2}{21,74} + \frac{(13 - 16,2)^2}{16,2} + \frac{(10 - 19,52)^2}{19,52} = 9,75$$

sigue una χ^2 con tres grados de libertad si es cierta la hipótesis nula con $e_{ij} > 5$ $\forall$ i, j; en caso contrario sería necesario agrupar filas o columnas contiguas.

El estadístico teórico $\chi^2_{0,10\,;\,3} = 6,251$

Como $\chi^2_3 = 9,75 > \chi^2_{0,10\,;\,3} = 6,251$ se cumple la región de rechazo, concluyendo que se rechaza la independencia, habiendo por tanto dependencia estadística entre el coeficiente intelectual y la alimentación.

6. Tres métodos de empaquetado de tomates fueron probados durante un período de cuatro meses; se hizo un recuento del número de kilos por 1000 que llegaron estropeados, obteniéndose los siguientes datos:

Meses	A	B	C	Total
1	6	10	10	26
2	8	12	12	32
3	8	8	14	30
4	9	14	16	39
Total	31	44	52	127

a) Observando simplemente los datos, ¿qué se puede inferir sobre el experimento?

b) Con un nivel de significación de 0,05, comprobar que los tres métodos tienen la misma eficacia.

Solución:

a) Con la simple observación de los datos, el empaquetado A parece ser el mejor, ya que es el que menos kilos de tomates estropeados tuvo. Ahora bien, esta situación puede ser engañosa, ya que hay que tener en cuenta el número de kilos que se empaquetaron.

Para tomar una decisión sobre si hay diferencia entre los diferentes métodos de empaquetado, se contrasta la hipótesis nula

H_0 : 'No existe diferencia entre los métodos de empaquetado'

b) La hipótesis nula H_0 : 'No existe diferencia entre los métodos de empaquetado'

Se acepta H_0 si: $\chi^2_{(k-1)\,.\,(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha;\,(k-1)\,.\,(m-1)}$

Se forma la tabla de contingencia 3 x 4 , donde $e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$

Empaquetado / Meses	A	B	C	Total
1	6 ($e_{11} = 6,35$)	10 ($e_{12} = 9,01$)	10 ($e_{13} = 10,62$)	26 (26)
2	8 ($e_{21} = 7,81$)	12 ($e_{22} = 11,09$)	12 ($e_{23} = 13,10$)	32 (32)
3	8 ($e_{31} = 7,32$)	8 ($e_{32} = 10,39$)	14 ($e_{33} = 12,28$)	30 (30)
4	9 ($e_{41} = 9,52$)	14 ($e_{42} = 13,51$)	16 ($e_{43} = 15,97$)	39 (39)
Total	31	44	52	127

$$e_{11} = \frac{26\,.\,31}{127} = 6,35 \qquad e_{21} = \frac{32\,.\,31}{127} = 7,81 \qquad e_{31} = \frac{30\,.\,31}{127} = 7,32 \qquad e_{41} = \frac{39\,.\,31}{127} = 9,52$$

$$e_{12} = \frac{26\,.\,44}{127} = 9,01 \qquad e_{22} = \frac{32\,.\,44}{127} = 11,09 \qquad e_{32} = \frac{30\,.\,44}{127} = 10,39 \qquad e_{42} = \frac{39\,.\,44}{127} = 13,51$$

$$e_{13} = \frac{26\,.\,52}{127} = 10,65 \qquad e_{23} = \frac{32\,.\,52}{127} = 13,10 \qquad e_{33} = \frac{30\,.\,52}{127} = 12,28 \qquad e_{43} = \frac{39\,.\,52}{127} = 15,97$$

Estadístico de contraste: $\chi^2_{(3-1)\,.\,(4-1)} = \chi^2_6 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{n_{ij}^2}{e_{ij}} - n = 128,24 - 127 = 1,24$

El estadístico teórico o esperado: $\chi^2_{0,05\,;\,6} = 12,592$

Siendo $\chi^2_6 = 1,24 < \chi^2_{0,05\,;\,6} = 12,592$, el estadístico observado es menor que el estadístico teórico o esperado, por tanto, no se cumple la región de rechazo, concluyendo que los tres métodos de empaquetado tienen la misma eficiencia.

7. Una empresa multinacional desea conocer si existen diferencias significativas entre sus trabajadores en distintos países en el grado de satisfacción en el trabajo- Para ello se toman muestran aleatorias simples de trabajadores, obteniendo los siguientes resultados:

	Satisfacción en el trabajo			
	Muy satisfecho	Satisfecho	Insatisfecho	Muy insatisfecho
España	200	300	300	100
Francia	300	400	350	150
Italia	350	300	250	150

¿Puede admitirse con un nivel de significación del 5% que la satisfacción en el trabajo es similar en los tres países?

Solución:

La hipótesis nula H_0: 'Las proporciones de los trabajadores con los distintos grados de satisfacción son iguales en los tres países'

Se acepta H_0:

$$\chi^2_{(k-1)\,.\,(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha\,;\;(k-1)\,.\,(m-1)}$$

Región de rechazo de la hipótesis nula: $R_{rechazo} = \left\{ \chi^2_{(k-1)\,.\,(m-1)} \geq \chi^2_{\alpha\,;\;(k-1)\,.\,(m-1)} \right\}$

Se forma la tabla de contingencia 3 x 4 donde cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\,;\,j=1,\cdots,m}$ tiene una frecuencia teórica o esperada $e_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$

	Satisfacción en el trabajo				Total
	Muy satisfecho	Satisfecho	Insatisfecho	Muy insatisfecho	
España	200 ($e_{11} = 242,86$)	300 ($e_{12} = 285,71$)	300 ($e_{13} = 257,14$)	100 ($e_{14} = 114,29$)	900 (900)
Francia	300 ($e_{21} = 323,81$)	400 ($e_{22} = 380,95$)	350 ($e_{23} = 342,86$)	150 ($e_{24} = 152,38$)	1200 (1200)
Italia	350 ($e_{31} = 283,33$)	300 ($e_{32} = 333,33$)	250 ($e_{33} = 300$)	150 ($e_{34} = 133,33$)	1050 (1050)
Total	850	1000	900	400	3150

Estadístico observado: $\chi^2_{(3-1)\,.\,(4-1)} = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{n_{ij}^2}{e_{ij}} - n =$

$$= \frac{200^2}{242,86} + \frac{300^2}{285,71} + \frac{300^2}{257,14} + \frac{100^2}{114,29} + \frac{300^2}{323,81} + \frac{400^2}{380,95} + \frac{350^2}{342,86} + \frac{150^2}{152,38} +$$

$$+ \frac{350^2}{283,33} + \frac{300^2}{333,33} + \frac{250^2}{300} + \frac{150^2}{133,33} - 3150 = 49,55$$

Estadístico teórico: $\chi^2_{0,05\ ;\ (3-1).(4-1)} = \chi^2_{0,05\ ;\ 6} = 12,592$

Como $\chi^2_6 = 49,55 > 12,592 = \chi^2_{0,05\ ;\ 6}$ se rechaza la hipótesis nula de homogeneidad de las tres muestras.
Es decir, la satisfacción en el trabajo de los empleados de los tres países es significativamente distinta.

8. Las compañías de seguros de automóviles suelen penalizar en sus primas a los conductores más jóvenes, con el criterio que éstos son más propensos a tener un mayor número de accidentes. En base a la tabla adjunta, con un nivel de significación del 5%, contrastar si el número de accidentes es independiente de la edad del conductor.

Edad del conductor	Número de accidentes al año				
	0	1	2	3	4
25 o menos	10	10	20	40	70
26 - 35	20	10	15	20	30
más de 36	60	50	30	10	5

Solución:

Hipótesis nula H_0: 'El número de accidentes sufridos por los conductores no depende de la edad del conductor'

Se acepta H_0:

$$\chi^2_{(k-1)\ .\ (m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha\ ;\ (k-1)\ .\ (m-1)}$$

Región de rechazo de la hipótesis nula: $R_{rechazo} = \left\{ \chi^2_{(k-1)\ .\ (m-1)} \geq \chi^2_{\alpha\ ;\ (k-1)\ .\ (m-1)} \right\}$

Se forma la tabla de contingencia 3 x 5 donde cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\ ;\ j=1,\cdots,m}$ tiene una frecuencia teórica o esperada en caso de independencia

$$e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Edad del conductor	Número de accidentes por año					$\sum_{j=1}^{m} n_{i\bullet}$
	0	1	2	3	4	
25 o menos	**10** $e_{11} = 33,75$	**10** $e_{12} = 26,25$	**20** $e_{13} = 24,37$	**40** $e_{14} = 26,25$	**70** $e_{15} = 39,37$	**150** (150)
26 - 35	**20** $e_{21} = 21,37$	**10** $e_{22} = 16,62$	**15** $e_{23} = 15,44$	**20** $e_{24} = 16,62$	**30** $e_{25} = 24,94$	**95** (95)
más de 36	**60** $e_{31} = 34,87$	**50** $e_{32} = 27,12$	**30** $e_{33} = 25,19$	**10** $e_{34} = 27,12$	**5** $e_{35} = 40,69$	**155** (155)
$\sum_{i=1}^{k} n_{\bullet j}$	**90**	**70**	**65**	**70**	**105**	**400**

$$e_{11} = \frac{150 \cdot 90}{400} = 33,75 \quad e_{12} = \frac{150 \cdot 70}{400} = 26,25 \quad e_{13} = \frac{150 \cdot 65}{400} = 24,37 \quad e_{14} = \frac{150 \cdot 70}{400} = 26,25 \quad e_{15} = \frac{150 \cdot 105}{400} = 39,37$$

$$e_{21} = \frac{95 \cdot 90}{400} = 21,37 \quad e_{22} = \frac{95 \cdot 70}{400} = 16,62 \quad e_{23} = \frac{95 \cdot 65}{400} = 15,44 \quad e_{24} = \frac{95 \cdot 70}{400} = 16,62 \quad e_{25} = \frac{95 \cdot 105}{400} = 24,94$$

$$e_{31} = \frac{155 \cdot 90}{400} = 34,87 \quad e_{32} = \frac{155 \cdot 70}{400} = 27,12 \quad e_{33} = \frac{155 \cdot 65}{400} = 25,19 \quad e_{34} = \frac{155 \cdot 70}{400} = 27,12 \quad e_{35} = \frac{155 \cdot 105}{400} = 40,69$$

Estadístico observado: $\chi^2_{(3-1)\,.\,(5-1)} = \chi^2_8 = \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{n_{ij}^2}{e_{ij}} - n =$

$$= \left(\frac{10^2}{33,75} + \frac{10^2}{26,25} + \frac{20^2}{24,37} + \frac{40^2}{26,25} + \frac{70^2}{39,37}\right) + \left(\frac{20^2}{21,37} + \frac{10^2}{16,62} + \frac{15^2}{15,44} + \frac{20^2}{16,62} + \frac{30^2}{24,94}\right) +$$

$$+ \left(\frac{60^2}{34,87} + \frac{50^2}{27,12} + \frac{30^2}{25,19} + \frac{10^2}{27,12} + \frac{5^2}{40,69}\right) - 400 = 143,51$$

Estadístico teórico: $\chi^2_{0,05\,;\;(3-1)\,.\,(5-1)} = \chi^2_{0,05;8} = 15,507$

Como $\chi^2_8 = 143,51 > 15,507 = \chi^2_{0,05\,;\,8}$ se rechaza la hipótesis nula de independencia entre la edad del conductor y el número de accidentes.
En consecuencia, la edad influye significativamente en el número de accidentes al año.

9. En dos ciudades, A y B, se observó el color del pelo y de los ojos de sus habitantes, encontrándose las siguientes tablas:

Ciudad A

Pelo / Ojos	Rubio	No Rubio
Azul	47	23
No azul	31	93

Ciudad B

Pelo / Ojos	Rubio	No Rubio
Azul	54	30
No azul	42	80

a) Hallar los coeficientes de contingencia de las dos ciudades.

b) ¿En cuál de las dos ciudades podemos afirmar que hay mayor dependencia entre el color del pelo y de los ojos?

Solución:

a) Se calculan los valores de la χ^2 correspondientes a las dos observaciones, siendo la frecuencia esperada $e_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$

Ciudad A

Pelo / Ojos	Rubio	No Rubio	Total
Azul	47 ($e_{11} = 28,14$)	23 ($e_{12} = 41,85$)	70 (70)
No azul	31 ($e_{21} = 49,85$)	93 ($e_{22} = 74,14$)	124 (124)
Total	78	116	194

$$e_{11} = \frac{70 \cdot 78}{194} = 28,14 \qquad e_{12} = \frac{70 \cdot 116}{194} = 41,85$$

$$e_{21} = \frac{124 \cdot 78}{194} = 49,85 \qquad e_{22} = \frac{124 \cdot 116}{194} = 74,14$$

Estadístico de contraste:

$$X^2_{(2-1)\,.\,(2-1)} = X^2_1 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{n_{ij}^2}{e_{ij}} - n = \frac{47^2}{28,14} + \frac{23^2}{41,85} + \frac{31^2}{49,85} + \frac{93^2}{74,14} - 194 = 33,07$$

El coeficiente de contingencia: $C_A = \sqrt{\dfrac{33,07}{33,07 + 194}} = 0,3816$

En la población B, la tabla de contingencia 2 x 2:

Ciudad B

Pelo / Ojos	Rubio	No Rubio	Total
Azul	54 (e_{11} = 39, 15)	30 (e_{12} = 44,85)	84 (84)
No azul	42 (e_{21} = 56, 85)	80 (e_{22} = 65, 15)	122 (122)
Total	96	110	206

$$e_{11} = \frac{84 \,.\, 96}{206} = 39,15 \qquad e_{12} = \frac{84 \,.\, 110}{206} = 44,85$$

$$e_{21} = \frac{96 \,.\, 122}{206} = 56,85 \qquad e_{22} = \frac{110 \,.\, 122}{206} = 65,15$$

Estadístico de contraste:

$$\chi^2_{(2-1)\,.\,(2-1)} = \chi^2_1 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{n_{ij}^2}{e_{ij}} - n = \frac{54^2}{39,15} + \frac{30^2}{44,85} + \frac{42^2}{56,85} + \frac{80^2}{65,15} - 206 = 17,82$$

El coeficiente de contingencia: $C_B = \sqrt{\dfrac{17,82}{17,82 + 206}} = 0,282$

b) Como el coeficiente de contingencia mide el grado de relación o dependencia entre las variables, afirmamos que en la población A hay mayor dependencia entre el color de los ojos y del pelo.

10. En una muestra aleatoria de personas se analizan algunos hábitos de la vida, habiendo recogido datos de las siguientes variables:

X_1= Estado general de salud: muy bueno (3), bueno (2), regular (1), malo (0)

X_2= Sexo: mujer (1), hombre (0)

X_3= Nivel del ejercicio diario: intenso (2), moderado (1), ninguno (0)

Realizadas las tablas de contingencia correspondientes, se calcularon los siguientes estadísticos para contrastar la asociación:

a) $\chi^2(X_1, X_2) = 8$ b) $\chi^2(X_2, X_3) = 4,5$ $\chi^2(X_1, X_3) = 6,1$

Con la información facilitada, a un nivel de significación del 5%, elaborar un diagnóstico para cada una de las parejas de variables.

Solución:

Calculando los p-valor(α_p) de cada estadístico se obtiene:

a) H_0 : X_1 e X_2 son independientes

En $\chi^2(X_1, X_2) = 8$ el número de grados de libertad es $(4-1)\times(2-1) = 3$

$\alpha_p = P(\chi^2_{p,3} \geq 8)$. Interpolando en la tabla de la Chi-cuadrado:

0,05	α_p	0,025
7,815	8	9,348

$$0,05-0,025 \longrightarrow 7,815-9,348$$
$$\alpha_p - 0,025 \longrightarrow 8-9,348$$

$$(\alpha_p - 0,025)\times(7,815-9,348) = (0,05-0,025)\times(8-9,348) \quad \mapsto \quad \alpha_p = 0,0469$$

Siendo $\alpha_p = 0,0469 < 0,05$ se rechaza la hipótesis nula, concluyendo que el estado general de salud está asociado al sexo.

b) H_0 : X_2 e X_3 son independientes

En $\chi^2(X_2, X_3) = 4,5$ el número de grados de libertad es $(2-1)\times(3-1) = 2$

$\alpha_p = P(\chi^2_{p,2} \geq 4,5)$. Interpolando en la tabla de la Chi-cuadrado:

0,90	α_p	0,10
0,211	4,5	4,605

$$0,90-0,10 \longrightarrow 0,211-4,605$$
$$\alpha_p - 0,10 \longrightarrow 4,5-4,605$$

$$(\alpha_p - 0,10)\times(0,211-4,605) = (0,90-0,10)\times(4,5-4,605) \quad \mapsto \quad \alpha_p = 0,119$$

Siendo $\alpha_p = 0,119 > 0,05$ se acepta la hipótesis nula, concluyendo que el sexo es independiente del nivel del ejercicio diario.

c) H_0 : X_1 e X_3 son independientes

En $\chi^2(X_1, X_3) = 6,1$ el número de grados de libertad es $(4-1)\times(3-1) = 6$

$\alpha_p = P(\chi^2_{p,6} \geq 6,1)$. Interpolando en la tabla de la Chi-cuadrado:

0,90	α_p	0,10
2,204	6,1	10,645

$$0,90-0,10 \longrightarrow 2,204-10,645$$
$$\alpha_p - 0,10 \longrightarrow 6,1-10,645$$

$$(\alpha_p - 0,10)\times(2,204-10,645) = (0,90-0,10)\times(6,1-10,645) \quad \mapsto \quad \alpha_p = 0,530$$

Siendo $\alpha_p = 0,530 > 0,05$ se acepta la hipótesis nula, concluyendo que el estado general de salud es independiente del nivel del ejercicio diario.

11. Para curar cierta enfermedad se sabe que existen cuatro tratamientos diferentes. Aplicados por separado a un grupo distinto de enfermos, se han observado los siguientes resultados:

Enfermo / Tratamientos	Curados	No curados	Total
A	60	23	83
B	46	10	56
C	70	17	87
D	54	30	84

¿Se puede considerar que la eficacia de los cuatro tratamientos es la misma con un nivel de confianza del 95 por 100?

Solución 1:

Se trata de un contraste de homogeneidad de cuatro muestras, con 83, 56, 87 y 84 personas, de las cuales hay, respectivamente, 60, 46, 70 y 54 personas curadas.

Se establece la hipótesis nula

H_0 : Los cuatro tratamientos (muestrales) son de la misma eficacia

Lo que lleva a afirmar que la proporción de personas curadas en cada muestra es $p = 230/310 = 0{,}742$ y las no curadas $q = 1 - 0{,}742 = 0{,}258$, donde $e_{i\bullet} = n_{i\bullet} \times p$

Enfermo / Tratamientos	Curados n_{i1}	No curados	$n_{i\bullet}$
A	60	23	83 ($e_{1\bullet}$= 61,58)
B	46	10	56 ($e_{2\bullet}$= 41,55)
C	70	17	87 ($e_{3\bullet}$= 64,55)
D	54	30	84 ($e_{4\bullet}$= 62,32)

$$\sum_{i=1}^{4} \frac{(n_{i1} - e_{i\bullet})^2}{n_{i\bullet}} = \frac{(60-61{,}58)^2}{83} + \frac{(46-41{,}55)^2}{56} + \frac{(70-64{,}55)^2}{87} + \frac{(54-62{,}32)^2}{84} = 1{,}55$$

$$\chi^2_{4-1} = \left(\frac{1}{p \times q}\right) \sum_{i=1}^{4} \frac{(n_{i1} - e_{i\bullet})^2}{n_{i\bullet}} = 1{,}55 \times \left(\frac{1}{0{,}742 \times 0{,}258}\right) = 8{,}09$$

Siendo $\chi^2_3 = 8,09 > 7,815 = \chi^2_{0,05\,;\,3}$ se rechaza la hipótesis nula, es decir, los tratamientos a efectos de curar a los pacientes, a un nivel $\alpha = 0,05$, tienen diferente eficacia.

Solución 2:

Se establece la hipótesis nula

H_0 : Los cuatro tratamientos (muestrales) son de la misma eficacia

Es una tabla de contingencia 4x2, con una frecuencia teórica $e_{ij} = p_{ij}.n = \dfrac{O_{i\bullet} \times O_{\bullet j}}{n}$

Enfermo / Tratamientos	Curados	No curados	$O_{i\bullet}$
A	60 (61,58)	23 (21,42)	83
B	46 (41,55)	10 (14,45)	56
C	70 (64,55)	17 (22,45)	87
D	54 (62,32)	30 (21,68)	84
$O_{\bullet j}$	230	80	310

$$e_{11} = \frac{83x230}{310} = 61,58 \qquad e_{21} = \frac{56x230}{310} = 41,55 \qquad e_{31} = \frac{87x230}{310} = 64,55 \qquad e_{41} = \frac{84x230}{310} = 62,32$$

$$e_{12} = \frac{83x80}{310} = 21,42 \qquad e_{22} = \frac{56x80}{310} = 14,45 \qquad e_{32} = \frac{87x80}{310} = 22,45 \qquad e_{42} = \frac{84x80}{310} = 21,68$$

$$\chi^2_{(4-1)\,.\,(2-1)} = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \frac{(60-61,58)^2}{61,58} + \frac{(23-21,42)^2}{21,42} + \frac{(46-41,55)^2}{41,55} +$$

$$+\frac{(10-14,45)^2}{14,45} + \frac{(70-64,55)^2}{64,55} + \frac{(17-22,45)^2}{22,45} + \frac{(54-62,32)^2}{62,32} + \frac{(30-21,68)^2}{21,68} = 8,09$$

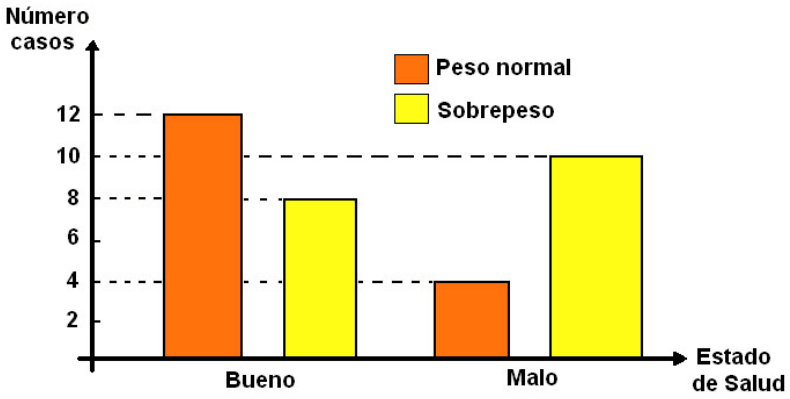
La expresión anterior se podía haber realizado de forma más sencilla con la igualdad:

$$\chi^2_{(4-1)\,.\,(2-1)} = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{O_{ij}^2}{e_{ij}} - n = \frac{60^2}{61,58} + \frac{23^2}{21,42} + \frac{46^2}{41,55} + \frac{10^2}{14.45} +$$

$$+\frac{70^2}{64,55} + \frac{17^2}{22,45} + \frac{54^2}{62,32} + \frac{30^2}{21,68} - 310 = 8,09$$

Como $\chi^2_3 = 8{,}09 > 7{,}815 = \chi^2_{0{,}05\,;\,3}$ se rechaza la hipótesis nula, es decir, los tratamientos a efectos de curar a los pacientes, a un nivel $\alpha = 0{,}05$, tienen diferente eficacia.

12. En el gráfico se presenta la evaluación del estado general de salud de una muestra de personas adultas mayores, según sea su peso normal o sobrepeso.



ESTADO GENERAL DE SALUD DE UN ADULTO MAYOR, SEGÚN PESO

Con los datos del gráfico, con un nivel de significación del 5%, analizar la existencia de una relación significativa entre el peso y el estado general de salud en el adulto mayor.

Solución:

a) Se trata de dos variables dicotómicas, con datos de frecuencia, pudiéndose aplicar una prueba de contraste de asociación con la Chi-cuadrado.

La hipótesis nula H_0 : El estado de salud y el peso son independientes

Llevando la información a una tabla de contingencia de 2×2

Estado de Salud	**Peso**		Total
	Normal	Sobrepeso	
Bueno	12 (9,41)	8 (10,59)	20 (20)
Malo	4 (6,59)	10 (7,41)	14 (14)
Total	16	18	34

La frecuencia observada $n_{21} = 4$ es menor que lo aconsejable en cada celda (≥ 5), lo que podría hacer pensar en una inestabilidad del cálculo.

Como la frecuencia esperada $e_{21} = 6{,}59$, todas las celdas cumplen con el mínimo aconsejable de 5 en su valor esperado. En la práctica se acepta hasta un 20% de las celdas que no cumplen con el requisito de que la frecuencia esperada sea ≥ 5

Se calculan los valores de la χ^2 correspondientes a las dos observaciones, siendo la frecuencia esperada $e_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$

$e_{11} = \dfrac{20 \,.\, 16}{34} = 9,41$ $\quad$ $e_{12} = \dfrac{20 \,.\, 18}{34} = 10,59$ $\quad$ $e_{21} = \dfrac{14 \,.\, 16}{34} = 6,59$ $\quad$ $e_{22} = \dfrac{14 \,.\, 18}{18} = 7,41$

Estadístico de contraste:

$$\chi^2_{(2-1)\,.\,(2-1)} = \chi^2_1 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{n_{ij}^2}{e_{ij}} - n = \frac{12^2}{9,41} + \frac{8^2}{10,59} + \frac{4^2}{6,59} + \frac{10^2}{7,41} - 34 = 3,27$$

Estadístico teórico: $\chi^2_{0,05,1} = 3,841$

Como $\chi^2_1 = 3,27 < 3,841 = \chi^2_{0,05,1}$ se acepta la hipótesis nula, concluyendo que el estado general de salud del adulto mayor no está asociado a su peso.

📖 Adviértase que como la muestra $n < 40$ se hace aconsejable el uso de la Chi-cuadrado con el factor de corrección de continuidad de Yates:

$$\text{Factor corrección} \begin{cases} n_{ij} < e_{ij} & \mapsto \quad n_{ij} + 0,5 \\ n_{ij} > e_{ij} & \mapsto \quad n_{ij} - 0,5 \end{cases}$$

Para una tabla de contingencia de 2×2 la corrección de Yates:

$$\chi^2_1 = \frac{n\left(\left|n_{11}.n_{22} - n_{12}.n_{21}\right| - \dfrac{n}{2}\right)^2}{n_{1\bullet}.n_{2\bullet}.n_{\bullet 1}.n_{\bullet 2}}$$ la corrección no es válida cuando $\left|n_{11}.n_{22} - n_{12}.n_{21}\right| \le \dfrac{n}{2}$

En general, la corrección de Yates se hace cuando el número de grados de libertad es 1.

En este caso, $\chi^2_1 = \dfrac{34\left(\left|12 \times 10 - 8 \times 4\right| - \dfrac{34}{2}\right)^2}{20 \times 14 \times 16 \times 18} = 2,13$

Como $\chi^2_1 = 2,13 < 3,841 = \chi^2_{0,05,1}$ se acepta la hipótesis nula.

La validez del contraste también se puede hacer con el p-valor (α_p):

$\alpha_p = P\left(\chi^2_{p,1} > 2,13\right) = 0,271$

0,90	α_p	0,10
0,0158	2,13	2,706

$0,90 - 0,10 \longrightarrow 0,0158 - 2,706$

$\alpha_p - 0,10 \longrightarrow 2,13 - 2,706$

$(\alpha_p - 0{,}10) \times (0{,}0158 - 2{,}706) = (0{,}90 - 0{,}10) \times (2{,}13 - 2{,}706) \quad \mapsto \quad \alpha_p = 0{,}271$

Al ser $\alpha_p = 0{,}271 > 0{,}05 = \alpha$ se rechaza la hipótesis nula, afirmando que el estado general de salud del adulto mayor es independiente de su peso.

13. Un experimento para investigar el efecto de vacunación de animales de laboratorio refleja la siguiente tabla:

Vacuna	**Animal laboratorio**	
	Enfermo	No Enfermo
Vacunado	9	42
No Vacunado	18	28

Con un nivel de significación de 0,05, ¿Es conveniente vacunar?.

Solución:

Hipótesis nula H_0 : Es independiente la vacuna de los animales enferemos

Vacuna	**Animal laboratorio**		Total
	Enfermo	No Enfermo	
Vacunado	9	42	51
No Vacunado	18	28	46
Total	27	70	97

En una tabla de contingencia de 2×2 se puede calcular la χ^2 de una forma sencilla recurriendo a las frecuencias observadas.

Estadístico observado: $\chi_1^2 = \dfrac{n\left(n_{11}.n_{22} - n_{12}.n_{21}\right)^2}{n_{1\bullet}.n_{2\bullet}.n_{\bullet 1}.n_{\bullet 2}} = \dfrac{97\left(9.28 - 42.18\right)^2}{51.46.27.70} = 5{,}5570$

El número de grados de libertad es $(2-1)\times(2-1) = 1$

Estadístico teórico: $\chi^2_{0{,}05,1} = 3{,}841$

Siendo $\chi_1^2 = 5{,}5570 > 3{,}841 = \chi^2_{0{,}05,1}$ se rechaza la hipótesis nula, es decir, la vacuna afecta a la enfermedad, con un nivel de significación $\alpha = 0{,}05$

14. Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46 **(a)**	50 **(b)**	96 **(a+b)**
B	85 **(c)**	59 **(d)**	144 **(c+d)**
Total	131 **(a+c)**	109 **(b+d)**	240 **(n)**

Se desea saber si el debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

Solución:

Se trata de una muestra pareada en una situación antes-después, con lo que es idóneo un contraste estadístico Chi-cuadrado de McNemar.

Sea la hipótesis nula H_0 : La intención de voto es la misma antes y después del debate

Estadístico muestral: $\chi^2_{McNemar} = \frac{(85-50)^2}{85+50} = 9{,}074$

Estadístico teórico: $\chi^2_{\alpha/2,1} = \chi^2_{0{,}025,1} = 5{,}024$

Como $\chi^2_{McNemar} = 9{,}074 > 5{,}024 = \chi^2_{0{,}025,1}$ se rechaza la hipótesis nula, concluyendo que la intención de voto cambió significativamente después del debate, con un nivel de significación del 5%.

15. Se desea analizar si los estudiantes de universidades privadas preferentemente son de los estratos económicos altos del país. Para ello, se ha tomado la siguiente muestra:

Universidades	Grupos socioeconómicos			
	Alto	Medio alto	Medio bajo	Bajo
Estado	13	17	4	3
Privadas	38	19	2	2

a) Para validar el análisis con un nivel de confianza del 95%, realizar un contraste por la razón de verosimilitud (test G).

b) Estudiar el grado de dependencia entre el tipo de universidad y el estrato socioeconómico.

Solución:

a) El test de contraste de independencias por la razón de verosimilitudes (test G) es una prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el de Pearson. Se distribuye asintóticamente como una variable aleatoria χ^2 con $(k-1).(m-1)$ grados de libertad.

Se define el estadístico $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$

Se acepta la hipótesis nula H_0 si $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) < \chi^2_{\alpha,(k-1).(m-1)}$

Hipótesis nula H_0 : El tipo de universidad es independiente del grupo socioecónomico

En un principio, la tabla presenta un 50% de celdas que no verifican que las frecuencias sean mayores que 5, teniendo que agrupar modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.

Universidades	Grupos socioeconómicos		
	Alto	Medio alto	Medio bajo - Bajo
Estado	13	17	7
Privadas	38	19	4

Se calculan los valores esperados de cada celda, donde $e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$

Universidades	Grupos socioeconómicos			Total
	Alto	Medio alto	Medio bajo - Bajo	
Estado	13 (19,26)	17 (13,59)	7 (4,15)	37 (37)
Privadas	38 (31,74)	19 (22,41)	4 (6,85)	61 (61)
Total	51	36	11	98

$$e_{11} = \frac{37 \cdot 51}{98} = 19,26 \qquad e_{12} = \frac{37 \cdot 36}{98} = 13,59 \qquad e_{13} = \frac{37 \cdot 11}{98} = 4,15$$

$$e_{21} = \frac{61 \cdot 51}{98} = 31,74 \qquad e_{22} = \frac{61 \cdot 36}{98} = 22,41 \qquad e_{23} = \frac{61 \cdot 11}{98} = 6,85$$

La frecuencia esperada $e_{13} = 4,15 < 5$, valor mínimo recomendado para la prueba. En un caso práctico se admite hasta un 20% de las celdas que no verifican este requisito, como ocurre en este caso.

En cada celda se calcula el valor de $n_{ij} \times \ln\left(\frac{n_{ij}}{e_{ij}}\right)$

Universidades	Grupos socioeconómicos			Total
	Alto	Medio alto	Medio bajo - Bajo	
Estado	−5,11	3,80	3,66	2,35
Privadas	6,84	−3,14	−2,15	1,55
Total	1,73	0,66	1,51	3,9

$$13 \times \ln\left(\frac{13}{19,26}\right) = -5,11 \qquad 17 \times \ln\left(\frac{17}{13,59}\right) = 3,80 \qquad 7 \times \ln\left(\frac{7}{4,15}\right) = 3,66$$

$$38 \times \ln\left(\frac{38}{31,74}\right) = 6,84 \qquad 19 \times \ln\left(\frac{19}{22,41}\right) = -3,14 \qquad 4 \times \ln\left(\frac{4}{6,85}\right) = -2,15$$

El estadístico observado $G = 2\sum_{i=1}^{2}\sum_{j=1}^{3} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) = 2 \times 3,9 = 7,8$

El número de grados de libertad es $(2-1).(3-1) = 2$

El estadístico teórico $\chi^2_{0,05,2} = 5,991$

Como $G = 7,8 > 5,991 = \chi^2_{0,05,2}$, se rechaza la hipótesis nula de independencia, concluyendo que el tipo de universidad está asociado al grupo socioeconómico.

La validez del contraste también se puede hacer con el p-valor (α_p):

$\alpha_p = P\left(\chi^2_{p,1} > 7,8\right) = 0,271$

0,025	α_p	0,02
7,378	7,8	7,824

$$0,025 - 0,02 \longrightarrow 7,378 - 7,824$$
$$\alpha_p - 0,02 \longrightarrow 7,8 - 7,824$$

$$(\alpha_p - 0,02) \times (7,378 - 7,824) = (0,025 - 0,02) \times (7,8 - 7,824) \mapsto \alpha_p = 0,02026$$

Al ser $\alpha_p = 0,02026 < 0,05 = \alpha$ se acepta la hipótesis nula, afirmando que el tipo de universidad depende del estrato socioeconómico.

b) El grado de contingencia mide el grado de relación o dependencia: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

$C = \sqrt{\frac{G}{G+n}} = \sqrt{\frac{7,8}{7,8+98}} = 0,2715$, hay una dependencia del 27,15%.

16. La tabla adjunta refleja un análisis de la obesidad en 14 sujetos. Con un nivel de significación de 0,05, se desea analizar si existen diferencias en la prevalencia de obesidad entre hombres y mujeres o si, por el contrario, el porcentaje de obesos no varía entre sexos.

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1 **(a)**	4 **(b)**	5 **(a+b)**
Hombres	7 **(c)**	2 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

Solución:

El *test exacto de Fisher* permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no cumple las condiciones necesarias para que la aplicación del test de la Chi-cuadrado sea idónea.

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exige que al menos el 80% de los valores esperados de las celdas sean mayores que 5. De este modo, en una tabla de contingencia de 2×2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

Si las dos variables que se están analizando son dicotómicas, y la frecuencia esperada es menor que 5 en más de una celda, no resulta adecuado aplicar el test de la χ^2, aunque sí el test exacto de Fisher.

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas 2×2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla observada. Cada uno de estas probabilidades se obtiene bajo la hipótesis de independencia de las dos variables que se están analizando.

La probabilidad asociada a los datos que han sido observados viene dada por:

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

La fórmula general de la probabilidad descrita deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales de filas y columnas de la tabla observada.

El valor de la p asociado al test exacto de Fisher puede calcularse sumando las probabilidades de las tablas que resulten menores o iguales a la probabilidad de la tabla que ha sido observada.

Cuando el planteamiento es bilateral, es decir, cuando la hipótesis alternativa asume la dependencia entre las variables dicotómicas, pero sin especificar de antemano en qué sentido se producen dichas diferencias, el valor de la p se multiplica por 2.

En este caso, planteando la hipótesis nula H_0 : El sexo y ser obeso son independientes

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1 **(a)**	4 **(b)**	5 **(a+b)**
Hombres	7 **(c)**	2 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{5!\,9!\,8!\,6!}{14!\,1!\,4!\,7!\,2!} = 0{,}0599$$

Las siguientes tablas muestran todas las posibles combinaciones de frecuencias que se pueden obtener con los mismos totales de filas y columnas:

Sexo	Obesidad		Total
	Sí	No	
Mujeres	4 **(a)**	1 **(b)**	5 **(a+b)**
Hombres	4 **(c)**	5 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$p = 0{,}2098$

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{5!\,9!\,8!\,6!}{14!\,4!\,1!\,4!\,5!} = 0{,}2098$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	2 **(a)**	3 **(b)**	5 **(a+b)**
Hombres	6 **(c)**	3 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$p = 0{,}2797$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	3 **(a)**	2 **(b)**	5 **(a+b)**
Hombres	5 **(c)**	4 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$p = 0{,}4196$

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{5!\,9!\,8!\,6!}{14!\,3!\,2!\,5!\,4!} = 0{,}4196$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	0 **(a)**	5 **(b)**	5 **(a+b)**
Hombres	8 **(c)**	1 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$p = 0,0030$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	5 **(a)**	0 **(b)**	5 **(a+b)**
Hombres	3 **(c)**	6 **(d)**	9 **(c+d)**
Total	8 **(a+c)**	6 **(b+d)**	14 **(n)**

$p = 0,0280$

Sumando las probabilidades de las tablas que son menores o iguales a la probabilidad de la tabla observada ($p = 0,0599$) se tiene:

$$p = 0,0599 + 0,0030 + 0,0280 = 0,0909$$

Siendo $p - valor = 0,0909 > 0,05$ se acepta la hipótesis nula, concluyendo que el sexo y el hecho de ser obeso son independientes, es decir, no existe asociación entre las variables en estudio, con un nivel de significación $\alpha = 0,05$

Otro método de calcular el p-valor consiste en sumar las probabilidades asociadas a aquellas tablas que sean más favorables a la hipótesis alternativa de los datos observados. La tabla extrema de los datos observados es la que no se observa ninguna mujer obesa, $p = 0,0030$

$$p = 0,0599 + 0,0030 = 0,0629$$

El SPSS para el cómputo del test de Fisher, calcula el p-valor correspondiente a la alternativa bilateral ($2p = 2 \times 0,0909 = 0,1818$) y el p-valor asociado a un planteamiento unilateral ($p = 0,0909$).