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

ESTADÍSTICA TEÓRICA: CHI-CUADRADO TABLAS DE APLICACIONES CONTINGENCIA

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	$\cdots$	x_i	$\cdots$	x_k
Frecuencia observada	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k
Frecuencia esperada	(e_1)	(e_2)	$\cdots$	(e_i)	$\cdots$	(e_k)

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 sean inferiores 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 CONTINGENCIA: CONTRASTE DE DEPENDENCIA - INDEPENDENCIA

Cuando se desea comparar dos caracteres (X, Y) en una misma población que admiten las modalidades: $X(x_1, x_2, \cdots, x_i, \cdots, x_k)$ e $Y(y_1, y_2, \cdots, y_j, \cdots, y_m)$, 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	$\cdots$	y_j	$\cdots$	y_m	$\sum_{i=1}^{k} n_{i\bullet}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{km}	$n_{k\bullet}$
$\sum_{j=1}^{m} n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet m}$	n

Hipótesis nula H_0 : **No existe diferencia entre las distribuciones empíricas de X e Y**

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

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

Agrupando frecuencias observadas y esperadas en la TABLA DE CONTINGENCIA $k \times m$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	$\sum_{i=1}^{k} n_{i\bullet}$
x_1	n_{11} (e_{11})	n_{12} (e_{12})	$\cdots$	n_{1j} (e_{1j})	$\cdots$	n_{1m} (e_{1m})	$n_{1\bullet}$
x_2	n_{21} (e_{21})	n_{22} (e_{22})	$\cdots$	n_{2j} (e_{2j})	$\cdots$ $\cdots$	n_{2m} (e_{2m})	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1} (e_{i1})	n_{i2} (e_{i2})	$\cdots$	n_{ij} (e_{ij})	$\cdots$	n_{im} (e_{im})	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1} (e_{k1})	n_{k2} (e_{k2})	$\cdots$	n_{kj} (e_{kj})	$\cdots$	n_{km} (e_{km})	$n_{k\bullet}$
$\sum_{j=1}^{m} n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet m}$	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 es necesario que todas las celdas tengan frecuencias esperadas mayores que cinco, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

El estadístico de contraste observado $\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(k-1)\cdot(m-1)}$ 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 sí: $\overbrace{\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} < \overbrace{\chi^2_{\alpha,(k-1).(m-1)}}^{\text{estadístico teórico}}$

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

- **CONTRASTE DE INDEPENDENCIA**

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

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

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

TABLAS CONTINGENCIA 2 x 2 y 2 x 3

Para las tablas de contingencia 2 x 2 y 2 x 3 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

$$\chi_1^2 = \frac{n(n_{11}\,.\,n_{22} - n_{12}\,.\,n_{21})^2}{n_{1\bullet}\,.\,n_{2\bullet}\,.\,n_{\bullet 1}\,.\,n_{\bullet 2}}$$

Se acepta H_o: $\chi_1^2 < \chi_{\alpha,1}^2$ **Se rechaza H_o: $\chi_1^2 \geq \chi_{\alpha,1}^2$**

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

$$\chi_2^2 = \frac{n}{n_{1\bullet}}\left[\frac{n_{11}^2}{n_{\bullet 1}} + \frac{n_{12}^2}{n_{\bullet 2}} + \frac{n_{13}^2}{n_{\bullet 3}}\right] + \frac{n}{n_{2\bullet}}\left[\frac{n_{21}^2}{n_{\bullet 1}} + \frac{n_{22}^2}{n_{\bullet 2}} + \frac{n_{23}^2}{n_{\bullet 3}}\right] - n$$

Se acepta H_o: $\chi_2^2 < \chi_{\alpha,2}^2$ **Se rechaza H_o: $\chi_2^2 \geq \chi_{\alpha,2}^2$**

Coeficiente de CONTINGENCIA

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

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

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:

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 x 2 la corrección de Yates:

$$\chi_1^2 = \frac{n\left(\left|n_{11}.n_{22} - n_{12}.n_{21}\right| - \frac{n}{2}\right)^2}{n_{1\bullet}.n_{2\bullet}.n_{\bullet 1}.n_{\bullet 2}}$$

La corrección no es válida cuando $\left|n_{11}.n_{22} - n_{12}.n_{21}\right| \leq \frac{n}{2}$

En general, 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 $(k-1)\,x\,(m-1)$ 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_o sí: $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) < \chi^2_{\alpha,\,(k-1).(m-1)}$

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 $\boxed{b}$ y $\boxed{c}$

Un individuo es clasificado en la celdilla $\boxed{b}$ si cambia de $+$ a $-$,

en la celdilla $\boxed{a}$ cuando la respuesta es $+$ antes y después,

en la celdilla $\boxed{d}$ cuando la respuesta es $-$ antes y después.

En el test de McNemar para la significación de cambios solamente interesa conocer las celdas $\boxed{b}$ y $\boxed{c}$ 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.

Hipótesis nula

H_0 : El tratamiento no induce cambios significativos en las respuestas

- Estadístico de contraste sí $b+c<20$:

  Se acepta H_o sí $\chi^2_{McNemar}=b<\chi^2_{\alpha/2,1}$

- Estadístico de contraste si $b+c\geq 20$:

Se acepta H_o sí $\chi^2_{McNemar} = \chi^2_1 = \frac{(b-c)^2}{b+c} < \chi^2_{\alpha/2,\,1}$

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:

Se acepta H_o sí $\chi^2_{McNemar} = \chi^2_1 = \frac{(|b-c|-1)^2}{b+c} < \chi^2_{\alpha/2,\,1}$

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:

- **El coeficiente ϕ alcanza su máximo valor sólo cuando una de las dos diagonales se ha vaciado.**
- **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: $\phi = \dfrac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad 0 \le \phi \le 1$

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

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

En una TABLA DE CONTINGENCIA de 2x2 es necesario que todas las celdas tengan frecuencias esperadas mayores que cinco, 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 2x2 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.

EJERCICIOS APLICACIONES DE LA CHI-CUADRADO

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 no se puede utilizar en este caso por no tratarse de muestras pareadas (antes-después).

Estableciendo la hipótesis nula:

H_0: La situación laboral de la mujer es independiente 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 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

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 $(e_i = 10)_{i=1,\cdots,4}$

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 **sí**

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

k ≡ Número intervalos

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 DISTRIBUCIÓN DE POISSON CON PARÁMETRO DESCONOCIDO

En un laboratorio se observó el número de partículas α que llegan a una determinada zona procedente 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:

Hipótesis nula H_0 : La distribución empírica se ajusta a la Poisson

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

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

$k \equiv$ Número intervalos $\quad p \equiv$ Número parámetros a estimar

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 $n_i \equiv$ 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	

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

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

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

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:**

$$\chi_3^2 = \sum_{i=1}^{5} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{5} \frac{n_i^2}{e_i} - n =$$

$$= \frac{120^2}{148,2} + \frac{200^2}{177,8} + \frac{140^2}{106,27} + \frac{20^2}{42,7} + \frac{12^2}{15,8} - 492 = 32,31$$

- **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:

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

La tabla refleja el número de accidentes mortales de tráfico que se producen en una carretera a lo largo de un período de tiempo.

Accidentes mortales por día	0	1	2	3	4	5
Número de días	132	195	120	60	24	9

¿Se ajustan los datos a una distribución de Poisson?. Utilizar un nivel de significación 0,05

Solución:

Hipótesis nula H_0 : La distribución empírica se ajusta a la Poisson

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

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

$k \equiv$ Número intervalos $p \equiv$ Número parámetros a estimar

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

Sea la variable aleatoria X = Número de accidentes mortales por día y $n_i \equiv$ Número de días

x_i	n_i	$x_i n_i$	$P(x_i = k) = p_i$
0	132	0	0,2466
1	195	195	0,3452
2	120	240	0,2417
3	60	180	0,1128
4	24	96	0,0395
5	9	45	0,0111
	n = 540	756	

$$\overline{x} = \lambda = \frac{\sum x_i n_i}{n} = \frac{756}{540} = 1,4$$

$$P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4} \quad k = 0, \cdots, 5$$

Las probabilidades con que llegan las partículas $k = 0, \cdots, 5$ se obtienen sustituyendo los valores de k en $P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4}$ o en las tablas con $\lambda = 1,4$

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
Fr	132 133,16	195 186,43	120 130,50	60 60,90	24 21,31	9 5,97

$e_1 = 540.0,2466 = 133,16$ $\quad e_2 = 540.0,3452 = 186,43$ $\quad e_3 = 540.0,2417 = 130,5$

$e_4 = 540.0,1128 = 60,90$ $\quad e_5 = 540.0,0395 = 21,31$ $\quad e_6 = 540.0,0111 = 5,97$

Dando lugar a una tabla de contingencia 1 x 6, no teniendo que agrupar columnas contiguas al no aparecer frecuencias esperadas menor que cinco.

Los grados de libertad son cuatro: $k - p - 1 = 6 - 1 - 1 = 4$

- **Estadístico de contraste:**

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

$$= \frac{132^2}{133,16} + \frac{195^2}{186,43} + \frac{120^2}{130,5} + \frac{60^2}{60,9} + \frac{24^2}{21,31} + \frac{9^2}{5,97} - 540 = 4,87$$

- **Estadístico teórico:** $\chi^2_{0,05\,;\,4} = 9,488$

El estadístico de contraste (bondad de ajuste) es menor que el estadístico teórico (9,488), por lo que se acepta la hipótesis nula, es decir, con un nivel de significación 0,05, los accidentes mortales de tráfico diarios en la carretera se ajustan a una distribución de Poisson.

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

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 ajustar los datos obtenidos a una distribución normal $N(\mu,\sigma)$ de parámetros desconocidos, se necesitan estimar los dos parámetros recurriendo a los estimadores máximo-verosímiles: $(\hat{\mu}=\overline{x}\ ,\ \hat{\sigma}^2=\sigma_x^2)$, donde la variable aleatoria X = Número de urgencias diarias.

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 sí

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

$k \equiv$ Número intervalos $p \equiv$ Número parámetros a estimar

- **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
		$\sum_{i=1}^{6} n_i = 350$	$\sum_{i=1}^{6} x_i\, n_i = 5075$	$\sum_{i=1}^{6} x_i^2 . n_i = 86437,5$

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

- **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	$e_i = p_i . n$	$(n_i - e_i)^2$	$(n_i - e_i)^2 / e_i$
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
	$n = 350$				$\sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = 4,57$

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

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

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

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

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

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

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

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

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

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

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

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

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

En el estadístico de contraste χ^2, el número de grados de libertad es $k-p-1=(n^0\text{ intervalos})-(n^0\text{ parámetros a estimar})-1=6-2-1=3$,

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 $\chi^2_{0,05\,;\,3}=7,815$

Siendo $\chi^2_3=4,57<\chi^2_{0,05\,;\,3}=7,815$, 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 $N(14,5;\,6,06)$.

La tabla refleja el número de accidentes mortales de tráfico que se producen en una carretera a lo largo de un período de tiempo.

Accidentes mortales por día	0	1	2	3	4	5
Número de días	132	195	120	60	24	9

¿Se ajustan los datos a una distribución de Poisson?. Utilizar un nivel de significación 0,05

Solución:

Hipótesis nula H_0 : La distribución empírica se ajusta a la Poisson

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

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

$k \equiv$ Número intervalos $p \equiv$ Número parámetros a estimar

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

Sea la variable aleatoria X = Número de accidentes mortales por día y $n_i \equiv$ Número de días

x_i	n_i	$x_i n_i$	$P(x_i = k) = p_i$
0	132	0	0,2466
1	195	195	0,3452
2	120	240	0,2417
3	60	180	0,1128
4	24	96	0,0395
5	9	45	0,0111
	n = 540	756	

$$\overline{x} = \lambda = \frac{\sum x_i n_i}{n} = \frac{756}{540} = 1,4$$

$$P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4} \quad k = 0, \cdots, 5$$

Las probabilidades con que llegan las partículas $k = 0, \cdots, 5$ se obtienen sustituyendo los valores de k en $P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4}$ o en las tablas con $\lambda = 1,4$

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
Fr	132 133,16	195 186,43	120 130,50	60 60,90	24 21,31	9 5,97

$e_1 = 540.0,2466 = 133,16$ $\quad e_2 = 540.0,3452 = 186,43$ $\quad e_3 = 540.0,2417 = 130,5$

$e_4 = 540.0,1128 = 60,90$ $\quad e_5 = 540.0,0395 = 21,31$ $\quad e_6 = 540.0,0111 = 5,97$

Dando lugar a una tabla de contingencia 1 x 6, no teniendo que agrupar columnas contiguas al no aparecer frecuencias esperadas menor que cinco.

Los grados de libertad son cuatro: $k - p - 1 = 6 - 1 - 1 = 4$

- **El estadístico de contraste:**

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

$$= \frac{132^2}{133,16} + \frac{195^2}{186,43} + \frac{120^2}{130,5} + \frac{60^2}{60,9} + \frac{24^2}{21,31} + \frac{9^2}{5,97} - 540 = 4,87$$

- **El estadístico teórico:** $\chi^2_{0,05\,;\,4} = 9,488$

El estadístico de contraste (bondad de ajuste) es menor que el estadístico teórico (9,488), por lo que se acepta la hipótesis nula, es decir, con un nivel de significación 0,05, los accidentes mortales de tráfico en la carretera se ajustan a una distribución de Poisson.

CONTRASTE DE HOMOGENEIDAD

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.

Hipótesis nula: H_0 : No existe diferencia entre hombres y mujeres respecto a la opinión.

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

O bien se acepta H_0 cuando $\chi^2_{(k-1)\,.\,(m-1)} < \chi^2_{\alpha\,;\ (k-1).(m-1)}$

Se forma una tabla de contingencia 2 x 3: En 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:** $e_{ij} = p_{ij}\,.\,n = \dfrac{n_{i\bullet}\ x\ n_{\bullet j}}{n}$**, 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	$n_{i\cdot}$
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
$n_{\cdot j}$	202	140	62	n = 404

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

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

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

$$\chi^2_2=\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\frac{(84-99,5)^2}{99,5}+\frac{(78-68,96)^2}{68,96}+\frac{(37-30,53)^2}{30,53}+$$

$$+\frac{(118-102,5)^2}{102,5}+\frac{(62-71,03)^2}{71,03}+\frac{(25-31,46)^2}{31,46}=9,76$$

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 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

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.

Hipótesis nula: H_0: Las dos variables analizadas son independientes

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

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: $e_{ij}=\dfrac{n_{i\bullet}\ x\ n_{\bullet j}}{n}$

	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

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

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

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

$$+\frac{13^2}{16,2}+\frac{10^2}{19,52}-950=9,75$$

O 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 $\chi^2_{(2-1)\,.\,(4-1)}=\chi^2_3$ 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.

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

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

En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\bullet}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\bullet j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se desea conocer si existe diferencia de opinión entre fumadores y no fumadores.

Solución:

Se establecen las hipótesis:

H_0: La opinión es independiente de su condición de fumador o no fumador

H_1: La opinión no es independiente de su condición de fumador o no fumador

Se acepta H_o sí: $\chi_c^2 = \overbrace{\sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} < \overbrace{\chi^2_{\alpha,(2-1).(4-1)}}^{\text{estadístico teórico}} = \chi^2_{0,05,3}$

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\bullet}$
Fumador	60 $e_{11}=28$	50 $e_{12}=32$	20 $e_{13}=36$	10 $e_{14}=44$	140 140
No Fumador	10 $e_{21}=42$	30 $e_{22}=48$	70 $e_{23}=54$	100 $e_{24}=66$	210 210
$n_{\bullet j}$	70	80	90	110	350

$$e_{11}=\frac{140\,.\,70}{350}=28 \quad e_{12}=\frac{140\,.\,80}{350}=32 \quad e_{13}=\frac{140\,.\,90}{350}=36 \quad e_{14}=\frac{140\,.\,110}{350}=44$$

$$e_{21}=\frac{210\,.\,70}{350}=42 \quad e_{22}=\frac{210\,.\,80}{350}=48 \quad e_{23}=\frac{210\,.\,90}{350}=54 \quad e_{24}=\frac{210\,.\,110}{350}=66$$

$$\chi_c^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{n_{ij}^2}{e_{ij}} - n =$$

$$= \frac{60^2}{28} + \frac{50^2}{32} + \frac{20^2}{36} + \frac{10^2}{44} + \frac{10^2}{42} + \frac{30^2}{48} + \frac{70^2}{54} + \frac{100^2}{66} - 350 = 133,46$$

Estadístico teórico: $\chi^2_{0,05,3} = 7,815$

Siendo $\chi_c^2 = 133,46 > \chi^2_{0,05,3} = 7,815$ **se rechaza la hipótesis nula, se acepta por tanto la hipótesis alternativa, pudiendo afirmar con una significación 0,05 que la opinión sobre el tabaco depende de sí es o no fumador.**

- **Coeficiente de contingencia:** $C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\frac{133,46}{133,46 + 350}} = 0,525$

El grado de dependencia es del 52,5% por lo que la asociación entre las variables es alta. En las tablas de contingencia $k \times k$ **el valor máximo de C es** $C_{máximo} = \sqrt{\frac{k-1}{k}}$

- **Coeficiente Phi:** $\phi = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$

El estadístico Phi mide el grado de asociación entre las variables.

- **Coeficiente V de Cramer:**

$$V_{Cramer} = \sqrt{\frac{\chi_c^2}{n.\,mín(k-1, m-1)}} = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$$

En las tablas de contingencia 2×2 **es idéntico al estadístico Phi, presenta el problema de subestimar el grado de asociación entre las variables.**

- **Test G de la razón de verosimilitud:** $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$

Se acepta la hipótesis nula H_o sí: $G = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) < \chi^2_{\alpha,(2-1).(4-1)}$

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\bullet}$
Fumador	60 $e_{11} = 28$ $g_{11} = 45,7$	50 $e_{12} = 32$ $g_{12} = 22,3$	20 $e_{13} = 36$ $g_{13} = -11,7$	10 $e_{14} = 44$ $g_{14} = -14,8$	140 140
No Fumador	10 $e_{21} = 42$ $g_{21} = -14,3$	30 $e_{22} = 48$ $g_{22} = -14,1$	70 $e_{23} = 54$ $g_{23} = 18,2$	100 $e_{24} = 66$ $g_{24} = 41,6$	210 210
$n_{\bullet j}$	70	80	90	110	350

$g_{11} = 60 \ln\left(\frac{60}{28}\right) = 45,7$ $g_{12} = 50 \ln\left(\frac{50}{32}\right) = 22,3$ $g_{13} = 20 \ln\left(\frac{20}{36}\right) = -11,7$ $g_{14} = 10 \ln\left(\frac{10}{44}\right) = -14,8$

$g_{21} = 10 \ln\left(\frac{10}{42}\right) = -14,3$ $g_{22} = 30 \ln\left(\frac{30}{48}\right) = -14,1$ $g_{23} = 70 \ln\left(\frac{70}{54}\right) = 18,2$ $g_{24} = 100 \ln\left(\frac{100}{66}\right) = 41,6$

$$G = 2\sum_{i=1}^{2}\sum_{j=1}^{4} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) =$$

$$= 2\left[45,7 + 22,3 - 11,7 - 14,8 - 14,3 - 14,1 + 18,2 + 41,6\right] = 145,475$$

El test G da la razón de verosimilitud es una Prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.

- **Coeficiente Lambda (λ) de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda, de es decir el número de aciertos que proporciona el conocer la distribución divido por el número de errores sin conocerla.**

$$\lambda_{YX} = \frac{\sum m_Y - M_Y}{n - M_Y}$$

$M_Y \equiv$ Frecuencia modal global

$\sum m_Y \equiv$ Suma de frecuencias modales

$n \equiv$ Número total de casos

También, $\lambda = \dfrac{E_1 - E_2}{E_1}$ $\begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$

Valores Lambda (λ) próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.
Dos variables son independientes cuando λ = 0. Sin embargo λ = 0 no implica independencia estadística.

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\bullet}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\bullet j}$	70	80	90	110	350

$$\lambda_{YX} = \frac{\sum m_Y - M_Y}{n - M_Y} = \frac{280 - 210}{350 - 210} = 0{,}5 \quad \begin{cases} M_Y \equiv 210 \\ \sum m_Y \equiv 60 + 50 + 70 + 100 = 280 \\ n \equiv 350 \end{cases}$$

$$\lambda_{YX} = \frac{E_1 - E_2}{E_1} = \frac{140 - 70}{140} = 0{,}5 \quad \begin{cases} E_1 = n - M_y = 350 - 210 = 140 \\ E_2 = n - \sum m_y = 350 - 280 = 70 \end{cases}$$

$$\lambda_{XY} = \frac{\sum m_X - M_X}{n - M_X} = \frac{160 - 110}{350 - 110} = 0{,}208 \quad \begin{cases} M_X \equiv 110 \\ \sum m_X \equiv 60 + 100 = 160 \\ n \equiv 350 \end{cases}$$

$$\lambda_{XY} = \frac{E_1 - E_2}{E_1} = \frac{240 - 190}{240} = 0{,}208 \quad \begin{cases} E_1 = n - M_X = 350 - 110 = 240 \\ E_2 = n - \sum m_X = 350 - 160 = 190 \end{cases}$$

Un Fumador que estuviera Muy en contra de fumar en lugares públicos acertaría 60 veces de 70, es decir fallaría en 10 ocasiones. Un fumador que estuviera en contra tendría $80 - 50 = 30$ errores.

- **Coeficiente Tau de Goodman y Kruskal: Al igual que el coeficiente Lambda (λ) es un coeficiente asimétrico, aunque a diferencia del Lambda parte de los errores cometidos al asignar aleatoriamente los casos a las categorías de la variable dependiente.**

$$\tau = \frac{E_1 - E_2}{E_1} \quad \text{donde } E_1 = \sum_{i=1}^{k}\left[\frac{(n - n_{i\bullet})\, n_{i\bullet}}{n}\right] \quad \text{y} \quad E_2 = \sum_{j=1}^{m}\sum_{i=1}^{k}\left[\frac{(n_{\bullet j} - n_{ij})\, n_{ij}}{n_{\bullet j}}\right]$$

◎ **Para conocer los errores sin conocer la distribución de la variable independiente:**

Se supone que en cada categoría se clasificaran erróneamente por azar un número de casos, que en cada categoría es igual al número de casos que no pertenecen a la misma.

$$E_1 = \sum_{i=1}^{k}\left[\frac{(n - n_{i\bullet})\, n_{i\bullet}}{n}\right]$$

$\begin{cases} n \equiv \text{número total de casos} \\ k \equiv \text{número de categorías de la variable} \\ n_{i\bullet} \equiv \text{frecuencia de la categoría i-ésima} \end{cases}$

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\bullet}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\bullet j}$	70	80	90	110	350

En la categoría de Fumadores de $n_{1\bullet} = 140$ de un total de $n = 350$ se cometerían $n - n_{1\bullet} = 350 - 140 = 210$ errores.
Intentando designar al azar los $n_{1\bullet} = 140$ casos de Fumadores se cometería un error promedio de: $\frac{(n - n_{1\bullet})}{n} \times n_{1\bullet} = \frac{350 - 140}{350} \times 140 = 84$

En la categoría de No Fumadores de $n_{2\bullet} = 210$ de un total de $n = 350$ se cometerían $n - n_{2\bullet} = 350 - 210 = 140$ errores.
Intentando designar al azar los $n_{2\bullet} = 210$ casos de No Fumadores se cometería un error promedio de: $\frac{(n - n_{2\bullet})}{n} \times n_{2\bullet} = \frac{350 - 210}{350} \times 210 = 84$

$$E_1 = \sum_{i=1}^{2}\left[\frac{(350 - n_{i\bullet})\, n_{i\bullet}}{350}\right] = 84 + 84 = 168$$

◎ **Para conocer los errores conociendo la distribución de la variable independiente:**

$$E_2 = \sum_{j=1}^{m}\sum_{i=1}^{k}\left[\frac{(n_{\bullet j} - n_{ij})\, n_{ij}}{n_{\bullet j}}\right]$$

$n_{ij} \equiv$ **frecuencia de cada celdilla en la categoría i-ésima variable dependiente**

$m \equiv$ **número de categorías de la variable independiente**

$n_{\bullet j} \equiv$ **total parcial de las categorias de la variable independiente**

Categoría con la opinión Muy en contra:

Fumadores: $\frac{(n_{\bullet 1} - n_{11})\, n_{11}}{n_{\bullet 1}} = \frac{(70 - 60)\, 60}{70} = 8{,}57$

No Fumadores: $\frac{(n_{\bullet 1} - n_{21})\, n_{21}}{n_{\bullet 1}} = \frac{(70 - 10)\, 10}{70} = 8{,}57$

Errores en la categoría $E_{21} = 8{,}57 + 8{,}57 = 17{,}14$

Categoría con la opinión En contra:

Fumadores: $\frac{(n_{\bullet 2} - n_{12})\, n_{12}}{n_{\bullet 2}} = \frac{(80 - 50)\, 50}{80} = 18{,}75$

No Fumadores: $\frac{(n_{\bullet 2} - n_{22})\, n_{22}}{n_{\bullet 2}} = \frac{(80 - 30)\, 30}{80} = 18{,}75$

Errores en la categoría $E_{22} = 18{,}75 + 18{,}75 = 37{,}5$

Categoría con la opinión A favor:

Fumadores: $\frac{(n_{\bullet 3} - n_{13})\, n_{13}}{n_{\bullet 3}} = \frac{(90 - 20)\, 20}{90} = 15{,}56$

No Fumadores: $\frac{(n_{\bullet 3} - n_{23})\, n_{23}}{n_{\bullet 3}} = \frac{(90 - 70)\, 70}{90} = 15{,}56$

Errores en la categoría $E_{23} = 15,56 + 15,56 = 31,12$

- **Categoría con la opinión Muy a favor:**

Fumadores: $\frac{(n_{\bullet 4} - n_{14})\, n_{14}}{n_{\bullet 4}} = \frac{(110 - 10)\, 10}{110} = 9,09$

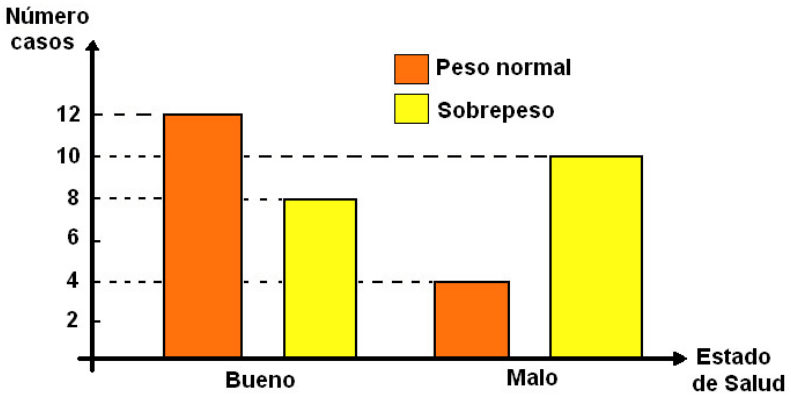
No Fumadores: $\frac{(n_{\bullet 4} - n_{24})\, n_{24}}{n_{\bullet 4}} = \frac{(110 - 100)\, 100}{110} = 9,09$

Errores en la categoría $E_{24} = 9,09 + 9,09 = 18,18$

$$E_2 = \sum_{j=1}^{4} \sum_{i=1}^{2} \left[\frac{(n_{\bullet j} - n_{ij})\, n_{ij}}{n_{\bullet j}} \right] = 17,14 + 37,5 + 31,12 + 18,18 = 103,94$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{168 - 103,94}{168} = 0,381$$

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

Analizar la existencia de una relación significativa entre el peso y el estado general de salud en el adulto mayor, con un nivel de significación del 5%,

Solución:

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		$n_{i\bullet}$
	Normal	Sobrepeso	
Bueno	12 9,41	8 10,59	20 20
Malo	4 6,59	10 7,41	14 14
$n_{\bullet j}$	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 χ^2 correspondientes a las dos observaciones, siendo la frecuencia esperada $e_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$

$$e_{11} = \frac{20 \,.\, 16}{34} = 9,41 \qquad e_{21} = \frac{14 \,.\, 16}{34} = 6,59$$

$$e_{12} = \frac{20 \,.\, 18}{34} = 10,59 \qquad e_{22} = \frac{14 \,.\, 18}{34} = 7,41$$

Estadístico de contraste:

$$\chi^2_{(2-1)\,.\,(2-1)} = \chi^2_1 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{n_{ij}^2}{e_{ij}} - n = \frac{12^2}{9,41} + \frac{8^2}{10,59} + \frac{4^2}{6,59} + \frac{10^2}{7,41} - 34 = 3,265$$

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

Como $\chi^2_1 = 3,265 < 3,841 = \chi^2_{0,05\,,\,1}$ 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 $\left|n_{11}\,.\,n_{22} - n_{12}\,.\,n_{21}\right| \leq \dfrac{n}{2}$

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

En este caso, $\chi_1^2 = \dfrac{34\left(\left|12\,x\,10 - 8\,x\,4\right| - \dfrac{34}{2}\right)^2}{20\,x\,14\,x\,16\,x\,18} = 2,125$

Como $\chi_1^2 = 2,125 < 3,841 = \chi_{0,05,1}^2$ **se acepta la hipótesis nula.**

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

$\alpha_p = P\left(\chi_{p,1}^2 > 2,125\right) = 0,273$

0,90	α_p	0,10
0,0158	2,125	2,706

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

$\alpha_p - 0,10 \longrightarrow 2,125 - 2,706$

$(\alpha_p - 0,10)\,x\,(0,0158 - 2,706) = (0,90 - 0,10)\,x\,(2,125 - 2,706) \;\mapsto\; \alpha_p = 0,273$

Al ser $\alpha_p = 0,273 > 0,05 = \alpha$ **se acepta la hipótesis nula, afirmando que el estado general de salud del adulto mayor es independiente de su peso.**

- **Test G de la razón de verosimilitud:** $G = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln\left(\dfrac{n_{ij}}{e_{ij}}\right) =$

$$= 2\left[12\ln\left(\frac{12}{9,41}\right) + 8\ln\left(\frac{8}{10,59}\right) + 4\ln\left(\frac{4}{6,59}\right) + 10\ln\left(\frac{10}{7,41}\right)\right] = 3,344$$

- **Coeficiente Phi:** $\phi = \sqrt{\dfrac{\chi_c^2}{n}} = \sqrt{\dfrac{3,265}{34}} = 0,310$

El estadístico Phi mide el grado de asociación entre las variables.

- **Coeficiente V de Cramer:**

$$V_{Cramer} = \sqrt{\frac{\chi_c^2}{n.\,mín(k-1\,,\,m-1)}} = \sqrt{\frac{3,265}{34.mín(2-1\,,\,2-1)}} = \sqrt{\frac{3,265}{34}} = 0,310$$

En tablas de contingencia $2\,x\,2$ **el estadístico Phi y V de Cramer tienen el mismo valor.**

- **Gamma de Goodman y Kruskal:** $\gamma = \dfrac{C-D}{C+D} = \dfrac{120-32}{120+32} = 0{,}579$

Estado de Salud	Peso		$n_{i\bullet}$
	Normal	Sobrepeso	
Bueno	12	8	20
Malo	4	10	14
$n_{\bullet j}$	16	18	34

Pares Concordantes: $C = \boxed{12}[10] = 120$

Pares Discordantes: $D = \boxed{8}[4] = 32$

Parejas empatadas en X: $T_X = \sum_{i=1}^{2} \dfrac{n_{i\bullet}(n_{i\bullet}-1)}{2} = \dfrac{1}{2}\left[20.19 + 14.13\right] = 281$

Parejas empatadas en Y: $T_Y = \sum_{j=1}^{2} \dfrac{n_{\bullet j}(n_{\bullet j}-1)}{2} = \dfrac{1}{2}\left[16.15 + 18.17\right] = 273$

- **Tau-C de Kendall:** $\tau_C = \dfrac{2.\text{mín}(k, m).(C-D)}{\text{mín}(k-1, m-1).n^2} = \dfrac{2.2.(120-32)}{34^2} = 0{,}304$

- **Tau-B de Kendall:** $\tau_B = \dfrac{C-D}{\sqrt{\left(\dfrac{n(n-1)}{2} - T_X\right)\left(\dfrac{n(n-1)}{2} - T_Y\right)}}$

$$\tau_B = \frac{120-32}{\sqrt{\left(\dfrac{34 \times 33}{2} - 281\right)\left(\dfrac{34 \times 33}{2} - 273\right)}} = 0{,}310$$

- **Lambda de Goodman y Kruskal:** $(X, Y) \equiv$ (Estado Salud , Peso)

$$\lambda_{YX} = \frac{\sum m_Y - M_Y}{n - M_Y} = \frac{22-20}{34-20} = 0{,}143 \qquad \begin{cases} M_Y \equiv 20 \\ \sum m_Y \equiv 12 + 10 = 22 \\ n \equiv 34 \end{cases}$$

$M_Y \equiv$ Frecuencia modal global $\quad \sum m_Y \equiv$ Suma de frecuencias modales

$$\lambda_{XY}=\frac{\sum m_x - M_x}{n - M_x}=\frac{22-18}{34-18}=0,250 \qquad \begin{cases} M_x \equiv 18 \\ \sum m_x \equiv 12+10=22 \\ n \equiv 34 \end{cases}$$

- **Tau de Goodman y Kruskal:**

Peso dependiente: $\tau_{YX}=\frac{E_1-E_2}{E_1}=\frac{16,47-14,89}{16,47}=0,096$

$$E_1=\sum_{i=1}^{2}\left[\frac{(n-n_{i\bullet})\,n_{i\bullet}}{n}\right]=\frac{(34-20)20}{34}+\frac{(34-14)14}{34}=16,47$$

$$E_2=\sum_{j=1}^{2}\sum_{i=1}^{2}\left[\frac{(n_{\bullet j}-n_{ij})\,n_{ij}}{n_{\bullet j}}\right]=$$

$$=\frac{(16-12)12}{16}+\frac{(16-4)4}{16}+\frac{(18-8)8}{18}+\frac{(18-10)10}{18}=14,89$$

Estado Salud dependiente: $\tau_{YX}=\frac{E_1-E_2}{E_1}=\frac{16,94-15,31}{16,94}=0,096$

$$E_1=\sum_{j=1}^{2}\left[\frac{(n-n_{\bullet j})\,n_{\bullet j}}{n}\right]=\frac{(34-16)16}{34}+\frac{(34-18)18}{34}=16,94$$

$$E_2=\sum_{j=1}^{2}\sum_{i=1}^{2}\left[\frac{(n_{i\bullet}-n_{ij})\,n_{ij}}{n_{i\bullet}}\right]=$$

$$=\frac{(20-12)12}{20}+\frac{(20-8)8}{20}+\frac{(14-4)4}{14}+\frac{(14-10)10}{14}=15,31$$

- **Coeficiente de Incertidumbre**

$$I(X)=\sum_{i=1}^{2}\frac{n_{i\bullet}}{n}\ln\left(\frac{n_{i\bullet}}{n}\right)=\frac{20}{34}\ln\left(\frac{20}{34}\right)+\frac{14}{34}\ln\left(\frac{14}{34}\right)=-0,677$$

$$I(Y)=\sum_{j=1}^{2}\frac{n_{\bullet j}}{n}\ln\left(\frac{n_{\bullet j}}{n}\right)=\frac{16}{34}\ln\left(\frac{16}{34}\right)+\frac{18}{34}\ln\left(\frac{18}{34}\right)=-0,691$$

$$I(XY)=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{n_{ij}}{n}\ln\left(\frac{n_{ij}}{n}\right)=\frac{12}{34}\ln\left(\frac{12}{34}\right)+\frac{8}{34}\ln\left(\frac{8}{34}\right)+\frac{4}{34}\ln\left(\frac{4}{34}\right)+\frac{10}{34}\ln\left(\frac{10}{34}\right)=$$

$$=-1,319$$

Coeficiente simétrico:

$$I=\frac{2\left[I(X)+I(Y)-I(XY)\right]}{I(X)+I(Y)}=\frac{2\left[-0,677-0,691+1,319\right]}{-0,677-0,691}=0,072$$

Estado de salud como variable dependiente:

$$I_{X/Y}=\frac{I(X)+I(Y)-I(XY)}{I(X)}=\frac{-0,677-0,691+1,319}{-0,677}=0,073$$

Peso como variable dependiente:

$$I_{Y/X}=\frac{I(X)+I(Y)-I(XY)}{I(Y)}=\frac{-0,677-0,691+1,319}{-0,691}=0,071$$

- **El coeficiente o índice de Kappa κ es una medida de concordancia propuesta por Cohen en 1960, se basa en comparar la concordancia observada en un conjunto de datos, respecto a lo que podría ocurrir por pura casualidad. Se puede calcular en tablas de cualquier dimensión, en el caso de tablas de 2×2 tiene algunas peculiaridades.**

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

Índice de Kappa: $\kappa=\frac{p_0-p_e}{1-p_e}$

$$p_0=\frac{1}{n}\sum_i n_{ii} \qquad p_e=\frac{1}{n^2}\sum_i n_{i\bullet} \times n_{\bullet i}$$

Donde p_0 es la proporción de concordancia observada y p_e es la proporción de concordancia esperada por azar.

Cuando $\kappa=1$ se da la máxima concordancia posible. El valor $\kappa=0$ indica que la concordancia observada es precisamente la que se espera por pura casualidad.

$$p_0 = \frac{1}{n}\sum_i n_{ii} = \frac{12+10}{34} = 0,647$$

$$p_e = \frac{1}{n^2}\sum_i n_{i\bullet} x\, n_{\bullet i} = \frac{20\,x\,16 + 14\,x\,18}{34^2} = 0,495$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0,647 - 0,495}{1 - 0,495} = 0,301$$

En el caso de más de dos evaluadores, clasificaciones, métodos, etc., Joseph L. Fleiss generalizó el método de Cohen, dando lugar a la Kappa de Fleiss.

En la tabla se refleja la edad de los empleados de una empresa y el grado de satisfacción en el trabajo, con un nivel de significación del 5%, contrastar si el grado de satisfacción en el trabajo no depende de la edad de los empleados.

Edad	Satisfacción en el trabajo				
	A	B	C	D	E
< 25	10	10	20	40	70
25 - 36	20	10	15	20	30
> 36	60	50	30	10	5

Solución:

Variables: X= 'edad de los empleados' e Y= 'satisfacción en el trabajo'

Hipótesis nula H_0: 'El grado de satisfacción en el trabajo no depende de la edad de los empleados'

Se acepta H_0:

$$\chi_c^2 = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{3}\sum_{j=1}^{5}\frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha;\,(3-1)\,.\,(5-1)}$$

Se forma la tabla de contingencia 3 x 5 donde cada frecuencia observada $(n_{ij})_{i=1,2,3\ ;\ j=1,\cdots,5}$ tiene una frecuencia teórica o esperada en caso de independencia $e_{ij} = \dfrac{n_{i\bullet} \times n_{\bullet j}}{n}$

Edad	Satisfacción en el trabajo					$n_{i\bullet}$
	A	B	C	D	E	
< 25	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)
25 - 36	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)
> 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)
$n_{\bullet j}$	90	70	65	70	105	400

$$e_{11}=\frac{150\,.\,90}{400}=33,75 \qquad e_{21}=\frac{95\,.\,90}{400}=21,37 \qquad e_{31}=\frac{155\,.\,90}{400}=34,87$$

$$e_{12}=\frac{150\,.\,70}{400}=26,25 \qquad e_{22}=\frac{95\,.\,70}{400}=16,62 \qquad e_{32}=\frac{155\,.\,70}{400}=27,12$$

$$e_{13}=\frac{150\,.\,65}{400}=24,37 \qquad e_{23}=\frac{95\,.\,65}{400}=15,44 \qquad e_{33}=\frac{155\,.\,65}{400}=25,19$$

$$e_{14}=\frac{150\,.\,70}{400}=26,25 \qquad e_{24}=\frac{95\,.\,70}{400}=16,62 \qquad e_{34}=\frac{155\,.\,70}{400}=27,12$$

$$e_{15}=\frac{150\,.\,105}{400}=39,37 \qquad e_{25}=\frac{95\,.\,105}{400}=24,94 \qquad e_{35}=\frac{155\,.\,105}{400}=40,69$$

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

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

$$+\left(\frac{60^2}{34,87}+\frac{50^2}{27,12}+\frac{30^2}{25,19}+\frac{10^2}{27,12}+\frac{5^2}{40,69}\right)-400=143,458$$

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

Como $\chi_8^2=143,458>15,507=\chi^2_{0,05\,;\,8}$ **se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo. En consecuencia, la edad influye significativamente en la satisfacción en el trabajo.**

🕮 ESTADÍSTICOS VARIABLES NOMINALES: FUERZA DE LA RELACIÓN

- **Coeficiente Phi:** $\phi=\sqrt{\frac{\chi_c^2}{n}}=\sqrt{\frac{143,51}{400}}=0,599$

El estadístico Phi mide el grado de asociación entre las variables.

- **Coeficiente V de Cramer:**

$$V_{Cramer}=\sqrt{\frac{\chi_c^2}{n\,.\,\text{mín}(k-1\,,\,m-1)}}=\sqrt{\frac{143,51}{400\,.\,\text{mín}(3-1\,,\,5-1)}}=\sqrt{\frac{143,51}{400\,.2}}=0,423$$

El estadístico V de Cramer es una medida simétrica que cuantifica la relación entre dos o más variables de la escala nominal. Quizás es el estadístico más utilizado.

Un valor del estadístico V de Cramer próximo a 0 indica la falta de asociación de las variables, mientras que próximo a 1 refleja mayor asociación entre las variables en estudio.

Como $V_{Cramer} = 0,423$ se detecta una relación moderada de las variables.

- **Coeficiente de contingencia:** $C = \sqrt{\dfrac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\dfrac{143,51}{143,51 + 400}} = 0,514$

El grado de dependencia es del 51,4% por lo que la asociación entre las variables es alta.

- **Test G de la razón de verosimilitud:** $G = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} \ln\left(\dfrac{n_{ij}}{e_{ij}}\right)$

Se acepta la hipótesis nula H_0 sí: $G = 2\sum_{i=1}^{3}\sum_{j=1}^{5} n_{ij} \ln\left(\dfrac{n_{ij}}{e_{ij}}\right) < \chi^2_{\alpha,\,(3-1)\,.\,(5-1)}$

	Satisfacción en el trabajo					$n_{i\bullet}$
Edad	**A**	**B**	**C**	**D**	**E**	
< 25	**10** $e_{11} = 33,75$ $g_{11} = -12,16$	**10** $e_{12} = 26,25$ $g_{12} = -9,65$	**20** $e_{13} = 24,37$ $g_{13} = -3,95$	**40** $e_{14} = 26,25$ $g_{14} = 16,85$	**70** $e_{15} = 39,37$ $g_{15} = 40,28$	**150** **(150)** **(31,37)**
25 - 36	**20** $e_{21} = 21,37$ $g_{21} = -1,33$	**10** $e_{22} = 16,62$ $g_{22} = -5,08$	**15** $e_{23} = 15,44$ $g_{23} = -0,43$	**20** $e_{24} = 16,62$ $g_{24} = 3,7$	**30** $e_{25} = 24,94$ $g_{25} = 5,54$	**95** **(95)** **(2,4)**
> 36	**60** $e_{31} = 34,87$ $g_{31} = 32,56$	**50** $e_{32} = 27,12$ $g_{32} = 30,59$	**30** $e_{33} = 25,19$ $g_{33} = 5,24$	**10** $e_{34} = 27,12$ $g_{34} = -9,98$	**5** $e_{35} = 40,69$ $g_{35} = -10,48$	**155** **(155)** **(47,93)**
$n_{\bullet j}$	**90**	**70**	**65**	**70**	**105**	**400** **(81,7)**

$$G = 2\sum_{i=1}^{3}\sum_{j=1}^{5} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) = 2.81,667 = 163,334 > 15,507 = \chi^2_{0,05;8}$$

Se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo, concluyendo que la edad influye significativamente en la satisfacción en el trabajo.

$$g_{11} = 10 \ln\left(\frac{10}{33,75}\right) = -12,16 \qquad g_{21} = 20 \ln\left(\frac{20}{21,37}\right) = -1,33 \qquad g_{31} = 60 \ln\left(\frac{60}{34,87}\right) = 32,56$$

$$g_{12} = 10 \ln\left(\frac{10}{26,25}\right) = -9,65 \qquad g_{22} = 10 \ln\left(\frac{10}{16,62}\right) = -5,08 \qquad g_{32} = 50 \ln\left(\frac{50}{27,12}\right) = 30,59$$

$$g_{13} = 20 \ln\left(\frac{20}{24,37}\right) = -3,95 \qquad g_{23} = 15 \ln\left(\frac{15}{15,44}\right) = -0,43 \qquad g_{33} = 30 \ln\left(\frac{30}{25,19}\right) = 5,24$$

$$g_{14} = 40 \ln\left(\frac{40}{26,25}\right) = 16,85 \qquad g_{24} = 20 \ln\left(\frac{20}{16,62}\right) = 3,7 \qquad g_{34} = 10 \ln\left(\frac{10}{27,12}\right) = -9,98$$

$$g_{15} = 70 \ln\left(\frac{70}{39,37}\right) = 40,28 \qquad g_{25} = 30 \ln\left(\frac{30}{24,94}\right) = 5,54 \qquad g_{35} = 5 \ln\left(\frac{5}{40,69}\right) = -10,48$$

El test G da la razón de verosimilitud es una Prueba de hipótesis que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.

🕮 MEDIDAS DE ASOCIACIÓN DE VARIABLES ORDINALES

■	Satisfacción en el trabajo					$n_{i\bullet}$
Edad	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\bullet j}$	90	70	65	70	105	400

Pares Concordantes:

$$
\begin{aligned}
C = {} & \boxed{10}[10+15+20+30+50+30+10+5] \\
& +10[15+20+30+30+10+5] \\
& +20[20+30+10+5] \\
& +40[30+5] \\
& +\boxed{20}[50+30+10+5] \\
& +10[30+10+5] \\
& +15[10+5] \\
& +20[5] \\
& =8175
\end{aligned}
$$

Pares Discordantes:

$$
\begin{aligned}
D = {} & \boxed{70}[20+10+15+20+60+50+30+10] \\
& +40[20+10+15+60+50+30] \\
& +20[20+10+60+50] \\
& +10[20+60] \\
& +\boxed{30}[60+50+30+10] \\
& +20[60+50+30] \\
& +15[60+50] \\
& +10[60] \\
& =35600
\end{aligned}
$$

- **La Gamma de Goodman y Kruskal γ mide la fuerza de asociación de los datos cuando las variables se miden en el nivel ordinal.**

$\gamma = 0$ indica la ausencia de asociación.

$$\gamma = \frac{C-D}{C+D} \quad -1 \le \gamma \le 1$$

$$\gamma = \frac{C-D}{C+D} = \frac{8175-35600}{8175+35600} = -0,626$$

- **El coeficiente de rango de Kendall (τ_C) a menudo se utiliza como un estadístico control en una prueba de hipótesis estadística para establecer si dos variables pueden considerarse estadísticamente dependientes.**

Es una prueba no paramétrica, ya que no se basa en suposiciones sobre las distribuciones de X o Y o la distribución de (X, Y).

Bajo la hipótesis nula de independencia de X e Y, la distribución muestral de Tau-C (τ_C) tiene un valor esperado de cero.

Para muestras pequeñas: $\tau_C = \dfrac{2.\text{mín}(k, m).(C-D)}{\text{mín}(k-1, m-1).n^2}$

En muestras grandes, se utiliza una aproximación a N(0, 1): $\tau_C = \dfrac{2(2n+5)}{9n(n-1)}$

$$\tau_C = \frac{2.\text{mín}(k, m).(C-D)}{\text{mín}(k-1, m-1).n^2} = \frac{2.\text{mín}(3, 5).(8175-35600)}{\text{mín}(3-1, 5-1).400^2} = \frac{2.3.(-27425)}{2.400^2} = -0,514$$

- **Parejas empatadas en X o en Y:**

$$T_X = \sum_{i=1}^{k} \frac{n_{i\bullet}(n_{i\bullet}-1)}{2} \qquad T_Y = \sum_{j=1}^{m} \frac{n_{\bullet j}(n_{\bullet j}-1)}{2}$$

$$T_X = \sum_{i=1}^{3} \frac{n_{i\bullet}(n_{i\bullet}-1)}{2} = \frac{1}{2}\left[150.149 + 95.94 + 155.154\right] = 27575$$

$$T_Y = \sum_{j=1}^{5} \frac{n_{\bullet j}(n_{\bullet j}-1)}{2} = \frac{1}{2}\left[90.89 + 70.69 + 65.64 + 70.69 + 105.104\right] = 16375$$

- **El coeficiente Tau-B de Kendall (τ_B) es una medida no paramétrica de la correlación para variables ordinales o de rangos que tiene en consideración los empates.**

El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza de la relación. Varía entre -1 y 1 según sea el sentido de la asociación entre las variables. Los valores mayores indican que la relación es más estrecha.

Cuando la tabla no es cuadrada este coeficiente no puede llegar a valer 1 dado que existirán más pares empatados en la variable que tenga más categorías.

$$\tau_B = \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_X\right)\left(\frac{n(n-1)}{2} - T_Y\right)}}$$

$$\tau_B = \frac{8175 - 35600}{\sqrt{(79800 - 27575)(79800 - 16375)}} = -0,477$$

- **El estadístico D de Somers establece si las variables ordinales son dependientes o independientes entre sí.**

El coeficiente D de Somers varía entre -1 y 1, es una medida asimétrica como el coeficiente Lambda, los dos valores que se pueden obtener de la tabla dependen de que se tome como independiente la variable X o Y.

Valores del estadístico D cercanos a 0 indican que no hay ninguna o muy poca asociación entre las variables.

D de Somers: $D_X = \frac{C - D}{\frac{n(n-1)}{2} - T_X}$ $\quad D_Y = \frac{C - D}{\frac{n(n-1)}{2} - T_Y}$

Número de pares: $\binom{n}{2} = \frac{n(n-1)}{2} = \frac{400(400-1)}{2} = 79800$

$$D_X = \frac{C - D}{\frac{n(n-1)}{2} - T_X} = \frac{8175 - 35600}{79800 - 27575} = -0,525$$

$$D_Y = \frac{C - D}{\frac{n(n-1)}{2} - T_X} = \frac{8175 - 35600}{79800 - 16375} = -0,432$$

🕮 MEDIDAS BASADAS EN EL ERROR PROPORCIONAL

- **Coeficiente Lambda (λ) de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda.**

Estadístico utilizado para determinar si usar los resultados de una de las variables puede utilizarse para predecir los resultados de la otra variable.

Valores Lambda (λ) próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.

Dos variables son independientes tienen λ = 0. Sin embargo λ = 0 no implica independencia estadística.

	Satisfacción en el trabajo					$n_{i\bullet}$
Edad	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\bullet j}$	90	70	65	70	105	400

$$\lambda_{YX} = \frac{\sum m_Y - M_Y}{n - M_Y} \quad \begin{cases} M_Y \equiv \text{Frecuencia modal global} \\ \sum m_Y \equiv \text{Suma de frecuencias modales} \\ n \equiv \text{Número total de casos} \end{cases}$$

También, $\lambda = \dfrac{E_1 - E_2}{E_1}$ $\begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$

$$\lambda_{YX} = \frac{\sum m_Y - M_Y}{n - M_Y} = \frac{250 - 155}{400 - 155} = 0{,}388 \quad \begin{cases} M_Y \equiv 155 \\ \sum m_Y \equiv 60 + 50 + 30 + 40 + 70 = 250 \\ n \equiv 400 \end{cases}$$

$$\lambda_{XY} = \frac{\sum m_x - M_x}{n - M_x} = \frac{160 - 105}{400 - 105} = 0,186 \quad \begin{cases} M_x \equiv 105 \\ \sum m_x \equiv 70 + 30 + 60 = 160 \\ n \equiv 400 \end{cases}$$

- **Tau de Goodman y Kruskal (τ) considera todas las categorías de respuesta y no únicamente la que contempla más casos entre dos variables nominales (variables cualitativas).**

El valor de Tau de Goodman y Kruskal (τ) se interpreta como el porcentaje que mejora el error al incluir la variable independiente en la predicción de los valores de la variable dependiente.

Se parece a la Lambda (λ), siendo su cálculo más complejo. Lo mismo que Lambda adopta valores entre 0 y 1, dónde 0 es independencia y 1 el total de dependencia.

$$\tau = \frac{E_1 - E_2}{E_1}$$

❶ Errores sin conocer la distribución de la variable independiente:

$$E_1 = \sum_{i=1}^{k} \left[\frac{(n - n_{i\bullet})\, n_{i\bullet}}{n} \right] \quad \begin{cases} n \equiv \text{número total de casos} \\ k \equiv \text{número de categorías de la variable} \\ n_{i\bullet} \equiv \text{frecuencia de la categoría i-ésima} \end{cases}$$

❷ Errores conociendo la distribución de la variable independiente:

$$E_2 = \sum_{j=1}^{m} \sum_{i=1}^{k} \left[\frac{(n_{\bullet j} - n_{ij})\, n_{ij}}{n_{\bullet j}} \right]$$

$n_{ij} \equiv$ frecuencia de cada celdilla en la categoría i-ésima variable dependiente

$m \equiv$ número de categorías de la variable independiente

$n_{\bullet j} \equiv$ total parcial de las categorias de la variable independiente

	Satisfacción en el trabajo					$n_{i\bullet}$
Edad	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\bullet j}$	90	70	65	70	105	400

$$E_1 = \sum_{i=1}^{3}\left[\frac{(n-n_{i\bullet})\,n_{i\bullet}}{n}\right] = \frac{(400-150)150}{400} + \frac{(400-95)95}{400} + \frac{(400-155)155}{400} = 261{,}125$$

$$E_2 = \sum_{j=1}^{5}\sum_{i=1}^{3}\left[\frac{(n_{\bullet j}-n_{ij})\,n_{ij}}{n_{\bullet j}}\right] = \frac{(90-10)10}{90} + \frac{(90-20)20}{90} + \frac{(90-60)60}{90}$$
$$+\frac{(70-10)10}{70} + \frac{(70-10)10}{70} + \frac{(70-50)50}{70}$$
$$+\frac{(65-20)20}{65} + \frac{(65-15)15}{65} + \frac{(65-30)30}{65}$$
$$+\frac{(70-40)40}{70} + \frac{(70-20)20}{70} + \frac{(70-10)10}{70}$$
$$+\frac{(105-70)70}{105} + \frac{(105-30)30}{105} + \frac{(105-5)5}{105}$$
$$= 206{,}93$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{261{,}125 - 206{,}93}{261{,}125} = 0{,}208$$ **edad variable dependiente**

Cuando la variable dependiente es la satisfacción en el trabajo:

$$E_1 = \sum_{j=1}^{5}\left[\frac{(n-n_{\bullet j})\,n_{\bullet j}}{n}\right] =$$

$$= \frac{(400-90)90}{400} + \frac{(400-70)70}{400} + \frac{(400-65)65}{400} + \frac{(400-70)70}{400} + \frac{(400-105)105}{400} = 317{,}125$$

$$E_2 = \sum_{i=1}^{3}\sum_{j=1}^{5}\left[\frac{(n_{i\bullet} - n_{ij})\, n_{ij}}{n_{i\bullet}}\right] =$$

$$= \frac{(150-10)10}{150} + \frac{(150-10)10}{150} + \frac{(150-20)20}{150} + \frac{(150-40)40}{150} + \frac{(150-70)70}{150}$$
$$+ \frac{(95-20)20}{95} + \frac{(95-10)10}{95} + \frac{(95-15)15}{95} + \frac{(95-20)20}{95} + \frac{(95-30)30}{95}$$
$$+ \frac{(155-60)60}{155} + \frac{(155-50)50}{155} + \frac{(155-30)30}{155} + \frac{(155-10)10}{155} + \frac{(155-5)5}{155}$$
$$= 285{,}38$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{317{,}125 - 285{,}38}{317{,}125} = 0{,}100$$ **satisfacción variable dependiente**

- **El Coeficiente de Incertidumbre es una medida de asociación basada en la reducción proporcional del error. Es una medida semejante a Lambda en cuanto a su concepción de la asociación de las variables, en relación a la capacidad predictiva y la disminución del error de dicha predicción.**

El coeficiente de incertidumbre (I) depende de toda la distribución y no sólo de los valores modales (caso de Lambda), varía entre 0 y 1, tomando el valor 0 en el caso total de independencia. Es más difícil de interpretar que Lambda.

Tiene versiones asimétricas (dependiendo de cual de las dos variables sea dependiente) y una simétrica (donde no se distingue entre variable dependiente e independiente).

La versión asimétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de una variable a partir de los de valores de la otra variable.

La versión simétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de cualquiera de las dos variables mediante la tabla de contingencia.

Se obtiene mediante la fórmula: $I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)}$

Para obtener $I_{X/Y}$ **basta con intercambiar los papeles de I(X) e I(Y).**

La versión simétrica: $I = \dfrac{2\left[I(X)+I(Y)-I(XY)\right]}{I(X)+I(Y)}$

donde:

$$I(X) = \sum_{i=1}^{k} \frac{n_{i\bullet}}{n} \ln\left(\frac{n_{i\bullet}}{n}\right) \qquad I(Y) = \sum_{j=1}^{m} \frac{n_{\bullet j}}{n} \ln\left(\frac{n_{\bullet j}}{n}\right) \qquad I(XY) = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right)$$

	Satisfacción en el trabajo					$i_{i\bullet}$
Edad	**A**	**B**	**C**	**D**	**E**	
< 25	**10** $i_{11} = -0{,}092$	**10** $i_{12} = -0{,}092$	**20** $i_{13} = -0{,}150$	**40** $i_{14} = -0{,}230$	**70** $i_{15} = -0{,}305$	**150** $-0{,}368$
25 - 36	**20** $i_{21} = -0{,}150$	**10** $i_{22} = -0{,}092$	**15** $i_{23} = -0{,}123$	**20** $i_{24} = -0{,}150$	**30** $i_{25} = -0{,}194$	**95** $-0{,}341$
> 36	**60** $i_{31} = -0{,}284$	**50** $i_{32} = -0{,}260$	**30** $i_{33} = -0{,}194$	**10** $i_{34} = -0{,}092$	**5** $i_{35} = -0{,}055$	**155** $-0{,}367$
$i_{\bullet j}$	**90** $-0{,}335$	**70** $-0{,}305$	**65** $-0{,}295$	**70** $-0{,}305$	**105** $-0{,}351$	**400**

$$i_{1\bullet} = \frac{150}{400}\ln\left(\frac{150}{400}\right) = -0{,}368 \qquad i_{2\bullet} = \frac{95}{400}\ln\left(\frac{95}{400}\right) = -0{,}341 \qquad i_{3\bullet} = \frac{155}{400}\ln\left(\frac{155}{400}\right) = -0{,}367$$

$$I(X) = \sum_{i=1}^{3} \frac{n_{i\bullet}}{n} \ln\left(\frac{n_{i\bullet}}{n}\right) = -0{,}368 - 0{,}341 - 0{,}367 = -1{,}076$$

$$i_{\bullet 1} = \frac{90}{400}\ln\left(\frac{90}{400}\right) = -0{,}335 \qquad i_{\bullet 2} = \frac{70}{400}\ln\left(\frac{70}{400}\right) = -0{,}305 \qquad i_{\bullet 3} = \frac{65}{400}\ln\left(\frac{65}{400}\right) = -0{,}295$$

$$i_{\bullet 4} = \frac{70}{400}\ln\left(\frac{70}{400}\right) = -0{,}305 \qquad i_{\bullet 5} = \frac{105}{400}\ln\left(\frac{105}{400}\right) = -0{,}351$$

$$I(Y) = \sum_{j=1}^{5} \frac{n_{\bullet j}}{n} \ln\left(\frac{n_{\bullet j}}{n}\right) = -0{,}335 - 0{,}305 - 0{,}295 - 0{,}305 - 0{,}351 = -1{,}591$$

$$i_{11} = \frac{10}{400}\ln\left(\frac{10}{400}\right) = -0{,}092 \qquad i_{21} = \frac{20}{400}\ln\left(\frac{20}{400}\right) = -0{,}150 \qquad i_{31} = \frac{60}{400}\ln\left(\frac{60}{400}\right) = -0{,}284$$

$$i_{12} = \frac{10}{400}\ln\left(\frac{10}{400}\right) = -0{,}092 \qquad i_{22} = \frac{10}{400}\ln\left(\frac{10}{400}\right) = -0{,}092 \qquad i_{32} = \frac{50}{400}\ln\left(\frac{50}{400}\right) = -0{,}260$$

$$i_{13}=\frac{20}{400}\ln\left(\frac{20}{400}\right)=-0,150 \quad i_{23}=\frac{15}{400}\ln\left(\frac{15}{400}\right)=-0,123 \quad i_{32}=\frac{30}{400}\ln\left(\frac{30}{400}\right)=-0,194$$

$$i_{14}=\frac{40}{400}\ln\left(\frac{40}{400}\right)=-0,230 \quad i_{24}=\frac{20}{400}\ln\left(\frac{20}{400}\right)=-0,150 \quad i_{34}=\frac{10}{400}\ln\left(\frac{10}{400}\right)=-0,092$$

$$i_{15}=\frac{70}{400}\ln\left(\frac{70}{400}\right)=-0,305 \quad i_{25}=\frac{30}{400}\ln\left(\frac{30}{400}\right)=-0,194 \quad i_{35}=\frac{5}{400}\ln\left(\frac{5}{400}\right)=-0,055$$

$$I(XY)=\sum_{i=1}^{3}\sum_{j=1}^{5}\frac{n_{ij}}{n}\ln\left(\frac{n_{ij}}{n}\right)=-2,463$$

Coeficiente de Incertidumbre, Satisfacción como variable dependiente:

$$I_{Y/X}=\frac{I(X)+I(Y)-I(XY)}{I(Y)}=\frac{-1,076-1,591+2,463}{-1,591}=0,128$$

Coeficiente de Incertidumbre, Edad como variable dependiente:

$$I_{X/Y}=\frac{I(X)+I(Y)-I(XY)}{I(X)}=\frac{-1,076-1,591+2,463}{-1,076}=0,190$$

Coeficiente de Incertidumbre simétrico:

$$I=\frac{2\left[I(X)+I(Y)-I(XY)\right]}{I(X)+I(Y)}=\frac{2\left[-1,076-1,591+2,463\right]}{-1,076-1,591}=0,153$$

H_0: Las variables son independientes

Pruebas significación estadística
- **Chi-cuadrado de Pearson**
- **Razón de verosimilitud Chi-cuadrado**

H_0: La asociación entre las variables es nula (son independientes)

Estadísticos Nominales Variables Cualitativas
- **Phi**
- **Coeficiente de Contingencia**
- **V de Cramer**
- **Lambda**
- **Coeficiente de Incertidumbre**
- **Q de Yule**

H_0: La asociación entre las variables es nula (son independientes)

Estadísticos Ordinales Variables Cuantitativas
- **Gamma de Goodman y Kruskal**
- **D de Somers**
- **Tau-B de Kendall**
- **Tau-C de Kendall**
- **Riesgo relativo**

Análogos a las medidas de asociación, aplicables a las variables que se computan en función de acuerdos-desacuerdos o concordancias-discrepancias

Pruebas de Concordancia
- **Índice de Concordancia**
- **Coeficiente Kappa de Cohen**

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	4	5
Hombres	7	2	9
Total	8	6	14

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 exigen 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.

Probabilidad asociada a los datos que han sido observados:

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

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 contraste bilateral asume que la hipótesis alternativa establezca la dependencia entre las variables dicotómicas, pero sin especificar de antemano en qué sentido se producen dichas diferencias.

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, $p = 0,0030$

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

SPSS para el cómputo del test de Fisher, calcula el p-valor correspondiente a un contraste bilateral ($p = 0,0909$) y el p-valor asociado a un contraste unilateral ($p = 0,0629$).

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	50	96
B	85	59	144
Total	131	109	240

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.

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)

Hipótesis nula

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

En esta prueba para la significación de cambios solo interesa conocer las celdas que presentan cambios (celdas b y c) y siendo $(b+c)$ el número de personas que cambiaron, de acuerdo con la hipótesis nula planteada se espera que $\left(\frac{b+c}{2}\right)$ casos cambien en una dirección y $\left(\frac{b+c}{2}\right)$ casos a otra dirección.

- **Estadístico de contraste sí $b+c<20$**

Se acepta H_o sí $\chi^2_{McNemar}=b<\chi^2_{\alpha/2,\,1}$

- **Estadístico de contraste sí $b+c \geq 20$:** $\chi^2_{McNemar} = \chi^2_1 = \dfrac{\left[|b-c|-1\right]^2}{b+c}$

La aproximación muestral a la distribución Chi-cuadrado llega a ser muy buena si se realiza una corrección por continuidad, considerando que se utiliza una distribución continua para aproximar una distribución discreta (binomial), por lo que se realiza la corrección de Yates.

Se acepta H_o sí $\chi^2_{McNemar} = \chi^2_1 = \dfrac{\left[|b-c|-1\right]^2}{b+c} < \chi^2_{\alpha/2,\,1}$

En este caso, $b+c = 50+85 = 135 > 20$

Estadístico muestral: $\chi^2_{McNemar} = \dfrac{\left[|50-85|-1\right]^2}{50+85} = 8{,}563$

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

Como $\chi^2_{McNemar} = 8{,}563 > 5{,}024 = \chi^2_{0{,}025,1}$ **se rechaza la hipótesis nula, concluyendo que la intención de voto cambió significativamente después del debate, con un nivel de significación del 5%.**

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

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 $(4-1)\text{x}(2-1) = 3$

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

0,05	α_p	0,025
7,815	8	9,348

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

$\alpha_p - 0,025 \longrightarrow 8 - 9,348$

$(\alpha_p - 0,025)\ \text{x}\ (7,815 - 9,348) = (0,05 - 0,025)\ \text{x}\ (8 - 9,348) \mapsto \alpha_p = 0,0469$

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 $(2-1)\text{x}(3-1) = 2$

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

0,90	α_p	0,10
0,211	4,5	4,605

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

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

$(\alpha_p - 0{,}10)\ x\ (0{,}211 - 4{,}605) = (0{,}90 - 0{,}10)\ x\ (4{,}5 - 4{,}605) \mapsto \alpha_p = 0{,}119$

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 $(4-1)x(3-1) = 6$

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

0,90	α_p	0,10
2,204	6,1	10,645

$0{,}90 - 0{,}10 \longrightarrow 2{,}204 - 10{,}645$

$\alpha_p - 0{,}10 \longrightarrow 6{,}1 - 10{,}645$

$(\alpha_p - 0{,}10)\ x\ (2{,}204 - 10{,}645) = (0{,}90 - 0{,}10)\ x\ (6{,}1 - 10{,}645) \mapsto \alpha_p = 0{,}530$

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.