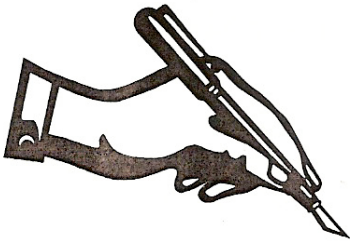
Estadística Teórica II



TABLAS DE CONTINGENCIA

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, rechazamos la hipótesis nula y se concluye 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

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

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	

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

$\lambda = 1{,}2$

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 la fórmula $P(x_i = k) = \frac{1{,}2^k}{k!} e^{-1.2}$, 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, tenemos que calcular las frecuencias esperadas $(e_i = np_i)$

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 $(e_6 = 3,05)$

$$e_1 = 492.0,3012 = 148,2 \qquad e_2 = 492.0,3614 = 177,8 \qquad e_3 = 492.0,2169 = 106,7$$
$$e_4 = 492.0,0867 = 42,7 \qquad e_5 = 492.0,0260 = 12,8 \qquad e_6 = 492.0,0062 = 3,05$$

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

En el test de la χ^2 hay que tener en cuenta las siguientes consideraciones:

- *Hay que tener en cuenta el número de modalidades que admite el carácter, ya que al haber más modalidades la χ^2 va siendo cada vez más grande.*
- *Para emplearlo correctamente es necesario que las frecuencias esperadas de las distintas modalidades no sea inferior a cinco.*

- *Si existe alguna modalidad que tenga una frecuencia esperada menor que cinco se agrupan dos o más modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.*
- *Si para obtener las frecuencias esperadas se necesitan hallar p parámetros entonces los grados de libertad de la χ^2 son (k - p) si son independientes, y (k – p - 1) si son excluyentes las modalidades.*
- *Se puede aplicar el test de la χ^2 en situaciones en las que se quiere decidir si una serie de observaciones se ajusta o no a una función teórica previamente determinada (binomial, Poisson, normal, o hipotética).*

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 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 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

1) 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
		$n = \sum_{i=1}^{6} n_i = 350$	$\sum_{i=1}^{6} x_i n_i = 5075$	$\sum_{i=1}^{6} x_i^2 n_i = 86437,5$

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

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

$$P(0 < x < 5) = P\left[\frac{0-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{5-14{,}5}{6{,}06}\right] = P(-2{,}39 < z < -1{,}57) =$$
$$= P(1{,}57 < z < 2{,}39) = P(z > 1{,}57) - P(z > 2{,}39) = 0{,}0582 - 0{,}00842 = 0{,}04978$$

$$P(5 < x < 10) = P\left[\frac{5-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{10-14{,}5}{6{,}06}\right] = P(-1{,}57 < z < -0{,}74) =$$
$$= P(0{,}74 < z < 1{,}57) = P(z > 0{,}74) - P(z > 1{,}57) = 0{,}2296 - 0{,}0582 = 0{,}1714$$

$$P(10 < x < 15) = P\left[\frac{10-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{15-14{,}5}{6{,}06}\right] = P(-0{,}74 < z < 0{,}08) =$$
$$= P(0{,}08 < z < 0{,}74) = 1 - P(z > 0{,}74) - P(z > 0{,}08) = 1 - 0{,}4681 - 0{,}2296 = 0{,}3023$$

$$P(15 < x < 20) = P\left[\frac{15-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{20-14{,}5}{6{,}06}\right] = P(0{,}08 < z < 0{,}91) =$$
$$= P(z > 0{,}08) - P(z > 0{,}91) = 0{,}4681 - 0{,}1814 = 0{,}2867$$

$$P(20 < x < 25) = P\left[\frac{20-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{25-14{,}5}{6{,}06}\right] = P(0{,}91 < z < 1{,}73) =$$
$$= P(z > 0{,}91) - P(z > 1{,}73) = 0{,}1814 - 0{,}0418 = 0{,}1396$$

$$P(25 < x < 30) = P\left[\frac{25-14{,}5}{6{,}06} < \frac{x-14{,}5}{6{,}06} < \frac{30-14{,}5}{6{,}06}\right] = P(1{,}73 < z < 2{,}56) =$$
$$= P(z > 1{,}73) - P(z > 2{,}56) = 0{,}0418 - 0{,}0052 = 0{,}0366$$

3) Se calculan las frecuencias esperadas, multiplicando las probabilidades por el número total $e_i = p_i . n$

4) 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,

$$\chi_3^2 = \sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = 4{,}57$$

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

Como $\chi_3^2 = 4{,}57 \;<\; \chi^2_{0{,}05\,;\,3} = 7{,}815$, 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.

Tenemos 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: $R = \left\{ \chi^2_{(k-1).(m-1)} \geq \chi^2_{\alpha\,;\,(k-1).(m-1)} \right\}$

Se forma una tabla de contingencia 2 x 3 de la siguiente forma: en cada frecuencia observada $(n_{ij})_{i=1,\cdots,k\,;\,j=1,\cdots,m}$ en la tabla de contingencia tenemos una frecuencia teórica o esperada e_{ij} que se calcula mediante la expresión: $e_{ij} = p_{ij}\, n = \dfrac{n_{x_i}\, n_{y_j}}{n}$, 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	Total (n_{x_i})
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

Total (n_{y_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: $\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \chi^2_{(k-1).(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - n$ (manera útil)

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

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, esto es, 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: H_0 : 'Las dos variables estudiadas son independientes'
H_1 : 'Existe dependencia entre las dos variables'

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

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

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:

$$e_{ij} = p_{ij}\,n = \frac{n_{x_i}\,n_{y_j}}{n}$$

	Coeficiente Intelectual				Total (n_{x_i})
	< 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

Total (n_{y_j})	276	255	190	229	950

$$e_{11} = \frac{869.276}{950} = 252,46 \quad e_{12} = \frac{869.255}{950} = 233,25 \quad e_{13} = \frac{869.190}{950} = 173,8 \quad e_{14} = \frac{869.229}{950} = 209,47$$

$$e_{21} = \frac{81.276}{950} = 23,53 \quad e_{22} = \frac{81.255}{950} = 21,74 \quad e_{23} = \frac{81.190}{950} = 16,2 \quad e_{24} = \frac{81.229}{950} = 19,52$$

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 $\chi_3^2 = 9,75 > \chi^2_{0,10\,;\,3} = 6,251$ 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; si 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) Sea la hipótesis nula H_0='No existe diferencia entre los métodos de empaquetado'.

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

Formamos la tabla de contingencia 3 x 4 , donde $e_{ij} = \dfrac{n_{x_i} \cdot n_{y_j}}{n}$

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

$$e_{11} = \frac{26.31}{127} = 6,35 \quad e_{21} = \frac{32.31}{127} = 7,81 \quad e_{31} = \frac{30.31}{127} = 7,32 \quad e_{41} = \frac{39.31}{127} = 9,52$$

$$e_{12} = \frac{26.44}{127} = 9,01 \quad e_{22} = \frac{32.44}{127} = 11,09 \quad e_{32} = \frac{30.44}{127} = 10,39 \quad e_{42} = \frac{39.44}{127} = 13,51$$

$$e_{13} = \frac{26.52}{127} = 10,65 \quad e_{23} = \frac{32.52}{127} = 13,10 \quad e_{33} = \frac{30.52}{127} = 12,28 \quad e_{43} = \frac{39.52}{127} = 15,97$$

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 $\chi^2_{0,05\,;\,6} = 12,592$

Como $\chi^2_6 = 1,24 < \chi^2_{0,05\,;\,6} = 12,592$, el estadístico observado es menor que el estadístico teórico o esperado, 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:

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

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

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 $e_{ij} = p_{ij}\,n = \dfrac{n_{x_i}\,n_{y_j}}{n}$

Satisfacción en el trabajo

	Muy satisfecho	Satisfecho	Insatisfecho	Muy insatisfecho	Total
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 en los tras países.

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.-

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

Se acepta H_0:

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

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

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

$$e_{ij} = p_{ij}\,n = \frac{n_{x_i}\,n_{y_j}}{n}$$

Número de accidentes por año

Edad del conductor	0	1	2	3	4	n_{x_i}
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)
n_{y_j}	90	70	65	70	105	400

donde,

$$e_{11} = \frac{150.90}{400} = 33,75 \qquad e_{21} = \frac{95.90}{400} = 21,37 \qquad e_{31} = \frac{155.90}{400} = 34,87$$

$$e_{12} = \frac{150.70}{400} = 26,25 \qquad e_{22} = \frac{95.70}{400} = 16,62 \qquad e_{32} = \frac{155.70}{400} = 27,12$$

$$e_{13} = \frac{150.65}{400} = 24,37 \qquad e_{23} = \frac{95.65}{400} = 15,44 \qquad e_{33} = \frac{155.65}{400} = 25,19$$

$$e_{14} = \frac{150.70}{400} = 26,25 \qquad e_{24} = \frac{95.70}{400} = 16,62 \qquad e_{34} = \frac{155.70}{400} = 27,12$$

$$e_{15} = \frac{150.105}{400} = 39,37 \qquad e_{25} = \frac{95.105}{400} = 24,94 \qquad e_{35} = \frac{155.105}{400} = 40,69$$

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

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

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

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; es decir, la edad influye significativamente en el número de accidentes al año.

COEFICIENTE DE CONTINGENCIA

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

Se pide:

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) El COEFICIENTE DE CONTINGENCIA es una medida de grado de relación o dependencia entre dos caracteres en una tabla de contingencia, dada por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{donde } C \leq 1$$

Cuánto mayor sea el valor de C más alto es el grado de dependencia entre dos caracteres (X, Y).

Tenemos que hallar primero los valores de la χ^2 correspondientes a las dos observaciones $e_{ij} = \frac{n_{x_i} \cdot n_{y_j}}{n}$. Para la población A, la tabla de contingencia 2 x 2:

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

$$e_{11} = \frac{70.78}{194} = 28,14 \quad e_{12} = \frac{70.116}{194} = 41,85 \quad e_{21} = \frac{124.78}{194} = 49,85 \quad e_{22} = \frac{124.116}{194} = 74,14$$

Estadístico de contraste:

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

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

Para 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

$$e_{11} = \frac{84.96}{206} = 39,15 \quad e_{12} = \frac{84.110}{206} = 44,85 \quad e_{21} = \frac{96.122}{206} = 56,85 \quad e_{22} = \frac{110.122}{206} = 65,15$$

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: $C_B = \sqrt{\frac{17,82}{17,82 + 206}} = 0,282$

a) 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.