Tema 9: Contrastes de hipótesis no paramétricos

1. A partir de los siguientes datos, que muestran el número de accidentes en un determinado regimiento del ejército durante 200 días elegidos al azar, contrastar si el número de accidentes diarios sigue una distribución de Poisson de parámetro 2.

Nº de accidentes	0	1	2	3	4	5	6	7
Nº de días	22	53	58	39	20	5	2	1

2. Una tela cuadrada tiene 60 defectos de fabricación. Con objeto de analizar la distribución de los defectos en la superficie de la tela, se ha dividido esta en 9 zonas cuadradas exactamente iguales, observándose los siguientes defectos en cada zona:

8	7	3
5	9	11
6	4	7

Contrastar, a partir de estos datos, si los defectos se distribuyen uniformemente en toda la superficie o, por el contrario, siguen algún patrón de ocurrencia.

3. Un modelo genético indica que la distribución de una población de hombres y mujeres, daltónicos o no, se ajusta a probabilidades de la forma:

	Hombres	*Mujeres*
No daltónicos	$(1-p)/2$	$(1-p^2)/2$
Daltónicos	$p/2$	$p^2/2$

Para comprobar esta teoría se examinaron 2000 individuos de la población, elegidos al azar, obteniéndose los siguientes resultados:

	Hombres	*Mujeres*
No daltónicos	894	1015
Daltónicos	81	10

Contrastar, mediante el test de la χ^2, si esta muestra concuerda con el modelo teórico.

4. Un laboratorio farmacéutico afirma que uno de sus productos confiere inmunidad a la picadura de insectos durante un tiempo exponencial de media 2.5 horas. Probado el producto en 20 sujetos, en un ambiente con gran número de mosquitos, los instantes (en horas) en que recibieron la primera picadura fueron:

0.01, 0.02 0.03, 0.23, 0.51, 0.74, 0.96, 1.17, 1.46, 1.62,
2.18, 2.25, 2.79, 3.45, 3.83, 3.92, 4.27, 5.43, 5.79, 6.34.

Usando el test de Kolmogorov-Smirnov, contrastar, a partir de estos datos, si puede aceptarse la afirmación del laboratorio.

5. Cierta comunidad ha modificado la procedencia del agua destinada al consumo doméstico. Se sabe que, con el antiguo suministro, la distribución de la cantidad de sodio por unidad de volumen de sangre de sus habitantes es simétrica alrededor de 3.24 gr. Tras cierto tiempo, se quiere comprobar si la modificación ha afectado a la concentración de sodio, en el sentido de que su distribución se haya trasladado o no. Para ello, se han realizado 15 análisis con los siguientes resultados (en gr. por unidad):

   2.37 2.95 3.4 2.64 3.66 3.18 2.72 3.61 3.87 1.97 1.66 3.72 2.10 1.83 3.03

   ¿Se puede afirmar, al nivel de significación 0.1, que la distribución de la cantidad de sodio no ha variado?

6. La siguiente tabla presenta las presiones sanguíneas sistólicas de 10 individuos antes y después de haber dejado la bebida.

A	140	165	160	160	175	190	170	175	155	160
D	145	150	150	160	170	175	160	165	145	170

   ¿Se puede afirmar a partir de los datos que el abandono de la bebida no disminuye la presión sanguínea? ¿Bajo qué hipótesis?

7. En cierta comunidad de E.E.U.U. se realizó un estudio para investigar si el sueldo anual de las familias influía en los hijos para la elección de los diferentes cursos de enseñanza secundaria (Preparatorio, General y Comercial). Para ello, se hizo una clasificación de los sueldos en cuatro niveles (I, II, III y IV), y se tomó una muestra aleatoria simple de 390 estudiantes, obteniéndose la siguiente tabla de frecuencias:

Sueldo / *Curso*	I	II	III	IV
Preparatorio	23	40	16	2
General	11	75	107	14
Comercial	1	31	60	10

   A la vista de los datos, decidir, al nivel de significación 0.01, si se acepta que el nivel económico familiar no influye en la decisión de los estudiantes a la hora de elegir curso.

8. En un estudio sociológico sobre el tema de la polución atmosférica se entrevistó a 40 residentes de cada una de tres zonas residenciales en Gran Bretaña. La siguiente tabla muestra las respuestas a la pregunta: ¿Hay problema de polución en su barrio?

Respuesta / *Zona residencial*	No	Sí	No sabe	No contesta
1	5	31	2	2
2	10	21	4	5
3	11	20	7	2

   Contrastar si las tres poblaciones de residentes pueden considerarse homogéneas con respecto a su opinión sobre la polución.

9. Para determinar si diferentes tipos de profesiones de los individuos activos de cierto colectivo afectan a la tensión arterial, se clasificó a los individuos en cuatro grupos, atendiendo a su profesión, y se midió la tensión a una muestra de individuos elegidos de forma aleatoria. Clasificando la tensión en los niveles "Bajo", "Normal" y "Alto", se obtuvo los siguientes resultados, que muestran el número de individuos de cada tipo de profesión con los distintos niveles de tensión:

Nivel de tensión / *Profesión*	Bajo	Normal	Alto
I	8	4	3
II	5	7	7
III	4	8	8
IV	5	7	8

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.

10. Para determinar si las calificaciones de los alumnos en selectividad son independientes de las calificaciones en bachiller, se eligió de forma aleatoria una muestra de alumnos, a los que se preguntó ambas calificaciones, obteniendo los siguientes resultados:

Selectividad / *Bachiller*	Suspenso	Aprobado	Notable	Sobresaliente
Aprobado	10	6	4	5
Notable	7	9	9	4
Sobresaliente	6	10	10	6

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.

Problema 1

A partir de los siguientes datos, que muestran el número de accidentes en un determinado regimiento del ejército durante 200 días elegidos al azar, contrastar si el número de accidentes diarios sigue una distribución de Poisson de parámetro 2.

Nº de accidentes:	0	1	2	3	4	5	6	7
Nº de días:	22	53	58	39	20	5	2	1

$$X = \textit{Número de accidentes diarios} \quad \rightarrow \quad \begin{array}{l} H_0 : X \rightarrow \mathcal{P}(2) \\ H_1 : X \nrightarrow \mathcal{P}(2) \end{array}$$

Resolvemos mediante el *test* χ^2 por tratarse de una distribución de tipo discreto.

- El conjunto de valores de la distribución teórica es $\mathbb{N} \cup \{0\}$ y, a la vista de las observaciones disponibles, particionamos dicho conjunto de la siguiente forma:

$$\mathbb{N} \cup \{0\} = \{0\} \cup \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\} \cup \{x \backslash x \geq 7\}.$$

Ahora calculamos las probabilidades que asigna la distribución teórica a cada clase y las correspondientes frecuencias esperadas:

A_i	0	1	2	3	4	5	6	≥ 7
p_i^0	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	0.0045
$np_i^0 = 200p_i^0$	27.067	54.134	54.134	36.089	18.045	7.218	2.406	0.907

Para que todas las frecuencias esperadas sean mayores o iguales que 5 agrupamos las tres últimas clases. En la siguiente tabla se expresan las correspondientes frecuencias esperadas y observadas:

A_i	0	1	2	3	4	≥ 5
np_i^0	27.067	54.134	54.134	36.089	18.045	10.531
n_i	22	53	58	39	20	8

- A continuación calculamos el valor del estadístico de contraste:

$$\chi^2_{exp} = \sum_{i=0}^{5} \frac{(n_i - np_i^0)^2}{np_i^0} = -n + \sum_{i=0}^{5} \frac{n_i^2}{np_i^0} = \frac{22^2}{27.06} + \frac{53^2}{54.14} + \ldots + \frac{8^2}{10.54} - 200 = 2.3031.$$

- Finalmente, calculamos el p-nivel asociado a los datos. Al haber seis categorías en la clasificación de las 200 observaciones, la distribución límite del estadístico χ^2, bajo H_0, es una $\chi^2(5)$:

$$p - \text{nivel} = P_{H_0}\left(\chi^2(N_1, \ldots, N_6) > 2.307\right) \in (0.8, 0.85).$$

Se concluye que *los datos no proporcionan evidencia para rechazar la hipótesis nula que el número de accidentes diarios en el regimiento sigue una distribución* $\mathcal{P}(2)$.

Problema 2

Una tela cuadrada tiene 60 defectos de fabricación. Con objeto de analizar la distribución de los defectos en la superficie de la tela, se ha dividido esta en 9 zonas cuadradas exactamente iguales, observándose los siguientes defectos en cada zona:

8	7	3
5	9	11
6	4	7

Contrastar, a partir de estos datos, si los defectos se distribuyen uniformemente en toda la superficie o, por el contrario, siguen algún patrón de ocurrencia.

Se trata de contrastar que la distribución de la variable *localización de los defectos en la superficie de la tela* es uniforme. Puesto que es una variable continua, se categoriza (en las 9 cuadrículas). Así, el problema se resuelve contrastando que la probabilidad de que cada defecto caiga en una cuadrícula específica es 1/9. Se puede describir este problema considerando la variable

$$X: \ \textit{cuadrícula en la que cae cada defecto},$$

con nueve categorías, y se trata de un problema de ajuste, que resolvemos mediante el *test* χ^2.

Bajo la hipótesis de distribución uniforme 60 defectos, el número esperado de defectos en cada zona es $np_i^0 = 60/9 \ (\geq 5), \quad i = 1, \ldots, 9$ y, por tanto, el estadístico chi-cuadrado toma el valor:

$$\chi^2_{exp} = \sum_{i=1}^{9} \frac{(n_i - np_i^0)^2}{np_i^0} = -60 + \frac{9}{60} \sum_{i=1}^{9} n_i^2 = 7.5.$$

Ya que la distribución de $\chi^2(N_1, \ldots, N_9)$ bajo H_0 es $\chi^2(8)$:

$$p - nivel: \ P_{H_0}\left(\chi^2(N_1, \ldots, N_9) > 7.5\right) \in (0.45, \ 0.5).$$

Por tanto, deducimos que *los datos no dan evidencia para rechazar la hipótesis de que los defectos se distribuyen al azar por toda la superficie.*

Problema 3

Un modelo genético indica que la distribución de una población de hombres y mujeres, daltónicos o no, se ajusta a probabilidades de la forma:

	Hombres	*Mujeres*
No daltónicos	$(1-p)/2$	$(1-p^2)/2$
Daltónicos	$p/2$	$p^2/2$

Para comprobar esta teoría se examinaron 2000 individuos de la población, elegidos al azar, obteniéndose los siguientes resultados:

	Hombres	*Mujeres*
No daltónicos	894	1015
Daltónicos	81	10

Contrastar, mediante el test de la χ^2, si esta muestra concuerda con el modelo teórico.

Puesto que p es desconocido, comenzamos estimándolo por máxima verosimilitud a partir de la información muestral, dada por el valor de las siguientes variables:

- N_1: Número de hombres no daltónicos en una muestra aleatoria de 2000.
- N_2: Número de hombres daltónicos en una muestra aleatoria de 2000.
- N_3: Número de mujeres no daltónicas en una muestra aleatoria de 2000.
- N_4: Número de mujeres daltónicas en una muestra aleatoria de 2000.

Ya que, bajo H_0, $(N_1, N_2, N_3) \to M(2000;\ (1-p)/2,\ p/2,\ (1-p^2)/2)$:

- $L_{n_1,n_2,n_3}(p) = PR_{2000}^{894,81,1015,10} \left(\frac{1-p}{2}\right)^{894} \left(\frac{p}{2}\right)^{81} \left(\frac{1-p^2}{2}\right)^{1015} \left(\frac{p^2}{2}\right)^{10} \propto p^{101}(1-p)^{894}(1-p^2)^{1015}$
- $\ln L_{n_1,n_2,n_3}(p) \propto 101\ln p + 894\ln(1-p) + 1015\ln(1-p^2) \Rightarrow \dfrac{d\ln L_{n_1,n_2,n_3}(p)}{dp} = \dfrac{101}{p} - \dfrac{894}{1-p} - \dfrac{2030p}{1-p^2}.$
- $\dfrac{d\ln L_{n_1,n_2,n_3}(p)}{dp} = 0 \Rightarrow 3025p^2 + 894p - 101 = 0 \Rightarrow p = \begin{cases} 0.087229251 \\ -0.382766441 \end{cases} \Rightarrow \widehat{p} = 0.087229251.$

Por tanto, el número de observaciones esperadas en cada uno de los cuatro grupos es:

	Hombres	Mujeres
No daltónicos	$2000 \times \frac{1-0.087229251}{2}$	$2000 \times \frac{1-0.087229251^2}{2}$
Daltónicos	$2000 \times \frac{0.087229251}{2}$	$2000 \times \frac{0.087229251^2}{2}$

	Hombres	Mujeres
No daltónicos	912.77	992.39
Daltónicos	87.23	7.61

El valor del estadístico de contraste y el p-nivel, calculado a partir de la $\chi^2(2)$, son:

$$\chi^2_{exp} = \frac{(894-912.77)^2}{912.77} + \frac{(81-87.23)^2}{87.23} + \frac{(1015-992.39)^2}{992.39} + \frac{(10-7.61)^2}{7.61} = 2.097 \Rightarrow P_{H_0}\left(\chi^2(N_i) > 2.097\right) \in (0.35, 0.4)$$

Por tanto, *no puede rechazarse el modelo teórico; esto es, las observaciones concuerdan con dicho modelo.*

Problema 4

Un laboratorio farmacéutico afirma que uno de sus productos confiere inmunidad a la picadura de insectos durante un tiempo exponencial de media 2.5 horas. Probado el producto en 20 sujetos, en un ambiente con gran número de mosquitos, los instantes (en horas) en que recibieron la primera picadura fueron:

$$\begin{array}{llllllllll} 0.01, & 0.02 & 0.03, & 0.23, & 0.51, & 0.74, & 0.96, & 1.17, & 1.46, & 1.62, \\ 2.18, & 2.25, & 2.79, & 3.45, & 3.83, & 3.92, & 4.27, & 5.43, & 5.79, & 6.34. \end{array}$$

Usando el test de Kolmogorov-Smirnov, contrastar, a partir de estos datos, si puede aceptarse la afirmación del laboratorio.

El problema planteado, notando F_0 la función de distribución de la exponencial de media 2.5. es:

$$H_0 : F_X = F_0$$
$$H_1 : F_X \neq F_0$$

siendo

$$X = \textit{tiempo de inmunidad a la picadura con el producto del laboratorio}$$

$x_{(i)}$	$F_n(x_{(i)})$	$F_0(x_{(i)})$	$F_n(x_{(i)}) - F_0(x_{(i)})$	$F_0(x_{(i)}) - F_n(x_{(i)}^-)$
0.01	0.05	0.003992011	0.046007989	0.003992011
0.02	0.10	0.007968085	0.092031915	−0.042031915
0.03	0.15	0.011928287	0.138071713	−0.088071713
0.23	0.20	0.087894850	0.112105150	−0.062105150
0.51	0.25	0.184537629	0.065462371	−0.015462371
0.74	0.30	0.256212572	0.043787428	0.006212572
0.96	0.35	0.318868573	0.031131427	0.018868573
1.17	0.40	0.373746476	0.026253524	0.023746476
1.46	0.45	0.442336754	0.007663246	0.042336754
1.62	0.50	0.476909087	0.023090913	0.026909087
2.18	0.55	0.581885517	−0.031885517	0.081885517
2.25	0.60	0.593430340	0.006569660	0.043430340
2.79	0.65	0.672412472	−0.022412472	0.072412472
3.45	0.70	0.748421447	−0.048421447	0.098421447
3.83	0.75	0.783896971	−0.033896971	0.083896971
3.92	0.80	0.791538311	0.008461689	0.041538311
4.27	0.85	0.818772114	0.031227886	0.018772114
5.43	0.90	0.886050510	0.013949490	0.036050510
5.79	0.95	0.901332533	0.048667467	0.001332533
6.34	1	0.920817503	0.079182497	−0.029182497

Se obtiene que el valor del estadístico es D_{exp} =0.13807171. Buscando en la tabla para $n = 20$ se tiene:

$$P(D_n > 0.23156) = 0.2 \Rightarrow p - nivel = P(D_n > 0.13807171) > 0.2.$$

Por tanto, *no tenemos evidencia para rechazar la hipótesis nula, es decir, aceptamos que la muestra proviene de la distribución exponencial de media 2.5.*

Problema 5

Cierta comunidad ha modificado la procedencia del agua destinada al consumo doméstico. Se sabe que, con el antiguo suministro, la distribución de la cantidad de sodio por unidad de volumen de sangre de sus habitantes es simétrica alrededor de 3.24 gr. Tras cierto tiempo, se quiere comprobar si la modificación ha afectado a la concentración de sodio, en el sentido de que su distribución se haya trasladado o no. Para ello, se han realizado 15 análisis con los siguientes resultados (en gr. por unidad):

2.37	2.95	3.4	2.64	3.66	3.18	2.72	3.61
3.87	1.97	1.66	3.72	2.10	1.83	3.03	

¿Se puede afirmar, al nivel de significación 0.1, que la distribución de la cantidad de sodio no ha variado?

Notando X= *Cantidad de sodio por unidad de volumen de sangre de los habitantes con el nuevo suministro*, el problema de contraste es

$$H_0 : M_X = 3.24$$
$$H_1 : M_X \neq 3.24.$$

X simétrica $\rightarrow$ *test de los rangos signados al nivel de significación* $\alpha = 0.1$:

$$\varphi = \begin{cases} 1 & T^+_{exp} \leq k \text{ o } T^+_{exp} \geq \frac{n(n+1)}{2} - k = 120 - k \\ 0 & k < T^+_{exp} < 120 - k \end{cases} \qquad \begin{array}{l} \textit{Tamaño} = 2P(T^+ \leq k) \leq 0.1 \rightarrow P(T^+ \leq k) \leq 0.05 \\ n = 15 \rightarrow P(T^+ \leq 30) = 0.047, \quad P(T^+ \leq 31) = 0.053 \Rightarrow k = 30. \end{array}$$

$$\varphi = \begin{cases} 1 & T^+_{exp} \leq 30 \text{ o } T^+_{exp} \geq 90 \\ 0 & 30 < T^+_{exp} < 90 \end{cases}$$

$d_i = -0.87,\ -0.29,\ 0.16,\ -0.6,\ 0.42,\ -0.06,\ -0.52,\ 0.37,\ 0.63,\ -1.27,\ -1.58,\ 0.48,\ -1.14,\ -1.41,\ -0.21$

$\lvert d_i \rvert$	0.06	0.16	0.21	0.29	0.37	0.42	0.48	0.52	0.6	0.63	0.87	1.14	1.27	1.41	1.58
Signo	−	+	−	−	+	+	+	−	−	+	−	−	−	−	−
Rango	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$T^+_{exp} = 2 + 5 + 6 + 7 + 10 = 30$ y, por tanto, *se rechaza la hipótesis nula al nivel de significación 0.1; el cambio en el suministro sí afecta a la cantidad de sodio en sangre.*

Como se ha rechazado H_0 y observamos un valor pequeño de T^+, parece lógico pensar que el valor de la mediana haya disminuido y, por tanto, planteamos el problema:

$$H_0 : M_X = 3.24$$
$$H_1 : M_X < 3.24$$

El test para resolver este problema, al nivel de significación 0, 1, es

$$\varphi = \begin{cases} 1 & T^+_{exp} \leq k \\ 0 & T^+_{exp} > k \end{cases} \rightarrow P(T^+ \leq k) \leq 0.1 \rightarrow \varphi = \begin{cases} 1 & T^+_{exp} \leq 36 \\ 0 & T^+_{exp} > 36. \end{cases}$$

Por tanto, la conclusión a partir de nuestros datos es que, *al nivel de significación* 0.1*, se rechaza la hipótesis nula a favor de que el cambio en el suministro ha disminuido la cantidad de sodio en sangre*.

Problema 6

La siguiente tabla presenta las presiones sanguíneas sistólicas de 10 individuos antes y después de haber dejado la bebida.

A	140	165	160	160	175	190	170	175	155	160
D	145	150	150	160	170	175	160	165	145	170

¿Se puede afirmar a partir de los datos que el abandono de la bebida no disminuye la presión sanguínea? ¿Bajo qué hipótesis?

Suponiendo que la forma de la distribución de la presión es la misma antes y depués de dejar la bebida, consideramos la variable

$X = A - D$=*Diferencia de la presión de un individuo antes y después de dejar la bebida,*

y, notando M_X a la mediana de X, el problema planteado es:

$$H_0 : M_X = 0$$
$$H_1 : M_X > 0.$$

Aunque las medidas suelan darse exactas por redondeo, la presión (y, por tanto, X) es una variable continua y podemos aplicar los tests de localización.

Test de los signos: Calculamos los valores observados de $X = A - D$; eliminamos el valor 0 (igual al hipotético de M_X) y trabajamos con 9 datos:

$$-5,\ 15,\ 10,\ 5,\ 15,\ 10,\ 10,\ 10,\ -10$$

$$T = \textit{Número de observaciones muestrales mayores que 0} \rightarrow T_{exp} = 7.$$

$$T \underset{H_0}{\rightarrow} B(9, 1/2) \Rightarrow P_{H_0}(T \geq 7) = 1 - 0.9102 \Rightarrow p - nivel = 0.0898.$$

Se acepta H_0 a niveles de significación inferiores a 0.0898, y se rechaza a niveles superiores.

Test de los rangos signados: Suponiendo que la distribución de X es *simétrica*:

$\lvert d_i\rvert = \lvert x_i - 0\rvert$	5	5	10	10	10	10	10	15	15	
Signo	+	−	+	+	+	+	−	+	+	
Rango	1.5	1.5	5	5	5	5	5	8	9	$1.5 + 5 + 5 + 5 + 5 + 8 + 9 = 38.5$

$$T^+ = \textit{Suma de rangos de diferencias positivas} \longrightarrow T^+_{exp} = 38.5.$$

$$p - nivel = P(T^+ \geq 38.5) = P(T^+ \geq 39) = 0.027.$$

Se acepta H_0 a niveles de significación inferiores a 0.027, y se rechaza a niveles superiores.

NOTA: Hay niveles de significación que conducen al rechazo con el test de los rangos signados y a la aceptación con el test de los signos. Ya que el primero es más potente, si se está razonablemente seguro de que la distribución de las diferencias es simétrica, éste es el que debe usarse.

Problema 7

En cierta comunidad de E.E.U.U. se realizó un estudio para investigar si el sueldo anual de las familias influía en los hijos para la elección de los diferentes cursos de enseñanza secundaria (Preparatorio, General y Comercial). Para ello, se hizo una clasificación de los sueldos en cuatro niveles (I, II, III y IV), y se tomó una muestra aleatoria simple de 390 estudiantes, obteniéndose la siguiente tabla de frecuencias:

Sueldo / *Curso*	*I*	*II*	*III*	*IV*
Preparatorio	23	40	16	2
General	11	75	107	14
Comercial	1	31	60	10

A la vista de los datos, decidir, al nivel de significación de 0.01, si se acepta que el nivel económico familiar no influye en la decisión de los estudiantes a la hora de elegir curso.

El diseño de la experiencia (tomar una muestra y observar cada una de las variables en cada uno de los individuos), corresponde a un test de independencia.

X : *cursos de enseñanza* $= A$ (Preparatorio), B (General), C (Comercial)

Y : *sueldo anual* $= I, II, III, IV$

$$H_0 : X, Y \text{ son independientes}$$

Sueldo / *Curso*	I	II	III	IV	Totales
Preparatorio	23 7.26923	40 30.32308	16 38.00769	2 5.4	81
General	11 18.57692	75 77.49231	107 97.13077	14 13.8	207
Comercial	1 9.15385	31 38.18462	60 47.86154	10 6.8	102
Totales	35	146	183	26	390

$E_{ij} > 5, \ \forall (i,j).$

$$\chi^2_{exp} = \frac{(23-7.26923)^2}{7.26923} + \frac{(40-30.32308)^2}{30.32308} + \cdots = 69.38933.$$

La distribución teórica del estadístico de contraste es $\chi^2(6)$ y $\chi^2_{exp} > \chi^2_{6,0.01} = 16.8119$.

Por tanto, *al nivel de significación 0.01, se rechaza la hipótesis de independencia: el sueldo de la familia influye en los hijos a la hora de elegir curso.*

Problema 8

En un estudio sociológico sobre el tema de la polución atmosférica se entrevistó a 40 residentes de cada una de tres zonas residenciales en Gran Bretaña. La siguiente tabla muestra las respuestas a la pregunta: ¿Hay problema de polución en su barrio?

Zona residencial \ *Respuesta*	*No*	*Si*	*No sabe*	*No contesta*
1	5	31	2	2
2	10	21	4	5
3	11	20	7	2

Contrastar si las tres poblaciones de residentes pueden considerarse homogéneas con respecto a su opinión sobre la polución.

El diseño de la experiencia (tomar una muestra de cada zona residencial y preguntar la opinión), corresponde a un test de homogeneidad.

X_i : *opinión sobre la polución en la zona i-ésima* = *No, Sí, No sabe, No contesta*, $i = 1, 2, 3$.

$$H_0 : X_1, X_2, X_3 \text{ tienen la misma distribución}$$

Zona residencial \ *Respuesta*	*No*	*Si*	*No sabe*	*No contesta*	Totales
1	5 8.66667	31 24	2 4.33333	2 3	40
2	10 8.66667	21 24	4 4.33333	5 3	40
3	11 8.66667	20 24	7 4.33333	2 3	40
Totales	26	72	13	9	**120**

$n_{i.} > 20, \quad i = 1, 2, 3$

$E_{ij} > 2$

El test es poco fiable, ya que el 50 % de las frecuencias esperadas es menor que 5. Sería conveniente elegir más individuos en cada muestra. Aún así, lo realizamos:

$$\chi^2_{exp} = \frac{(5 - 8.6667)^2}{8.6667} + \frac{(31 - 24)^2}{24} + \cdots + \frac{(2 - 3)^2}{3} = 10.39103.$$

La distribución teórica del estadístico de contraste es $\chi^2(6)$ y el p-nivel es

$$P_{H_0}(\chi^2(N_{ij}) > 10.39103) \approx 0.10912. \ (EXCEL)$$

Se acepta (con reservas) la hipótesis de homogeneidad para cualquier nivel de significación inferior a 0.10912. Esto es, los habitantes de las tres zonas opinan de igual forma. JUNTAMOS CATEGORÍAS

Respuesta / *Zona residencial*	*No*	*Si*	*No sabe/ No contesta*	Totales
1	5 8.66667	31 24	4 7.33333	40
2	10 8.66667	21 24	9 7.33333	40
3	11 8.66667	20 24	9 7.33333	40
Totales	26	72	22	**120**

$n_{i.} > 20, \quad i = 1, 2, 3$

$E_{ij} > 5$

$$\chi^2_{exp} = \frac{(5-8.6667)^2}{8.6667} + \frac{(31-24)^2}{24} + \cdots + \frac{(9-7.33333)^2}{7.33333} = 7.74.$$

La distribución teórica del estadístico de contraste es $\chi^2(4)$ y el p-nivel es

$$P_{H_0}(\chi^2(N_{ij}) > 7.74) = 0.10158. \text{ EXCEL}$$

Se acepta (con reservas) la hipótesis de homogeneidad para cualquier nivel de significación inferior a 0.1. Esto es, los habitantes de las tres zonas opinan de igual forma.

Problema 9

Para determinar si diferentes tipos de profesiones de los individuos activos de cierto colectivo afectan a la tensión arterial, se clasificó a los individuos en cuatro grupos, atendiendo a su profesión, y se midió la tensión a una muestra de individuos elegidos de forma aleatoria. Clasificando la tensión en los niveles "Bajo", "Normal" y "Alto", se obtuvo los siguientes resultados, que muestran el número de individuos de cada tipo de profesión con los distintos niveles de tensión:

Profesión \ Nivel de tensión	*Bajo*	*Normal*	*Alto*
I	8	4	3
II	5	7	7
III	4	8	8
IV	5	7	8

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.

El diseño de la experiencia (tomar una muestra y observar cada una de las variables en cada uno de los individuos), corresponde a un test de independencia.

X : *nivel de tensión* = *Bajo, Normal, Alto.*

Y : *nivel de profesión* = *I, II, III, IV.*

$$H_0 : X, Y \text{ independientes}$$

Profesión \ *Tensión*	*Bajo*	*Normal*	*Alto*	Totales
I	8 4.45946	4 5.27027	3 5.27027	15
II	5 5.64865	7 6.67568	7 6.67568	19
III	4 5.94595	8 7.02703	8 7.02703	20
IV	5 5.94595	7 7.02703	8 7.02703	20
Totales	22	26	26	**74**

$E_{ij} > 2$ y sólo $E_{11} < 5$ $(20\,\%(12) = 2.4)$

$$\chi^2_{exp} = \frac{(8-4.45946)^2}{4.45946} + \frac{(4-5.27027)^2}{5.27027} + \cdots + \frac{(8-7.02703)^2}{7.02703} = 5.39271.$$

La distribución teórica del estadístico de contraste es $\chi^2(6)$ y el p-nivel es

$$P_{H_0}(\chi^2(N_{ij}) > 5.3927) \approx 0.4945. \; (EXCEL)$$

Se acepta H_0: el nivel de tensión y la profesión son independientes.

Problema 10

Para determinar si las calificaciones de los alumnos en selectividad son independientes de las calificaciones en bachiller, se eligió de forma aleatoria una muestra de alumnos, a los que se preguntó ambas calificaciones, obteniendo los siguientes resultados:

Selectividad / Bachiller	*Suspenso*	*Aprobado*	*Notable*	*Sobresaliente*
Aprobado	10	6	4	5
Notable	7	9	9	4
Sobresaliente	6	10	10	6

¿Qué conclusión acerca del problema planteado se obtiene a la vista de estos datos? Especificar las hipótesis nula y alternativa que se contrastan.

El diseño de la experiencia (tomar una muestra y observar cada una de las variables en cada uno de los individuos), corresponde a un test de independencia.

X : *nota en bachiller* = *Aprobado, Notable, Sobresaliente.*

Y : *nota en selectividad* = *Suspenso, Aprobado, Notable, Sobresaliente.*

$$H_0 : X, Y \text{ independientes}$$

Selectividad / *Bachiller*	*Suspenso*	*Aprobado*	*Notable*	*Sobresaliente*	Totales
Aprobado	10 6.68605	6 7.26744	4 6.68605	5 4.36047	25
Notable	7 7.75581	9 8.43023	9 7.75581	4 5.05814	29
Sobresaliente	6 8.55814	10 9.30233	10 8.55814	6 5.58140	32
Totales	23	25	23	15	**86**

$$E_{ij} > 2 \text{ y sólo } E_{13} < 5 \; (20\,\%(12) = 2.4)$$

$$\chi^2_{exp} = \frac{(10 - 6.68605)^2}{6.68605} + \frac{(6 - 7.26744)^2}{7.26744} + \cdots + \frac{(6 - 5.58140)^2}{5.58140} = 4.66092.$$

La distribución teórica del estadístico de contraste es $\chi^2(6)$ y el p-nivel es

$$P_{H_0}(\chi^2(N_{ij}) > 4.66092) \approx 0.588. \; (EXCEL)$$

Se acepta la hipótesis de independencia entre las notas de bachiller y las de selectividad.