TEMA 9: Contrastes no paramétricos

9.1. PROBLEMA DE BONDAD DE AJUSTE

9.1.1. TEST χ^2 DE PEARSON $\rightarrow X$ *variable cualitativa,* $X = A_1, \ldots, A_k$

$$H_0 : P(X = A_i) = p_i^0, \quad \forall i = 1, \cdots, k.$$
$$H_1 : P(X = A_i) \neq p_i^0 \text{ para algún } i = 1, \cdots, k.$$

- Se toma una muestra de n observaciones independientes de X.
- N_i : *Número de observaciones muestrales en la categoría* A_i, $i = 1, \ldots, k$.

$$\chi^2(N_1, \ldots, N_k) = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \underset{n \to +\infty,\ H_0}{\xrightarrow{L}} \chi^2(k-1)$$

Región de rechazo (test de tamaño α*)*$^{(*)}$	$p-$*nivel*$^{(*)}$
$\chi^2_{exp} \geq \chi^2_{k-1;\ \alpha}$ $P_{H_0}\left(\chi^2(N_1, \ldots, N_k) \geq \chi^2_{k-1;\ \alpha}\right) = \alpha$	$P_{H_0}\left(\chi^2(N_1, \ldots, N_k) \geq \chi^2_{exp}\right)$

$^{(*)}$ Tamaño y p-nivel aproximados por la distribución asintótica de $\chi^2(N_1, \ldots, N_k)$ bajo H_0.
Requisitos mínimos: $np_i^0 \geq 5, \quad \forall i = 1, \ldots, k.$

9.1.2. TEST DE KOLMOGOROV-SMIRNOV $\rightarrow X$ *variable aleatoria continua*

$$H_0 : P_X = P_0.$$
$$H_1 : P_X \neq P_0.$$

- $(X_1, \ldots, X_n)$ muestra aleatoria simple de X.
- $F^*_{X_1,\ldots,X_n}$ *función de distribución muestral.*

$$D(X_1, \ldots, X_n) = \sup_{x \in \mathbb{R}} \left|F^*_{X_1,\ldots,X_n}(x) - F_0(x)\right| \underset{H_0}{\longrightarrow} Z \text{ de Kolmogorov}$$

Región de rechazo (test de tamaño α*)*	$p-$*nivel*
$D_{exp} \geq d_\alpha$ $P_{H_0}\Big(D(X_1, \ldots, X_n) \geq d_\alpha\Big) = \alpha$	$P_{H_0}\Big(D(X_1, \ldots, X_n) \geq D_{exp}\Big)$

$$x_i \neq x_j,\ i \neq j \Rightarrow D_{exp} = D(x_1, \cdots, x_n) = \max\left\{\max_{x_i}\left[F^*(x_i) - F_0(x_i)\right],\ \max_{x_i}\left[F_0(x_i) - F^*(x_i^-)\right]\right\}.$$

$$F^* := F^*_{x_1,\ldots,x_n}$$

9.2. PROBLEMA DE LOCALIZACIÓN

$(X_1, \ldots, X_n)$ muestra aleatoria simple de X, con mediana M_X

$$H_0 : M_X = m \quad H_1 : M_X \neq m \qquad H_0 : M_X = m \quad H_1 : M_X > m \qquad H_0 : M_X = m \quad H_1 : M_X < m$$

9.2.1. TEST DE LOS SIGNOS DE FISHER $\rightarrow$ *X variable aleatoria continua*

$T(X_1, \ldots, X_n) =$ *Número de observaciones muestrales* X_i *mayores que* $m \underset{H_0}{\longrightarrow} B(n, 1/2)$.

Alternativa	*Región de rechazo (n.s.* α*)*	$p -$ *nivel*
$H_1 : M_X > m$	$T_{exp} \geq k$ $P_{H_0}(T(X_1, \ldots, X_n) \geq k) \leq \alpha$	$P_{H_0}(T(X_1, \ldots, X_n) \geq T_{exp})$
$H_1 : M_X < m$	$T_{exp} \leq k$ $P_{H_0}(T(X_1, \ldots, X_n) \leq k) \leq \alpha$	$P_{H_0}(T(X_1, \ldots, X_n) \leq T_{exp})$
$H_1 : M_X \neq m$	$T_{exp} \leq k$ ó $T_{exp} \geq n - k$ $P_{H_0}(T(X_1, \ldots, X_n) \leq k) \leq \alpha/2$	$2P_{H_0}(T(X_1, \ldots, X_n) \leq T_{exp})$ si $T_{exp} \leq n/2$ $2P_{H_0}(T(X_1, \ldots, X_n) \geq T_{exp})$ si $T_{exp} \geq n/2$

En cualquier caso, el punto crítico, k, se elige ajustándose lo más posible al nivel de significación.

9.2.2. TEST DE LOS RANGOS SIGNADOS DE WILCOXON $\rightarrow$ *X v.a. continua y simétrica*

Se asignan rangos a $D_i = X_i - m$, $i = 1, \ldots, n$, según orden creciente de $|D_1|, \ldots, |D_n|$

$T^+(X_1, \ldots, X_n) =$ *Suma de los rangos correspondientes a los* D_i *positivos*$^{(*)}$

Alternativa	*Región de rechazo (n.s.* α*)*	$p -$ *nivel*
$H_1 : M_X > m$	$T^+_{exp} \geq k$ $P_{H_0}(T^+(X_1, \ldots, X_n) \geq k) \leq \alpha$	$P_{H_0}(T^+(X_1, \ldots, X_n) \geq T^+_{exp})$
$H_1 : M_X < m$	$T^+_{exp} \leq k$ $P_{H_0}(T^+(X_1, \ldots, X_n) \leq k) \leq \alpha$	$P_{H_0}(T^+(X_1, \ldots, X_n) \leq T^+_{exp})$
$H_1 : M_X \neq m$	$T^+_{exp} \leq k$ ó $T^+_{exp} \geq \frac{n(n+1)}{2} - k$ $P_{H_0}(T^+(X_1, \ldots, X_n) \leq k) \leq \alpha/2$	$2P_{H_0}(T^+(X_1, \ldots, X_n) \leq T^+_{exp})$ si $T^+_{exp} \leq n(n+1)/4$ $2P_{H_0}(T^+(X_1, \ldots, X_n) \geq T^+_{exp})$ si $T^+_{exp} \geq n(n+1)/4$

De nuevo, el punto crítico, k, se elige ajustándose lo más posible al nivel de significación.

(*) La distribución de $T^+(X_1, \ldots, X_n)$ bajo H_0, (simétrica alrededor de $\frac{n(n+1)}{4}$) está tabulada para $n \leq 15$. Para $n > 15$ puede usarse la siguiente aproximación:

$$T^+(X_1, \ldots, X_n) \approx \mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

9.3. PROBLEMA DE INDEPENDENCIA

TEST χ^2 DE INDEPENDENCIA $\rightarrow X = A_1, \ldots, A_m, \quad Y = B_1, \ldots, B_k$ *variables cualitativas*

$$H_0 : P(X = A_i, Y = B_j) = P(X = A_i)P(Y = B_j), \quad \forall i = 1, 2, \ldots, m; \ j = 1, 2, \ldots, k.$$

$$H_1 : P(X = A_i, Y = B_j) \neq P(X = A_i)P(Y = B_j), \quad \text{para algún par } (i, j).$$

- Se toma una muestra de n observaciones independientes de (X, Y).
- N_{ij} : *Número de observaciones muestrales en* A_i *y* B_j, $\quad i = 1, \ldots, m; \ j = 1, \ldots, k.$

X \ Y	B_1	$\cdots$	B_j	$\cdots$	B_k	Totales
A_1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1k}	$N_{1.}$
$\vdots$	$\cdots$		$\cdots$		$\cdots$	$\vdots$
A_i	N_{i1}	$\cdots$	N_{ij}	$\cdots$	N_{ik}	$N_{i.}$
$\vdots$	$\cdots$		$\cdots$		$\cdots$	$\vdots$
A_m	N_{m1}	$\cdots$	N_{mj}	$\cdots$	N_{mk}	$N_{m.}$
Totales	$N_{.1}$	$\cdots$	$N_{.j}$	$\cdots$	$N_{.k}$	n

$$\chi^2(N_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}} \underset{n\to+\infty,\ H_0}{\xrightarrow{L}} \chi^2((m-1)(k-1))$$

Región de rechazo (test de tamaño α*)*$^{(*)}$	$p-$*nivel*$^{(*)}$
$\chi^2_{exp} \geq \chi^2_{(m-1)(k-1);\ \alpha}$ $P_{H_0}\left(\chi^2(N_{ij}) \geq \chi^2_{(m-1)(k-1);\ \alpha}\right) = \alpha$	$P_{H_0}\left(\chi^2(N_{ij}) \geq \chi^2_{exp}\right)$

$^{(*)}$ Tamaño y p-nivel aproximados por la distribución asintótica de $\chi^2(N_{ij})$ bajo H_0.
Requisitos mínimos: $n_{i.}n_{.j}/n \geq 2, \quad \forall i = 1, \ldots, m, \ j = 1, \ldots, k$ y no más del 20 % menores que 5.

9.4. PROBLEMA DE HOMOGENEIDAD

TEST χ^2 DE HOMOGENEIDAD $\rightarrow$ $X = A_1, \ldots, A_k$ *variable cualitativa analizada en* m *poblaciones*

$X_1, \ldots, X_m$ *variables que describen* X *en las distintas poblaciones.*

$$H_0 : P(X_1 = A_j) = P(X_2 = A_j) = \cdots = P(X_m = A_j), \quad j = 1, 2, \cdots, k.$$

$$H_1 : P(X_i = A_j) \neq P(X_{i'} = A_j) \text{ para algún par } (i, i') \text{ y algún } j = 1, \ldots, k.$$

- Se toman m muestras independientes de observaciones independientes de $X_1, \ldots, X_m$.
- N_{ij} : *Número de observaciones de la muestra* i*-ésima en la categoría* A_j, $j = 1, \ldots, k$.

Categorías / Muestras	A_1	$\cdots$	A_j	$\cdots$	A_k	Totales
1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1k}	$n_{1.}$
$\vdots$	$\cdots$		$\cdots$		$\cdots$	$\vdots$
i	N_{i1}	$\cdots$	N_{ij}	$\cdots$	N_{ik}	$n_{i.}$
$\vdots$	$\cdots$		$\cdots$		$\cdots$	$\vdots$
m	N_{m1}	$\cdots$	N_{mj}	$\cdots$	N_{mk}	$n_{m.}$
Totales	$N_{.1}$	$\cdots$	$N_{.j}$	$\cdots$	$N_{.k}$	n

$$\chi^2(N_{ij}) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i.} N_{.j}}{n}\right)^2}{\frac{n_{i.} N_{.j}}{n}} \underset{\substack{n_i \to +\infty \\ \forall i=1,\ldots m, \ H_0}}{\xrightarrow{L}} \chi^2((m-1)(k-1)).$$

Región de rechazo (test de tamaño α*)*$^{(*)}$	$p - nivel^{(*)}$
$\chi^2_{exp} \geq \chi^2_{(m-1)(k-1);\ \alpha}$ $P_{H_0}\left(\chi^2(N_{ij}) \geq \chi^2_{(m-1)(k-1);\ \alpha}\right) = \alpha$	$P_{H_0}\left(\chi^2(N_{ij}) \geq \chi^2_{exp}\right)$

$^{(*)}$ Tamaño y p-nivel aproximados por la distribución asintótica.
Requisitos mínimos: $n_{i.} \geq 20$, $n_{i.} n_{.j}/n \geq 2$, $\forall i = 1, \ldots, m$, $j = 1, \ldots, k$ y no más del 20 % menores que 5.

EJERCICIOS RESUELTOS

Ejemplo 9.1: *Se recoge una muestra aleatoria simple de 30 tornillos producidos por cierta máquina y se mide su longitud, obteniéndose:*

10.39	10.66	10.12	10.32	10.25	10.52	10.83	10.72	10.28	10.35
10.46	10.54	10.23	10.18	10.62	10.49	10.61	10.64	10.29	10.78
10.81	10.34	10.75	10.41	10.53	10.31	10.47	10.43	10.57	10.74

Contrastar si estos datos avalan que la distribución de la longitud de los tornillos es normal.

Solución: A partir de 30 observaciones independientes de la v.a. $X :=$ *Longitud de los tornillos*, se pretende contrastar que $X \to \{\mathcal{N}(\mu, \sigma^2);\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$.

Ya que la distribución hipotética depende de parámetros no especificados, aplicamos el *test* χ^2, estimando previamente estos parámetros por máxima verosimilitud:

$$\widehat{\mu} = \overline{x} = 10.488 \qquad \widehat{\sigma}^2 = \frac{\sum_{i=1}^{30} (x_i - \overline{x})^2}{30} = 0.038204 \quad (\widehat{\sigma} = 0.195458).$$

Ahora adaptamos el test χ^2 para contrastar $H_0 : X \to \mathcal{N}(10.488,\ 0.038204)$.

En primer lugar se particiona el conjunto de valores de la distribución hipotética ($\mathbb{R}$) en k subconjuntos, $A_1, \ldots, A_k$, tales que el número esperado de observaciones en cada uno de ellos sea, al menos, 5; esto es, si $\widehat{p}_i^0 = P_{H_0}(X \in A_i),\ i = 1, \ldots, k$, ha de ser $30\widehat{p}_i^0 \geq 5$ o, equivalentemente, $\widehat{p}_i^0 \geq 1/6,\ i = 1, \ldots, k$. Entonces, con objeto de tener el mayor número de subconjuntos, tomamos 6 intervalos con probabilidad igual a 1/6, que se determinan a partir de los cuantiles $Q_{1/6},\ Q_{2/6}, \cdots, Q_{5/6}$ de la distribución a contrastar:

$$P_{H_0}\left(X \leq Q_{i/6}\right) = P\left(Z \leq \frac{Q_{i/6} - 10.488}{0.195458}\right) = \frac{i}{6}, \quad i = 1, \cdots, 5, \text{ siendo } Z = \frac{X - 10.488}{0.195458} \to \mathcal{N}(0,1).$$

La siguiente tabla presenta los cuantiles, los intervalos de la partición, y las frecuencias esperadas y observadas en cada uno, que proporcionan el valor del estadístico de contraste:

$Q_{i/6}$	A_i	$n\widehat{p}_i^0$	N_i
10.2989	$(-\infty, 10.2989]$	5	6
10.4038	$(10.2989, 10.4038]$	5	5
10.4880	$(10.4038, 10.4880]$	5	4
10.5722	$(10.4880, 10, 5722]$	5	5
10.6771	$(10.5722, 10.6771]$	5	4
	$(10.6771, +\infty)$	5	6

$$\widehat{\chi}^2_{exp} = \sum_{i=1}^{6} \frac{(N_i - n\widehat{p}_i^0)^2}{n\widehat{p}_i^0} = \frac{(6-5)^2}{5} + \frac{(5-5)^2}{5} + \cdots + \frac{(6-5)^2}{5} = 0.8.$$

Ya que no se especifica nivel de significación, calculamos el p-nivel asociado a los datos, teniendo en cuenta que la distribución teórica del estadístico de contraste bajo H_0 es $\chi^2(3)$, ya que estamos trabajando con seis conjuntos y se han estimado dos parámetros:

$$p - nivel = P_{H_0}\left(\widehat{\chi}^2(N_1, \ldots, N_6) \geq 0.8\right) \approx 0.85.$$

Puesto que este valor es muy grande, se acepta H_0; esto es, *los datos no aportan evidencia para rechazar la hipótesis de que la longitud de los tornillos tiene distribución normal.* ■

Ejemplo 9.2: *Se supone que el tiempo de reacción a un determinado compuesto se distribuye según una* $\mathcal{N}(10.5;\ 0.15^2)$. *Contrastar si los siguientes datos, obtenidos en un muestreo aleatorio simple de 10 individuos a los que se ha administrado el compuesto, proporcionan evidencia para rechazar esta hipótesis:*

10.39 10.66 10.12 10.32 10.25 10.52 10.83 10.72 10.28 10.35.

Solución: La variable observada, de la que se quiere contrastar su distribución, y la hipótesis a contrastar son:

$$X = \textit{Tiempo de reacción al compuesto} \qquad H_0 : X \to \mathcal{N}(10.5;\ 0.15^2).$$

Puesto que la distribución hipotética es de tipo continuo y está totalmente especificada, podemos aplicar el test de Kolmogorov-Smirnov. Para ello, disponemos los datos en orden creciente, $x_{(1)} < \cdots < x_{(10)}$, y calculamos las diferencias $F^*(x_{(i)}) - F_0(x_{(i)})$ y $F_0(x_{(i)}) - F^*(x_{(i)}^-)$, que proporcionan el valor del estadístico de contraste:

- $F_0(x_{(i)}) = P_{H_0}(X \leq x_{(i)}) = P\left(Z \leq \dfrac{x_{(i)} - 10.5}{0.15}\right), \quad Z \to \mathcal{N}(0,1),$
- $F^*(x_{(i)}) = \dfrac{i}{10},$
- $F^*(x_{(i)}^-) = \dfrac{i-1}{10}.$

$x_{(i)}$	$F^*(x_{(i)})$	$F_0(x_{(i)})$	$F^*(x_{(i)}) - F_0(x_{(i)})$	$F_0(x_{(i)}) - F^*(x_{(i)}^-)$
10.12	0.1	0.00565	0.09435	0.00565
10.25	0.2	0.04779	0.15221	−0.05221
10.28	0.3	0.07123	0.22877	−0.12877
10.32	0.4	0.11507	0.28493	−0.18493
10.35	0.5	0.15865	0.34135	−0.24135
10.39	0.6	0.23168	**0.36832**	−0.26832
10.52	0.7	0.55303	0.14697	−0.04697
10.66	0.8	0.85694	−0.05694	**0.15694**
10.72	0.9	0.92877	−0.02877	0.12877
10.83	1	0.98609	0.01391	0.08609

$$D_{exp} = \max\left\{\max_{x_i}\left[F^*(x_i) - F_0(x_i)\right],\ \max_{x_i}\left[F_0(x_i) - F^*(x_i^-)\right]\right\} = 0.36832.$$

La tabla del estadístico de Kolmogorov-Smirnov para $n = 10$ proporciona el p-nivel asociado a los datos:

$$P_{H_0}\left(D(X_1,\ldots,X_{10}) \geq 0.36866\right) = 0.1 \ \Rightarrow\ p\text{−}nivel = P_{H_0}\left(D(X_1,\ldots,X_{10}) \geq 0.36832\right) > 0.1,$$

y concluimos que *los datos no proporcionan evidencia para rechazar la hipótesis de que la distribución del tiempo de reacción es* $\mathcal{N}(10.5,\ 0.15^2)$, *para cualquier nivel de significación menor o igual que 0.1.* ■

Ejemplo 9.3: *Una empresa que tradicionalmente comenzaba su actividad diaria a las 9 h. ha cambiado su horario para abrir a las 8 h. y se pregunta si ello ha afectado significativamente al retraso de sus empleados. Es aceptable pensar que la forma de la distribución de los retrasos no ha variado con el cambio de horario, pero se teme que se haya desplazado a la derecha, lo cual supondría un incremento del tiempo perdido. Se sabe que la mediana de los retrasos de los empleados era inicialmente de 5 minutos. Con el cambio de horario se selecciona a 12 empleados y se observa, en determinados días, los siguientes retrasos (en minutos):*

$$2.5,\quad 1.2,\quad 7,\quad 1.8,\quad 8.3,\quad 6.8,\quad 5.2,\quad 3.4,\quad 4.7,\quad 6.2,\quad 9.1,\quad 5.,2$$

A partir de estos datos, contrastar la hipótesis de que la distribución de los retrasos no ha variado con el cambio de horario.

Solución: Ya que la forma de la distribución de los retrasos es la misma antes y después del cambio de horario, si hay diferencia entre ellas, será debida a su localización, que puede determinarse por el valor de la mediana. Se trata, por tanto de un problema de localización de la mediana de la variable:

$$X:\ \textit{Retraso de los empleados con el cambio de horario},$$

y ya que el temor del empresario es que la distribución se haya desplazado a la derecha con el cambio de horario, el problema de contraste es:

$$H_0 : M_X = 5$$
$$H_1 : M_X > 5.$$

Como la variable X es de tipo continuo, puede aplicarse el *test de los signos* que, para el problema planteado, *rechaza H_0 si hay un alto número de observaciones muestrales mayores que 5*:

$$T(X_1,\ldots,X_{12}):\ \textit{Número de observaciones muestrales mayores que 5} \underset{H_0}{\longrightarrow} B(12,1/2).$$

El valor de este estadístico para la muestra observada es $T_{exp} = 7$ y, por tanto, el p-nivel de los datos observados es:

$$p - nivel = P_{H_0}\left(T(X_1,\ldots,X_{12}) \geq 7\right) = 0.3872.$$

Puesto que este valor es grande, se acepta H_0; esto es, *los datos no aportan evidencia para decidir que el cambio de horario desplaza a la derecha la distribución de los retrasos.* ■

Ejemplo 9.4: *A partir de los datos del Ejemplo 9.3, y suponiendo que la distribución de los retrasos es simétrica, contrastar la hipótesis de que ésta no varía con el cambio de horario.*

Solución: La hipótesis de simetría en la distribución de los retrasos permite usar el *test de los rangos signados*, basado en el estadístico

$$T^+(X_1,\ldots,X_{12}):\ \textit{Suma de los rangos correspondientes a } D_i \textit{ positivos.}$$

Ya que la hipótesis alternativa es $H_1: M_X > 5$, se rechazará H_0 si los datos proporcionan un alto valor de $T^+(X_1,\ldots,X_{12})$. Calculamos dicho valor:

$$d_i = x_i - 5: \quad -2.5,\ -3.8,\ 2,\ -3.2,\ 3.3,\ 1.8,\ 0.2,\ -1.6,\ -0.3,\ 1.2,\ 4.1,\ 0.2$$

$$|d_i| = |x_i - 5|: \quad 2.5,\ 3.8,\ 2,\ 3.2,\ 3.3,\ 1.8,\ 0.2,\ 1.6,\ 0.3,\ 1.2,\ 4.1,\ 0.2$$

$\lvert d_i\rvert$ ordenados	**0.2**	**0.2**	0.3	**1.2**	1.6	**1.8**	**2**	2.5	3.2	**3.3**	3.8	**4.1**
Signo	+	+	−	+	−	+	+	−	−	+	−	+
Rango	**1**	**2**	3	**4**	5	**6**	**7**	8	9	**10**	11	**12**

$$T^+_{exp} = 1+2+4+6+7+10+12 = 42.$$

Calculamos ahora el p-nivel, usando la tabla del estadístico de Wilcoxon para $n = 12$:

$$p - nivel = P_{H_0}\left(T^+(X_1,\ldots,X_{12}) \geq 42\right) = 0.425.$$

Se concluye, como con el test de los signos, que *los datos no dan evidencia para decidir que el cambio de horario aumenta la mediana de la distribución de los retrasos.* ■

Ejemplo 9.5: *Para estudiar si el grupo sanguíneo de los individuos tiene relación con la predisposición a la diabetes, se han seleccionado al azar 400 sujetos a los que se ha determinado el grupo sanguíneo y el nivel de glucosa en sangre en idénticas condiciones experimentales. Clasificando la segunda medida en tres niveles, los resultados han sido:*

Nivel de glucosa / *Grupo*	*Bajo*	*Medio*	*Alto*
O	137	86	35
A	42	23	11
B	19	17	7
AB	14	7	2

Contrastar, al nivel de significación 0.05, si ambas variables son independientes.

Solución: Aplicamos el test χ^2 para contrastar la independencia de las variables *grupo sanguíneo* y *nivel de glucosa en sangre*.

Puesto que las variables están clasificadas en 4 y 3 categorías, respectivamente, el estadístico de contraste tiene distribución $\chi^2(6)$ bajo H_0. Por lo tanto, al nivel especificado, se rechazará H_0 si los datos dan un valor $\chi^2_{\text{exp}} \geq \chi^2_{6;\ 0.05} = 12.5916$.

Construimos la tabla de contingencia con las frecuencias observadas y las esperadas $(n_{i.}n_{.j}/n)$, y calculamos χ^2_{exp}:

Nivel de glucosa / Grupo	Bajo	Medio	Alto	Totales
O	137 136.74	86 85.785	35 35.475	258
A	42 40.28	23 25.27	11 10.45	76
B	19 22.79	17 14.2975	7 5.9125	43
AB	14 12.19	7 7.6475	2 3.1625	23
Totales	212	133	55	400

Observamos que todas las frecuencias esperadas son mayores que 2 y sólo una (menos del 20 %) menor que 5; por lo tanto, podemos aplicar el test χ^2. Calculamos el valor del estadístico:

$$\chi^2_{exp} = \frac{(137-136.74)^2}{136.74} + \frac{(86-85.785)^2}{85.785} + \cdots + \frac{(2-3.1625)^2}{3.1625} = 2.41.$$

Puesto que $\chi^2_{exp} < \chi^2_{6;\ 0.05}$, se deduce que *los datos no dan evidencia para rechazar la hipótesis de independencia entre el grupo sanguíneo y el nivel de glucosa al nivel de significación 0.05*. De hecho, el p-nivel asociado a los datos es

$$P_{H_0}\left(\chi^2(N_{ij}) > 2.41\right) \in (0.85, 0.9),$$

lo que que avala perfectamente la hipótesis nula. ■

Ejemplo 9.6: *Contrastar, a partir de los resultados de la siguiente tabla, si los distintos grupos sanguíneos se presentan con la misma frecuencia en tres grupos étnicos diferentes:*

Grupo sanguíneo / *Raza*	*O*	*A*	*B*	*AB*
1	32	11	7	2
2	47	13	17	9
3	23	7	9	6

Solución: Se trata de contrastar la homogeneidad de las tres poblaciones frente a la variable *grupo sanguíneo* (cualitativa, con cuatro categorías).

Para resolver este problema construimos la tabla de contingencia con los valores observados, los totales y los valores esperados bajo la hipótesis de homogeneidad, $n_{i.}n_{.j}/n$.

Grupo sanguíneo / Raza	O	A	B	AB	Totales
1	32 28.98	11 8.8	7 9.377	2 4.83	52
2	47 47.93	13 14.57	17 15.5	9 7.99	86
3	23 25.08	7 7.62	9 8.15	6 4.18	45
Totales	102	31	33	17	183

Observamos que todos los tamaños muestrales son mayores que 20, que todas las frecuencias esperadas son mayores que 2 y sólo dos (menos del 20 %) menores que 5; por lo tanto, podemos aplicar el test χ^2. Calculamos el valor del estadístico de contraste:

$$\chi^2_{exp} = \frac{(32-28.98)^2}{28.98} + \frac{(11-8.8)^2}{8.8} + \cdots + \frac{(6-4.18)^2}{4.18} = 4.691,$$

y determinamos el p-nivel, teniendo en cuenta que la distribución teórica del estadístico de contraste bajo H_0 es $\chi^2(6)$:

$$0.55 < P_{H_0}\left(\chi^2(N_{ij}) > 4.691\right) < 0.6.$$

Por tanto, se acepta la hipótesis de homogeneidad; esto es, *los datos avalan que las tres razas presentan los distintos grupos sanguíneos con la misma frecuencia.* ■