Análisis de datos categóricos

Universidad de Medellín
Ciencia y Libertad

Ya se estudiaron muchas pruebas de hipótesis, todas estas situaciones de pruebas presentaron una característica común: necesitaban de ciertos supuestos respecto a la población.

Ejemplo: las pruebas *t* y las *F*, requieren el supuesto de que la población está distribuida normalmente.

Debido a que tales pruebas dependen de postulados sobre la población y sus parámetros, se denominan **pruebas paramétricas**.

En la práctica surgen muchas situaciones en las cuales simplemente no es posible hacer de forma segura ningún supuesto sobre el valor de un parámetro o sobre la forma de la distribución poblacional. Las **pruebas no paramétricas o libres de distribución**, son procedimientos estadísticos que pueden utilizarse para contrastar hipótesis cuando no son posibles los supuestos respecto a los parámetros o a las distribuciones poblacionales.

Distribución chi-cuadrado (X^2)

Una de las herramientas no paramétricas más útiles es la prueba chi-cuadrado (X^2) la cual es toda una familia de distribuciones, existe una distribución para cada grado de libertad. Sus dos aplicaciones más comunes son:

1. Pruebas de bondad de ajuste
2. Pruebas de independencia.

Pruebas de bondad de ajuste

La prueba de bondad de ajuste se utiliza para determinar si la distribución de los valores en la población se ajusta a una forma en particular planteada como hipótesis.

H_o: la distribución poblacional es uniforme.
H_A: la distribución poblacional no es uniforme.

Si el ajuste es razonablemente cercano, puede concluirse que sí existe la forma de distribución planteada como hipótesis.

La prueba toma la siguiente forma:

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Donde:

O_i es la frecuencia de los eventos observados en los datos muestrales.

E_i es la frecuencia de los eventos esperados si la hipótesis nula es correcta

K es el número de categorías o clases

La prueba tiene $K - m - 1$ grados de libertad, donde m es el número de parámetros a estimar.

Prueba para un ajuste uniforme

Chris Columbus, director de mercadeo de Seven Seas, tiene la responsabilidad de controlar el nivel de existencias para cuatro tipos de botes vendidos por su firma. En el pasado ha ordenado nuevos botes bajo la premisa de que los cuatro tipos son igualmente populares y la demanda de cada tipo es la misma. Sin embargo, recientemente las existencias se han vuelto más difíciles de controlar, y Chris considera que debería probar su hipótesis respecto a una demanda uniforme. Sus hipótesis son:

H_o: la demanda es uniforme para los cuatro tipos de datos.
H_A: la demanda no es uniforme para los cuatro tipos de datos.

Para probar la hipótesis, Chris selecciona una muestra de n = 48 botes vendidos durante los últimos meses. Si la demanda es uniforme, puede esperar que 48/4=12 botes de cada tipo se vendan. A continuación se muestra esta expectativa junto con el número de cada tipo que en realidad de vendió.

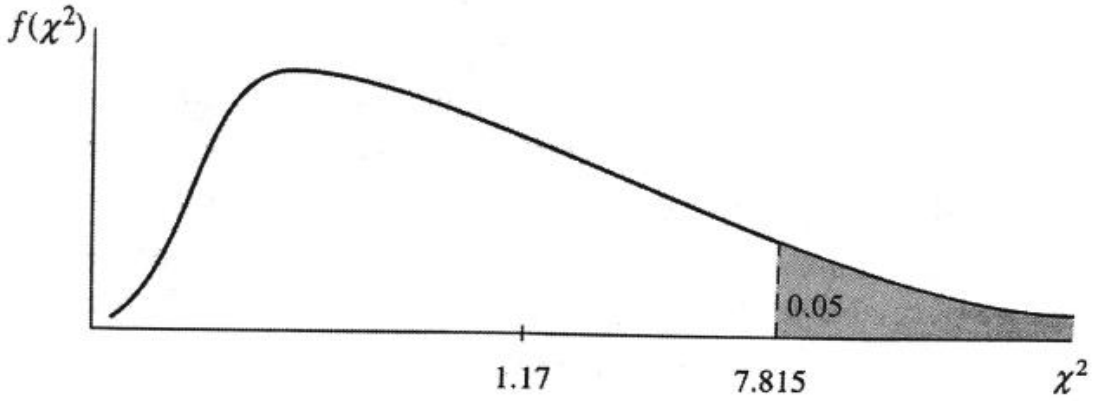
Tipo de bote	Ventas observadas (O_i)	Ventas esperadas (E_i)
Pirates' Revenge	15	12
Jolly Roger	11	12
Bluebeard's Treasure	10	12
Ahab's Quest	12	12
	48	48

Chris debe determinar ahora si los números vendidos realmente en cada una de las categorías K = 4 está lo suficientemente cerca de lo que se esperaría si la demanda fuese uniforme.

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(15-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(10-12)^2}{12} + \frac{(12-12)^2}{12} = 1.17$$

El valor 1.17 se compara con un valor crítico de X^2 tomado de la tabla de la distribución. Debido a que no existen parámetros que tengan que estimarse, m = 0, por lo tanto hay k – 1 = 3 grados de libertad. Si Chris deseara probar al nivel del 5%, se encontraría en la tabla con $X^2_{0.05,3} = 7.815$.

Regla de decisión: "No rechazar si $X^2 \leq 7.815$. Rechazar si $X^2 > 7.815$".



Gracias a que 1.17 < 7.815, la hipótesis nula de que la demanda es uniforme no se rechaza. Las diferencias entre lo que se observó en realidad O_i, y lo que Chris esperaba observar si la demanda fuera la misma para los cuatro tipos de botes E_i, no son lo suficientemente grandes como para refutar la hipótesis nula. Las diferencias no son significativas y pueden atribuirse simplemente a un error de muestreo.

Prueba de ajuste de un parámetro específico

Existen muchos casos en los cuales las frecuencias se prueban contra un patrón determinado, en el cual las frecuencias esperadas no son todas iguales. En su lugar deben determinarse así:

$$E_i = np_i$$

Donde

n es el tamaño de la muestra

p_i es la probabilidad de cada categoría como se especificó en la hipótesis nula

El John Dillinger National Bank, en New York, trata de seguir una política de extender un 60% sus créditos a empresas comerciales, un 10% a personas naturales y un 30% a prestatarios extranjeros.

Para determinar si la política se está siguiendo, Jay Hoover, vicepresidente de mercadeo, selecciona aleatoriamente 85 créditos que se aprobaron recientemente. Encuentra que 62 de tales créditos se otorgaron a negocios, 10 a personas naturales, y 13 a prestatarios extranjeros. Al nivel del 10%, ¿parece que el patrón de cartera deseado de preserva?

H_o: se mantuvo el patrón deseado: 60% son créditos comerciales, 10% son préstamos personales y 30% son créditos extranjeros.

H_A: el patrón deseado no se mantuvo.

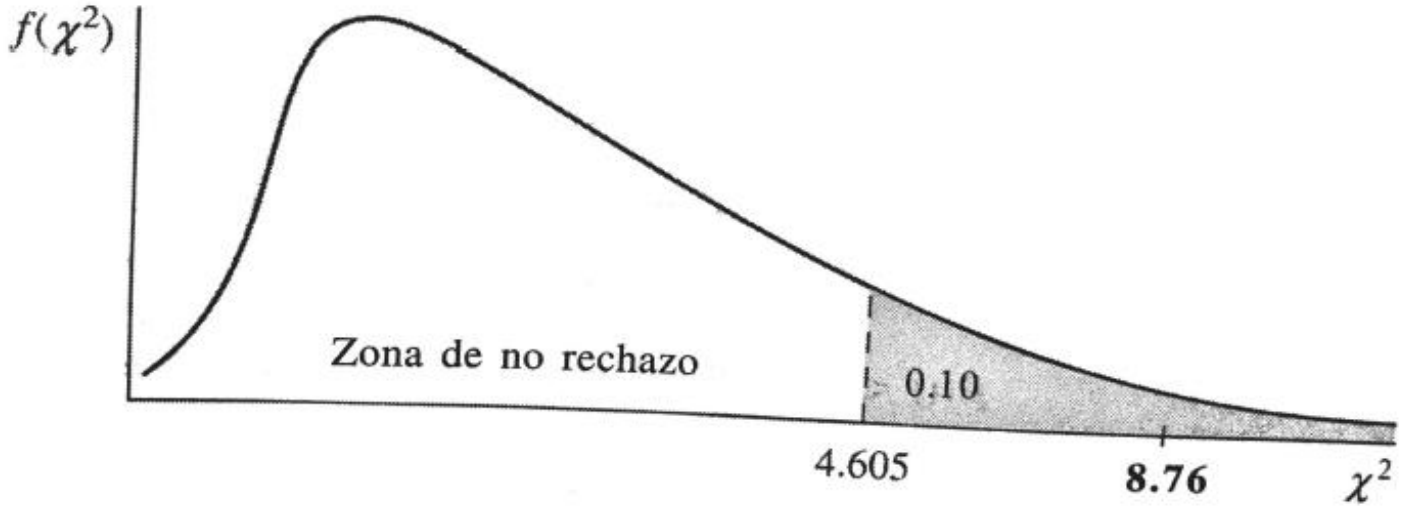
Tipo de crédito	Ventas observadas (O_i)	Ventas esperadas (E_i)
Comercial	62	$E_i = np_i$ = 85*0.6 = 51
Personal	10	$E_i = np_i$ = 85*0.1 = 8.5
Extranjero	13	$E_i = np_i$ = 85*0.3 = 25.5
	85	85

El valor de X^2 es:

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(62-51)^2}{51} + \frac{(10-8.5)^2}{8.5} + \frac{(13-25.5)^2}{25.5} = 8.76$$

Como no se estimaron parámetros m = 0. Con α = 10 y k = 3 categorías de crédito, los grados de libertad son k – m - 1 = 3 – 0 – 1 = 2. El señor Hoover encuentra en la tabla un valor crítico de $X^2_{0.10,2}$= 4.605.

Regla de decisión: "No rechazar si $X^2 \leq 4.605$. Rechazar si $X^2 > 4.605$".



Como lo muestra la figura, la hipótesis nula debería rechazarse debido a que 8.76 > 4.605.

Las diferencias entre lo que el señor Hoover observó y lo que esperaba observar si el patrón de crédito deseado se alcanzaba eran demasiado grande como para ocurrir por simple azar. Existe solo un 10% de probabilidad de que una muestra de 85 créditos seleccionados aleatoriamente pudiera producir las frecuencias observadas aquí demostradas, si el patrón deseado en la cartera de crédito del banco se estuviera manteniendo.

Prueba de normalidad

Las especificaciones para la producción de tanques de aire utilizados en inmersión requieren que los tanques se llenen a una presión promedio de 600 psi. Se permite una desviación estándar de 10 psi. Las especificaciones de seguridad permiten una distribución normal en los niveles de llenado.

Usted acaba de ser contratado por Aqua Lung, su primera tarea es determinar si los niveles de llenado se ajustan a una distribución normal. Aqua Lung está seguro de que la media de 600 psi y la desviación estándar de 10 psi prevalecen. Solo queda probar la naturaleza de la distribución. Para esto se mide n = 1000 tanques y se halla la siguiente distribución:

PSI	Frecuencia real (Oi)
0 y por debajo de 580	20
580 y por debajo de 590	142
590 y por debajo de 600	310
600 y por debajo de 610	370
610 y por debajo de 620	128
620 y por encima	30
	1000

Las hipótesis son:

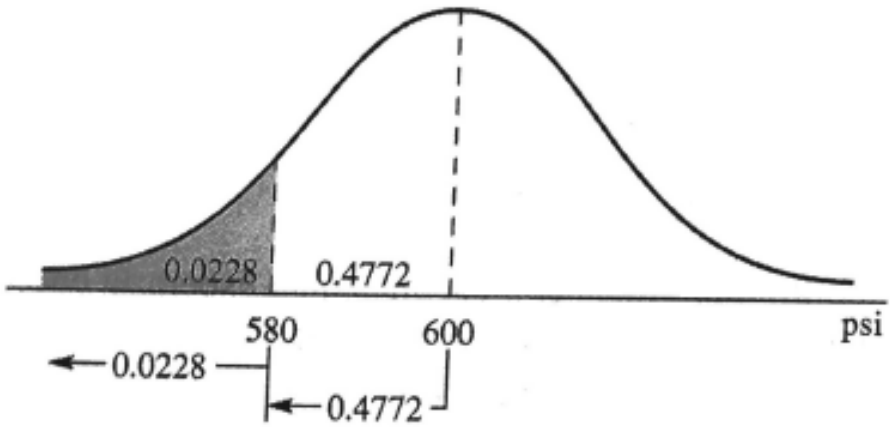
H_o: los niveles de llenado están distribuidos normalmente.
H_A: los niveles de llenado no están distribuidos normalmente.

Para determinar las frecuencias esperadas, se deben calcular las probabilidades de que los tanques seleccionados aleatoriamente tengan los niveles de contenido en los intervalos presentados en la tabla.

Probabilidad de que un tanque caiga en el primer intervalo es P(0 < X < 580)

$$Z = \frac{X - \mu}{\sigma} = \frac{580 - 600}{10} = -2 \; para \; un \; área \; de \; 0.4772 \; (tabla)$$

P(0 < X < 580) = 0.5 – 0.4772 = 0.0228

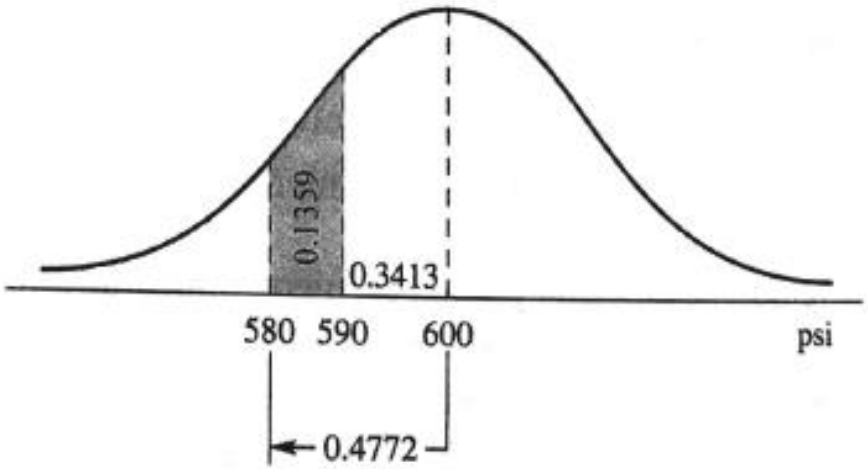


Probabilidad de que un tanque caiga en el primer intervalo es P(580 < X < 590)

$$Z1 = \frac{X - \mu}{\sigma} = \frac{580 - 600}{10} = -2 \; para \; un \; área \; de \; 0.4772 \; (tabla)$$

$$Z2 = \frac{X - \mu}{\sigma} = \frac{590 - 600}{10} = -1 \; para \; un \; área \; de \; 0.3413 \; (tabla)$$

P(580 < X < 590) = 0.4772 – 0.3413 = 0.1359



Universidad de Medellín
Ciencia y Libertad

PSI	Frecuencia real (O_i)	Probabilidades (p_i)	Frecuencia esperada (E_i) $E_i = np_i$
0 y por debajo de 580	20	0.0228	22.8
580 y por debajo de 590	142	0.1359	135.9
590 y por debajo de 600	310	0.3413	341.3
600 y por debajo de 610	370	0.3413	341.3
610 y por debajo de 620	128	0.1359	135.9
620 y por encima	30	0.0228	22.8
	1000	1	1000

El valor de X^2 es:

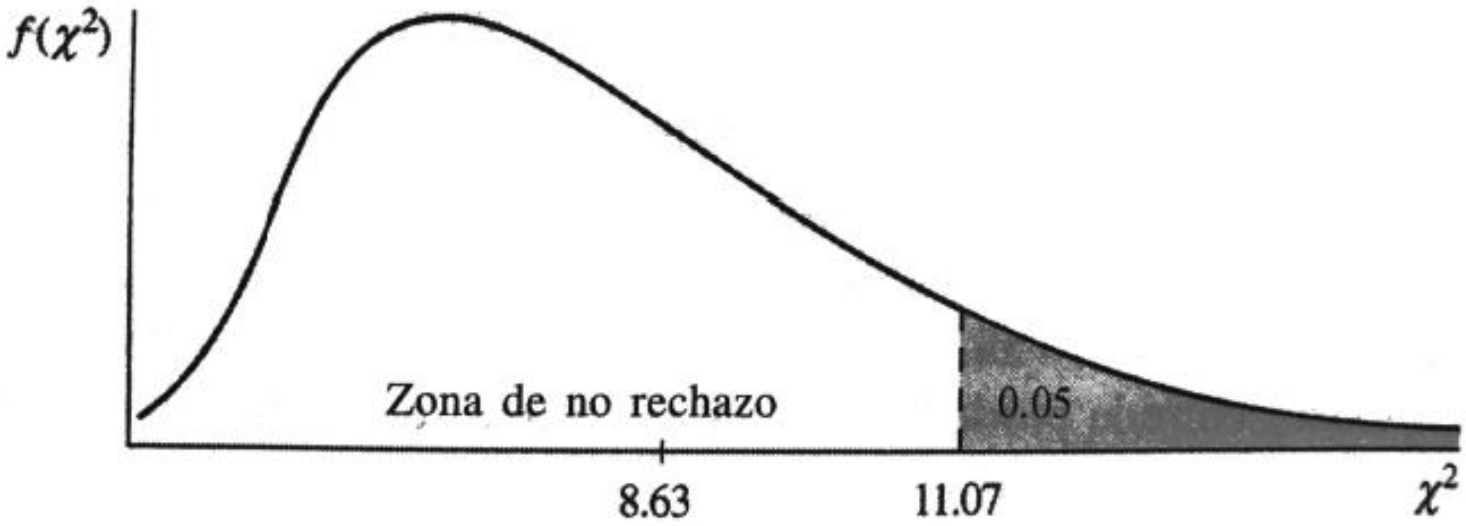
$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(20-22.8)^2}{22.8} + \frac{(142-135.9)^2}{135.9} + \frac{(310-341.3)^2}{341.3}$$
$$+ \frac{(370-341.3)^2}{341.3} + \frac{(128-135.9)^2}{135.9} + \frac{(30-22.8)^2}{22.8} = 8.63$$

Se desea probar la hipótesis al nivel del 5%. Debido a que tanto la media poblacional como la desviación estándar son dadas y no tienen que estimarse,

m = 0. Existen k 0 6 clases en la tabla de frecuencias, de manera que los grados de libertad son k – 1 = 5. Se encuentra en la tabla un valor crítico de $X^2_{0.05,5} = 11.07$.

Regla de decisión: "No rechazar si $X^2 \leq 11.07$. Rechazar si $X^2 > 11.07$".



La hipótesis nula no debería rechazarse. Las diferencias entre lo que se observó y lo que se espera observar si los contenidos estuvieran distribuidos normalmente con una media de 600 psi y una desviación estándar de 10 psi pueden atribuirse al error de muestreo.

Si la media poblacional y la desviación estándar no fueran conocidas, se hubieran tenido que estimar de los datos muestrales. Entonces m sería 2 y los grados de libertad serían k – 2 – 1 o 6 – 2 – 1 = 3.

La prueba chi-cuadrado de bondad de ajuste es confiable solo si todo E_i es por lo menos 5. Si una muestra tiene un $E_i < 5$, debe combinarse con clases adyacentes para garantizar que todas las categorías $E_i \geq 5$.

Si en el ejercicio anterior se hubiera seleccionado una muestra de tan solo n = 100 en lugar de 1000 tanques de inmersión, el Ei para la primera clase hubiera sido $E_i = (100)(0.0228) = 2.28$ en lugar de 22.8. Esta primera clase se combinaría con la segunda clase de manera que $E_i \geq 5$. Por lo anterior también se tendrían que unir la clase 5 y la 6. Esto genera que los grados de libertad se reduzcan de manera considerable.

Tablas de contingencia. Una prueba de independencia

Chi-cuadrado también permite la comparación de dos atributos para determinar si existe una relación entre ellos.

Wilma Keeto es la directora de investigación de productos en Dow Chemical. En su proyecto actual, la señorita Keeto debe determinar si existe alguna relación entre la clasificación de efectividad que los consumidores asignan a un nuevo insecticida y el sitio (urbano o rural) en el cual se utiliza. De los 100 consumidores a quienes se les practicó la encuesta, 75 vivían en zonas urbanas y 25 en zonas rurales.

Atributo A - Clasificación	**Atributo B - Ubicación**		
	Urbano	Rural	Total
Por encima del promedio	20	11	31
Promedio	40	8	48
Por debajo del promedio	15	6	21
Total	75	25	100

La tabla tiene f = 3 filas y c = 2 columnas. Existen fc = 6 celdas en la tabla.

La señorita Keeto desea comparar el atributo B (ubicación) con el atributo A (clasificación del producto). Sus hipótesis son:

H_o: la clasificación y la ubicación son independientes.
H_A: la clasificación y la ubicación no son independientes.

Si la ubicación no tiene ningún impacto en la clasificación efectiva, entonces el porcentaje de residentes urbanos que clasificaron el producto “por encima del promedio” debería ser igual al porcentaje de residentes rurales que clasificaron el producto “por encima del promedio”.

Según la tabla anterior, el 31% de todos los 100 usuarios clasificaron el producto “por encima del promedio”. Por ello, el 31% de los 75 residentes urbanos y el 31% de los 25 residentes rurales deberían dar esta clasificación si la clasificación y la ubicación son independientes.

Estos valores de (75)(0.31) = 23.3 y (25)(0.31) = 7.75 dan las frecuencias esperadas E_i para cada celda, $E_i = np_i$.

Atributo A	**Atributo B**		
	Urbano	Rural	Total
Por encima del promedio	O_i =20 E_i = 23.3	O_i =11 E_i = 7.75	31
Promedio	O_i =40 E_i = 36	O_i =8 E_i = 12	48
Por debajo del promedio	O_i =15 E_i = 15.8	O_i =6 E_i = 5.25	21
Total	75	25	100

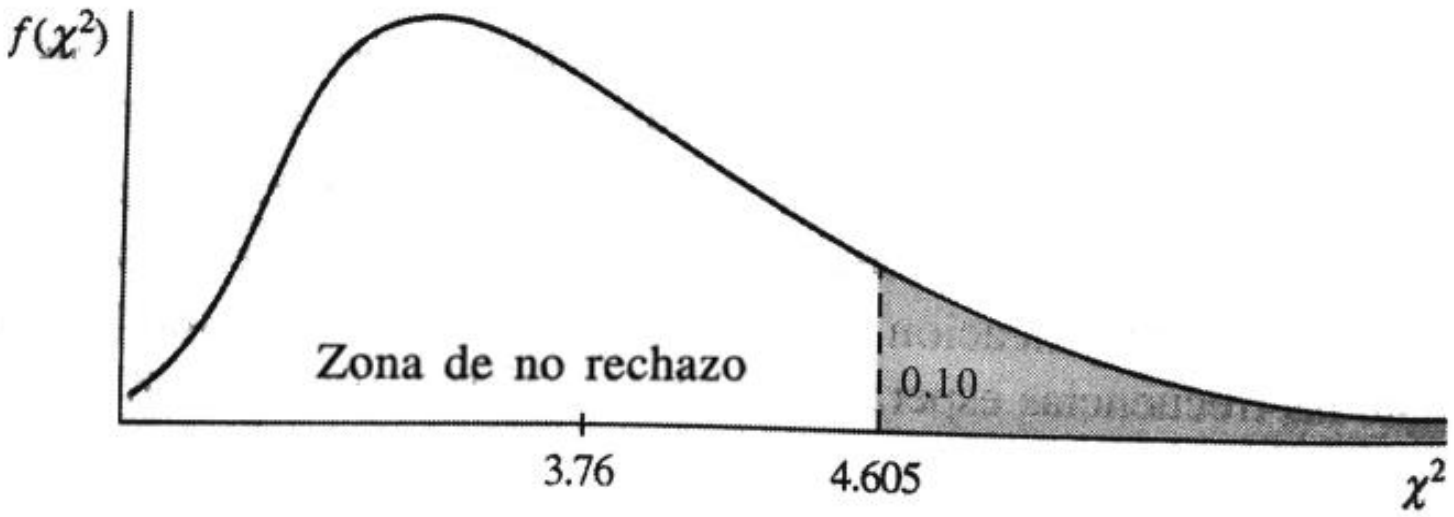
Probar la hipótesis requiere una comparación de O_i y E_i sobre las fc = 6 celdas utilizando la ecuación:

$$X^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

$$X^2 = \frac{(20 - 23.3)^2}{23.3} + \frac{(11 - 7.75)^2}{7.75} + \frac{(40 - 36)^2}{36} + \frac{(8 - 12)^2}{12} + \frac{(15 - 15.8)^2}{15.8} + \frac{(6 - 5.25)^2}{5.25} = 3.76$$

La prueba tiene (f-1)(c-1) = (3 – 1) (2 – 1) = 2 grados de libertad. Si Wilma fija α = 10%, $X^2_{0.10,2} = 4.605$.

Regla de decisión: "No rechazar si $X^2 \leq 4.605$. Rechazar si $X^2 > 4.605$".



La hipótesis nula no debería rechazarse. Se puede concluir que parece que la clasificación y la ubicación son independientes.

Existen otras pruebas no paramétricas, algunas de ellas son:

- La prueba de signo
- La prueba de rachas
- La prueba U de Mann-Whitney
- La prueba de correlación de rangos de Spearman
- La prueba de Kruatal-Waills

Ejercicio 11.1. Hedonistic Auto Sales desea determinar si existe alguna relación entre el ingreso de los clientes y la importancia que dan al precio de los automóviles de lujo. Los gerentes de la compañía desean probar la hipótesis de que:

H_o: ingreso e importancia del precio son independientes.
H_A: ingreso e importancia del precio no son independientes.

Atributo A – nivel de importancia	**Atributo B - Ingreso**			**Total**
	Bajo	Medio	Alto	
Grande	83	62	37	182
Moderado	52	71	49	172
Poco	63	58	63	184
Total	198	191	149	538

Utilice un alfa de 1%

Ejercicio 11.2. El vicepresidente de operaciones del First National Bank argumenta que los tres tipos de crédito – créditos para autos, créditos a estudiantes y créditos para propósitos generales – se conceden a los clientes según un patrón tal que la mitad son para propósitos generales y el resto se dividen de manera equitativa entre los tipos restantes. Para probar su hipótesis, se recolectaron 200 créditos recientes y se encuentra que 55 fueron créditos para autos, 47 para estudiantes y el resto para propósitos generales. ¿Qué se puede concluir a un nivel del 5%?