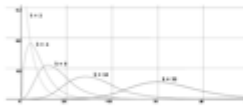


Probabilidad y Estadística

Actividades de Aprendizaje

Conceptos y definiciones de esta clase:



La Distribución chi cuadrado
Frecuencias observadas y
esperadas

Pruebas de independencia
Pruebas de bondad de
ajuste

1. La Distribución chi cuadrado

1.1 Introducción y Fórmula

La distribución χ^2 (que se pronuncia "chi" o "ji" cuadrado) es una distribución de probabilidad continua que suele utilizarse para analizar la asociación de variables cualitativas o los grados de libertad de una de ellas frente a otra. Hay diversas formas de cálculo para esta distribución, siendo una de las más generalizadas la propuesta por Karl Pearson, quien es considerado el padre de la estadística moderna.

Es una distribución que usualmente se utiliza cuando se trata con sujetos (o sucesos) que son clasificados en base a una categorización, como por ejemplo la incidencia del género, nacionalidad o edad para una determinada situación, el análisis de gustos, marcas y preferencias.

Es así como, en general, la preparación para esta prueba comienza con el diseño de un formulario adecuado, como por ejemplo una tabla de doble entrada. A continuación, se muestra un ejemplo:

	Hombres	Mujeres	Totales
Sí			
No			
Total			

Así, por ejemplo, podrá analizarse si el sexo de una persona es un factor determinante para la respuesta. En términos prácticos, esta podría ser una hipótesis se buscará refutar o reforzar mediante el análisis de esta distribución.

Para terminar de cerrar esta idea, pensemos que, frente a la situación planteada, contaremos con dos posiciones extremas, a saber:

	Hombres	Mujeres	Totales
Sí	50%	50%	
No	50%	50%	
Total	100%	100%	

En este caso, vemos que el sexo no ha influido en la respuesta. El 50% de los varones y mujeres encuestados respondió afirmativamente y el otro 50% negativamente", por lo tanto, el valor de chi cuadrado que obtendremos será muy bajo y no podremos rechazar la hipótesis que planteamos.

	Hombres	Mujeres	Totales
Sí	100%		
No		100%	
Total	100%	100%	

Este otro caso, también extremo, es el contrario del anterior. Vemos que la distinción por sexo tiene total influencia en la respuesta. El 100% de los varones respondieron positivamente y el 100% de las mujeres respondieron negativamente. Obtendremos un valor muy alto de chi cuadrado y podremos rechazar la hipótesis de que la influencia del sexo es producto del azar.

1.2 Frecuencias observadas y esperadas

En una distribución chi cuadrado, lo que hacemos en definitiva es comparar entre dos tipos de frecuencias,

- Las denominadas **frecuencias observadas** (o **empíricas**), que son las que observamos, y
- Las **frecuencias esperadas** (o **teóricas**), que son las más probables en el caso de que no haya relación.

Para relacionarlo con lo visto en el punto anterior, diremos que un valor pequeño de χ^2 nos indica que no hay relación entre el indicador y su posible incidencia en los resultados (las frecuencias que observamos se parecen mucho a las teóricas, a las que tendríamos en caso de no asociación o no diferencia).

En cambio, un valor grande de χ^2 indica que sí hay relación (las frecuencias que observamos se apartan mucho de las teóricas).

1.3 ¿Cuándo utilizar la distribución χ^2 ?

Esta distribución se utiliza en dos tipos de situaciones, que se denominan así:

- **Pruebas de independencia**, cuando hay dos criterios de clasificación, como en el ejemplo anterior, con una tabla de doble entrada, donde observábamos si los hombres y mujeres pensaban distinto o no. Es decir, si existía o no algún tipo de interdependencia respecto del género.
- La prueba clásica de este tipo es la de conocer si existe interdependencia entre el color de cabello y el color de ojos, ejercicio que se encuentra explicado en muchos textos. Dejamos en manos del lector revisar la bibliografía para encontrarlo.
- **Pruebas de bondad de ajuste**, cuando tenemos un solo criterio de clasificación (un conjunto subdividido en varias categorías) y deseamos determinar si existe una diferencia significativa entre una distribución de frecuencias observada y una distribución teórica para describir a la distribución observada. Por ejemplo, saber si la edad de las personas que asisten a un determinado club sigue una distribución normal o no.

Desarrollemos con mas detalle cada una de las pruebas.

1.4 Pruebas de independencia

En esta sección aprenderemos cómo se utiliza la distribución χ^2 para comparar dos atributos o características y determinar si mantienen interdependencia.

Consideraciones:

- Para calcular el número de grados de libertad de una prueba de independencia chi-cuadrado se multiplica el número de filas (menos uno) por el número de columnas (menos 1)

$$\text{grados de libertad} = (\text{nro filas} - 1) \times (\text{nro columnas} - 1)$$

- La expresión para calcular la frecuencia esperada f_e para cualquier celda de una tabla de contingencia viene dada por:

$$f_e = \frac{\text{total de la fila} \times \text{total de la columna}}{\text{total de observaciones}}$$

A los totales de fila y columna también se los conoce como sumas marginales.

- El valor calculado de χ^2 se obtiene mediante la siguiente fórmula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- La prueba de independencia siempre es de una cola, con la región de rechazo en la cola superior de la distribución chi-cuadrado. **(1)**
- Las frecuencias esperadas deben ser de 5 o más para todas las categorías, aunque en algunos casos puede aceptarse una tolerancia de hasta 25%. Frente a esta situación, una opción es recategorizar la muestra, indicando menos niveles.

Pasos generales para las pruebas de independencia:

1. Establecer dos Hipótesis distintas ya que por lo mencionado en **(1)** el valor encontrado podrá estar incluido o no en la zona de rechazo. A estas Hipótesis las llamaremos: H_0 y H_a
2. Seleccionar el nivel de significancia.
3. Calcular los grados de libertad.
4. Hallar los valores esperados mediante las sumas marginales.
5. Examinar si el problema es factible de analizar mediante χ^2 . Es un paso que muchas veces se omite pero que es sumamente importante. La tolerancia es que las celdas con un valor esperado menor que cinco no superen el 25% del total de celdas.
6. Determinar el valor teórico de χ^2 (que depende del nivel de significancia y de los grados de libertad) y determinar la región de aceptación y rechazo. Este paso se realiza mediante el uso de una tabla de la distribución teórica de χ^2 o bien mediante una calculadora online.
7. Encontrar el valor calculado de χ^2
8. Analizar la región en donde se encuentra el χ^2 calculado y sacar conclusiones. Si χ^2 cae en la región de rechazo, se rechazará la H_0 .

Ejemplo 1

Supongamos que se desea saber si el sexo de una persona está relacionado con lo que opinan sobre la calidad del servicio ofrecido por un restaurante, para lo cual se han consultado 128 personas a la salida del lugar, entre hombres y mujeres, y se les ha pedido que pongan una calificación al servicio otorgado. La siguiente tabla resume las opiniones:

Observados	Calidad del Servicio				
	Malo	Regular	Muy bueno	Excelente	Sumas
Mujer	6	15	16	9	46
Hombre	31	32	16	3	82
	37	47	32	12	128

Seguiremos los pasos establecidos para la resolución del ejercicio. Entonces de acuerdo al enunciado del problema planteamos dos hipótesis distintas, para ver si existe o no interdependencia.

Paso 1

H_0 : El sexo de una persona es independiente de su opinión sobre el servicio.

H_a : Hay relación entre el sexo y lo que opina.

Paso 2

Habitualmente un nivel de significancia del cinco por ciento ($\alpha=0.05$) funciona de manera adecuada. Un nivel de significancia de 0.05 indica un riesgo de 5% de concluir que existe una asociación entre las variables cuando no hay una asociación real.

Paso 3.

Grados de libertad: $(4 - 1) \times (2 - 1) = 3$ (explicado en el apartado 1.4, donde dice "consideraciones")

Paso 4

Hallamos los valores esperados, utilizando las sumas marginales (suma de la fila y suma de la columna). Por ejemplo, para la primera celda (Mujer y Malo), el valor se calcula con la suma de la fila "mujer" (46,0) y la columna "malo" (37) y su valor esperado nos da:

$$f_e = \frac{\text{total de la fila} \times \text{total de la columna}}{\text{total de observaciones}} = \frac{46 \times 37}{128} \cong 13.3$$

Lo mismo hacemos para todas las celdas.

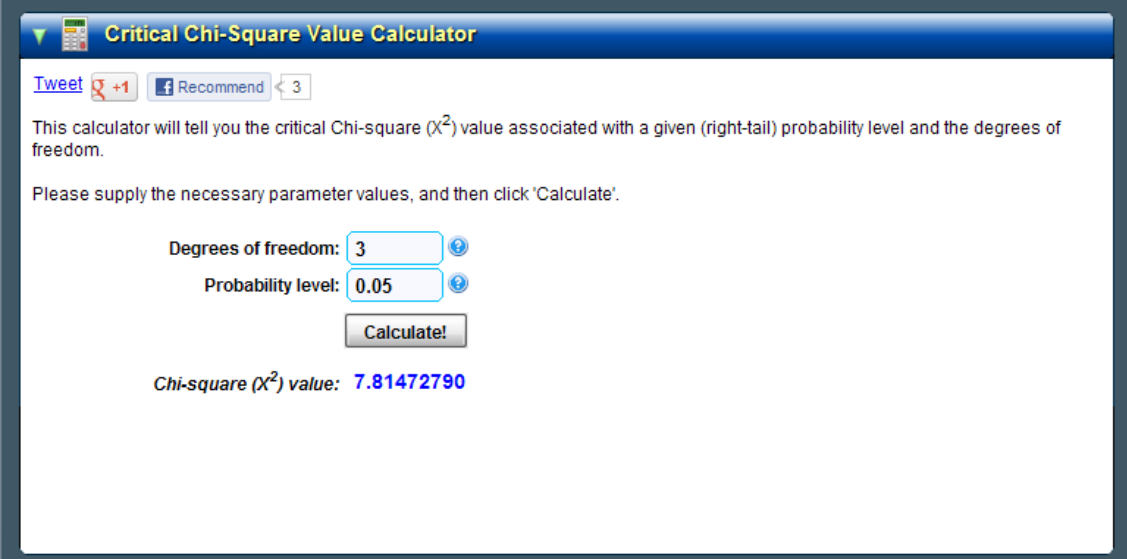
Esperados	Calidad del Servicio				
	Malo	Regular	Muy bueno	Excelente	Sumas
Mujer	13,3	16,9	11,5	4,3	46,0
Hombre	23,7	30,1	20,5	7,7	82,0

Paso 5

Como solamente una frecuencia esperada dio un resultado menor que 5, el análisis puede realizarse sin problemas.

Paso 6

Hallamos el valor teórico de χ^2 (lo calculamos volcando los datos en la siguiente página <http://www.danielsoper.com/statcalc3/calc.aspx?id=12>)



Critical Chi-Square Value Calculator

Tweet +1 Recommend 3

This calculator will tell you the critical Chi-square (χ^2) value associated with a given (right-tail) probability level and the degrees of freedom.

Please supply the necessary parameter values, and then click 'Calculate'.

Degrees of freedom:

Probability level:

Chi-square (χ^2) value: **7.81472790**

Paso 7

Hallamos el valor calculado de χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(6 - 13.3)^2}{13.3} + \frac{(15 - 16.9)^2}{16.9} + \frac{(16 - 11.5)^2}{11.5} + \dots + \frac{(3 - 7.7)^2}{7.7} = 17.3$$

Paso 8

Como el resultado calculado está por encima del esperado, entra en su área de rechazo, entonces se rechaza H_0 y se concluye que el sexo de la persona debe tener alguna influencia sobre su opinión.

1.5 Pruebas de bondad de ajuste

Como ya fue mencionado anteriormente, se trata de una prueba estadística que permite determinar si existe una diferencia significativa entre una distribución de frecuencias observadas y una distribución teórica basada en una hipótesis que describe la distribución observada.

Si la diferencia entre las distribuciones de los sucesos observados y de los esperados es demasiado grande para poderla atribuir a un error de muestreo, se

tiene que llegar a la conclusión de que la población presenta una distribución distinta de la especificada en la hipótesis nula.

La prueba de bondad de ajuste siempre es de una cola, con la región de rechazo en la cola superior de la distribución chi cuadrado.

Los pasos generales son:

1. Establecer H_0 y H_a
2. Seleccionar el nivel de significancia.
3. Establecer los grados de libertad, que se calculan como el número de categorías (o clases) menos el número de parámetros a estimar menos uno. Debemos entender los grados de libertad como los valores que podemos considerar (o elegir) libremente.
4. Se calcula χ^2 al igual que antes, pero teniendo en cuenta que las frecuencias observadas son las correspondientes a los sucesos en los datos muestrales, mientras que las esperadas son las deducidas mediante la hipótesis. En este caso aplicaremos la distribución de Poisson para su calculo
5. Examinar si el problema es factible de analizar mediante χ^2 . Es un paso que muchas veces se omite pero que es sumamente importante. La tolerancia es que las celdas con un valor esperado menor que cinco no superen el 25% del total de celdas.
6. Se obtiene el χ^2 teórico y se comparan ambos para sacar conclusiones.

Ejemplo 2

Se desea saber si la cantidad de errores encontrados en las pruebas de tarjetas SIM fabricadas por una determinada máquina siguen una distribución de Poisson, para lo cual se toma una muestra de 50 tarjetas, y se las analiza arrojando los siguientes datos:

Errores	Frecuencia Observada
0	27
1	13
2	8
3 o más	2

Vamos a realizar una prueba de bondad de ajuste para ver si estos datos muestran evidencia suficiente de que hay una distribución Poisson.

H_0 : La distribución es Poisson.

H_a : La distribución no es Poisson.

Como desconocemos la media de la distribución Poisson la calcularemos

$$\lambda = \frac{0 \times 27 + 1 \times 13 + 2 \times 8 + 3 \times 2}{50} = 0.7$$

Con el parámetro 0.7 conocido calculamos la probabilidad de cada valor para una distribución de Poisson, por ejemplo, para 0 error será:

$$P(0; 0.7) = \frac{e^{-0.75} 0.7^0}{0!} = 0.497$$

Luego calculamos cada frecuencia multiplicando este número por el total de observaciones (n=50) y los volcamos en una tabla.

En el caso de 0 error entonces será: $0.497 \cdot 50 = 24.85$

Y así sucesivamente con los demás errores, vamos completando la tabla.

Errores	Probabilidad	Frec. Esperada	Frec. Observada
0	0.497	24.85	27
1	0.348	17.4	13
2	0.122	6.1	8
3 o más	0.028	1.4	2

Observemos que la última frecuencia observada es menor que 5. Entonces procedemos a combinar las últimas dos filas.

Errores	Probabilidad	Frec. Esperada	Frec. Observada
0	0.497	24.85	27
1	0.348	17.4	13
2 o más	0.150	7.5	10

Los grados de libertad son igual a $3-1-1=1$ puesto que la media fue estimada a partir de los datos del problema.

Luego calculamos el valor teórico de χ^2 , como en el ejercicio anterior y obtenemos el valor 3.84

Ahora hallamos el valor calculado χ^2 , es decir el estadístico de prueba que nos da:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(27 - 24.85)^2}{24.85} + \frac{(13 - 17.4)^2}{17.4} + \frac{(10 - 7.5)^2}{7.5} = 2.13$$

Comparando, **como 2.13 es menor que 3.84, no se rechaza la hipótesis nula y se concluye que la distribución en los errores de las tarjetas SIM sigue una distribución de Poisson.**

Situación Problemática 1

Pruebe la hipótesis de que la distribución de frecuencia de las duraciones de baterías para notebooks dadas en la siguiente tabla, se puede aproximar mediante una distribución normal con media $\mu = 3.5$ y desviación estándar $\sigma = 0.7$. Utilice un $\alpha = 0.05$.

Límites de clase	Frecuencias Observadas
Menor que 1.75	3
De 1.75 a 2.25	2
De 2.25 a 2.75	5
De 2.75 a 3.25	16
De 3.25 a 3.75	11
De 3.75 a 4.25	6
De 4.25 a 4.75	4