Revista Ciencias 2017; 1:13-17 / ISSN 2521-537X

PRUEBA CHI-CUADRADO EN LA ESTADÍSTICA NO PARAMÉTRICA

CHI-SQUARE TEST IN NON-PARAMETRIC STATISTICS

Luis Andrés Amaya Cedrón[1]

RESUMEN

En este trabajo de investigación, la parte de la estadística que se aplica a la investigación es la fórmula o modelo estadístico Chi-Cuadrado, la cual se aplicará a diversos casos en la estadística no paramétrica. Se estudiará primero el uso de la prueba Chi-Cuadrado, para la prueba de Bondad de Ajuste en una variable. Luego estudiaremos el uso de la prueba Chi-Cuadrado, para la prueba de Homogeneidad entre dos variables y finalmente se estudiará la prueba Chi-Cuadrado, para la prueba de Dependencia entre dos variables.

Concluimos que es importante la aplicación de este modelo Chi-cuadrado, y los resultados obtenidos a partir del análisis y conclusión podrán ser de gran utilidad para tomar decisiones.

Palabras claves: Chi-cuadrado, Estadística no Paramétrica.

ABSTRACT

In this research work, part of the statistics applied to the research is the Chi-Square statistical formula or model, which will be applied to various cases in non-parametric statistics. We will first study the use of the Chi-Square test, for the test of Goodness of Fit in a variable. Then we will study the use of the Chi-Square test for the homogeneity test between two variables, and finally the Chi-Square test will be studied for the Dependency test between two variables.

We conclude that the application of this Chi-square model is important, and the results obtained from the analysis and conclusion can be very useful for making decisions.

Keywords: Chi-square, Non-Parametric Statistics

INTRODUCCIÓN

El término estadística proviene del latín *statisticum collegium* 'consejo de Estado' y de su derivado italiano *statista* 'hombre de Estado o político'. En 1749, el alemán Gottfried Achenwall comenzó a utilizar la palabra alemana statistik para designar el análisis de datos estatales. Por lo tanto, los orígenes de la estadística están relacionados con el gobierno y sus cuerpos administrativos.

En ESTADÍSTICA PARAMÉTRICA se asume que la población de la cual la muestra es extraída es **NORMAL o aproximadamente normal**. Esta propiedad es necesaria para que la prueba de hipótesis sea válida. Sin embargo, en un gran número de casos no se puede determinar la distribución original ni la distribución de los estadísticos por lo que en realidad no tenemos parámetros a estimar, tenemos solo distribuciones que comparar, esto se llama ESTADÍSTICA NO-PARAMÉTRICA.

Además hay pruebas en la que existen supuestos sobre las distribuciones poblacionales de la media muestral y del valor de

1 Doctor en Administración de la Educación.
Docente de la Facultad de Ciencias - Universidad Nacional Jorge Basadre Grohmann.

la media poblacional. En el caso de que uno de sus supuestos no se cumpla, las técnicas paramétricas (si no son robustas) generarán resultados erróneos y por ende las conclusiones de sus hipótesis serán inválidas. Las técnicas estadísticas no paramétricas ofrecen menor rigidez con respecto a sus condiciones que las técnicas paramétricas, aunque sacrificando para ello su potencia de explicación.

Son procedimientos estadísticos que poseen ciertas propiedades bajo supuestos generales y sin importar la población de la cual los datos han sido obtenidos. La mayoría de las veces estos supuestos se refieren, por ejemplo, a la simetría o continuidad de la distribución poblacional. La inferencia no paramétrica constituye un campo muy amplio que va desde las equivalencias no paramétricas de las pruebas paramétricas existentes hasta llegar a las estimaciones de punto e intervalo de constantes poblacionales que no pueden ser llevadas a modelos paramétricos por su complejidad (percentiles, deciles, etc.). El rápido desarrollo de las técnicas no paramétricas ha sido en parte por las siguientes razones:

Las técnicas no paramétricas hacen supuestos muy generales respecto a la distribución de probabilidad que siguen los datos. En particular, dejan de lado el supuesto de normalidad en una población.

Son aplicables cuando la teoría de normalidad no puede ser utilizada, por ejemplo cuando no se trabaja con magnitudes de observaciones.

Análisis no paramétrico.

Se denominan pruebas no paramétricas aquellas que no presuponen una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre (distribution free). En la mayor parte de ellas los resultados estadísticos se derivan únicamente a partir de procedimientos de ordenación y recuento, por lo que su base lógica es de fácil comprensión. Cuando trabajamos con muestras pequeñas ($n < 10$) en las que se desconoce si es válido suponer la normalidad de los datos, conviene utilizar pruebas no paramétricas, al menos para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal.

Ventajas de los Métodos No Paramétricos

1. Los métodos no paramétricos pueden ser aplicados a una amplia variedad de situaciones porque ellos no tienen los requisitos rígidos de los métodos paramétricos correspondientes. En particular, los métodos no paramétricos no requieren poblaciones normalmente distribuidas.
2. Diferente a los métodos paramétricos, los métodos no paramétricos pueden frecuentemente ser aplicados a datos no numéricos, tal como el género de los que contestan una encuesta.
3. Los métodos no paramétricos usualmente involucran simples computaciones que los correspondientes en los métodos paramétricos y son por lo tanto, más fáciles para entender y aplicar.

Desventajas de los Métodos No Paramétricos

1. Los métodos no paramétricos tienden a perder información porque datos numéricos exactos son frecuentemente reducidos a una forma cualitativa.
2. Las pruebas no paramétricas no son tan eficientes como las pruebas paramétricas, de manera que con una prueba no paramétrica generalmente se necesita una evidencia más fuerte (así como una muestra más grande o mayores diferencias) antes de rechazar una hipótesis nula.

Cuando los requisitos de la distribución de una población son satisfechos, las pruebas no paramétricas son generalmente menos eficientes que sus contrapartes paramétricas, pero la reducción de eficiencia puede ser compensada por un aumento en el tamaño de la muestra.

Según (Aviles Garay, José, 2001) Para realizar análisis no paramétricos debe partirse de las siguientes consideraciones:

1. La mayoría de estos análisis no requieren de presupuestos acerca de la forma de la distribución poblacional. Aceptan distribuciones no normales.
2. Las variables no necesariamente deben estar medidas en un nivel para intervalos o de razón, pueden analizar datos nominales u ordinales. De hecho, si se quiere aplicar análisis no paramétricos a datos de intervalos o razón, estos deben ser resumidos a

categorías discretas (a unas cuantas). Las variables deben ser categóricas.

Métodos o pruebas estadísticas no paramétricas más utilizadas.

- La ji cuadrada o x^2
- Los coeficientes de correlación en independencia para tabulaciones cruzadas.
- Los coeficientes de correlación por rangos ordenados de Spearman y Kendall.

MATERIALES

En el presente trabajo se utilizaron materiales bibliográficos acerca de Estadística, Estadística Inferencial y No paramétrica, etc. también se usó material bibliográfico sobre el software estadístico Minitad.

Métodos

Se realizó el estudio bibliográfico acerca de Estadística, Estadística descriptiva, Inferencial y No paramétrica, y sus aplicaciones a la investigación. Las aplicaciones (ejemplos) son procesados con el software Minitad 16; PRUEBA CHI-CUADRADO PARA UNA VARIABLE:

1. Prueba de bondad de ajuste

Estas pruebas permiten verificar que la población de la cual proviene una muestra, tiene una distribución especificada o supuesta.

Sea X: variable aleatoria poblacional

f0(x) la distribución (o densidad) de probabilidad especificada o supuesta para X

Se desea probar la hipótesis:

Ho: $f(x) = f_0(x)$

En contraste con la hipótesis alterna:

H_1: $f(x) \neq f_0(x)$ (negación de Ho)

Frecuencias esperadas (teóricas), a las que denotaremos por $e_1, e_2, e_3, \ldots, e_k$. Se cumplirá:

$$e_1 + e_2 + e_3 + \ldots + e_k = n$$

	Frecuencia O	Frecuencia E
Clase 1	o_1	e_1
Clase 2	o_2	e_2
...	...	...
Clase k	o_k	e_k
Total	n	n

El estadístico de contraste será:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Ejemplo:

Se ha tomado una muestra aleatoria de 40 baterías y se ha registrado su duración en años. Estos resultados se los ha agrupado en 7 clases en el siguiente cuadro

i	Clase (Duración)	Frec. Observada o_i
1	1.45-1.95	2
2	1.95-2.45	1
3	2.45-2.95	4
4	2.95-3.45	15
5	3.45-3.95	10
6	3.95-4.45	5
7	4.45-4.95	3

Verificar con 5% de significancia que la duración en años, de las baterías producidas por este fabricante tiene duración distribuida normalmente con media 3.5 y desviación estándar 0.7.

Solución

Sea X: duración en años (variable aleatoria continua)

Hipótesis

Ho: X tiene una distribución Normal (μ=3.5, σ=0.7)

Ha: No Ho

α=0.05

Cálculo de la probabilidad a cada intervalo

p_1=0.0136; p_2=0.0532; p_3=0.135; p_4=0.2575; p_5=0.2675 p_6=0.175; p_7=0.075

Cálculo de las frecuencias esperadas

$e_1 \approx 0.5$; $e_2 \approx 2.1$; $e_3 \approx 5.4$; $e_4 \approx 10.3$; $e_5 \approx 10.7$; $e_6 \approx 7$; $e_7 \approx 3.5$

Resumen de resultados

Duración (años)	F. Observa (o_i)	F. Esperada (ei)
1.45-1.95	2	0.5
1.95-2.45	1	2.1
2.45-2.95	4	5.4
2.95-3.45	15	10.3
3.45-3.95	10	10.7
3.95-4.45	5	7
4.45-4.95	3	3.5

Es necesario que se cumpla la condición $\forall i, \quad e_i \geq 5$ por lo que se deben agrupar clases adyacentes. Como resultado se tienen cuatro clases **k=4**

Duración (años)	F. Observa (o_i)	F. Esperada (ei)
1.45-2.95	7	8.5
2.95-3.45	15	10.3
3.45-3.95	10	10.7
3.95-4.95	8	10.5

Ahora se puede definir la región de rechazo de Ho
Observemos que en este ejemplo la media y la desviación estándar de la distribución normal no se estimaron, sino que están propuestas, de donde r = 0

$$\alpha = 0.05, \quad v = k - r - 1 = 4 - 0 - 1 = 3$$
$$\Rightarrow \chi^2_{0.05} = 7.815; \quad \text{Tomado de la tabla } \chi^2$$

Rechazamos H_0 si $\chi^2 > \chi^2_{0.05} = 7.815$

Cálculo del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)}{e_i} = \left[\frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} \right] = 3.05$$

Decisión

Como 3.05 no es mayor a 7.815, se dice que no hay evidencia suficiente para rechazar el modelo propuesto para la población.

PRUEBA CHI-CUADRADO PARA DOS VARIABLES:

A. Prueba de Independencia y prueba de homogeneidad

Concretamente, usaremos el estadístico:

$$\chi^{2^*} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}; \text{ con (m-1)(k-1) grados de libertad.}$$

Supongamos que deseamos establecer si hay homogeneidad entre la proporción de aprobados en la misma clase de matemáticas, es igual tanto para estudiantes que provienen de escuelas públicas como de escuelas privadas, si hay relación entre las variables tipo de escuela superior y la aprobación de la primera clase de matemáticas que toma el estudiante en la universidad, usando los datos de 20 estudiantes que se muestran abajo

Estudiante	1	2	3	4		5	6	7	8	9	10	11	12	13	14
Escuela	Pr	Pr	Pu	Pr		Pu	Pu	Pu	Pr	Pu	Pr	Pu	Pr	Pu	Pr
Aprueba	si	no	no	si		si	no	si	si	si	si	si	no	no	si

De donde obtenemos el cuadro de doble entrada:

	Aprueba			
Escuela	No	Si		
Privado	3	7	10	n_1
Pública	5	5	10	n_2
Total	8	12	**20**	
	n_1	n_1		

Solución:

Para la prueba de homogeneidad.

H_0 = La proporción de aprobados en la primera clase de matemáticas es igual tanto para estudiantes que provienen de escuela pública como de escuela privada.
H_a = La proporción de aprobados en la primera clase de matemáticas no es la misma para ambos tipos de escuela.

Hallamos las probabilidades y *frecuencias esperadas*:

Usando: $p_i = n_j / n$; Por lo tanto:

$e_{ij} = n_i \cdot n_j / n$; para *i=1, 2; j=1,2*

Para i=1; y j=1, 2.

$$p_1 = n_1 / n = \frac{8}{20} = 0.4 \Rightarrow e_{11} = n_1 \cdot n_1 / n = 10 * \frac{8}{20} = 4$$

$$p_1 = n_2 / n = \frac{12}{20} = 0.6 \Rightarrow e_{12} = n_1 \cdot n_2 / n = 10 * \frac{12}{20} = 6$$

Para i=2; y j=1, 2.

$$p_2 = n_1 / n = \frac{8}{20} = 0.4 \Rightarrow e_{21} = n_2 \cdot n_1 / n = 10 * \frac{8}{20} = 4$$

$$p_2 = n_2 / n = \frac{12}{20} = 0.6 \Rightarrow e_{22} = n_2 \cdot n_2 / n = 10 * \frac{12}{20} = 6$$

Estudiante	15	16	17	18	19	20
Escuela	Pr	Pu	Pr	Pu	Pu	Pr
Aprueba	si	no	no	si	no	si

Obtenemos:

	Aprueba			
Escuela	No	Si		
Privado	3(4)	7(6)	**10**	n_1
Pública	5(4)	5(6)	**10**	n_2
Total	**8**	**12**	**20**	
	n_1	n_1		

Las frecuencias esperadas bajo homogeneidad son las representadas entre paréntesis.
El estadístico del contraste será:

$$\chi^{2^*} = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{2} \frac{(n_{i1} - e_{i1})^2}{e_{i1}} + \frac{(n_{i2} - e_{i2})^2}{e_{i2}}$$

$$= \frac{(n_{11} - e_{11})^2}{e_{11}} + \frac{(n_{12} - e_{12})^2}{e_{12}} + \frac{(n_{21} - e_{21})^2}{e_{11}} + \frac{(n_{22} - e_{22})^2}{e_{22}}$$

$$\chi^{2^*} = \frac{(3-4)^2}{4} + \frac{(7-6)^2}{6} + \frac{(5-4)^2}{4} + \frac{(5-6)^2}{6} = 0.25 + 0.166 + 0.25 + 0.166 = 0.8333$$

$$\chi^{2^*} = 0.8333$$

Con (2-1)(2-1)=1 grados de libertad.

Este valor $\chi^{2^*} = 0.8333$ del estadístico Ji-cuadrado es menor que el valor para el nivel de significación del 5%, por lo tanto debemos concluir que existe Independencia $\chi^2_{0.05}(1) = 5.0239$

Es decir, la hipótesis nula de Independencia es aceptada y se concluye de que, no hay asociación entre el tipo de escuela de donde proviene el estudiante y el resultado que obtiene en la primera clase de matemáticas.

DISCUSIÓN DE RESULTADOS

La importancia de esta parte de estudio se ve reflejada en sus aplicaciones y sus resultados están respaldados, para: *Prueba de bondad de Ajuste*, desde las págs. 18 al 32. *Prueba de Homogeneidad* conforme al ejemplo 5.8, 5.9 y 5.10 en las págs. 40, 46 y 50 respectivamente. Prueba de Independencia, desde la págs. 51 al 64.

Finalmente, los resultados son contrastados con todo libro de estadística y de estadística aplicada a la investigación, además consideramos que este trabajo de investigación es un aporte que permitirá contribuir a futuras investigaciones de estudiantes de universidades e Instituciones educativas

CONCLUSIONES

General

- Se aplicó la prueba chi-cuadrado en la estadística NO paramétrica

Específicos:

- Se Aplicó la prueba Chi-Cuadrado, para la prueba de Bondad de Ajuste en una variable.
- Se Aplicó la prueba Chi-Cuadrado, para la prueba de Homogeneidad entre dos variables
- Se Aplicó la prueba Chi-Cuadrado, para la prueba de Dependencia entre dos variables

REFERENCIAS BLIBLIOGRÁFICAS

- AVILA ACOSTA, R. (2001). Metodología de la investigación. Lima - perú: Estudios y Ediciones R.A.
- De la Horra Navarro J. (2003): "Estadística Aplicada". Ediciones Díaz de Santos. España.
- GALAN MORENO, M. (11 de Abril de 2006). Complemento de Matematicas. Apuntes de estadistica descriptiva. España: ETSA.
- GARCIA MANCILLA, H. (2011). Estadistica descriptiva e inferencial I. España: Colegio de Bachilleres.
- GONZÁLEZ RODRÍGUEZ, B. (2013). Estadística descriptiva. Laguna, España.
- MOORE D. S. (2000): "Estadística Aplicada Básica". Antoni Bosch Editor S.A. España.
- NAVIDI WILLIAM (2006): "Estadística para Ingenieros y Científicos". Ed. McGraw-Hil.
- ORELLANA, L. (1 de MARZO de 2001). ESTADÍSTICA DESCRIPTIVA. Recuperado el 1 de mARZO de 2015, de ESTADÍSTICA: http://www.hacienda.go.cr/cifh/sidovih/cursos/material_de_apoyo-F-C-CIFH/1MaterialdeapoyocursosCIFH/4Estad%C3%ADsticaBasica/Estadisticadescriptiva-LillianaOrellana.pdf
- WAYNE, D. (1988). Estadistica con Aplicaciones a la ciencias sociales y a la educación. México D.F: McGraw - Hill.

Otros

" https://www.youtube.com/watch?v=vYcJ8MUEogg
" http://www.slideshare.net/olivaresmtro/formulacion-de-hipotesis
" http://www.monografias.com/trabajos57/hipotesis-investigacion/hipotesis-investigacion2.shtml
" http://academic.uprm.edu/eacuna/miniman8sl.pdf