

USO DE R-STUDIO EN MODELOS NO PARAMÉTRICOS Y DE REGRESIÓN.

PRESENTA: DULCE MARÍA REYES VARELA

TUTORA: DRA. SOFÍA VILLERS GÓMEZ



ACTIVIDAD DE APOYO A LA DOCENCIA

Este trabajo es un material utilizado como herramienta para apoyar a los docentes y guiar a los alumnos de la clase de Modelos no Paramétricos y de Regresión, correspondiente al sexto semestre de la licenciatura de Actuaría, éste se implementó como una herramienta llamada Bookdown, el cuál por medio del software R-Studio se plasmó la parte teórica, seguida de un ejercicio paso por paso y la respectiva réplica en R-Studio, dándole al alumno todas la herramientas y facilidades para que su aprendizaje sea de manera más didáctica.

Modelos no paramétricos y de Regresión

Sofía Villers Gómez

Dulce María Reyes Varela

Prefacio

Al llegar al curso de Modelos No Paramétricos y de Regresión ya hemos cursado Inferencia estadística en la que estudiamos una serie de métodos de estimación puntual (método de momentos, estimadores de máxima verosimilitud), además aprendimos a evaluar dichos estimadores para encontrar los mejores. Sin embargo, en éste enfoque estadístico se tiene la desventaja de que siempre se trabaja con muestras aleatorias basadas en el supuesto de que siguen cierta distribución conocida, más adelante los conocerán en los ejercicios prácticos.

Algunos de los problemas que tienen las pruebas de hipótesis es que suponen que las observaciones disponibles para el estadístico provienen de distribuciones cuya forma exacta es conocida, aún cuando los valores de algunos parámetros sean desconocidos. En otras palabras, se supone que las observaciones provienen de una cierta familia paramétrica de distribuciones y que se debe hacer inferencia estadística acerca de los valores de los parámetros de dicha familia, comunmente la media, la varianza y en otros casos la proporción.

Mostramos la portada de nuestro Bookdown, dónde damos una breve introducción al Curso de Modelos no Paramétricos y de Regresión.

Este se compone de los siguientes grandes temas, divididos en 32 capítulos:

- Pruebas Binomiales
- Pruebas de Rango
- Tablas de Contingencia
- Bondad de Ajuste
- Regresión Lineal Simple y
- Regresión Lineal Múltiple

- Nuestro objetivo es proporcionar a los alumnos las herramientas necesarias para el desarrollo del curso.
- Reforzar las bases teóricas con contenido electrónico complementado con herramientas de R-Studio.

Ejemplificaremos la función y uso de nuestro Bookdown con el Capítulo 16 de Bondad de Ajuste con la Prueba de la Ji-cuadrada.

Objetivos

- Proporcionar a los alumnos, herramientas suficientes para el curso Modelos no paramétricos y de Regresión.
- Reforzar las bases teóricas con contenido electrónico completado con herramientas de R-Studio.
- Dar continuidad al material para el curso *Modelos no paramétricos y de Regresión*.

Este libro fue escrito con **bookdown** usando **RStudio**.

Esta versión fue escrita con:

Licencia

This work is licensed under a **Creative Commons Attribution-ShareAlike 4.0 International License**.



This is a human-readable summary of (and not a substitute for) the license. Please see <https://creativecommons.org/licenses/by-sa/4.0/legalcode> for the full legal text.

PRUEBAS DE BONDAD DE AJUSTE

Capítulo 16.- Prueba de la Ji-cuadrada

16.1 Datos

Sea X_1, \dots, X_n m. a. de tamaño " n " que proviene de una distribución $F(x)$ desconocida.

Cada una de las variables se pueden acomodar en alguna clase " k " (o categoría)

Cada uno de nuestros capítulos contiene los datos con los que se trabajará cada prueba

16.2 Hipótesis

H_0 : Los datos siguen una distribución $F_0(x)$.

vs

H_a : Los datos no siguen una distribución $F_0(x)$.

- Donde $F_0(x)$ es la distribución que se propone.

Es decir:

H_0 : $P[X \text{ pertenezca a la categoría } j] = P_j$, para toda $j = 1, \dots, k$.

vs

H_a : $P[X \text{ pertenezca a la categoría } j] \neq P_j$, para alguna $j = 1, \dots, k$.

Para cada prueba a utilizar se formularán dos Hipótesis, la hipótesis nula (H_0) y la hipótesis alternativa (H_a).

16.3 Estadístico de Prueba

El estadístico de prueba que ocuparemos es:

$$Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j}, \quad \text{Donde } Q \sim \chi^2_{(k-1)}.$$

Observaciones

- Q es estable cuando el número de observaciones en cada categoría debe ser mayor a 5.
- En caso de que alguna categoría tenga menos de 5 observaciones colapsamos las categorías.
- Si desconocemos los parámetros los estimamos con la muestra (*EMV*, *Momentos*), con esto se van perdiendo grados de libertad.
- $Q \sim \chi^2_{(k-1-r)}$ Donde k =Número de clases o intervalos, r =Número de parámetros estimados.

Regla de decisión.

Rechazamos H_0 si $Q > q_{teórica}$.

- Cuando $Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j} > \chi^2_{(k-1)}(1 - \alpha)$.

Enseguida se tiene el estadístico de prueba, que se calcula a partir de los datos de la muestra y se utiliza para poder determinar si se rechaza o no nuestra hipótesis nula.

La regla de decisión como su nombre nos indica nos ayuda a tomar la decisión de acuerdo a ciertos criterios de en qué casos se rechazará la hipótesis nula.

Después de mostrar la teoría se le proporciona al alumno un ejemplo para que el alumno pueda aplicar todos sus conocimientos...

16.4 Ejemplo

Un gobierno local tiene registros del número de niños y el número de hogares en el área. Se sabe que el número promedio de niños por hogar es 1.40. Se sugiere que el número de niños por hogar se pueda modelar por una distribución Poisson con parámetro 1.40. Para probar esta hipótesis se toma una muestra de 1000 hogares; los resultados se muestran en la siguiente tabla:

Número de niños	0	1	2	3	4	5+
Número de hogares	273	361	263	78	21	4

NOTA: Como tenemos una categoría con observaciones menores a 5 colapsamos la categoría.

Número de niños	0	1	2	3	4+
Número de hogares	273	361	263	78	25

Paso 1 Prueba a utilizar **Prueba Ji-cuadrada**.

Paso 2 Planteamiento de hipótesis:

H_0 : El número de niños por hogar sigue una distribución Poisson (1.40).

vs

H_a : El número de niños por hogar no sigue una distribución Poisson (1.40).

Necesitamos buscar la Distribución de una variable aleatoria Poisson, con parámetro $\lambda = 1.40$.

Nota En R-Studio tenemos `dpois(c(0,1,2,3,4),1.40)` para la primera categoría.

Nos ayudaremos de la siguiente tablita:

Número de niños	Número de hogares	$P_i = \text{dpois}(0 : 4, 1.4)$	$e_i = n \times P_i$
0	273	0.2465	247
1	361	0.3452	345
2	263	0.2416	242
3	78	0.1127	113
4	25	0.0394	39
$n = 1000$		$.9854 \approx 1$	$985.4 \approx 1000$

Paso 3 Estadístico de prueba.

$$Q = \sum_{j=0}^k \frac{(f_j - e_j)^2}{e_j} = \sum_{j=0}^4 \frac{(f_j - e_j)^2}{e_j} =$$
$$= \frac{(273 - 247)^2}{247} + \frac{(361 - 345)^2}{345} + \frac{(263 - 242)^2}{242} + \frac{(78 - 113)^2}{113} + \frac{(25 - 39)^2}{39}.$$
$$T = 21.4605.$$

Y los pasos a seguir para resolver nuestro ejemplo son:

- Paso 1: Prueba a utilizar...

Prueba Ji-cuadrada

- Paso 2: Planteamiento de las hipótesis correspondientes a nuestro ejemplo...

H_0 = El número de niños por hogar sigue una distribución Poisson (1.40)

vs

H_a = El número de niños por hogar no sigue una distribución Poisson (1.40)

- Paso 3: Estadístico de prueba

$$Q = \sum_{j=0}^k \frac{(f_j - e_j)^2}{e_j}$$

Paso 4 Regla de decisión.

Rechazamos H_0 si $Q > q_{teórica}$.

- Cuando

$$Q = \sum_{j=1}^k \frac{(f_j - e_j)^2}{e_j} > \chi_{(k-1)}^2(1 - \alpha).$$

Ocuparemos $\alpha = 0.05$.

$$\chi_{(k-1)}^2(1 - \alpha) = \chi_{(5-1)}^2(1 - 0.05) = \chi_{(4)}^2(.95).$$

$$Q = 21.4605 > 9.48 = \chi_{(4)}^2(.95).$$

∴ Rechazamos H_0 .

Paso 5 Conclusión.

∴ Existe evidencia estadística suficiente para decir que el número de niños por hogar no sigue una Distribución Poisson (1.40).

- Paso 4 Regla de decisión: Si el Estadístico de prueba es mayor al cuantil correspondiente a nuestra prueba entonces rechazamos H_0 . **Se rechaza H_0 .**
- Paso 5: Conclusión: Aquí se le debe recordar al alumno que se debe terminar dándole respuesta a los que se plantea en el ejemplo.
Existe evidencia estadística suficiente para decir que el número de niños por hogar no sigue una distribución Poisson (1.40).

16.5 Ejemplo en R-Studio

Ahora haremos la réplica en R.

```
#Número de hogares
observados =c(273,361,263,78,21,4)
#Matriz de frecuencias observadas
tabla=matrix(observados,nrow=1)
#Número de niños
num_ninos=c("0","1","2","3","4","5+")
#Asigna nombres a la tabla de cada categoría
dimnames(tabla)=list(NULL,num_ninos)
tabla
```

```
      0    1    2    3    4 5+
[1,] 273 361 263 78 21  4
```

Y para concluir el capítulo...

Se replica el ejercicio anterior con nuestra herramienta R-Studio, aquí se puede hacer de manera sencilla el ejercicio, ya que R nos da la facilidad de hacer los cálculos más fáciles y amigables para el alumno.

Ya que al momento de ir avanzando como lo es en Bondad de Ajuste será de gran ayuda poder almacenar nuestros datos y que R nos pueda calcular el estadístico de una manera sencilla.

16.6 Ejercicios

1. Se lanza un dado 600 veces se obtuvieron los siguientes resultados.

Número del dado	1	2	3	4	5	6
Observaciones	87	96	108	89	122	98


¿El dado está balanceado (es decir, los datos tienen distribución uniforme con proba 1/6)? Use $\alpha = 0.10$

2. Cierta banca otorga crédito a las personas con una tasa preferencial, de tal manera que los acreditados pueden pagar en cualquier momento desde que piden el préstamo hasta 8 semanas posteriores para que les sea respetada la tasa preferencial. Se seleccionaron aleatoriamente a 1,000 personas y observó su comportamiento de pago, generando de esta manera la siguiente tabla de frecuencia:

Semana	Créditos Pagados
Menos de 1 semana	64
$1 \leq x < 2$	191
$2 \leq x < 3$	283
$3 \leq x < 4$	241
$4 \leq x < 5$	140
$5 \leq x < 6$	51
$6 \leq x < 7$	25
$7 \leq x < 8$	4
8 semanas o más	1

Probar que el pago de estos créditos, sigue una distribución binomial con parámetros $n = 10$ y $p = 0.25$.

Finalmente terminando cada capítulo se le proporciona al alumno 2 o más problemas relacionados al tema para que pueda resolverlos.



El material se pudo implementar en los semestres 2020-1 de manera presencial y en el 2020-2 de manera tanto presencial como virtual debido a la contingencia sanitaria, cabe mencionar que el formato que se ocupó fue en PDF compartido a los alumnos y divididos en cada uno de los temas del plan de estudios y el formato para la implementación de R-Studio fue en HTML, ya que éste se puede descargar en sus computadoras y revisarse sin necesidad de estar conectado a internet.