8. ANÁLISIS DE DATOS CATEGÓRICOS

Dr. Edgar Acuña
academic.uprm.edu/eacuna

Departmento de Matematicas
Universidad de Puerto Rico- Mayaguez

Introducción

Datos categóricos representan atributos o categorías. Cuando se consideran dos variables categoricas, entonces los datos se organizan en tablas llamadas tablas de contingencia o tablas de clasificacion cruzada.

Primero se discutiran la relación entre las variables que definen las filas y las columnas de tablas de contingencia y luego se estudian medidas que dan una idea del grado de asociación entre las dos variables categóricas.

Finalmente se estudiara la prueba de bondad de ajuste que permite ver si un conjunto de datos sigue una distribución conocida agrupando previamente los datos en categorías.

Tabla de contingencia

Una ***Tabla de contingencia*** con r filas y c columnas tiene la siguiente forma

		VAR B					
		B_1	B_2	B_3	…	B_c	Total
VAR A	A_1	O_{11}	O_{12}	O_{13}		O_{1c}	R_1
	A_2	O_{21}	O_{22}	O_{23}		O_{2c}	R_2
	A_3	O_{31}	O_{32}	O_{33}		O_{3c}	R_3
	…					…	
	A_r	O_{r1}	Or_2	O_{r3}	…	O_{rc}	R_r
	Total	C_1	C_2	C_3	…	C_c	n

Tablas de contingencia

O_{ij} es el número de sujetos que tienen las características A_i y B_j a la vez.
R_i (i = 1,…,r) es la suma de la *i-ésima* fila de la tabla. Es decir, es el total de sujetos que poseen la característica A_i.
C_j (j = 1,…,c) es la suma de la *j-ésima* columna de la tabla. Es decir, es el total de sujetos que poseen la característica B_j.
n representa el total de observaciones tomadas.
¿Existirá o no relación entre las variables A y B?, es decir, si A y B son o no independientes. A y B serán independientes si cada entrada de la tabla es igual al producto de los totales marginales dividido entre el número de datos. Esto es si cumple,

$$O_{ij} = \frac{R_i C_j}{n}$$

Ejemplo de una tabla 2x2

A: El estudiante graduando consigue trabajo, B: Sexo del graduando. Uno puede estar interesado en comparar la proporción de mujeres graduandas que consiguen trabajo con la proporción de varones graduandos que consiguen trabajo. Consideremos ahora la tabla:

	B1(Mujer)	B2 (Varon)	Total
A1(no)	10	6	16
A2(si)	5	16	21
Total	15	22	37

Notar que los valores de la segunda fila están en sentido contrario a los de la primera fila. O sea hay un efecto en la variable A al cambiar los valores de B, en consecuencia aquí sí hay relación entre las variables. La fórmula de independencia no se cumple para ninguna de las entradas. Por otro lado las proporciones de los valores de la variable A no son los mismos en cada columna. Por ejemplo, para A1 (no) las proporciones son 10/15 =.67 versus 6/22=.27.

Pruebas de Independencia y Homogeneidad

Cuando consideramos que los valores de nuestra tabla han sido extraídos de una población, entonces nos interesaría probar las siguientes dos hipótesis:

La **prueba de Independencia**, que se efectúa para probar si hay asociación entre las variables categóricas A y B, y

La **prueba de Homogeneidad**, que es una generalización de la prueba de igualdad de dos proporciones. En este caso se trata de probar si para cada nivel de la variable B, la proporción con respecto a cada nivel de la variable A es la misma.

Pruebas de Independencia y Homogeneidad

Las hipótesis de independencia son:
Ho: No hay asociación entre las variables A y B (hay independencia)
Ha: Sí hay relación entre las variables A y B

Las hipótesis de Homogeneidad son:
Ho: Las proporciones de cada valor de la variable A son iguales en cada columna.
Ha: Al menos una de las proporciones para cada valor de la variable A no son iguales en cada columna

Pruebas de Independencia y Homogeneidad

Ambas hipótesis se prueban usando una prueba de Ji-Cuadrado:

$$\sum_{i=}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

donde O_{ij} es la frecuencia observada de la celda que está en la fila i, columna j, $E_{ij}=\frac{R_i C_j}{n}$ es la frecuencia esperada de la celda (i, j).

La frecuencia esperada es aquella que debe ocurrir para que la hipótesis nula sea aceptada.

La prueba estadística se distribuye como una Ji-Cuadrado con $(r\text{-}1)(c\text{-}1)$ grados de libertad.

La hipótesis Nula se rechaza si $\chi^2_{cal} > \chi^2_{1-\alpha}$, donde es el nivel de significancia o equivalentemente si el "p-value" es menor que 0.05.

Tablas de Contingencia en Python

Se usa la función chi2_contingency del submódulo stats de Python

Hay dos situaciones.

La primera de ellas es cuando los datos están dados en dos columnas, o sea como si hubiesen sido las contestaciones a dos preguntas de un cuestionario.

La segunda de ellas es cuando los datos estan resumidos. En este caso hay que usar la funcion pivot_table antes de aplicar chi2_contingency

Ejemplo 8.1

Ejemplo 8.1. Usando los datos del ejemplo 3.16 (academic.uprm.edu/eacuna/eje316.txt), establecer si hay relación entre las variables tipo de escuela superior y el resultado (aprueba o no aprueba), de la primera clase de matemáticas que toma el estudiante en la universidad, basados en los resultados de 20 estudiantes.

Solución:

Para la prueba de Independencia las hipótesis son:

Ho: No hay relación entre el tipo de escuela y el resultado obtenido en la primera clase de Matemáticas.

Ha: Si hay relación entre ambas variables.

Para la prueba de homogeneidad las hipótesis son:

Ho: La proporción de aprobados en la primera clase de matemáticas es igual tanto para estudiantes que provienen de escuela pública como de escuela privada.

Ha: La proporción de aprobados en la primera clase de matemáticas no es la misma para ambos tipos de escuela.

Tabulated statistics: escuela, aprueba

Rows: escuela Columns: aprueba

	no	si	All
priv	3	7	10
	4	6	10
publ	5	5	10
	4	6	10
All	8	12	20
	8	12	20

Cell Contents: Count
Expected count

Pearson Chi-Square = 0.833, DF = 1, P-Value = 0.361
Likelihood Ratio Chi-Square = 0.840, DF = 1, P-Value = 0.359

* NOTE * 2 cells with expected counts less than 5

Interpretación: *Como el "P-value" es mayor que .05 se puede concluir que la hipótesis nula de Independencia entre las variables es aceptada. O sea no hay asociación entre el tipo de escuela de donde proviene el estudiante y el resultado que obtiene en la primera clase de matemáticas.*

Por otro lado, la hipótesis nula de homogeneidad también es aceptada y se concluye de que, la proporción de estudiantes que aprueban el curso de matemáticas es la misma para estudiantes de escuela pública y escuela privada.

Ejemplo 8.2. Usar los datos del ejemplo 3.17, para tratar de establecer si hay relación entre el Sexo del entrevistado y su opinión.

Sexo	Opinion	Conteo
male	si	10
male	no	20
male	abst	30
female	si	15
female	no	31
female	abst	44

Solución: Las hipótesis correspondientes son:
Ho: No hay asociación entre el sexo del entrevistado y su opinión, y
Ha: Si hay relación entre las variables.

Tabulated statistics: Sexo, Opinion

Using frequencies in Conteo

Rows: Sexo Columns: Opinion

	abst	no	si	All
female	44	31	15	90
	44.40	30.60	15.00	90.00
male	30	20	10	60
	29.60	20.40	10.00	60.00
All	74	51	25	150
	74.00	51.00	25.00	150.00

Cell Contents: Count
Expected count

Pearson Chi-Square = 0.022, DF = 2, P-Value = 0.989
Likelihood Ratio Chi-Square = 0.022, DF = 2, P-Value = 0.989

Interpretación: *Como el "P-value" es mayor que .05, la conclusión en este caso es que la hipótesis nula es aceptada o sea no hay relación entre el sexo y la opinión del entrevistado.*

Medidas de Asociación

Asumiendo que se rechaza la hipótesis Nula.
Ho: No hay relación entre las variables de la tabla,
entonces el próximo paso es determinar el grado de asociación de las dos variables categóricas, para ello se usan las llamadas medidas de asociación como:

El Coeficiente de Contingencia:

El Coeficiente de Cramer:

El Coeficiente de Contingencia:

Se define por $C = \sqrt{\frac{\chi^2}{n + \chi^2}}$

donde χ^2 es el valor calculado de la prueba de Ji-Cuadrado y
n es el número de datos.
El valor de C varía entre 0 y 1.
C = 0, significa que no hay asociación entre las variables.
C > .30, indica una buena asociación entre las variables.
Sin embargo hay que tomar en consideración también el tamaño de la tabla. Es mas adecuado para tablas con mas de 4 columnas y filas.

Desventajas

El coeficiente de contigencia no alcanza el valor de uno aún cuando las dos variables sean totalmente dependientes.
Su valor tiende a aumentar a medida que el tamaño de la tabla aumenta.

Ejemplo 8.4. Calcular el coeficiente de contingencia para la siguiente tabla, donde se trata de relacionar las variables: asistir a servicios religiosos y faltar a clases.

Rows: va a igl Columns: falta a clases

	de vez e	frecuent	nunca	All
de vez e	78	119	140	337
	75.56	103.44	158.01	337.00
frecuent	106	90	296	492
	110.31	151.01	230.68	492.00
nunca	68	136	91	295
	66.14	90.55	138.31	295.00
All	252	345	527	1124
	252.00	345.00	527.00	1124.00

Chi-Square = 86.842, DF = 4, P-Value = 0.000

coef-conting

0.267807

Interpretación:

No existe una buena asociación entre asistir a la iglesia y faltar a clases.

El Coeficiente de Cramer

Se calcula por $V = \sqrt{\frac{\chi^2}{nt}}$

donde *t* es el menor de los números *r*-1 y *c*-1, aquí *r* representa el número de filas y *c* el número de columnas.

V= 0 entonces, no hay asociación entre las variables.

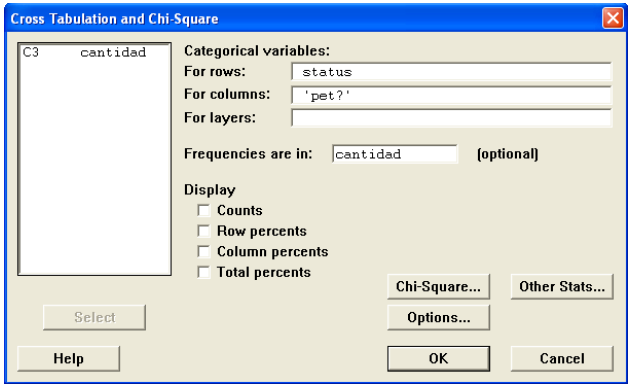
V >.30 indica un cierto grado de asociación entre las variables.

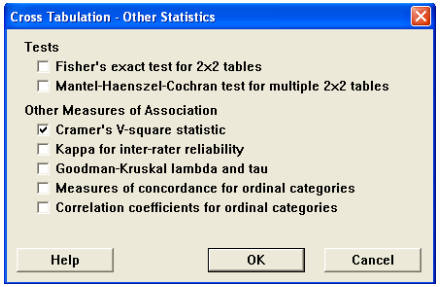
El coeficiente de Cramer si alcanza un máximo de 1.

En el ejemplo anterior el coeficiente de Cramer es la raiz cuadrada de 86.842/(1124)2 dando .1965, lo que reafirma que no existe buena asociación entre las variables.

Ejemplo 8.5

Calcular el coeficiente de Cramer para la siguiente tabla, donde se trata de relacionar las variables: sobrevivir a un ataque cardíaco y tener mascota ("pet").





Tabulated Statistics

Rows: status Columns: pet?

	no	si	All
muere	11	3	14
	5.93	8.07	14.00
vive	28	50	78
	33.07	44.93	78.00
All	39	53	92
	39.00	53.00	92.00

Chi-Square = 8.851, DF = 1, P-Value = 0.003

En este caso r =2 y c =2, luego t es el menor de

r-1=1 y c-1=1, asi t=1

Cramer's V-square 0.0962075

De donde: V = 0.310173

Interpretación: Se concluye que existe buena asociación entre tener mascota y sobrevivir a un ataque cardíaco.

Prueba de Bondad de Ajuste

Aquí se trata de probar si los datos de una muestra tomada siguen una cierta distribución predeterminada. Los *n* datos tomados deben estar divididos en categorías

Categoría	1	2	3	…	K	
Frecuencia observada	Obs_1	Obs_2	Obs_3		Obs_k	n

Se asume que las probabilidades p_{io}, de caer en la categoría *i* deben ser conocidas. Por ejemplo, la probabilidad de nacer en enero es 1/12, de nacer en febrero es 1/12 y asi sucesivamente hasta el mes de diciembre.

Prueba de Bondad de Ajuste

Las hipótesis a considerar son las siguientes:

Ho: $p_1 = p_{10}, p_2 = p_{20} = \ldots = p_k = p_{k0}$, es decir, los datos siguen la distribución deseada, y

Ha: al menos una de las p_i es distinta de la probabilidad dada p_{i0}.

La prueba estadística es: $\sum_{i=1}^{k} \frac{(Obs_i - np_{io})^2}{np_{io}}$

donde p_{io} representa la proporción deseada en la *i-ésima* categoría, Obs_i la frecuencia observada en la categoría i y n es el tamaño de la muestra.

La prueba estadística se distribuye como una Ji-Cuadrado con k-1 grados de libertad donde, k es el número de categorías.

Si el valor de la prueba estadística es mayor que $\chi^2_{1-\alpha}$ se rechaza la hipótesis nula.

Ejemplo 8.6

Los siguientes datos representan los nacimientos por mes en PR durante 1993 (http://academic.uprm.edu/eacuna/nacimientosPR.txt). Probar si hay igual probabilidad de nacimiento en cualquier mes del año.

Usar un nivel de significación del 5%.

5435 4830 5229 4932 5052 5072 5198 5712 6126 5972 5748 5936

Solución:

Ho: Hay igual probabilidad de nacer en cualquier mes del año ($p_1 = p_2 = \ldots = p_{12} = 1/12 = .083$).

H_a: En algunos meses hay mas probabilidad de nacer que en otros..

Para hacer la prueba de bondad de ajuste en Python se usa la función stats.chisquare

Los resultados se muestran en el siguiente slide

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Nacidos

Using category names in Mes

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
ene	5435	0.0833333	5436.83	0.0006
feb	4830	0.0833333	5436.83	67.7318
mar	5229	0.0833333	5436.83	7.9448
abr	4932	0.0833333	5436.83	46.8759
may	5052	0.0833333	5436.83	27.2395
jun	5072	0.0833333	5436.83	24.4818
jul	5198	0.0833333	5436.83	10.4917
ago	5712	0.0833333	5436.83	13.9266
sep	6126	0.0833333	5436.83	87.3580
oct	5972	0.0833333	5436.83	52.6783
nov	5748	0.0833333	5436.83	17.8090
dic	5936	0.0833333	5436.83	45.8295

N	DF	Chi-Sq	P-Value
65242	11	402.368	0.000

El valor de la prueba estadistica resulta ser 402.369 y el P-value es .000. Luego, se concluye que se rechaza la hipótesis nula, es decir que en algunos meses hay mayor probabilidad de nacimiento que en otros.

Ejemplo 8.7

Según el último censo se sabe que la distribución porcentual del estado marital de las personas adultas en los Estados Unidos es como sigue

De acuerdo al censo de 2000, en Puerto Rico se tiene la siguiente distribución de personas adultas por estado marital:

Se desea establecer si la distribución del estado marital en Puerto Rico, es igual a la de los Estados Unidos. Usar un nivel de significación del 5%.

Solución:

Ho: Los datos tomados en Puerto Rico siguen la misma distribución de la de Estados Unidos, y

Ha: Los datos no siguen la misma distribución.

Estados Unidos

Soltero	Casado	Viudo	Divorciado	Separado
27.1	54.4	6.6	9.7	2.2

Puerto Rico

Soltero	Casado	Viudo	Divorciado	Separado
211352	567694	16542	104206	31071

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: obs.PR

Using category names in Status

Category	Test Observed	Proportion	Expected	Contribution to Chi-Sq
Soltero	811291	0.30	743645	6153
Casado	1279628	0.40	991527	83711
Viudo	198553	0.12	297458	32886
Divorciado	189346	0.18	446187	147847

N	DF	Chi-Sq	P-Value
2478818	3	270598	0.000

Interpretación: Al ser el P-vaue=.0000 , se rechaza la hipótesis nula y se concluye que la distribución del estado marital en Puerto Rico es distinta a la de Estados Unidos.