ANÁLISIS DE DATOS NO NUMERICOS

Jorge Galbiati Riesco

ESCALAS DE MEDIDA CATEGORICAS

Los datos categóricos son datos que provienen de resultados de experimentos en que sus resultados se miden en escalas categóricas. Medir en una escala categórica consiste en observar el resultado de un experimento y asignarle una clase o categoría, de entre un número finito de clases posibles. Esta escala es no numérica, y puede ser categórica ordinal, es decir, sus categorías tienen un orden natural, o en caso contrario la escala es categórica nominal.

EJEMPLO 1. Una encuesta reveló las opiniones de un grupo de personas respecto de mayores limitaciones en la venta de armas de fuego. Entre otras, se midieron las siguientes dos variables: Una es el grado de restricción esperado por el encuestado, en la venta de armas de fuego. La otra es el nivel educacional del encuestado.

La variable grado de restricción esperado tiene los valores:

a) Mucho más restringida,
b) moderadamente más restringida,
c) tal como está,
d) moderadamente menos restringida,
e) mucho menos restringida.

La variable nivel educacional tiene los valores:

a) Hasta 4º Básico,
b) básico completo,
c) hasta segundo medio,
d) medio completo,
e) estudios universitarios incompletos,
f) titulado universitario.

El tipo de análisis que se suele hacer con datos categóricos consiste en determinar el tipo de asociación existente entre pares de variables, lo que se denomina cruzar las variables. Las

posibilidades son, que no haya asociación alguna, en tal caso se dice que las variables son independientes, o que haya diferentes grados de asociación.

En el caso de independencia entre dos variables, el valor que tome una de ellas no predispone el valor de la otra. En el Ejemplo 1, podría ser que el grado de restricción esperado sea independiente de la otra variable, nivel educacional. Es decir, cualquiera sea su nivel educacional, la probabilidad de que opine que la venta de armas debiera ser mucho más restringida, es la misma. Lo mismo ocurrirá con las otras categorías.

Si dos variables no son independientes, están asociadas, y el grado de asociación no es único. Puede haber diversos grados de asociación. Si hay asociación, quiere decir que algunos valores de una de las variables predispones a que la otra variable tome ciertos valores de la otra variable, más que otros. Esta predisposición es mayor cuanto mayor es el grado de asociación.

EJEMPLO 2. Se hizo un estudio de niños de 10 a 12 años, consistente en experimentar la efectividad de dos métodos de higiene bucal en la prevención de caries, el método A y el método B. Después de un año, se observó el desarrollo de caries. El resultado observado se clasificó en tres categorías: Bajo, moderado, alto.

Si los niños con el tratamiento A tienden a tener desarrollo de caries moderados o altos, mientras que los niños con tratamiento B tienden a tener bajo desarrollo de caries, entonces hay un cierto grado de asociación.

En resumen, la independencia entre dos variables es total, no tiene grados de intensidad. Si no hay independencia, las variables tienen asociación, que puede tener diversos grados.

La independencia es una propiedad simétrica, en el sentido de que si una variable es independiente de una segunda variable, la segunda es independiente de la primera. Lo mismo ocurre con cualquier grado de asociación.

TABLAS DE CONTINGENCIA

El primer paso en el cruce de dos variables categóricas, para medir el grado de asociación entre ellas, es construir una tabla de contingencia, que consta de un cuadro en que una de las variables se representa en el lado izquierdo, y la otra en la parte superior, cada una con todos sus respectivos valores. El cuadro contiene, en cada casilla, el conteo del número de casos en cada una de las combinaciones de valores de ambas variables.

Además, se muestran los totales verticales (por columnas) en la parte inferior, y los totales horizontales (por filas), en el lado derecho. Estas, por aparecer en los márgenes, se denominan frecuencias marginales.

En el extremo inferior derecho, se muestra el total de casos, N, que corresponde a la suma de las frecuencias marginales fila, o a la suma de las frecuencias columna, que son iguales.

EJEMPLO 3. Supóngase que en el Ejemplo 2, participaron en el estudio un total de 200 niños. Con los resultados obtenidos, se construyó la siguiente tabla de contingencia:

TABLA DE FRECUENCIAS OBSERVADAS		Desarrollo de caries			
		Bajo	Moderado	Alto	Totales
Tratamiento	A	8	40	34	82
	B	84	22	12	118
	Totales	92	62	46	200

Podemos observar que al tratamiento A le corresponden más casos con desarrollo de caries moderado y alto, mientras que al tratamiento B le corresponden más casos de bajo desarrollo de caries.

En la tabla se muestran las sumas por columnas, que son las frecuencias distintos grados de desarrollo de caries, y los totales por filas, que son las frecuencias de niños con cada uno de los dos tipos de tratamientos. La suma de los totales, tanto por fila como por columna, son iguales a 200, el total de casos.

EL ESTADÍSTICO JI-CUADRADO COMO MEDIDA DE ASOCIACIÓN DE VARIABLES

Como primer paso en el cálculo de una medida del grado de asociación entre las dos variables, se debe construir una tabla de frecuencias esperadas, que es una tabla que muestra las frecuencias que habrían si fuera cierto que ambas variables son independientes. En tal caso, la proporciones en las casillas de todas las filas (o columnas) son proporcionales. En contraste con la tabla de contingencia, que también toma el nombre de tabla de frecuencias observadas.

La tabla de frecuencias esperadas se construye de la siguiente forma; la frecuencia esperada e_{ij} de la casilla correspondiente a la fila i y a la columna j, está dada por la fórmula

$$e_{ij} = \frac{(frecuencia\ marginal\ \ fila\ i) \times (frecuencia\ marginal\ columna\ \ j)}{Total\ \ de\ observaciones}$$

Si calculamos las frecuencias marginales de la tabla de frecuencias esperadas, sumando las filas y las columnas, se podrá observar que son iguales a las frecuencias marginales de la tabla de frecuencias observadas.

Si ambas variables son independientes, la tablas de frecuencias esperadas y observadas serán parecidas. Si difieren, entonces hay asociación entre la variable fila y la variable columna. Mientras más difieren las dos tablas, mayor será el grado de asociación entre las variables.

EJEMPLO 4. Se calculará la tabla de frecuencias esperadas, a partir de la tabla de frecuencias observadas del Ejemplo 3, sobre el estudio de prevención de caries.

TABLA DE FRECUENCIAS OBSERVADAS		Desarrollo de caries			
		Bajo	Medio	Alto	Totales
Tratamiento	A	8	40	34	82
	B	84	22	12	118
	Totales	92	62	46	200

Esta tabla se construye multiplicando las frecuencias de la fila y la columna respectiva, y dividiendo por el total. De esta forma, la frecuencia esperada correspondiente al tratamiento A y al desarrollo de caries bajo, es igual a 92x82/200 = 37.72. Así se construye toda la tabla, que da los siguientes valores, redondeados a un decimal:

TABLA DE FRECUENCIAS ESPERADAS		Desarrollo de caries			
		Bajo	Medio	Alto	Totales
Tratamiento	A	37.7	25.4	18.9	82
	B	54.3	36.6	27.1	118
	Totales	92	62	46	200

Si comparamos esta tabla con la de valores observados, del Ejemplo 3, vemos que son muy diferentes, lo que indica que no hay independencia, sino que hay asociación entre las variables.

EJEMPLO 5. Supóngase que se aplicó la encuesta del Ejemplo 1, acerca de armas de fuego, a una muestra de 1000 personas elegidas al azar y con los datos obtenidos se construyó una tabla de contingencia. Recordar que las variables de interés, y que se van a cruzar son:

Grado de restricción esperado con los valores: (a) Mucho más restringida, (b) moderadamente más restringida, (c) tal como está, (d) moderadamente menos restringida, y (e) mucho menos restringida,

nivel educacional con los valores: (a) hasta 4º Básico, (b) básico completo, (c) hasta segundo medio, (d) medio completo, (e) estudios universitarios incompletos, y (f) titulado universitario.

Supóngase que la tabla de contingencia es la siguiente, con los totales por fila y por columna (frecuencias marginales) es

TABLA DE FRECUENCIAS OBSERVADAS		Nivel educacional						
		a	b	c	d	e	f	Totales
Grado de restricción deseado	a	6	5	7	23	40	22	103
	b	7	16	22	36	73	37	191
	c	20	40	49	74	118	73	373
	d	18	15	22	44	82	56	237
	e	10	7	10	25	21	23	96
	Totales	61	83	109	202	334	211	1000

La tabla de frecuencias esperadas se construye multiplicando las frecuencias de la fila y la columnas respectivas, y dividiendo por el total. Así la frecuencia esperada correspondiente a grado de restricción (a) mucho más restringida, y nivel educacional (a) hasta 4º Básico, es igual a 61x103/1000 = 6.30. De esta forma cubrimos toda la tabla, que da los siguientes valores, redondeados a un decimal:

TABLA DE FRECUENCIAS ESPERADAS		Nivel educacional						
		a	b	c	d	e	f	Totales
Grado de restricción esperado	a	6.3	8.5	11.2	20.8	34.4	21.7	103
	b	11.7	15.9	20.8	38.6	63.8	40.3	191
	c	22.8	31.0	40.7	75.3	124.6	78.7	373
	d	14.5	19.7	25.8	47.9	79.2	50.0	237
	e	5.9	8.0	10.5	19.4	32.1	20.3	96
	Totales	61	83	109	202	334	211	1000

A simple vista no es posible determinar si las tablas se parecen, en tal caso las variables serían independientes, o si difieren, y habría asociación entre ellas.

Entonces lo que falta es una medida que refleje el grado en que difieren estas dos tablas, que será una medida del grado de asociación. Esta medida es la denominada estadístico ji-cuadrado, en símbolos, χ^2, que se define como sigue

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad (1)$$

en que o_{ij} es la frecuencia observada de la casilla i,j (fila i , columna j), e_{ij} es la frecuencia esperada de la casilla i,j. Las sumas se extienden a través de todas las filas y columnas.

Se puede ver que este indicador es cero si ambas tablas son idénticas, es decir, si hay independencia entre las dos variables.

En la medida que difieran ambas tablas, será más grande el estadístico ji-cuadrado, lo que será indicación de que hay más asociación entre las variables.

Hay otra fórmula que da el mismo resultado, pero que es más fácil de calcular:

$$\chi^2 = \sum_i \sum_j \frac{o_{ij}^{\ 2}}{e_{ij}} - N \qquad (2)$$

donde N es el total de observaciones.

EJEMPLO 6. Calculamos el estadístico ji-cuadrado para las tablas de frecuencias observada y esperada del estudio de prevención de caries, del Ejemplo 4. Usaremos la fórmula (2), de modo que

$\chi^2 = 8^2/37.7 + 40^2/25.4 + 34^2/18.9 + 84^2/54.3 + 22^2/36.3 + 12^2/27.1 - 200 = 74.46$

No es muy claro si este es un valor pequeño o grande.

EJEMPLO 7. Ahora calcularemos el valor del estadística ji-cuadrado para las tablas del Ejemplo 5, donde se cruzan las variables grado de restricción esperado en la venta de armas, y nivel educacional. también usando la fórmula (2). Entonces

$\chi^2 = 6^2/6.3 + 5^2/8.5 + 7^2/11.2 + ... + 23^2/20.3 - 1000 = 25.02$

Este valor es mucho más pequeño que el valor del estadístico ji-cuadrado del Ejemplo 6, pero tampoco podemos decir si es pequeño o grande.

La limitación que tiene el estadístico ji-cuadrado como medida de asociación, es que está acotado inferiormente por cero, pero no tiene cota superior. Por lo tanto es difícil evaluar si su valor es grande o pequeño.

Se definen otros índices, en base al estadístico ji-cuadrado.

PRUEBAS DE HIPÓTESIS DE INDEPENDENCIA

Con el estadístico ji-cuadrado se pueden efectuar pruebas de hipótesis para confirmar si hay asociación entre las dos variables que se están cruzando. Esta prueba se denomina prueba ji-cuadrado.

Las hipótesis que se van a poner a prueba son:

H_0: Hay independencia entre las variables.
H_1: No hay independencia.

Para hacer la prueba, se debe comparar el estadístico con el valor obtenido de la Tabla Ji-cuadrado correspondiente. Para obtener el valor de tabla, se calcula el parámetro grados de libertad, que es el producto (número de filas –1) * (número de columnas –1)
Este valor se busca en la fila correspondiente de la tabla ji-cuadrado, que se muestra más adelante.

Si el estadístico ji-cuadrado es mayor que el valor de la tabla, se rechaza la hipótesis H_0, y por lo tanto, se concluye que no hay independencia entre las dos variables. Si no es mayor,

se asume que no hay evidencia muestral para rechazar esa hipótesis, por lo tanto se asume que si hay independencia entre las variables.

Siempre que se hace una prueba de hipótesis, es posible rechazar erróneamente la hipótesis de inedependencia, siendo que es verdadera. Se puede cuantificar la probabilidad de cometer este tipo de error. Esta probabilidad se denomina nivel de significación de la prueba. No es posible eliminar la probabilidad de este error, pero se espera que sea pequeña.

La tabla siguiente corresponde a un nivel de significación de 0.05 (probabilidad de rechazar erróneamente la hipótesis H_0). Hay tablas más completas, que entregan otras probabilidades de rechazar H_0 erróneamente, sin embargo, el valor mayormente aceptado es 0.05 o 5%.

TABLA JI-CUADRADO

Grados de libertad	Ji-cuadrado	Grados de libertad	Ji-cuadrado	Grados de Libertad	Ji-cuadrado
1	3.841	11	19.68	21	32.67
2	5.992	12	21.03	22	33.92
3	7.815	13	22.36	23	35.17
4	9.488	14	23.68	24	36.42
5	11.07	15	25.00	25	37.65
6	12.59	16	26.30	26	38.89
7	14.07	17	27.59	27	40.11
8	15.51	18	28.87	28	41.34
9	16.92	19	30.14	29	42.56
10	18.31	20	31.41	30	43.77

Una precaución que se debe tomar con las pruebas ji-cuadrado es que frecuencia esperada en cada casilla sea de a lo menos 5. En caso contrario, el estadístico ji-cuadrado se estará distorsionado, y el nivel de significación no será el correcto.

EJEMPLO 8. En el caso del desarrollo de caries, Ejemplo 6, los grados de libertad son 1*2 = 2.
La tabla nos da el valor 5.992. Vemos que el valor del estadístico ji-cuadrado de 74.46, más grande con el valor de tabla, por lo tanto rechazamos la hipótesis de independencia, y

concluimos que hay asociación entre ambas variables, el tipo de tratamiento y el grado de desarrollo de caries.

En el caso del grado de restricción de armas esperado y el nivel educacional, el estadístico ji-cuadrado es de 25.02, con 4*5 = 20 grados de libertad. La tabla nos entrega un valor de

31.41, por lo tanto no se rechaza la hipótesis de independencia. Se acepta que las variables grado de restricción esperado en la venta de armas, y nivel educacional, son independientes. Una de las variables no es condicionante de la otra.

OTRAS MEDIDAS DE ASOCIACIÓN

Como alternativa a efectuar una prueba ji-cuadrado, se puede simplemente cuantificar el grado de asociación, utilizando alguna medida de asociación adecuada. O puede servir como complemento a la prueba, que sólo concluye si hay o no asociación, pero no dice cuánta asociación.

Se dispone de tres medidas, todas basadas en el estadístico ji-cuadrado.

La primera medida de asociación es el coeficiente ϕ, definido como

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad (3)$$

en que N es el total de observaciones. El coeficiente ϕ es mayor que 0, y aunque es mucho menor que el estadístico ji-cuadrado, no está acotado superiormente. Puede ser mayor que uno.

Otra medida de asociación es el coeficiente de contingencia, que se define como

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \qquad (4)$$

Este coeficiente toma valores entre 0 y 1, sin embargo nunca alcanza el valor 1. Su máximo depende del número de filas y columnas. Por ejemplo, en tablas de 4 filas por 4 columnas, su valor máximo es de 0.87.

Por último, está el coeficiente V de Cramer,

$$V = \sqrt{\frac{\chi^2}{N \times (k-1)}} \qquad (5)$$

en que k es el mínimo entre el número de filas y el número de columnas.

Este coeficiente está acotado entre 0 y 1, y puede alcanzar ambas cotas, por lo tanto es el mejor de las medidas de asociación, por ser más fácil de interpretar.

Si hay dos filas o dos columnas, los coeficientes ϕ y V de Cramer son iguales.

EJEMPLO 7. Calcularemos los tres índices para los datos del estudio de prevención de caries en niños, a partir del estadístico ji-cuadrado calculado en el Ejemplo 6:

Coeficiente ϕ

$$\phi = \sqrt{\frac{74.46}{200}} = 0.610$$

Coeficiente de contingencia

$$C = \sqrt{\frac{74.46}{74.46 + 200}} = 0.521$$

Coeficiente V de Cramer

$$V = \sqrt{\frac{74.46}{200 \times (2-1)}} = 0.612$$

En este caso hay dos filas, por eso coinciden los coeficientes ϕ y V de Cramer. Recordar que este último toma valores entre 0 y 1, por lo tanto el valor 0.612 se ve suficientemente grande como para concluir que hay asociación entre las variables tratamiento y desarrollo de caries. Recordar que la prueba de hipótesis dio como resultado que no hay independencia entre las variables.

EJEMPLO 8. Ahora calculamos los tres índices para los datos en que se cruzan las variables grado de restricción esperado en la venta de armas, y nivel educacional, a partir del estadístico ji-cuadrado calculado en el ejemplo 7:

El coeficiente ϕ es $\phi = \sqrt{\frac{25.02}{1000}} = 0.1580$

El coeficiente de contingencia es $C = \sqrt{\frac{25.02}{25.02 + 1000}} = 0.156$

Y el coeficiente V de Cramer es V = $\sqrt{\frac{25.02}{1000 \times (5-1)}}$ = 0.079

Ahora los valores son bastante bajos, en particular el coeficiente V de Cramer. Por lo tanto se puede concluir que no hay, o hay muy poca asociación entre las variables grado de restricción en la venta de armas, esperado por el encuestado, y su nivel educacional. Esto es consistente con el resultado de la prueba de hipótesis, que dio independencia entre as variables.

---

1) Se aplica una encuesta a un grupo de 80 personas. La preguntas 10 y la pregunta 17 tienen dos posibles respuestas, "si" y "no". Se construye una tabla de doble entrada para cruzar estas dos preguntas, y dió el siguiente resultado:

		Pregunta 17	
		si	no
Pregunta 10	si	3	41
	no	34	2

Es decir, 3 personas respondieron "si" a la pregunta 10 y "si" a la pregunta 17, etc. Más allá de los resultados numéricos, ¿que se puede decir, en general, respecto de la relación entre estas dos preguntas?

2) Se aplica una encuesta a un grupo de 50 personas. Se construyó una tabla de contingencia para cruzar las preguntas 17 y 22. Luego se calculó el estadístico ji-cuadrado, y en base a él se calculó el coeficiente V de Cramer, que resultó ser igual 0.92. ¿Que se puede decir respecto de la relación entre las preguntas 17 y 22?

3) Se aplica una encuesta a un grupo de 67 personas. La preguntas 4 y la pregunta 8 tienen dos posibles respuestas, "si" y "no". Se construye una tabla de doble entrada para cruzar estas dos preguntas, y dio el siguiente resultado:

		Pregunta 8	
		si	no
Pregunta 4	si	24	16
	no	18	9

Es decir, 24 personas respondieron "si" a la pregunta 4 y "si" a la pregunta 8, etc. Más allá de los resultados numéricos, ¿que se puede decir, en general, respecto de la relación entre estas dos preguntas?