Divulgaciones

Notas Estadísticas 3: Más medidas de asociación

RECIBIDO: 29/6/06
ACEPTADO: 10/10/06

Vaira, S. • Carrera, E. de • Contini, L.

Departamento de Matemática, Área Estadística. Facultad de Bioquímica y Ciencias Biológicas. Universidad Nacional del Litoral. CC. 242. Ciudad Universitaria, Paraje El Pozo. (3000) Santa Fe.

RESUMEN: Es muy común en las Ciencias Experimentales y Sociales trabajar con variables que sólo expresan cualidad o atributo, son las llamadas categóricas o de atributo. Éstas frecuentemente son relacionadas de a pares, en forma de tabla o gráfico que permiten visualizar la dependencia o no entre ellas. Al momento de establecer el grado de relación existente entre dos variables categóricas no es suficiente observar las frecuencias, ya sean absolutas, relativas o porcentuales, de una tabla de contingencia, porque no establecen la fuerza de asociación entre ellas, razón por la cual es necesario utilizar medidas que determinen la fuerza de la relación, acompañadas de su nivel de significancia.

PALABRAS CLAVE: tablas de contingencia, estadístico χ^2 (Chi-cuadrado), medidas de asociación

SUMMARY: *Statistical notes 3: More association measurements.*
Vaira S, Carrera E, Contini L.

When dealing with Experimental or Social Sciences, it is usual to work with variables that express only quality or attributes. These are called category or attribute variables. They are usually viewed as pairs, shown in tables or graphs that allow dependency between them to be made overt. When trying to establish the degree to which two category variables are related, noticing frequencies (absolute, relative or percentual) from a table of contingencies is not enough, because the degree of association between the variables cannot be established. For this reason, it is necessary to use measurements that could reveal the strength of their relationship, together with their significance.

KEY WORDS: Contingency table, χ^2 statistic, measures of association

Introducción

La medida de asociación entre variables continuas más usada es el coeficiente de correlación lineal de Pearson (ρ), cuyo cálculo, usos e interpretaciones del mismo se realizó en Notas Estadísticas 2: "El coeficiente de correlación lineal" (1).

Es muy común en las Ciencias Experimentales y Sociales trabajar con variables que sólo expresan cualidades, pertenecen a ellas las llamadas variables categóricas o de atributo, que también se subdividen en dos grandes grupos: nominales y ordinales.

Son ejemplos de variables nominales las mediciones de la sensibilidad de una bacteria a un cierto antibiótico, cuya respuesta es si o no; la clasificación en los distintos grupos sanguíneos; en accidentología es común referenciar la parte del cuerpo donde se ha producido el daño: cabeza, brazos, piernas, cuello, columna vertebral y cuántas divisiones de ella quiera hacer el investigador. La clasificación de un ser humano en fumador o no fumador según el criterio de la OMS (Organización Mundial de la Salud), es también una variable categórica y nominal. Otros ejemplos son: procedencia, lugar de nacimiento, raza, género o el tipo de tratamiento aplicado a una enfermedad.

En las primeras investigaciones sólo se realizaban algunas tablas de frecuencias o gráficos, pero en la actualidad se requiere de técnicas estadísticas un poco más complejas. Prueba de su importancia actual es que la estadística ha desarrollado una serie de test de hipótesis y medidas de asociación que involucra el análisis de este tipo de variables.

También las variables ordinales son muy utilizadas, variables que sin ser cuantitativas establecen un orden en sus categorías. Son algunos ejemplos de este tipo de variables los siguientes: los estadios del desarrollo de un tumor en grado I, grado II y grado III; los niveles educativos alcanzados por un individuo en primario, secundario, universitario y cuarto nivel, el grado de avance en una carrera universitaria, entre otros. La estadística también contempla para este tipo de variables tanto pruebas de hipótesis como medidas de asociación.

Además, se pueden relacionar variables nominales y ordinales a través de estadísticos desarrollados para tal fin.

Tablas de contingencia para variables categóricas

Al trabajar con variables categóricas, en particular al momento de relacionar las variables de a pares, se suelen agrupar los datos en tablas de doble entrada, conocidas como tablas de contingencia, donde cada entrada representa a una variable o criterio de clasificación. Como resultado se tiene una tabla de distribución conjunta de frecuencias, donde cada casillero representa la cantidad de objetos o individuos que cumplen con ambos criterios. La Tabla 1 muestra un ejemplo de la relación entre dos variables: "Clasificación de pacientes en dos grupos, uno con artritis reumatoidea de aparición tardía (ARAT) y el otro con polimialgia reumática PMR" y "Valores de anticuerpos de citrulina altos (mayores que 20) y bajos".

En ella 14 pacientes tienen ARAT y valores de CCP elevados, esto representa el 28% de la muestra y 5 de los 50 pacientes (10%) tienen PMR y valores altos de CCP. También en forma descriptiva se puede decir que 14 de 24 pacientes con ARAT tienen valores altos de citrulina. Se suele acompañar a esta tabla con un gráfico de barras comparativas como se muestra en el Figura 1.

Por supuesto que al tener varias variables categóricas en lugar de relacionar dos de ellas y tener dos criterios de clasificación para organizar la tabla de contingencia se pueden tener más, pero esto complica la lectura y análi-

sis posterior de los datos, pues el problema en vez de ser bivariado es multivariado, y en algunos casos las tablas de contingencia son segmentadas por los niveles de una tercer variable que se quiere relacionar con las dos primeras, pero este problema será dejado para otro estudio (2 - 4).

Prueba de independencia: estadístico χ^2 (Chi-cuadrado)

El grado de relación existente entre dos variables categóricas no puede ser establecido simplemente observando las frecuencias de una tabla de contingencia, aunque en lugar de las frecuencias absolutas uno identifique el porcentaje de participación de cada celda de la tabla, esos porcentajes tampoco son indicadores de la fuerza de asociación entre las variables involucradas (2, 3). Para determinar esta fuerza de asociación se comienza con una prueba de independencia y luego se debe elegir una medida de asociación de acuerdo a la naturaleza de las variables involucradas, además de tener en cuenta el tipo de diseño estadístico que se utilizó, ya que es importante saber si las muestras son o no pareadas. Todas estas características definen el estadístico a utilizar y luego debe incluirse el nivel de significancia con el que se desea trabajar.

Para probar la hipótesis de independencia entre dos variables categóricas, cuya información se reúne en una tabla de contingencia, se calcula el valor del estadístico X^2 (equis al cuadrado) de la siguiente manera:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \qquad (A)$$

donde e_{ij} es la frecuencia esperada, n_{ij} es la frecuencia observada en cada celda y

$$e_{ij} = \frac{(\text{total de la fila i}) x (\text{total de la columna j})}{\text{número total de casos (n)}}$$

Este estadístico sigue el modelo de distribución de probabilidad x^2 chi – cuadrada con υ = (número de filas-1)x(número de columnas-1) grados de libertad, la distribución fue propuesto por Karl Pearson en 1911 (5), el mismo que desarrolló el coeficiente de correlación lineal para variables continuas que lleva su nombre en el análisis de regresión y correlación lineal. Para valores *"grandes de este estadístico* X^2" se rechaza la hipótesis de independencia concluyendo que las variables están relacionadas (2 - 4) y si "X^2 *se acerca a cero*" se concluye que las variables son independientes.

Para que las probabilidades de la distribución χ^2 constituyan una buena aproximación del estadístico χ^2 es necesario ver que se cumplan algunas condiciones; entre ellas que las frecuencias esperadas no sean demasiado pequeñas, ya que debe interpretarse con mucha cautela la significancia del estadístico (3-4, 6).

Además de este estadístico chi-cuadrado habitualmente se muestra el estadístico denominado razón de verosimilitud (Fisher, 1924; Neyman y Pearson, 1928) (7) que se obtiene mediante la llamada:

$$\textit{Razón de verosimilitud} = 2\sum_i \sum_j n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

donde ln es el logaritmo natural.

Se trata de un estadístico asintóticamente equivalente a X^2 y es muy utilizado para estudiar la relación entre variables categóricas, particularmente en el contexto de los modelos log-lineales (2, 6, 8).

Si la tabla de contingencia se construye con dos variables dicotómicas (tablas 2x2), hay dos estadísticos asociados a ellas:

- el *Chi-cuadrado de Pearson con corrección de continuidad deYates* (1934) y
- el *estadístico exacto de Fisher* (1935) (5, 7).

El primero de ellos consiste en restarle la cantidad 0,5 al numerador de X^2 dado en (A), para muestras pequeñas este estadístico se

ajusta mejor a la distribución χ^2, pero no existe un concenso generalizado de que sea así (4, 8-9). El segundo permite obtener significancias exactas en lugar de asintóticas, muy utilizado para muestras pequeñas.

Medidas de asociación para variables nominales

El estadístico X^2, junto a su distribución aproximada a una Chi- cuadrada permite contrastar la hipótesis de independencia entre dos variables estudiadas, pero no nos dice nada acerca de la fuerza de la asociación lineal entre ellas. El valor del estadístico aumenta, en general, cuando aumenta el tamaño de la muestra. Si se multiplica, por ejemplo, todas las casillas de una tabla por 10 el estadístico quedará también multiplicado por 10. Por esta razón, para estudiar el grado de relación existente entre dos variables se utilizan medidas de asociación que intentan cuantificar ese grado de relación eliminando el efecto del tamaño muestral.

Un forma de resolver este problema y poder comparar tablas de contingencia con distinto tamaño muestral es dividir el valor de X^2 por el tamaño muestral, se genera así el coeficiente de asociación:

$$CA = \frac{X^2}{n}$$

donde n es el tamaño de la muestra. Un inconveniente de este coeficiente de asociación es que puede tomar cualquier valor. Si se quiere un coeficiente con la propiedad de estar en el intervalo [0; 1] se debe utilizar el *coeficiente de contingencia C* definido como:

$$C = \sqrt{\frac{X^2}{X^2 + n}}$$

Este estadístico tomará el valor cero si hay perfecta independencia entre las variables, de lo contrario se acercará al valor 1.

El *coeficiente phi* (ϕ) se obtiene de la siguiente manera:

$$\phi = \sqrt{\frac{X^2}{n}}$$

se utiliza en especial en tablas 2x2 y varía entre 0 y 1 como el coeficiente de contingencia C (8-11).

Medidas de asociación para variables ordinales

La naturaleza de la variables nominales hace que en la construcción de las medidas de asociación no se pueda hablar de la dirección de la relación, a diferencia de las correspondientes a variables ordinales que sí tendrá sentido de hablar de dirección de la relación. Por lo tanto en datos ordinales se podrá decir que una relación es positiva si a valores bajos de una variable se asocia con los valores bajos de la segunda variable y altos con altos, de manera similar si la relación es negativa.

Al recordar que el coeficiente de correlación de Pearson (ρ) está fuertemente influenciado por outliers, desigualdad de variancias, no normalidad de los datos y no linealidad; su competidor más fuertemente usado es el coeficiente de correlación de Spearman (ρ_s), que en lugar de relacionar dos variables (X e Y) a partir de las mediciones originales lo hace utlizando los rangos (ranks), es decir, la posición que ocupan las observaciones y no los valores observados. Si en particular X e Y son de naturaleza ordinal es el coeficiente adecuado para medir la fuerza de la asociación entre ellas. Además se sabe que esta medida varía en el intervalo [-1, 1], claramente da una respuesta a la fuerza de la asociación entre variables de naturaleza ordinal y el signo, la dirección de la relación (11-14).

Otra medida muy utilizada es la Gamma (γ) de Goodman y Kruskal, este estadístico no se basa en la comparación directa entre las frecuencias observadas y esperadas de las celdas, sino que se construye teniendo en cuenta si caen

mayores frecuencias en la "diagonal" o fuera de ésta en una tabla de contingencia (13, 15).

Conclusiones

Con este artículo se trata de mostrar que hay diferentes medidas de asociación e independencia que se utilizan para relacionar variables categóricas. Cada una de ellas tiene en cuenta características relativas al diseño del experimento, tamaño de la muestra, naturaleza de las variables observadas que deben ser tenidas en cuenta al momento de seleccionarlas.

Ante un mismo problema puede haber más de una medida para inferir acerca de la fuerza de la asociación entre dos variables.

En esta presentación no se han mencionado todas las medidas de asociación que se pueden utilizar para diferentes variables quedando por ejemplo, τ (tau) de Kendall, coeficiente κ (kappa) de concordancia, índices de riesgo, coeficiente de incertidumbre, entre otros fuera de la discusión.

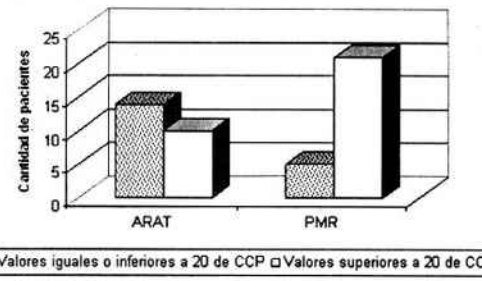
Agradecimientos

Agradecemos al Doctor Sergio Paira y a su grupo de trabajo por habernos permitido emplear los datos de sus investigaciones en enfermedades reumáticas como ejemplo para este artículo.

Tabla 1: Tabla de contingencia o distribución de frecuencias conjuntas

		Enfermedad		
		ARAT	PMR	Total
CCP	Alta	14	5	19
	Baja	10	21	31
Total		24	26	n=50

Figura 1: Diagrama de barras comparativas entre dos grupos de pacientes.



Bibliografía

1. Carrera, E.; Vaira, S.; Contini, L. 2004. Notas Estadísticas 2: El coeficiente de correlación lineal. Revista FABICIB. **8:** 255-259.

2. Fleiss J. L. 2003. "Statistical Methods for rates and proportions". 3ra ed. John Wiley & Sons. (New York).

3. Altman D. G. 1997. "Practical statistics for medical research". Chapman & Hall. (London).

4. Ten Have, T.; Becker, M. 1995. Multivariate Contingency tables and the analysis of exchangeability. Biometrics. **51** (3): 1001-1016.

5. Hacking, I. 1995. "El surgimiento de la probabilidad". Editorial Gedisa. Barcelona. (Traducción de la obra: The emergente of probability. Cambridge University Press. 1975).

6. Armitage P., Berry G. 1999. "Estadística para la investigación biomédica". Harcourt Brace. (Barcelona).

7. Fisher, R. 1976. "Sigma. El mundo de las matemáticas". James Newman (editor). Tomo 3. Ediciones Grijalbo. (Barcelona).

8. Conover, W. 1998. "Practical nonparametric statistic" John Willey & Sons. (New York). 143-330.

9. Agresti, A. 1996. "An introduction to categorical data analysis". John Willey & Sons. (New York).

10. SPSS Base Apllications Guide and Advanced Models™, 1999. SPSS Inc. version 10,0 (Chicago).

11. Dawson-Saunders B, Trapp, R. G. 1996. "Bioestadística Médica". 2ª ed. Editorial el Manual Moderno. (México).

12. Chen, M.; Kianifard, F. 1999. Application of Goodman-Kruskal's Gamma for ordinal data, in comparing Several ordered treatments: A different approach. Biometrical Journal. **41:** 491-498.

13. Molinero, L. 2004. Asociación de variables cualitativas nominales y ordinales. www.she-lelha.org/stat1.htm

14. Betensky, R. A. and Rabinowitz, D. 1999. Maximally selected _2 statistics for k×2 tables. Biometrics. **55:** 317-320.

15. Miller, R. and Siegmund, D. (1982). Maximally selected Chi square statistics. Biometrics **48:** 1011-1016.