

**FORMACIÓN MÉDICA****Como un reloj olvidado. Descripción de variables cualitativas.***Molina Arias M.**Hospital Infantil Universitario La Paz.***Resumen**

El análisis de las variables cualitativas suele limitarse a escasos índices como sus frecuencias relativas y acumuladas o la moda. Sin embargo, existen índices diseñados para estudiar la variabilidad de este tipo de variables, a semejanza de las variables cuantitativas, que suelen recibir más atención. Se describen algunos de estos índices que miden la variabilidad de las variables cualitativas.

Introducción

El análisis de las variables cualitativas suele limitarse a escasos índices como sus frecuencias relativas y acumuladas o la moda. Sin embargo, existen índices diseñados para estudiar la variabilidad de este tipo de variables, a semejanza de las variables cuantitativas, que suelen recibir más atención. Se describen algunos de estos índices que miden la variabilidad de las variables cualitativas.

No me gusta el final del verano. Empiezan los días con mal tiempo, me levanto totalmente de noche y anochece cada vez más temprano. Y, por si fuera poco, se aproxima el engorroso momento del cambio de hora.

Además de las molestias del cambio y del tedio de estar dos o tres días recordando la hora que es y la que podría ser de no haber cambiado, hay que proceder a cambiar un montón de relojes de forma manual. Y, por mucho que te esfuerces en cambiarlos todos, siempre te dejas alguno con la hora vieja. No te pasa con el reloj de la cocina, que miras siempre para saber cómo de rápido tienes que desayunar, o con el del coche, que te mira fijamente todas las mañanas. Pero seguro que hay alguno que no cambias. Incluso, alguna vez me ha pasado, que me doy cuenta cuando al siguiente cambio de hora veo que no lo necesita porque lo dejé sin cambiar en la vez anterior.

Estos relojes olvidados me recuerdan un poco a las variables categóricas o cualitativas.

Pensaréis que, una vez más, me he olvidado de tomar la pastilla esta mañana, pero no. Todo tiene su razonamiento. Cuando terminamos un estudio y tenemos ya los resultados, lo primero que hacemos es una descripción de los mismos para,

después, pasar a hacer todo tipo de contrastes, si viene al caso.

Pues bien, las variables cualitativas siempre se menosprecian cuando aplicamos nuestros conocimientos de estadística descriptiva. Habitualmente nos limitamos a clasificarlas y hacer tablas de frecuencia con las que calcular algunos índices como su frecuencia relativa o acumulada, dar alguna medida representativa como la moda y poco más. Con su representación gráfica ya nos esforzamos un poco más, con diagramas de barras o de sectores, pictogramas y otros inventos parecidos. Y, por último, nos aplicamos un poco más cuando relacionamos dos variables cualitativas mediante una tabla de contingencia.

Sin embargo, nos olvidamos de la variabilidad, algo que nunca haríamos con una variable cuantitativa. Las variables cuantitativas son como ese reloj de la pared de la cocina que nos mira directamente a los ojos cada mañana y que no consiente que lo dejemos fuera de hora. Por eso, recurrimos a esos conceptos que entendemos tan bien como la media y la varianza o la desviación típica. Pero el que no conozcamos la forma de medir de forma objetiva la variabilidad de las variables cualitativas o categóricas, ya sean nominales u ordinales, no quiere decir que no exista. Para este fin, se han desarrollado diversos índices de diversidad, que algunos autores distinguen como índices de dispersión, variabilidad y disparidad. Vamos a ver algunos de ellos, cuyas fórmulas podéis ver en el recuadro adjunto, para que podáis disfrutar de la belleza del lenguaje matemático.

Los dos índices más conocidos y utilizados para medir la variabilidad o diversidad son el índice de Blau (o de Hirschman-Herfindal) y el índice de entropía (o de Teachman). Ambos

tienen un significado muy similar y, de hecho, están correlacionados linealmente.

El índice de Blau cuantifica la probabilidad de que dos individuos tomados al azar de una población estén en diferentes categorías de una variable (siempre que el tamaño de la población sea infinito o el muestreo se realice con reemplazo). Su valor mínimo, cero, indicaría que todos los miembros están en la misma categoría, con lo que no habría variedad. Cuanto mayor sea su valor, más dispersos entre las diferentes categorías de la variable estarán los componentes del grupo. Este valor máximo se alcanza cuando los componentes se distribuyen de manera igual entre todas las categorías (sus frecuencias relativas son iguales). Su valor máximo sería $(k-1)/k$, con lo que es función de k (el número de categorías de la variable cualitativa) y no del tamaño de la población. Este valor tiende a 1 al aumentar el número de categorías (para decirlo de forma más correcta, cuando k tiende a infinito).

Veamos algunos ejemplos para aclararnos un poco. Si os fijáis en la fórmula del índice de Blau, el sumatorio de los cuadrados de las frecuencias relativas en una población totalmente homogénea valdrá 1, con lo que el índice valdrá 0. Solo habrá una categoría con frecuencia 1 (el 100%) y el resto con frecuencia cero.

Como hemos dicho, aunque los sujetos se distribuyan de forma similar en todas las categorías, el índice aumenta al aumentar el número de categorías. Por ejemplo, si hay cuatro categorías con una frecuencia de 0,25, el índice de Blau valdrá 0,75 ($1 - (4 \times 0,25^2)$). Si hay cinco categorías con una frecuencia de 0,2, el índice valdrá 0,8 ($1 - (5 \times 0,2^2)$). Y así sucesivamente.

Como ejemplo práctico, imaginad una enfermedad en la que hay diversidad desde el punto de vista genético. En una ciudad A tienen el genotipo 1 el 85% de los enfermos y el genotipo 2 el 15%. El índice de Blau valdrá $1 - (0,85^2 + 0,15^2) = 0,255$. A la vista de este resultado podremos decir que, aunque no es homogénea, el grado de heterogeneidad no es muy alto.

Ahora imaginad una ciudad B con un 60% de genotipo 1, un 25% de genotipo 2 y un 15% de genotipo 3. El índice de Blau valdrá $1 - (0,6^2 + 0,25^2 + 0,15^2) = 0,555$. Claramente, el grado de heterogeneidad es mayor entre los enfermos de la ciudad B que entre los de A. Los más listillos me diréis que eso ya se veía sin calcular el índice, pero tenéis que tener en cuenta que son ejemplos muy sencillos para no echar las bilis calculando. En los estudios de la vida real, más complejos, no suele ser tan evidente y, en cualquier caso, siempre es más objetivo cuantificar la medida que quedarnos con nuestra impresión subjetiva.

Este índice podría usarse también para comparar la diversidad de dos variables diferentes (siempre que tenga sentido hacerlo) pero, el hecho de que su valor máximo dependa del número de categorías de la variable, y no del tamaño de la muestra o de la población, cuestiona su utilidad para comparar la diversidad de variables con diferente número de categorías. Para evitar este problema el índice de Blau puede normalizarse dividiéndolo por su máximo, obteniéndose así el índice de variación cualitativa. Su significado es, lógicamente, el mismo que el del índice de Blau y su valor oscila entre 0 y 1. Así, podremos usar cualquiera de los dos si comparamos la diversidad de dos variables con el mismo número de categorías, pero será más correcto usar el índice de variación cualitativa si las

variables tienen un número de categorías diferente.

El otro índice, algo menos famoso, es el índice de Teachman o de entropía, cuya fórmula también os adjunto. Muy brevemente diremos que su valor mínimo, que es cero, indica que no hay diferencias entre los componentes en la variable de interés (la población es homogénea). Su valor máximo puede estimarse como el valor negativo del logaritmo neperiano del inverso del número de categorías ($-\ln(1/k)$) y se alcanza cuando todas las categorías tienen la misma frecuencia relativa (la entropía alcanza su valor máximo). Como veis, muy parecido al de Blau, que es mucho más sencillo de calcular que el de Teachman.

Para ir acabando esta entrada, el tercer índice del que os quiero hablar hoy nos indica, más que la variabilidad de la población, la dispersión que sus componentes tienen respecto al valor más frecuente. Esto puede medirse mediante la razón de variación, que indica el grado en que los valores observados no coinciden con el de la moda, que es la categoría más frecuente. Como con los anteriores, también os dejo la fórmula en el recuadro adjunto.

Para no desentonar con los anteriores, su valor mínimo también es cero y se obtiene cuando todos los casos coinciden con la moda. Cuanto más bajo el valor, menos dispersión. Cuanto más baja sea la frecuencia absoluta de la moda, más se aproximará a 1, el valor que indica máxima dispersión. Creo que este índice es muy sencillito, así que no le vamos a dedicar más atención.

Y hemos llegado al final. Espero que a partir de ahora prestemos más atención al análisis descriptivo de los resultados de las variables cualitativas. Claro que habría que completarlo con una descripción gráfica adecuada utilizando

los archiconocidos diagramas de barras o de sectores (las tartas) y otros menos conocidos como los diagramas de Pareto. Pero esa es otra historia...



Figura 1. Variación variables cualitativas

Bibliografía

– Turbany J, Barrios M, Carrera MJ. Estadística descriptiva univariante y bivalente. En: Peró M, Leiva D, Guardia J, Solanas A, eds. Estadística aplicada a las ciencias sociales mediante R y R-Commander. Ibergarceta Publicaciones SL. Madrid, 2012: 165-228.

– Solanas A, Selvam RM, Navarro J, Leiva D. Some common indices of group diversity: upper boundaries. Psychol Rep. 2012;111:777-96. ([PubMed](#))

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en octubre de 2019.