EXTENDIENDO EL CONCEPTO DE DISPERSIÓN ESTADÍSTICA A VARIABLES DE CARÁCTER CUALITATIVO

José Antonio Camúñez Ruiz, Mª Dolores Pérez Hidalgo y Francisco Javier Ortega Irizo

Universidad de Sevilla

Resumen

El estudio de la variabilidad en caracteres categóricos rara vez es abordado. A partir de un enfoque menos usado sobre variabilidad en cuantitativas, el de la disparidad, distinto al de la dispersión que, por ejemplo, proporciona la varianza, se propone la construcción de coeficientes de medida de variabilidad en variables cualitativas. La sencillez y proximidad de los mismos permiten que sean abordados en un curso introductorio de estadística descriptiva. Con ejemplos sencillos se introducen las medidas y, también, el profesor amplía la idea que el alumno tiene sobre variabilidad.

Comenzamos con ejemplos de variables dicotómicas en primer lugar, y de más de dos categorías después, introducimos las medidas de disparidad en las respuestas de los individuos. Aquí no hay diferencias cuantitativas, dado que no hay número en las respuestas, hay respuestas diferentes, disparidades, que nuestros alumnos captan y "miden" mediante los coeficientes propuestos.

Conseguimos que una mayoría de nuestros estudiantes acepten la variabilidad de los datos en su doble faceta, dispersión y disparidad, y así, que capten la razón de ser de la estadística. También, con esto se llena uno de los vacíos tradicionales de la enseñanza de esta disciplina.

Abstract

The study of the variability in categorical characters rarely addressed. From least-used approach to quantitative variation in the disparity, different from the dispersion, for example, provides the variance coefficients construction variability measure proposed qualitative variables. The simplicity and proximity of them allow them to be addressed in an introductory course in descriptive statistics. With simple examples are introduced measures and also the teacher extends the idea that the student has on variability

We begin with examples of dichotomous variables first, and after more than two categories, we introduce the measures of disparity in the responses of individuals. There are no quantitative differences, since there is no number on the responses; there are different answers, disparities, our students capture and "measured" by the coefficients proposed.

We got a majority of our students accept the variability of the data in his dual role, dispersion and disparity, and thus to capture the rationale behind the statistics. Also, with this fills one of the gaps in the traditional teaching of this discipline.

1. Introducción

Las variables cualitativas o categóricas siempre han ocupado un mínimo espacio en los cursos introductorios de estadística. Se suelen definir, clasificar en nominales u ordinales, introducir la moda como una medida representativa y, en el caso de las ordinales, alguna medida similar a la mediana. También, representarlas gráficamente, siendo en este aspecto donde, quizás, encontramos más variedad de propuestas: diagramas de barras, de sectores, pictogramas, y una pluralidad de gráficos cuyo nivel de sofisticación depende, casi, de la imaginación de la persona interesada.

Prácticamente, nuestro trabajo en el aula se reduce a lo que acabamos de citar en el caso del estudio de una variable categórica aislada. Después, al tratar con dos variables cualitativas relacionadas entre si, las tablas de contingencia, con sus medidas asociadas, amplían un poco la visión sobre este tipo de estadísticas.

Desde luego, la variabilidad, tan profusamente estudiadas en cuantitativas, no es tratada en general en las categóricas, dando la sensación, entonces, de que este tipo de medidas no existe. Es claro que esa idea de variabilidad alrededor de la media, significado habitual que damos a varianza o desviación típica, no tiene sentido. Se suele usar el término "dispersión" para esta forma de variabilidad.

Pero hay otra manera de entender la variabilidad, la que se detiene en el análisis comparativo de respuestas donde la comparación se reduce a igualdad o desigualdad de las mismas, sin pararse en medir la magnitud de esa desigualdad. Podemos usar en este caso el término "disparidad". Estas medidas, que se emplean aunque con menos frecuencia en variables cuantitativas, pueden extenderse a las cualitativas, pues la disparidad existe siempre que se manifiesten opiniones distintas. O sea, la variabilidad

existe en las categóricas (no tendría sentido cualquier estudio estadístico si no fuese así). Creemos que es algo que debemos inculcar a nuestro alumnos y que, si es posible, construir medidas o indicadores de dichas variabilidad.

En este trabajo presentamos un par de medidas sencillas para casos categóricos (aunque en concepto podríamos hablar de una sola, dado que la diferencia entre ambas es la misma que la existente entre varianza y cuasivarianza), a las que proponemos llamar "coeficientes de disparidad".

En algunos trabajos hemos comprobado la utilidad de estas medidas que, acompañada de lo intuitivas que resultan, creemos, deben ser medidas que engrosen el contenido de una asignatura dedicada a Estadística Descriptiva.

2. Variabilidad en cuantitativas: dispersión y disparidad

En variables cuantitativas nos encontramos como primeras medidas de dispersión la varianza y la cuasivarianza. Gini (1912), cuando estudia la variabilidad entre las cuantitativas distingue dos tipos de variables: las que se definen como un sólo valor real, μ, pero que al ser medido se producen diferentes mediciones debido a los errores asociados a las mismas, por lo que los valores observados u observaciones efectuadas son de la forma $x_i = \mu + \varepsilon_i$, y las que presentan distintas modalidades cuantitativas que van surgiendo con las repetidas observaciones de las variables. Pues bien, para el primer tipo, Gini (1912) propone medidas del tipo de las citadas anteriormente, o sea, medidas de dispersión alrededor de la media (siendo ésta el valor real de la variable), mientras que para las del segundo formula medidas que recojan todas las posibles diferencias, por parejas, entre los valores observados. Serían, pues, medidas construidas a partir de los siguientes agregados: $\sum_i \sum_j \left(x_i - x_j\right)^2$, $\sum_i \sum_j \left|x_i - x_j\right|$, (las distancias entre observaciones son medidas mediante diferencias al cuadrado o diferencias en valor absoluto) donde este autor apuesta más por el segundo que por el primero, pues la que propuso es la conocida como media de las diferencias:

$$\Delta = \frac{\sum_i \sum_j \left|x_i - x_j\right|}{n\left(n-1\right)}.$$

Para el primer agregado es fácil demostrar la siguiente igualdad:

$$\sum_i \sum_j \left(x_i - x_j\right)^2 = 2n \sum_i \left(x_i - \overline{X}\right)^2 .$$

De alguna forma, esta igualdad genera conciliación, tanto sobre la varianza como sobre la cuasivarianza, entre las dos formas de observar la dispersión desde los dos tipos de variables, según Gini (1912).

En todas las medidas citadas hasta ahora la variabilidad depende de dos factores, del número de valores diferentes que nos encontremos y de la distancia entre los mismos (influida por la magnitud de los correspondientes valores). Dos valores, x_i y x_j, que estén muy separados entre sí, por ser dos cantidades muy distintas, aportan mucho peso a la hora de calcular la dispersión mediante cualquiera de esas medidas. Serían éstas las que al principio hemos llamado "medidas de dispersión".

Ahora, podemos plantearnos la variabilidad sólo desde el punto de vista de la disparidad, del número de posibles parejas de componentes distintos que se pueden formar, lo que depende del número de valores distintos que presente una variable, sin tener en cuenta la magnitud de dichos valores. Así, bajo este punto de vista se nos ocurre dos posibles medidas a las que podemos llamar "coeficientes de disparidad" (Perry y Kader, 2005):

$$D_1 = \frac{\sum_i \sum_j c\left(x_i, x_j\right)}{n^2} \text{ y } D_2 = \frac{\sum_i \sum_j c\left(x_i, x_j\right)}{n\left(n-1\right)}, \text{ con } c\left(x_i, x_j\right) = \begin{cases} 1, \text{ si } x_i \neq x_j \\ 0, \text{ si } x_i = x_j \end{cases} .$$

Por tanto, el numerador de estos coeficientes cuenta el número de disparidades que encontramos entre los valores de la variable y, como se ha dicho, no tiene en cuenta la magnitud de dichos valores ni, por tanto, la distancia entre los mismos. Cada disparidad la cuenta dos veces, pues contamos la de x_i con x_j y la de x_j con x_i.

Hemos de destacar que estas dos medidas tienen carácter de coeficiente o indicador, por dos razones: no depende de las unidades de la variable y su recorrido es menor estricto que 1, en la primera, y menor o igual que 1 en la segunda. Podemos escribir: $0 \leq D_1 \leq \frac{n-1}{n} < 1$ y $0 \leq D_2 \leq 1$. Cuando no hay disparidad, cuando todas las observaciones coinciden, ambos coeficientes toman el valor cero. Cuando se produce la máxima disparidad, cuando todas las observaciones son distintas, el primero toma el

valor $\frac{n-1}{n}$ y el segundo el valor 1. En este aspecto, podríamos decir que se trata de medidas relativas de variabilidad.

3. Midiendo la variabilidad en categóricas: coeficientes de disparidad

De las dos formas de medir la variabilidad, comentadas en el apartado anterior, la primera basada en las distancias no es aplicable en variables categóricas. Supongamos el caso más sencillo, una variable de carácter dicotómico donde las dos posibles respuestas son representadas por A y B. Esas respuestas no están definidas por magnitudes numéricas (salvo que codifiquemos arbitrariamente) por lo que no podemos medir la distancia entre A y B, o sea, no podemos construir "medida de dispersión" para esta variable. Lo que si podemos hacer es comparar las respuestas de los individuos y ver si las mismas coinciden o no. Por tanto, los dos coeficientes de disparidad introducidos para cuantitativas serían perfectamente válidos en las cualitativas y esas son las medidas de variabilidad que proponemos para las mismas.

En el caso de una variable categórica con dos posibles respuestas, si p_1 es la proporción de respuestas correspondientes a la primera categoría, o sea, $p_1 = \frac{n_1}{n}$, con n_1 número de veces que aparece la primera respuesta, y si p_2 es la proporción para la segunda respuesta, $p_2 = \frac{n_2}{n}$, podemos escribir el primer coeficiente de disparidad como:

$$\boxed{D_1 = 2 \cdot p_1 \cdot p_2},$$

o sea, 2 veces la varianza de una variable aleatoria Bernoulli.

En general, se consigue otra expresión más para el cálculo de este coeficiente:

$$\boxed{D_1 = 1 - p_1^2 - p_2^2}.$$

Para el caso de tres posibles respuestas obtenemos:

$$\boxed{D_1 = 1 - p_1^2 - p_2^2 - p_3^2}.$$

A partir de ejemplos analizados para dos o tres posibles respuestas de una variable cualitativa nos resulta relativamente fácil establecer diferentes expresiones para el primer coeficiente de disparidad: Si una variable categórica tiene *k* posibles respuestas o categorías y si disponemos de un número finito de observaciones, *n*, y si

$n_1, n_2, ..., n_i, ..., n_k$ representan la frecuencia con que aparece cada una de las categorías con, naturalmente, $n_1 + n_2 + ... + n_i + ... + n_k = n$, llamamos $p_i = \frac{n_1}{n}, \quad i = 1, 2, ..., k$, o sea, la proporción de respuestas que corresponde a la categoría *i* entre las observaciones. Entonces, podemos escribir para el primer coeficiente de disparidad las siguientes expresiones:

$$D_1 = 2\sum_{i<j} p_i p_j \, .$$

$$D_1 = \sum_{i=1}^{k} p_i \left(1 - p_i\right).$$

$$D_1 = 1 - \sum_{i=1}^{k} p_i^2 \, .$$

4. Conclusiones

El concepto de variabilidad es más amplio de lo que habitualmente se explica en los libros de texto y en clase. En variables cuantitativas, además de la idea de dispersión, en general ligada a la desviación respecto a la media, podemos introducir por ejemplo la de disparidad, que conduce a medidas sencillas e intuitivas. La distinción entre el "cuánto" y "con que frecuencia" es la base de la separación entre dispersión y disparidad. Aunque el "cuánto se diferencian los datos" no se puede medir en variables categóricas, sí podemos contar "con qué frecuencia son distintas las respuestas". Por tanto, medidas relacionadas con la disparidad son posibles en variables cualitativas. Creemos que dichas medidas, a las que hemos llamado "coeficientes de disparidad", por su naturalidad y sencillez, deben ser abordadas en un curso de introducción a la estadística descriptiva llenando así uno de los vacíos tradicionales de la enseñanza de esta disciplina.

Bibliografía

BLASIUS, J. & GREENACRE, M. (1998). *Visualization of Categorical Data.* Academic Press.

GINI, C. W. (1912). "Variability and Mutability, contribution to the study of statistical distributions and relations." *Estudi Economico-Giuricici della R. Universita de Cagliari.*

GORDON, T. (1986) “Is the standard deviation tied to the mean?” *Teaching Statistics*, 8(2), 67-70.

KADER, G. D. & PERRY, M. (2007). “Variability for Categorical Variables”. *Journal of Statistics Education*.

PERRY, M. & KADER, G. (2005). “Variation as Unalikeability.” *Teaching Statistics*, *27*(2), 58-60.