Pecvnia Monográfico 2012, pp. 1-20

MIDIENDO LA VARIABILIDAD EN CARACTERES CUALITATIVOS / *MEASURING VARIABILITY IN QUALITATIVE CHARACTERISTICS*

Jesús Basulto Santos[1]
basulto@us.es
José Antonio Camúñez Ruiz[1]
camunez@us.es
Francisco Javier Ortega Irizo[1]
fjortega@us.es
María Dolores Pérez Hidalgo[1]
mdperez@us.es

Universidad de Sevilla

Resumen

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

Palabras clave: Variables cualitativas o categóricas; Variabilidad; Dispersión; Disparidad.

Abstract

The study of variability in categorical characteristics is rarely discussed. From a less used viewpoint of variability in quantitative variables, as it is the one of dissimilarity, which is different from the dispersion that, for example, the variance provides, we propose the construction of two coefficients that measure the variability in qualitative or categorical variables, which we call

[1] Departamento de Economía Aplicada I. Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Avda. Ramón y Cajal, 1, 41018-Sevilla.

coefficients of dissimilarity. Simple examples are provided to introduce the measures, so that the teacher can also evaluate the idea students have about variability, dispersion and dissimilarity.

Keywords: Qualitative or categorical variables; Variability; Dispersion; Dissimilarity.

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. La media aritmética, la que presenta mayores posibilidades de manipulación algebraica, la más conocida y utilizada, la medida por antonomasia en variables cuantitativas, no dispone de su equivalente entre las categóricas.

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, (cualidad de variable, según el diccionario de la Real Academia Española) tan profusamente estudiada en cuantitativas, no es tratada en general en las categóricas, dando la sensación, entonces, de que este tipo de

1. INTRODUCTION

Qualitative or categorical variables have always been residually dealt with in introductory statistics courses. These courses usually include their definition, classification into nominal or ordinal variables, the presentation of the mode as a representative measure and, in the case of the ordinal variables, other kind of measures similar to the median. They are also graphically represented, following a variety of options: bar chart, pie chart, pictograms, and a diversity of charts whose level of sophistication, it can be said, depends on the imagination of the person in question. The arithmetic mean, which presents the largest possibility of being algebraically manipulated, which is the most known and used and the measure *par excellence* in the case of quantitative variables, has no counterpart in the case of the categorical ones.

From a practical point of view, our work in the classroom is reduced to what we have just mentioned in the case of the study of a separate categorical variable. After that, when dealing with two related categorical variables, the use of contingency tables and their associated measures allow spreading a bit the idea of this kind of statistics.

Certainly, variability –defined as the quality of variable, according to the Academy of Spanish Language (RAE)–, which has been so profusely studied in the case of quantitative variables, is not usually dealt with for the categorical ones, which seems to mean that this type of

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" (desemejanza, desigualdad y diferencia de unas cosas respecto de otras, según el diccionario de la Real Academia Española). 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 nuestros alumnos y que, si es posible, construir medidas o indicadores de dicha 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", y las aplicamos a ejemplos sencillos que nos permiten observar, en el aula, si la percepción de variabilidad que tienen nuestros estudiantes es coherente con la que mide estos coeficientes.

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.

measures does not exist. It is clear that the idea of variability around the mean, which is the usual meaning given to the variance or standard deviation, makes no sense in the case of categorical variables. For this kind of variability, the term 'dispersion" is generally used.

However, there is another way of understanding variability, which is the one that focuses on the comparative analysis of responses, where the comparison is reduced to their similarity or disparity, but it does not deal with measuring the amount of disparity. In this case, the term 'dissimilarity' can be used (which is defined as disparity, inequality or difference of some things with regard to others by the Academy of Spanish Language). These measures, which are used for quantitative variables but less frequently, can be spread to the qualitative ones, since dissimilarity exists as long as there are different options. That is, variability exists in the case of categorical variables (otherwise, any statistical analysis would make no sense). We think this is something we must instil in our students and, if possible, construct measures or indicators of the above-mentioned variability.

In this paper we present a couple of simple measures for categorical cases (although strictly speaking we could talk about only one, since the difference between them is the same as the one between variance and quasivariance), which are proposed to be named as 'dissimilarity coefficients' and we apply them in simple examples which allow us to observe in the classroom if the perception of variability our students have is coherent with the one these coefficients measure.

We have checked in some papers the usefulness of these measures, which, together with the fact of being so intuitive, must widen the contents of a subject in Descriptive Statistics.

Dado que en variables categóricas la proporción de respuestas en un sentido u otro es uno de los primeros cálculos que realizamos y que, la idea de proporción enlaza con la de probabilidad para el caso de variables aleatorias, terminamos analizando la similitud entre una de las medidas propuestas y la varianza de una variable probabilística dicotómica tipo Bernoulli.

2. VARIABILIDAD EN CUANTITATIVAS: DISPERSIÓN Y DISPARIDAD

En variables cuantitativas nos encontramos como primeras medidas de dispersión la varianza y la cuasivarianza, cuyas definiciones recordamos: $S^2 = \frac{\sum_i (x_i - \bar{X})^2}{n}$ y $S_c^2 = \frac{\sum_i (x_i - \bar{X})^2}{n-1}$, respectivamente. 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 $x_i = \mu + \varepsilon_i$ (habla de variables relacionadas con la medición en astronomía), 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 siguien-

Provided in categorical variables the proportion of responses in one and another sense is one of the first computations that are carried out, this idea of proportion is connected with the one of probability, so we conclude with analysing the similarity between one of the proposed measures and the variance of a Bernoulli dichotomous random variable.

2. VARIABILITY IN QUANTITATIVE: DISPERSION AND DISSIMILARITY

In quantitative variables, the first measures of dispersion are variance and quasi-variance, whose expressions are reminded: $S^2 = \frac{\sum_i (x_i - \bar{X})^2}{n}$ and $S_c^2 = \frac{\sum_i (x_i - \bar{X})^2}{n-1}$, respectively. When Gini (1912) studies variability in quantitative variables, he distinguishes between two types of variables: those which are defined as an only real value, μ, but when this is measured there are different measurements due to mistakes associated to the former, so the observed values or observations are in the form of $x_i = \mu + \varepsilon_i$ (actually, he talks about variables related to the measurement in astronomy); and those which present different qualitative categories that arise with the repeated observations of the variables. In this context, for the first type, Gini (1912) proposes measures which are similar to the ones we have mentioned above, that is, measures of dispersion around the mean (which is the real value of the variable), whereas for the second group he formulates measures that include all the possible pairwise differences among observed values. They would be, therefore, measures that would be constructed from the following expressions:

tes agregados: $\sum_i\sum_j (x_i - x_j)^2$, $\sum_i\sum_j |x_i - x_j|$, (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:

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

Para el primer agregado es fácil demostrar la siguiente igualdad:

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

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).

Podemos construir dos nuevas medidas usando el agregado del primer miembro de la igualdad anterior, a los que podemos llamar, por ejemplo, "promedios cuadráticos de diferencias por pares" y que definimos a continuación:

$$V^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n^2} \text{ y } V_c^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n(n-1)}$$

La igualdad de arriba nos permite escribir: $V^2 = 2 \cdot S^2$, $V_c^2 = 2 \cdot S_c^2$.

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).

$\sum_i\sum_j (x_i - x_j)^2$, $\sum_i\sum_j |x_i - x_j|$, (the distances among observations are measured as squared differences or differences in absolute value), although this author banks on the first one rather than the second, since he proposed the measure known as differences mean:

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

For the first expression it is easy to prove the following equality:

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

Somehow, this equality makes agreement come, on both the variance and the quasivariance, about the two ways of observing the dispersion from both types of variables, according to Gini (1912).

We can construct two new measures using the first side in the previous equality, which can be called, for example, 'squared means of pairwise differences" and which are defined as follow:

$$V^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n^2} \text{ and } V_c^2 = \frac{\sum_i\sum_j (x_i - x_j)^2}{n(n-1)}.$$

The above equality allows us to write the expressions as: $V^2 = 2 \cdot S^2$, $V_c^2 = 2 \cdot S_c^2$.

In all the aforementioned measures, variability depends on two factors: on the number of different values that can be found and on the distance among them

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 ocurren dos posibles medidas a las que podemos llamar "coeficientes de disparidad" (Perry y Kader, 2005):

(affected by the magnitude of the respective values). Two values, x_i and x_j, which are very separated from each other, as they are two very different quantities, present a lot of weight in order to calculate dispersion through any of those measures. These would be what at the beginning we have called 'measures of dispersion'.

In this point, we can consider variability from the viewpoint of dissimilarity, of the number of possible pairs of different components, which depends on the number of different values a variable presents, without taking into account their magnitude. Thus, from this point of view, two measures can be defined, which can be called 'coefficients of dissimilarity' (Perry and Kader, 2005):

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

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. Téngase presente que en una muestra tamaño *n*, si todos los valores observados son distintos, el número total

Therefore, the numerator in both coefficients counts the number of dissimilarities that are found among the values of the variable and, as it has already been said, it does not take into account the magnitude of the values nor the distance among them. Every dissimilarity is counted twice, since it is counted the dissimilarity between x_i and x_j, and the one between x_j with x_i.

It must be emphasised that these two measures present the nature of coefficient or indicator for two reasons: they do not depend on the variable units and their range is less than 1, in the first measure, and less than or equal to 1, in the second one. It must also be considered that in a sample of size *n*, if all observed values are different, the total

de posibles parejas que se pueden formar, de x_i con x_j y de x_j con x_i, es n^2.

A ese número restamos las parejas del tipo (x_i, x_i), que son n en total, nos queda como número máximo de parejas con componentes distintos $n^2 - n = n(n-1)$. Podemos escribir:

$0 \le D_1 \le \frac{n-1}{n} < 1$ y $0 \le D_2 \le 1$. 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.

Mostramos ejemplos ilustrativos sencillos:

Ejemplo 1: La variable X toma 5 valores siendo todos distintos, $X : \{1, 2, 3, 4, 5\}$.

La media aritmética es 3.

Calculamos en primer lugar las "medidas de dispersión" comentadas arriba.

number of possible pairs x_i with x_j and x_j with x_i) to be formed n^2. From that number we subtract the pairs of the form (x_i, x_i), which are n in total, so the highest number of pairs with different components is $n^2 - n = n(n-1)$. We can write: $0 \le D_1 \le \frac{n-1}{n} < 1$ and $0 \le D_2 \le 1$. When there is no disparity, when all the observations coincide, both coefficients take the value zero. When there is no dissimilarity, that is, when all observations are the same, both coefficients equal zero. When there is the highest dissimilarity, that is, when all observations are different, the first coefficient equals $\frac{n-1}{n}$ and the second one equals 1. In this regard, it could be said that they are relative measures of variability.

Next some simple illustrative examples are shown.

Example 1: Variable X presents 5 different values, $X : \{1, 2, 3, 4, 5\}$.

The arithmetic mean is 3.

We first calculate the aforementioned 'measures of dispersion'.

Tabla 1. Cálculo de las desviaciones al cuadrado respecto de la media
Table 1. Calculation of squared differences around the mean

x_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4
	Total	10

Entonces, $S^2 = \frac{10}{5} = 2$ y $S_c^2 = \frac{10}{4} = 2'5$.

Para la "media de las diferencias" construimos la siguiente tabla:

Then, $S^2 = \frac{10}{5} = 2$ and $S_c^2 = \frac{10}{4} = 2'5$.

For the 'differences mean' we construct the following table:

Tabla 2. Cálculo de las diferencias por parejas en valor absoluto
Table 2. Calculation of squared pairwise differences in absolute value

$\lvert x_i - x_j \rvert$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	2	3	4	10
2	1	0	1	2	3	7
3	2	1	0	1	2	6
4	3	2	1	0	1	7
5	4	3	2	1	0	10
					Total	40

$\Delta = \frac{40}{5 \cdot 4} = 2$. Para los promedios cuadráticos de diferencias por pares: / *For the squared means of pairwise differences*:

Tabla 3. Cálculo de las diferencias cuadráticas por parejas
Table 3. Calculation of squared pairwise differences

$(x_i - x_j)^2$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	4	9	16	30
2	1	0	1	4	9	15
3	4	1	0	1	4	10
4	9	4	1	0	1	15
5	16	9	4	1	0	30
					Total	100

$$V^2 = \frac{100}{5^2} = 4, \; V_c^2 = \frac{100}{5 \cdot 4} = 5.$$

Calculamos por último los "coeficientes de disparidad" / We finally calculate the 'coefficients of dissimilarity':

Tabla 4. Cálculo de las disparidades por parejas
Table 4. Calculation of pairwise dissimilarities

$c(x_i, x_j)$						
x_i \ x_j	1	2	3	4	5	Suma de cada fila *Total in row*
1	0	1	1	1	1	4
2	1	0	1	1	1	4
3	1	1	0	1	1	4
4	1	1	1	0	1	4
5	1	1	1	1	0	4
					Total	20

Por tanto / *Therefore* $D_1 = \frac{20}{25} = 0'8$ y / *and* $D_2 = \frac{20}{20} = 1$.

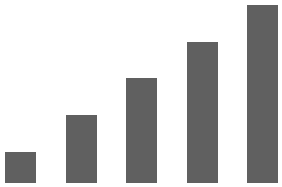
Estamos en un caso de máxima disparidad, todos los valores observados de la variable son distintos.

No hay cosa mejor para visualizar la variabilidad que observar los propios valores mediante alguna asociación geométrica, sobre todo cuando, como en este caso, tenemos pocos valores observados. Construimos, entonces, cinco barras cuyas longitudes son proporcionales a la magnitudes de los datos, y proponemos a los estudiantes su observación para que comparen con otras variables también representadas. Advertimos sobre la posible confusión de este gráfico con el diagrama de barras habitual en estadística. Aquí, y en los ejemplos que siguen, la longitud de cada barra no representa la frecuencia absoluta de un valor de la variable, sino que es el propio valor, y observamos en el conjunto cómo de diferentes son entre sí los valores observados.

This is the case of highest dissimilarity, since all observed values of the variable are different.

A better way of visualising variability is to observe the values through some geometric association, especially when there are few observed values, like in this case. Then, we draw five bars, whose height is proportional to the magnitudes of data and we suggest that our students observe them in order to compare with other variables which have also been presented. We warn them about mistaking this chart for the usual bar chart in statistics. In this example and for the following ones, every bar height does not represent the absolute frequency for a value of the variable, but the value itself, and we observe how different to each other the observed values are within the set.

Gráfico 1. Visualización de 5 valores observados

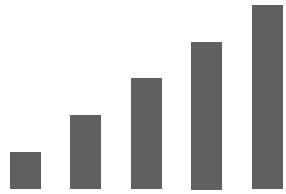


Ejemplo 2. La variable *X* toma también 5 valores distintos, $X:\{1,3,5,7,9\}$. La diferencia con la anterior está en la magnitud de los mismos. Procedemos con los mismos cálculos y representamos de manera similar al anterior. En la tabla resumen que ponemos más abajo (Tabla 5) aparecen los valores de los estadísticos de dispersión y de disparidad.

Ejemplo 3. La variable X toma estos cinco valores $X:\{1,1,3,5,5\}$. Aquí se da más paridad o, quizás mejor, menos disparidad. Resumiremos en la tabla. Igual haremos con el último de los cuatro ejemplos.

Ejemplo 4. La variable X toma estos cinco valores $X:\{1,1,5,9,9\}$.

Graph 1. Visualisation of 5 observed values

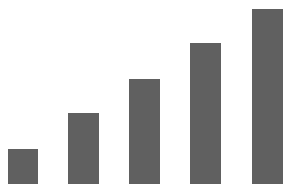


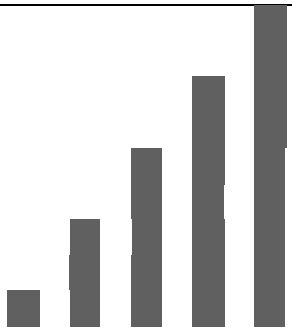
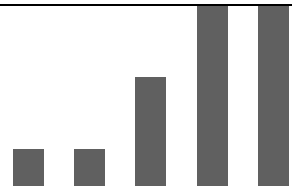
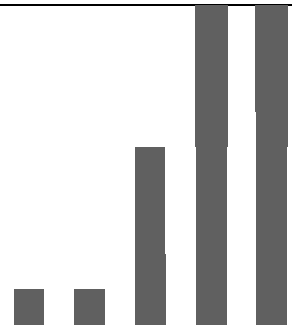
Example 2: Variable *X* also presents 5 different values, $X:\{1,3,5,7,9\}$. The difference with the previous one is their magnitude. We proceed with the same computations and present data in a similar way. In the summary table below (Table 5) all measures of dispersion and dissimilarity are presented.

Example 3. Variable X presents these 5 values, $X:\{1,1,3,5,5\}$. In this case there is more similarity, or more correct, less dissimilarity. This will also be summarised in the table, as it will be for the last example.

Example 4. Variable X presents these 5 values, $X:\{1,1,5,9,9\}$.

Tabla 5. Cuadro resumen de las medidas de variabilidad para los cuatro ejemplos / *Table 5. Summary Table of measures of variability for the four examples*

Ejemplo / Example	S^2	S_c^2	Δ	V^2	V_c^2	D_1	D_2
	2	2'5	2	4	5	0'8	1

	8	10	4	16	20	0'8	1
	3'2	4	2'4	6'4	8	0'64	0'8
	12'8	16	4'8	25'6	32	0'64	0'8

Comparamos entre sí los ejemplos:

- *Ejemplo* 1 y *Ejemplo* 3: mayor dispersión en el 3 y mayor disparidad en el 1. Podemos decir que en el ejemplo 3 hay mayor dispersión que en el 1 y, sin embargo, menor disparidad.
- *Ejemplo* 1 y *Ejemplo* 2: mayor dispersión en el 2 que en el 1 (las diferencias en cuanto a sus magnitudes son mayores en los valores observados en el 2 que en el 1) y la misma disparidad.
- *Ejemplo* 3 y *Ejemplo* 4: mayor dispersión en el 4 que en el 3 (las diferencias en cuanto a sus magnitudes son mayores en los valores observados en el 4 que en el 3) y la misma disparidad.

A comparison between examples is carried out:

- *Example* 1 and *Example* 3: higher dispersion in 3 and higher dissimilarity in 1. It can be said that in example 3 there is more dispersion than in 1 but less dissimilarity.
- *Example* 1 and *Example* 2: higher dispersion in 2 than in 1 (the differences in the magnitude are higher in the observed values in 2 than in 1) and same dissimilarity.
- *Example* 3 and *Example* 4: higher dispersion in 4 than in 3 (the differences in the magnitude are higher in the observed values in 4 than in 3) and same dissimilarity.

- De los cuatro ejemplos, el de mayor dispersión es el 4 y, sin embargo, es uno de los de menor disparidad.
- De los cuatro ejemplos, el de menor dispersión es el 1 y, sin embargo, es uno de los de mayor disparidad.

Por tanto, hemos de distinguir entre lo que es el "cuánto" de lo que es "con qué frecuencia", o sea, la distinción entre medidas basadas en la distancia (dispersión) de las más simples basadas en la disyuntiva entre igualdad o no igualdad (disparidad). Es interesante intentar captar la percepción que nuestros estudiantes tienen de la variabilidad mediante el ejercicio sencillo de mostrar representaciones similares a las anteriores para que se manifiesten sobre cuál presenta mayor o menor 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 una "medida de dispersión" para esta variable. Lo que sí 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.

- Out of the four examples, the highest dispersion is in 4 and, however, it is one of the examples where there is less dissimilarity.
- Out of the four examples, the smallest dispersion is in 1 and, however, it is one of the examples where there is more dissimilarity.

Therefore, we have to distinguish between what is "how much" of what is "with what frequency", or, the distinction between measures based on the distance (dispersion) of the simplest stocks in the dilemma between equality or not equality (disparity). It is interesting to try to catch the perception that our students have of the variability by means of the exercise simple to show representations similar to the previous ones in order that they demonstrate on which he presents major or minor variability.

3. MEASURING VARIABILITY IN CATEGORICAL CHARACTERISTICS: COEFFICIENTS OF DISSIMILARITY

Out of the two ways of measuring variability which were presented in the previous section, the first one based on distances is not applicable to categorical variables. Let us figure out the simplest case, a dichotomous variable where the two possible responses are represented by A and B. These responses are not defined as numerical magnitudes (unless they are arbitrarily codified), so we cannot measure the distance between A and B, that is, we cannot construct a 'measure of dispersion' for this variable. What we can do is to compare the individuals' responses and observe whether they are the same or not. Therefore, both coefficients of dissimilarity which were presented for quantitative variables could also be valid in the case of qualitative ones, and they are the measure of variability we propose for them.

Planteamos tres ejemplos de variables dicotómicas en los que, para los tres casos, requerimos las respuestas de 6 individuos. Visualizamos las respuestas y calculamos los dos coeficientes de disparidad en cada uno de los tres casos:

Ejemplo 1: $X:\{A,B,B,B,B,B\}$.

Ejemplo 2: $X:\{A,A,A,B,B,B\}$.

Ejemplo 3: $X:\{A,A,B,B,B,B\}$.

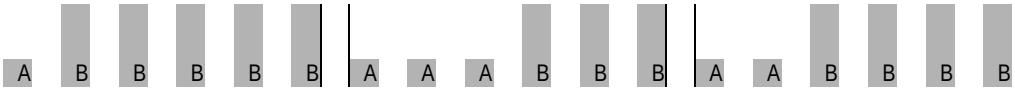
We set out three examples of dichotomous variables where the response of 6 individuals is needed. We show the responses and calculate both coefficients of dissimilarities for each of the three cases:

Example 1: $X:\{A,B,B,B,B,B\}$.

Example 2: $X:\{A,A,A,B,B,B\}$.

Example 3: $X:\{A,A,B,B,B,B\}$.

Gráfico 2. Visualización de tres variables categóricas con dos posibles respuesta cada una / *Graph 2. Visualisation of three categorical variables with two possible responses*



Ejemplo 1/*Example 1* Ejemplo 2/*Example 2* Ejemplo 3/*Example 3*

Calculamos los coeficientes para los tres ejemplos consecutivamente, usando tablas de disparidades similares a los ejemplos de cuantitativas, donde hemos sombreado las "cajas" donde aparecen 1 (disparidades).

We calculate the coefficients for the three examples consecutively, using tables of dissimilarities which are similar to the ones used in the examples of quantitative variables. The cells in the tables where the value is 1 (there is dissimilarity) have been shaded.

Tabla 6. Disparidades del *Ejemplo* 1 / *Table 6. Dissimilarities in Example 1*

$c(x_i, x_j)$							
	A	B	B	B	B	B	Suma de cada fila *Total in row*
A	0	1	1	1	1	1	5
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
B	1	0	0	0	0	0	1
						Total	10

$$D_1 = \frac{10}{6^2} = 0'277\,, \qquad D_2 = \frac{10}{6\cdot 5} = 0'333\,.$$

Tabla 7. Disparidades del *Ejemplo* 2 / *Table 7. Dissimilarities in Example 2*

$c(x_i, x_j)$							
	A	A	A	B	B	B	Suma de cada fila *Total in row*
A	0	0	0	1	1	1	3
A	0	0	0	1	1	1	3
A	0	0	0	1	1	1	3
B	1	1	1	0	0	0	3
B	1	1	1	0	0	0	3
B	1	1	1	0	0	0	3
						Total	18

$$D_1 = \frac{18}{6^2} = 0'5\,, \qquad D_2 = \frac{18}{6 \cdot 5} = 0'6\,.$$

Tabla 8. Disparidades del *Ejemplo* 3 / *Table 8. Dissimilarities in Example 3*

$c(x_i, x_j)$							
	A	A	B	B	B	B	Suma de cada fila *Total in row*
A	0	0	1	1	1	1	4
A	0	0	1	1	1	1	4
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
B	1	1	0	0	0	0	2
						Total	16

$$D_1 = \frac{16}{6^2} = 0'444\,, \qquad D_2 = \frac{16}{6 \cdot 5} = 0'533\,.$$

Según estos coeficientes, la variable categórica donde hay menor variabilidad (en el sentido de disparidad), es la del *Ejemplo* 1, y la de mayor, la del *Ejemplo* 2.

En el *Ejemplo* 1, para el primer coeficiente podemos escribir, observando las dimensiones de las cajas donde aparecen 1:

According to these coefficients, the categorical variable where there is the smallest variability (in the sense of dissimilarity) is the one in *Example* 1, whereas the highest one is in *Example* 2.

In *Example* 1, observing the cells where a value 1 is present, the first coefficient can be written as:

$$D_1 = \frac{1 \cdot 5 + 5 \cdot 1}{6^2} = \frac{2 \cdot 1 \cdot 5}{6^2} = 2 \cdot \frac{1}{6} \cdot \frac{5}{6}$$

Obsérvese que la primera fracción, $\frac{1}{6}$, es la proporción de respuestas A que encontramos en esa variable categórica, mientras que la segunda, $\frac{5}{6}$, es la de respuestas B. Por tanto, 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 (la misma relación que la existente entre varianza y cuasivarianza, por una parte, y los dos promedios cuadráticos de diferencias por pares, por la otra, en el caso cuantitativo). Podíamos evitar ese 2 si contásemos las disparidades de una pareja una sola vez. Como ya se ha comentado, en los coeficientes propuestos contamos la disparidad de x_i con x_j y la de x_j con x_i. A nivel práctico bastaría con dividir por 2 esos coeficientes. Ahora bien, al hacerlo cambiaríamos el recorrido de ambos. Por ejemplo, D_2, en lugar de tomar valores entre 0 y 1, los tomaría entre 0 y 0'5, como ocurre con los posibles valores de la varianza de una distribución Bernoulli.

Alguna manipulación más es posible:

The first fraction, $\frac{1}{6}$, is the proportion of A responses that are found in that categorical variable, whereas the second one, $\frac{5}{6}$, is the proportion of B responses. Therefore, in the case of a categorical variable with two possible responses, if p_1 is the proportion of responses corresponding to the first category, that is, $p_1 = \frac{n_1}{n}$, with n_1 being the number of times the first responses is found, and if p_2 is the proportion for the second response, that is, $p_2 = \frac{n_2}{n}$, then, the first coefficient of dissimilarity can be written as:

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

That is, twice the variance of a Bernoulli random variable (the same relationship as the one existing between variance and quasivariance, on the one hand, and both squared pairwise differences means, on the other, in the case of quantitative variables). We could eliminate '2' in the above expression by counting every pairwise dissimilarity only once. As it was already said, in the proposed coefficients, the dissimilarity of x_i with x_j and the one of x_j with x_i is counted. At a practical level, it would be enough to divide both coefficients by 2. Nevertheless, by doing that their range would be modified. For instance, D_2, instead of being between 0 and 1, would vary from 0 to 0,5, as it happens with the possible values of the variance in a Bernoulli distribution.

Some more manipulation is possible:

La suma de "unos" que aparece en cada tabla se podría construir así (mirar zonas sombreadas):

The addition of 'ones' in each table could be calculated as (observe the shaded areas):

$$n_1 \cdot n_2 + n_2 \cdot n_1 = n_1 (n - n_1) + n_2 (n - n_2)$$

Por tanto, / *Therefore*,

$$D_1 = \frac{n_1 (n - n_1) + n_2 (n - n_2)}{n^2} = \frac{n_1}{n} \cdot \frac{n - n_1}{n} + \frac{n_2}{n} \cdot \frac{n - n_2}{n}.$$

O sea, ese coeficiente se puede escribir también como:

That is, the coefficient can also be written as:

$$\boxed{D_1 = p_1 (1 - p_1) + p_2 (1 - p_2)}.$$

$$\boxed{D_1 = p_1 (1 - p_1) + p_2 (1 - p_2)}.$$

Aún otra expresión más. El número de "unos" que hay en la caja también se puede calcular restando al total de celdas de la tabla el número de "ceros". Así, en el *Ejemplo* 3 sería $16 = 6^2 - 2^2 - 4^2$. Por tanto,

$D_1 = \frac{6^2 - 2^2 - 4^2}{6^2} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$. En general, otra expresión más para el cálculo de este coeficiente es:

There is also another possible expression. The number of 'ones' in the table can also be calculated by subtracting the number of 'zeros' from the total number of cells in the table. Thus, in *Example* 3, it would be $16 = 6^2 - 2^2 - 4^2$. Therefore, $D_1 = \frac{6^2 - 2^2 - 4^2}{6^2} = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$. In general, another expression for the computation of this coefficient is:

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

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

A continuación planteamos otro ejemplo de variable categórica en el que hay tres posibles respuestas, A, B y C, de una cuestión planteada a 8 individuos, dando como resultado la siguiente estadística ya agrupada por respuestas:

$X : \{A, B, B, C, C, C, C, C\}$. Visualizamos estas respuestas, pero en este caso evitamos la utilización de la longitud como elemento distintivo de las respuestas con el objeto de que las mismas pueden ejercer impacto visual ajeno a lo buscado.

We next set out another example of categorical variable where there are three possible responses, A, B and C, to a question posed to 8 individuals, resulting in the following statistics, which have already been grouped according to the responses:

$X : \{A, B, B, C, C, C, C, C\}$. We show these responses but in this case we avoid the use of the height as a distinguishing element among the responses, since they can have a visual impact different to the proper one.

Gráfico 3. Visualización de una variable categórica con tres posibles respuestas / *Graph 3. Visualisation of a categorical variable with three possible responses*

Calculamos para esta variable los dos coeficientes de disparidad. En primer lugar, la tabla de disparidades: / We calculate both coefficients of dissimilarity for this variable, We first present the table of dissimilarities:

Tabla 9. Disparidades de variable categórica con tres posibles respuestas ***Table 9. Dissimilarities of a categorical variable with three possible responses***

$c(x_i, x_j)$

	A	B	B	C	C	C	C	C	Suma de cada fila *Total in row*
A	0	1	1	1	1	1	1	1	7
B	1	0	0	1	1	1	1	1	6
B	1	0	0	1	1	1	1	1	6
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
C	1	1	1	0	0	0	0	0	3
								Total	34

Entonces,

$D_1 = \frac{34}{8^2} = 0'531$ y $D_2 = \frac{34}{8 \cdot 7} = 0'607$. Si observamos la tabla, el número de "unos" que hay en la misma es la suma del número de celdas contenidos en la tres cajas enmarcadas y sombreadas (la de A con B, la de A con C, y la de B con C) que, a su vez, están duplicadas. O sea, $34 = 2\left(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5\right)$. Por tanto,

$$D_1 = \frac{2(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5)}{8^2} = 2\left(\frac{1}{8} \cdot \frac{2}{8} + \frac{1}{8} \cdot \frac{5}{8} + \frac{2}{8} \cdot \frac{5}{8}\right).$$

Entonces, si p_1, p_2 y p_3 son las proporciones de individuos que escogen cada una de las tres respuestas, tenemos:

$$\boxed{D_1 = 2\left(p_1 p_2 + p_1 p_3 + p_2 p_3\right)}$$

Then,

$D_1 = \frac{34}{8^2} = 0'531$ and $D_2 = \frac{34}{8 \cdot 7} = 0'607$. If we observe the table, the number of 'ones' that exists in the same one is the sum of the number of cells contained in three boxes framed and shaded (her of A with B, her of A with C, and her of B with C) that, in turn, are duplicated. Or, $34 = 2\left(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5\right)$. Therefore,

$$D_1 = \frac{2(1 \cdot 2 + 1 \cdot 5 + 2 \cdot 5)}{8^2} = 2\left(\frac{1}{8} \cdot \frac{2}{8} + \frac{1}{8} \cdot \frac{5}{8} + \frac{2}{8} \cdot \frac{5}{8}\right).$$

If p_1, p_2 and p_3 are the proportions of individuals that who choose each of three responses, respectively, then:

$$\boxed{D_1 = 2\left(p_1 p_2 + p_1 p_3 + p_2 p_3\right)}$$

También, el número de "unos" de la caja anterior puede ser determinado mediante $34 = 1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)$, por lo que el coeficiente sería,

The number of 'ones' in the table above can also be calculated as $34 = 1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)$, so the coefficient would be:

$$D_1 = \frac{1\cdot(8-1)+2\cdot(8-2)+5\cdot(8-5)}{8^2} = \frac{1}{8}\cdot\frac{8-1}{8}+\frac{2}{8}\cdot\frac{8-2}{8}+\frac{5}{8}\cdot\frac{8-5}{8} = \frac{1}{8}\left(1-\frac{1}{8}\right)$$

$$+\frac{2}{8}\left(1-\frac{2}{8}\right)+\frac{5}{8}\left(1-\frac{5}{8}\right).$$ En general, / *In general,*

$$\boxed{D_1 = p_1(1-p_1)+p_2(1-p_2)+p_3(1-p_3)}.$$

Por último, la suma de "unos" puede calcularse también restando al total de celdas de la caja, 8^2, el total de ceros que hay en ella. O sea, $34 = 8^2-1^2-2^2-5^2$. Por tanto,

Finally, the sum of 'ones' can also be calculated by subtracting from the total number of cells in the table, 8^2, the number of 'zeros' in it. That is, $34 = 8^2-1^2-2^2-5^2$. Therefore,

$$D_1 = \frac{8^2-1^2-2^2-5^2}{8^2} = 1-\left(\frac{1}{8}\right)^2-\left(\frac{2}{8}\right)^2-\left(\frac{5}{8}\right)^2.$$ En general,

$$D_1 = \frac{8^2-1^2-2^2-5^2}{8^2} = 1-\left(\frac{1}{8}\right)^2-\left(\frac{2}{8}\right)^2-\left(\frac{5}{8}\right)^2.$$ In general,

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

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

A partir de los 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 $p_i = \frac{n_1}{n}$, $i = 1,2,...,k$, o sea, la proporción de respuestas que corresponde a la categoría i entre las observaciones.

From the examples above for two or three possible responses of a qualitative variable it is relatively easy to define different expressions for the first coefficient of dissimilarity: If there is a categorical variable with k possible responses o categories, if there are a finite number of observations, n, and if $n_1, n_2, ..., n_i, ..., n_k$, represent the frequency of appearance of every category, with $n_1 + n_2 + ... + n_i + ... + n_k = n$, then $p_i = \frac{n_1}{n}$, $i = 1,2,...,k$, is the proportion of responses in the category i for the observations.

Entonces, podemos escribir para el primer coeficiente de disparidad las siguientes expresiones: / Then, the first coefficient of dissimilarity can be written as:

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

$$D_1 = \sum_{i=1}^{k} p_i (1-p_i) \ ,$$

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

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 qué 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. La estadística existe al existir variabilidad dentro de un carácter medido en una población y dicho carácter puede ser cuantitativo o cualitativo. Es función del usuario de la estadística poder medir dicha variabilidad. La visualización de ejemplos simples por parte de los alumnos permitirá al profesor la captación de las ideas que sobre variabilidad tienen los mismos.

4. CONCLUSIONS

The concept of variability is wider than the one that is usually discussed in literature and classroom. In the case of quantitative variables, apart from the idea of dispersion, in general related to the deviation around the mean, it can be introduced, for example, the idea of dissimilarity, which results in simple and intuitive measures. The difference between 'when' and 'how frequently' is the base to distinguish between dispersion and dissimilarity. Even though 'how much data are different' cannot be measured in categorical variables, it is possible to count 'how frequently responses are different'. Therefore, measures related to dissimilarity are possible to be defined in qualitative variables. We think that these measures, which we have called 'coefficients of dissimilarity', due to their naturalness and simplicity, must be dealt with in a descriptive statistics introductory course, filling in this way one of the gags in the teaching of this subject. Statistics exists because variability inside a characteristic measured on a population exists, and that characteristic can be quantitative or qualitative. To measure that variability is a role corresponding to the user of statistics. The visualization of simple examples by students will allow the teacher to catch the idea they have on variability.

BIBLIOGRAFÍA/REFERENCES

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons.

Blasius, J. y Greenacre, M. (1998). *Visualization of categorical data.* San Diego (CA): 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. y Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2), 1-17.

Loosen, F., Lioen, M. y Lacante, M. (1985). The standard deviation: Some drawbacks to an intuitive approach. *Teaching Statisctics*, *7*(1), 2-5.

Perry, M. y Kader, G. (2005). Variation as unalikeability. *Teaching Statistics*, *27*(2), 58-60.