Deviation, inequality, polarization : some measures of social diversity

Author(s): Escobar, Modesto
Date 1997
Type Working Paper
Series Estudios = Working papers / Instituto Juan March de Estudios e Investigaciones, Centro de Estudios Avanzados en Ciencias Sociales 1997/102
City: Madrid
Publisher: Centro de Estudios Avanzados en Ciencias Sociales

Deviation, inequality, polarization: Some measures of social diversity

It can be hardly denied that the issue of inequality is crucial in the study of society. The concept of diversity can already be found in pre-sociological theories, from the hierarchical conceptions of Plato to Rousseau's discussion of the origin of inequality. Later, the founding fathers of the discipline were obviously concerned with the study of divisions in society, formulated in the concepts of class in Marx, status in Weber and mechanic/organic solidarity in Durkheim. Thus, Nisbet considers status to be one of the five key ideas of the sociological tradition. Since similarities and inequalities within and between societies must be an object of study for sociologists, precise tools are needed to measure social diversity, not only in order to study society but also from a strictly methodological point of view. This paper presents some results of the measure of variation, one of the most controversial issues in most empirical studies of dispersion, including the measure of inequality, which plays an important role in economics and sociology.

All too often, descriptive statistics appears to be the science of percentages and averages, because many general works use these statistics to simplify and illustrate their arguments. However, in order to obtain a complete picture of the structure of a society, the distribution of the variables is as important as their means. In the same way, advanced statistics is essentially based on the concept of variance. In general, most of the techniques employed to explain causal relationships between variables use the term "percentage of explained variance".

Before presenting a number of statistics whose main purpose is to quantify diversity within populations analyzed through samples, it is perhaps necessary to clarify the nature of three related, but different, concepts used in the study of diversity: deviation, inequality and polarization.

If individuals in the population are distant from a certain point of reference, such as the arithmetic mean, we speak of *deviation*. This means that there is a benchmark for comparing all individuals. The further they are from this point, the larger the measure of deviation. When we speak of *inequality*, there is no single point of comparison; rather, each individual has to be

compared with all the others in the population. And finally, the concept of *polarization* (or the opposite concept of concentration) indicates the extent to which the values of the variables are close or far away from the two extremes of the distribution. It is obvious that these three notions of diversity are closely related to each other, and that so too are the statistics used to measure them; an increase in a measure of deviation in most cases appears to correspond to an increase in the measure of inequality. As a result it is very often the case that a certain statistic is used to take into account some aspects that could be explained in a better way by a different kind of statistic.

In this paper I propose a family of statistics to treat the subject of social diversity from the perspective of examining measures of both deviation and inequality.

One of the most frequently used measures in complex statistical analysis is that of variance, which is the average of the squared deviations of the values of the variable around the central mean. The specific properties of this statistic may be appreciated by imagining a situation in which four people are eating four indivisible chickens in a restaurant, so the average of chickens is one (see Table 1). Apart from equal treatment, whereby each person eats his/her chicken, there could be four cases of inequality:

(a) Only one person does not eat his/ her chicken, so another eats two,

(b) Two people do not eat at all. In this case, two situations are possible:

(b1) The other two people eat two chickens each, or

(b2) The third person eats one chicken and the fourth eats three.

(c) Three people do not eat at all and the fourth one eats all the chickens.

In the case of equality there is null variance: since everyone is at the central point, there is no variation. In case a) two people deviate one unit from the mean. As there are four people in total, the variance is a half unit. In case (b1) the variance is one since they all deviate one unit from the mean. Cases (b2) and (c) are more interesting. In (b2) one person eats two more

chickens than the mean, so the square of his/her deviation is four and the average of the total deviations yields a variance of one and a half (equal to six over four). In the final situation the variance is equal to three, as the square of the deviation around the mean for the person who eats the four chickens is nine.

Table 1. *Distibutions of inequality for four units*

	Equality	Inequality			
		a	b	c	d
First person	1	2	2	3	4
Second person	1	1	2	1	0
Third person	1	1	0	0	0
Forth person	1	0	0	0	0
Mean	1,0	1,0	1,0	1,0	1,0
Variance	0,0	0,5	1,0	1,5	3,0
St. Deviation	0,0	0,7	1,0	1,2	1,7
Mean Dev.	0,0	0,5	1,0	1,0	1,5
C.of variation.	0,0	0,7	1,0	1,2	1,7

This example illustrates two characteristics of variance. First, that this increases considerably for variables with values which are very distant from the mean (the case where a single person eats the four chickens). Secondly, the variance is measured in a magnitude that is the square of the magnitude used for the original values of the variable. In this example, the statement that the variance is three means that, on average, people deviate from the mean three "square chickens". For this reason, it is often better to work with the original magnitude of measurement instead of with its square, which is why it is very useful to give the square root of the variance, known as standard deviation. Thus, three "square chickens" become 1.7 chickens.

The mathematical properties of variance are as follows:

(1) It is always greater than or equal to zero.

(2) It can be obtained from the difference between the average of squares and the square mean.

(3) If a variable is constant, its variance is equal to zero.

(4) If a constant C is added to a variable, the new variable has the same variance as the old one.

(5) If a variable is multiplied by a constant C, the new variable has a variance that is equal to C^2 the variance of the old variable.

(6) The variance of the sum of two variables is equal to the sum of the variance minus twice the covariance between both variables.

One drawback of variance is that it is an absolute coefficient as measured in a magnitude that is the square of the magnitude used for the original values of the variable. As explained above, one way of avoiding this problem is to use the standard deviation. However, two difficulties remain even when this is done. First, it is not easy to conclude whether the variable is well or badly distributed; in other words, there are no points of comparison to evaluate the degree of dispersion of the variable (e.g., what exactly does it mean to say that the standard deviation of the variable is equal to three or to one and a half?). Secondly, it is impossible to compare the dispersion of two variables measured in different magnitudes (e.g., if, in our example, drinks were studied , could we say that this second variable has greater dispersion than the first?).

All these problems can be overcome through the use of measures of relative (or according to Weisberg [1986]) normalized) variation, such as the well-known Pearson coefficient of variation. The aim of this paper is to analyze the problems of using this measure and to propose another coefficient which is free of these drawbacks.

Relative measures are those which are not expressed in any magnitude. Of these the best known are obviously percentages and proportions. Both are obtained with a ratio of similar quantities, and therefore, the magnitude of measurement is eliminated. For instance, when dividing people by people we obtain a magnitude comparable to the one resulting from dividing pesetas by pesetas, even if people are not comparable to pesetas. It is important to note that in a ratio of different magnitudes, both must be given, whereas this is not the case in a ratio of equal magnitudes as the result is a relative measure. A clear example is speed: when the number of

miles driven is divided by the time, the result is given in miles per hour. However, when dividing votes for a political party by the total number of votes cast in the election, the resulting quantity is not expressed in any magnitude.

Most statistics textbooks include a relative measure of variance, known as the Pearson coefficient of variation. This is defined as the ratio between the standard deviation and the mean. In order to see an example of its application, let us consider two variables in a population —age and number of children— and assume we are interested in stating which variable shows greater variation. The first is measured in years, the second people. It is clear that the corresponding means, variances or standard deviations are not comparable. For instance, consider the case of the Spanish capital, Madrid (see Table 2). The average age of the inhabitants of Madrid is 38.2, the standard deviation is 22.1, implying that it could somehow be said that the average deviation per inhabitant is of 22 years. As for the second variable (number of children), women over 15 have, on average, 1.5 children, with a standard deviation of 1.7. This standard deviation may seem to be quite high when compared with that for the variable age. However, when we use coefficients of variation, 58% for age and 116% for number of children, we reach the right conclusion -that is, the variable age has a smaller deviation than the variable number of children. It is important to note that the second figure, 116%, should make us wonder about the possible limits of this coefficient of variation.

Table 2. *Age and children in two Spanish cities*

	CITY									
	Madrid					Salamanca				
	N	Mean	St. Dev	CV	CV_b	N	Mean	St. Dev	CV	CV_b
AGE	3.113.818	38,2	22,1	58%	46%	162.737	37,2	22,6	61%	47%
CHILDREN	1.385.577	1,5	1,7	116%	56%	70.917	1,7	1,9	114%	59%

Source: INE.Sociodemografic Poll (1991). Elaborated by the author.

If these statistics are compared with those for another Spanish city, say Salamanca, the population is a bit younger and on average, women have more children. It is also possible to compare the measures of absolute and relative deviation in the two cities. For the variable age,

both the standard deviation and the coefficient of variation are smaller than in Madrid (which means that there will be fewer children and fewer elderly people). For the variable number of children, the absolute magnitude indicates greater dispersion in Salamanca whereas the relative magnitude shows greater dispersion in Madrid. It turns out that it is better to use this last statistic for the comparisons.

This coefficient of variation has the following properties:

(1) It is equal to zero for constant variables.

(2) It may take negative values for variables with negative means.

(3) It will take extremely large values if the mean is close to zero, and be infinite if the mean is zero.

(4) If a (positive) constant C is added to a variable, the coefficient decreases.

(5) If a variable is multiplied by a constant C, the coefficient does not change.

(6) The maximum value of the coefficient for a positive variable, that is, one without negative values, is $\sqrt{n-1}$.

It follows that it is advisable not to use the coefficient of variation for variables that take negative values or with means close to zero. Actually, this coefficient can only be used for those variables with values bounded from below at zero, the extreme situation being the case where all individuals except one have the value zero and the other one has the value $n \times \bar{x}$. To return to the example of the four chickens, this would happen when one person eats all the chickens and the other three eat none at all. In this case, the standard deviation is 3 and, since the mean is equal to one, the coefficient of variation is also 3, which yields a value of 173% in percentage terms.

One step that can be taken to prevent the maximum value of the coefficient of variation changing with the distribution would be to divide it by its maximum value $\sqrt{n-1}$. This would give a value of 100% instead of 173%, as the square root of n-1 is equal to the square root of 3. The value of 100% would indicate that this distribution has the maximum possible variance.

Thus, the adjusted coefficient of variation would be defined as

$$CV_{ad} = \frac{s}{\bar{x}\sqrt{n-1}} \quad (1)$$

However, this measure is very weak, especially when n is large and the hypothesis of an unlimited value for the variable cannot be sustained. For instance, in a poll of 1000 individuals with an average age of 35 and standard deviation of 20 years, when calculating the maximum coefficient of variation, it must be assumed that the age of 999 individuals is zero and the age of the other individual is 34965 = 999 x 35, which is of course completely unfeasible.

Another measure of relative variation could be obtained by dividing the standard deviation by the maximum possible deviation, as long as the variable takes values in a bounded interval, or in other words, as long as the range of the distribution is known. In this case, the most unfavorable situation would be that in which the variable takes the minimum value in half of the cases and the maximum value in the other half. If we assume that nobody can eat more than two chickens, then the range will be two and the situation in which two people eat two chickens each and the other two eat none will give maximum variance (Weisberg 1986:53). In this case the mean would be

$$\bar{X} = \frac{X_{\max} - X_{\min}}{2} \quad (2)$$

And the variance with maximum polarization would be as follows:

$$Var_{max} = \frac{(X_{max} - X_{min})^2}{4} \quad (3)$$

Nevertheless, this calculation of maximum deviation suffers from a major limitation: it assumes a constant mean, even if the mean of the distribution was not the equidistant point from the maximum and minimum value. As a result, this measure tends to overestimate the maximum variance of the distributions.

In this paper I try to develop a measure of maximum variance which takes into account the sample mean in variables with values in a bounded interval, and has all the desirable properties of the measures presented above.

Let us consider a bounded variable X which takes more than two values. The case of maximum variance would be that with precisely two values, the limits of the interval. Let us denote the mean of X from the proportions of the cases of the minimum and maximum values:

$$\left.\begin{aligned} p_{\min} + p_{\max} &= 1 \\ \overline{X} &= p_{\min} X_{\min} + p_{\max} X_{\max} \end{aligned}\right\} \tag{4}$$

So, it is easy to calculate the maximum variance

$$Var_{\max} = (X_{\max} - \overline{X})^2 p_{\max} + (X_{\min} - \overline{X})^2 p_{\min} \tag{5}$$

which, using basic algebra, can be simplified as follows:

$$Var_{max} = (X_{max} - \overline{X})(\overline{X} - X_{min}) \tag{6}$$

Table 3. *Comparison of maximum variances*

Media=4	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$	**Media=3**	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme left	1	0,50	0,50	4,5	Extreme left	1	0,67	0,67	2,667
Extreme right	7	0,50	3,50	4,5	Extreme right	7	0,33	2,33	5,333
			μ= 4,00	s^2= 9,00				μ= 3,00	s^2= 8,00
Max. var. (W)=	9				Max. var. (W)=	9			
Max. Variance=	**9**				Max. Variance=	**8**			

The use of this measure may be illustrated by taking a simple case where the sociological identification of an individual interviewed in an opinion poll is measured on a scale from 1 to 7. If the sample mean was 4, that is, the middle point, the situation of maximum polarization (or maximum variance) would arise if 50% of the people questioned answer 1 and the other 50% answer 7. In this case, formula 2 would give a maximum dispersion of 9 (Table 3). The distribution given on the left side of Table 3 shows that all the opinions deviate three points from

the mean in absolute terms. The distribution on the right side corresponds to a sample mean of 3; in this case, two thirds of the population would opt for 1 and only a third for 7, with a maximum dispersion of 8.

Once the maximum variance is known, the maximum standard deviation can be obtained by simply calculating the square root. It then becomes possible to obtain two interrelated measures: a proportion of the *bounded variance*, PV_b, which is the ratio of the empirical variance to the maximum variance conditional on the empirical mean, and a *bounded coefficient of variation*, CV_b, which is the ratio between the corresponding standard deviations. That is to say,

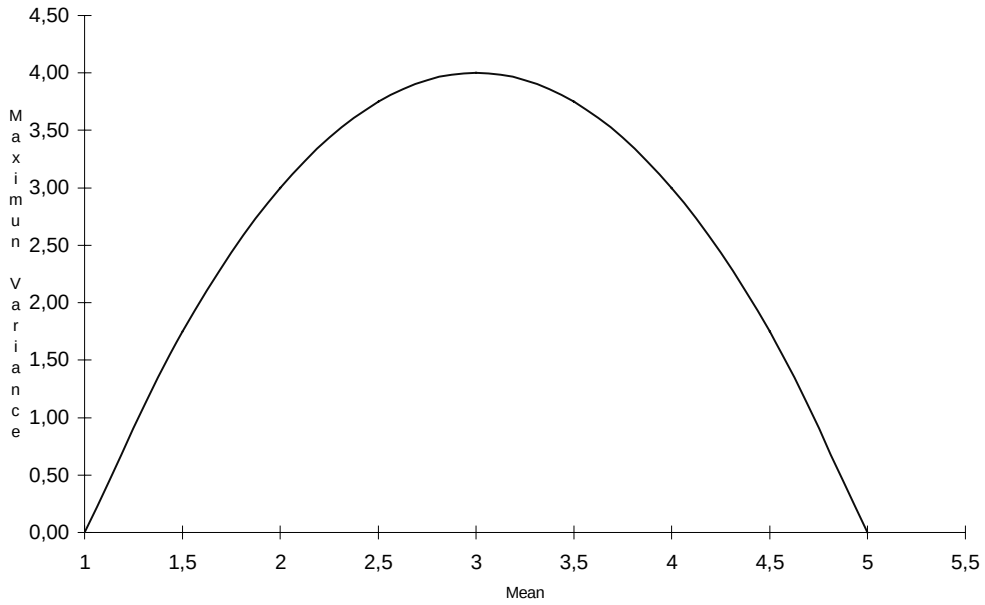
$$PV_a = \frac{s^2}{(\overline{X} - X_{\min})(X_{\max} - \overline{X})}$$
$$CV_a = \frac{s}{\sqrt{(\overline{X} - X_{\min})(X_{\max} - \overline{X})}} = \sqrt{PV_a} \qquad (7)$$

These two measures have the desirable property of taking values between 0 and 1. The minimum value is zero when the variable has zero variance, and equal to one when the variable takes only two values that are precisely the limits of the distribution. Moreover, they can be easily obtained from each other as (7) suggests.

Both measures, in particular the bounded coefficient of variation, given that it generally takes higher values, are useful for comparing the variability between measures with different scales of measurement. To return to the example of age and number of children in Madrid and Salamanca, the coefficient of variation was equal to 116% for the variable number of children, and to 58% for the variable age. Using the bounded coefficient of variation,[1] the percentages are 56% and 46% respectively in Madrid, and 59% and 47% in Salamanca, figures that correspond more closely to the general notion of percentage. Hence these results confirm the scant variation between the two cities.

[1] The minimum value is zero for both variables, whereas the maximum value is assumed to be 8 for number of children and 98 for age.

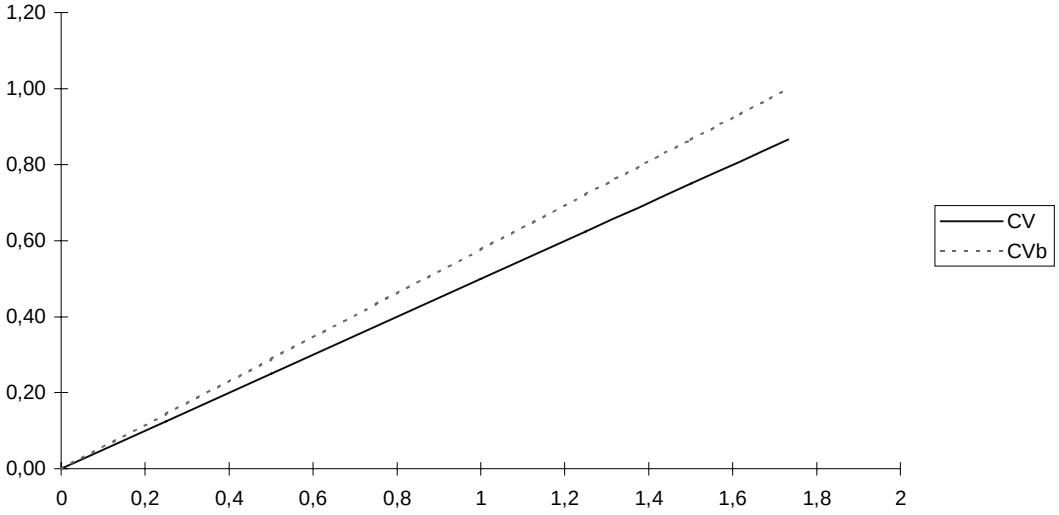
Figure 1.- Maximum variance by its mean



Moreover, unlike all the other measures of dispersion, these new coefficients have the attractive property of being invariant with respect to linear transformations. In other words, these coefficients do not change when a constant is added to the variable or the variable is multiplied by a constant. This is not the case for the standard deviation (or variance), which changes when the distribution is multiplied by a constant, or for the coefficient of variation, which varies with the addition of a constant. This property is particularly advantageous in the field of social sciences, in which there is a need for measures that are independent of the magnitude of the scale, since most measurement scales are arbitrary. For example, why measure ideology on a scale from 1 to 10, rather than 0 to 20? Likert scales, in which the items —usually measured from 1 to 5— are often inverted by means of the formula X'=6 - X and the traditional coefficient of variation changes depending on whether or not it is inverted, is a clear example.

The variability of this index in the context of 5-point Likert scales is analyzed in Figure 1. To start with, the denominator of the proportion of variance, that is, the maximum variance, is a parabolic function that is inverted with respect to the mean of the variable. It is logical to think that if on a scale of 1 to 5 the mean is 1, then the maximum variance will be zero since the variable must take the value 1 as it cannot take any value below 1. In the same way, the

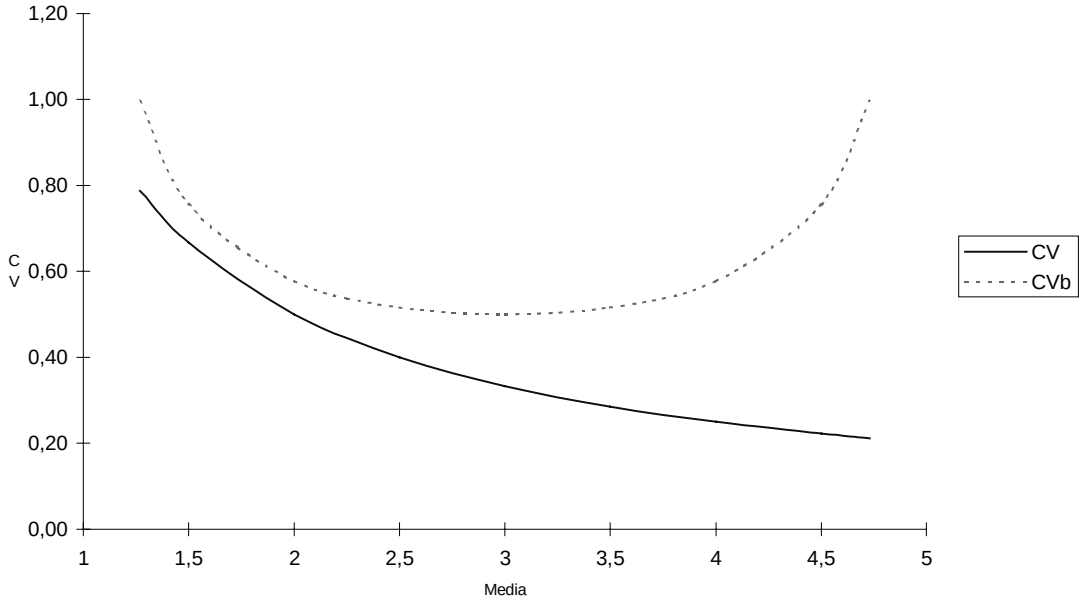
Figure 2.- Coefficients of variation by variance (Constant mean=2)



maximum variance will also be zero if the mean is 5, as the variable cannot take any value above 5. The maximum variance in the distribution corresponds to a mean of 3. In this case, 50% of the individuals will have the minimum value (1), and the other 50% will have the maximum value (5), which implies a variance of 4. With intermediate means, the maximum variance must be between 0 and 4. Thus, for example, if the sample mean is equal to 2, only with 75% of 1s and 25% of 5s (this is the only possible combination of 1s and 5s that yields a mean of 2) we would obtain maximum variance (equal to 3). If the mean were 4, the situation would be symmetrical with a maximum variance also equal to 3.

Figures 2 and 3 compare the bounded coefficient of variation and Pearson coefficient of variation. If the mean is constant, both coefficients show a similar pattern with respect to variance, the bounded coefficient usually being slightly higher. However, their evolution with respect to the mean is quite different. Thus, the range of oscillation is lower for the bounded coefficient, and while the Pearson coefficient decreases as the mean increases, the bounded coefficient oscillates symmetrically.

Figure 3.- Coefficients of variation by mean (Constant variance=1)



In order to facilitate the interpretation of these statistics, Table 4 presents eight hypothetical distributions of ideology with their corresponding means, standard deviations, and percentages of empirical deviations, grouped in pairs. The first pair corresponds to the case in which all the individuals in the population have the same ideology (4 in distribution A, 3 in distribution B). This implies variance equal to zero, a maximum deviation of 9 and 8 respectively, and a proportion of bounded variance, PV_b, also equal to zero. The situation described by distribution C has a mean of 4, with half the individuals located on the extreme left and the other half on the extreme right, that is, the population is split equally between the limit values of the distribution. In this case, the variance is maximum, so the PV_b is 100%. The same PV_b is obtained in distribution D, as although the mean is not 4, the variable only takes the limit values. The other four distributions represent less extreme situations. In cases E and F, 50% of the individuals are at the mean, and the rest are split equally at the limit values of the distribution. The proportion of bounded variation is now equal to 50%, and the bounded coefficient of variation is equal to 71%. In cases G and H, 50% of the individuals are again at the mean, but the rest take not only limit but also intermediate values. Thus, this bounded coefficient will decrease in a proportion that is always lower than the number of individuals with intermediate values. Thus, in G, 46%

of cases with respect to E have moved away from the limit values with a diminution in the CV_b of 21 percentage points, and the same is true for H where only 10% of the cases take the limit values and so, in comparison to F, 40% of the individuals have moved away from these limit values.

Table 4. *Simulated distributions of ideology*

Distribution A	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Center	4	1,00	4,00	0
			μ= 4,00 s^2=	0,00
PV_b= 0%		PH_a=		
CV_b= 0%				

Distribution B	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Center-left	3	1,00	3,00	0,00
			μ= 3,00 s^2=	0,00
PV_b= 0%				
CV_b= 0%				

Distribution C	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,50	0,50	4,5
Extreme Right	7	0,50	3,50	4,5
			μ= 4,00 s^2=	9,00
PV_b= 100%				
CV_b= 100%				

Distribution D	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,67	0,67	2,67
Extreme Right	7	0,33	2,33	5,33
			μ= 3,00 s^2=	8,00
PV_b= 100%				
CV_b= 100%				

Distribution E	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,25	0,25	2,25
Center	4	0,50	2,00	0
Extreme Right	7	0,25	1,75	2,25
			μ= 4,00 s^2=	4,50
PV_b= 50%				
CV_b= 71%				

Distribution F	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,33	0,33	1,33
Center-left	3	0,50	1,50	0,00
Extreme Right	7	0,17	1,17	2,67
			μ= 3,00 s^2=	4,00
PV_b= 50%				
CV_b= 71%				

Distribution G	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,02	0,02	0,22
Left	2	0,23	0,45	0,90
Center	4	0,50	2,00	0,00
Right	6	0,23	1,35	0,90
Extreme Right	7	0,02	0,17	0,22
			μ= 4,00 s^2=	2,25
PV_b= 25%				
CV_b= 50%				

Distribution H	x_i	f_i	$x_i f_i$	$(x_i-\mu)^2 f_i$
Extreme Left	1	0,06	0,06	0,24
Left	2	0,31	0,62	0,31
Center-left	3	0,50	1,50	0,00
Right	6	0,09	0,54	0,81
Extreme Right	7	0,04	0,28	0,64
			μ= 3,00 s^2=	2,00
PV_b= 25%				
CV_b= 50%				

The bounded coefficient of variation in distributions of probability

An important question regarding the use and interpretation of these coefficients is the value they adopt for different distributions of probability. Let us consider three of the most widely used distributions in statistics:

The distribution function of a uniform distribution in [a,b] is

$$f(x) = \frac{1}{b-a} \tag{8}$$

The parameters a and b are respectively the minimum and maximum possible values of the distribution.

Its mean and variance are equal to

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ Var(X) &= \frac{(b-a)^2}{12} \end{aligned} \tag{9}$$

Thus the maximum variance will be equal to

$$Var_{\max} = \frac{(b-a)^2}{4} \tag{10}$$

And, therefore, the proportion of bounded variation and bounded coefficient of variation will take the following values :

$$\begin{aligned} PV_a &= \frac{1}{3} = 33.3\% \\ CV_a &= \sqrt{\frac{1}{3}} = 57.7\% \end{aligned} \tag{11}$$

The probability function of a binomial distribution with parameters n and p is

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x} \tag{12}$$

The minimum and maximum possible values of this distribution are respectively 0 and n. Its mean and variance are equal to

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1 - p) \end{aligned} \tag{13}$$

Thus, the maximum variance will be equal to

$$Var_{\max} = (np - 0)(n - np) = n^2 p(1 - p) \tag{14}$$

And, therefore, the proportion of bounded variation and bounded coefficient of variation are functions of n, the number of times the binomial experiment is repeated with

$$\begin{aligned} PV_a &= \frac{npq}{n^2 pq} = \frac{1}{n} \\ CV_a &= \sqrt{\frac{1}{n}} \end{aligned} \tag{15}$$

The density function of the standard normal density (μ=0, σ=1) is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} \tag{16}$$

The calculation of the bounded coefficients is more problematic, since the normal distribution does not take values in a bounded interval but can in fact take any real value. To escape this problem, we may restrict the values of the distribution to the interval with limits above and below three times the standard deviation. Since this assumption does not apply in a mere 0.3% of the cases, the impact of the contribution of these exceptional cases to the final statistic can be presumed to be slight. Thus, the calculation of the bounded coefficients for the normal distribution is standard, with a PV_b of 11% and a CV_b of 33% .

$$CV_a = \sqrt{\frac{1}{(0-(-3))(3-0)}} = \sqrt{\frac{1}{9}} = 33\% \qquad (17)$$

These results can be used as a benchmark to compare the variability of the distributions used in practice. The distributions with a coefficient of bounded variation below 33% will show less deviation than the normal distribution, those with a coefficient of over 33% will show more deviation than the normal distribution, and when the coefficient exceeds 57.7% we may speak of greater deviation than the uniform distribution.

The limits of other coefficients of diversity

This is obviously not the only coefficient of diversity which ranges between 0 and 1,as a number of other such indexes may be found in the literature.

One of these is the well-known *Gini index*, an index of concentration which is widely used to measure economic inequality. It can be calculated through two different but interrelated formulae: one takes the sum of the difference between percentages of cases (p_i) and percentages of quantities ($q_i = x_i * p_i$), while the other consists of the sum of the absolute values of the differences between each value of the variable and the rest.

$$G = \frac{\sum_{i=1}^{k-1} P_i - Q_i}{\sum_{i=1}^{k-1} P_i}$$

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |X_i - X_j|}{2\overline{X}(n-1)n}, \; i \neq j \qquad (18)$$

As long as the variable takes positive values, this index takes values between 0 and 1. However, its values account for a different kind of deviation. Thus, although when all p_i are equal to q_i, that is to say, when the variable takes the same value for all individuals, the Gini index is equal to the minimum value (0) and so are the coefficients of variation, the situation changes for

the maximum value (1). The Gini index is only equal to 1 if the variable takes two different values and the minimum is 0, while the proportion of bounded variance does not require the minimum value to be zero in order to be equal to 1. Moreover, the Gini index adjusts quite badly when the variable takes negative values and, like the Pearson coefficient, it does not change when the variable is multiplied by a constant but it does change when a constant is added to the variable.

The same corrections made above for variance and standard deviation can be applied to this index. If the minimum value of the variable is other than zero,[2] the maximum value of the Gini coefficient would be obtained by

$$G_{\max} = 1 - \frac{x_{\min}}{\bar{x}} \tag{19}$$

The adjusted coefficient is defined as

$$G_a = \frac{G}{G_{\max}} \tag{20}$$

And this has the following properties:

1) It takes values between 0 and 1.

2) If a constant is added to a variable, the adjusted coefficient does not change.

Another index of dispersion with values between 0 and 1 is the *index of qualitative variation*, which is obtained from

$$IVC = \frac{(1 - \sum_{1}^{k} p_i^2)}{(k-1)k} \tag{21}$$

[2] It is assumed that there are no negative values.

This index only makes sense for non-metric variables. It measures the degree of concentration of individuals in certain categories of the variable and does not therefore show the distance between the values of the variable. The coefficient is equal to zero if in 100% of the cases the variable has the same value, and is equal to 1 if all the values of the variable have the same frequency. The disadvantage of this index is that it cannot be used correctly in the case of variables where the distance between different values matters. Thus, for a variable which takes the values strongly in favor of, in favor of , against and strongly against, the result would be the same if the only values found were strongly in favor of and strongly against as if the only values found were strongly in favor of and in favor of, when greater dispersion obviously exists in the first case.

Another coefficient that can be applied to nominal variables is that of *entropy*. This measure, which comes from the theory of information (Kripendorff 1986), represents the amount of uncertainty provided by a variable. If all the cases belong to the same category, then there is no uncertainty; if all the cases are equally split among the different categories, then there is total uncertainty.

The coefficient of entropy is given by

$$H = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \tag{22}$$

Its values are in the interval between the extremes of 0 and log_2(n). Therefore, a normalized coefficient with limits between 0 and 1 may be obtained by dividing by log_2(n). However, like the index of qualitative variation, this coefficient is not sensitive to the values that the variables may take; as a result, while it can be applied to nominal variables, it cannot be used in the case of the variables we are interested in, that is, quantitative variables with values in a bounded interval.

However, the statistical literature offers us another coefficient which is very suitable for the study of inequality of distributions: the *Theil index*. This coefficient is obtained from the coefficient of entropy as

$$H = \sum_{i=1}^{n} q_i \log_2 \frac{1}{q_i} \tag{23}$$

where the index of entropy is calculated from proportions of quantities rather than proportions of the variable.

Since the more unequal the distribution of the variable, the greater H will be, it must be inverted so that the Theil index takes values between 0 and log_2(n).

$$\begin{aligned} H' &= \log_2 n - H \\ &= \sum_{i=1}^{n} q_i \log_2 \frac{q_i}{p_i} \end{aligned} \tag{24}$$

Once again, this index can be normalized by dividing by log_2(n)

$$H'_N = \frac{H'}{\log_2 n} \tag{25}$$

It is important to note that the maximum value which H' can take for a variable with a mean of $\overline{X}$ and values in a bounded interval is

$$H'_{max} = \frac{X_{\min}(X_{\max} - \overline{X})\log_2 \frac{X_{\min}}{\overline{X}} + X_{\max}(\overline{X} - X_{\min})\log_2 \frac{X_{\max}}{\overline{X}}}{\overline{X} * (X_{\max} - X_{\min})} \tag{26}$$

It follows, therefore, that if the maximum, minimum and mean values of the variables are known, it is possible to obtain a coefficient with values between 0 and 1, in the same way as we obtained the bounded coefficient of variation. That is to say, by dividing the Theil index by this maximum value. The result is known as the *bounded Theil index*.

$$H'_a = \frac{H'}{H'_{max}} \tag{27.}$$

Let us take now a real example which illustrates the advantages of using this proportion instead of the (normalized) coefficient of entropy. This example uses data from the 1991 Census to calculate the distribution of the number of children given birth by women over 15 in Salamanca (see Table 5).

The average number of children per woman in Salamanca is 1.7. With this mean, a minimum value equal to 0 and a maximum value equal to 8,[3] the bounded Theil index is .44, which means that given the characteristics of the distribution, the value of this coefficient is 44% of the possible maximum value. If we had used the normalized Theil index, the result would have been .02. A detailed analysis of the distribution suggests that we should opt for the first measure rather than the second one.

Table 5. *Distribution of number of children among women over 15*

	xi	fi			Bounded Indexes
			Mean	1,7	
Salamanca	0	29497	Variance	3,6	34,3%
Mujeres >15	1	7861	St. Dev.	1,9	58,6%
	2	14379	ARA	2,3	34,3%
	3	8119	Theil	0,69	44,2%
	4	4871	Gini	0,39	38,9%
	5	2943	MD Gini	1,3	38,9%
	6	1624	Entropy	1,7	14,9%
	8	1624	Skewness	1,2	
		70917	Kurtosys	1,27	
			DM	0,81	30,6%

Source: INE.Sociodemografic Poll (1991). Elaborated by the author.

Finally, we should mention another statistical formula, *the average of ratios of advantage*, which is defined as the average of the ratios between quantity (q_i) and population (p_i) for every

[3] The final row of the distribution of frequencies represents the case of 7 or more children. The value 8 would be the mid point of this interval. A different value would have very little effect on the bounded Theil index. For example, note that H'_max=2.09 if X_max=7, =2.27 if X_max=8, =2.42 if X_max=9, =2.55 if X_max=10, =2.67 if X_max=11, and that when X_max takes any of these values, PH only varies between .14 and .15.

sector (i) of the distribution. In other words, the average of ratios of advantage is simply the Theil index without logarithms:

$$R_m = \sum_{i=1}^{n} q_i \frac{q_i}{p_i} \tag{28}$$

This coefficient varies between 1 (when the variable is split equally among all the cases), and n when the variable takes the same value in 100% of cases. For this reason, the normalized coefficient of the average of ratios of advantage has been defined as:

$$R_m^N = \frac{R_m - 1}{n - 1} \tag{29}$$

However, for the same reason discussed above with respect to the coefficient of variation, it is not possible to consider this adjustment for those variables with values in a bounded interval where the hypothesis of an unlimited value is meaningless. Hence, here we propose a different measure derived from the maximum value for this coefficient, which depends on the maximum, minimum and mean value of the distribution under consideration. This is

$$R_{m_{\max}} = \frac{(\overline{X}\, X_{\min}) + (\overline{X}\, X_{\max}) - (X_{\max}\, X_{\min})}{\overline{X}} \tag{30}$$

Thus, we may calculate the proportion of the average of ratios of advantage as the ratio

$$\frac{R_m}{R_{m_{\max}}} \tag{31}$$

As the minimum value which these two quantities may take is 1, it would be better to subtract one from both the numerator and denominator, and consider the ratio

$$RM_a = \frac{R_m - 1}{R_{m_{\max}} - 1} \tag{32}$$

And, surprisingly, we find the following algebraic relationship:

$$RM_a = PV_a \tag{33}$$

This means, for instance, in the last example, that the proportion of bounded variance was 34.3% and the proportion of average of ratios of advantage is also 34.3%.

Applications

The utility and advantages of the coefficients proposed here can be illustrated by applying them to two empirical examples.

The aim of the first example is to show how measures of inequality may often produce contradictory results, and therefore how particular care must be taken in choosing the appropriate measure if we wish to avoid misleading conclusions.

Table 6 has been obtained from the socio-demographic poll (1991), whose sample size allows us to perform a study of inequality in education by cohorts without sample errors. It is important to note that careful attention must be paid to the possible problems for the interpretation of the results resulting from differential mortality. The analysis of means shows a quite considerable increase in the number of years of schooling amongst Spaniards in the cohort aged over 75 years and the cohort of those aged between 25 and 29 years: from three and a half years to slightly over nine years. The increase is greater for women -from 3.1 to 9.5 years- than for men -from 3.5 to 8.9 years. In other words, in those cohorts born before 1956 men had more years of schooling than women, and this situation was reversed among younger cohorts.

Table 6. *Years in school by COHORT and gender*

	TOTAL POPULATION				MEN				WOMEN			
COHORT	Mean	St.D.	CV	CV_b	Mean	St.D.	CV	CV_b	Mean	St.D.	CV	CV_b
1962-1966 (25 - 29)	9,2	3,4	37,0%	34,1%	8,9	3,3	37,1%	33,2%	9,5	3,6	37,9%	36,0%
1957-1961 (30 - 34)	8,0	4,2	52,2%	42,8%	8,0	4,1	51,3%	41,8%	8,1	4,3	53,1%	43,8%
1952-1956 (35 - 39)	7,1	4,2	59,2%	43,9%	7,3	4,2	57,5%	43,6%	6,9	4,2	60,9%	44,2%
1947-1951 (40 - 44)	6,4	4,0	62,5%	42,9%	6,5	4,1	63,1%	43,8%	6,2	3,8	61,3%	41,1%
1942-1946 (45 - 49)	5,8	3,9	67,2%	43,0%	6,1	4,1	67,2%	44,5%	5,5	3,7	67,3%	41,4%
1937-1941 (50 - 54)	4,9	3,6	73,5%	41,9%	5,2	3,7	71,2%	42,2%	4,6	3,3	71,7%	39,2%
1932-1936 (55 - 59)	4,5	3,5	77,8%	41,9%	4,8	3,6	75,0%	42,1%	4,2	3,3	78,6%	40,5%
1927-1931 (60 - 64)	4,2	3,2	76,2%	39,3%	4,4	3,4	77,3%	41,0%	4,0	3,0	75,0%	37,5%
1922-1926 (65 - 69)	4,1	3,2	78,0%	39,6%	4,3	3,3	76,7%	40,2%	3,9	3,0	76,9%	37,9%
1917-1921 (70 - 74)	3,9	3,1	79,5%	39,1%	4,3	3,3	76,7%	40,2%	3,5	2,9	82,9%	38,2%
BEF. 1916 (75 AND OVER)	3,5	3,1	88,6%	40,8%	3,9	3,2	82,1%	40,4%	3,1	2,8	90,3%	38,7%

Source: INE.Sociodemografic Poll (1991). Elaborated by the author.

We can see that the standard deviation systematically increases, except in the case of the cohort aged 25 to 29. This group may include people still registered as students, either undergraduate or postgraduate, and therefore, we expect an increase in both its mean and standard deviation. This constant increase in the standard deviation is mainly due to the increase in the percentage of students at university, since if the mean rises above the mid point in a scale (the scale here has limits of 0 and 20 years, so the mid point is equal to 10), then the standard deviation also increases.

The coefficient of variation shows a dramatic decrease in the deviation in years of education from 88.6% for the cohort aged 75 years or over to the (possibly undervalued) 37% for the cohort aged 25 to 29 years. As the mean increases considerably more than the standard deviation, the coefficient of variation decreases. However, can we affirm that inequality in education has decreased as consistently and significantly as these figures suggest? Which statistic should be believed: the standard deviation or the coefficient of variation? The answer to this second question seems clear: it is better to use a relative coefficient than an absolute coefficient for comparisons. However, it could also be argued that as one coefficient measures absolute deviation and the other relative deviation, the decision to choose one or another depends on what exactly we want to compare.

The bounded coefficient of variation proposed here produces different results. This statistic shows a certain stability in inequality in education over the different cohorts. And it also shows

that if we abstract the mean of years of schooling from the analysis, this inequality does not decrease during the early years of the francoist dictatorship (1939-1975) but actually increased with respect to previous generations.

The bounded coefficient of variation can also be used to compare men and women: for those cohorts born before 1952, the variation in years of schooling is greater for men than for women. As most women left school at an early age, inequality is low for this group. Only with the incipient incorporation of upper middle-class women into middle and high levels of education in the 1960s does inequality among women begin to overtake that among men.

It could be argued that the fact that the bounded coefficient of variation does not vary among the different cohorts reveals that it is not very sensitive to deviation and, therefore, that it is not a very useful measure of deviation. I now wish to demonstrate that this is not the case, by considering a cross section of time series data for the same variable, as compiled by Mas, Pérez, Uriel and Serrano (1995) (see Table 7).

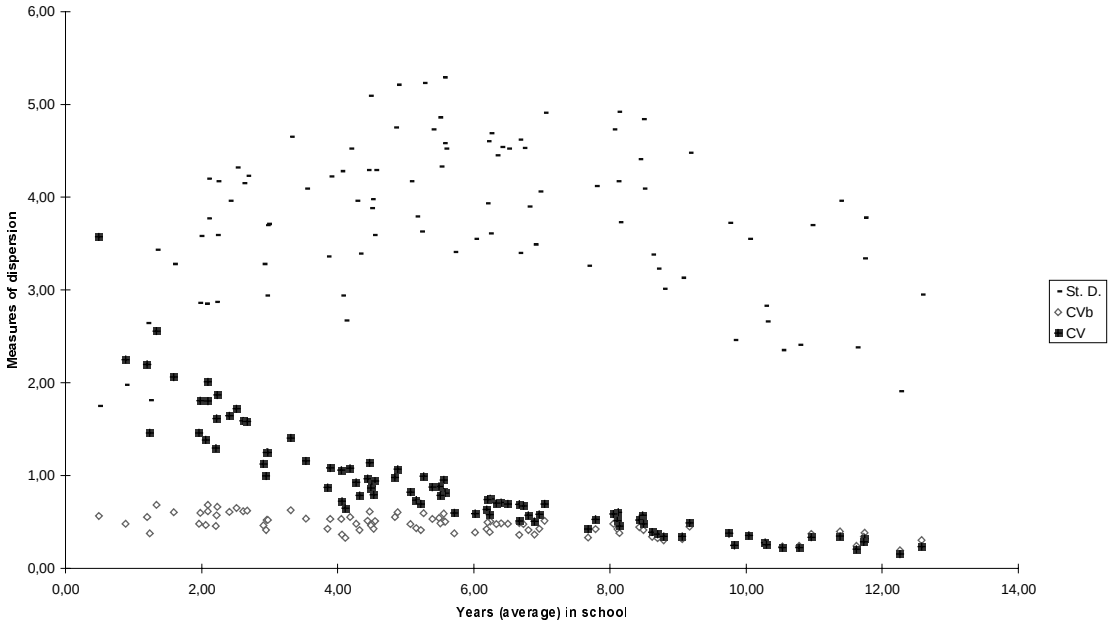
Table 7.- *Years in School. Spain. 1964-1992*

Year	Mean	St. D.	CV_b	CV
1964	4,9	2,7	31%	55%
1965	4,9	2,7	32%	55%
1966	4,9	2,7	32%	55%
1967	5,0	2,7	32%	55%
1968	5,0	2,8	32%	55%
1969	5,1	2,8	32%	55%
1970	5,2	2,9	33%	56%
1971	5,3	3,0	34%	56%
1972	5,4	3,0	34%	57%
1973	5,5	3,1	35%	57%
1974	5,6	3,2	35%	57%
1975	5,6	3,2	36%	57%
1976	5,8	3,3	36%	57%
1977	5,8	3,3	36%	57%
1978	5,9	3,3	37%	57%
1979	5,9	3,3	37%	56%
1980	6,0	3,4	37%	57%
1981	6,1	3,5	38%	57%
1982	6,2	3,5	38%	56%
1983	6,4	3,6	39%	56%
1984	6,5	3,6	39%	56%
1985	6,6	3,7	39%	56%
1986	6,7	3,7	39%	55%
1987	6,8	3,7	40%	55%
1988	6,9	3,8	40%	55%
1989	7,1	3,9	40%	55%
1990	7,2	3,9	40%	54%
1991	7,2	3,9	40%	54%
1992	7,3	3,9	41%	53%

Source: Mas, Pérez, Uriel y Serrano (1995)

The Table gives the mean, standard deviation, the bounded coefficient of variation and the classical Pearson coefficient for the whole population for each year. In this case, one obviously expects to find that deviation in years of schooling will increase as the new generations spend on average much longer in education than the mean. The bounded coefficient increases in a monotonic form from 31% to 41% over the 29 years considered in the analysis, and this increase is always less than one percentage point. In contrast, the classic Pearson coefficient does not follow a regular pattern.

Figure 4.- Measures of dispersion by years in school (n= 94 countries)



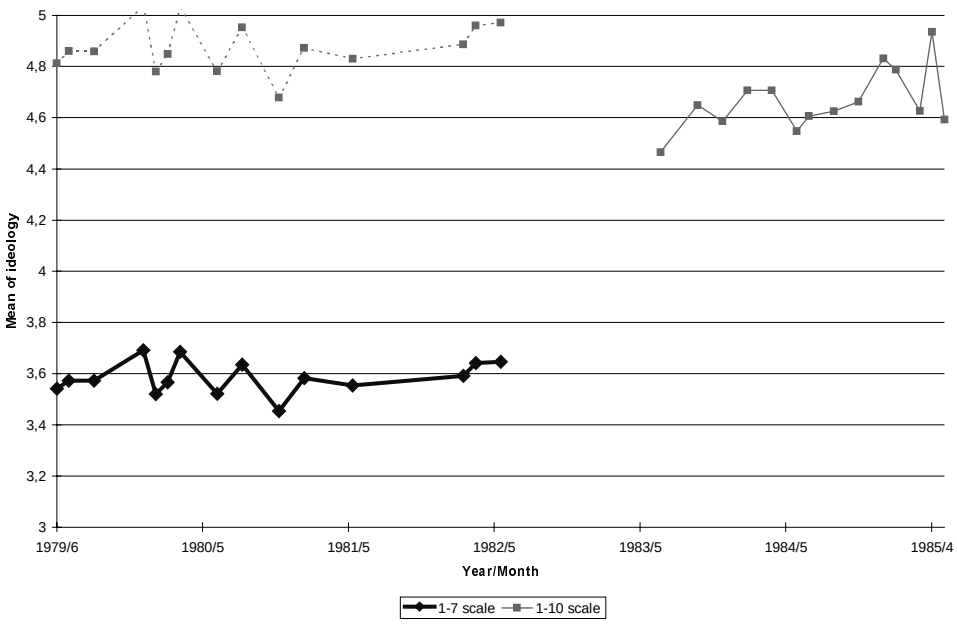
Source: Ram (1990). Elaborated by the author.

In this context it is interesting to analyze the data published by Ram (1990) in *The Review of Economics and Statistics* in 1990. In this paper he studies the empirical relationship between the means and standard deviations of years of schooling in 94 countries (see Figure 4). Since Ram uses a quadratic regression equation and the standard deviation for variables with values in a bounded interval decreases when the mean is between the limits of the distribution, his data may lead to misleading conclusions. Thus, to say that in developing countries an increase in the average level of education implies an increase in inequality in education is not strictly accurate. In fact, although there is an initial increase in inequality, as soon as the average years of schooling exceeds the mid point in the scale, there will be a decrease in inequality.

Moreover, the results would be different in terms of relative deviation. If we plot the coefficients of variation for the same countries in the same graph, it can be seen that (relative) inequality dramatically decreases inversely with respect to the mean. If instead we plot the bounded coefficients of variation (Figure 4), inequality in education now increases with the mean

following a tenuous inverse linear relationship that is not so clear at the beginning because of the upward bias of the coefficient of variation when the mean is close to zero.

Figure 5.- Trends of ideology



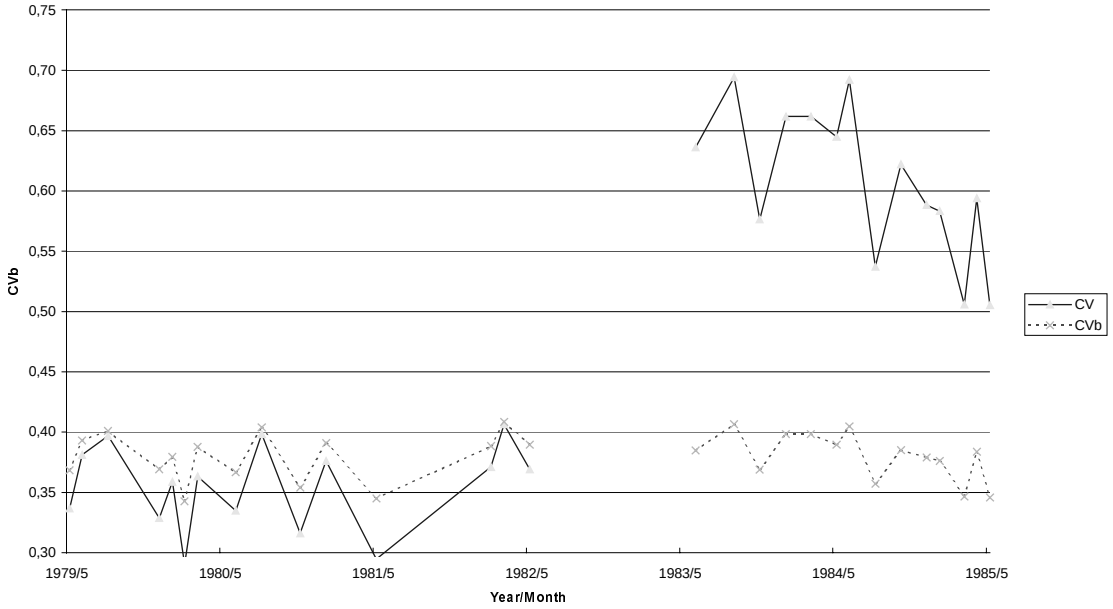
Source: CIS (1979-1985). Elaborated by the author.

The second example is taken from 29 studies carried out by the Spanish Center of Sociological Research (CIS). In these polls individuals were asked about their ideology. The measurement scales used a range from 1 to 7 in the first 15 studies (until June 1982) and from 1 to 10 in the last 14 studies. We wish to analyze the effect of these scales on the mean and coefficients of variation. Figure 5 shows the evolution of the mean. The thick line represents ideology measured before July 1983 on the 7-point scale. The thin line depicts the ideology measured after this date on the 10-point scale. The dotted line is the mean obtained before July 1983 with the 7-point scale[4].

[4] To do so, we have performed a linear transformation with parameters a=-0.5, b=1.5, so that both distributions have the same limits.

As was to be expected, there is a fictitious gap between the thin and thick lines due to the artificial change in the measurement scale. However, when the dotted line is compared to the thin one, we can see a slight decline to below the mean. This can be explained in two ways: firstly, by arguing that under the socialist government a large proportion of the population identified with the left; or, secondly, and more plausibly, that the 10-point scale confuses the Center (5) and Center-left as it is impossible to identify with the ideological position represented by that corresponding to the actual mid point on the scale (5.5).

Figure 6.- Trends of the coefficients of variation of ideology



Source: CIS (1979-1985). Elaborated by the author.

The main question here is to analyze the dispersion of the distribution. Our initial hypothesis is that neither the change in the political situation, nor the change in the measurement scale, should have any effect on the ideological polarization in the population. Figure 6 compares both coefficients of variation. The Pearson coefficient of variation is very sensitive to the change in the scale used; thus, it ranges in the [.3,.4] interval until 1982 and in the [.5,.7] interval after 1983. In contrast, the bounded coefficient of variation remains in the [.3,.4] range in both periods. Both coefficients follow a similar pattern in the second period, with wider variations for the Pearson coefficient. The difference between both periods raises suspicions with respect to the

coefficient of variation. The bounded coefficient of variation, on the contrary, seems to be more reliable when comparing distributions with different measurement scales.

In short, the paper began by highlighting the importance of the study of inequality in sociology. Its aim was to identify a family of relative measures of dispersion of use for comparing diversity over time and space. Beginning with the variance, a number of widely used statistics —the Gini coefficient, the Theil index and the average of ratios of advantage- were discussed, and their limit values calculated. They were then used as a denominator to obtain an adjusted measure with limits between 0 and 1, unlike most relative coefficients of dispersion in the statistical literature.

The idea, that can be generalized to a series of important statistics for the study of deviation, inequality and polarization, is robust in the sense that when applied to two different statistics it yields the same result. Two examples - inequality in education and polarization in ideology- have illustrated the practical advantages of using these coefficients. Further research is nonetheless required. This should focus in particular on the calculation of the standard error of these statistics and on the analysis of their welfare functions; as Atkinson (1970) has stated, all measures of inequality implicitly embrace a conception of justice.

REFERENCES

Atkinson, A.B. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2:244-263.

Blalock, H.M. 1991. *Understanding Social Inequality.* Newbury Park: Sage.

Bosch, A., C. Escribano, and I. Sánchez. 1989. *Evolución de la desigualdad y la pobreza en España. Estudio basado en las Encuestas de Presupuestos Familiares 1973-1974 y 1980-1981.* Madrid: INE.

Bossert, W., and A. Pfingsten. 1990. "Intermediate Inequality: Concepts, Indices, and Welfare Implications." *Mathematical Social Sciences* 19: 117-134.

Cortés, F., and R.M. Rubalcaba. 1984. *Técnicas estadísticas para el estudio de la desigualdad social.* México: El Colegio de México.

Coulter, P.B. 1984. "Distinguishing Inequality and Concentration." *Political Methodology* 10: 323-335.

Esteban, J.M., and D. Ray. 1993. "El concepto de polarización y su medición." *I Simposio sobre Igualdad y Distribución de la Renta y Riqueza*. Madrid: Fundación Argentaria.

Jacobson, H.I. 1969. "The Maximum Variance of Restricted Unimodal Distributions". *Ann. Math. Statist.* 10: 1746-1752.

Johnson, N.L., and C.A. Rogers. 1951. "Inequalities on moments of unimodal distributions." *Ann. Math Statist* 22: 433-439.

Leabo, D.A. 1976. *Basic Statistics.* Homewood: Irwin.

Kendall, M.G., and A. Stuart. 1973. *The Advanced Theory of Statistics.* London: Griffin.

Kolm, S.C. 1976. "Unequal Inequalities I." *Journal of Economic Theory* 12: 416-442.

_____ 1976. "Unequal Inequalities II." *Journal of Economic Theory* 13: 82-111.

Kotz, S., N.L. Johnson, and C.B. Read. 1983. *Encyclopedia of Statistical Science*. 8 vols. New York: John Wiley.

Krippendorf, K. 1986. *Information Theory.* Newbury Park: Sage.

Mas, M., F. Pérez, E. Uriel and L. Serrano. 1995. *Capital Humano. Series Históricas, 1964-1992.* Fundación Bancaja.

Ram, R. 1990. "Educational Expansion and Schooling Inequality: International Evidence and Some Implications." *The Review of Economics and Statistics:* 266-273.

Rayner, S.C.W. 1975. "Variance Bounds". *Indian Journal of Statistics* 37:135-138.

Ruiz Castillo, J. 1993. "Distribución personal de la renta: medición empírica y juicios de valor." *I Simposio sobre Igualdad y Distribución de la Renta y Riqueza*. Madrid: Fundación Argentaria.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North Holland Publishing.

Waldman, L.K. 1976. "Measures of party systems' properties: The number and sizes of parties." *Political Methodology* 3: 199-214.

Weisberg, H.F. 1986. *Central Tendency and Variability.* Newbury Park: Sage