AUSTRIAN JOURNAL OF STATISTICS
Volume 39 (2010), Number 4, 341–348

Interesting Properties of Variation Characteristics

Czesław Stepniak
Institute of Mathematics, University of Rzeszów, Poland

Abstract: As yet the statistical literature does not provide sufficient information on the upper and lower bounds of variation characteristics in descriptive statistics. Our aim is to fill this gap and to give precise answers to some questions regarding the problem.

Zusammenfassung: Bislang bietet die statistische Literatur keine ausreichende Information über obere und untere Schranken von Variationscharakteristika in der deskriptiven Statistik. Unser Ziel ist es diese Lücke zu schließen und präzise Antworten auf einige Fragen bezüglich dieses Problems zu geben.

Keywords: Descriptive Statistics, Location Characteristic, Variation Characteristic, Coefficient of Variation, Gini Index.

1 Introduction

At some university faculties, involving economy, descriptive statistics appear in a separate course treated as an introduction to statistical inference. It involves, among others, some measures of location, variation and asymmetry for collected data. However, as yet, the academic books in the subject do not provide sufficient information on the upper and lower bounds of these characteristics, or the information provided is not precise (in terms “usually”, “the most often” etc.). The aim of this paper is to fill this gap and give a precise answer to these questions.

The second problem concerns the statistical teaching and training. Our experience shows that some elegant formulae, being simple for interpretation, are not always convenient for computation and vice versa. So we suggest to work with two equivalent forms: one for theoretical consideration and the other one for computational purposes, respectively. In this way the consideration will be clear while the computation will be less faulty and time-consuming. In fact the whole work can be done by using a simple calculator. A special attention is given to the mean absolute deviation and to the Gini index.

2 Data Characteristics and their Attributes

Descriptive statistics refer to a finite sequence of observations, say $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, called data. The aim of descriptive statistics is a visual (i.e. graphical) and quantitative presentation of the data. In this paper we focus on the second one.

A real number $x_0 = f(x_1, \ldots, x_n)$ expressing a desired property of the data is said to be a *data characteristic*. In consequence any f may be a potential characteristic of a respective kind (for instance of location, variation or asymmetry). The members of each class may be chosen in two ways:

1. by defining some desirable attributes and then considering all functions for which the attributes are met,

2. by choosing from the characteristics being in a common use, for instance, from the sample moments, order statistics, or their possible combinations.

With regards to the first way, it seems that any reasonable data characteristic should be invariant with respect to a permutation of the observations, i.e. it should satisfy the initial condition

(I) $f(x_{i_1},\dots,x_{i_n}) = f(x_1,\dots,x_n)$

for any permutation $x_{i_1},\dots,x_{i_n}$ of the numbers $x_1,\dots,x_n$. We shall assume that the condition (I) is met for all data characteristics considered in this paper.

3 Location Characteristics

The characteristics of this type refer to a central point of the data. Attributes of such characteristics may be introduced either directly, or indirectly. Let us start from the first one. For a *location characteristic* the following attributes are desirable:

(L.1) $\min\{x_1,\dots,x_n\} \le f(x_1,\dots,x_n) \le \max\{x_1,\dots,x_n\}$ for all x

(L.2) $f(x_1+a,\dots,x_n+a) = f(x_1,\dots,x_n)+a$ for all x and for any a (transition)

(L.3) $f(cx_1,\dots,cx_n) = cf(x_1,\dots,x_n)$ for all x and for any positive c (positive homogeneity).

Of course the properties (I) and (L.1) – (L.3) do not determine the function f uniquely. Not only the usual mean and median, but also some other functions of order statistics possess these attributes. The set of potential location characteristics may be decreased by some additional attributes, for instance by

(L.4) $f(x_1+y_1,\dots,x_n+y_n) = f(x_1,\dots,x_n)+f(y_1,\dots,y_n)$

for arbitrary data $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$ (additivity).

We note that the conditions (L.1) – (L.4) are not independent because (L.1) and (L.4) imply (L.2).

Theorem 1. The unique data characteristic satisfying conditions (I), (L.1), (L.3) and (L.4) is the data mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

Proof. By (L.4), (L.3) and (I) we get the chain of equalities

$$\begin{aligned} f(x_1,\dots,x_n) &= f(x_1,0,\dots,0)+f(0,x_2,0,\dots,0)+\cdots+f(0,\dots,0,x_n)\\ &= x_1f(1,0,\dots,0)+x_2f(0,1,0,\dots,0)+\cdots+x_nf(0,\dots,0,1)\\ &= f(1,0,\dots,0)\sum_{i=1}^{n}x_i\,. \end{aligned}$$

In particular $f(a,\dots,a) = naf(1,0,\dots,0)$.

On the other hand, by (L.1), $f(a,\dots,a)=a$ and, therefore, $f(1,0,\dots,0)=1/n$. By setting this to the above equation we get the desired result. □

4 Characteristics Minimizing the Common Distance From Observations

For given data $x = (x_1, \ldots, x_n)$ and a given number x_0 the common distance from x_0 to x may be defined as the Euclidean distance between the points $P = (x_1, \ldots, x_n)$ and $P_0 = (x_0, \ldots, x_0)$ in $\mathbb{R}^n$, i.e. $\sqrt{(x_1 - x_0)^2 + \cdots + (x_n - x_0)^2}$.

Since $\sqrt{x}$ is nondecreasing in x, the common distance from x_0 to x is minimal, if and only if, $\sum_{i=1}^n (x_i - x_0)^2$ is minimal. In order to determine the point x_0 realizing this minimum we shall consider the function

$$f(\alpha) = \sum_{i=1}^n (x_i - \alpha)^2 .$$

Theorem 2. The function f attains its unique minimum for $\alpha = \bar{x}$.

Proof. We get the chain of equations

$$\begin{aligned}\sum_{i=1}^n (x_i - \alpha)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \alpha)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \alpha) \sum_{i=1}^n (x_i - \bar{x}) + n(\bar{x} - \alpha)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \alpha)^2 ,\end{aligned}$$

because $\sum_{i=1}^n (x_i - \bar{x}) = 0$. This implies the desired result. □

It seems that the common distance defined above, although easy to work with, is not natural. Instead, one can consider the usual (Euclidean) total distance from x_0 to all points $x_1, \ldots, x_n$ on the x-axis, i.e. $\sum_{i=1}^n |x_i - x_0|$.

We shall show that the minimum of this distance is realized if x_0 is a median of the data $x = (x_1, \ldots, x_n)$. This property of a median is known in the literature as minimizing the mean absolute deviation (MAD). The MAD of a median was derived in many papers and books, among others in Bickel and Doksum (1977) and Norton (1984), but these proofs were rather complicated. A first simple and elegant proof of this fact is due to Joag-Dev (1989). To present it let us consider the function

$$h(\beta) = \sum_{i=1}^n |x_i - \beta| . \tag{1}$$

Denote by $x_{(i)}$, $i = 1, \ldots, n$, the i-th order statistic, i.e. the i-th element in $(x_{(1)}, \ldots, x_{(n)})$ formed from the data $(x_1, \ldots, x_n)$ by sorting its elements in nondecreasing order, i.e. with $x_{(1)} \leq \cdots \leq x_{(n)}$.

Theorem 3. The function h attains its minimum for any β satisfying the condition $\beta = x_{((n+1)/2)}$ if n is odd, and β belongs to the closed interval $\left[x_{(n/2)}, x_{(n/2+1)}\right]$ if n is even.

Proof. Let us consider the intervals $I_i = \left[x_{(i)}, x_{(n-i+1)}\right]$ for $i = 1, \ldots, k$, where k is the integer part of $n/2$. Then $h(\beta)$ may be presented as

$$h(\beta) = \begin{cases} \sum_{i=1}^{k}(|x_{(i)} - \beta| + |x_{(n-i+1)} - \beta|), & \text{if } n \text{ is even,} \\ \sum_{i=1}^{k}(|x_{(i)} - \beta| + |x_{(n-i+1)} - \beta|) + |x_{((n+1)/2)} - \beta|, & \text{if } n \text{ is odd.} \end{cases}$$

Thus by the triangular inequality we get $h(\beta) \geq \sum_{i=1}^{k}(x_{(n-i+1)} - x_{(i)})$ with the equality, if and only if, $\beta \in I_k$ if n is even, while $\beta = x_{((n+1)/2)}$ if n is odd. This completes the proof of the Theorem. □

We note that in the even case the point β minimizing the function h is not uniquely determined. However, the values of the function (1) in all the points are the same. One of these points is the usual median $\operatorname{med}(x)$ defined by

$$\operatorname{med}(x) = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) .$$

As a direct consequence of Theorem 3 we get the following:

Corollary 1. The minimal value of the function h is given by the formula

$$\min h(\beta) = \sum_{i>(n+1)/2} x_{(i)} - \sum_{i<(n+1)/2} x_{(i)} ,$$

where $x_{(i)}$ is the i-th order statistic of the data x.

5 Variation Characteristics

Any nonnegative invariant function $f = f(x_1, \ldots, x_n)$ may be considered as a *variation characteristic* for the data $x = (x_1, \ldots, x_n)$ if it possesses the following attributes:

(V.1) $f(x_1 + a, \ldots, x_n + a) = f(x_1, \ldots, x_n)$ for all a (transition invariance)

(V.2) $f(cx_1, \ldots, cx_n) = |c| f(x_1, \ldots, x_n)$ for all c (nonnegative homogeneity).

From (V.1) and (V.2) one can derive the following property

(V.3) $f(a, \ldots, a) = 0$ for any scalar a.

One can easily verify that, among others, the following characteristics satisfy the above requirements:

standard deviation $s(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$,

mean absolute deviation $d(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \operatorname{med}(x)|$,

half average deviation between observations $g(x) = \frac{1}{2n^2}\sum_{i,j=1}^{n}|x_i - x_j|$.

The property (V.2) means that the characteristic f depends strongly on the scale, i.e. on the measurement unit of the observations. Very often we are interested in variation characteristics which are independent on the scale, i.e. instead of (V.2), they satisfy the condition

(V.4) $f(cx_1, \ldots, cx_n) = f(x_1, \ldots, x_n)$ for all $c \neq 0$.

To meet this condition we introduce the following notion.

Definition 1. Data $x = (x_1, \ldots, x_n)$ is said to be nonnegative if all observations x_i, $i = 1, \ldots, n$, are nonnegative and at least one of them is positive.

In this section we shall assume that any data under consideration is nonnegative. For such a data a characteristic independent on the scale may be obtained from arbitrary variation characteristic by dividing it over the data mean $\bar{x} > 0$. Applying this rule to the above variation characteristics we get the following characteristics independent on the scale:

coefficient of variation $v(x) = \frac{s(x)}{\bar{x}}$,

relative mean absolute deviation $r(x) = \frac{d(x)}{\bar{x}}$,

relative half average deviation between observations

$$G(x) = \frac{1}{2n^2\bar{x}} \sum_{i,j=1}^{n} |x_i - x_j| \,. \tag{2}$$

Now we are going to present an interesting relation between G and the well known Gini index. The Gini index is usually introduced in the following way.

For nonnegative data $x = (x_1, \ldots, x_n)$ consider its cumulate sums $s_1, \ldots, s_n$, i.e. the sequence of numbers defined by the formula $s_k = s_k(x) = \sum_{i=1}^{k} x_{(i)}$, where $x_{(i)}$ is the i-th order statistic.

Definition 2. Denote by P_1 the area of the convex polygon with vertices $(0,0), (1, s_1), \ldots, (n, s_n)$ and by P the area of the triangle with vertices $(0,0), (n,0)$ and (n, s_n). The ratio P_1/P is called *Gini index* of the data x.

We observe that if all observations x_i, $i = 1, \ldots, n$, are equal then $P_1 = 0$, and, in consequence, the Gini index is 0. On the other hand, if $x_1 = \cdots = x_{n-1} = 0$ and $x_n > 0$ then the Gini index is equal to $(n-1)/n$. Thus, we get the following conclusion.

Conclusion. The Gini index takes the values from the interval $[0, (n-1)/n]$.

However, the above definition of Gini index is not convenient for computation. We shall derive a more useful formula and show that it coincides with G.

Theorem 4. The Gini index coincides with the sample characteristic G defined by (2) and it can be expressed in the form

$$G(x) = \frac{\sum_{i=1}^{n} (2i - n - 1) x_{(i)}}{n \sum_{i=1}^{n} x_{(i)}} \,. \tag{3}$$

Proof. We observe that

$$2P_1 = ns_n - 2\sum_{i=1}^{n-1} s_i - s_n$$

and

$$2P = n \sum_{i=1}^{n} x_i \,.$$

Now by the identity

$$\sum_{i=1}^{n-1} s_i = \sum_{i=1}^{n} (n-i) x_{(i)}$$

we get

$$2P_1 = (n-1)\sum_{i=1}^{n} x_{(i)} - 2\sum_{i=1}^{n} (n-i)x_{(i)} = \sum_{i=1}^{n} (2i-n-1)x_{(i)} \, .$$

Therefore the Gini index may be presented in the form (3) and it remains to show that it coincides with G defined by (2). In fact we only need to verify that

$$\sum_{i=1}^{n} (2i-n-1)x_{(i)} = \frac{1}{2}\sum_{i,j=1}^{n} |x_i - x_j| \, . \tag{4}$$

We shall prove it by induction with respect to n.

For $n = 2$ the formula (4) may be verified directly. Now suppose it holds for $n = k$. We shall show that it also holds for $n = k+1$. Without loss of generality one can assume that $x_{(k+1)} = x_{k+1}$. Then, by the inductive assumption,

$$\begin{aligned} \sum_{i=1}^{k+1} (2i-(k+1)-1)x_{(i)} &= \sum_{i=1}^{k} (2i-k-1)x_{(i)} + kx_{(k+1)} - \sum_{i=1}^{k} x_{(i)} \\ &= \frac{1}{2}\sum_{i,j=1}^{k} |x_i - x_j| + \sum_{i=1}^{k} |x_{(k+1)} - x_{(i)}| \\ &= \frac{1}{2}\sum_{i,j=1}^{k+1} |x_i - x_j| \, . \end{aligned}$$

This implies the identity (4) and, in consequence, the statement of Theorem 4. □

By simple operations on the formula (3) we get two additional expressions for the Gini index G. We shall collect all these results in the form of the following theorem.

Theorem 5. For any nonnegative data $x = (x_1, \ldots, x_n)$ the following identities hold

$$\begin{aligned} \frac{1}{2n^2\bar{x}} \sum_{i,j=1}^{n} |x_i - x_j| &= \frac{\sum_{i=1}^{n} (2i-n-1)x_{(i)}}{n\sum_{i=1}^{n} x_i} \\ &= \frac{2\sum_{i=1}^{n} i x_{(i)}}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n} \\ &= \frac{2\sum_{i=1}^{n} i(x_{(i)} - \bar{x})}{n^2\bar{x}} \, . \end{aligned}$$

Proof. The first identity was just proved in Theorem 4 and the second one is evident. For the third one we only need to observe that $2\bar{x}\sum_{i=1}^{n} i = (n+1)\sum_{i=1}^{n} x_{(i)}$. □

6 Bounds of some Variation Characteristics

In this section we shall restrict our attention to nonnegative data, i.e. to the case when all observations are nonnegative and at least one of them is positive. A natural question is whether the coefficients $v(x), r(x)$ and $G(x)$ presented in Section 5 are normalized, i.e. whether they belong to the interval $[0,1]$ and whether the values 0 and 1 are attained. The question about the first coefficient was negatively answered by Stepniak (2007). This result may be stated in the following form:

Theorem 6. The attainable upper bound for the coefficient of variation $v(x)$ for nonnegative data $x=(x_1,\dots,x_n)$ is $\sqrt{n-1}$.

Here we shall present an alternative proof of this theorem. For this aim we need an auxiliary result.

For arbitrary integers k and l, such that $1\le k<l\le n$ and a positive ε less or equal to $x_{(l)}-x_{(k)}$ (if such exists), define the operation $y(x)=y(x;k,l,\varepsilon)=(y_1,\dots,y_n)$, where

$$y_i=\begin{cases} x_{(i)}, & \text{if } i\neq k,l,\\ x_{(k)}+\varepsilon, & \text{if } i=k,\\ x_{(l)}-\varepsilon, & \text{if } i=l.\end{cases}$$

It is evident that $\bar{y}=\bar{x}$. We shall prove

Lemma 1. Under the above assumption, $\sum_{i=1}^n (y_i-\bar{y})^2\le\sum_{i=1}^n (x_i-\bar{x})^2$.

Proof. We observe that

$$\begin{aligned}\sum_{i=1}^n (x_i-\bar{x})^2-\sum_{i=1}^n (y_i-\bar{y})^2 &= x_{(k)}^2+x_{(l)}^2-(x_{(k)}+\varepsilon)^2-(x_{(l)}-\varepsilon)^2\\ &= 2\varepsilon(x_{(l)}-x_{(k)}-\varepsilon)\,.\end{aligned}$$

Now the desired result follows by the assumption about ε. □

Proof of Theorem 6. For nonnegative data $x=(x_1,\dots,x_n)$ with mean $\bar{x}$ consider the auxiliary data $t=(t_1,\dots,t_n)=(0,\dots,0,n\bar{x})$. It is clear that $\bar{t}=\bar{x}$.

First we shall show that $\sum_{i=1}^n (t_i-\bar{t})^2=n(n-1)\bar{x}^2$. Notice that

$$\sum_{i=1}^n (t_i-\bar{t})^2=(n-1)\bar{x}^2+(n\bar{x}-\bar{x})^2=(n-1)\bar{x}^2+(n-1)^2\bar{x}^2=n(n-1)\bar{x}^2\,.$$

Now let us observe that x may be obtained from t by a finite number of operations of type $y=y(x;k,l,\varepsilon)$. Thus, by Lemma 1,

$$v(x)=\frac{s(x)}{\bar{x}}\le\frac{s(t)}{\bar{x}}=\frac{\sqrt{(n-1)\bar{x}^2}}{\bar{x}}=\sqrt{n-1}\,,$$

the upper bound follows. □

From Theorem 6 we get the following corollary.

Corollary 2. The ratio $v(x)/\sqrt{n-1}$ is a normalized coefficient of variation.

Now one can ask about the upper bound of the relative mean absolute deviation $r(x)$. It is clear that for given n and a positive constant c the ratio

$$\frac{\frac{1}{n}\sum_{i=1}^{n}|x_i - c|}{c}$$

is not bounded within all data $x = (x_1, \dots, x_n)$. However, we get

$$r(x) = \frac{1}{n\bar{x}}\sum_{i=1}^{n}|x_i - \mathrm{med}(x)| = \frac{1}{n\bar{x}}\left(\sum_{i>(n+1)/2} x_{(i)} - \sum_{i<(n+1)/2} x_{(i)}\right) \leq \frac{1}{n\bar{x}}\sum_{i=1}^{n} x_i = 1\,.$$

In this way we have proved

Theorem 7. The lower bound of $r(x)$ equals zero and is attained if $x_1 = \cdots = x_n$ while the upper bound equals 1 and is attained e.g. if $x_1 = \cdots = x_{n-1} = 0$ and $x_n > 0$, respectively.

Thus the characteristic $r(x)$ is normalized.

7 Conclusions

We have demonstrated some properties and connections between sample characteristics. These results can be treated as the first step towards formalization of descriptive statistics.

Acknowledgements

The author thanks the reviewer for his suggestions concerning some improvements in the presentation of this paper.

References

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics*. San Francisco: Holden-Day.

Joag-Dev, K. (1989). MAD property of median. A simple proof. *The American Statistician*, *43*, 26-27.

Norton, R. M. (1984). The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Statistician*, *38*, 135-136.

Stepniak, C. (2007). An effective characterization of Schur-convex functions with applications. *Journal of Convex Analysis*, 103-108.

Authors addresses:

Czesław Stepniak
Institute of Mathematics
University of Rzeszów
Rejtana 16 A
35-959 Rzeszów, Poland

E-mail: stepniak@umcs.lublin.pl