ANALYSIS OF REPARTITION FORM IN DISTRIBUTIONS OF FREQUENCIES

Associate Professor Ph.D. Elisabeta R. ROŞCA
„Stefan cel Mare" University of Suceava, Romania
elisabetar@seap.usv.ro

Abstract:

The paper „The Analysis of Repartition Form in the Distributions of Frequencies" presents an introduction in the study of distribution of frequencies, which have some characteristic features by the nature and the group degree of data, by which existing classifications and variation series, unidimension and multidimension series etc. The paper emphasizes that the statistical analysis of series of distributions is based on the comparison of the empirical distributions with theoretical distributions, on the comparison of parameters of theoretical distributions with the computed values by using the empirical distributions data. The paper presents the indicators of interquartilical and interdecilical variation, the principal methods of asymmetry analysis such as the graphic method and the analytic method based on the estimation and the interpretation of asymmetry coefficients established as relations between the indicators of central tendency or between the individual values and central tendency indicators . Also are presented some aspects concerning the computing and analysis the indicators of peakedness/arching.

Key words: interquartilical and interdecilical variation, symmetrical and asymmetrical series, coefficients of asymmetry, absolute and relative densities of frequencies, degrees of peakedness.

INTRODUCTION

In the analysis of the economic and social phenomena from the whole used statistical methods are remarked the group method which allows to form the *statistic series*. The nature and the systematization degree of the presented data are elements which differentiate the statistic series, those than are the object of this study being the distributions of frequencies. The analysis of distributions of frequencies has a particular character after how these presents the variation of a qualitative characteristic under the form of classifications or the variation of a quantitative characteristic under the form of variation series, after how these result from a simple group forming one dimension series or from a combined group forming multi dimension series. Generally, the analysis of statistic series are based on the theoretical distributions, elaborated on the basis of a hypothesis of distribution of frequencies so that to be possible to establish mathemat ics relations well determined between the studied values and their frequencies of apparition, interpreted as a function of probability and also based on empirical distributions obtained from the observed data using the absolute and relative frequencies, the comparison of the two forms being achieved on the base of parameters of the theoretical distribution.

The statistical analysis of series of distributions consists of the determination of typical values of series, the explanation of the level differences between the individual terms and typical values with the aim of interpretation the form and the degree of variation of the studied characteristic and realization of the dynamical and territorial comparisons. The problems determined by the variability, the distribution form, the homogeneity, the independence or interdependence of series terms are analysed by estimation and interpretation of the average, variation and asymmetry indicators, by analysis of variance etc. In the analysis of distributions of frequencies the asymmetry study is a part from the problems of form of distributions analysis, near by the indicators of interquartilical and interdecilical variation and the indicators of peakedness/arching. [6], [3]

INTERQUARTILICAL AND INTERDECILICAL V ARIATION

In a perfect symmetrical series the three basic indicators of centrale tendency: the average - $\bar{x}$, the mode - M_o, the median - M_e occupied the same place, between they being a relatio n of equality, such as:

$$\bar{x} = M_o = M_e \tag{1}$$

If we compute the deviations between the values of position means and the median we can interpret the tendency of distribution for frequencies of characteristic variants. [1], [4]

In the series in which are compute the values of quartiles, the deviation between the inferior quartile and the median is equal with the deviation between the superior quartile and median and inside of them there are 50% from the number of individual values.

In a perfect symmetrical series this equality can be wrote:

$$M_e - Q_1 = Q_3 - M_e \tag{2}$$

In this situation the arithmetic means of the two extreme quartiles is equal with the value of the second quartile or with the median of series, such as:

$$\overline{Q} = \frac{Q_1 + Q_2}{2} = Q_2 = M_e \tag{3}$$

When the two relations didn't verified, which we can write: $M_e - Q_1 \neq Q_3 - M_e$ and respective $\overline{Q} = M_e$ result that the series presents a certain degree of interquartilical variation which must be statistical measured.

The indicators of interquartilical variation can be computed using absolute and relative indicators.

The interquartilical deviation (Q_d) is computed as a mean of the two deviations of the extreme quartiles comparative with the centrale quartile such as:

$$Q_d = \frac{(M_e - Q_1) + (Q_3 - M_e)}{2} = \frac{Q_3 - Q_1}{2} \tag{4}$$

As an absolute indicator, the interquartilical deviation is expressed in the unit of measure of the studied characteristic and it isn't used in the direct comparison between more statistic series. In this last aim in statistics can be compute the interquartilical deviation coefficient.

The interquartilical deviation coefficient V_q is computed as a ratio between the interquartilical variation and the median value such as :

$$V_q = \frac{Q_d}{M_e} = \frac{\frac{Q_3 - Q_1}{2}}{2M_e} \tag{5}$$

This coefficient take values in the interval [0, 1] and it is appreciated that the interquartilical deviation is more significant as its value is increased.

When the series presents a great degree of asymmetry is nec essary to compute also the interquartilical variation.

The indicators of interdecilical variation are based on the same grounds as in case of interquartilical variation, which means that in a perfect symmetrical series the distances between the extreme deciles and median are equal such as:

$$M_e - D_1 = D_9 - M_e \qquad (6)$$

The interdecilical deviation is equal with the arithmetical average of the deviation of extreme deciles to the centrale quartile of series such as:

$$V_d = \frac{D_d}{M_e} = \frac{\frac{D_9 - D_1}{2}}{M_e} \qquad (7)$$

As a rule, the computing of interdecilical variation is applied for the statistical series with a great number of groups and with an obvious tendency of asymmetry.

The value of this coefficient is sub-unit and positive as in case of interquartilical var iation.

SYMMETRICAL AND ASYMMETRICAL SERIES

In the statistics, the series are the result of application of group method according one or more characteristics, obtaining two lines of data, in which the first line represents the variation of group characteristic and the second line is the result of appeared frequencies centralization or of the values of other characteristic with which the first is correlated. In these conditions, the statistics series can be considered a mathematical function in which the s ystematized values of frequencies or the values of the characteristics are dependent on the values of group characteristics. Being a mathematical function, the statistical series permits the graphic representation and the analysis of distribution form of frequencies according with their line of distribution or in comparison with theoretical distributions.

The practice of economical and social statistics had demonstrated that there are symmetrical distributions, easy asymmetrical distributions and distributi ons with pronounced tendency of asymmetry.

In case of a symmetrical variation in comparison with the central value of characteristic, the compensation of the deviations is made also on the whole series and inside of it as a result of the appeared frequencies of variants are equal on the both sides of central value (Figure 1a). If the appeared frequencies of variants didn't follow this regularity this means that the series presents an asymmetrical tendency to the greatest values of characteristic (Figure 1b) or to the smaller values of the characteristic (Figura 1c). For the description of the degree of asymmetry are compared the values of the three indicators of central tendency: the average, the median and the mode, as they result from Figure 1. [2], [5]

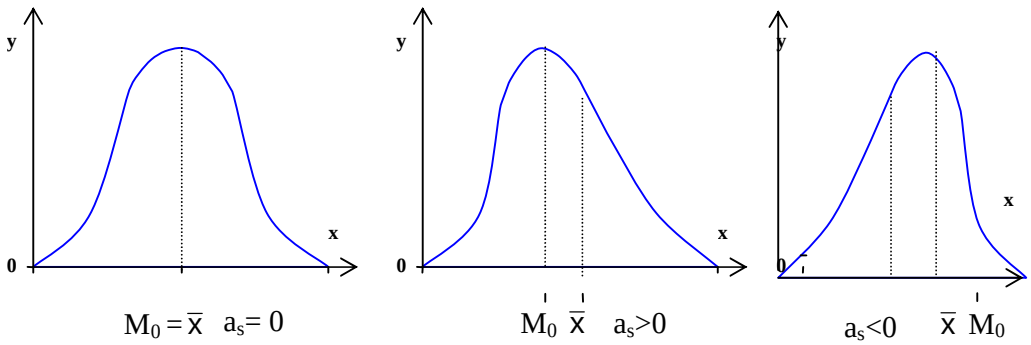

Figure no. 1. Symmetrical and asymmetrical distributions

a) Symmetrical distribution — Distribution is skewed to the right — Distribution is skewed to the left

The three presented cases emphasize that the degree of representation of the average is increasing for a symmetrical distribution and it has a more reduced field of variation of characteristic. The knowledge of the intensity of asymmetry supposes the application of the anal ytic method by the computing of specific indicators and their interpretation according with the principal properties of normal distributions.

THE ESTIMATION AND THE INTERPRETATION OF THE ASYMMETRY COEFFICIENTS

The analysis of distribution form is achievi ng using the graphic method or the analytic method and by computing the coefficients of asymmetry. In the case of application of graphic method are used the histogram, the polygon of frequencies and the cumulative curve of frequencies. The polygon of frequencies is used for the graphic representation of distributions of frequencies. The graphic representation supposes that on the axis Ox will be represented the centres of intervals which in the case of equal intervals is distributed at equal distances on th e axis. In this case on the graphic will be represented the line which unite the points with co -ordinates given by the values of centres of intervals and their frequencies. [3], [6], [10]

If the polygon of frequencies will be designed on the basis of histo gram, than it is obtained by union the centres of columns. If the group intervals are unequal than the frequencies are reduced proportional and than the results are represented on the graphic by the dimension of intervals.

A histogram for an equal interval s distribution was used for the data concerning the employment with professional status employee, by age groups in Romania in 2006 from the Table 1 and was obtained the graphic from Figure 1.

Table no 1. Employment (professional status employee), by age g roups in Romania, in 2006

Groups of employment by age	Employment (n_i) [persons]	Cumulative relative frequencies [%]	
		Increasing	Decreasing
0	1	2	3
15-24	481.026	481.026	6.167.000
25-34	1.874.768	2.355.794	5.685.974
35-44	1.825.432	4.181.226	3.811.206
45-54	1.566.418	5.747.644	1.985.774
55-64	407.022	6.154.666	419.356
65 ani şi peste	12.334	6.167.000	12.334

Souce: *Romanian Statistical Yearbook*, National Institute of Statistics, Bucharest, 2007

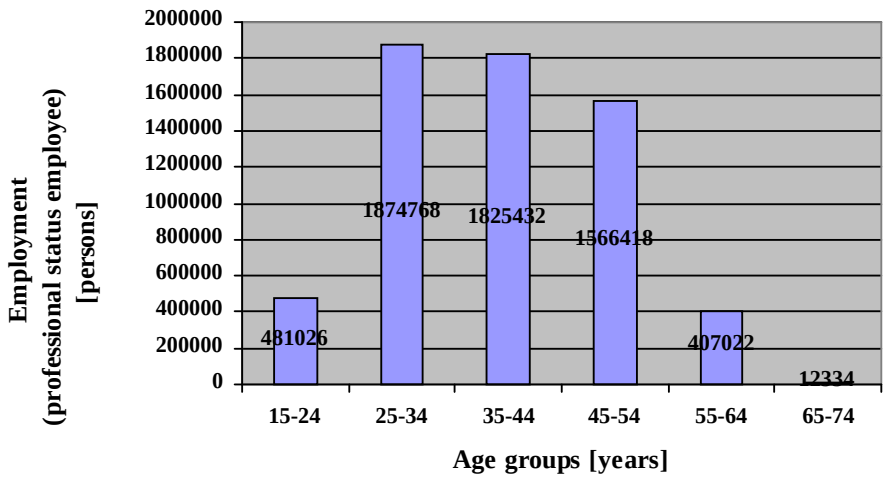

Figure no. 1. Employment (professional status employee), by age groups in Romania, in 2006

For the same data, the drawing of cumulative curve of frequencies supposes the successive totalizing of frequencies in both senses of distribution, the result being the graphic from Figure 2.

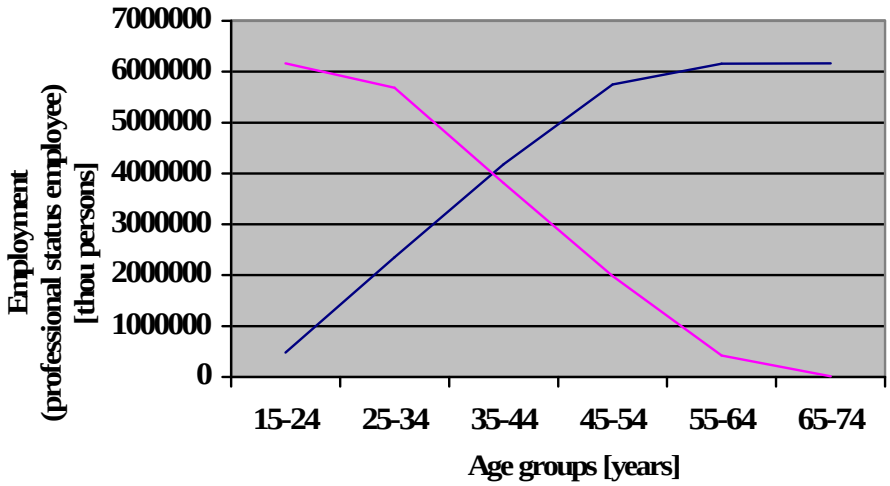

Figure no. 2. Cumulative curve of frequencies for employment (professional status employee), by age groups in Romania, in 2006

By the graphic drawing obtains a suggestive image about the asymmetry but not a quantitative value as a measure of the degree of asymmetry. [12] To establish the type of asymmetry but without to give emphasis to intensity of this we are computing the centrate moment of three orders using the formula:

$$\mu_3 = \frac{\sum_{i=1}^{12}(x_i - \bar{x})^3 \cdot n_i}{\sum_{i=1}^{12} n_i} \qquad (8)$$

In comparison with the sign of the centrate moment of three order we can appreciate the type of asymmetry. So we can distinguish three cases:

- if $\mu_3 \pi 0$ than the distribution is skewed to the right;
- if $\mu_3 = 0$ than the distribution is a symmetrical one (for a symmetrical distribution all the centrate moments of odd order are null);
- if $\mu_3 \phi 0$ than the distribution is skewed to the left.

For computing the intensity of asymmetry are used also the indicators of asymmetry, expressed in absolute and relative dimensions. A first information concerning with the degree and the sense of asymmetry is based on the computing values for the density of distribution of frequencies, computing by comparison each of frequency and the dimension of proper interval of variation.

The absolute density of frequencies (d_a) is obtained as a report between the absolute frequency (n_i) at the size of interval (h), according with the formula:

$$d_a = \frac{n_i}{h} \qquad (9)$$

The relative density of frequencies (d_r) is obtained as a report between the relative frequency (n_i^*) and the size of interval (h), according with the formula:

$$d_r = \frac{n_i^*}{h} \tag{10}$$

On the base of density of distribution we can establish analogies with the density of probabilities, considered as a theoretical model for the empirical density. If the values of these indicators show an increased tendency to the central value of characterist ic, this means that the distribution is with normal tendency and the average is a representative value for the most characteristic values. The necessity of estimation of these indicators appears especially for the distributions on great or unequal interval s.

The asymmetry of a distribution will be much great as the difference between the arithmetic average and the mode will be bigger and inverse. [7], [12] In the case of one mode perfect symmetrical distribution, the two indicators are equal such as: $\bar{x} = M_O$. This means that we can form a first image about the degree of asymmetry for a distribution by comparison the arithmetic average with the mode such as:

$$\Delta = \bar{x} - M_0 \tag{11}$$

where:

$$\begin{cases} \Delta \phi\ 0\_\ for\_distribution\_skewed\_to\_the\_right \\ \Delta \pi\ 0\_\ for\_distribution\_skewed\_to\_the\_left \\ \Delta = 0\_\ for\_symmetrical\_distribution \end{cases}$$

The principal drawback of this absolute asymmetry indicator is the expression in the same units of measure with the characteristics of distributions so that it can't be used for comparison the degree of asymmetry of some distributions expressed in different units of m easure.

Frequently, for the interpretation the asymmetry are used the indicators proposed by Pearson, which can measure how oblique is the distribution. [3], [8]

These are:

- *the coefficient of asymmetry* (C_{as}), which is computed as a repo rt between the absolute asymmetry $(\bar{x} - M_0)$ and standard deviation (σ), according with the formula:

$$C_{as} = \frac{\bar{x} - M_0}{\sigma} \tag{12}$$

The coefficient of asymmetry C_{as} can take values between -1 and +1 and how smaller it is in absolute value so smaller is the asymmetry. In a perfect symmetrical distribution, C_{as} is zero because the average coincides with the value of mode. If the average is bigger than the mode, the coefficient of asym metry takes values between 0 and +1, so the distribution is skewed to the right and if the mode is bigger than the average, the coefficient of asymmetry take values between -1 and 0, so the distribution is skewed to the left.

- *the coefficient of asymmetry* $(C_{as}^{'})$, which is computed as a report between the deviation of the median from the average taken three times and the standard deviation, according with the formula:

$$C_{as}^{'} = \frac{3(\bar{x} - M_e)}{\sigma} \quad (13)$$

This formula is used for easy asymmet rical distributions in which for a great number of cases the following formula which defines *the asymmetry in absolute size* is real:

$$As = \bar{x} - Mo \text{ or } \bar{x} - Mo = 3 \cdot (\bar{x} - Me) \quad (14)$$

The asymmetry coefficient $C_{as}^{'}$ takes values in the interval [-3, +3] and shows the greatest degree of asymmetry at the values zero.

The estimating the asymmetry coefficients proposed by Karl Pearson are based on the relation between the three indicators of central tendency: the average, the medi an and the mode. But the asymmetry can be analysed used also other average indicators of position, so that there are other methods to measure the asymmetry. Also, for taking into consideration the influence of quartiles in the asymmetry study is used the *Bowley's coefficient*, estimated as a ratio of the difference between deviances of two extreme quartiles from the median and the sum of these, by the relation:

$$A_s = \frac{q_2 - q_1}{q_2 + q_1} \quad (15)$$

where: $q_2 = Q_3 - M_e$; $q_1 = M_e - Q_1$.

Bowley appreciated that if the value of this coefficient is near 0,1 than the series has a moderate asymmetry and if the value is over 0,3, the series has a pronounced asymmetry.

The coefficient of asymmetry Yulle-Kendall is an indicator estimated on the base of the three quartiles of the series using the relation:

$$A_s^{'} = \frac{Q_1 + Q_3 - 2\bar{x}}{Q_3 - Q_1} \quad \text{or} \quad A_s^{'} = \frac{(Q_3 - \bar{x}) - (\bar{x} - Q_1)}{(Q_3 - \bar{x}) + (\bar{x} - Q_1)} \quad (16)$$

The value of coefficient is contained in the interval [-1, +1] and its interpretation is similar with the value of asymmetry coefficient C_{as}.

The Bowley' coefficient of asymmetry can be computed also on the base of two deciles or two centiles placed at equal distance from median, for example D_1 and D_9 or C_{10} and C_{90} or C_1 and C_{99} etc.

In the case of using the intercentilical deviations equal distanced from the median the indicator of asymmetry will be:

$$S_k = \frac{(C_{90} - Me) - (Me - C_{10})}{C_{90} - C_{10}} \quad \text{o} \quad S_k = \frac{C_{10} + C_{90} - 2Me}{C_{90} - C_{10}} \quad (17)$$

T. Kelley has proposed a coefficient of asymmetry with a most important theoretical signification computed on the base of centiles C_{10} and C_{90} after the following formula:

$$S_k = C_{50} - \frac{C_{90} + C_{10}}{2} \quad (18)$$

For the measuring of the degree of asymmetry are also used the coefficients β_1, proposed by Karl Pearson and γ_1, proposed by Ronald A. Fisher, which are based on the central moment of the series.(1) [6], [12]

The calculus formula for the coefficient β_1 is:

$$\beta_1 = \frac{\mu_3'^2}{\mu_2'^3} \tag{19}$$

where:

$$\mu_2' = \frac{\sum(x-\bar{x})^2}{n} \text{ (centred moment of order 2)} \tag{20}$$

$$\mu_3' = \frac{\sum(x-\bar{x})^3}{n} \text{ (centred moment of order 3)} \tag{21}$$

The statistical interpretation of this coefficient is based on the properties of normal distribution. In the symmetrical distribution, the odd centred moments are equals with zero, but for the asymmetrical distributions its values indicate the size of asymmetry. So, as the value of β_1 is deviated from zero, the degree of asymmetry is bigger.

The coefficient γ_1 is obtained starting from β_1 by formula:

$$\gamma_1 = \pm\sqrt{\beta_1} \tag{22}$$

or

$$\gamma_1 = \frac{\mu_3}{\sigma^3}, \text{ with } \sigma^3 = \sqrt{\mu_2^3}\,. \tag{23}$$

The interpretation of the value of coefficient γ_1 is the following:

- if $\gamma_1 \phi 0$ than the coefficient indicates the asymmetry of left because the sum of partial positive differences $(x_i - \bar{x})$ taken in absolute value is predominant;
- if $\gamma_1 = 0$, than the coefficient indicates symmetry, because the positive and negative sums $(x_i - \bar{x})$ are equals;
- if $\gamma_1 \pi 0$, than the coefficient indicates the asymmetry of right, because the sum of negative differences $(x_i - \bar{x})$ taken in absolute value is bigger than the sum of positive differences.

These two methods: the graphic method or the analytic method, permit the representati on, the establishing of the sense and the intensity of asymmetry in the distributions of frequencies.

THE PEAKEDNESS/ARCHING IN DISTRIBUTIONS OF FREQUENCIES

The description of the degree of peakedness in distributions of frequencies is made by comparison with the graphic of mesokurtic curve. [3], [9], [11] So, a distribution has a great degree of peakedness if a great variation of the studied characteristic trains a small variation of frequencies and inverse. This reasoning is represented in Figure 3.

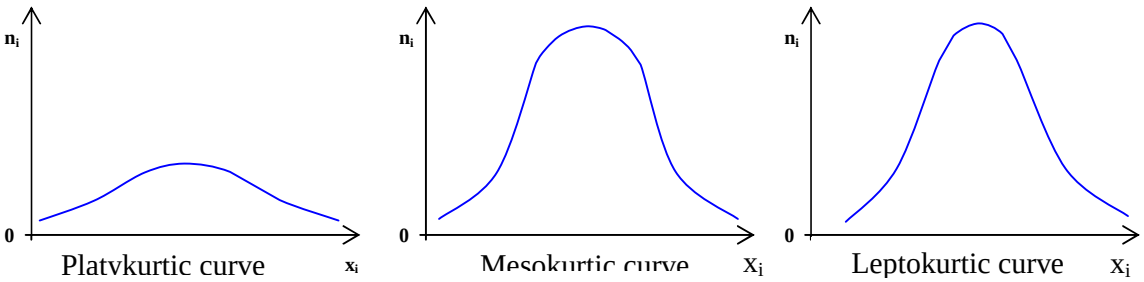

Figure no. 3. Three degrees of peakedness

The platykurtic curve is an approximation of the graphic representation of Student distribution (known as t distribution too). The Student rule is used for the description the samples distributions in the case of sample of small size $(n \pi 30)$ and the graphic representation of this distribution take the form of normal curve with tendency of peakedness.

The leptokurtic curve is an approximation of the graphic representation of Fisher – Snedecor distribution.

For the measure the degree of peakedness of the curves of frequencies are used the indicators β_2 and γ_2. *The Pearson's coefficient* β_2 is computed with the formula:

$$\beta_2 = \frac{\mu_4^{'4}}{\mu_2^{'2}} \tag{24}$$

where:

$\mu_4^{'}$ represents the centrate moment of order four and is determined with the formula:

$$\mu_4^{'} = \frac{\sum (x - \bar{x})^4}{n} \tag{25}$$

For a normal distribution the value of coefficient β_2 is equal with 3 and the interpretation of the degree of peakedness of the curve is made after how the computed value, for an empirical series is bigger or smaller than 3 (as much the value of coefficient is smaller so the curve of freque ncies is much platykurtic.

The Ronald A. Fisher's coefficient γ_2 is computed using the formula:

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4^{'4}}{\mu_2^{'2}} \tag{26}$$

The interpretation of the coefficient γ_2 in the description the degree of peakedness for the curves of frequencies is the following:

- if $\gamma_2 = 0$, than the distribution is normal;
- if $\gamma_2 \phi 0$, than the distribution is leptokurtic.

The computing the coefficients β_2 and γ_2 as measures of peakedness is difficult about the operations involved, so that in many case is made an analysis of the graphic representations of distributions of frequencies (the histogram, the polygon of frequencies etc.

CONCLUSIONS

It can be ascertained that the statistical analysis of asymmetry is based on the analysis of the evolution shape of empirical data using the graphic representation, on the basis of histogram and frequencies polygon and the comparison of th is with the normal distribution graphic and also on the calculus and interpretation of different indicators of asymmetry, which are based on the relation between the central tendency (mean, median, mode, quartiles) or between the individual values and the central tendency indicators. The conclusions of application of these methods are refer to the form of distribution of frequencies, the sense of asymmetry (at the left or at the right) and the intensity of asymmetry (distributions more asymmetric or distrib utions near by normal).

NOTES:

(1) In statistics are used three types of moments:
a) initial moments of different orders, then the origin point of deviations is equal with zero:

$M_k = \frac{\sum x_i^k}{n}$ for simple series and $M_k = \frac{\sum x_i^k n_i}{\sum n_i}$ for the series with frequencies;

where:
M_k represents the initial moment of order k. For $k = 1, M_1 = \bar{x}$.
b) ordinary moments, when the origin point of deviations is considered an arbitrary value A.
The formulas of computing are:

$\mu_k = \frac{\sum (x_i - A)^k}{n}$ for simple series and $\mu_k = \frac{\sum (x_i - A)^k \cdot n_i}{\sum n_i}$ for the series with frequencies.

c) centrate moments, when the origin point of deviations is the arithmetic average of series. The formulas of computing are:

$\mu_k^{'} = \frac{\sum (x_i - \bar{x})^k}{n}$ for simple series and $\mu_k^{'} = \frac{\sum (x_i - \bar{x})^k \cdot n_i}{\sum n_i}$.

So: $\mu_1^{'} = 0$ and $\mu_2^{'} = dispersia(\sigma^2)$.

BIBLIOGRAPHY:

[1] Andrei, T., Stancu, S., Pele, D. T., *Statistica. Teorie şi aplicaţii*, Editura Economică, Bucureşti, 2002

[2] Anghelache, C., *Statistică generală*, Editura Economică, Bucureşti, 1999

[3] Baron, T., Biji, E. M., Tövissi, L., et al., *Statistică teoretică şi economică*, Editura Didactică şi Pedagogică – R.A., Bucureşti, 1997

[4] Bădiţă, M., Baron, T., Korka, M., *Statistica pentru afaceri*, Editura Eficient, Bucureşti, 1998

[5] Biji, M., Biji, E. M., Lilea, E., Anghelache, C., *Tratat de statistică aplicată*, Editura Economică, Bucureşti, 2002

[6] Biji, E. M., Lilea, E., Roşca, R. E., Vătui, M., *Statistică aplicată în economie*, Editura Universal Dalsi, Bucureşti, 2000

[7] Isaic-Maniu, Al., Mitruţ, C., Voineagu, V., *Statistica pentru managementul afacerilor*, Editura Economică, Bucureşti, 1999

[8] Jaba, E., *Statistică*, Editura Economică, Bucureşti, 1998

[9] Keller, G., Warrack, B., Bartel, H., *Statistics for Management and Economics. A Systematic Approach*, Wadaworth Publishing Company, Belmont, California, 1988

[10] Levin, I., R., *Statistics for management*, Fourth Edition, Published by Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987

[11] Loftus, G. R., Loftus, F. E., *Essence of Statistics*, Second Edition, Published by Alfred A. Knopf, Inc., New York, 1988

[12] Ţarcă, M., *Tratat de statistică aplicată*, Editura Didactică şi Pedagogică R. A., Bucureşti, 1998

* * * *Romanian Statistical Yearbook*, National Institute of Statistics, Bucharest, 2007