ECREG STUDIES
Economic and Regional Studies
Studia Ekonomiczne i Regionalne
ISSN 2083-3725 Volume 8, No. 2, 2015

www.ers.edu.pl

ORIGINAL ARTICLE ARTYKUŁ

SOME NOTES ON THE ASYMMETRY OF EMPIRICAL DISTRIBUTIONS

KILKA UWAG O ASYMETRII ROZKŁADÓW EMPIRYCZNYCH

Mirosława Wesołowska-Janczarek
Pope John Paul II State School of Higher Education in Biała Podlaska
Państwowa Szkoła Wyższa im. Jana Pawła II w Białej Podlaskiej

Wesołowska-Janczarek M. (2015), *Kilka uwag o asymetrii rozkładów empirycznych/ Some notes on the asymmetry of empirical distributions.* Economic and Regional Studies, vol. 8, no.2, pp. 80-84.

Summary: This paper discusses different aspects of asymmetry of empirical distributions. An attempt was made to clarify the definition of such distributions and to identify some of the problems associated with commonly used skewness coefficients of As and γ and their interpretation and those yet requiring further research.

Keywords: empirical distribution, asymmetry of the distribution, measures of asymmetry

Streszczenie: W pracy rozważane są różne aspekty asymetrii rozkładów empirycznych. Podjęto próbę sprecyzowania określenia takich rozkładów oraz wskazano pewne problemy związane z wykorzystywanymi powszechnie współczynnikami skośności As i γ oraz ich interpretacją a wymagające jeszcze dalszych badań.

Słowa kluczowe: rozkład empiryczny, asymetria rozkładu, miary asymetrii

Introduction

Both within the economy as well as in other fields of knowledge the conducted studies often focus on the determination of the structure of the data obtained in the course of research for a feature or for many features of interest to the researcher. Features, as we know, can be of various types. In this study, we will be dealing with measurable characteristics. Their values collected during the study, to the extent possible, with all the elements of the entire population under consideration, that is, within the full study, ought to be placed in order creating a stemplot also called the empirical distribution.

A division row can be point or interval, and its graphical display is respectively a spot chart or histogram. Each sectional division row with sections of equal intervals can changed into a point one by defining the middle values of the sections. Therefore, further considerations will apply to the point distribution rows of c different values of the form:

Wstęp

Zarówno w ekonomii jak i w innych dziedzinach wiedzy prowadzone badania często koncentrują się na określeniu struktury danych uzyskanych w trakcie badań dla cechy lub wielu cech interesujących badającego. Cechy jak wiadomo mogą być różnego typu. Tutaj będziemy się zajmować cechami mierzalnymi. Ich wartości zebrane w trakcie badania, o ile to jest możliwe, u wszystkich elementów całej rozważanej zbiorowości, czyli w badaniu pełnym, należy uporządkować tworząc szereg rozdzielczy zwany też rozkładem empirycznym.

Szereg rozdzielczy może być punktowy lub przedziałowy, a ich graficznym obrazem jest odpowiednio wykres punktowy lub histogram. Każdy szereg rozdzielczy przedziałowy o przedziałach jednakowej długości można sprowadzić do punktowego wyznaczając wartości środkowe przedziałów. Dlatego też dalsze rozważania będą dotyczyły punktowych szeregów rozdzielczych o c różnych wartościach postaci:

Address for correspondence: *prof. dr hab. Mirosława Wesołowska-Janczarek, Pope John Paul II State School of Higher Education in Biała Podlaska, Sidorska 95/97, 21-500 Biała Podlska, Poland; phone: +48 83 344-99-05; e-mail: wesolowska.janczarek@gmail.com*
Full text PDF: *www.ers.edu.pl; Open-access article.*

Indexation: *Index Copernicus Journal Master List ICV 2014: 70.81 (6.96); Polish Ministry of Science and Higher Education 2014: 4 points.*

Values x_i	Numbers n_i
x_1	n_1
x_2	n_2
...	...
x_c	n_c
Total	n

Wartości x_i	Liczebności n_i
x_1	n_1
x_2	n_2
...	...
x_c	n_c
Suma	n

We also accept further the indications: $\bar{x} = \frac{1}{n}\sum_{i=1}^{c} x_i n_i$ is a mean value Me - median (middle value of an ordered data set), D - also called dominant (most common value in the data set) . Considerations cover the conventional unimodal distributions.

When examining the data structure various measures are determined, among which there are: core position, diversity and variability, asymmetry and flattening and concentration. The most commonly used are the measure of the position and diversity, but also while calculating other ones we may obtain interesting information. We will proceed to dealing with the problem of asymmetry of empirical distributions.

Przyjmujemy też dalej oznaczenia: $\bar{x} = \frac{1}{n}\sum_{i=1}^{c} x_i n_i$ jest wartością średnią, Me – medianą (wartością środkową uporządkowanego zbioru danych), D – dominantą zwaną też modą (wartością najczęściej występującą w zbiorze danych). Rozważania dotyczą typowych rozkładów jednomodalnych.

Przy badaniu struktury danych wyznacza się różne miary, wśród których wyróżnia się miary położenia, zróżnicowania lub zmienności, asymetrii oraz spłaszczenia i koncentracji. Najczęściej używane są miary położenia i zróżnicowania, ale też obliczając pozostałe można uzyskać ciekawe informacje. W dalszym ciągu będziemy się zajmować problemem asymetrii rozkładów empirycznych.

The asymmetry of the statistical distribution in literature

Usually, upon introducing certain concepts one starts by giving their definition. Even a general review of academic books, most recommended for the subject of statistics does not provide the definition of the asymmetry of the distribution. One can find the definition of a symmetric distribution, and then any distribution that does not meet the criteria of this definition will be asymmetric distribution. Most often, however, the concept of asymmetry is introduced in a descriptive fashion or it is illustrated graphically.

Jozwiak and Podgórski (2012), p. 49 give the following definition of the empirical symmetrical distribution: "We say that the empirical distribution is symmetric if each feature value $x_i < \bar{x}$ corresponds to the same value $x_j > \bar{x}$ and $\bar{x} - x_i = x_j - \bar{x}$ and $n_i = n_j$ ". This means that equally distant values from the middle to the right and left must occur equally frequently. If it is not the case, then the distribution is considered to be asymmetrical.

On the other hand Zeliaś (2000), pp. 65, starting with equal mean, median and dominant states that "in the asymmetric distributions the values of these characteristics differ, and the differences between them are greater, when the empirical distribution of the variable under consideration differs more and more from the symmetrical distribution."

Sobczak (2010), (2005) and (2006) by introducing asymmetry states that it indicates whether the "overwhelming number of units making up for the researched population has a higher or lower feature values than the average level." This can be understood both in such a way that the number $x_i < \bar{x}$ and $x_j > \bar{x}$ is not the same as well as the corresponding numbers of n_i and n_j are different.

Asymetria rozkładu w literaturze statystycznej

Zwykle wprowadzając pewne pojęcie zaczyna się od podania jego definicji. Już pobieżny przegląd podręczników akademickich, najczęściej polecanych do przedmiotu statystyka nie przynosi definicji asymetrii rozkładu. Można znaleźć definicję symetrycznego rozkładu, a wtedy każdy rozkład, który nie spełnia warunku podanego w tej definicji, będzie rozkładem asymetrycznym. Najczęściej jednak pojęcie asymetrii wprowadza się w sposób opisowy słowny lub ilustrowany graficznie.

Jóźwiak i Podgórski (2012) str. 49 podają następującą definicję symetrycznego rozkładu empirycznego: „Mówimy, że rozkład empiryczny jest symetryczny, jeżeli każdej wartości cechy $x_i < \bar{x}$ odpowiada taka sama wartość $x_j > \bar{x}$, że $\bar{x} - x_i = x_j - \bar{x}$ oraz $n_i = n_j$". Oznacza to, że wartości jednakowo odległe od średniej na prawo i lewo muszą występować tak samo często. Jeżeli tak nie jest, to rozkład uznaje się za asymetryczny.

Natomiast Zeliaś (2000) str. 65 wychodząc z równości średniej, mediany i dominanty stwierdza, że „w rozkładach asymetrycznych wartości tych charakterystyk różnią się między sobą, a różnice między nimi są tym większe, im empiryczny rozkład badanej zmiennej bardziej odbiega od symetrycznego".

Sobczak (2010), (2005) i (2006) wprowadzając asymetrię stwierdza, że wskazuje ona, czy „przeważająca liczba jednostek tworzących badaną zbiorowość ma wartości cechy wyższe lub niższe od przeciętnego poziomu". Można to rozumieć zarówno w ten sposób, że liczba wartości $x_i < \bar{x}$ jak i $x_j > \bar{x}$ nie jest taka sama, jak też i odpowiadające im liczebności n_i i n_j są różne.

Starzyńska (2002) by introducing the concept of an asymmetric distribution based it on a comparison of the average value with the dominant. Other authors, like Zeliaś, also starting from symmetric distribution for which $\bar{x} = Me = D$ state that each distribution wherein there is no equality of the average, median and dominant is asymmetric. As we can see, not all authors take the average value as a focal point in the considerations of the empirical distribution asymmetry.

It is also worth to recall that in the book by Koronacki and Mielniczuk (2001) p. 21 on the asymmetry of the distribution it is said that "the histogram values on the right side of the dominant decrease much more slowly than on the left side.". The values of the histogram are to be understood as the highnees of rectangles, and the slower decrease in these values is often associated with a longer "tail" of the graph.

To summarize this information, one can say that a definition of an empirical definition of asymmetric distribution has not yet been formulated.

On some issues related to the distribution asymmetry measures

In the subject literature, one can find several different measures of asymmetry of the distribution. The most commonly used are the coefficients of skewness or asymmetry expressed formulas:

$$As = \frac{\bar{x} - D}{S}$$

and

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{c} (x_i - \bar{x})^3 n_i}{S^3}.$$

The latter one is based on the third central distribution point. The expression in the numerator of the first of these formulas $\bar{x}$ – D is called an indicator of asymmetry. Markings used herein are consistent with the previously entered ones where $\bar{x}$ is average, D - dominant, and the standard deviation of the feature is marked as S . The values of these coefficients, as you may read, for example, in the book by Sobczyk (2005), are generally included in the range <-1, + 1>, but their values are not equal (As ≠ γ) in absolute value, and sometimes they different in signs. This is a problem because the characters are to point to left-sided asymmetry when the asymmetry coefficient is negative or right-sided asymmetry in case of positive asymmetry coefficient. It is also worth noting, as reported by Zeliaś (2000), pp. 67, that if the "asymmetry is not too strong," then the absolute value of the asymmetry coefficient As is a number between 0 and 2.

It is assumed that the right-handed asymmetry, which usually means "long right tail of the distribution graph" occurs when the relation $\bar{x} > Me > D$ or left-handed, when the "long tail graph is left," and

Starzyńska (2002) wprowadzając pojęcie rozkładu asymetrycznego opiera go na porównaniu wartości średniej z dominantą. Inni autorzy, podobnie jak Zeliaś, także zaczynając od rozkładu symetrycznego, dla którego $\bar{x} = Me = D$ stwierdzają, że każdy rozkład w którym nie zachodzi równość średniej, mediany i dominanty jest asymetryczny. Jak widać nie wszyscy autorzy jako punkt centralny w rozważaniach asymetrii rozkładu empirycznego przyjmują wartość średnią.

Warto tu jeszcze przypomnieć, że w książce Koronackiego i Mielniczuka (2001) str. 21 o asymetrii rozkładu mówi się gdy: „wartości histogramu po prawej stronie mody (dominanty) maleją znacznie wolniej niż po lewej jej stronie". Pod wartościami histogramu należy rozumieć wysokości prostokątów, a powolniejsze zmniejszanie się tych wartości często związane jest z dłuższym „ogonem" wykresu.

Podsumowując te informacje można stwierdzić, że dotychczas nie została sformułowana definicja asymetrycznego rozkładu empirycznego.

O pewnych problemach dotyczących miar asymetrii rozkładu

W literaturze przedmiotu można znaleźć kilka różnych miar asymetrii rozkładu. Najczęściej używane są współczynniki skośności lub asymetrii wyrażane wzorami:

$$As = \frac{\bar{x} - D}{S}$$

oraz

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{c} (x_i - \bar{x})^3 n_i}{S^3}.$$

Ten ostatni oparty jest na trzecim momencie centralnym rozkładu. Wyrażenie w liczniku pierwszego z tych wzorów $\bar{x}$ – D nazywane jest wskaźnikiem asymetrii. Użyte tu oznaczenia są zgodne z wcześniej wprowadzonymi czyli $\bar{x}$ jest średnią, D – dominantą, a S- odchyleniem standardowym cechy. Wartości tych współczynników, jak można przeczytać na przykład w książce Sobczyka (2005), na ogół zawierają się w przedziale < -1,+1>, ale ich wartości nie są równe (As ≠ γ) co do wartości bezwzględnej, a czasami też różnią się znakiem. Jest to problem, gdyż znaki mają wskazywać na lewostronną asymetrię, gdy współczynnik asymetrii jest ujemny lub prawostronną asymetrię przy dodatnim współczynniku asymetrii. Warto tu jeszcze dodać, jak podaje Zeliaś (2000) str. 67, że gdy „asymetria nie jest zbyt silna", to wartość bezwzględna współczynnika asymetrii As jest liczbą z przedziału < 0,2 >.

Przyjmuje się, że asymetria prawostronna, co zwykle oznacza „dłuższy prawy ogon wykresu rozkładu" jest wtedy, gdy zachodzi relacja $\bar{x} > Me > D$ lub lewostronna, gdy „dłuższy jest lewy ogon wykresu" i wtedy zachodzi relacja $\bar{x} < Me < D$. Można

then the relation $\bar{x} < Me < D$ takes place. It can, in many distributions however be noted that the median and the mode are equal and the distribution will also not be symmetric. Such a distribution can be for example:

x_i	0	1	2	3	4	5
n_i	10	23	15	5	5	2

where n = 60, D = 1 = Me, $\bar{x}$ = 1,63, As = 0,4931, γ = 0,8605. This is the distribution of the right-sided asymmetry where $\bar{x} > Me = D$.

Consider another example of the empirical distribution:

x_i	1	2	3	4	5
n_i	2	6	12	7	3

where n = 30, $\bar{x}$ = 3,1, D = 3 = Me and As = 0,0958 while γ = –0,0246. Through this also $\bar{x} > Me = D$. Is it a right handed asymmetry as shown As or left-handed ace as the sign of the coefficient γ would indicate.

Let's consider more on what do the asymmetry factors depend. Let us consider the following examples of empirical distributions:

x_i	1	2	3	4	5		y_i	1	3	5	7	9
n_i	5	10	20	12	3		n_i	5	10	20	12	3

When calculating the coefficients of As and γ of both these examples, we obtain: As = –0,0385 and γ = –0,1344. It can be seen that the values of the coefficients of As and γ do not depend on the value of the features here. Is that all?

It is worth to point out that only if $\bar{x} = D$ it is possible to find the values of x_i and x_j satisfying the condition of the definition of the symmetry of the distribution of the already quoted book by Jozwiak and Podgórski (2012), ie. $\bar{x} - x_i = x_j - \bar{x}$. If x ≠ D thus $\bar{x} < D$ or $\bar{x} > D$ such values that are the same distance from the average to the mean value x_i and x_j are non-existent. Usually, if the number of observations also $x_i < \bar{x}$ is greater than the number of observations $x_j > \bar{x}$, then $\bar{x} > D$, and if the number of observations $x_i < \bar{x}$ is less than the number of observations $x_j > \bar{x}$ is $\bar{x} < D$. But one cannot say that it affects different coefficient of the considered factors. Finding the answer to the question of what is causing the different signs, the factors considered and the occurrence of other suggested here doubts remains an open question.

An attempt to clarify the determination of the asymmetry of the empirical distribution

As already mentioned, the asymmetry of the empirical distribution has not yet been defined. Is it possible to try to clarify this concept?

When taking up this challenge it can be assumed that:

jednak w wielu rozkładach stwierdzić, że mediana i dominanta mogą być równe i rozkład też nie będzie symetryczny. Takim rozkładem może być na przykład:

x_i	0	1	2	3	4	5
n_i	10	23	15	5	5	2

gdzie n = 60, D = 1 = Me, $\bar{x}$ = 1,63, As = 0,4931, γ = 0,8605. Jest to rozkład o prawostronnej asymetrii gdzie $\bar{x} > Me = D$.

Rozważmy kolejny przykład rozkładu empirycznego:

x_i	1	2	3	4	5
n_i	2	6	12	7	3

gdzie n = 30, $\bar{x}$ = 3,1, D = 3 = Me oraz As = 0,0958 natomiast γ = –0,0246. Tym razem też $\bar{x} > Me = D$. Czy jest to prawostronna asymetria jak pokazuje As czy lewostronna jak wskazywałby znak współczynnika γ.

Zastanówmy się jeszcze nad tym od czego zależą współczynniki asymetrii. Rozważmy następujące przykłady rozkładów empirycznych:

x_i	1	2	3	4	5		y_i	1	3	5	7	9
n_i	5	10	20	12	3		n_i	5	10	20	12	3

Obliczając współczynniki As i γ w obu tych przykładach otrzymujemy: As = –0,0385 oraz γ = –0,1344. Widać, że wartości współczynników As i γ nie zależą tu od wartości cechy. Czy tak jest zawsze?

Warto jeszcze zwrócić uwagę na to, że tylko wtedy, gdy $\bar{x} = D$ można znaleźć wartości x_i oraz x_j spełniające warunek z definicji symetrii rozkładu w cytowanej już książce Jóźwiak i Podgórski (2012) czyli $\bar{x} - x_i = x_j - \bar{x}$. Jeżeli x ≠ D czyli $\bar{x} < D$ lub $\bar{x} > D$ takich wartości równoodległych od średniej x_i i x_j nie ma. Zwykle też jeżeli liczba obserwacji $x_i < \bar{x}$ jest większa od liczby obserwacji $x_j > \bar{x}$, to $\bar{x} > D$, a jeśli liczba obserwacji $x_i < \bar{x}$ jest mniejsza od liczby obserwacji $x_j > \bar{x}$ to $\bar{x} < D$. Nie można jednak stwierdzić, że to wpływa na różne znaki rozważanych współczynników. Znalezienie odpowiedzi na pytanie co powoduje występowanie różnych znaków rozważanych współczynników i wyjaśnienie innych zasugerowanych tu wątpliwości pozostaje nadal sprawą otwartą.

Próba sprecyzowania określenia asymetrii empirycznego rozkładu

Jak już wcześniej wspomniano asymetria rozkładu empirycznego dotychczas nie została zdefiniowana. Czy można spróbować sprecyzować to pojęcie?

Podejmując tę próbę można przyjąć, że:

Empirical distribution will be called asymmetric if the following condition is met: the average value of $\bar{x}$ is different from the dominant D.

This condition also shows that

1. the number of different values $x_i < \bar{x}$ is not equal to the number of different values $x_j > \bar{x}$,
2. with the same number of $x_i < \bar{x}$ and $x_j > \bar{x}$ and corresponding to them numbers n_i and n_j are not the same,
3. Total number of observations smaller than the average is different from the number of all observations greater than the average.

Let us also note that the median has the least important impact on the asymmetry. It is therefore not important whether the median and the mode are equal or not.

Issues raised in this paper require further research.

Rozkład empiryczny będzie nazywany asymetrycznym, jeśli spełniony jest następujący warunek: wartość średnia $\bar{x}$ jest różna od dominanty D.

Z warunku tego wynika też, że

1. liczba różnych wartości $x_i < \bar{x}$ nie jest równa liczbie różnych wartości $x_j > \bar{x}$,
2. przy jednakowej liczbie wartości $x_i < \bar{x}$ i $x_j > \bar{x}$ odpowiadające im liczebności n_i i n_j nie są takie same,
3. liczba wszystkich obserwacji mniejszych od średniej jest różna od liczby wszystkich obserwacji większych od średniej.

Zauważmy jeszcze, że na asymetrię najmniejszy wpływ ma mediana. Nie jest też ważne, czy mediana i dominanta są równe czy też nie.

Poruszone w tej pracy problemy wymagają jeszcze dalszych badań.

References/ Literatura:

1. Jóźwiak J., Podgórski J. (2012), *Statystyka od podstaw*, Wyd. VII, Polskie Wydawnictwo Ekonomiczne, Warszawa .
2. Koronacki J., Mielniczuk J. (2001), *Statystyka dla kierunków technicznych i przyrodniczych,* Wydawnictwo Naukowo-Techniczne, Warszawa.
3. Sobczyk M. (2010), *Statystyka opisowa*, Wydawnictwo C.H.Beck, Warszawa .
4. Sobczyk M. (2005), *Statystyka*, Wyd. IV zmienione. Wydawnictwo Naukowe PWN. Warszawa.
5. Sobczyk M. (2006), *Statystyka aspekty praktyczne i teoretyczne*, Wydawnictwo UMCS, Lublin .
6. Starzyńska W. (2002), *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, Warszawa.
7. Zeliaś A.(2000), *Metody statystyczne,* Polskie Wydawnictwo Ekonomiczne, Warszawa.

Submitted/ Zgłoszony: September/ wrzesień 2014
Accepted/ Zaakceptowany: November/ listopad 2014