Proceedings of the 2019 International Conference "ECONOMIC SCIENCE FOR RURAL DEVELOPMENT" No 52
Jelgava, LLU ESAF, 9-10 May 2019, pp. 150-156
DOI: 10.22616/ESRD.2019.116

THE STATISTICAL ANALYSIS OF POLISH FOOD ENTERPRISES: - NONPARAMETRIC APPROACH

Aleksandra Baszczynska[1], PhD
[1]Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz, Poland

Abstract. Statistical analysis of Polish food enterprises is done to present economic situation of agro-food industry in Poland. In analysis, nonparametric approach is chosen as effective and simple method of studying variables in populations. This approach is widely used, especially when additional information about regarded variable is not available (as often happens in economic researches).

Two nonparametric estimation methods are taken into consideration: kernel density estimation and bootstrap confidence interval. The special emphasis is taken on choosing the proper values of parameters in kernel density estimation and choosing the most effective bootstrap interval among these presented in literature. The study concerns applying basic descriptive statistics and nonparametric estimation of number of employees and revenues total of Polish food enterprises, using kernel method for estimating the density function and bootstrap confidence interval for median of regarded variable. Results and conclusions from the study can be useful for the users of nonparametric methods in economic researches.

The main research aim of the paper is to present and examine some statistical procedures that can be used in the analysis of economic situation of chosen enterprises connected strictly with food production. The good properties of regarded methods allow compering some regions of country to indicate these regions where there are friendly conditions for the food production enterprises, including the natural character of region (rural or urban area).

Key words: Polish food enterprises, nonparametric methods, kernel density estimation, bootstrap interval.
JEL code: C13, C14, Q10.

Introduction

Nonparametric statistical methods become more and more popular and widely used because of their simplicity and good properties, not only in economic but also in technical and natural researches. Classical statistical methods that are based on assumption that data are generated by known family of distribution (for example family of normal distributions) in most cases cannot be used because of not fulfilling this assumption. In many cases there is No additional information about regarded variables. This information can be connected with knowledge of random variable distribution (exact form of this variable's distribution). When the distribution of underlying observations cannot be taken to be of certain form (for example normal one) the nonparametric (distribution-free) methods are the only ones are used in statistical analysis.

In study, two nonparametric methods are chosen and applied. They are both estimation procedures and they are both of general nature. In most cases, the researcher can get, using these methods, sufficient information of regarded phenomenon. Basic characteristics of variables are achievable from this kind of nonparametric analysis. Sometimes they can play the introductionary role in wide statistical analysis and the results of applying these methods are the base of further detailed procedures, indicating the essential direction of analysis.

Two nonparametric estimation methods are taken into consideration: kernel density estimation and bootstrap confidence interval for median.

Kernel density estimator is used to observe the distribution of the random variable across its support (Kvam, 2007). A few of basic properties, such as asymmetry, modality, dispersion can be detected in this way. Kernel density estimator is defined in the following way (e.g.: Wand, Jones, 1995; Baszczynska, 2016; Ghosh, 2018; Gramacki, 2018):

[1] E-mail adress: aleksandra.baszczynska@uni.lodz.pl

DOI: 10.22616/ESRD.2019.116

$$\widehat{f}(x) = \frac{1}{n\boldsymbol{h}_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\boldsymbol{h}_n}\right) \quad (1)$$

where: $x_1, \dots, x_n$ is the realization of the sample $X_1, \dots, X_n$; $i = 1, \dots, n$; K is the kernel function and $\boldsymbol{h}_n$ is the smoothing parameter. Kernel function is chosen, in most cases, to be symmetric about zero, unimodal density function, e.g. gaussian kernel, Epanechnickov kernel, triangular kernel, box kernel (e.g. Silverman, 1986; Baszczynska, 2016). Gaussian kernel is density function for normally distributed random variable with mean 0 and variance 1.

Epanechnickov kernel has the form (e.g. Härdle, 1991):

$$K_E(x) = \begin{cases} \frac{3}{4}(1 - x^2) & |x| \leq 1 \\ 0 & |x| > 1. \end{cases} \quad (2)$$

Smoothing parameter decides about the spread of kernel and the following conditions are assumed: $\boldsymbol{h}_n \to 0$, $n\boldsymbol{h}_n \to \infty$ as $n \to \infty$. Selectors of the smoothing parameter most often used, are the following (e.g. Baszczynska, 2016): reference rule, over-smoothed rule, least squares cross-validation, biased cross-validation, plug-in selector. Reference rule is based on optimization of asymptotic mean integrated squared error for kernel density estimator with the assumption of the normal density with the same scale as the estimated density. Least squares cross-validation belongs to automatic smoothing parameter selectors and is based on optimization of mean integrated squared error for kernel density estimator with the „leave-one-out" density estimator.

Bootstrap's statistical procedures are based on the idea of resampling the sample itself with replacement. This technique is used in analysis of estimator's statistical accuracy, testing hypothesis and in confidence intervals (e.g. Efron, 1993; Shao, Tu, 1995; Davison, Hinkley, 1997; Domanski et al., 1998; Hutson, 1999; Domanski, Pruska, 2000; Chernik, 2008).

Methods for bootstrap confidence intervals for parameter θ are the following (e.g.: Hall, 1988; Baszczynska, Pekasiewicz, 2008; Chernik, LaBudde, 2011):

- normal approximated interval: $\left[\hat{\theta} - \sigma u_{(1-\alpha)}; \hat{\theta} + \sigma u_{(1-\alpha)}\right]$, where $u_{(1-\alpha)} = \Phi^{-1}(1 - \alpha)$,
- basic percentile method: $\left[\theta^*_{\left(\frac{\alpha}{2}\right)}; \theta^*_{\left(1-\frac{\alpha}{2}\right)}\right]$, where $\theta^*_{\left(1-\frac{\alpha}{2}\right)}$ denotes $1 - \frac{\alpha}{2}$ percentile of bootstrapped coefficients θ^*,
- bias corrected percentile method, adjusts for bias in the bootstrap distribution,
- bias corrected and accelerated percentile method,
- Studentized confidence interval: $\left[\hat{\theta} - \sigma t^*_{\left(1-\frac{\alpha}{2}\right)}; \hat{\theta} - \sigma t^*_{\left(\frac{\alpha}{2}\right)}\right]$, where $t^*_{\left(1-\frac{\alpha}{2}\right)}$ denotes percentile of the bootstrapped Student's t-test.

Regarded nonparametric methods can be widely applied in statistical analysis of the situation of Polish food enterprises. The food production sector is treated as one of the fastest-growing branches that mostly affect economic development in Poland. The analysis of economic-financial situation of food enterprises should be made using the appropriate statistical methods. Nonparametric procedures seem to be well chosen because of special character of variables considered in analysis of economic phenomenon of this kind. In many cases the researcher is not able to get additional information about variable. It is caused not only by the lack of historical statistical data but also by

situation of an attempt of describing quite new phenomenon or phenomenon with high frequency of changes.

In the study the data of number of employees and revenues total of Polish food enterprises in voivodships in 2017 are used. The data were obtained from data base EMIS using the following criteria:

regions and countries: Poland; sector: food production; companies [access date: 12.12.2018]. In analysis two variables are taken into consideration: number of employees and revenues total in companies in each voivodships in Poland. All calculations are made using software program Matlab 2016b.

Research results and discussion

First stage of statistical analysis is devoted to descriptive statistics. Some chosen descriptive statistics methods are used to get very general information of considered phenomenon. The results are presented in Table 1-2.

Table 1

Results of descriptive statistics for numbers of employees for Polish food enterprises in voivodships

Voivodship	Max	Mean	Coefficient of Variation for Standard Deviation	Median	Coefficient of Variation for Semi-interquartile Range	Skewness
Dolnoslaskie	1207	35.86	0.37	5	0.40	7.40
Kujawsko-Pomorskie	1500	71.33	0.46	19.5	0.69	4.88
Lubelskie	1900	76.57	0.37	10	0.41	5.46
Lubuskie	560	34.96	0.50	10	0.52	4.02
Lodzkie	1300	59.95	0.43	10	0.42	4.71
Malopolskie	4832	65.26	0.23	10	0.46	12.61
Mazowieckie	7000	73.83	0.23	5	0.41	12.39
Opolskie	800	55.71	0.48	10	0.43	4.01
Podkarpackie	550	50.58	0.60	10	0.35	2.84
Podlaskie	3480	98.54	0.27	15	0.60	7.52
Pomorskie	3900	58.50	0.26	10	0.42	13.62
Slaskie	1200	52.46	0.43	10	0.42	4.74
Swietokrzyskie	470	54.71	0.51	10	0.40	2.65
Warminsko-Mazurskie	1600	76.83	0.34	10	0.42	5.00
Wielkopolskie	1840	62.92	0.40	10	0.41	6.08
Zachodniopomorskie	958	46.26	0.46	10	0.41	4.85

Source: author's calculations

Table 2

Results of descriptive statistics for revenues total of Polish food enterprises in voivodships

Voivodship	Max (mln zl)	Mean (mln zl)	Coefficient of Variation for Standard Deviation	Median (mln zl)	Coefficient of Variation for Semi-interquartile Range	Skewness
Dolnoslaskie	876.75	39.38	0.37	10.22	0.56	6.02
Kujawsko-Pomorskie	2370.92	74.94	0.27	8.50	0.43	6.71
Lubelskie	596.01	57.44	0.54	21.31	1.08	3.09
Lubuskie	268.63	61.99	0.91	35.96	0.86	1.50
Lodzkie	2151.58	102.77	0.35	20.48	0.62	5.14
Malopolskie	698.33	73.95	0.57	19.05	0.48	2.71
Mazowieckie	4238.11	151.85	0.38	31.45	0.52	6.37
Opolskie	1477.03	87.60	0.31	13.45	0.65	4.28
Podkarpackie	299.540	32.42	0.52	6.43	0.52	2.77
Podlaskie	3420.90	140.95	0.27	13.01	0.54	5.47
Pomorskie	3671.92	75.93	0.24	9.55	0.46	9.18
Slaskie	1039.96	83.73	0.54	25.73	0.70	3.42
Swietokrzyskie	379.40	43.75	0.52	18.16	1.0	3.01
Warminsko-Mazurskie	1226.60	40.29	0.29	5.83	0.46	7.37
Wielkopolskie	1169.12	76.06	0.45	13.95	0.49	3.66
Zachodniopomorskie	448.26	42.38	0.49	9.03	0.43	3.33

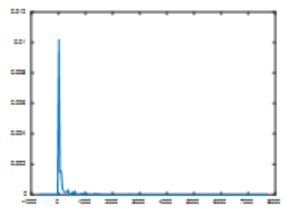
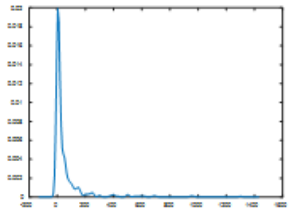
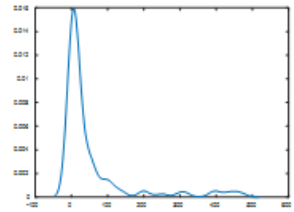
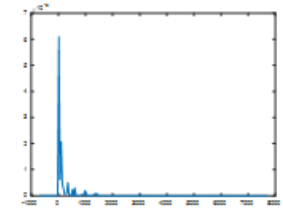
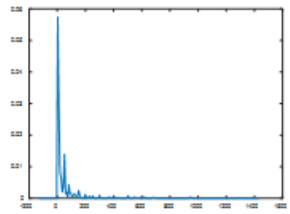
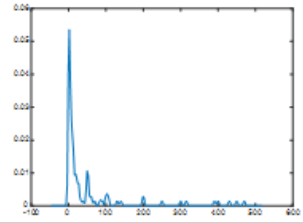
Source: author's calculations

The results of applying descriptive statistics show that both in case of number of employees and revenues total we can observe the lack of symmetry of regarded variables. The big differences between values of mean and median nearly for each voivodship and comparing with the maximum values of variable indicate that the assumption of normality, that is necessary in mostly used statistical procedures, cannot be accepted. This indicates that classical parametric approach is inadmissible. So, the next stage of analysis is based on nonparametric approach.

The kernel density estimators are used to catch more detailed analysis. The kernel density estimators for numbers of employees for chosen voivodships are presented in Figure 1. The kernel density estimators for revenues total for chosen voivodships are presented in Figure 2.

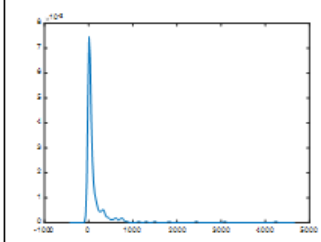
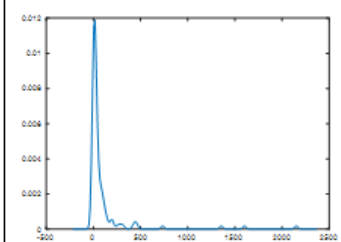
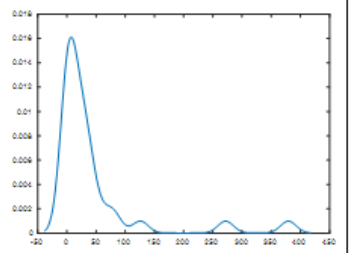
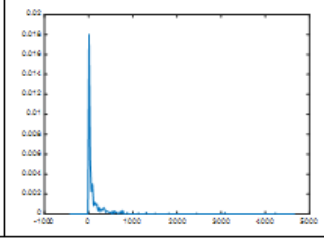
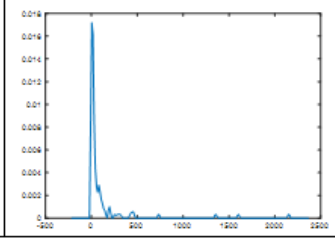
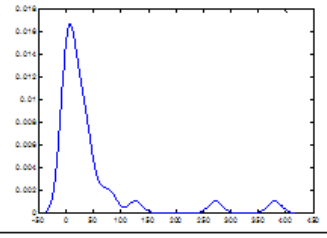
In kernel density estimation gaussian kernel and reference rule for choosing the smoothing parameter (type I) and Epanechnickov kernel and least squares cross-validation for choosing the smoothing parameter (type II) are used. In study the kernel density estimators are calculated for all variables in all voivodships. Mazowieckie, Lodzkie and Swietokrzyskie voivodships are chosen just as examples in presenting the results of second stage of statistical analysis.

DOI: 10.22616/ESRD.2019.116

Type of kernel density estimator	Mazowieckie voivodship	Lodzkie voivodship	Swietokrzyskie voivodship
I			
II			

Source: author's calculations

Fig. 1. **Kernel density estimators for numbers of employees in Polish food enterprises**

Type of kernel density estimator	Mazowieckie voivodship	Lodzkie voivodship	Swietokrzyskie voivodship
I			
II			

Source: author's calculations

Fig. 2. **Kernel density estimators for revenues total in Polish food enterprises**

Basing on the results of density estimation it can be stated that regarded variables are characterized strong asymmetry and in most cases unimodality. So, in next stage of statistical analysis – estimating the mean of variable - the decision of using nonclassical measure of central tendency is made. The bootstrap confidence intervals for medians for regarded variables are taken into consideration. The confidence coefficient is set to be 0.95 and the number of repetition is set to be 10000. The following types of bootstrap confidence intervals are presenting: normal approximated interval, basic percentile method, bias corrected percentile method, bias corrected and accelerated percentile method and Studentized confidence interval.

Tables 3 presents results for bootstrap confidence intervals for median of numbers of employees in Polish food enterprises for chosen voivodships.

Table 3

Bootstrap confidence intervals for median of numbers of employees in Polish food enterprises

Type of bootstrap confidence interval	Mazowieckie voivodship	Lodzkie voivodship	Swietokrzyskie voivodship
Normal approximated interval	(4.91, 5.09)	(8.92, 10.97)	(4.23, 14.28)
Basic percentile method	(4.91, 5.09)	(9.83, 10.09)	(5.00, 18.00)
Bias corrected percentile method	(4.91, 5.09)	(9.83, 10.09)	(5.00, 16.00)
Bias corrected and accelerated percentile method	(4.91, 5.09)	(9.83, 10.09)	(5.00, 16.00)
Studentized confidence interval	(4.91, 5.09)	(9.83, 10.09)	(5.92, 14.18)

Source: author's calculations

It can be noticed that for voivodships characterized by strong asymmetry, the bootstrap confidence interval for median has rather short length of intervals – example: Mazowieckie voivodship. For voivodships characterized by short range of variable, the bootstrap confidence interval for median, in most cases, is characterized by bigger length of intervals with example of Swietokrzyskie.

The results for bootstrap confidence intervals for median of revenues total in Polish food enterprises for chosen voivodships is presented in Table 4.

Table 4

Bootstrap confidence intervals for median of revenues total in Polish food enterprises

Type of bootstrap confidence interval	Mazowieckie voivodship	Lodzkie voivodship	Swietokrzyskie voivodship
Normal approximated interval	(21.64, 41.58)	(13.42, 27.44)	(4.24, 34.50)
Basic percentile method	(20.86, 41.00)	(14.03, 28.25)	(4.57, 32.63)
Bias corrected percentile method	(21.16, 41.00)	(14.03, 28.25)	(4.57, 32.63)
Bias corrected and accelerated percentile method	(20.86, 40.26)	(14.03, 28.74)	(4.57, 32.63)
Studentized confidence interval	(19.12, 40.56)	(11.91, 26.73)	(-9.09, 33.24)

Source: author's calculations

Conclusions, proposals, recommendations

1) The quantitative analysis of Polish food enterprises, especially taking into regard employment and financial situation of enterprises, should be done using statistical procedures appropriate to the character of regarded variables. Nonparametric statistical methods can play the significant role in this process.
2) The proposed nonparametric approach in analysis of the economic variable, consisting of three stages: general descriptive analysis based rather on order statistics, nonparametric estimation of density function and bootstrap confidence interval for median is simply to use and easy to interpret. It can be used even by researcher without big experience because almost No assumption connected to regarded variables has to be fulfilled.
3) Different kernel functions and different values of smoothing parameters in kernel density estimation indicate the same characteristic feature of considered variables. In all regarded cases of applying type II of density estimator (Epanechnickov kernel function and least squares cross-

validation method of choosing the smoothing parameter), the estimator is under-smoothed, which can cause loss of significant information of variable. Even in situations where we have asymmetric distribution of variable the simplest method (gaussian kernel and reference rule) for choosing kernel parameters works quite well.

4) The approach of bootstrap confidence intervals gives a lot of information. But it should be noted that when the range of variable is quite big all regarded intervals are the same. It can be treated as drawback of the procedure.
5) The shape of the kernel density estimators indicates very clearly the character of region where the food production companies come from. For both rural and urban regions, the modality of kernel density estimators for chosen economic characteristics is of the same type but the asymmetry is quite different.
6) It can be also noticed that the length of bootstrap intervals for median (for numbers of employees and for revenues total) is much more bigger for the rural regions.
7) Presented statistical methods can be widely applied in comparisons for different countries as well as for regions in one country indicating the needs and possibility of development of analysed regions.

Bibliography

1. Baszczynska, A. (2016). *Parametr wygladzania w estymacji jadrowej funkcji gestosci dla zmiennych losowych w badaniach ekonomicznych (Smoothing Parameter in Kernel Density Estimation for Random Variables in Economic Researches)*. Wydawnictwo Uniwersytetu Lodzkiego. Lodz. pp. 15-113.
2. Baszczynska, A.. Pekasiewicz, D. (2008). Bootstrap Confidence Intervals for Population Mean in Case of Asymmetric Distributions of Random Variables. *Acta Universitatis Lodziensis Folia Oeconomica*. No. 216. Wydawnictwo Uniwersytetu Lodzkiego. Lodz. pp. 9-20.
3. Chernik, M. R. (2008). Bootstrap Methods: A Guide for Practitioners and Researchers. Hoboken New Jersey: John Wiley & Sons Ltd. pp. 26-78.
4. Chernik, M. R., LaBudde, R. A. (2011). An Introduction to Bootstrap Methods with Applications to R. Hoboken New Jersey: John Wiley & Sons Ltd. pp. 76-98.
5. Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press. pp. 191-202.
6. Domanski, Cz., Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne. Warszawa. pp. 260-274.
7. Domanski, Cz., Pruska, K., Wagner, W. (1998). *Wnioskowanie statystyczne przy nieklasycznych zalozeniach*. Wydawnictwo Uniwersytetu Lodzkiego. Lodz. pp. 109-148.
8. Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall. pp. 153-199.
9. Ghosh, S. (2018). *Kernel Smoothing. Principles. Methods and Applications*. Hoboken New Jersey: John Wiley & Sons Ltd. pp. 40-94.
10. Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Studies in Big Data. Volume 37. Springer International Publishing AG 2018. pp. 25-80.
11. Hall, p. (1988). Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics*. Volume 16. No. 3. pp. 927-953.
12. Härdle, W. (1991). *Smoothing Techniques. With implementation in S*. New York Berlin Heidelberg London: Springer-Verlag. pp. 44-48.
13. Hutson, A. (1999). Calculating Nonparametric Confidence Intervals for Quantiles Using Fractional Order Statistics. *Journal of Applied Statistics*. 26:3.pp343-353.
14. Kvam, P.H,. Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering*. Wiley Series in Probability and Statistics. New Jersey Hoboken: John Wiley & Sons. Inc. pp. 205-219.
15. Shao, J., Tu, D. (1995). *The* Jackknife and Bootstrap. New York: Springer-Verlag. pp. 129-140.
16. Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall. pp. 43-61.
17. Wand, M., Jones, C. (1995). *Kernel Smoothing*. London: Chapman and Hall. p. 11.