Analysis of the properties of selected inequality measures based on quantiles with the application to the Polish income data

Alina Jędrzejczak[1], Dorota Pekasiewicz[2]

Abstract

Quantiles of income distributions are often applied to the estimation of various inequality and poverty characteristics. The most popular synthetic inequality measures, including the Gini and Pietra indices, are based on the Lorenz curve, but also simple quantile ratios or quantile dispersion ratios can be utilized to compare incomes of different population groups. Some other measures of income inequality have been constructed using differences (or ratios) between population and income quantiles. The concentration curve and corresponding synthetic concentration coefficient proposed by Zenga, are also defined in terms of quantiles. In the paper, selected inequality measures based on deciles and quintiles are considered. The main objective was to compare statistical properties of different estimation methods for quantiles, including Bernstein and Huang –Brill estimators, with the classical quantile estimator based on a relevant order statistic. Several Monte Carlo experiments have been conducted to assess biases and mean squared errors of quantile estimators for different sample sizes under the lognormal or Dagum distributions assumed as a population model. The results of the experiments have been used to the estimation of inequality measures in Poland.

Keywords: *income, inequality, estimator, bias, mean squared error*

JEL Classification: C13, C15

1 Introduction

The early development of statistical tools for income inequality measurement dates back to the end of the XIXth century, when the works of Vilfredo Pareto came out, but the literature on the evaluation and measurement of economic inequality has remarkably grown over the last few decades. For a long time income inequality measures have been used mainly for descriptive purposes. Rapid development of sample surveys after the second world war, as well as the growing demand for high-quality estimates at low levels of aggregation, made it necessary to study the sampling properties of inequality measures. Nevertheless, in many applications the estimates of inequality measures are still presented without any information about their precision, which must be the basis for further statistical inference e.g. statistical hypothesis testing and interval estimation. The problem can be neglected to some extent when

[1] Corresponding author: University of Łódź/Institute of Statistics and Demography, Department of Statistical Methods, 90-214 Łódź, 41/43 Rewolucji 1905, e-mail: jedrzej@uni.lodz.pl.

[2] University of Łódź/Institute of Statistics and Demography, Department of Statistical Methods, 90-214 Łódź, 41/43 Rewolucji 1905, e-mail: pekasiewicz@uni.lodz.pl.

we consider the overall population or the sample size is large enough to apply the asymptotic theory; one should be conscious however, that for heavy-tailed income distributions the sufficient sample size can be very large indeed. For some population divisions (by age, occupation, family type or geographical area) estimators of inequality measures can be seriously biased and their standard errors can be far beyond the values that can be accepted by social policy-makers for making reliable policy decisions (Jędrzejczak, 2012). The situation seems even more complicated for nonlinear sample statistics including numerous inequality indices based on quantiles.

The paper addresses the problem of statistical properties of the estimators of popular inequality measures based on quantiles. After a brief description of such measures (section 2), selected quantile estimators have been introduced (section 3). The next section comprises the results of Monte Carlo experiments which have been conducted to assess biases and mean squared errors of quantile estimators and their functions. In the last part of the paper we present the application of quantile-based inequality indices to the Polish Household Budget Survey (HBS) data.

2 Statistical inequality measures based on quantiles

Distribution quantiles of a random variable *X* which is identified as a household or personal income, or the estimators of these quantiles, have been applied to the construction of simple inequality indices as quintile dispersion ratio and decile dispersion ratio (Panek, 2011).

The quintile dispersion ratio has the following form:

$$\mathrm{W}_{20:20}^{(1)} = \frac{Q_{0.8}}{Q_{0.2}}, \tag{1}$$

where $Q_{0.8}$, $Q_{0.2}$ are quintiles, respectively, the fourth and the first.

The quintile dispersion ratio can also be defined as the ratio of the sum of incomes of the richest 20 percent of the population to the sum of incomes of the poorest 20 percent:

$$\mathrm{W}_{20:20}^{(2)} = \frac{\sum_{i \in GK_5} x_i}{\sum_{i \in GK_1} x_i}, \tag{2}$$

where GK_j is *j*-th quintile group.

The measure (2) can be interpreted as the ratio of the average income of the richest 20 percent of the population to the average income of the poorest 20 percent of the population and it is usually calculated on the basis of equivalised income.

Similar ratios can also be calculated for other quantiles, for instance deciles or percentiles (95th and 5th) of income distributions. Using the first and ninth decile we can obtain the following decile dispersion ratio:

$$\mathrm{W}_{10:10}^{(1)} = \frac{Q_{0.9}}{Q_{0.1}}, \tag{3}$$

where $Q_{0.9}$, $Q_{0.1}$ are deciles, respectively, the ninth and the first

and

$$\mathrm{W}_{10:10}^{(2)} = \frac{\sum_{i \in GD_{10}} x_i}{\sum_{i \in GD_1} x_i}, \tag{4}$$

where GD_j is *j*-th decile group.

The reciprocal of the decile dispersion ratio defined by (4) takes values from the interval [0,1] and is called the dispersion index for the end portions of the distribution:

$$\mathrm{K}_{1:10} = \frac{\sum_{i \in GD_1} x_i}{\sum_{i \in GD_{10}} x_i} = \frac{1}{\mathrm{W}_{10:10}^{(2)}}. \tag{5}$$

If the index $\mathrm{K}_{1:10}$ is closer to the 1, the inequality is lower (mean incomes in the extremal decile groups are the same).

The examples of more sophisticated inequality measures are Gini and Zenga indices. The popular Gini index is not considered in this paper. The synthetic Zenga index is based on the concentration curve that can be considered point concentration measure, and thus becomes sensitive to changes at every “point” of income distribution. The Zenga point measure of inequality is based on the relation between income and population quantiles (Zenga, 1990; Greselin et al. 2012):

$$\mathrm{Z}_p = \frac{x_p^* - x_p}{x_p^*} = 1 - \frac{x_p}{x_p^*}, \tag{6}$$

where $x_p = F^{-1}(p)$ denotes the population *p*-quantile and $x_p^* = Q^{-1}(p)$ is the corresponding income quantile. Therefore the Zenga approach consists of comparing the abscissas at which *F*(*x*) and *Q*(*x*) take the same value *p*.

Zenga synthetic inequality index is defined as simple arithmetic mean of point concentration measures $\mathrm{Z}_p, p \in \langle 0,1 \rangle$.

3 Selected quantile estimators

Let X be a continuous random variable with distribution function F and let $Q_p = F^{-1}(p)$ be the p-quantile of the random variable X, where $p \in (0,1)$. If F is continuous and strictly increasing distribution function, the p^{th} quantile always exists and is uniquely determined.

The well-known estimator of the quantile Q_p is the statistic:

$$\hat{Q}_p = F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}, \tag{7}$$

where $F_n(x)$ is empirical distribution obtaining on the basis of a n-element random sample $X_1, X_2, ..., X_n$.

The problem of quantile estimation has a very long history. In the subject literature numerous nonparametric (distribution-free) quantile estimators have been presented. Their particular expressions depend on the underlying empirical distribution function definition.

Classical quantile estimator obtained for the distribution $F_n(x) = \frac{card\{1 \leq j \leq n : x_i \leq x\}}{n}$ for $x \in R$ is defined by the following formula:

$$\hat{Q}_p = \begin{cases} X_{(np)}^{(n)}, & \text{for } np \in N, \\ X_{([np]+1)}^{(n)}, & \text{for } np \notin N, \end{cases} \tag{8}$$

where $X_{(k)}^{(n)}$ is an order statistic of rank k.

Among other estimators of quantiles Q_p, we can mention the standard estimator, Huang-Brill estimator, Harrel-Davis estimator and Bernstein estimator, to name only a few (Huang and Brill, 1999; Harrell and Davis, 1982; Zieliński, 2006).

By means of the empirical distribution *level crossing*, which has the following form:

$$F_n(x) = \sum_{i=1}^{n} w_{n,i} I_{(-\infty,x\rangle}\left(x_{(i)}^{(n)}\right), \tag{9}$$

where

$$w_{n,i} = \begin{cases} \frac{1}{2}\left[1 - \frac{n-2}{\sqrt{n(n-1)}}\right] & \text{for } i = 1, n, \\ \frac{1}{\sqrt{n(n-1)}} & \text{for } i = 2, 3, ..., n-1, \end{cases}$$

we obtain the Huang-Brill estimator of the p^{th} quantile Q_p:

$$\hat{Q}_p^{HB} = X_{(b)}^{(n)}, \tag{10}$$

where

$$b = \left[\sqrt{n(n-1)} \left(p - \frac{1}{2}\left[1 - \frac{n-2}{\sqrt{n(n-1)}} \right] \right) \right] + 2. \tag{11}$$

It can easily be noticed that for $p = 0.5$ the estimator of the quantile $Q_{0.5}$ is the order statistic $X^{(n)}_{\left(\left[\frac{n}{2}\right]+1\right)}$.

Another interesting quantile estimator is the Bernstein estimator given by:

$$\hat{Q}_p^{Brs} = \sum_{i=1}^{n} \left[\binom{n-1}{i-1} p^{i-1}(1-p)^{n-i} \right] X^{(n)}_{(i)}. \tag{12}$$

More examples of quantile estimators can be found in the papers of Pekasiewicz (2015) and Zieliński (2006).

4 Analysis of Monte Carlo experiments

The main objective of the Monto Carlo experiments conducted in the study was to assess the properties of selected estimators of quantiles. We were especially interested in their biases and sampling variances, the components of their sampling errors. The following estimators have been taken into consideration: the classical quantile estimator (8), Huang-Brill estimator (10) and Bernstein estimator (12). The estimators presenting the best performance were further applied to evaluate the quantile-based inequality measures for income distributions in Poland.

In the experiments two different probability distributions were utilized as population models: two-parameter lognormal distribution, $LG(\mu,\sigma)$, defined by the following density function:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right), x > 0 \tag{13}$$

and three-parameter Dagum distribution $D(p,a,b)$, known also as the Burr type-III distribution, with the density function of the form (Kleiber and Kotz, 2003):

$$f(x) = ab^{-ap} p x^{ap-1} \left(1 + \left(\frac{x}{b} \right)^{a} \right)^{-p-1}, x > 0. \tag{14}$$

The sets of parameters of both theoretical distributions were established on the basis of real income data coming from Polish HBS and administrative registers, comprising large variety of subpopulations differing in the level of income inequality, which have been observed over the last two decades. The sample sizes were fixed for each variant as *n*=500; *n*=1000, *n*=2000. The number of repetitions of Monte Carlo experiment was *N*=20 000 (Białek, 2013). The simulated sample spaces were used to assess, for each estimator, its empirical bias and standard error.

Distribution	***p***	$\hat{Q}_p$		$\hat{Q}_p^{HB}$		$\hat{Q}_p^{Brs}$	
		BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$LG(8.0, 0.6)$	0.1	-0.087	3.240	0.254	3.248	0.132	3.165
	0.2	-0.079	2.718	0.139	2.726	0.108	2.669
	0.3	-0.039	2.504	0.133	2.511	0.095	2.481
	0.7	0.089	2.528	-0.082	2.521	0.042	2.469
	0.8	-0.077	2.712	-0.077	2.712	0.047	2.680
	0.9	-0.131	3.245	-0.131	3.245	0.041	3.169
$LG(8.3, 0.8)$	0.1	-0.097	4.350	0.359	4.373	0.302	4.220
	0.2	-0.088	3.581	0.195	3.592	0.177	3.571
	0.3	-0.057	3.336	0.176	3.346	0.134	3.271
	0.7	0.169	3.338	-0.061	3.324	0.108	3.280
	0.8	-0.099	3.620	-0.099	3.620	0.070	3.510
	0.9	-0.116	4.339	-0.116	4.339	0.089	4.208
$D(0.7,3.6,3800)$	0.1	-0.182	3.923	0.313	3.916	0.086	3.803
	0.2	-0.068	2.800	0.141	2.776	0.069	2.741
	0.3	-0.105	2.349	0.114	2.346	0.000	2.303
	0.7	0.010	2.054	-0.080	2.049	0.043	2.013
	0.8	-0.085	2.298	-0.078	2.287	0.032	2.256
	0.9	-0.083	2.984	0.116	2.991	0.121	2.915
$D(0.7,2.8,3800)$	0.1	-0.156	5.073	0.368	5.069	0.221	4.493
	0.2	-0.112	3.580	0.232	3.589	0.082	3.509
	0.3	-0.080	3.015	0.144	2.991	0.062	2.958
	0.7	0.137	2.652	-0.063	2.681	0.073	2.599
	0.8	-0.084	2.956	-0.077	2.935	0.069	2.900
	0.9	-0.133	3.846	-0.112	3.848	0.147	3.774

Table 1. Properties of selected quantile estimators for sample sizes *n*=1000.

Table 1 presents the results of the calculations for three quantile estimators: classical, Huang-Brill, and Bernstein, each of the following orders: *p*=0.1; 0.2; 0.3; 0.7; 0.8; 0.9. In particular, the table shows the relative biases and relative root mean squared errors (in %) of these estimators obtained for predefined population models- lognormal and Dagum - differing across the experiments in the overall inequality levels. The similar experiments for Gini and Zenga ratios were reported in Jędrzejczak (2015).

Analysing the results of the calculations it becomes obvious that the Bernstein estimator performs better than its competitors- its root mean squared errors (RMSE) are much smaller than those observed for the other quantile estimators and its relative biases (BIAS) are also smaller, especially when the quantiles of higher orders are taken into regard. The bias and RMSE of Huang-Brill estimator are similar to the respective values for the classical quantile estimator. It is worth noting that for all cases biases are rather negligible so the total errors are dominated by sampling variances. In general, the estimation errors are higher for extremal quantile orders, for the heavy-tailed Dagum model and they also tend to increase as income inequality increases. The three types of quantile estimators mentioned above were then used to the simulation study concerning income inequality measures: $W_{10:10}^{(1)}$ and $W_{20:20}^{(1)}$ given by the formulas (1) and (3). The properties of decile dispersion ratios have been demonstrated in table 2. The results obtained for quintile dispersion ratios show similar regularities.

Distribution	$\mathbf{W_{10:10}^{(1)}}$ **(stand.)**		$\mathbf{W_{10:10}^{(1)}}$ **(Huang-Brill)**		$\mathbf{W_{10:10}^{(1)}}$ **(Bernstein)**	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$LG(8.0, 0.6)$	0.065	4.327	-0.324	4.304	0.017	4.191
$LG(8.1, 0.7)$	0.084	5.088	-0.273	5.028	0.019	4.926
$LG(8.3, 0.8)$	0.124	5.815	-0.352	5.766	0.037	5.615
$D(0.7, 3.6, 3800)$	0.162	4.702	-0.211	4.651	0.082	4.543
$D(0.8, 3.0, 3200)$	0.097	5.179	-0.266	5.193	0.039	5.009
$D(0.7, 2.8, 3800)$	0.181	6.003	-0.298	5.948	0.066	5.800

Table 2. Properties of Decile Dispersion Ratio based on quantile estimators for *n*=1000.

5 Application

The inequality measures based on deciles and quintiles, as well as the Zenga indices, have been applied to the inequality analysis in Poland by macroregion (NUTS1), based on HBS sample 2014. To obtain the reliable estimates of these coefficients we used the Bernstein quantile estimator which turned out to have the highest precision (tables 1 and 2). Basic characteristics of the HBS sample, divided by macroregions, are presented in table 3, while table 4 contains the results of the approximation of the empirical distributions by means of the Dagum model. We can observe very high consistency of the empirical distributions with the theoretical ones (see table 4 and figures 1 – 2).

Macroregion	Number of households	Minimum	Maximum	Average	Standard Deviation
I	8046	11.00	155017.49	4240.21	3790.53
II	7433	12.50	37152.00	3634.03	2179.59
III	6246	10.00	84032.90	3461.45	2876.23
IV	5658	3.00	43493.45	3772.15	2611.00
V	3971	1.67	37200.00	3591.07	2337.83
VI	5575	9.00	126739.54	3646.44	3225.72
Total	36929	1.67	155017.49	3755.33	2959.95

Table 3. Numerical characteristics of income in macroregions.

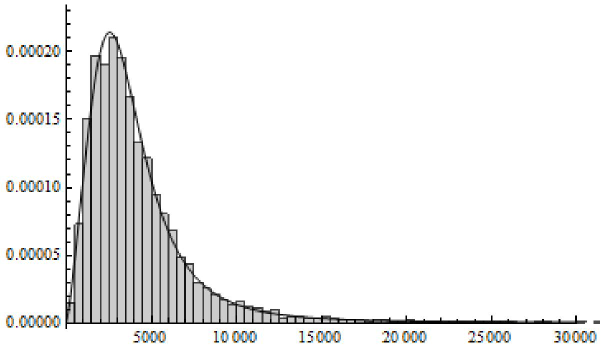

Fig. 1. Income distributions for macroregions I and fitting by means of the Dagum model.

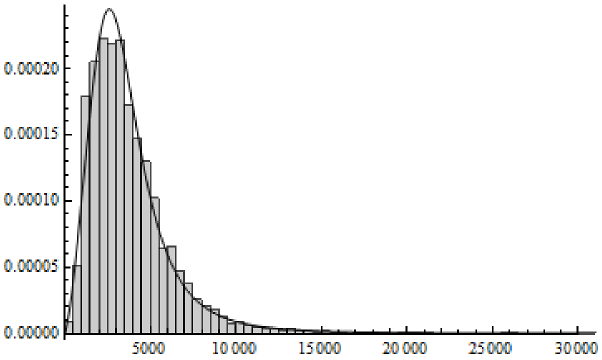

Fig. 2. Income distributions for macroregions IV and fitting by means of the Dagum model.

Macroregion	Dagum distribution parameters			Overlap
	p	***a***	***b***	**measure**
I	0.790	2.804	3839.630	0.982
II	0.669	3.618	3800.167	0.970
III	0.756	3.051	3286.467	0.971
IV	0.743	3.233	3687.076	0.964
V	0.722	3.301	3587.800	0.970
VI	0.718	3.158	3544.934	0.979
Total	0.747	3.125	3611.017	0.975

Table 4. Approximation of income distributions for macroregions.

The basic results of inequality analysis have been outlined in table 5. The estimated values of quintile and decile share ratios, as well as the values of synthetic Zenga inequality indices, indicate the central macroregion (I) as the one with the highest income inequality level. It is particularly evident for extremal income groups, i.e. income of the richest 10 percent of households is 12 times bigger than the income of the poorest 10 percent. The southern macroregion (II) presents the lowest values of all inequality measures except for the *K* index.

Macroregion	$\mathbf{W}^{(1)}_{20:20}$	$\mathbf{W}^{(2)}_{20:20}$	$\mathbf{W}^{(1)}_{10:10}$	$\mathbf{W}^{(2)}_{10:10}$	$\mathbf{K}_{1:10}$	Zenga
I	3.049	6.939	5.494	12.085	0.083	0.386
II	2.595	4.962	4.283	7.577	0.132	0.269
III	2.904	6.147	4.927	9.908	0.101	0.348
IV	2.750	5.577	4.742	8.614	0.116	0.308
V	2.789	5.375	4.536	8.172	0.122	0.295
VI	2.828	6.039	4.814	9.841	0.102	0.347
Total	2.819	5.916	4.843	9.526	0.105	0.338

Table 5. Inequality measures for macroregions.

Conclusion

Analysis of income and wage distribution is strictly connected with the estimation of inequality and poverty measures based on quantiles. Therefore, for income data coming usually from sample surveys it becomes crucial to use the quantile estimators presenting satisfying statistical properties. In the paper, the Huang-Brill and Bernstein estimators have been proposed and analysed from the point of view of their sampling errors under several income distribution models. In the simulations studies the properties of these estimators have been compared with the classical one which is most often applied in practice. The results of the calculations reveal that the Bernstein estimator performs better than its competitors- its root mean squared errors (RMSE) are much smaller than those observed for the other quantile estimators and its relative biases (BIAS) are also smaller, especially when the quantiles of higher orders are taken into regard. Consequently, the Bernstein estimator has been applied to the estimation of various inequality measures in regions NUTS 1 in Poland.

References

Białek, J. (2014). Simulation Study of an Original Price Index Formula. *Communications in Statistics, Simulation and Computation, 43(2),* 285–297.

Greselin, F., Pasquazzi, L., & Zitikis, R. (2012). Contrasting the Gini and Zenga indices of economic inequality. *Journal of Applied Statistics, 40(2),* 282–297.

Harrell, F. E., & Davis, C. E. (1982). A New Distribution-Free Quantile Estimator. *Biometrika, 69,* 635–640.

Huang, M. L., & Brill, P. H. (1999). A Level Crossing Quantile Estimation Method. *Statistics & Probability Letters, 45*, 111–119.

Jędrzejczak, A. (2012). Estimation of Standard Errors of Selected Income Concentration Measures on the Basis of Polish HBS. *International Advances in Economic Research, 18(3)*, 287-297.

Jędrzejczak, A. (2015). Asymptotic Properties of Some Estimators for Gini and Zenga Inequality Measures: a Simulation Study. *Statistica & Applicazioni*, *13,* 143-162.

Kleiber, C., & Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, Hoboken.

Panek, T. (2011). *Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru*. Szkoła Główna Handlowa, Warszawa.

Pekasiewicz, D. (2015). *Statystyki pozycyjne w procedurach estymacji i ich zastosowania w badaniach społeczno-ekonomicznych*. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Zieliński, R. (2006). Small-Sample Quantile Estimators in a Large Nonparametric Model. *Communications in Statistics Theory and Methods, 35,* 1223–1241.

Zenga, M. (1990). Concentration Curves and Concentration Indices Derived from Them. *Income and Wealth Distribution, Inequality and Poverty,* Springer -Verlag, Berlin, 94–110.