tween a supply and demand of a labour. The active state policy on a labour market is expedient for sharing in view of distinctive elements and the reasons of occurrence of existing disproportions which demands the differential strategic approach. All these problems it is possible to solve in the different ways, and one of them, — use of the complex approach to a statistical estimation of functioning of a labour market in regions of the country.

Bibliography

1. *Блинова Т. В.*, *Русановский В. А.* Экономическая политика, структура занятости и безработицы в российских регионах. — М.: РПЭИ, 2002. — 46 с.

2. *Єріна А. М.* Статистичне моделювання та прогнозування: навч.посібник. — К.: КНЕУ, 2001. — 170 с.

3. *Карышев М. Ю.* Социальная безопасность России: региональный аспект статистической оценки// Вопросы статистики. — 2003. — № 2. — С. 41—46.

4. *Чернюк Л. Г.*, *Клиновий Д. В.* Економіка та розвиток регіонів (областей) України. Навчальний посібник. — Київ: ЦУЛ, 2002. — 644 с.

Eva Sodomová,
Ľubica Sipková,
Branislav Pacák,
University of Economics in Bratislava

MODELS OF HOUSEHOLDS' IN THE SLOVAK REPUBLIC

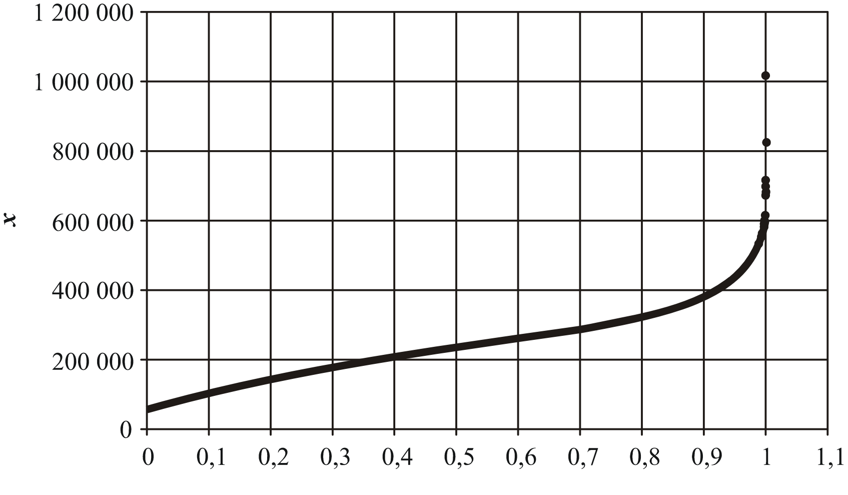
1. Introduction

The article includes the methods and the results of the statistical analysis and modelling of distribution of incomes of households in the Slovak Republic in 2002. INPUT empirical data are the sampling data about the yearly real net income of 1563 households, found by Statistical Office of the Slovak Republic. Empirical distribution of sampling data and its description is the basis of the modelling of the probability distribution of the household incomes.

2. Graphical analysis of the sampling data

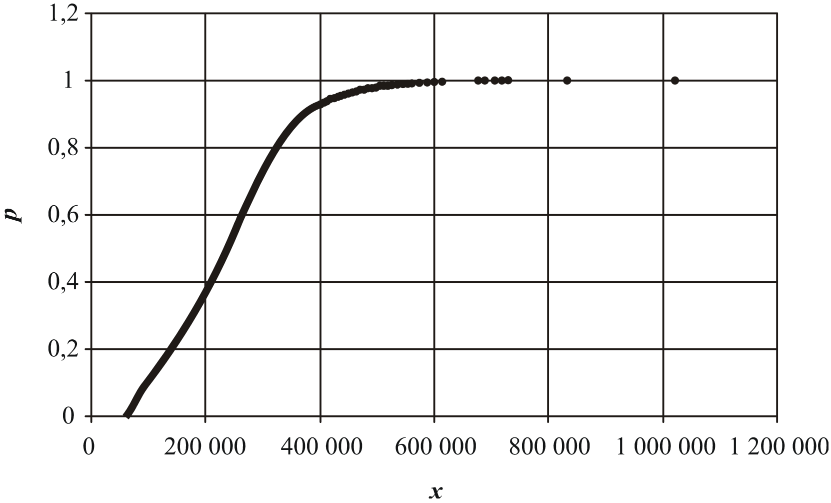
The first step of the data description is graphical analysis. We will analyze the variable *X* (*CP*) — *net income of household in the Slovak Republic in 2002*. Before we can sensibly model a set of data, we need to have a clear perception of it, otherwise, we will find ourselves imposing our views of how it ought to behave, rather than finding out how it does behave. The best way to develop this perception and feel for the data is graphically, with the support of some numerical summaries of the properties seen on the graphs. Let as look at a set of data to illustrate a number of different approaches.

The original data *x* has been sorted by magnitude. This is a fundamental step in the types of analysis that will be developed in this article. With each ordered value, x, we will associate a probability p, indicating, that the x lies in proportion to p, of the way through the data. At first guess, we would associate the rth order observation, denoted by $x_{(r)}$, with $p_r = r/n$, $n = 1, 2, ..., 1566$. The range of values we would expect for p is $(0,1)$, The value of r/n goes from $1/1566$ to 1, i.e.; it is not symmetrical. Hence, to get the p values symmetrically placed in the interval $(0,1)$, we use the formula $p_r = (r - 0{,}5)/n$. This formula corresponds to breaking the interval $(0,1)$ into 1566 equal sections and using the midpoint of each. Thus, we have pairs of values describing the data as $(x_{(r)}, p_r)$. The value of *x* for any *p* is referred to as the *sample p — quantile.*



Source: own calculations

Figure 1. Empirical quantile function



Source: own calculations

Figure 2. Empirical distribution function

There are two natural plots of such data. First (figure 1), of $x_{(r)}$ against p_r (empirical quantile function) and second (Figure 2), of p_r against $x_{(r)}$ (empirical distribution function), which are just the same plot with the axes interchanged. However, a look at these plots shows that different features stand out more clearly in different plots. The difference in slope, between high values and low values of x, looks to be a much more natural feature of the data.

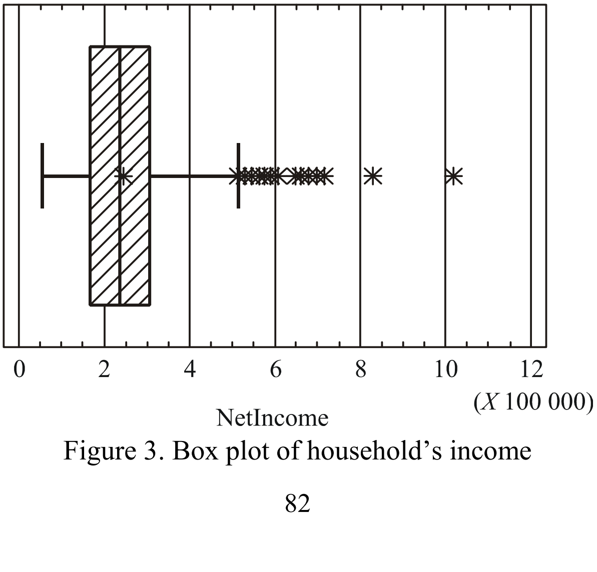


Figure 3. Box plot of household’s income

The fact that the length of the right «whisker» is approximately twice as long as the left one, and 29 high values, two of which exceed even the level of three times the quartile range, give a clear evidence of a long right tail of the distribution.

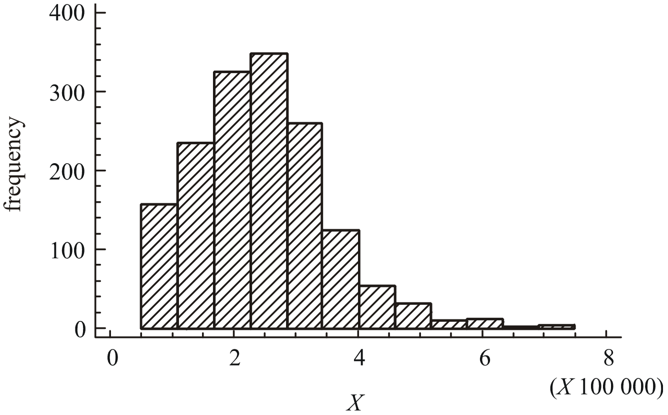


Figure 4. Histogram of household income

Frequency histogram of empirical distribution (Figure 4) proves the conclusions obtained from the box-and-whisker plot. It clearly shows a cut off bottom side of the distribution.

The quantitative expression of characteristic properties, detected in graphs, is completed by estimation of the descriptive statistics of the sample.

```
Count = 1566                          Lower quartile = 166663.0
Average = 244284.0                    Upper quartile = 306433.0
Median = 237407.0                     Interquartile range = 139770.0
Mode =                                Skewness = 1.04975
Variance = 1.20118E10                 Stnd. skewness = 16.9592
Standard deviation = 109598.0         Kurtosis = 2.99699
Minimum = 54325.0                     Stnd. kurtosis = 24.2089
Maximum = 1.0206E6                    Coeff. of variation = 44.8652 %
Range = 966275.0                      Sum = 3.82549E8

Quantiles for CP                      Octiles for CP
1.0 % = 65235.0                       12.5 % = 125460.0
5.0 % = 85176.0                       25.0 % = 166663.0
10.0 % = 107901.0                     37.5 % = 202124.0
25.0 % = 166663.0                     50.0 % = 237407.0
50.0 % = 237407.0                     62.5 % = 270816.0
75.0 % = 306433.0                     75.0 % = 306433.0
90.0 % = 377556.0                     87.5 % = 358387.0
95.0 % = 432913.0                     95.0 % = 432913.0
99.0 % = 584366.0                     99.0 % = 584366.0
```

Figure 5. Descriptive statistics of the net income distribution

3. Models of the Population

We are often seeking to achieve some theoretical model that has a structure similar to that shown by the sampling data. We want to link the sample features to corresponding features in the populations from which our sample comes. We therefore need to turn to ways of describing a population.

We will describe five models for the random variation in population. Our definitions kept to the case where the random variable is continuous.

1. The *Cumulative Distribution Function*, CDF, denoted by $F(x)$, is defined as

$$F(x) = P(X \leq x) = p\ .$$

The plot of $\mathrm{F}(\mathrm{x})$ corresponds to the sample plot of p against x.

2. The *Probability Density Function*, PDF, denoted by $f(x)$, defines the relation

$$f(x) = \frac{dF(x)}{dx}$$

is the derivative of the CDF.

3. The *Quantile Function*, QF, denoted by $Q(p)$, expresses the p-quantile x_p as a function of p:

4. $x_p = Q(p)$ is the value of x for which $p = P(X \leq x_p) = F(x_p)$.[22]

5. The definitions of the QF and the CDF can by written for any pairs of values (x, p) as $x = Q(p)$ and $p = F(x)$. These functions are thus simple inverses of each other, provided that they are both continuous increasing functions. Thus, we can also write $Q(p) = F^{-1}(p)$ and $F(x) = Q^{-1}(x)$. For sample data, the plot of $Q(p)$ is the plot of x against p.

In the same way that the CDF can be differentiated to give the PDF, we can use the derivative of the QF.

6. The *Quantile Density Function*, QDF, is defined as

7. $q(p) = \dfrac{dQ(p)}{dp}$ for $0 \leq p \leq 1$.

4. Modelling of net income of households using goodness of fit tests

8. The general conclusion from the sampling is that distribution of the variable *X*— net incomes of households is *skew and long tailed.* We have used Pearson χ^2 goodness of fit test to find the suitable distribution that can be used to describe the variation in household's net income. The STATGRAPHICS *Plus* package contains a few on the kinds of distribution, but not one of them is perfect as a probability model for the variable *X*. The best of all, is the Weibull distribution with maximum likelihood

[22] The first paper to systematically develop quantile functions was by Parzen (1979).

estimation of the parameters. The results of χ^2 test and Kolmogorov-Smirnov (K-S) test present the table on Figure 5. Figure 6 shows the fitted Weibull density probability plot, superimposed on the histogram on Figure 4, as a result of Statgraphics' procedure *Distribution fitting.*

```
Goodness-of-Fit Tests for CP

                               Chi-Square Test
----------------------------------------------------------------------------
               Lower          Upper       Observed      Expected
               Limit          Limit      Frequency     Frequency    Chi-Square
----------------------------------------------------------------------------
         at or below        50000,0              0         27,86         27,86
             50000,0        95833,3            122         96,97          6,46
             95833,3       141667,0            143        169,90          4,26
            141667,0       187500,0            239        225,08          0,86
            187500,0       233333,0            256        248,72          0,21
            233333,0       279167,0            273        237,75          5,23
            279167,0       325000,0            233        199,73          5,54
            325000,0       370833,0            129        148,54          2,57
            370833,0       416667,0             72         98,12          6,95
            416667,0       462500,0             39         57,61          6,01
            462500,0       508333,0             23         30,07          1,66
            508333,0       554167,0             14         13,93          0,00
above       554167,0                            20          8,72         14,59
----------------------------------------------------------------------------
Chi-Square = 82,2094 with 10 d.f.   P-Value = 1,84963E-13

Estimated Kolmogorov statistic DPLUS = 0,0415772
Estimated Kolmogorov statistic DMINUS = 0,0270346
Estimated overall statistic DN = 0,0415772
Approximate P-Value = 0,00899895
```

Figure 6. Results of Goodness-of-Fit Tests

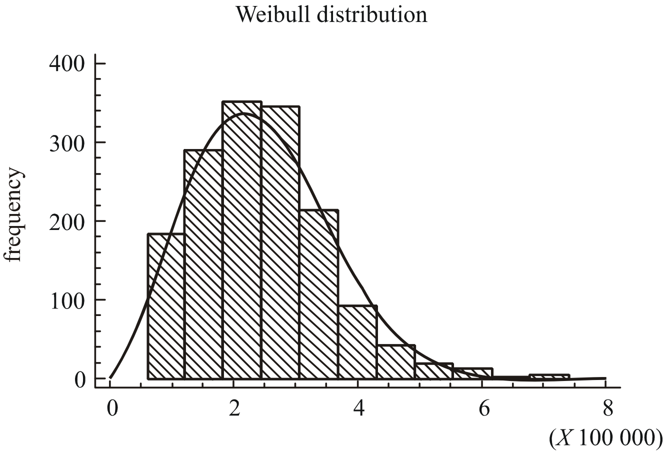


Figure 7. Fitted Weibull density probability plot

In K-S test, we do not reject a hypothesis of the tested distribution on a level of significance approximately 0.01, because p-value = 0.00899895. Results of χ^2 — test (Figure 6), high value of the statistic χ^2= 82.2094 and a very low p-value = 1.84963E-13 lead us to reject the hypothesis of the Weibull distribution. If we compare the observed and expected frequencies, we see that the reason for obtaining such a high value of χ^2 is the significant difference between observed and expected frequencies in intervals with lowest and highest incomes. This conclusion is also supported by Figure 7.

We can try to find a flexible model of distribution which also better models the intervals of lowest and highest incomes using quantile functions.

5. Modelling of net income of households with quantile functions

The basis for the approach to statistical modelling based on quantile function models is that quantile functions can be added and, in the right circumstances, multiplied to obtain new quantile functions. Quantile density functions can also be added together to derive new quantile density functions.

Using a modelling kit based on quantile function models, we have elaborated a four parametric composite *Weibull-Paretovo distribution*, denoted **WPD(λ,η,β,γ)** for further study. Its quantile function is

$$Q(p) = \lambda + \eta \left\{ (1-p)\left[-\ln(1-p)\right]^{\beta} + p\left[\frac{1}{(1-p)^{\gamma}}\right] \right\},$$

$$\text{kde} \quad 0 < p < 1, \beta > 0, \gamma > 0, \tag{1}$$

λ is the parameter of location, η is the scale parameter, β a γ are the parameters of shape.

Using approximate generalized Öztürk and Dale method for estimation of parameters[23] with the help of the optimalization program Solver in Excel we have composed the following quantile probability function:

$\mathbf{WPD_{OD}}$(51700,63612470; 194314,53494018; 0,68722619; 0,21221575).

To put the funded parameters to expression (1), we have obtained the analytical form of the model:

$$Q(p) = 51700{,}63612470 + 194314{,}53494018 \cdot$$

$$\cdot \left\{ (1-p)\left[-\ln(1-p)\right]^{0{,}68722619} + p\left[\frac{1}{(1-p)^{0{,}21221575}}\right] \right\}.$$

[23] Gilchrist 2000, p. 198, [9].

By minimizing the distributional least absolutes criterion (again in Solver of Excel), we obtained the model

WPDA(53935,08339853; 186704,81277537; 0,65685205; 0,23298934)

with analytical form by (1)

$$Q(p) = 53935{,}08339853 + 186704{,}81277537 \cdot .$$

$$\cdot \left\{ (1-p)\left[-\ln(1-p)\right]^{0{,}65685205} + p\left[\frac{1}{(1-p)^{0{,}23298934}}\right] \right\}.$$

6. Validation of models

Validation is the process of deciding whether an identified and fitted model is indeed valid for the data and the situation under consideration. There are many aspects of validation, but here we concentrate only on those of direct relevance to modelling with quantiles.

The natural plot to start with is the fit-observation diagram or Q—Q plot of the validation data against the corresponding fitted values.

As seen on Q — Q plot (Figure 8), both Weibull-Pareto quantile distributions fit well with the empirical distribution of income. Depending on a selected method of parameters estimation, the two models fit better in different parts of the distribution.

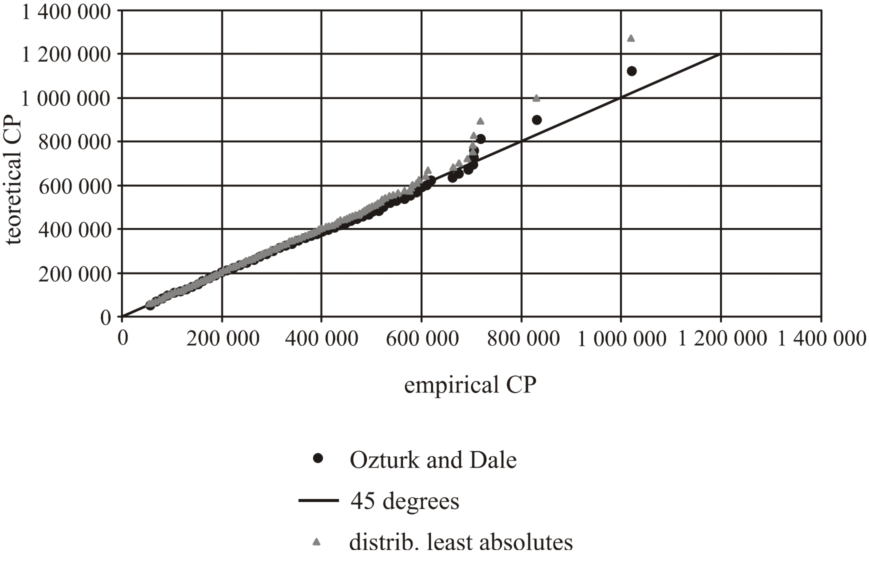


Figure 8. Q—Q plot of Weibull-Pareto distributions

Sometimes the look of two plots is very similar and it is useful to have some numerical measures to supplement the analysis. The suitable measure of the quality of the fit is the correlation between the fitted values and the actual values. In the case of both models WPD, their values are high (0,9963 a 0,9982) signalling very strong linear correlation.

The most often method of testing of proposed models against validation data is the goodness-of-fit test modified for quantile models.

Suppose for simplicity, we divide the p-axis into m equal sections using $p_j = j/m$, j =1, 2, ... , m-1; p_0 = 0, p_m = 1. If $x_j = Q(p_j)$, $x_{j-1} = Q(p_{j-1})$ and n_j is the number of observations in the new data set lying in the interval $\left(x_{j-1}, x_j \right\rangle$, than the expected value of n_j is *n/m* for all *j*. This fact is used to construct the test statistic C where

$$C = \sum_{j=1}^{m} \frac{\left(n_j - \frac{n}{m} \right)^2}{\frac{n}{m}} \qquad \text{j = 1, 2, ... , m.}$$

This statistic has approximately a χ^2 distribution and in this case it has *m*-1 degrees of freedom. If the new data is very different from the fitted model, the value of C will be larger than indicated by a χ^2 distribution.

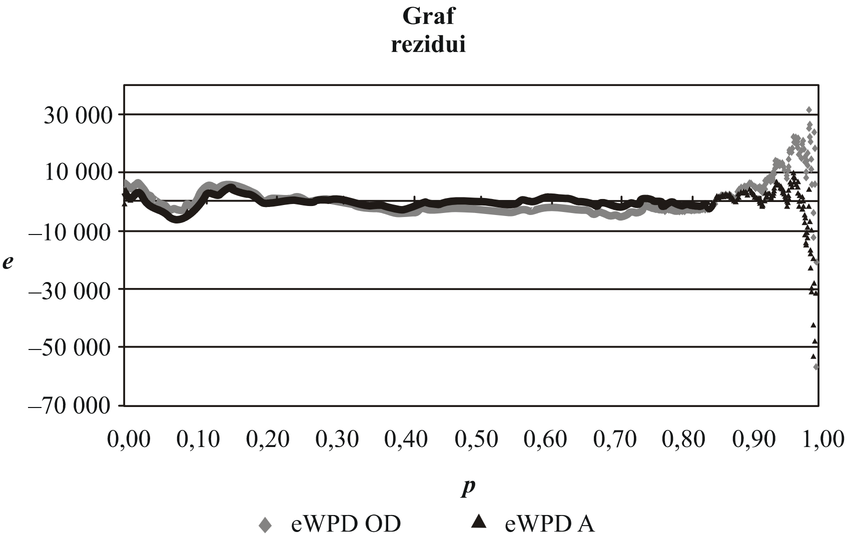


Figure 9. Residual plots of Weibull-Pareto distributions

There is C = 28.99 for model WPD_A and C = 28, 64 for WPD_{OD}. If level of significance $\alpha = 0.05$, than critical value of χ^2 distribution of statistic C — $\chi^2_{0,95}(19) = 30,14$ confirm the good fit for the both models of WPD.

Plot of residuals (Figure 9) shows that WPD_A model is better fit for 80 % values of household income. Only 2 % of values exceed 10 thousand SK.

Table 1

COMPARISON OF SUMMARY STATISTICS

Summary statistics	Empirical data	WPD_{OD}	WPD_A
Minimum value	54 325	52 532	54 938
Lower percentile	65 388	61 722	64 733
Tenth percentile	107 956	108 792	111 364
Lower quartile	166 663	165 241	165 621
Median	237 407	239 865	237 112
Upper quartile	358 324	308 089	305 196
Ninetieth percentile	377 394	371 291	373 608
Upper percentile	584 179	569 548	600 799
Maximum value	1 020 600	1 124 060	1 271 774
Variance of residuals		44 080 072	104 020 307
Std. Dev. of residuals		6 639	10 199
Coefficient of Correlation	x	0,9982	0;9963

We can compare numerical statistics of the empirical data and of both models WPD_A and WPD_{OD} by table 1 to review the quality of the models.

Summary

The article includes the methods and the results of the statistical analysis and modelling of income distribution of households in the Slovak Republic in 2002. The first step of the data description is the graphical analysis. The quantitative expression of characteristic properties, detected in the graphs, is completed by estimation and interpretation of the descriptive statistics of the sample.

Results from the analysis of the sampling data — the distribution of the variable CP — yearly real net income of the households is a positive asymmetrical long-tailed distribution.

Chi — square and Kolmogorov-Smirnov goodness-of— fit tests of the empirical distribution with more of theoretical distributions confirm the best of all is Weibull distribution with maximum likelihood parameters estimation, except the intervals of the lowest and the highest values of incomes.

Two fitting distributions from the class of the Weibull-Pareto, which are good models for intervals of the lowest and the highest values of incomes, were found by the methods of the modelling with quantile distributions.

Bibliography

1. *Arnold, B. C.*, *Balakrishnan, N.*, and *Nagaraja, H. N.*: A First Course in Order Statistics, John Wiley and Sons, New York, 1992.

2. *Balakrishnan, N.* — COHEN, A. C.*:* Order Statistics and Inference, Academic Press, San Diego, 1991.

3. *Champernowne, D. G.,*: A Model of Income Distribution, Economic Journal 63, June, 318—351, 1953.

4. *Cheng CH. — Parzen, E.*: Unified estimators of smooth quantile and quantile density functions, Journal of Statistical Planning and Inference 59, 291—307, 1997

5. *Chipman, J. S.*: The Theory and Measurement of Income Distribution, Advances in Econometrics, 4, 135—165, 1985

6. *Dagum, C.*: Income Distribution Models, The Encyclopaedia of Statistical Science, 4, 27—34, 1984.

7. *Fowlkes, E. B.*: A Folio of Distributions, A collection of Theoretical Q-Q plots, Marcel Deckker, New York, 1987.

8. *Gilchrist, W. G.*: Modelling with quantile distribution functions, Journal of Applied Statistics, Vol.24, No. 1, 113—122, 1997.

9. *Gilchrist, W. G.*: Statistical modelling with quantile functions, Chapman & Hall, 2000.

10. *Moors, J. J. A*, *Wagemakers, R. Th. A.*, *Coenen, V. M. J.*, *Heuts, R. M. J.*, *Janssens, M. J. B. T.*: Characterizing systems of distributions by quantile measures, Statistica Neerlandica, Vol. 50, nr. 3, 417—430, 1996.

11. *Oztürk, A.* and *Dale, R. F.*: A study of fitting the generalized lambda distribution to solar radiation data, Journal Appl. Meteor., 21, 995, 1982.

12. *Pacáková, V.*: Aplikovaná poistná štatistika**.** ELITA, Bratislava, 2000.

13. *Pacáková, V.a kol.*: Štatistika pre ekonómov. IURA EDITION, Bratislava, 2003.

14. *Pacáková, V.-Sodomová, E.*: Modelling with Quantile Distribution Functions, Ekonomika a informatika, 1/2003, 30—44, Bratislava, 2003.

15. *Parzen, E.*: Nonparametric statistical data modelling, Journal of Amer. Statist. Assoc., 74, 105—121, 1979

16. *Ramberg, J. — Dudewicz, E. — Tadikamalla, P. — Mykytka, E.*: A propability distribution and its uses in fitting data, Technometrics, 21(2), pg. 201—214, 1979.

17. *Sipková, Ľ.*: Zovšeobecnené lambda rozdelenie a odhad jeho parametrov, Ekonomika a informatika, 1/2004, 107—128, Bratislava 2004.

Antonina Sidorova

INFLUENCE OF HUMAN SERVICES ON PARAMETERS' OPTIMIZATION OF POPULATION REPRODUCTION REGIME

The condition of human services in Donetsk region is analyzed at standpoint of economic, social and demographic functions execution. The article denotes that deformation in human services structure and small volume consumption of some kind of services keeps back the society development, makes worse demographic situation both in state and regions, and aggravates the problem of labour potential forming. The quantitative evaluation of human services on the indices of population reproduction of Donetsk region is performed in the article.

The process of formation of market economy in Ukraine has changed greatly the structure of consumer market as a whole and the services industry structure in particular. Donetsk region based human services enterprises can serve as an example to prove the statement. One of the largest Ukrainian regions, Donetsk region occupies the central place in social and economical development of the country. Approximately 10 % of country population lives there. The region consists of 27 cities and 17 areas. Donetsk city is a regional center with more than 1 million inhabitants.

Considerable manufacturing potential of the region determines its high urbanization level. The region owns more than one fifth of Ukrainian main production-manufacturing funds and supplies 20 % of manufacturing produce of Ukraine.

At the present moment Donetsk region experiences the process of modern human service enterprises network formation. Thus, for example, the total amount of the services provided showed a 16.8 % increase in terms of factual prices from 2002 to 2003. Such a positive dynamics is a result of an overall improvement in Ukrainian economy, reformation of forms of ownership, implementation of market mechanisms. Besides this, such factors as price policy changes and faults — an increase of transportation, communication, public services tariffs etc — also determine the growth of amount of the services provided, that has resulted in a deviation of factual services consumption structure.

The structure of services consumption is characterized by two considerable disproportions. Transportation (55.4 %), postal and communication services (12.8 %), real estate operations (7.3 %) and educa-