Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(6), 1–23 https://doi.org/10.59139/ws.2023.06.1
Studia metodologiczne / Methodological studies
Zgłoszony/submitted: 12.10.2022, poprawiony/revised: 21.04.2023, zaakceptowany/accepted: 21.05.2023

Modelling income distributions based on theoretical distributions derived from normal distributions

Piotr Sulewski,[a] Marcin Szymkowiak[b]

Abstract. In income modelling studies, such well-known distributions as the Dagum, the lognormal or the Zenga distributions are often used as approximations of the observed distributions. The objective of the research described in the article is to verify the possibility of using other type of distributions, i.e. asymmetric distributions derived from normal distribution (ND) in the context of income modelling. Data from the 2011 EU-SILC survey on the monthly gross income *per capita* in Poland were used to assess the most important characteristics of the discussed distributions. The probability distributions were divided into two groups: I – distributions commonly used for income modelling (e.g. the Dagum distribution) and II – distributions derived from ND (e.g. the SU Johnson distribution). In addition to the visual evaluation of the usefulness of the analysed probability distributions, various numerical criteria were applied: information criteria for econometric models (such as the Akaike Information Criterion, Schwarz's Bayesian Information Criterion and the Hannan-Quinn Information Criterion), measures of agreement, as well as empirical and theoretical characteristics, including a measure based on quantiles, specifically defined by the authors for the purposes of this article. The research found that the SU Johnson distribution (Group II), similarly to the Dagum distribution (Group I), can be successfully used for income modelling.

Keywords: income modelling, EU-SILC, normal distribution, SU Johnson distribution, Dagum distribution

JEL: C13, C15, C55, D31

Modelowanie rozkładu dochodów z wykorzystaniem rozkładów teoretycznych wywodzących się z rozkładu normalnego

Streszczenie. W badaniach nad modelowaniem dochodów do aproksymacji ich rozkładów bardzo często wykorzystuje się takie znane rozkłady, jak Daguma, log-normalny czy Zengi. Celem badania omawianego w artykule jest sprawdzenie możliwości posłużenia się innymi

[a] Akademia Pomorska w Słupsku, Instytut Nauk Ścisłych i Technicznych, Polska / Pomeranian University in Słupsk, Institute of Exact and Technical Sciences, Poland. ORCID: https://orcid.org/0000-0002-0788-6567. Autor korespondencyjny / Corresponding author, e-mail: piotr.sulewski@apsl.edu.pl.

[b] Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Polska / Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Poland.
ORCID: https://orcid.org/0000-0003-3432-4364. E-mail: marcin.szymkowiak@ue.poznan.pl.

typami rozkładów, tj. rozkładami asymetrycznymi wywodzącymi się z rozkładu normalnego (ND), w kontekście modelowania dochodów. Najważniejsze charakterystyki rozpatrywanych rozkładów określono na podstawie danych z badania EU-SILC 2011 dotyczących miesięcznego dochodu brutto na mieszkańca w Polsce. Rozkłady prawdopodobieństwa podzielono na dwie grupy: I – rozkłady powszechnie stosowane do modelowania dochodów (np. rozkład Daguma) i II – rozkłady wywodzące się z ND (np. rozkład SU Johnsona). Oprócz wizualnej oceny przydatności analizowanych rozkładów prawdopodobieństwa zastosowano kryteria liczbowe, takie jak: kryteria informacyjne dla modeli ekonometrycznych (Akaike Information Criterion, Schwarz's Bayesian Information Criterion oraz Hannan-Quinn Information Criterion), miary zgodności oraz charakterystyki empiryczne i teoretyczne, w tym specjalnie zdefiniowana na potrzeby artykułu autorska miara wykorzystująca kwantyle. Jak wynika z badania, rozkład SU Johnsona (II grupa), może być – tak jak rozkład Daguma (I grupa) – z powodzeniem wykorzystany do modelowania dochodów.

Słowa kluczowe: modelowanie dochodów, EU-SILC, rozkład normalny, rozkład SU Johnsona, rozkład Daguma

1. Introduction

Finding a proper model of income distribution and, consequently, examining and explaining income inequality, has been a task undertaken by economists since the times of Pareto. One of the main directions of research in the area of income distribution is the search for a theoretical model describing empirical income distributions.

Firstly, a theoretical model simplifies the analysis. When a small number of parameters is used, the estimation of different characteristics of the distribution and studying the properties of these characteristics expressed as the functions of certain parameters of the theoretical distribution become possible. Secondly, a well-fitted theoretical model allows the prediction of income distributions across different domains, both in time and space. It can be used e.g. in small area estimation as part of the model-based approach (Pratesi, 2016). Thirdly, approximations of empirical income distributions based on appropriately chosen theoretical distributions can compensate for irregularities resulting from the data collection method.

Many authors who study income distribution propose a set of economic, econometric, stochastic and mathematical properties considered as criteria used to select a particular mathematical model of income distribution. The final choice of the model depends on the degree to which it is capable of satisfying these criteria (Jędrzejczak & Trzcińska, 2018).

Aitchison and Brown (1957) defined the fundamental properties that form the most representative model of any stochastic process that generates an income distribution. The same issue was analysed by Dagum (1977) and Metcalf (1972). The suggested properties characterise a desirable income distribution model, including its foundation, interpretation, flexibility and inferential properties (Jędrzejczak,

2006). Among the most important features is the convergence to the Pareto principle for high income groups, a small number of finite moments of a distribution (heavy tails), goodness-of-fit (GoF) for the whole range of a distribution, a simple interpretation of parameters, and simplicity (or a small number of parameters).

Normal distribution (ND) is certainly the best known representative of the family of distributions used in statistical theory and practice to model distributions of phenomena defined as positive or non-negative, characterised as symmetric (e.g. people's weight, height, etc.). Obviously, ND is not used to model income distributions, which are asymmetric, i.e. most values are clustered around the left tail, whereas the right tail is considerably longer (a positively skewed distribution).

The objective of the study is to verify the possibility of using asymmetric distributions derived from ND in the context of income distribution modelling.

2. Literature review

As mentioned in the Introduction, there are many probability distributions describing non-negative or positive variables that can be used to approximate income distributions. This family includes the following distributions: lognormal (LOG), defined by Gaddum (1945), Birnbaum-Saunders (BS), defined by Birnbaum and Saunders (1969), Dagum (DA), defined by Dagum (1977), beta prime (BPr) and inverse gamma (IG), defined by Johnson et al. (1995), Singh-Maddala (SM), defined by Singh and Maddala (1976), Zenga (Z), defind by Zenga (2010), Pareto type IV (PIV), defined by Pareto (1895), generalised gamma (GG), defined by Stacy (1962) and generalised beta of the second kind (GB2), defined by McDonald (1984). For more information on the PIV, GG and GB2 distribution families, see the Appendix.

Distributions such as Pareto, lognormal, gamma or Dagum were used to approximate income distributions in the Polish population (see e.g. Kordos, 1968, 1973; Kot, 1999, 2000; Lange, 1967; Vielrose, 1960; Wiśniewski, 1934). Research on income and wage distribution in Poland has confirmed that the Dagum, Singh-Maddala and Zenga distributions are particularly well-fitted to empirical data (see e.g. Brzeziński, 2013; Jędrzejczak, 1993, 2006; Jędrzejczak & Trzcińska, 2018; Łukasiewicz & Orłowski, 2004; Ostasiewicz, 2013; Salamaga, 2016; Trzcińska, 2020, 2022; Wałęga & Wałęga, 2021).

To achieve the aim of this article, a competitive family of distributions based on ND is needed. For this purpose, ND can be 'plasticised' by adding a 'plasticising' parameter to the cumulative distribution function (CDF) of the ND, by 'plasticising' the formula located in the exponential function or through a combination of distributions.

The first way was introduced by Azzalini (1985), who added a skewness parameter to the CDF of ND and initiated a very interesting family of ND-derived

distributions. This family includes distributions such as: skew-normal (SN), defined by Azzalini (1985), skew-generalised normal (SGN), defined by Arellano-Valle et al. (2004), flexible skew-normal (FSN), defined by Ma and Genton (2004), two-piece skew-normal (TPSN), defined by Kim (2005), power-normal (PN), defined by Gupta and Gupta (2008), generalised Balakrishnan skew-normal (GBSN), defined by Yadegari et al. (2008), Balakrishnan skew-normal (BSN), defined by Sharafi and Behboodian (2008), extended skew-generalised normal (ESGN1), defined by Choudhury and Abdul (2011), extended skew-generalised normal (ESGN2), defined by Venegas et al. (2011), skew-flexible normal (SFN), defined by Gómez et al. (2011), Kumaraswamy-normal (KN), defined by Cordeiro and de Castro (2011), flexible skew-generalised normal (FSGN1), defined by Nekoukhou et al. (2013), extended skew-generalised normal (ESGN3), defined by Kumar and Anusree (2015), flexible skew-generalised normal (FSGN2), defined by Bahrami and Qasemi (2015) and shape skew-generalised normal (SSGN), defined by Rasekhi et al. (2017).

The second approach is to 'plasticise' the formula in the exponential function of the ND. A family derived in this way includes the following distributions defined in a real domain: symmetrical sinh-normal (S-N), defined by Rieck and Nedelman (1991), SU defined by Johnson (1949), SC and expnormal (EN), defined by Sulewski (in press).

The third way involves a combination of at least two normal distributions which can fit more characteristics that the sample data might contain. Behboodian (1970) describes the conditions under which a combination of two normal distributions, called the compound normal (CN) distribution, is unimodal.

3. Income modelling – theoretical probability distributions

In the study, theoretical probability distributions used for income modelling are divided into two groups: Group I – consisting of distributions strictly dedicated to income modelling and Group II – mainly including ND-derived distributions, with ND being their special case. This group also consists of three distributions from the Johnson family, namely: SU, SC and EN. Some parameter values of these distributions are very similar to ND, but ND is not a special case of these distributions. It is possible to calculate the measure of the similarity of ND with the $\varphi(x;\mu,\sigma)$ PDF (probability density function) to the distribution with the $g(x;\boldsymbol{\Theta})$ PDF, where $\boldsymbol{\Theta}$ is the vector of parameters. This similarity measure (S_M) is provided by Sulewski (2019):

$$S_M(\mu,\sigma,\boldsymbol{\Theta}) = \int_{-\infty}^{\infty} \min[\varphi(x;\mu,\sigma),\quad g(x;\boldsymbol{\Theta})]dx. \tag{1}$$

As mentioned above, the main objective of the study is to verify the possibility of using asymmetric distributions derived from ND in the context of income distribution modelling (Group II) and to compare it with the properties of well-known distributions strictly intended for this purpose (Group I).

When analysing the structure and properties of distributions, one can group them into systems. The systems for income distributions include the Dagum, Pearson, D'Addairo, Burr and Johnson distribution. However, these systems are not always separate. For example, the lognormal distribution belongs to the D'Addairo and Johnson systems, while the gamma distribution to the Pearson and D'Addairo systems, the Dagum distribution belongs to the Dagum, Pearson and Burr systems and the Pareto distribution to the Dagum and D'Addairo systems;, whereas the Singh-Maddala distribution to the Dagum and Burr systems (Jędrzejczak & Pekasiewicz, 2020).

There is no need to select all distribution systems for a Monte Carlo simulation to assess their properties; only certain members need to be chosen. The use of distributions which happen to be special cases of other distributions is also convenient. PIV, GG and GB2 are examples of such distribution groups, with the exception of SM, which is a special case of GB2 and, as mentioned earlier, recommended for income modelling.

Group I distributions includes the LOG as a member of the D'Addairo and Johnson systems, BS as a member of the Johnson system, BPr and IG as members of the Pearson system, DA as a member of the Dagum system, SM as a member of the Dagum and Burr systems, Z as the youngest distribution in this group and the PIV, GG and GB2 families.

Group I distributions intentionally includes not only the most popular distributions for income modelling (e.g. Dagum, Singh-Maddala, Zenga), but also distributions never or very rarely used for income modelling (e.g. BS, BPr, IG).

Distributions derived from ND (Group II) include SN, SGN, ESGN1, ESGN2, ESGN3, SSGN, PN, SFN, FSN, FSGN1, FSGN2, KN, BSN, and GBSN. The Johnson system distributions, such as SU, SC and EN also belong to ND-derived distributions, because their measure of similarity to ND calculated with (1), as shown in the Appendix, exceeds 96%.

Let $\varphi(x)$ and $\Phi(x)$ be PDF and CDF of $N(0,\ 1)$, respectively. Let $\mu \in R$ be the position parameter, $\sigma > 0$ the scale parameter and α, β, γ the shape parameters. Probability distributions are described by PDF and CDF. The appropriate formulas for the probability distributions (either PDF or CDF were presented depending on the distribution and their simplicity) included in the empirical part of the article are presented in the Appendix.

Group II distributions, after eliminating distributions that are special cases of other ones, includes: ESGN1, ESGN2, ESGN3, SSGN, PN, FSGN1, FSGN2, KN, GBSN, TPSN, SU, and EN.

In summary, 22 distributions (10 from Group I and 12 from Group II) were used in the empirical study to model income.

4. Goodness-of-fit measures

In order to select the best theoretical models described in Section 3 and in the Appendix for income modelling, the authors assessed them in a two-step procedure. In the first step, the GoF measures were applied to assess the consistency of the estimated models with the empirical data from the EU-SILC. In the second step, the very popular theoretical characteristics of models selected in the first step were compared with the empirical characteristics computed for the EU-SILC data (see Section 5 for more details and Table 5 for comparison).

It is a well-known problem in studies of income distribution that in the case of large samples, GoF statistical tests lead to the rejection of the null hypothesis, even if the studied model describes the empirical distributions very well (Kunte & Gore, 1992). Therefore, to assess the properties of the distributions used for income modelling, the authors decided not to use statistical tests, but rather to apply other numerical measures: the information criteria, GoF measures, as well as empirical and theoretical characteristics, including a specifically defined measure using quantiles.

Let $M(\boldsymbol{\theta})$ be a model with a $\boldsymbol{\theta}$ vector of parameters used for describing the distribution of income. Let $f_M(x;\boldsymbol{\theta})$ and $F_M(x;\boldsymbol{\theta})$ be the respective PDF and CDF of this model. Let $x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(n)}$ be a sample of size n. Our goal is to estimate the unknown values of parameters $\boldsymbol{\theta}$ by using the maximum likelihood estimation (MLE) method. The likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_M(x_i^*;\boldsymbol{\theta}) \tag{2}$$

Then the log-likelihood function is defined as

$$l(\boldsymbol{\theta}) = \ln\big(L(\boldsymbol{\theta})\big) = \sum_{i=1}^{n} \ln[f_M(x_i^*;\boldsymbol{\theta})]. \tag{3}$$

Formulas of derivatives $dl/d\boldsymbol{\Theta}$ have complex forms. In practice, it is not necessary to calculate them. It is better to maximise the log-likelihood function using mathematical software instead of struggling with a system of complicated non-linear equations that may have extraneous roots.

To avoid any local maxima of the log-likelihood function, the optimisation routine is run repeatedly each time from different starting values that are widely scattered in the parameter space. The maximum likelihood estimates of parameters $\boldsymbol{\Theta}$ can be easily calculated, e.g. in the R software (R Core Team, 2021) using the *fitdistr* function (package MASS), or in Mathcad. Information criteria, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Hannan-Quinn Information Criterion ($HQIC$) are used for the model comparisons. Let us recall that:

$$AIC = -2l + 2p, \quad BIC = -2l + p\ln(n), \quad HQIC = -2l + 2p\ln(\ln(n)), \quad (4)$$

where l is the log-likelihood function (3), n is the sample size and p is the number of model parameters.

Let k be the number of intervals in which n individual values are grouped. Let n_i and $\hat{n}_i$ $(i = 1, 2, \dots, k)$ be the observed and the estimated counts of the i-th interval, respectively, then $w_i = n_i/n$ and $\hat{w}_i = \hat{n}_i/n$ $(i = 1, 2, \dots, k)$ represent empirical and theoretical frequencies. Estimated counts $\hat{n}_i$ are given by

$$\hat{n}_i = \begin{cases} nF_M(x_i; \boldsymbol{\Theta}) & i = 1 \\ n[F_M(x_i; \boldsymbol{\Theta}) - F_M(x_{i-1}; \boldsymbol{\Theta})] & i = 2, 3, \dots, k, \end{cases} \quad (5)$$

where x_i $(i = 1, 2, \dots, k-1)$, $x_k = \infty$ are the upper bounds of the k intervals.

The first GoF measure, proposed by Egon Vielrose, a Polish economist, demographer and statistician, is calculated using the following simple formula (Vielrose, 1960):

$$W_p = \sum_{i=1}^{k} \min(w_i, \hat{w}_i). \quad (6)$$

The higher the value of W_p, the better the consistency of the compared distributions.

The second and the third GoF measures are Mortara's A_1 index and quadratic Pearson's A_2 index defined as (Zenga et al., 2012)

$$A_1 = \frac{1}{n}\sum_{i=1}^{k}|n_i - \hat{n}_i|\,, A_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}}, \tag{7}$$

respectively.

The smaller the value of A_i $(i = 1, 2)$, the better the consistency of the compared distributions.

The fourth GoF measure takes into account a coefficient based on the relative difference between the mean value of empirical distribution $\bar{X}$ and the expected value of theoretical distribution $E(X)$ (Jędrzejczak & Pekasiewicz, 2020):

$$A_3 = \frac{|\bar{X} - E(X)|}{E(X)} \cdot 100\%. \tag{8}$$

The smaller the A_3 value, the better the consistency of the compared distributions.

The fifth GoF measure, relative error RE, calculated by comparing the theoretical and empirical characteristics, TCH and ECH is expressed as follows:

$$RE = \frac{|TCH - ECH|}{TCH} \cdot 100\%. \tag{9}$$

The smaller the RE value, the better the consistency of the compared distributions.

A number of characteristics can be selected: the mean (A), the lower quartile (Q_1), the median (Q_2), the upper quartile (Q_3), the standard deviation (SD) or the coefficient of variation (CoV). A model is considered to be well-fitted if the differences between empirical and theoretical characteristics are less than 5%.

Let x_p^* and x_p $(0 < p < 1)$ be empirical and theoretical p-th quantiles, respectively. The last GoF measure, specifically defined for the purpose of this study, is the QM given by

$$QM = \sum_{i=1}^{19}|x_{0.05i} - x_{0.05i}^*|. \tag{10}$$

The smaller the QM value, the better the consistency of the compared distributions.

The GoF measures listed above can be divided into three classes. The first class includes information criteria (IC) (4), the second one the GoF coefficients (GOFC) (6)–(8), and the third one the characteristics (CH) represented by classic measures from A to CoV, plus the new measure proposed by the authors (10).

5. Data

The empirical analysis is based on data for gross monthly income *per capita* in Poland from the 2011 edition of the EU Statistics on Income and Living Conditions (EU-SILC) survey. EU-SILC is a non-obligatory, representative questionnaire of individual households, where the data are collected in face-to-face interviews. The main objective of the survey is to supply the European Union with comparable data on the living conditions of the population.

EU-SILC is the basic source of information used for the calculation of indicators, including those related to income, poverty and social exclusion for the EU member states. Data from the survey are used to produce various income statistics, such as the average yearly equivalised disposable income *per capita*, selected measures of the diversification of the average disposable income (e.g. the Gini coefficient, S80/S20), at-risk-of-poverty thresholds, the relative at-risk-of-poverty rate or selected measures of average disposable income distribution in the Polish regions (Główny Urząd Statystyczny [GUS], 2021).

6. Results

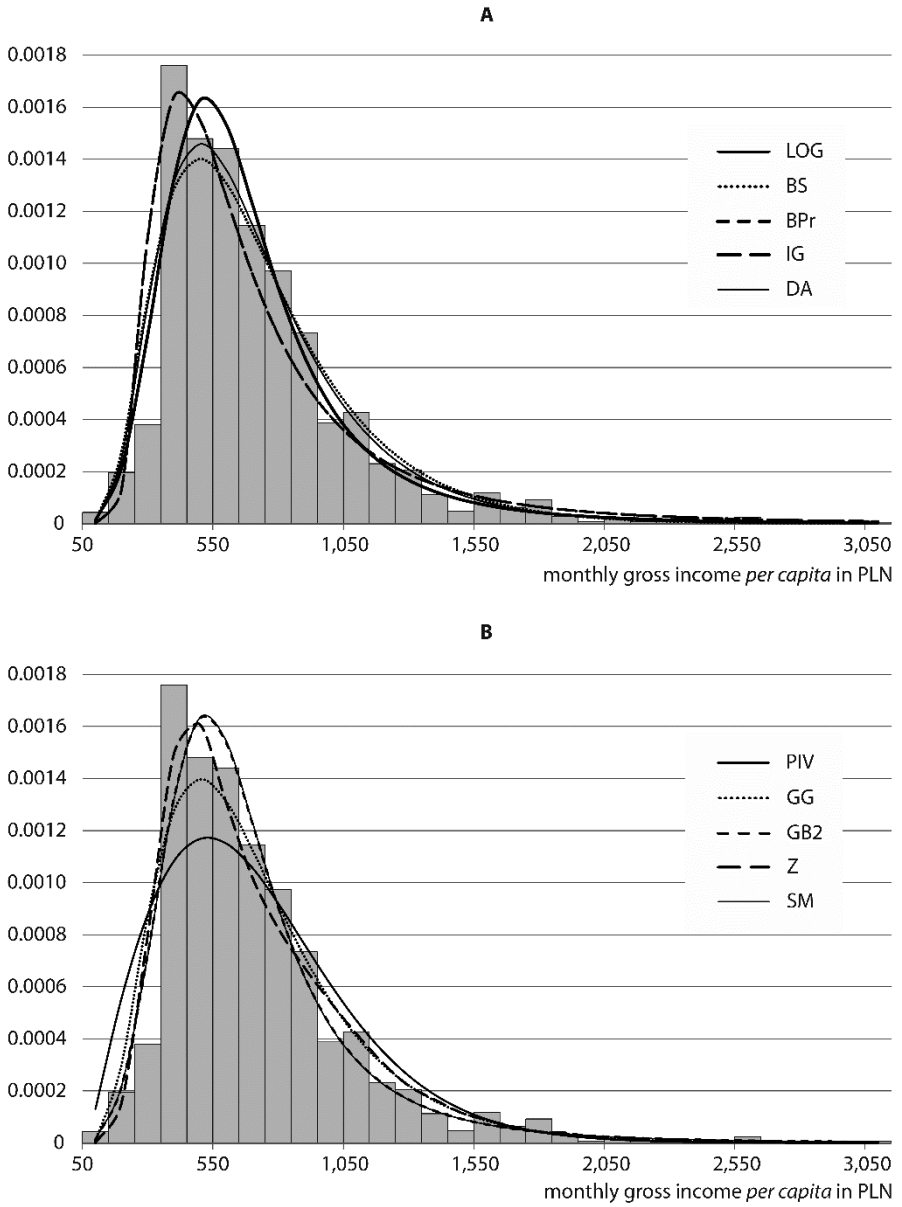
Figures 1 and 2 show histograms and estimated PDFs for the analysed models from Group I and Group II, respectively. In order to improve their legibility, only income values below PLN 3,000 are displayed.

Estimated PDFs for the BPr and IG models are almost identical (see Figure 1A). The situation is similar for the DA, SM and GB2 models (see Figure 1B). The above is also confirmed by the results presented in Tables 1 and 3.

The estimated PDFs for the ESGN1 and SSGN models are almost identical (see Figure 2B), which is also confirmed by the results provided in Tables 2 and 4.

When the estimated PDFs are very similar in shape, additional numerical measures are necessary. The first group of numerical measures consists of the values of information criteria AIC, BIC and $HQIC$ (4). Tables 1 and 2 display values of the MLEs and the information criteria for the Group I and Group II of models, respectively. These models are sorted by AIC values (the first three models in appropriate tables are in bold).

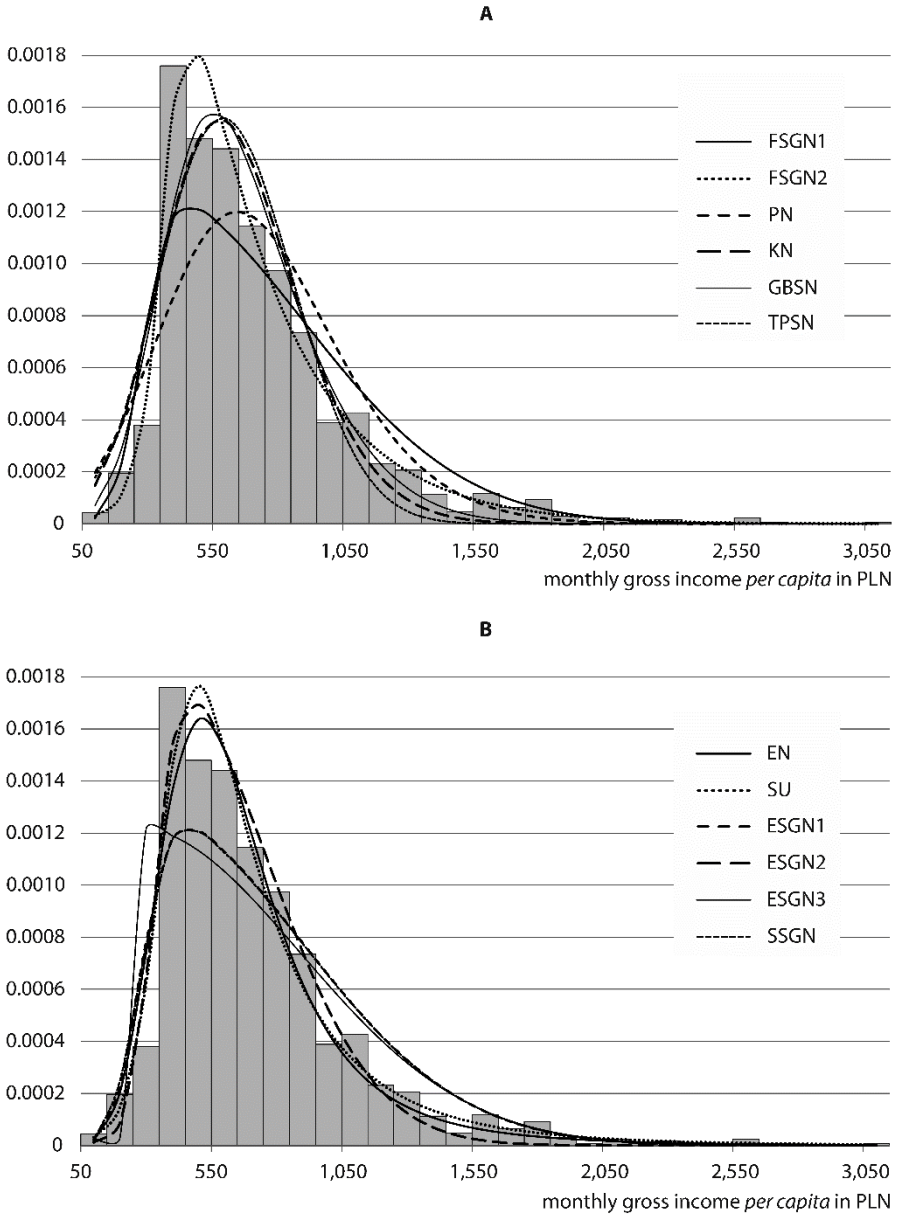
Figure 1. Histograms and probability density functions of distributions from Group I



Note. Distributions: LOG – lognormal, BS – Birnbaum-Saunders, BPr – beta prime, IG – inverse gamma, DA – Dagum, PIV – Pareto type IV, GG – generalised gamma, GB2 – generalised beta of the second kind, Z – Zenga, SM – Singh-Maddala.

Source: authors' work based on EU-SILC data.

Figure 2. Histograms and probability density functions of distributions from Group II



Note. Distributions: FSGN – flexible skew generalised normal, PN – power-normal, KN – Kumaraswamy-normal, GBSN – generalised Balakrishnan skew normal, TPSN – two-piece skew-normal, EN – expnormal, SU – SU Johnsona, ESGN – extended skew generalised normal, SSGN – shape skew generalised normal.

Source: authors' work based on EU-SILC data.

As the figures above indicate, the DA, SM and GB2 models stand out in the Group I distribution. The SU, FSGN2 and EN stand out in the Group II distribution. According to the information criteria, SU and FSGN2 (which are ND-derived distributions) produce better models of income distribution than DA, SM and GB2, which are typically used for this purpose. The AIC ranking in both distribution groups is the same as the BIC and $HQIC$ rankings.

Table 1. Values of maximum likelihood estimation and information criteria – Group I distributions

Model	Estimates of $\boldsymbol{\theta}$	AIC	BIC	$HQIC$
DA	$\hat{\sigma} = 550.3921,\ \hat{\alpha} = 3.2984,\ \hat{\beta} = 1.0904$	143,925.1	143,946.8	143,932.4
SM	$\hat{\sigma} = 549.6120,\ \hat{\alpha} = 3.5085,\ \hat{\beta} = 0.9129$	143,925.4	143,947.1	143,932.8
GB2	$\hat{\sigma} = 3.3767,\ \hat{\alpha} = 549.0722,\ \hat{\beta} = 1.0574,\ \hat{\gamma} = 0.9639$	143,926.9	143,955.8	143,936.7
LOG	$\hat{b} = 0.4787,\ \hat{c} = 6.4631,\ \hat{a} = -58.4635$	144,226.5	144,248.1	144,233.8
Z	$\hat{\beta} = 665.4616,\ \hat{\alpha} = 6.1377,\ \hat{\gamma} = 7.6694$	144,305.7	144,327.4	144,313.1
GG	$\hat{\sigma} = 1.9195,\ \hat{\alpha} = 0.4740,\ \hat{\beta} = 15.4100$	144,465.5	144,487.2	144,472.8
BS	$\hat{\alpha} = 0.4757,\ \hat{\sigma} = 674.6981,\ \hat{\mu} = -85.2903$	144,508.8	144,530.5	144,516.1
PIV	$\hat{\mu} = 24.7489,\ \hat{\sigma} = 2{,}147.9490,\ \hat{\gamma} = 0.5191,\ \hat{\alpha} = 8.9888$	145,390.3	145,419.2	145,400.1
BPr	$\hat{\alpha} = 2{,}219.1921,\ \hat{\beta} = 3.3177,\ \hat{\sigma} = 0.7331$	145,816.2	145,837.9	145,823.6
IG	$\hat{\alpha} = 3.3021,\ \hat{\sigma} = 1{,}617.5991$	145,832.9	145,847.4	145,837.8

Note. As in Figure 1.

Source: authors' work based on EU-SILC data.

Table 2. Values of maximum likelihood estimation and information criteria – Group II distributions

Model	Estimates of $\boldsymbol{\theta}$	AIC	BIC	$HQIC$
SU	$\hat{c} = -1.5371,\ \hat{\delta} = 1.2759,\ \hat{a} = 275.6729,\ \hat{b} = 188.6396$	143,805.7	143,834.6	143,815.5
FSGN2	$\hat{\mu} = 329.8192,\ \hat{\sigma} = 2{,}744.0519,\ \hat{\alpha} = 27.6787,$ $\hat{\beta} = 97.1796,\ \hat{\gamma} = 7.7981$	143,825.0	143,861.1	143,837.2
EN	$\hat{a}_1 = 1{,}249,\ \hat{b}_1 = 701.716,\ \hat{a}_2 = 10.747,\ \hat{b}_2 = -1.187,$ $\hat{c} = 2.582$	144,159.0	144,195.1	144,171.2
SSGN	$\hat{\mu} = 226.1337,\ \hat{\sigma} = 618.5770,\ \hat{\alpha} = 14.9750,$ $\hat{\beta} = 2.1659,\ \hat{\gamma} = -0.1215$	146,141.2	146,177.3	146,153.4
FSGN1	$\hat{\mu} = 227.8910,\ \hat{\sigma} = 617.1702,\ \hat{\alpha} = 7.3362,$ $\hat{\beta} = 1.8201,\ \hat{\gamma} = 6.2170$	146,142.6	146,178.7	146,154.8
ESGN1	$\hat{\mu} = 228.2867,\ \hat{\sigma} = 617.0445,\ \hat{\alpha} = 7.3236,$ $\hat{\beta} = 0.0312,\ \hat{\gamma} = 0.0028$	146,142.6	146,178.7	146,154.9
GBSN	$\hat{\mu} = 97.382,\ \hat{\sigma} = 480.465,\ \hat{\alpha} = 1.954,\ \hat{\beta} = 4.315,$ $\hat{\gamma} = 0.015$	147,061.6	147,097.7	147,073.9
ESGN3	$\hat{\mu} = 200.929,\ \hat{\sigma} = 639.164,\ \hat{\alpha} = 98.292,$ $\hat{\beta} = 2{,}468.553,\ \hat{\gamma} = -286.549$	147,164.7	147,200.8	147,176.9
PN	$\hat{\mu} = -278.3487,\ \hat{\sigma} = 594.8178,\ \hat{\alpha} = 11.2871$	147,842.2	147,863.9	147,849.5
KN	$\hat{\mu} = 109.364,\ \hat{\sigma} = 339.288,\ \hat{\alpha} = 4.928,\ \hat{\beta} = 0.703$	148,116.5	148,114.1	148,111.1
ESGN2	$\hat{\mu} = 278.806,\ \hat{\sigma} = 426.31,\ \hat{\alpha} = 5.819,\ \hat{\beta} = 1.076,$ $\hat{\gamma} = 3.825$	149,350.3	149,386.4	149,362.6
TPSN	$\hat{\mu} = 535.082,\ \hat{\sigma} = 270.331,\ \hat{\alpha} = 1.633,\ \hat{\beta} = 1.687$	153,299.6	153,328.5	153,309.4

Note. As in Figure 2.

Source: authors' work based on EU-SILC data.

Tables 3 and 4 present the values of the W_p and A_i $(i = 1, 2, 3)$ measures described in Section 4, with rankings for Group I and II distributions, respectively. The analysed models are sorted according to the values of final ranking R, which is based on the sum of sub-rankings $R()$.

As can be seen in Tables 3 and 4, models based on DA, GB2 and Z stand out in the Group I distribution. The SU, FSGN2 and EN stand out in the Group II distribution. According to the GoF measures, the best (highest) values of W_p were obtained for the FSGN2, LOG and SU models. The best (lowest) values of A_1 are achieved for the LOG, FSGN2 and SU models. The best (lowest) values of A_2 are archived for the DA, GB2 and SM models. The best (lowest) values of A_3 are for the Z, BS and SU models.

Table 3. The final ranking R of GoF measures – Group I distributions

Model	W_p	$R(W_p)$	A_1	$R(A_1)$	A_2	$R(A_2)$	A_3	$R(A_3)$	R
DA	0.983	3	0.033	3	0.066	1	0.145	3	1
GB2	0.983	3	0.034	4	0.067	2	0.239	4	2
Z	0.983	3	0.034	4	0.134	6	0.024	1	3
LOG	0.994	1	0.013	1	5.863	7	0.766	8	4
SM	0.983	3	0.035	7	0.068	3	0.308	5	5
BS	0.983	3	0.034	4	>1.000	10	0.043	2	6
GG	0.987	2	0.027	2	2.516	8	0.492	7	7
BPr	0.979	8	0.041	8	0.106	4	5.200	9	8
IG	0.979	8	0.042	9	0.107	5	5.290	10	9
PIV	0.968	10	0.063	10	391.900	9	0.440	6	10

Note. As in Figure 1.

Source: authors' work based on EU-SILC data.

Table 4. The final ranking R of GoF measures – Group II distributions

Model	W_p	$R(W_p)$	A_1	$R(A_1)$	A_2	$R(A_2)$	A_3	$R(A_3)$	R
SU	0.990	2	0.020	1	0.128	1	0.060	1	1
FSGN2	0.994	1	0.020	1	0.292	2	0.604	2	1
EN	0.973	3	0.049	3	2.029	5	4.910	5	3
GBSN	0.943	5	0.114	4	1.072	3	9.450	10	4
SSGN	0.927	6	0.146	5	1.478	4	6.940	6	5
FSGN1	0.927	6	0.147	6	>1.000	8	7.030	7	6
ESGN1	0.926	8	0.147	6	>1.000	8	7.070	8	7
PN	0.970	4	0.601	11	>1.000	8	1.270	3	8
ESGN3	0.916	10	0.153	8	20.204	6	4.480	4	9
KN	0.920	9	0.154	9	>1.000	8	14.295	11	10
TPSN	0.865	11	0.266	10	>1.000	8	19.060	12	11
ESGN2	0.120	12	0.880	12	180.361	7	8.199	9	12

Note. As in Figure 2.

Source: authors' work based on EU-SILC data.

Tables 5 and 6 show numerical characteristics, i.e. the mean (A), the lower quartile (Q_1), the median (Q_2), the upper quartile (Q_3), the standard deviation (SD) and the coefficient of variation (CoV) calculated for the top five models according to the numerical measures presented in Tables 1 and 3 (Group I distributions) and in Tables 2 and 4 (Group II distributions). These empirical characteristics are compared with the theoretical ones using percentage relative errors RE. The analysed models are sorted according to the value of final ranking R based on the sum of sub-rankings $R()$. The top three models are in bold.

The results presented in Tables 5 and 6 indicate that models based on DA, GB2 and SM stand out in the Group I distribution, while those based on SU, EN and FSGN2 stand out in the Group II distribution. The best model for the mean are Z, SU and DA, for the lower quartile are Z, LOG and SU, for the median are DA, GB2 and SM, for the upper quartile are EN, FSGN2 and SU, for the standard deviation are DA, SU and GB2, for the coefficient of variation are SU, DA and GB2.

Table 7 shows the values of the last GoF measure (10), specifically defined for this study, calculated for the models from Tables 6 and 7. The models are sorted according to the values of the proposed QM measure. The models that stand out in relation to the QM values are DA, GB2 and SM.

Table 5. Empirical (ECH) and theoretical characteristics with sub-rankings $R()$ and final ranking R – Group I distributions

Specification	ECH	DA	GB2	Z	SM	LOG
A	665.46	666.43	667.06	665.30	667.52	660.40
		(0.14)	(0.24)	(0.02)	(0.31)	(0.76)
$R(A)$	.	2	3	1	4	5
Q_1	400.53	413.63	413.83	404.05	414.12	405.69
		(3.27)	(3.32)	(0.88)	(3.39)	(1.29)
$R(Q_1)$	.	3	4	1	5	2
Q_2	570.76	570.51	570.28	572.02	570.06	582.58
		(0.04)	(0.08)	(0.22)	(0.12)	(2.07)
$R(Q_2)$	.	1	2	4	3	5
Q_3	801.06	791.34	790.72	842.23	789.64	826.88
		(1.21)	(1.29)	(5.14)	(1.43)	(3.22)
$R(Q_3)$	.	1	2	5	3	4
SD	435.26	452.94	458.17	372.25	464.27	364.81
		(4.06)	(5.26)	(14.47)	(6.66)	(16.19)
$R(SD)$	.	1	2	4	3	5
CoV	0.65	0.68	0.69	0.56	0.70	0.55
		(4.61)	(5.69)	(13.85)	(7.08)	(15.08)
$R(CoV)$	.	1	2	4	3	5
R	.	1	2	3	4	5

Note. Theoretical characteristic with *RE* values (in %) in parentheses. A – the mean, Q_1 – the lower quartile, Q_2 – the median, Q_3 – the upper quartile, *SD* – the standard deviation, *CoV* – the coefficient of variation.

Source: authors' work based on EU-SILC data.

Table 6. Empirical (ECH) and theoretical characteristics with sub-rankings $R()$ and the final ranking R – Group II distributions

Specification	ECH	SU	EN	FSGN2	GBSN	SSGN
A	665.46	664.99	643.64	660.90	608.83	715.09
		(0.08)	(3.28)	(0.69)	(8.51)	(7.45)
$R(A)$	.	1	3	2	5	4
Q_1	400.53	413.15	420.52	413.49	413.19	423.10
		(3.16)	(5.00)	(3.24)	(3.17)	(5.64)
$R(Q_1)$	.	1	4	3	2	5
Q_2	570.76	562.03	583.38	562.06	574.29	643.42
		(1.54)	(2.20)	(1.53)	(0.61)	(12.72)
$R(Q_2)$	.	3	4	2	1	5
Q_3	801.06	792.82	795.83	794.45	769.52	937.81
		(1.03)	(0.66)	(0.83)	(3.94)	(17.06)
$R(Q_3)$	.	3	1	2	4	5
SD	435.26	414.21	354.56	348.04	272.39	378.89
		(4.84)	(18.55)	(20.05)	(37.42)	(12.96)
$R(SD)$	.	1	3	4	5	2
CoV	0.65	0.62	0.55	0.53	0.45	0.53
		(4.15)	(15.23)	(18.92)	(31.23)	(18.46)
$R(CoV)$	.	1	2	4	5	3
R	.	1	2	2	4	5

Note. As in the Table 5.

Source: authors' work based on EU-SILC data.

Table 7. Values of the GoF measure QM – Group I and Group II distributions

Specification	DA	GB2	SM	SU	FSGN2	Z	LOG	EN	GBSN	SSGN
Group	I	I	I	II	II	I	I	II	II	II
QM	182.02	187.07	195.64	198.94	251.08	304.73	320.95	382.42	763.28	1,386.41

Note. As in Figure 1 and Figure 2.

Source: authors' work based on EU-SILC data.

The final collective model ranking, taking into account the results presented in Tables 1–7, is provided in Table 8. The final collective ranking of models is based on information criteria (4), GoF coefficients (6)–(8) and analysed characteristics including (9) and the measure proposed by the authors (10). It is worth noting that among the top three distributions are two ND-derived distributions, i.e. SU and FSGN2; the SU is slightly better than the Dagum distribution in terms of the analysed measures. ND-derived distributions (Group II), especially SU, can be a good alternative to well-known distributions, which are typically used to model income (Group I).

Table 8. Collective model ranking list for information criteria (IC), GoF coefficients (GOFC) and characteristics (CH) – Group I and Group II distributions

Specification	Group	IC	GOFC	CH	Sum	*R*
SU	II	1	3	2	6	1
DA	I	3	4	1	8	2
FSGN2	II	2	2	4	8	2
GB2	I	5	5	3	13	4
Z	I	8	1	6	15	5
SM	I	4	7	5	16	6
LOG	I	7	6	9	22	7
EN	II	6	9	7	22	7
GBSN	II	10	8	8	26	9
SSGN	II	9	10	10	29	10

Note. As in Figure 1 and Figure 2.

Source: authors' work based on EU-SILC data.

7. Conclusion

As far as the authors have been able to establish, this article is the first attempt to apply ND-derived distributions with real domain to the problem of model income.

The results obtained in this study confirm what has often been documented in the literature, namely that the Dagum model is particularly useful for income modelling (see e.g. Jędrzejczak, 1993, 2006; Trzcińska, 2020). However, as the authors have demonstrated with real income data and by applying different numerical measures, the family of income models with ND-derived distributions (Group II) can compete with distributions typically used for this purpose (Group I). As evidenced by the collective model ranking in Table 8, the SU Johnson distribution, representing the group of ND-derived distributions, can serve as an interesting alternative for income modelling. It is also worth emphasizing that the SU distribution can be an interesting model for variables with positive and negative values (e.g. corporate profits), which cannot be modelled with the common distributions (e.g. Dagum).

It is also worth noting that a longer series of income data from EU-SILC could help capture the dependencies between errors resulting from the use of certain families of distributions to approximate the characteristics of the empirical distribution depending on the phases of the business cycle. This could provide a more systematic assessment of a given family of distributions. Moreover the ND-derived distributions, especially SU and FSGN2, could be used to model income of more homogeneous groups, with a negligible distribution asymmetry, e.g. people with disabilities. These issues will be the subject of a study the authors plan to undertake in the nearest future.

Acknowledgements

The work of Marcin Szymkowiak has been developed as part of a project entitled 'Indirect estimation of disability in the 2011 census', which was financed by Poland's National Science Centre from a grant awarded by virtue of decision no. DEC-2013/11/B/HS4/01472.

References

Aitchison, J., & Brown, J. A. C. (1957). *The Lognormal Distribution: with special reference to its uses in economics*. Cambridge University Press.

Arellano-Valle, R. B., Gómez, H. W., & Quintana, F. A. (2004). A new class of skew-normal distributions. *Communications in Statistics – Theory and Methods*, *33*(7), 1465–1480. https://doi.org/10.1081/STA-120037254.

Azzalini, A. (1985). A Class of Distributions which Includes the Normal Ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Bahrami, W., & Qasemi, E. (2015). A Flexible Skew-Generalized Normal Distribution. *Journal of Statistical Research of Iran JSRI*, *11*(2), 131–145. https://doi.org/10.18869/acadpub.jsri.11.2.131.

Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics*, *12*(1), 131–139. https://doi.org/10.2307/1267357.

Birnbaum, Z. W., & Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability*, *6*(2), 637–652. https://doi.org/10.2307/3212003.

Brzeziński, M. (2013). Parametric Modelling of Income Distribution in Central and Eastern Europe. *Central European Journal of Economic Modelling and Econometrics*, *5*(3), 207–230. https://doi.org/10.24425/cejeme.2013.119261.

Choudhury, K., & Abdul, M. M. (2011). Extended skew generalized normal distribution. *METRON*, *69*(3), 265–278. https://doi.org/10.1007/BF03263561.

Cordeiro, G. M., & de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, *81*(7), 883–898. https://doi.org/10.1080/00949650903530745.

Dagum, C. (1977). A New Model of Personal Income Distribution: Specification and Estimation. *Économie Appliquée*, *30*(3), 413–436.

Gaddum, J. H. (1945). Lognormal distributions. *Nature*, *156*, 463–466. https://doi.org/10.1038/156463a0.

Główny Urząd Statystyczny. (2021). *Dochody i warunki życia ludności Polski – raport z badania EU-SILC 2019*. https://stat.gov.pl/en/topics/living-conditions/living-conditions/incomes-and-living-conditions-of-the-population-in-poland-report-from-the-eu-silc-survey-of-2019,1,12.html.

Gómez, H. W., Elal-Olivero, D., Salinas, H. S., & Bolfarine, H. (2011). Bimodal Extension Based on the Skew-Normal Distribution with Application to Pollen Data. *Environmetrics*, *22*(1), 50–62. https://doi.org/10.1002/env.1026.

Gupta, R. C., & Gupta, R. D. (2008). Analyzing skewed data by power normal model. *TEST*, *17*(1), 197–210. https://doi.org/10.1007/s11749-006-0030-x.

Jędrzejczak, A. (1993). Application of the Dagum distribution in the analysis of income distributions in Poland. *Acta Universitatis Lodziensis. Folia Oeconomica*, (131), 103–112.

Jędrzejczak, A. (2006). The characteristic of theoretical income distributions and their application to the analysis of wage distributions in Poland by regions. *Acta Universitatis Lodziensis. Folia Oeconomica*, (196), 183–198.

Jędrzejczak, A., & Pekasiewicz, D. (2020). *Teoretyczne rozkłady dochodów gospodarstw domowych i ich estymacja*. Wydawnictwo Uniwersytetu Łódzkiego.

Jędrzejczak, A., & Trzcińska, K. (2018). Application of the Zenga distribution to the analysis of household income in Poland by socio-economic group. *Statistica & Applicazioni*, *16*(2), 123–140. https://doi.org/10.26350/999999_000015.

Johnson, N. L. (1949). System of frequency curves generated by methods of translation. *Biometrika*, *36*(1/2), 149–176. https://doi.org/10.2307/2332539.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (vol. 2, 2nd ed.). John Wiley & Sons.

Kim, H.-J. (2005). On a class of two-piece skew-normal distributions. *Statistics. A Journal of Theoretical and Applied Statistics*, *39*(6), 537–553. https://doi.org/10.1080/02331880500366027.

Kordos, J. (1968). *Metody matematyczne badania i analizy rozkładów dochodów ludności*. Główny Urząd Statystyczny.

Kordos J. (1973). *Metody analizy i prognozowania rozkładów płac i dochodów ludności*. Państwowe Wydawnictwo Ekonomiczne.

Kot, S. M. (Ed.). (1999). *Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji*. Wydawnictwo Naukowe PWN.

Kot, S. M. (2000). *Ekonometryczne modele dobrobytu*. Wydawnictwo Naukowe PWN.

Kumar, C. S., & Anusree, M. R. (2015). On an extended version of skew generalized normal distribution and some of its properties. *Communications in Statistics – Theory and Methods*, *44*(3), 573–586. https://doi.org/10.1080/03610926.2012.739251.

Kunte, S., & Gore, A. P. (1992). The paradox of large samples. *Current Science*, *62*(5), 393–395.

Lange, O. (1967). *Wstęp do ekonometrii* (4th ed.). Państwowe Wydawnictwo Naukowe.

Łukasiewicz, P., & Orłowski, A. (2004). Probabilistic Models of Income Distributions. *Physica A: Statistical Mechanics and its Applications*, *344*(1–2), 146–151. https://doi.org/10.1016/j.physa.2004.06.106.

Ma, Y., & Genton, M. G. (2004). Flexible Class of Skew-Symmetric Distributions. *Scandinavian Journal of Statistics*, *31*(3), 459–468.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, *52*(3), 647–663. https://doi.org/10.2307/1913469.

Mead, M., Nassar, M. M., & Dey, S. (2018). A Generalization of Generalized Gamma Distributions. *Pakistan Journal of Statistics and Operation Research*, *14*(1), 121–138. https://doi.org/10.18187/pjsor.v14i1.1692.

Metcalf, C. E. (1972). *An Econometric Model of Income Distribution*. Markham Publishing Company.

Nekoukhou, V., Alamatsaz, M. H., & Aghajani, A. H. (2013). A flexible skew-generalized normal distribution. *Communications in Statistics – Theory and Methods*, *42*(13), 2324–2334. https://doi.org/10.1080/03610926.2011.599003.

Ostasiewicz, K. (2013). Adekwatność wybranych rozkładów teoretycznych dochodów w zależności od metody aproksymacji. *Przegląd Statystyczny. Statistical Review*, *60*(4), 499–522.

Pareto, V. (1895). La legge della domanda. *Giornale Degli Economisti*, *10*(6), 59–68.

Pratesi, M. (Ed.). (2016). *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons. https://doi.org/10.1002/9781118814963.

R Core Team. (2021). *The R Project for Statistical Computing*. https://www.R-project.org/.

Rasekhi, M., Hamedani, G. G., & Chinipardaz, R. (2017). A flexible extension of skew generalized normal distribution. *METRON*, *75*(1), 87–107. https://doi.org/10.1007/s40300-017-0106-2.

Rieck, J. R., & Nedelman, J. R. (1991). A Log-Linear Model for the Birnbaum-Saunders Distribution. *Technometrics*, *33*(1), 51–60. https://doi.org/10.2307/1269007.

Salamaga, M. (2016). Badanie wpływu metody estymacji teoretycznych modeli rozkładu dochodów na jakość aproksymacji rozkładu dochodów mieszkańców Krakowa. *Zeszyty Naukowe UEK*, *3*(951), 63–79. https://doi.org/10.15678/ZNUEK.2016.0951.0305.

Sharafi, M., & Behboodian, J. (2008). The Balakrishnan skew-normal density. *Statistical Papers*, *49*(4), 769–778. https://doi.org/10.1007/s00362-006-0038-z.

Singh, S. K., & Maddala, G. S. (1976). A Function for Size Distribution of Income. *Econometrica*, *44*(5), 963–970. https://doi.org/10.2307/1911538.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, *33*(3), 1187–1192.

Stacy, E. W., & Mihram, G. A. (1965). Parameter Estimation for a Generalized Gamma Distribution. *Technometrics*, *7*(3), 349–358. https://doi.org/10.2307/1266594.

Sulewski, P. (2019). Modified Lilliefors Goodness-of-fit Test for Normality. *Communications in Statistics – Simulation and Computation*, *51*(3), 1199–1219. https://doi.org/10.1080/03610918.2019.1664580.

Sulewski, P. (in press). New Members of The Johnson Family of Probability Distributions: Properties and Application. *REVSTAT – Statistical Journal*.

Trzcińska, K. (2020). Analysis of Household Income in Poland Based on the Zenga Distribution and Selected Income Inequality Measure. *Folia Oeconomica Stetinensia*, *20*(1), 421–436. https://doi.org/10.2478/foli-2020-0025.

Trzcińska, K. (2022). An Analysis of Household Income in Poland and Slovakia Based on Selected Income Models. *Folia Oeconomica Stetinensia*, *22*(1), 287–301. https://doi.org/10.2478/foli-2022-0014.

Venegas, O., Sanhueza, A. I., & Gómez, H. W. (2011). An extension of the skew-generalized normal distribution and its derivation. *Proyecciones. Journal of Mathematics*, *30*(3), 401–413. https://doi.org/10.4067/S0716-09172011000300007.

Vielrose, E. (1960). *Rozkład dochodów według wielkości*. Polskie Wydawnictwo Gospodarcze.

Wałęga, A., & Wałęga, G. (2021). Self-employment and over-indebtedness in Poland: Modelling income and debt repayments distribution. *EBER Entrepreneurial Business and Economics Review*, *9*(4), 51–65. https://doi.org/10.15678/EBER.2021.090404.

Wiśniewski, J. (1934). *Rozkład dochodów według wysokości*. Instytut Badania Koniunktur Gospodarczych i Cen.

Yadegari, I., Gerami, A., & Khaledi, M. J. (2008). A generalization of the Balakrishnan skew-normal distribution. *Statistics & Probability Letters*, *78*(10), 1165–1167. https://doi.org/10.1016/j.spl.2007.12.001.

Zenga, M. M. (2010). Mixture of Polisicchio's truncated Pareto distributions with beta weights. *Statistica & Applicazioni*, *8*(1), 3–25. https://www.vitaepensiero.it/scheda-articolo_digital/michele-zenga/mixture-of-polisicchios-truncated-pareto-distributions-with-beta-weights-999999_2010_0001_0002-151302.html.

Zenga, M. M., Pasquazzi, L., & Zenga, M. (2012). First applications of a new three-parameter distribution for non-negative variables. *Statistica & Applicazioni*, *10*(2), 131–149. https://statisticaeapplicazioni.vitaepensiero.it/scheda-articolo_digital/leo-pasquazzi-mariangela-zenga-michele-zenga/first-applications-of-a-new-three-parameter-distribution-for-non-negative-variables-999999_2012_0002_0037-151352.html.

Appendix

Let $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ be the beta function defined by the gamma function and

$$v(x;\beta,k) = \begin{cases} 0.5\sqrt{\beta k}(1-k)^{-1}x^{-1.5} & x \in [\beta k, \beta/k] \\ 0 & \text{otherwise.} \end{cases}$$

The PDF or CDF for Group I distributions are as follows:

$$F_{LOG}(x;c,b,a) = \Phi\left[\frac{\ln(x-a)-b}{c}\right] \ (x > a;\ a,b \in R, c > 0);$$

$$F_{BS}(x;\alpha,\sigma,\mu) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{x-\mu}{\sigma}} - \sqrt{\frac{\sigma}{x-\mu}}\right)\right] \ (x > \mu;\ \alpha > 0);$$

$$f_{BPr}(x;\alpha,\beta,\sigma) = \frac{\left(\frac{x}{\sigma}\right)^{\alpha-1}\left(1+\frac{x}{\sigma}\right)^{-\alpha-\beta}}{\sigma B(\alpha,\beta)} \ (x \geq 0;\ \alpha,\beta > 0);$$

$$f_{IG}(x;\alpha,\beta) = \frac{\beta^{\alpha}x^{-\alpha-1}}{\Gamma(\alpha)}\exp\left(-\frac{\beta}{x}\right) \ (x > 0;\ \alpha > 0, \beta > 0);$$

$$F_{D}(x;\sigma,\alpha,\beta) = \left[1+\left(\frac{x}{\sigma}\right)^{-\alpha}\right]^{-\beta} \ (x > 0;\ \alpha,\beta > 0);$$

$$f_{Z}(x;\beta,\alpha,\gamma) = \int_0^1 v(x;\beta,k)\frac{k^{\alpha-1}(1-k)^{\gamma-1}}{B(\alpha,\gamma)}dk \ \left(x \geq 0;\ \beta,\alpha,\gamma > 0,\ k \in (0,1)\right);$$

$$f_{SM}(x;\sigma,\alpha,\gamma) = \frac{\alpha\gamma}{\sigma}\left(\frac{x}{\sigma}\right)^{\alpha-1}\left[1+\left(\frac{x}{\sigma}\right)^{\alpha}\right]^{-\gamma-1} \ (x \geq 0;\ \alpha,\gamma > 0);$$

$$F_{PIV}(x;\mu,\sigma,\gamma,\alpha) = 1 - \left[1+\left(\frac{x-\mu}{\sigma}\right)^{\frac{1}{\gamma}}\right]^{-\alpha} \ (x \geq \mu;\ \gamma,\alpha > 0);$$

$$f_{GG}(x;\sigma,\alpha,\beta)=\frac{\alpha}{\sigma\Gamma(\beta)}\left(\frac{x}{\sigma}\right)^{\alpha\beta-1}\exp\left[-\left(\frac{x}{\sigma}\right)^{\alpha}\right]\ \ (x\geq 0;\ \beta>0);$$

$$f_{GB2}(x;\alpha,\sigma,\beta,\gamma)=\frac{\alpha}{\sigma B(\beta,\gamma)}\left(\frac{x}{\sigma}\right)^{\alpha\beta-1}\left[1+\left(\frac{x}{\sigma}\right)^{\alpha}\right]^{-(\beta+\gamma)}\ (x\geq 0;\ \alpha,\beta,\gamma>0).$$

Special cases of PIV, GG and GB2 are presented in Tables A1–A3, respectively.

Table A1. Special cases of PIV

Name	μ	σ	γ	α
Pareto type: I	σ	.	1	.
II	.	.	1	.
III	.	.	.	1
Lomax	0	.	1	.

Source: authors' work based on Jędrzejczak and Pekasiewicz (2020).

Table A2. Special cases of GG

Name	σ	α	β
Exponential	.	1	1
Gamma	.	1	.
Weibull	.	.	1
Chi-square	2	2	$0.5n,\ n\in N$
Chi	$\sqrt{2}$	2	$0.5n,\ n\in N$
Rayleigh	$\sigma\sqrt{2}$	2	1
Maxwell-Boltzmann	$\sigma\sqrt{2}$	2	1.5

Source: authors' work based on Stacy and Mihram (1965).

Table A3. Special cases of GB2

Name	σ	α	β	γ
Singh–Maddala (Burr XII)	.	.	1	.
Dagum (Burr III)	.	.	.	1
Beta type II	.	1	.	.
Standard Burr XII	1	.	1	.
Standard Burr III	1	.	.	1
Standard Beta type II	1	1	.	.
Fisk (log-logistic)	.	.	1	1
Lomax (Pareto type II)	.	1	1	.
Inverse Lomax	.	1	.	1
Paralogistic	.	h	1	.
Inverse paralogistic	.	a	.	.
Fisher	v/u	$u/2$	$v/2$	1

Source: authors' work based on Mead et al. (2018).

Let a_1, a_2, b_1, b_2 be multipurpose parameters, c the semi-fraction parameter then

$$T(h,a)=\frac{1}{2\pi}\int_0^a \frac{e^{-0.5h^2(1+x^2)}}{1+x^2}dx \quad (h,a\in R).$$

The PDF or CDF for the distributions derived from ND are as follows:

$$f_{SN}(x;\mu,\sigma,\alpha)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\ (\alpha\in R,\ SN(\mu,\sigma,0)=N(\mu,\sigma);$$

$$f_{SGN}(x;\mu,\sigma,\alpha,\beta)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{\alpha(x-\mu)}{\sqrt{\sigma^2+\beta(x-\mu)^2}}\right)\ (\beta\geq 0,\ \alpha\in R),$$
$$SGN(\mu,\sigma,\alpha,0)=SN(\mu,\sigma,\alpha),\quad SGN(\mu,\sigma,0,0)=N(\mu,\sigma);$$

$$f_{ESGN1}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left[\frac{\alpha\frac{x-\mu}{\sigma}}{\sqrt{1+\beta\left(\frac{x-\mu}{\sigma}\right)^2+\gamma\left(\frac{x-\mu}{\sigma}\right)^4}}\right]\ (\beta,\gamma\geq 0,\ \alpha\in R),$$

$$ESGN1(\mu,\sigma,\alpha,\beta,0)=SGN(\mu,\sigma,\alpha,\beta),\quad ESGN1(\mu,\sigma,0,\beta,\gamma)=N(\mu,\sigma),$$
$$ESGN1(\mu,\sigma,\alpha,0,0)=SN(\mu,\sigma,\alpha);$$

$$f_{ESGN2}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{4}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\int_{-\infty}^{\frac{\alpha(x-\mu)}{\sqrt{\sigma^2+\beta(x-\mu)^2}}}\varphi(t)\Phi\left(\frac{-\sqrt{\beta}\gamma t(x-\mu)}{\sqrt{\sigma^2+\beta(x-\mu)^2+\sigma^2\gamma^2}}\right)dt,$$
$$\alpha,\gamma\in R,\quad \beta\geq 0,\quad ESGN2(\mu,\sigma,0,\beta,0)=N(\mu,\sigma),$$
$$ESGN2(\mu,\sigma,\alpha,0,\gamma)=SN(\mu,\sigma,\alpha),\quad ESGN2(\mu,\sigma,\alpha,\beta,0)=SGN(\mu,\sigma,\alpha,\beta);$$

$$f_{ESGN3}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{2}{\sigma(\gamma+2)}\varphi\left(\frac{x-\mu}{\sigma}\right)\left\{1+\gamma\Phi\left[\frac{\alpha(x-\mu)}{\sqrt{\sigma^2+\beta(x-\mu)^2}}\right]\right\},$$
$$\beta\geq 0,\gamma\geq -1,\quad \alpha\in R,\quad ESGN3(\mu,\sigma,\alpha,\beta,0)=ESGN3(\mu,\sigma,0,\beta,\gamma)=N(\mu,\sigma);$$

$$f_{SSGN}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left[\frac{\alpha\frac{x-\mu}{\sigma}}{\sqrt{1+\beta\left|\frac{x-\mu}{\sigma}\right|^{2\gamma}}}\right]\ (\beta\geq 0,\ \gamma\neq 0,\ \alpha\in R),$$
$$\beta=0\Longrightarrow\gamma=1;\quad \alpha=0\Longrightarrow\beta=0,\ \gamma=1;\quad SSGN(\mu,\sigma,0,0,1)=N(\mu,\sigma),$$
$$SSGN(\mu,\sigma,\alpha,0,1)=SN(\mu,\sigma,\alpha),\quad SSGN(\mu,\sigma,\alpha,\beta,1)=SGN(\mu,\sigma,\alpha,\beta),$$
$$SSGN(\mu,\sigma,\alpha,\beta,2)=ESGN1(\mu,\sigma,\alpha,0,\gamma);$$

$$f_{SFN}(x;\mu,\sigma,\alpha,\beta)=\frac{1}{\sigma[1-\Phi(\beta)]}\varphi\left(\frac{|x-\mu|}{\sigma}+\beta\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\ (\alpha,\beta\in R),$$
$$SFN(\mu,\sigma,0,0)=N(\mu,\sigma),\quad SFN(\mu,\sigma,\alpha,0)=SN(\mu,\sigma,\alpha);$$

$$f_{FSN}(x;\mu,\sigma,\alpha,\beta)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left[\alpha\frac{x-\mu}{\sigma}+\beta\left(\frac{x-\mu}{\sigma}\right)^3\right]\ (\alpha,\beta\in R),$$
$$FSN(\mu,\sigma,0,0)=N(\mu,\sigma),\quad FSN(\mu,\sigma,\alpha,0)=SN(\mu,\sigma,\alpha);$$

$$f_{FSGN1}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left[\frac{\alpha(x-\mu)+\frac{\gamma}{\sigma^2}(x-\mu)^3}{\sqrt{\sigma^2+\beta(x-\mu)^2}}\right]\ (\alpha,\gamma\in R,\ \beta\geq 0),$$
$$FSGN1(\mu,\sigma,0,0,0)=FSGN1(\mu,\sigma,0,\beta,0)=N(\mu,\sigma),$$

$$FSGN1(\mu,\sigma,\alpha,0,0)=SN(\mu,\sigma,\alpha),\quad FSGN1(\mu,\sigma,\alpha,\beta,0)=SGN(\mu,\sigma,\alpha,\beta),$$
$$FSGN1(\mu,\sigma,\alpha,0,\gamma)=FSN(\mu,\sigma,\alpha,\beta);$$

$$f_{FSGN2}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{1}{\sigma[1-\Phi(\gamma)]}\varphi\left(\frac{|x-\mu|}{\sigma}+\gamma\right)\Phi\left[\frac{\alpha(x-\mu)}{\sqrt{\sigma^2+\beta(x-\mu)^2}}\right](\alpha,\gamma\in R,\ \beta\geq 0),$$
$$FSGN2(\mu,\sigma,\alpha,0,0)=SN(\mu,\sigma,\alpha),\quad FSGN2(\mu,\sigma,\alpha,\beta,0)=SGN(\mu,\sigma,\alpha,\beta),$$
$$FSGN2(\mu,\sigma,\alpha,0,\gamma)=SFN(\mu,\sigma,\alpha,\beta);$$

$$f_{KN}(x;\mu,\sigma,\alpha,\beta)=\frac{\alpha\beta}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{\alpha-1}\left\{1-\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]^{\alpha}\right\}^{\beta-1}\quad(\alpha,\beta>0),$$
$$|KN(\mu,\sigma,1,1)=N(\mu,\sigma),\ \ KN(\mu,\sigma,2,1)=SN(\mu,\sigma,1);$$

$$f_{BSN}(x;\mu,\sigma,\alpha,\beta)=\frac{\varphi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\beta}}{\int_{-\infty}^{\infty}\varphi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\beta}dx}\quad(\alpha\in R,\ \ \beta=1,2,\dots),$$
$$BSN(\mu,\sigma,\alpha,0)=N(\mu,\sigma),\ BSN(\mu,\sigma,\alpha,1)=SN(\mu,\sigma,\alpha);$$

$$f_{GBSN}(x;\mu,\sigma,\alpha,\beta,\gamma)=\frac{\varphi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\beta}\left[1-\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\gamma}}{\int_{-\infty}^{\infty}\varphi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\beta}\left[1-\Phi\left(\alpha\frac{x-\mu}{\sigma}\right)\right]^{\gamma}dx}\quad(\alpha\in R,\ \ \beta,\gamma=1,2,\dots),$$
$$GBSN(\mu,\sigma,\alpha,\beta,0)=BSN(\mu,\sigma,\sigma,\alpha,\beta),\quad GBSN(\mu,\sigma,\alpha,0,0)=N(\mu,\sigma),$$
$$GBSN(\mu,\sigma,\alpha,1,0)=SN(\mu,\sigma,\alpha);$$

$$f_{TPSN}(x;\mu,\sigma,\alpha,\beta)=\frac{\phi\left(\frac{x-\mu}{\sigma},0,1\right)\Phi\left(\alpha\left|\frac{x-\mu}{\sigma}+\beta\right|,0,1\right)}{\sigma\left[I(\alpha>0)-2T\left(\frac{\alpha\beta}{\sqrt{1+\alpha^2}},\frac{1}{\alpha}\right)\right]}\quad(\alpha,\beta\in R),$$
$$TPSN(\theta,\sigma,0,0)=N(\theta,\sigma);$$

$$F_{SU}(x;\alpha,\beta,\mu,\sigma)=\Phi\left[\alpha+\beta\operatorname{asinh}\left(\frac{x-\mu}{\sigma}\right);0,1\right],$$
$SU(0,3.223,0,2.939)$, according to (1), is similar to the $N(0,0.916)$ in 98.66%;

$$F_{SC}(x;\alpha,\mu,\sigma)=\Phi\left[\alpha+2\operatorname{sinh}\left(\frac{x-\mu}{\sigma}\right);0,1\right](\alpha\in R),\ SC(0,\mu,\sigma),$$
according to (1), is similar to the $N(\mu,0.5\sigma)$ in 96.66%;

$$F_{EN}(x;a_1,b_1,a_2,b_2,c)=$$
$$=\Phi\left[c-\exp\left(\frac{a_1-x}{b_1}\right)+\exp\left(\frac{x-a_2}{b_2}\right)\right]\quad(a_1,a_2\in R;\quad b_1,b_2,\mathrm{c}>0),$$

$EN(a_1,b_1,a_1,b_1,0)$, according to (1), is similar to the $N(a_1,0.5b_1)$ in 96.66%,
$$EN(a_1,b_1,a_1,b_1,c)=SC(a_1,b_1,c).$$