The Impact of Interval Choice in Grouped Frequency Tables on Statistical Modelling

Adam Čabla[1] | *Prague University of Economics and Business, Prague, Czechia*

Received 30.4.2024 (revisions received 16.8.2024, 4.12.2024), **Accepted** (reviewed) 26.2.2025, **Published** 13.6.2025

Abstract

This paper examines the adequacy of grouped (interval) frequency tables for statistical modelling. Inspired by the chosen real-world data structure, the research question is: Can accurate modelling be achieved with the given grouping schemes without a significant loss of accuracy compared to the original data? To answer this, simulations based on log-normal distributions and various levels of grouping detail were conducted. The results show that large sample sizes enable accurate estimates even with low detailed censoring, provided the model aligns with the data-generating process. However, the mismatch between fitted and real distribution can introduce an additional bias, which can be reduced with detailed right-tail intervals. Therefore, it is recommended to consider this when choosing intervals for grouped frequency tables.[2]

Keywords	***DOI***	***JEL code***
Grouped frequency table, censoring, log-normal distribution, parametric estimate	*https://doi.org/10.54694/stat.2024.24*	*C24, C13, C15*

INTRODUCTION

One option for presenting data on a given topic is to offer it in the form of frequency tables, which show the frequencies of specific values. Frequency tables are suitable for categorical and discrete variables because they compress the information effectively, reducing it to a few rows rather than thousands. However, for continuous variables, this approach only slightly reduces the number of rows due to the abundance of unique values. For this type of variable, grouped frequency tables are often chosen instead. These tables convey information about non-overlapping intervals of the values and their frequencies, whether absolute or relative, within the original data. This approach is, however, connected to the loss of information, as it uses information on the interval the value belongs to instead of the exact value. This loss of information is sometimes intentional due to privacy reasons.

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business. W. Churchill Sq. 4, 130 67 Prague 3, Czechia. E-mail: adam.cabla@vse.cz. ORCID: <https://orcid.org/0000-0001-8549-801X>.

[2] Article based on the *AMSE 2022 Conference* contribution.

An example of this type of data chosen as the basis for this simulation study are tables provided by the Ministry of Labor and Social Affairs (MoLSA, 2024) or the Czech Statistical Office on earnings in wage and salary spheres (CZSO, 2023e). The original data were obtained through the Average Earnings Information System (AEIS), which provides structural statistics on individual employees' earnings and other personal information such as age, sex, and educational attainment. In this survey, gross earnings cover all wages and salaries, including bonus pay or reimbursements of pay (CZSO, 2023d).

Thus, freely available information from the original data is provided as a grouped frequency table, leading to the research question of this paper: How and to what extent does the grouping influence the precision of statistical modelling outputs compared to original data?

To answer this question, a set of simulations has been run to mimic the process from data collection to publication in the form of the grouped frequency table and distribution fits. The focus is on the parametric fit of log-normal distribution and the influence of the intervals on this parametric fit. Do these bands influence the fits compared to exact (originally collected) data, how, and to what extent? If so, can the problem be alleviated using different bands?

The results stress the problem in the statistical modelling based on this type of data and recommend how the data can be provided in such a way as to alleviate the possible issues.

The paper is organised as follows: The methodology part provides details on data supplied by the CZSO used for the simulation set-up, introduces characteristics of interest, discusses censoring and the maximum likelihood estimate of censored data, and finally describes the simulation process and settings used. Then, results are presented and discussed with a focus on the three main findings: observed consistency of maximum likelihood estimator, observed bias of the fit under miss-specified distribution, and observed bias-variance trade-off, all under censoring schemes with different levels of detail for comparison.

Throughout the process, the *R* (R Core Team, 2023) was used with packages *dplyr* (Wickham et al., 2023) for data manipulation, *fitdistrplus* (Delignette-Muller, 2015) to fit the parameters and *ggplot2* (Wickham, 2016) with *ggpubr* (Kassambara, 2023) packages to create figures.

1 LITERATURE SURVEY

In this paper, the research question outlined in the introduction is answered by the example of log-normal distribution. This distribution frequently results from multiplicative processes and is prevalent across numerous scientific fields, including medicine, ecology, microbiology, linguistics, and social sciences (Limpert et al., 2001). This distribution is also commonly applied as a model for incomes, for example, in Bartošová (2006), which serves as the motivation for this paper. The application can be expanded by using the three-parameter version, where the third parameter defines the distribution's support, often determined theoretically or through empirical data. For example, a support value of 250 was selected to provide a better fit in earnings modelling and predictions based on the data from AEIS in Marek et al. (2018). However, the two-parametric version is used in the presented article due to its regularity and the possibility of automatic global maximum likelihood estimation in the simulation setting.

The research question focuses on comparing fits between the grouped frequency data and the original data. The estimators and their properties using grouped frequency data have been studied. O'Neil and Wells (1972) utilise both the two-parameter and three-parameter versions of the log-normal distribution to fit the data from car insurance claims. One of their results recommends using logarithmic grouping to improve the fit of log-normal distribution, especially for smaller numbers of intervals, to achieve efficiency close to the original data. They use, however, the method of scoring, which is not a maximum likelihood estimator due to the computing limitations of the time.

From an estimation perspective, the maximum likelihood estimator for the log-normal distribution, based on the grouped frequency data, is both existent and unique, as Xia et al. (2009) find. Xiao et al. (2016) provide a general overview of estimation methods for grouped data across three distributions,

including a comparison involving the log-normal distribution. In their concluding remarks, the authors recommend using maximum likelihood or minimum distance estimators for small sample sizes. Shirazi et al. (2022), in their simulation study, compare the maximum likelihood estimator with the EM and MCEM methods. They conclude that even for small sample sizes, the three estimators yield similar results regarding their accuracy for univariate and bivariate estimates of grouped normal data.

Consistency is a standard property of the maximum likelihood estimator for many underlying probability distributions. In the case of interval-censored data, which is a broader category containing grouped data as a special case, this has been shown theoretically by Korobeynikov (2012) and observed using simulations, e.g. for Weibull distribution (El-Sagheer, 2018; Vittal and Phillips, 2007). Thus, one can expect that the errors of estimate will decrease with increasing sample sizes for the correctly identified distribution.

A specific method for estimating parameters of log-normal distribution is proposed in research not published in the scientific literature (Potharst, 2022). This method is applicable only if the grouped frequency data of the right tail are available, resulting in truncation. Still, this problem is of marginal importance to the research question of this paper.

So far, the literature review has covered mostly the properties of the maximum likelihood estimator of the log-normal distribution with a focus on the grouped data, assuming the distribution is correctly specified. Literature on the misspecification of log-normal distribution has also existed, as shown by Tarima et al. (2013). The paper concludes that the estimator of population mean combining log-normal parametric and non-parametric estimators can be recommended for small-to-medium-sized samples if the log-normal assumption is questionable.

However, the literature survey has not found the question on comparing fits on the original versus grouped data under the misspecified model. It thus opens the opportunity to explore the topic.

2 METHODS

2.1 Real-world motivation for simulation set-up

Based on AEIS individual observations, the MoLSA and the CZSO provide information about the percentage of people across earning bands and the CZ-NACE sections, and some joined divisions in the case of (MoLSA, 2024). These earning bands are non-overlapping intervals that form a grouped frequency table. The CZSO additionally provides data on these earning bands by sex (CZSO, 2023a). Besides the detailed earning bands, the CZSO also provides less detailed earning bands for the earnings divided by other variables – Unit size, Age intervals, Education attainment levels, ISCO major groups, intervals of lengths of service and Cohesion area. The information on the bands used by the CZSO is summarised in Table 1.

Table 1 Share of positive answers to job search questions and item-response probabilities

Less detailed bands	Detailed bands
Up to 16	Up to 18
16–60 in steps of 2	18–23
	23–27
	27–36 in steps of 3
	36–48 in steps of 6
60–100 in steps of 5	48–60
Over 100	Over 60

Source: CZSO, Trexima, own work

Individual observations can be reconstructed using the provided grouped frequency tables along the sample size. For example, table A15 from the year 2022 (CZSO, 2023b) shows percentages of the joint distribution of educational attainment and gross monthly earning bands; the percentage of the people with *higher education* having gross monthly earnings over 60 000 is 7.18 from circa 3 633 200 observations (retrievable, for example, from table A23 (CZSO, 2023c)). Hence, approximately 263 017 individuals were in this joint category in the original dataset.

This way, a sample closely resembling the original individual earnings data can be constructed for statistical modelling, given the grouped frequency table and the total number of observations.

2.2 Characteristics of interest

In the simulations, the log-normal distribution with probability density function:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left(-\frac{\left(\ln(x) - \mu\right)^2}{2\sigma^2} \right) \text{for} \begin{cases} x \in (0;\infty) \\ \mu \in (-\infty;\infty) \\ \sigma \in (0;\infty) \end{cases}, \tag{1}$$

is fitted. The distribution has explainable parameters: μ is the location parameter, with $\exp(\mu)$ being the median of the distribution, and σ is a scale parameter of the log-transformed distribution, and its functions determine, e.g. moment skewness.

$$\text{Skewness} = \left(\exp\left(\sigma^2\right) + 2\right)^2 \sqrt{\exp\left(\sigma^2\right) - 1}\,, \tag{2}$$

moment kurtosis, coefficient of variation:

$$\text{CV}(X) = \sqrt{\exp\left(\sigma^2\right) - 1}, \tag{3}$$

or Gini coefficient:

$$\text{G} = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1, \tag{4}$$

where: _() is the cumulative distribution function of standard normal distribution. Variance and other quantiles are functions of both parameters (Johnson et al., 1994).

The variables of interest are the relative absolute differences between the computed characteristics obtained using the two parametric fits, where *fit*1 is fit without censoring, and *fit*2 is with the presence of censoring according to the preselected scheme. Relative absolute differences compare the magnitudes of the errors:

$$\text{Relative absolute difference} = \frac{|fit2 - fit1|}{fit1}. \tag{5}$$

2.3 Censoring

A defining feature of the individual observations derived from the grouped frequency tables is that they represent intervals rather than exact values.

Generally, the value x_i of the i-th observation falls within the interval $(l_i; u_i]$. When $l_i = u_i$, the observation is complete (exact) with the value u_i. If $l_i = -\infty$, the observation is left-censored with interval $(-\infty; u_i)$. If $u_i = \infty$, the observation is right censored with interval $(l_i; \infty)$. The most general case is interval-censored observation, in which we know both values of l_i and u_i (Bogaerts et al., 2020; Chen et al., 2012; Liu, 2012).

In the case of the grouped frequencies, for all the individual observations in (re)constructed samples, we know intervals in which the value lies or whether the value is over the known value. Thus, the data that can be reconstructed from such tables are a mixture of interval-censored and right-censored observations, sometimes also called doubly-censored.

The censoring mechanism used is straightforward here, placing each value within a prespecified interval. In the simulations, censoring is thus noninformative, meaning the distribution of censoring values does not provide any additional information on the distribution of the original values and vice versa. The assumption of non-informativeness is important in statistical inference, validating standard estimating techniques (Chen et al., 2012).

If all the original values are presented in the frequency tables, truncation need not be considered, as it is defined as completely omitting values of the chosen interval. This should not be mistaken for censoring, in which all the original values are provided in the censored version (Liu, 2012).

For example, in the data from the CZSO, all values fall within the pre-selected earning band, indicating that the data are censored, not truncated. This censoring scheme is noninformative, as it does not provide any additional information on the underlying distribution.

Various estimating techniques may be used for parameter estimates of interval-censored data. Still, the most commonly applied is the maximum likelihood estimator (Chen et al., 2012), which is used in this paper. This estimator is obtained by numerically maximising the likelihood function in general form:

$$L(\boldsymbol{x};\mu,\sigma)=\prod_{i=1}^{n}\left[F(u_i;\mu,\sigma)-F(l_i;\mu,\sigma)\right], \tag{6}$$

where: μ and σ are parameter values to be manipulated, $F()$ denotes the cumulative distribution function of log-normal distribution, and l_i and u_i are lower and upper bounds of the interval, in which the actual value of i-th observation x_i lies, that is $l_i \_\_\_ x_i \_\_\_ u_i$ (Bogaerts et al., 2020).

2.4 Mixture of log-normal distributions

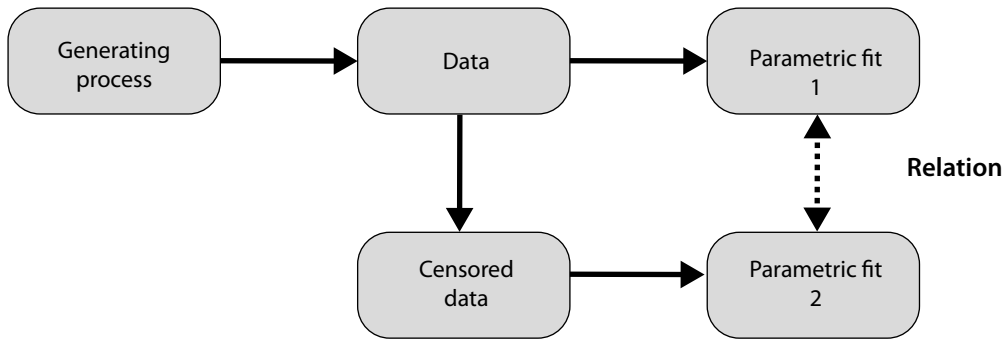
One cannot generally assume that the underlying distribution is purely log-normal. An alternative approach is to use a finite mixture of distributions. This approach assumes that the data originate from different sub-populations, regardless of whether the specific sub-population for each observation is known or unknown. The overall distribution is thus a convex combination of individual component distributions and their associated probabilities, known as mixing weights (McLachlan and Peel, 2000).

When the sub-population affiliation of an observation is unknown, the parameters of distributions and the mixing weights can be estimated using the EM algorithm. On the other hand, if the sub-group affiliation is known, the parameters for each specific sub-group distribution can be estimated directly. The relative frequencies of the sub-groups then can serve as estimators of the mixing weights (Dempster et al., 1977).

2.5 Simulation – general setting

This simulation study is inspired by the current real-world data structure on earnings provided by the CZSO and the MoLSA, aiming to derive some general insights. The approach involves using a parametric probability model to generate data akin to the original uncensored dataset and then computing the parametric fit of the chosen distribution. Subsequently, a prespecified censoring scheme is applied to all generated observations, and the parametric fit of the same distribution is calculated. This process mirrors the use of information from grouped frequency tables. The primary focus is on the relative absolute difference between the two fits of the same underlying data, which simulates the contrast between original uncensored and censored data possibly retrievable from the grouped frequency table. The process workflow is illustrated in Figure 1.

Figure 1 Simulation workflow (source)



Source: Own work

The phrase *parametric fit* is used intentionally here. The original real-world data motivating this simulation do not represent a random sample from any well-defined population, and it is unsustainable to assume a single underlying probability distribution. Thus, in a real-world application, there is no clear estimand. Therefore, parametric fit is the most appropriate term.

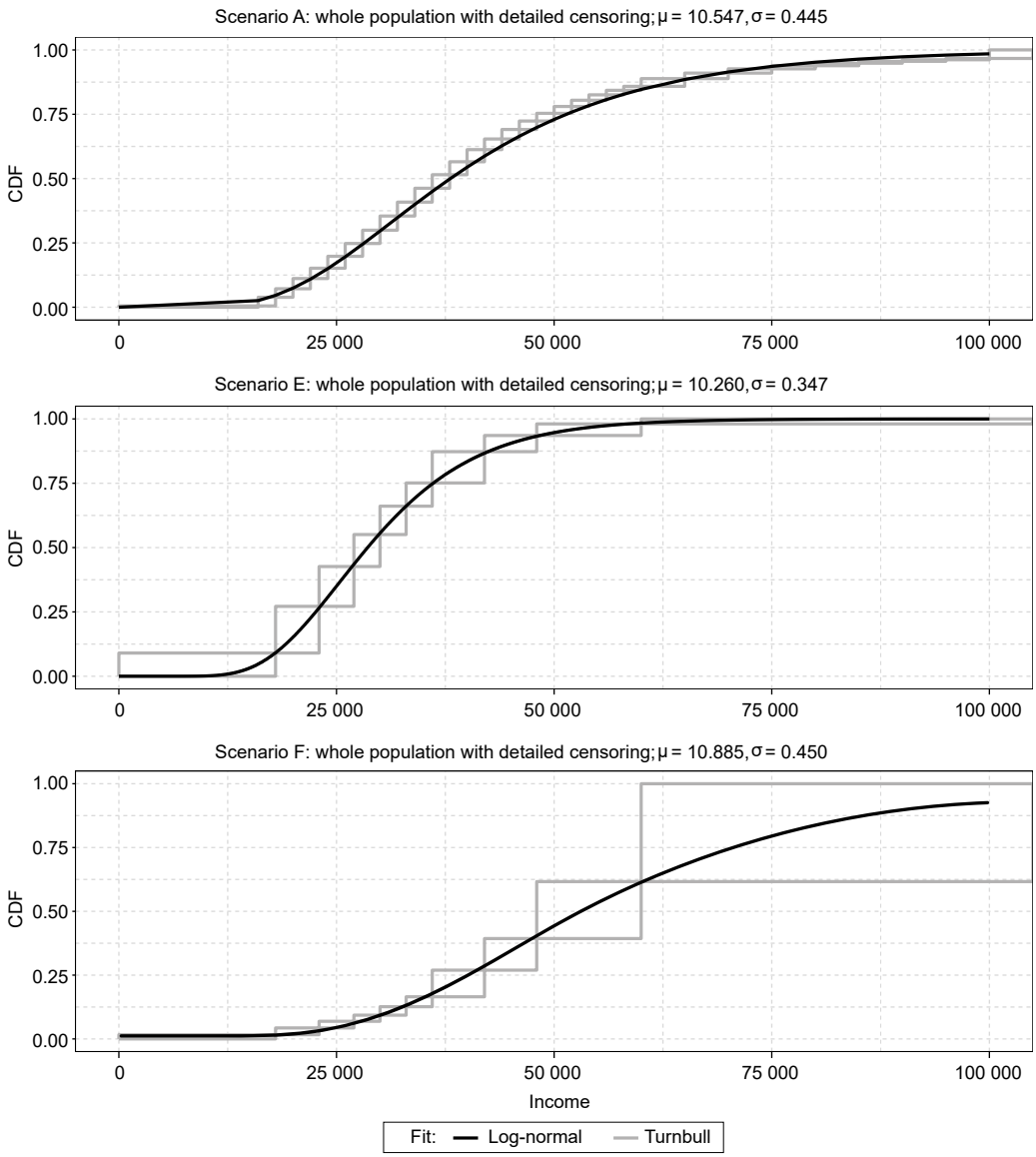
2.6 Simulation – details

The simulation process is divided into two stages. In the first stage, steps 1 to 3, the values for generating random samples were found. The second stage, steps 4 to 8, describes the simulation process itself.

2.6.1 Setting values for simulation

- **Step 1:** Data were retrieved from the CZSO and the MoLSA websites: general distribution of all observations across detailed earning bands (CZSO, 2023a); the distribution of observations by educational attainments and less detailed earning bands (CZSO, 2023b); the distribution of observations across the CZ-NACE sections and some joined divisions and detailed earning bands (MoLSA, 2024).
- **Step 2:** The retrieved contingency tables were transformed into individual data samples as described in Chapter 2.1, focusing on one variable at a time. This resulted in three datasets: all observations with no other variable other than earning band; all observations with earning bands and educational attainment; all observations with earning bands and the CZ-NACE section, and some joined divisions.
- **Step 3a:** Log-normal distributions were fitted to the individual data of general distribution and distribution of the lowest and the highest educational attainment using the maximum likelihood estimation. These fits, plotted in Figure 2, serve as the basis for three simulation scenarios (A, E and F), where the generating process is a log-normal distribution with the obtained parameters. The scenarios based on educational attainment are specifically selected to vastly differ from each other, allowing the observation of how the same censoring scheme on the different but correctly identified distributions affects the results.
- **Step 3b:** Similarly to Step 3a, log-normal fits were calculated across all the 36 available CZ-NACE sections and joined divisions. These fits serve as a basis for the other three simulation scenarios (B, C and D) using a mixture of log-normal distribution based on obtained parameters and mixing weights equivalent to relative frequencies of observations in the data. The mixture, shown in Figure 3, resembles the fitted log-normal distribution with a heavier tail. All the parametric values are in Table A,1, and the visualisation of the mixture's individual components' distributions are shown in Figure A1.

Figure 2 Parametric estimates of log-normal distributions used in scenarios A, E and F compared to the rectangles representing Turnbull estimates of empirical cumulative distribution functions (Turnbull, 1976)

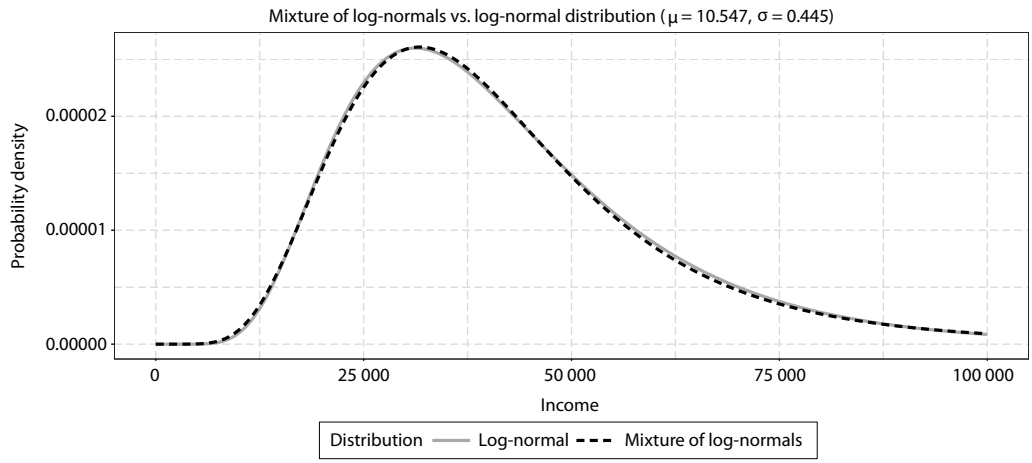


Source: CZSO, Trexima, own work

2.6.2 Simulation

- **Step 4a:** A random sample of the specified sample size is generated from the log-normal distribution with parameters determined by the scenario (A, E or F).
- **Step 4b:** Observations of the specified sample size are randomly assigned to the CZ-NACE section or joined divisions subgroups, using probabilities equivalent to the observed relative frequencies. For each observation, a random value is generated from the log-normal distribution, with parameters equal to those fitted for the specific subgroup.

Figure 3 Estimate of the mixture of log-normal distributions estimated for the CZ-NACE sections and joined divisions compared to the estimate of log-normal distribution for the whole dataset



Source: CZSO, Trexima, own work

- **Step 5:** The log-normal distribution is fitted in the generated random sample, and the resulting values, referred to as *fit1*, are documented.
- **Step 6:** The generated random sample is censored according to the preselected scheme. Three schemes are used: one that replicates more detailed earning bands provided by the CZSO (scenarios A and B), and another follows the less detailed bands (scenarios C, E and F). In scenario D, more detail is added in the right tail of the distribution compared to the detailed scheme of the CZSO – the values over 100 000 are additionally censored using 20 000 wide bands up to 200 000, and only values over 200 000 are right censored.
- **Step 7:** A log-normal distribution is fitted to the censored random sample, and the obtained values, referred to as *fit2,* are recorded.
- **Step 8:** Characteristics are calculated from the fits and compared using relative absolute differences.

In Table 2, a summary of the six simulation scenarios is provided. The term "General" population indicates that the scenario aligns closely with the observed distribution of all the observations from the CZSO-provided data, either using a log-normal distribution or a mixture of log-normal distributions.

Table 2 Short description of scenarios

Scenario	Population	Generating process	Bands
A	General	Log-normal	Detailed
B	General	Mixture of log-normal	Detailed
C	General	Mixture of log-normal	Less detailed
D	General	Mixture of log-normal	More detailed in the right tail
E	The lowest education	Log-normal	Less detailed
F	The highest education	Log-normal	Less detailed

Source: CZSO, Trexima, own work

Scenarios A, E and F employ the log-normal distribution as a generating process, making them somewhat idealistic. In these cases, fitting the model is equivalent to estimating the correctly identified distribution. Scenarios E and F differ in the proportions of observations in the highest interval, allowing a comparison to show how this difference impacts the estimates of the correctly specified distribution.

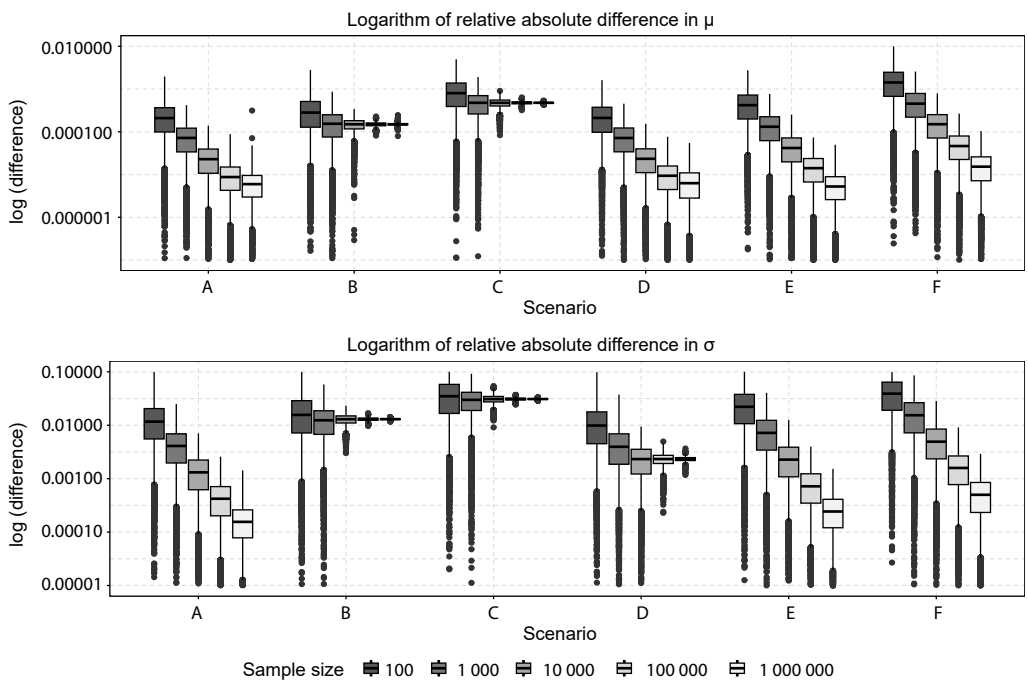
Scenarios B, C and D, which use a mixture of log-normal distributions, present more realistic conditions. Here, the estimated probability distribution does not exactly match the true underlying distribution, even though it is quite similar. The primary difference among these scenarios lies in the level of detail provided in the right tail of the censoring scheme. This demonstrates the effect of the censoring scheme on the fit of misspecified distribution.

Each scenario was simulated for the sample sizes 100, 1 000, 10 000, 100 000 and 1 000 000. Steps 4–8 in each scenario and each sample size were repeated 10 000 times.

3 RESULTS

Figure 4 graphically presents the results as boxplots of relative absolute differences in fitted parameters, plotted on logarithmic y-scale due to the results spanning four orders of magnitude. The second part of the results examines the association between parametric and characteristic fits, offering a different perspective.

Figure 4 Boxplots of relative absolute differences for fitted parameters of log-normal distribution on a logarithmic scale



Source: CZSO, Trexima, own work

In scenarios using the log-normal distribution as the generating process (A, E and F), the relative absolute differences converge to zero. Scenario F shows the slowest convergence, using the highest educational attainment for the generating process and featuring the largest share of observations in the top-right censoring interval (> 60 000). For sample sizes exceeding 100 000 observations, the mean relative absolute differences in the parametric values are below 0.05% for μ and 0.1% for σ.

In scenarios where a mixture of log-normal distributions is used as the generating process (B, C, and D), convergence to a constant value occurs, meaning the censoring scheme induces bias. The least detailed censoring scheme (C) induces the largest bias, which nearly diminishes for the most detailed scheme (D) with additional intervals introduced to the distribution's right tail. For scenarios with bias (B and C), there is notably faster reduction of variance.

Tables 3 and 4 detail the mean relative absolute differences for the two parameters. These converge to 0.015%, 0.048% and 0.001% for parameter μ and 1.3%, 3.1 % and 0.23% for parameter σ across the three scenarios, where the generating process is a mixture of log-normal distributions.

Table 3 Mean relative absolute differences of μ parameter for various scenarios and sample sizes (in %)

Sample size / scenario	A	B	C	D	E	F
100	0.026	0.037	0.097	0.026	0.050	0.173
1 000	0.008	0.017	0.050	0.009	0.016	0.055
10 000	0.003	0.015	0.048	0.003	0.005	0.018
100 000	0.001	0.015	0.048	0.001	0.002	0.006
1 000 000	0.001	0.015	0.048	0.001	0.001	0.002

Source: CZSO, Trexima, own work

Table 4 Mean relative absolute differences of σ parameter for various scenarios and sample sizes (in %)

Sample size / scenario	A	B	C	D	E	F
100	1.46	2.15	4.68	1.31	2.68	5.83
1 000	0.484	1.34	3.06	0.487	0.853	1.83
10 000	0.154	1.30	3.10	0.250	0.270	0.585
100 000	0.049	1.30	3.10	0.233	0.085	0.185
1 000 000	0.018	1.30	3.10	0.233	0.028	0.058

Source: CZSO, Trexima, own work

Results in Table 5 illustrate the magnitude of errors translated from parameters to various characteristics for the sample size of 100 000 observations. In examining "realistic" scenarios (B, C, and D), the location characteristics, especially the median, generally have average relative errors below 1%. Variability and shape characteristics typically exhibit average relative absolute errors ranging from 1% to 2% for detailed censoring and from 3% to 5% for less detailed censoring schemes. With the proposed more detailed censoring scheme applied to the right tail, average relative absolute errors fall between 0.2% and 0.3%, closer in magnitude to errors seen when using the log-normal distribution as the generating process.

Table 5 Mean relative absolute differences of selected characteristics parameter for various scenarios and sample size of 100 000 observations (in %)

Characteristic / scenario	A	B	C	D	E	F
Median	0.011	0.158	0.502	0.012	0.017	0.060
E(X)	0.013	0.431	1.14	0.054	0.014	0.096
S.D.(X)	0.062	1.86	4.52	0.313	0.092	0.299
Moment skewness	0.061	1.64	3.89	0.296	0.097	0.233
Moment kurtosis	0.076	2.10	4.91	0.382	0.085	0.292
95% quantile	0.035	1.14	2.81	0.181	0.041	0.194
Coefficient of variation	0.062	1.44	3.42	0.313	0.090	0.205
Gini coefficient	0.047	1.26	2.99	0.225	0.083	0.179

Source: CZSO, Trexima, own work

Figures A2 and A3 depict the associations between the fits of censored and uncensored data for parameters μ and σ across different scenarios. While the mean relative absolute differences are highest for scenario C, wherein a mixture of log-normal distributions is censored with the least detailed censoring scheme, conversely, the association is the weakest for scenario F, where the log-normal distribution is used with the same scheme. The disparity likely results from the bias-variance trade-off; scenario C displays high bias yet very low variance, while scenario F shows the opposite pattern. In Table A2, correlation coefficients for parameters and characteristics confirm this pattern. Under scenario F, the correlation coefficient of variations is only 0.315, compared to 0.639 for scenario C. These correlations are notably lower than those observed in other scenarios, where correlation coefficients are typically 0.9 or higher.

4 DISCUSSION

Bringing the results together reveals several key insights. First, when the actual generating process matches the distribution used for a parametric fit, the parameters and characteristics calculated from the censored data converge to those obtained from the original data. This confirms the consistency of the maximum likelihood estimator for the log-normal distribution under non-informative interval censoring. The speed of convergence depends on the level of censoring detail, corroborating the existing research on the parametric estimation theory for censored data (El-Sagheer, 2018; O'Neil and Wells, 1972; Shirazi et al., 2022; Vittal and Phillips, 2007; Xiao et al., 2016). With sample sizes typically observed in the CZSO data ranging from tens to hundreds of thousands, the mean relative absolute difference is negligible, staying below 0.3% for various characteristics.

Second, when the generating process does not exactly match the distribution used for the parametric fit, the parameters and characteristics calculated from the censored data do not converge to those obtained from the original data. In this case, fitting a log-normal distribution to a mixture of log-normal distributions with heavier tail systematically underestimates values. Convergence then occurs to specific constants depending on the censoring scheme.

The bias of the parametric fit of censored data relative to the fit of the original data is dependent on the censoring scheme´s detail; less detail induces more bias. The variance of the relative absolute

differences decreases sharply and more rapidly than in the previous type of scenarios. In a typical CZSO censoring scheme, the bias can be up to 5% for certain characteristics, leading to significant distortion.

However, the bias can be mitigated by introducing more detailed intervals in the distribution's right tail. This distortion in the parametric fit differs from the error caused by the probability model misspecification, arising specifically from censoring itself.

Third, while the errors are generally low for probability distribution fits matching their generating processes, high variance occurs with the least detailed censoring scheme. This results in very low correlation coefficients between parametric fits for original and censored data if the tail is cut at 60 000, resulting in a large share of data being above this threshold.

Relating these results to actual data provided by the CZSO, it becomes clear that fitting log-normal distribution to the data for higher earnings groups, such as the most educated or some CZ-NACE sections or joined divisions, likely results in underestimation compared to what could be observed using the original data. Providing more detailed information about frequencies in the distribution's right tail could help mitigate this statistical modelling issue.

CONCLUSION

This study evaluated the possible magnitude of the loss of precision in the parametric fit of a log-normal distribution when using grouped frequency tables instead of the original data. This precision is influenced by various assumptions and settings, some of which were explored here.

When assuming the data-generating process is a mixture of log-normal distributions, grouped frequency tables introduce additional bias in the parametric fit. For income data and, more generally, right-skewed heavy-tailed data, the level of detail in the right tail of the distribution is important. This bias stems not from the data itself but from the choice of censoring schemes or intervals used in the grouped frequency table. In the specific setting used in the simulations in this paper, this led to characteristics calculated being several per cent lower when distributional data placed a significant proportion of the observations in the highest, right-censored interval.

Furthermore, assuming the data-generating process is a log-normal, the provided simulations confirm consistency properties of maximum likelihood estimation under a non-overlapping mixture of interval and right censoring. The speed of convergence depends on the granularity of the censoring scheme.

The findings indicate that enhancing the granularity of the data provided in grouped frequency tables, particularly in the highest interval, can significantly improve the statistical modelling precision. Therefore, data providers should consider providing sufficient detail in this interval whenever feasible.

References

BARTOŠOVÁ, J. (2006). Logarithmic-Normal Model of Income Distribution in the Czech Republic. *Austrian Journal of Statistics*, 35(2–3): 215–221.

BOGAERTS, K., KOMAREK, A., LESAFFRE, E. (2020). *Survival Analysis with Interval-Censored Data*. 1st Ed. Routledge.

CHEN, D.-G., SUN, J., PEACE, K. E. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications* [online]. 1st Ed. Chapman and Hall/CRC. <https://doi.org/10.1201/b12290>.

CZSO. (2023a). *A11 Percentage of employees by gross monthly earnings band and by sex* [online]. Prague: Czech Statistical Office. <https://csu.gov.cz/docs/107508/771577f2-fb27-3b2d-85c1-921b9ed1af68/11002623a11.pdf?version=1.0>.

CZSO. (2023b). *A15 Percentage of employees by gross monthly earnings band and by educational attainment* [online]. Prague: Czech Statistical Office. <https://csu.gov.cz/docs/107508/70afe636-6fef-253b-c8ba-e4157ce7e974/11002623a15.pdf?version=1.0>.

CZSO. (2023c). *A23 Number of employees and their average gross monthly earnings by age and by educational attainment* [online]. Prague: Czech Statistical Office. <https://csu.gov.cz/docs/107508/41be1d6c-b7ce-8a94-6426-50b4772905d3/11002623a23.pdf?version=1.0>.

CZSO. (2023d). *Introduction* [online]. Prague: Czech Statistical Office. <https://csu.gov.cz/docs/107508/79cfc37d-eec8-356c-0502-11f27c0a9879/11002623ua.pdf?version=1.0>.
CZSO. (2023e). *Structure of Earnings Survey – 2022* [online]. Prague: Czech Statistical Office. <https://csu.gov.cz/produkty/structure-of-earnings-survey-2022>.
DELIGNETTE-MULLER, M. L. (2015). fitdistrplus: An R Package for Fitting Distributions [online]. *Journal of Statistical Software*, 64(4): 1–34. <https://doi.org/10.18637/jss.v064.i04>.
DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm [online]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1): 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
EL-SAGHEER, R. M. (2018). Estimation of parameters of Weibull–Gamma distribution based on progressively censored data [online]. *Statistical Papers*, 59(2): 725–757. <https://doi.org/10.1007/s00362-016-0787-2>.
JOHNSON, N. L., KOTZ, S., BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions, Vol 1.* 2nd Ed. Wiley.
KASSAMBARA, A. (2023). *ggpubr: 'ggplot2' Based Publication Ready Plots* (Version 0.6.0) [Computer software]. <https://CRAN.R-project.org/package=ggpubr>.
KOROBEYNIKOV, A. (2012). Consistency of Parametric MLE Under Mixed Case Interval Censoring [online]. *Communications in Statistics – Simulation and Computation*, 41(7): 1083–1092. <https://doi.org/10.1080/03610918.2012.625811>.
LIMPERT, E., STAHEL, W. A., ABBT, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues [online]. *BioScience*, 51(5): 341. <https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2>.
LIU, X. (2012). 1.4.2 Left censoring, interval censoring, and left truncation. In: *Survival analysis: Models and applications*, John Wiley & Sons Ltd., 6–7.
MAREK, L., VRABEC, M., BERKA, P. (2018). Ex-post Verification of Prediction Models of Wage Distributions [online]. *Applications of Mathematics and Statistics in Economics 2018 Conference Proceedings.* <https://www.amse-conference.eu/old/2018/wp-content/uploads/2018/10/Marek-Vrabec-Berka.pdf>.
MCLACHLAN, G. J., PEEL, D. (2000). *Finite Mixture Models.* 1st Ed. Wiley & Sons Ltd.
MOLSA. (2024). *Analýza vývoje příjmů a výdajů domácností ČR v roce 2022 a predikce na další období.* Prague: Ministry of Labour and Social Affairs.
O'NEIL, B., WELLS, W. T. (1972). Some Recent Results in Lognormal Parameter Estimation Using Grouped and Ungrouped Data. *Journal of the American Statistical Association*, 67(337): 76–80.
POTHARST, R. (2022). *Estimating the parameters of a lognormal distribution using grouped censored data* [online]. <https://doi.org/10.13140/RG.2.2.33629.28643>.
R CORE TEAM. (2023). *R: A Language and Environment for Statistical Computing* (Version 4.3.0 Already Tomorrow) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org>.
SHIRAZI, Z. A., DA SILVA, J. P. A. R., DE SOUZA, C. P. E. (2022). Parameter estimation for grouped data using EM and MCEM algorithms [online]. *Communications in Statistics – Simulation and Computation*, 53(8). <https://doi.org/10.1080/03610918.2022.2108843>.
TARIMA, S. S., VEXLER, A., SINGH, S. (2013). Robust Mean Estimation Under a Possibly Incorrect Log-Normality Assumption [online]. *Communications in Statistics – Simulation and Computation*, 42(2): 316–326. <https://doi.org/10.1080/03610918.2011.643850>.
TURNBULL, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data [online]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 38(3): 290–295. <https://doi.org/10.1111/j.2517-6161.1976.tb01597.x>.
VITTAL, S., PHILLIPS, R. (2007). Uncertainty Analysis of Weibull Estimators for Interval-Censored Data [online]. *2007 Proceedings – Annual Reliability and Maintainability Sympsoium*, 292–297. <https://doi.org/10.1109/RAMS.2007.328067>.
WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* [online]. Springer-Verlag. <https://ggplot2.tidyverse.org>.
WICKHAM, H., ROMAIN, F., LIONEL, H., MÜLLER, K., VAUGHAN, D. (2023). *dplyr: A Grammar of Data Manipulation* (Version 1.1.2) [R]. <https://CRAN.R-project.org/package=dplyr>.
XIA, J., MI, J., ZHOU, Y. (2009). On the Existence and Uniqueness of the Maximum Likelihood Estimators of Normal and Lognormal Population Parameters with Grouped Data [online]. *Journal of Probability and Statistics*, 2009(1): 310575. <https://doi.org/10.1155/2009/310575>.
XIAO, X., MUKHERJEE, A., XIE, M. (2016). Estimation procedures for grouped data – a comparative study [online]. *Journal of Applied Statistics*, 43(11): 2110–2130. <https://doi.org/10.1080/02664763.2015.1130801>.

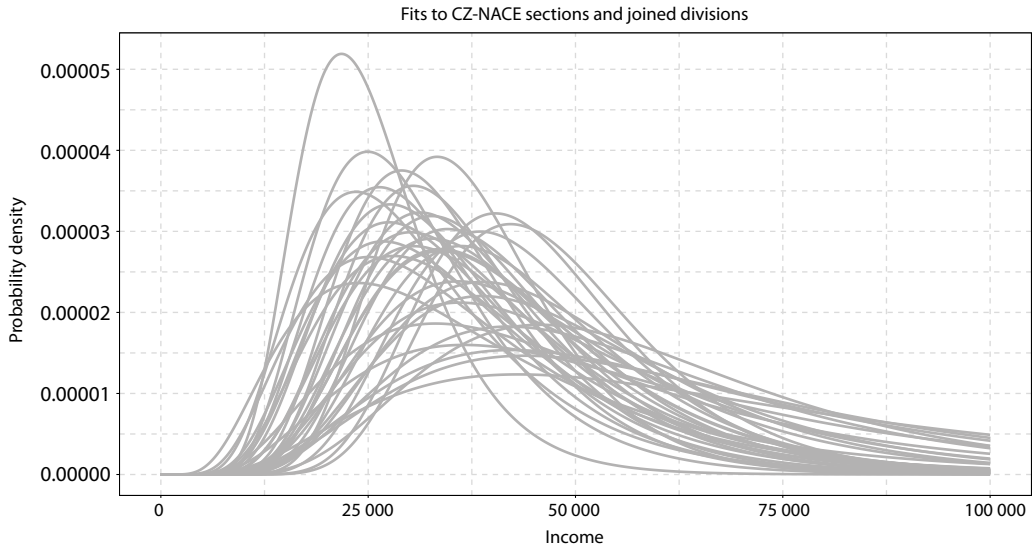
ANNEX

Table A1 Parameters and mixture proportions used in simulation settings

(sub)group	μ	σ	Mixture proportion (%)
Population	10.547	0.445	-
Lowest education	10.260	0.347	-
Highest education	10.885	0.450	-
CZ-NACE: A	10.395	0.345	2.3
CZ-NACE: B	10.691	0.294	0.5
CZ-NACE: CA (C10–C12)	10.403	0.426	2.4
CZ-NACE: CB (C13–C15)	10.264	0.374	0.7
CZ-NACE: CC (C16–C18)	10.388	0.399	1.8
CZ-NACE: CD (C19)	10.735	0.293	0.0
CZ-NACE: CE (C20)	10.669	0.427	0.8
CZ-NACE: CF (C21)	10.747	0.427	0.3
CZ-NACE: CG (C22–C23)	10.546	0.389	3.3
CZ-NACE: CH (C24–C25)	10.523	0.359	4.5
CZ-NACE: CI (C26)	10.584	0.397	1.1
CZ-NACE: CJ (C27)	10.580	0.394	2.5
CZ-NACE: CK (C28)	10.578	0.358	3.0
CZ-NACE: CL (C29–C30)	10.725	0.400	4.8
CZ-NACE: CM (C31–C33)	10.507	0.424	2.1
CZ-NACE: D	10.916	0.430	0.9
CZ-NACE: E	10.487	0.369	1.3
CZ-NACE: F	10.412	0.464	5.3
CZ-NACE: G	10.465	0.472	12.9
CZ-NACE: H	10.479	0.408	6.5
CZ-NACE: I	10.100	0.334	2.6
CZ-NACE: JA (J58–J60)	10.840	0.578	0.5
CZ-NACE: JB (J61)	10.936	0.533	0.5
CZ-NACE: JC (J62–J63)	11.054	0.619	2.5
CZ-NACE: K	10.978	0.537	1.9
CZ-NACE: L	10.398	0.518	1.2
CZ-NACE: MA (M69–M71)	10.715	0.555	2.9
CZ-NACE: MB (M72)	10.863	0.467	0.6
CZ-NACE: MC (M73–75)	10.433	0.593	0.9
CZ-NACE: N	10.261	0.441	4.0
CZ-NACE: O	10.666	0.328	7.6
CZ-NACE: P	10.631	0.365	7.7
CZ-NACE: QA (Q86)	10.711	0.468	5.8
CZ-NACE: QB (Q87–Q88)	10.500	0.293	2.1
CZ-NACE: R	10.443	0.347	1.2
CZ-NACE: S	10.335	0.395	1.1

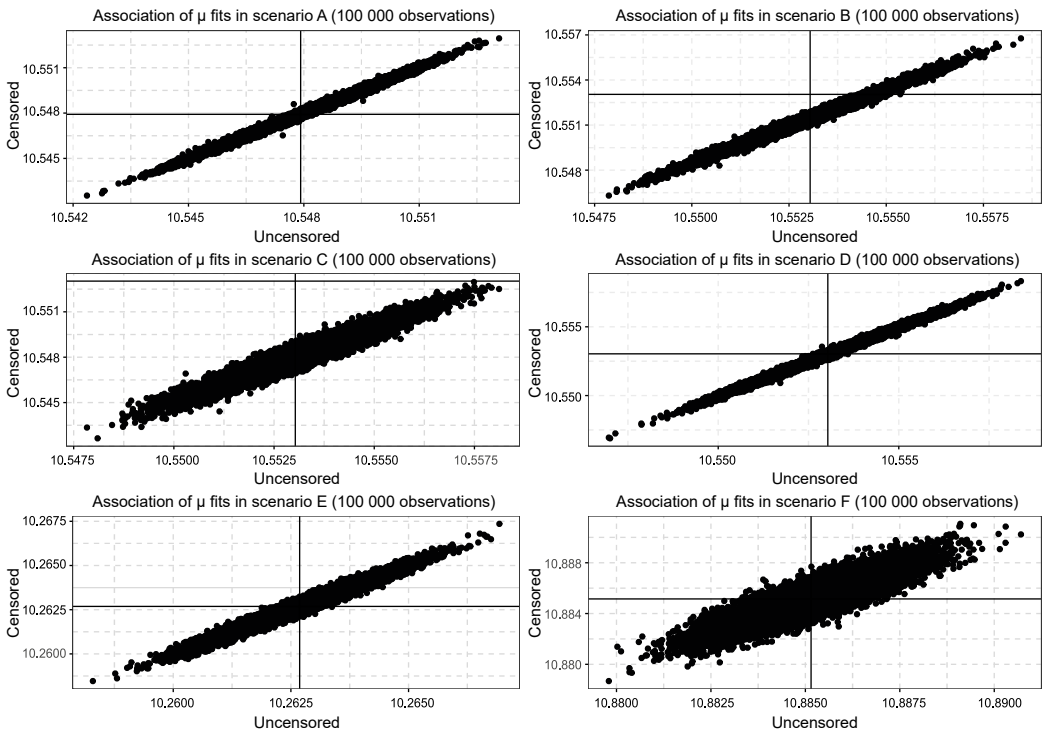
Source: Authors' computation

Figure A1 Estimated log-normal distributions of the individual components used for the mixture (parameters in Table A1)



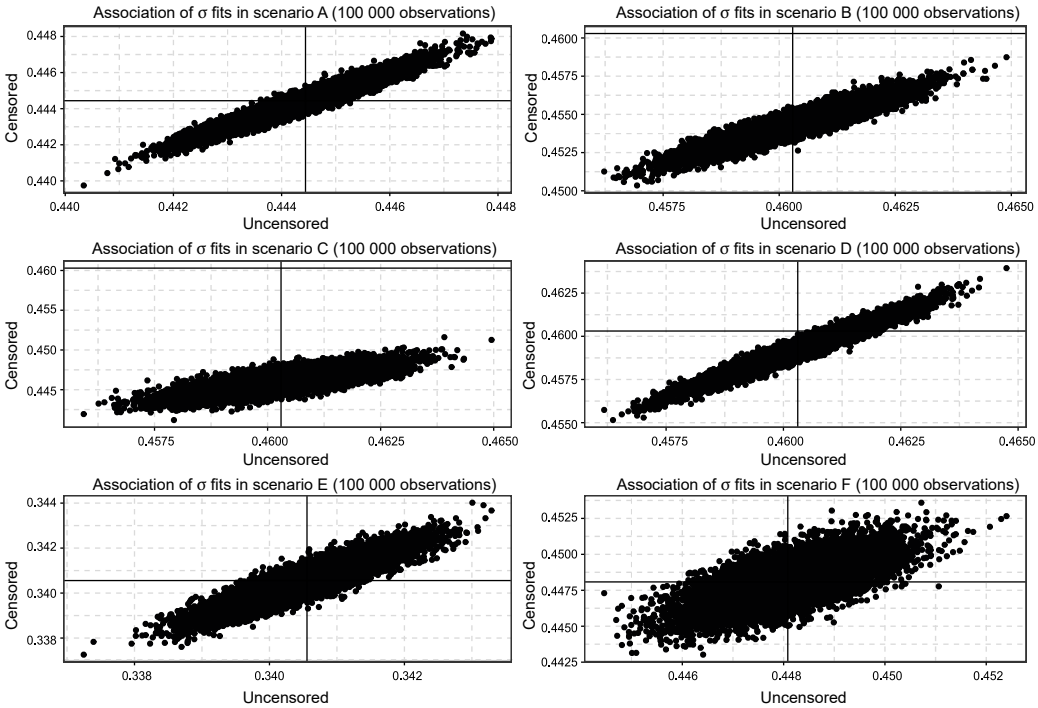
Source: CZSO, Trexima, own work

Figure A2 Association of μ fits across different scenarios (solid lines represent the mean of uncensored parametric fit)



Source: CZSO, Trexima, own work

Figure A3 Association of σ fits across different scenarios (solid lines represent the mean of uncensored parametric fit)



Source: CZSO, Trexima, own work

Table A2 Correlation coefficients of parametric fits for selected characteristics under various scenarios and sample size of 100 000 observations

Characteristic / scenario	A	B	C	D	E	F
μ	0.995	0.992	0.966	0.996	0.982	0.885
σ	0.964	0.928	0.775	0.972	0.905	0.691
Median	0.995	0.992	0.966	0.996	0.982	0.885
E(X)	0.994	0.982	0.911	0.997	0.987	0.781
S.D.(X)	0.972	0.938	0.776	0.981	0.928	0.655
Moment skewness	0.964	0.928	0.775	0.972	0.905	0.691
Moment kurtosis	0.964	0.928	0.775	0.972	0.905	0.691
95% quantile	0.98	0.951	0.804	0.987	0.955	0.668
Coefficient of variation	0.952	0.895	0.639	0.966	0.9	0.315
Gini coefficient	0.964	0.928	0.775	0.972	0.905	0.691

Source: CZSO, Trexima, own work