STATISTICS IN TRANSITION new series, September 2022
Vol. 23, No. 3, pp. VII–XII

sciendo

From the Editor

It is with great pleasure that we present our readers with the September issue consisting of 12 articles arranged, as usual, in three sections: *Original research articles*, *Other articles*, and *Research communicates and letters*. A wide spectrum of topics is discussed in these papers by authors from a large group of countries: USA, India, Iran, Algeria, Saudi Arabia, Sri Lanka, Nigeria, Thailand, and Poland.

Research articles

The issue starts with the paper by **Jacek Białek**, **Tomasz Panek**, and **Jan Zwierzchowski** entitled ***Assessing the effect of new data sources on the Consumer Price Index: a deterministic approach to uncertainty and sensitivity.*** The authors discuss the use of alternative sources of data about prices (scanned and scraped data) in the analysis of price dynamics with selecting the appropriate formula of the price index at the elementary group (5-digit) level as the one of the greatest challenges of the official statistics. The empirical study was based on data for February and March 2021, while scanner and scraped data about selected categories of food products were obtained from one retail chain operating hundreds of points of sale in Poland and selling products online. It was found that the choice of a data source has the most significant impact on the final value of the index at the elementary group level, while the choice of the aggregation formula used to consolidate different data sources is of secondary importance. The results indicate that consumer price indices calculated for the elementary groups of interest are characterised by a relatively low robustness to changes of a data source about consumer prices and by a relatively high robustness to changes of index formulas used for calculating price indices at the level of subgroups and elementary groups. For all elementary groups of interest, the first assumption, i.e. the choice of a given data source, has the biggest impact on the final value of the price index. Which index formula is used for aggregating indices for subgroups into elementary group indices has much less influence on the final results. The effect of choosing a particular index for aggregating indices derived from different sources is negligible.

Yeil Kwon in the article entitled ***A comparison of the method of moments estimator and maximum likelihood estimator for the success probability in the Fibonacci-type probability distribution*** shows that a Fibonacci-type probability distribution provides the probabilistic models for establishing stopping rules associated with the number of consecutive successes. It can be interpreted as a generalized version

of a geometric distribution. After revisiting the Fibonacci-type probability distribution to explore its definition, moments and properties, the authors proposed numerical methods to obtain two estimators of the success probability: the method of moments estimator (MME) and maximum likelihood estimator (MLE). The ways both of them performed were compared in terms of the mean squared error. A numerical study demonstrated that MLE tends to outperform MME for most of the parameter space with various sample sizes. A Fibonacci-type probability distribution can be employed to determine the probabilistic behaviour of a random variable N defined by the number of Bernoulli trials with a success probability p until there are k-consecutive successes. To compare MME with MLE, the authors used the computational methods to obtain MLE by approximating the maximum likelihood function using the pmf of N defined recursively. The result of the simulation discloses that, for both MLE and MME, the biases are considerably smaller than the variances under all of the values of p and the sample sizes, indicating that the variance explains the majority of MSE.

Mriganka Mouli Choudhury, **Rahul Bhattacharya**, and **Sudhansu S. Maiti** discuss ***Estimation of P(X ≤ Y) for discrete distributions with non-identical support***. The Uniformly Minimum Variance Unbiased (UMVU) and the Maximum Likelihood (ML) estimations of $R = P(X \leq Y)$ and the associated variance are considered for independent discrete random variables X and Y. Assuming a discrete uniform distribution for X and the distribution of Y as a member of the discrete one parameter exponential family of distributions, theoretical expressions of such quantities are derived. Similar expressions are obtained when X and Y interchange their roles and both variables are from the discrete uniform distribution. A simulation study is carried out to compare the estimators numerically. A real application based on demand-supply system data is provided. The UMVU and ML estimations of $P(X \leq Y)$ considering a discrete uniform distribution to represent stress and/or strength were discussed. However, an assumption of equal (but unknown) probability for stress and/or strength is less practical. Consequently, the authors intend further development with a general class of distributions to model stress and/or strength, allowing non-identical and parameter dependent supports.

In the next paper, ***Interval shrinkage estimation of parameter of exponential distribution in presence of outliers under loss functions,*** **Parviz Nasiri** examines estimators based on an interval shrinkage with equal weights point shrinkage estimators for all individual target points $\theta^- \in (\theta 0, \theta 1)$ for exponentially distributed observations in the presence of outliers drawn from a uniform distribution. The estimators obtained from both shrinkage and interval shrinkage were compared, showing that the estimators obtained via the interval shrinkage method perform better. The symmetric and asymmetric loss functions were also used to calculate the estimators. Finally, a numerical study and illustrative examples were provided to

describe the results. It is shown that the interval shrinkage estimator is better than the shrinkage estimator. Using different loss functions can also improve the performance of the estimator. The proposed method can be extended for Bayesian interval shrinkage estimation and other positive data distributions as well as for the presence of outliers from other distributions.

Adam Szulc focuses on ***Polish inequality statistics reconsidered: are the poor really that poor?*** The author critically analyses the data problem typically present in studies of income inequality. According to most empirical studies based on household surveys and considering the European standards, the recent income inequality in Poland is moderate and decreased significantly after reaching its peaks during the first decade of the 21st century. These findings were challenged by Brzeziński et al. (2022), who placed Polish income inequality among the highest in Europe. Such a conclusion was possible when the household survey data and information on personal income tax are analysed jointly. In this study the above-mentioned findings are explored using 2014 and 2015 data employing additional corrections to the household survey incomes. Incomes of the poorest people are replaced by their predictions made on a large set of well-being correlates, using the hierarchical correlation reconstruction. Applying this method together with the corrections based on Brzeziński's et al. results reduces the 2014 and 2015 revised Gini indices, still keeping them above the values obtained with the use of the survey data only. It seems that the hierarchical correlation reconstruction offers more accurate proxies to the actual low incomes, while matching tax data provides better proxies to the top incomes. The results of the present study only partly confirm findings by Brzeziński et al. (2022) on the serious underestimation of the Polish inequality indices. Corrections of the 2014 and 2015 survey income data applied to both tails of the distribution also results in inequality growth, however not so high and not for all types of inequality measures.

In their paper ***New polynomial exponential distribution: properties and applications,*** **Abdelfateh Beghriche**, **Halim Zeghdoudi**, **Vinton Raman**, and **Sarra Chouia** discuss the general concept of the XLindley distribution. Forms of density and hazard rate functions are investigated in the manuscript. Moreover, precise formulations for several numerical properties of distributions are derived. Extreme order statistics are established using stochastic ordering, the moment method, the maximum likelihood estimation, entropies and the limiting distribution. The authors demonstrate the new family's adaptability by applying it to a variety of real-world datasets, and a suggested family of distributions with only one parameter. Moments, distribution function, characteristic function, failure rate, stochastic order, maximum likelihood approach, and method of moments were among the properties studied. The Lindley and Zeghdoudi distributions lack the flexibility needed to examine and model many forms of data related to lifetime data and survival analysis. The NPED

distribution, on the other hand, is adaptable, straightforward, and simple to use. The novel distribution was used to evaluate two real data sets and was compared to existing distributions (Lindley, exponential, Zeghdoudi, Exponential Exponential and Xgamma). The comparison's findings support the NPED distribution's quality adjustment. The authors anticipate that the new distribution family will entice many additional life data, reliability analysis, and actuarial science applications, and it can employ a more general distribution with two parameters in future experiments.

Varathan Nagarajah's paper ***An improved ridge type of estimator for logistic regression*** demonstrates how to overcome the effect of multicollinearity in logistic regression. The proposed estimator is called a modified almost unbiased ridge logistic estimator (MAURLE). It is obtained by combining the ridge estimator and the almost unbiased ridge estimator. In order to assess the superiority of the proposed estimator over the existing estimators, theoretical comparisons based on the mean square error and the scalar mean square error criterion are presented. A Monte Carlo simulation study is carried out to compare the performance of the proposed estimator with the existing ones. Finally, a real data example is provided to support the findings. The superiority conditions for the proposed estimator with the existing MLE, LRE, and AURLE estimators are derived with respect to the MSE and SMSE criterions. Further, from the real data application and the Monte Carlo simulation study the authors notice that the proposed estimator performs well compared to MLE, LRE, and AURLE when the multicollinearity among the explanatory variables is high.

Sergiusz Herman examines ***Impact of restrictions on the COVID-19 pandemic situation in Poland*** focusing on the question of how the lockdown introduced in Poland affected the spread of the pandemic in the country. The study used the synthetic control method to this end. The analysis was carried on the basis of data from the Local Data Bank and a government website on the state of the epidemic in Poland. Results show that lockdown is an efficient tool that curbs the spread of the COVID-19 pandemic in Poland. Thanks to it, the number of new cases in the analysed region (in Poland) has diminished significantly (a drop of 9500 cases in three weeks was observed). Such a conclusion can be drawn from the performance of the placebo-in-space and placebo-in-time analyses. The research included the construction of the synthetic region that well illustrated the tendency of the pandemic development (in the Warmińsko-Mazurskie region) before the lockdown. After that the virus spread trajectories started to differ considerably.

In the next paper, **Wachirapond Permpoonsinsup** and **Rapin Sunthornwat** present ***Modified exponential time series model with prediction of total Covid-19 cases in Belgium, Czech Republic, Poland, and Switzerland***. The outbreaks in Belgium, the Czech Republic, Poland and Switzerland entered the second wave and was

exponentially increasing between July and November, 2020. Consequently, the authors estimated the compound growth rate, to develop a modified exponential time-series model compared with the hyperbolic time-series model, and the optimal parameters for the models based on the exponential least-squares, three selected points, partial-sums methods, and the hyperbolic least-squares for the daily COVID-19 cases in Belgium, the Czech Republic, Poland and Switzerland. The speed and spreading power of COVID-19 infections were obtained by using derivative and root-mean-squared methods, respectively. The optimal forecasting model with the estimated parameters was selected based on having the lowest RMSPE and the highest 2R. The results show that the exponential least-squares method was the most suitable for the parameter estimation. The compound growth rate of COVID-19 infection was the highest in Switzerland, and the speed and spreading power of COVID-19 infection were the highest in Poland between July and November, 2020. In conclusion, the authors maintain that the exponential least-squares method was relatively the most appropriate method for parameter estimation for the modified exponential time-series model for the daily COVID-19 cases, in all four countries.

Ahmad Aijaz, S. Qurat ul Ain's, Ahmad Afaq, and **Rajnee Tripathi's** article ***Poisson area-biased Ailamujia Distribution with applications in environmental and medical science***. The authors used compounding to develop a new distribution. A new Poisson area-biased Ailamujia distribution has been formulated to analyse count data. It was created by combining two distributions: the Poisson and area-biased Ailamujia distributions, using the compounding technique, to analyse count data. Several distributional properties of the formulated distribution have been studied. The distribution's ageing characteristics were determined and expressed explicitly. A variety of diagrams were used to demonstrate the characteristics of the probability mass function (pmf) and the cumulative distribution function (cdf). The parameter of the developed model was estimated using the well-known maximum likelihood estimation approach. Finally, two data sets were employed to demonstrate the effectiveness of the investigated distribution. It was shown that the Poisson area-biased Ailamujia distribution provides an appropriate fit for the two count data sets.

Other articles

The 38th Multivariate Statistical Analysis 2019, Lodz. Conference Papers. In the paper ***Triads or tetrads? Comparison of two methods for measuring the similarity in preferences under incomplete block design,*** **Artur Zaborski** compares two incomplete methods for measuring the similarity of preferences, i.e. the triad method and the tetrad method. These methods can be used whenever similarities are measured on an ordinal scale. They have been compared in terms of their labour intensity and ability to map the known structure of objects, even when all pairs of objects

in subgroups are not equal. The article indicates the possibility of reducing the number of sets presented to respondents in such a way that each pair of objects appears just as often, but less than their potential maximum number. In the example for 9 objects it was shown that scaling based on 8 tetrads gave a good solution. It was also demonstrated that the choice of the incomplete sets has no significant effect on the results of nonmetric multidimensional preference scaling, even when all pairs of objects cannot be presented equally frequently. This conclusion is particularly relevant for the creation of reduced sets when the number of objects does not allow to fulfil the condition of an equal number of pairs. The analysis indicated that the tetrad method can be used if each pair of objects appears in sets at least once, while for the method of triads each pair should appear at least twice.

Research Communicates and Letters

The section presents the paper by **Michael C. Ugwu** and **Mbanefo S. Madukaife** entitled ***Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage***. The authors introduce a new sampling design, namely a two-stage cluster sampling, where probability proportional to size with replacement (PPSWR) is used in the first stage unit and ranked set sampling in the second in order to address the issue of marked variability in the sizes of population units concerned with first stage sampling. An unbiased estimator of the population mean and total has been obtained, as well as the variance of the mean estimator. The authors calculated the relative efficiency of the new sampling design to the two-stage cluster sampling with simple random sampling in the first stage and ranked set sampling in the second stage. The results demonstrated that the new sampling design is more efficient as it produces a better estimator for estimating the population mean than a similar design built with simple random sampling in the first stage and ranked set sampling in the second stage units under the condition of significant variation in the sizes of the first stage units.

Włodzimierz Okrasa
Editor