SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 143

SUM OF GAMMA AND NORMAL DISTRIBUTION

Grzegorz SITEK

University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics;
grzegorz.sitek@ue.katowice.pl, ORCID: 0000-0002-7191-8631

Purpose: The article shows how to model audit errors using mixtures of probability distribution.
Design/methodology/approach: In financial accounting, data about the economic activities of a given firm is collected and then summarized and reported in the form of financial statements. Auditing, on the other hand, is the independent verification of the fairness of these financial statements. An item in an audit sample produces two pieces of information: the book (recorded) amount and the audited (correct) amount. The difference between the two is called the error amount. The book amounts are treated as values of a random variable whose distribution is a mixture of the distributions of the correct amount and the true amount contaminated by error. The mixing coefficient is equal to the proportion of the items with non-zero errors amounts.
Findings: The sum of normal and gamma distribution can be useful for modeling audit errors.
Originality/value: In this paper, the method of moments is proposed to estimate mixtures of probability distribution, and we derive a formulation of the probability distribution of the sum of a normally distributed random variable and one with gamma distribution. This research could be useful in financial auditing.

Keywords: mixture of probability, distribution, statistical auditing, sum of gamma and normal distribution, accounting error.

Category of the paper: empirical, scientific research.

Introduction

In probability theory, an exponentially modified Gaussian (exGaussian) distribution describes the sum of independent normal and exponential random variables. An exGaussian random variable Z may be expressed as $Z = X + Y$, where X and Y are independent, X is Gaussian with mean μ and variance σ^2 and Y is exponential of rate β. It has a characteristic positive skew from the exponential component. ExGaussian distribution is used as a theoretical model for the shape of chromatographic peaks (see Grushka, 1972), cellular biology (Golubev, 2010) and microarray preprocessing (Silver, 2009). A Gaussian minus exponential distribution

has been suggested for modelling option prices (Carr, 2009). Greene (1990) derived a formulation of the probability distribution of the difference between a normally distribution random variable and one with gamma distribution. Wywiał (2016, 2018) described a model of two Poisson distributions and a mixture of gamma probability distributions respectively. The sum of normal and gamma distribution can be useful for modeling audit errors.

Letting V_i denote the book amount of the ith item in the account $V = \sum_{i=1}^{N} V_i$ called the population book amount, at regular periods, an auditor samples n line items from the account and compares them against correct amounts. Therefore, let X_i denote the audited amounts for the ith line item and let $\varepsilon_i = V_i - X_i$ denote the error amount. The fundamental problem is the problem of constructing confidence limits for mean or totals in finite populations, when the underlying distribution is highly skewed and contains a substantial proportion of zero values. This situation is often encountered in statistical applications such as statistical auditing, reliability and insurance. The most distinctive feature of accounting data is the large proportion of line items without error, while an audit sample may not yield any nonzero error amounts. For analyses of such data, which most observations are zero, the classical interval estimation of the total error amount based on the asymptotic normality of the sampling distribution is not reliable. Johnson, Leitch and Neter (1981) observed that some of the accounts receivable have a J-shaped taint distribution with negative skewness. There are several distributions that also exhibit the same form of the distribution observed in accounting populations. These include the Gamma, Log-normal, Weibull, and Beta distributions. The error rates are usually very low, which render many existing statistical procedures inappropriate for estimating and hypothesis testing of error rates and error amounts.

There are two main types of audit tests for which the acquisition of information can profitably make use of statistical sampling. The first audit test, collecting data to determine the rate of procedural errors of a population of transations is called a compliance test. The second, collecting data for evaluating the aggregate monetary error in the stated balance, is called a substantive test of details. Inference on the total error amount is usually based on confidence intervals. Of course, they are related to testing problems. The decision-making process in auditing is treated as a problem of testing statistical hypotheses about admissibility of the total or the mean accounting errors. This approach lets us control not only significance level (risk of incorrect rejection), but also probability of the type II error appearing (risk of incorrect acceptance).

Substantive tests of details are concerned with the examination of the correctness of recorded monetary values in a financial statement. These tests provide direct evidence about the accuracy of total recorded monetary values. The auditor either applies substantive tests of detail extensively, or applies compliance tests to see if reliance on those controls are efficient and effective in reducing the tendency of material error in accounts. In compliance tests, the variable of interest is an error rate (proportion of transactions for which the internal control operates wrongly). Samples of transactions are used to make inferences about the error rate.

Many of the statistical methods adopted for quality control have been utilised in compliance testing. These methods are often referred to in the auditing context as attribute sampling (Robert, 1978). Based on the auditor's understanding of the accounting and internal control system, the attributes that indicate performance of a control, as well as possible conditions of deviation, are identified, e.g. failure to obtain suitable authorization for a purchase order, which does not necessarily lead to a monetary loss. The auditor generally makes a preliminary assessment of the rate of error he/she expects to find in the population to be tested and the level of control risk. This assessment is based on the auditor's prior knowledge or the examination of a small number of items from the population. The preliminary assessment is used by the auditor to design the audit sample and to determine the sample size.

An account, such as accounts receivable or inventory, is a population of N units known as line items. The dollar amounts that are recorded are called book amounts. Book value is the value recorded for accounts or financial statements. A sample is a selection of some, but not all, of the accounts. The information gathered from the sampled accounts is used to make inferences about the population. The only way to obtain the total value of the accounts is to audit all accounts, but this would be very costly. Not having to verify the information on all accounts to make these inferences reduces the cost in calculating the quantity of interest. A sampling approach is considered statistical if the selection of sampling items are random, each item having a calculated probability of being selected. Inferences about the population parameters may be made from the sample statistics. Random sampling enables the auditor to project sample results mathematically and to state, with measurable precision and confidence, the estimated rate of deviation in the population under audit (compliance audit sampling), or the estimated monetary misstatement in the population (substantive audit sampling).

The most important benefit which statistical sampling offers is reduction of the risk of overauditing or underauditing. The auditor's ultimate desire is to plan audits in a way that minimizes the total expected cost of performing the audit procedures while also giving a fair opinion on the financial statement. Sampling is therefore important in meeting these requirements. There are two audit procedures for which statistical sampling has been utilized. These are compliance and substantive tests. Statistical sampling in auditing seeks to assist auditors in using random selection methods and statistical evaluation techniques in testing, whether for compliance or substantive purposes. The objective is to reduce the risk of biased selection and quantify the sampling confidence level achieved.

In simple random sampling (SRS), a sample of line items of the fixed size n is drawn one by one with the same probability but without replacement. That is, each draw is carried out among items that have not already been chosen. There are thus $\frac{N!}{(N-n)!}$ samples each consisting of a combination of n of the N line items, and each such sample item has the probability $\frac{(N-n)!}{N!}$ of being selected. Robert (1978) gives a detailed account of the simple random sampling in auditing.

Stratified random sampling in auditing consists of dividing the auditing population into strata according to the sizes of the recorded amounts, and then selecting sampling items from each stratum independently by simple random sampling without replacement. Cyert, Hinckley, and Monteverde (1960) introduced the idea of achieving greater sampling efficiency through stratified sampling in auditing. A number of methods of stratifying audit populations effectively have been suggested, for example, by Arkin (1974) and Robert (1978).

Probability proportional to size sampling (PPS) is sampling with unequal probabilities of selecting items. If items with larger values are relatively more important, then sampling with probability proportional to size will be useful. With PPS selection a sample of line items are taken in such a way that the inclusion probability, π_i is proportional to v_i, that is $\pi_i = \frac{nv_i}{V}$, provided that $v_i < \frac{v}{n}$. This design implicitly stratifies the sample by recorded amount. PPS was originally developed in survey sampling theory by Hansen and Hurwitz (1943) for selection of clusters of unequal size. In the auditing context, this method and its variations are referred to as Monetary Unit Sampling (MUS), also known as Dollar Unit Sampling. The idea of using individual monetary values as the sampling units was suggested by Deming (1960). The basic concept of monetary unit sampling in auditing was developed independently, first by van Heerden (1961) and later by Stringer (1963). van Heerden suggested that an account balance or the line item could be regarded as a cluster of monetary units being either correct or in error. Monetary unit sampling was made popular by the work of Anderson and Teitlebaum (1973).

Sum of gamma and normal distribution

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two continuous random variables with density functions $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{y})$, respectively. Assume that both $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{y})$ are defined for all real numbers. Then the convolution $\boldsymbol{f} * \boldsymbol{g}$ of $\boldsymbol{f}$ and $\boldsymbol{g}$ is the function given by:

$$(\boldsymbol{f} * \boldsymbol{g})(\boldsymbol{z}) = \int_{-\infty}^{\infty} \boldsymbol{f}(\boldsymbol{z}-\boldsymbol{y})\boldsymbol{g}(\boldsymbol{y})\boldsymbol{dy} = \int_{-\infty}^{\infty} \boldsymbol{g}(\boldsymbol{z}-\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})\boldsymbol{dx} \quad (1)$$

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two independent random variables with density functions $\boldsymbol{f_X}(\boldsymbol{x})$ and $\boldsymbol{f_Y}(\boldsymbol{y})$ defined for all $\boldsymbol{x}$. Then the sum $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$ is a random variable with density function $\boldsymbol{f_Z}(\boldsymbol{z})$, where $\boldsymbol{f_Z}$ is the convolution of $\boldsymbol{f_X}$ and $\boldsymbol{f_Y}$. Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two independent random variables. Let $\boldsymbol{X}$ be a random variable with gamma density:

$$\boldsymbol{f_X}(\boldsymbol{x}) = \frac{\boldsymbol{\beta^\alpha}}{\boldsymbol{\Gamma(\alpha)}} \boldsymbol{x^{\alpha-1}} \boldsymbol{e^{-\beta x}} \boldsymbol{for\ x > 0} \quad (2)$$

Let $\boldsymbol{Y}$ be a random variable with normal density:

$$\boldsymbol{f_Y(y) = \frac{e^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}} \tag{3}$$

The density function for the sum $\boldsymbol{Z = X + Y}$ is given by

$$\boldsymbol{f_Z(z) = \frac{\beta^\alpha \int_0^\infty x^{\alpha-1} e^{(-\beta x)-\frac{(-\mu-x+z)^2}{2\sigma^2}}\, dx}{\sqrt{2\pi}\sigma\Gamma(\alpha)}} \tag{4}$$

Plancade (2012) introduced gamma-normal convolution to model the background correction of Illumina BeadArrays. We use Wolfram Mathematica to calculate the integral, and we give here only the result:

$$\boldsymbol{f_Z(z) = \frac{2^{\frac{\alpha-3}{2}-\frac{1}{2}}\beta^\alpha\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}-\frac{\alpha}{2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}}{\sqrt{\pi}\sigma\Gamma(\alpha)}\left(\frac{\sqrt{2}\Gamma\left(\frac{\alpha}{2}\right){}_1F_1\left(\frac{\alpha}{2};\frac{1}{2};\frac{\left(\beta\sigma^2-z+\mu\right)^2}{2\sigma^2}\right)}{\sqrt{\frac{1}{\sigma^2}}} + 2\Gamma\left(\frac{\alpha+1}{2}\right)(-\beta\sigma^2-\mu+z)\,{}_1F_1\left(\frac{\alpha+1}{2};\frac{3}{2};\frac{\left(\beta\sigma^2-z+\mu\right)^2}{2\sigma^2}\right)\right)} \tag{5}$$

where ${}_1\boldsymbol{F_1}$-is Kummer's confluent hypergeometric function.

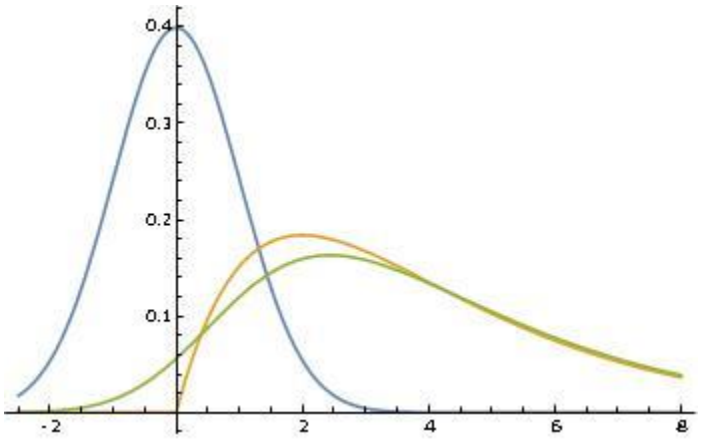

Figure 1. The density function for the sum of gamma and standard normal.

Sum of exponential and normal distribution

Let X be a random variable with exponential density with parameter β and Y be a variables with normal density. The density function for the sum $Z = X + Y$ is given by:

$$f_Z(x) = \frac{1}{2}\beta e^{\frac{1}{2}\beta(\beta\sigma^2+2\mu-2x)}\operatorname{erfc}\left(\frac{\beta\sigma^2+\mu-x}{\sqrt{2}\sigma}\right) \tag{6}$$

where $erf(x) = 1 - erfc(x) = \frac{2\int_0^x e^{-t^2}\,dt}{\sqrt{\pi}}$

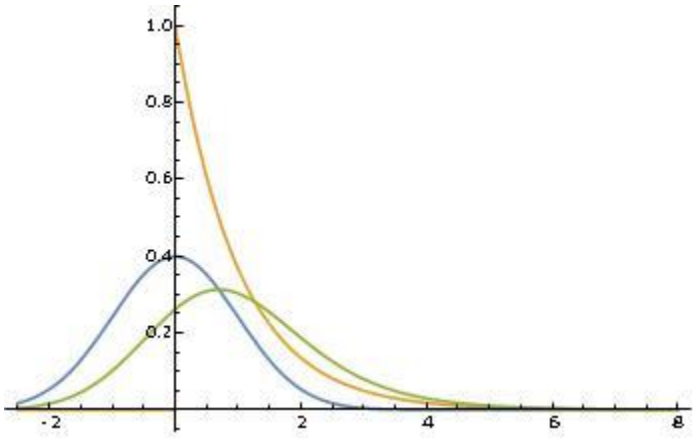

Figure 2. The density function for the sum of exponential(1) and standard normal.

The CDF function for the sum $Z = X + Y$ is given by:

$$F_Z(t) = \frac{1}{2}\left(\operatorname{erf}\left(\frac{t-\mu}{\sqrt{2}\sigma}\right) - e^{\frac{1}{2}\beta(\beta\sigma^2+2\mu-2t)}\operatorname{erfc}\left(\frac{\beta\sigma^2+\mu-t}{\sqrt{2}\sigma}\right) + 1\right) \tag{7}$$

Mixture of gamma and sum of gamma and normal probability

Wywiał (2018) proposed the following model based on mixtures of distributions. The pharse "mixture of distributions" usually refers to a situation in which the ith of k underlying distributions is chosen with probability $p_i, i = 1, \ldots, k$. The selection probabilities are usually unknown and the number of underlying distributions k may be fixed or a random. In general, the word ‘mixture’ refers to a convex combination of distributions or random variables. The following mixtures of probability distributions seem to be useful: mixtures of gamma distributions, Pareto, lognormal or mixtures of Pearson's type distributions with positive skewness. From another point of view, the types of distributions of true accounting amounts and accounting amounts contaminated with errors do not have be the same. The probability density of the observed accounting amounts is a mixture of density $f_0(x)$ of the true amounts and density $f_1(x)$ of the amounts contaminated by errors.

$$f(x) = (1-p)f_0(x) + pf_1(x) \tag{8}$$

Let $X \sim G(\alpha, \beta)$ and $Y \sim N(\mu, \sigma)$ be independent and $Z = X + Y$.

$$f(x|\alpha, \beta, \mu, \sigma) = (1-p)f_0(x|\alpha, \beta) + pf_1(x|\alpha, \beta, \mu, \sigma) \tag{9}$$

where $f_1(x|\alpha, \beta, \mu, \sigma)$ is the density of the variable Z and

$$f_0(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} x > 0 \tag{10}$$

Let X and Y be independent where Y is the accounting error. Hence $Z = X + Y$, $V = (1-W)X + WZ$, where $W = 1$ when an accounting error occurs $P(W = 1) = p$ and $W = 0$ when it does not occur $P(W = 0) = 1 - p$. The basic moments of the random variable V are:

$$E(V) = (1-p)E(V|W = 0) + pE(V|Z = 0) = (1-p)E(X) + pE(Z) \tag{11}$$

$$C_2(V) = p(1-p)(E(Z) - E(X))^2 + pC_2(Z) + (1-p)C_2(X) \tag{12}$$

$$C_3(V) = p(1-p)(1-2p)(E(Z) - E(X))^3 - 3p(1-p)(E(Z) - E(X)))C_2(X) + 3p(1-p)(E(Z) - E(X)))C_2(Z) + pC_3(Z) + (1-p)C_3(X) \tag{13}$$

$$C_4(V) = p(1-p)(3p^2 - 3p + 1)(E(Z) - E(X)) + 6p(1-p)^2(E(Z) - E(X))^2 C_2(Z) + +6p^2(1-p)C_2(X) - 4p(1-p)(E(Z) - E(X))C_3(Z) + 4p(1-p)(E(Z) - E(X))C_3(X) + pC_4(Z) + (1-p)C_4(X) \tag{14}$$

where $C_r(V) = E(V - E(V))^r$. Based on expressions (11)-(12), we obtain:

$$E(V) = \frac{\alpha}{\beta} + p\mu \; C_2(V) = p(1-p)\mu^2 + p\sigma^2 + \frac{\alpha}{\beta^2} \tag{15}$$

We can estimate parameters of the distribution mixture by means of the method of moments.

Let $\tau = E(\bar{V} - \bar{X})$ be the expected mean accounting error. Audit purpose is inference on τ or pn the expected total accounting $N\tau$. In particular, when we assume that τ_0 is the admissible mean accounting error, then the inference reduces to testing the following hypothesis:

$$H_0: \tau \leq \tau_0 \;\; H_1: \tau > \tau_0 \tag{16}$$

The basic idea of any sort of hypothesis test is to compare the observed value of a test statistic, say $\hat{\tau}$, with the distribution that it would follow if the null hypothesis were true. The null is then rejected if $\hat{\tau}$ is sufficiently extreme relative to this distribution. In most cases of interest to econometricians, however, the distribution of the test statistic we use is not known. We therefore have to compare $\hat{\tau}$ with a distribution that is only approximately correct. In consequence, the test may overreject or underreject.

Inference based of sample moments

We assume that $\mu = 0$. In this situation, expressions (11)-(15) lead to following results:

$$\begin{cases} E(V) = \frac{\alpha}{\beta} \\ V(V) = \frac{\alpha+\beta^2 p\sigma^2}{\beta^2} \\ C_3(V) = \frac{2\alpha}{\beta^3} \\ C_4(V) = \frac{3(\alpha^2+p\beta^4\sigma^4+2\alpha(1+p\beta^2\sigma^2))}{\beta^4} \end{cases} \tag{17}$$

The solution of this system provides estimators of parameters α, β, σ^2 and p.

$$\begin{cases} \alpha = \frac{\sqrt{2}(E(V))^{3/2}}{\sqrt{C_3(V)}} \\ \beta = \frac{\sqrt{2}\sqrt{E(V)}}{\sqrt{C_3(V)}} \\ \sigma^2 = \frac{6(E(V))^{3/2}C_2(V)C_3(V)-3\sqrt{2}(E(V))^2(C_3(V))^{3/2}+2\sqrt{2}E(V)\sqrt{C_3(V)}(6(C_2(V))^2-C_4(V))+}{6\sqrt{E(V)}(E(V)C_3(V)-2(C_2(V))^2)} \\ \frac{\sqrt{E(V)}(6(C_3(V))^2-4C_2(V)C_4(V))+6\sqrt{2}C_2(V)(C_3(V))^{3/2}}{6\sqrt{E(V)}(E(V)C_3(V)-2(C_2(V))^2)} \\ p = \frac{3(\sqrt{2}(E(V))^{3/2}\sqrt{C_3(V)}(12(C_2(V))^3-4C_2(V)C_4(V)+3(C_4(V))^2)+3(E(V))^3C_3(V)^2)+}{D} \\ \frac{3(2(E(V))^2C_3(V)(C_4(V)-9(C_2(V))^2)))+}{D} \\ \frac{3(6\sqrt{2}\sqrt{E(V)}(C_2(V))^2(C_3(V))^{3/2}+4E(V)C_2(V)(C_2(V)C_4(V)-3C_3(V)^2)))}{D} \end{cases} \tag{18}$$

where $D = 9(E(V))^3(C_3(V))^2 + 12(E(V))^2C_3(V)(C_4(V) - 6(C_2(V))^2) + E(V)(4(C_4(V))^2 - 72C_2(V)(C_3(V))^2) - 18(C_3(V))^3$ and $D \neq 0$.

More details about inference on mixtures of probability distributions can be found in the book by McLachlan and Peel (2000), where, e.g., the well-known EM algorithm is used to evaluate the maximum likelihood estimators.

Conclusion

In this paper it was shown how to estimate parameters of distribution mixtures by means of the method of moments when the expected value of normal distribution is 0. In the general case, estimators of the maximum likelihood method and the method of moments are usually the solutions of the systems of non-linear equations. In order to calculate those solutions, some numerical methods have to be used.

Acknowledgement

This paper is the result of a grant supported by the *National Science Centre, Poland*, no. 2016/21/B/HS4/00666.

References

1. Anderson, R.J. and Teitlebaum, A.D. (1973). Dollar-unit Sampling. *Canadian Chartered Accountant, April*, 30-39.
2. Arkin, H. (1984). *Handbook of Sampling for Auditing and Accounting*. New York: McGraw Hill.
3. Carr, P., Madan D.B. (2009). Saddlepoint Methods for Option Pricing. *The Journal of Computational Finance, 13, 1,* 49-61.
4. Cyert, R.M., Hinckley, G.M., Monteverde, R.J. (1960). Statistical Sampling in the Audit of the Air Force Motor Vehicle Inventory. *The Accounting Review, 35*, 667-673.
5. Deming, D.W. (1960). *Sampling Design in Business Research*. New York: Wiley.
6. Golubev, A. (2010). Exponentially modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. *Journal of Theoretical Biology, 262(2).* doi: org/10.1016/j.jtbi.2009.10.005.
7. Greene, W.H. (1990). A gamma-distributed stochastic fronier model. *Journal of Econometrics, 46.* North-Holland, 141-163.
8. Grushka, E. (1972). Characterization of Exponentially Modified Gaussian Peaks in Chromatography. *Analytical Chemistry. 44(11),* 1733-1738.
9. Hansen, M.H., Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics, 14,* 332-362.

10. Johnson, J.R., Leitch, R.A., Neter, J. (1981). Characteristics of Errors in Accounts Receivables and Inventory Audits. *Accounting Review, 58*, 270-293.
11. McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. New York:Wiley.
12. Plancade, S., Rozenholc, Y., Lund, E. (2012). Generalization of the normal- exponential model: exploration of a more accurate parameterisation for the signal distribution on Illumina BeadArrays. *BMC Bioinformatics, 13(329)*. doi.org/10.1186/1471-2105-13-329.
13. Roberts, D. (1978). *Statistical Auditing*. New York: American Institute of Certified Public Accountants.
14. Silver, J. (2009). Microarray background correction: maximum likelihood estimation for the normal-exponential convolution model. *Biostatistics, 10(2),* 352-363.
15. Stringer, K.W. (1963). *Practical Aspects of Statistical Auditing*. Preceeding of Business and Economic Statistics Section of the American Statistical Association, 405-41.
16. Wywiał, J.L. (2016). *Contributions to Testing Statistical Hypotheses in Auditing*. Warsaw: PWN, 91-95.
17. Wywiał, J.L. (2018). Application of two gamma distributions mixture to financial auditing. Sankhyã, B. *The Indian Journal of Statistics*. doi: org/10.1007/s13571-018-0154-5.