Pakistan Journal of Statistics and Operation Research

Approximations to the Moments of Order Statistics for Normal Distribution

Asuman Yılmaz[1*], Mahmut Kara[2]

* Corresponding Author

1.Faculty of Economics and Administrative Sciences, Department of Econometrics, Turkey, asumanduva@yyu.edu.tr

2. Faculty of Economics and Administrative Sciences, Department of Econometrics, Turkey, asumanduva@yyu.edu.tr

Abstract

Order statistics occupy an important place in statistical theory. They have an important place in many fields of applied statistics such as goodness of fit tests and parameter estimation. In addition, it is necessary to find the expected values of these order statistics in these application areas. However for some probability distributions, these expected values are very difficult to find such as the standard normal distribution. So the problem of finding the expected values of the order statistics in statistical theory is of importance. In this study, two novel approximation methods are proposed for the expected values of the order statistics of the standard normal distribution. Also, the true values with previously given approximations, simulation results and our proposed approximations are compared by using mean square error (MSE), mean absolute error (MAE) and maximum error (ME) criteria. Furthermore, to evaluate the performances of all approximation methods, we compute the differences between exact values and approximation values. Then, the plot of these differences against the exact values is given. Based on both the plots and the comparison results, novel approximations fit the true values better than the other approximations presented in this paper.

Key Words: Order statistics, Standard normal distribution, Expected value

Mathematics Subject Classification: 62E15 ,62E17, 62E20.

1. Introduction

The normal distribution is one of the most important probability distributions. It is of considerable importance in many application areas such as chemistry, economics, engineering, financial risk management, genetics, environmental sciences, accelerated life testing, medicine, reliability theory, and so on. It has been extensively studied from different perspectives by many researchers, see D'Agostino (2017), Du, Fan, and Wei (2022), Evans and Hastings (2022), Balakrishnan and Nevzorov (2003), Atangana and Gómez-Aguilar (2017), Liu et al. (2020); Guan et al. (2019), Saha et al. (2022). This distribution is also referred to as the Gaussian distribution. If a random variable X has the normal distribution with parameters mean μ and variance σ^2 then X is denoted as $X \sim N(\mu,\sigma^2)$ in shorthand. A normal distribution with $\mu = 0$ and $\sigma^2 = 1$, that is, $Z \sim N(0,1)$, is also known as the standard normal (SN) distribution. Let $Z_1, Z_2, ..., Z_n$ be a random sample such that $Z \sim N(0,1)$. Then, its probability density function (pdf) and cumulative density function (cdf) are given by.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty \tag{1}$$

and

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \phi(t)dt, \tag{2}$$

respectively.

Let Z_i, $i=1,2,...,n$ be the ordered random variables and $Z_{(1)} < Z_{(2)} < ... < Z_{(n)}$ be the ordered statistics. The order statistics are very important in statistics. There is a huge literature considering the order statistics under different scenarios, see Jäntschi (2020), Lando et al. (2021), Pahins, Ferreira et al. (2019), Sharma et al. (2020), Fowlie, et al. (2020), Triantafyllou (2018). They are widely used in statistical theory and practice, such as statistical inference, modelling goodness-of-fit tests, reliability theory, economics and operations research, hypothesis testing, and describing data in the context of *L*-moment and TL-moment estimators, see Zhao and Balakrishnan (2015), Headrick and Pant (2012). Therefore, it is necessary to find the expected values (EVs) of order statistics in these application areas. However, for some probability distributions, these EVs are very difficult to find such as the standard normal distribution.

The EV of the ith order statistics of a random sample from an SN distribution is given by

$$E\left(Z_{(i)}\right) = \frac{n!}{(n-i)!(i-1)!}\int_{-\infty}^{\infty} z\left(\Phi(x)\right)^{i-1}\left(1-\Phi(x)\right)^{n-i}\phi(z)\,dz.$$

(3)

Evaluation of the EV for the order statistics given in (3) is problematic, since it is not easy to integrate $\Phi(z)$ analytically. In this situation, it is impossible or very difficult to calculate many statistics, such as the difference between the minimum and the maximum order statistics of different sample sizes. Thus, numerous researchers have tried to find analytical solutions to the EV of equation (3), see Jones (1948), Ruben (1956), Bose and Gupta (1959), and David (1963). Their method is accurate enough but it is tedious and fails for samples $n > 5$, see Harter (1996). Then, Chen and Tyle (1999) presented an expression to approximate of the EVs of the least-order statistics and the greatest-order statistics in the SN distribution. Theoretical integration of this equation without the aid of a computer is intractable. Therefore, a numerical integration method is necessary. Many writers tabulated different sample sizes with different decimal places by using numerical integration, see Parrish (1992), Fisher and Yates (1943), Hastings et al. (1947), Harter (1961). These tables only contain selected sample sizes. Moreover, it is quite time-consuming to achieve accurate numerical integration. Therefore, these are not very practical and have limited accuracy in most cases. For more details see, Chen and Tyle (1999). Many approximation formulas have been proposed to approximate the EV of the order statistics of the SN distribution in the current literature. Blom's (1958) and Filliben's (1975) approximations are the most well-known among them. These approximations draw attention especially, in the different studies in the last decades, see Olivera and Heard (2019), Sulewski, (2021), Boylan and Byung (2012), Cho (2021), Sulewski (2021), Kim (2018), Górecki at al. (2020), Pakyari (2021). However, they may fail in some cases, such as calculating the EV of the order statistics for arbitrary sample size. This situation has motivated us to study the approximation expressions for the EVs of the order statistics of the SN distribution.

Therefore, in this study, we proposed two new approximations in addition to the Blom and Filliben approximations, and called them Proposed-1 and Proposed-2. The main aim of the present paper is to determine the best approximation method among them for the EV of the order statistics from the SN distribution. Therefore, we compared the performance of these approximation methods with exact values. We believe that if an accurate approximation formula is found, considerable improvements in statistical calculations can be achieved.

The remainder of this article is as follows: In section 2, Blom's approximations, Filliben's approximation, and our proposed approximations are given. In section 3, the exact means are compared with Blom's models, Filliben's model, simulation results, and our proposed models. Concluding remarks are presented at the end of the paper.

2. Approximation of $E\left(Z_{(i)}\right)$

In this section, some approximation models are discussed.

2.1. Blom's approximations

Blom's approximation is proposed by Blom (1958) to calculate the approximate ratio of the ith order statistic of a sample of size n from the SN distribution. This method is as given below:

$$E(Z_{(i)}) = \Phi^{-1}\left(\frac{i-\alpha}{n-2\alpha+1}\right).$$

(4)

Afterward, Blom presented as a table the value of α necessary to give the accurate value of $E\left(Z_{(i)}\right)$ for $i=1(1)\left[\frac{1}{2}n\right]$ when various sample sizes ranging from, $n=2$ to 20. Blom noticed that α forever lies in the interval $(0.33, 0.50)$. He proposed the use of $\alpha = 0.375$ as a comprise value.

Thus, the first approximation method (Blom-1) proposed by Blom for the expected values of the order statistics of the SN distribution can be written as follows:

$$M_{i1} = E\left(Z_{(i)}\right) \approx \Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right).$$

(5)

In estimating $E\left(Z_{(i)}\right)$ for $n \le 400$, to minimize the maximum error, the value of α should be chosen even smaller than 0.375, because the estimate of $E\left(Z_{(i)}\right)$ changes in α more sensitive for small values of n than for large values. The maximum error in estimating $E\left(Z_{(i)}\right)$ is minimized by choosing $\alpha = 0.363$. This gives a maximum error of 0.018, which is hardly satisfactory. Then, the substitution of $\alpha = 0.363$ into equation (4), Blom's second approximation (Blom-2) is found as:

$$M_{i2} = E\left(Z_{(i)}\right) \approx \Phi^{-1}\left(\frac{i-0.363}{n+0.2740}\right).$$

(6)

For more details about this method, see Harter (1961); and Blom (1958).

2.3 Filliben's approximation

Filliben (1975) suggested giving the following formula:

$$M_{i3} = E\left(Z_{(i)}\right) \approx \begin{cases} \Phi^{-1}\left(1-0.5^{1/n}\right), & i=1 \\ \Phi^{-1}\left((i-0.3175)/(n+0.365)\right), & i=2,3,\dots,n-1 \\ \Phi^{-1}\left(0.5^{1/n}\right), & i=n. \end{cases}$$

(7)

This approximation was used to approximate the median of the ith order statistic from the normal distribution for use in normal probability plots, see Pirouzi and Holmquist (2007).

2.4 Our approximations

In this study, we proposed two different approximation methods for the EVs of the order statistics from the SN distribution. The expected values of the first and last order statistics of the SN distribution are very easily computed. In the first of these methods, we use a similar procedure as in Pirouzi and Holmquist (2007). We thus minimize

$$Q(\alpha) = \sum_{i=2}^{n-1}\left(M_i - \Phi^{-1}\left(\frac{i-\alpha}{n-2\alpha+1}\right)\right)^2, \quad (8)$$

where, M_i represents the exact values given in Harter (1961). For the minimization of equation (8), we used the numerical algorithm "nlinfit" in Matlab. The optimal values of α for some specific sample sizes are given in Table 1. We suggested the use of $\alpha = 0.3923$ as an average value based on these sample sizes. So, the first approximation (proposed-1) is as follows:

$$\begin{cases} \Phi^{-1}\left(\frac{0.618}{n}-0.00008\right), & i=1 \\ \Phi^{-1}\left(\frac{i-0.3923}{n+0.2154}\right), & i=2,\ldots,n-1 \\ \Phi^{-1}\left(1.00008-\frac{0.618}{n}\right), & i=n \end{cases}$$

(9)

Table 1: Some values of α

n	5	10	20	35	50	75	100	130	150	200
α	0.3461	0.3618	0.3648	0.3883	0.4050	0.3950	0.4105	0.4127	0.4165	0.4198

To obtain the second approximation of the expected values of the order statistics from the SN distribution, we follow the following steps:

Step1: we take the sample size $n=\left[5(1)40,\ 45(5)100, 110(10)200,\ 220,\ 250,\ 300,\ 400\right]$.

Then, the values of α corresponding to each value of n are obtained by minimizing equation (8).

Step 2: we determine the appropriate method by plotting the α values against the sample size n . Then, we chose the logarithmic regression model given below:

$$y_i=\beta_0+\beta_1 \ln x_i+\varepsilon_i, \qquad (10)$$

where, y_i and x_i represent the α values and sample sizes n, respectively.

Step 3: The least squares estimators of the parameters β_0 and β_1 are obtained from equation (10). Thus, an approximate expression for α based on n is obtained as follows:

$$\alpha \approx 0.347+0.014\ln n \qquad (11)$$

By taking the approximate expression given in equation (11) instead of $\alpha=0.3923$, a second approximation (proposed-2) for the expected values of the order statistics from the SN distribution is given as follows:

$$M_{i5}=E\left(Z_{(i)}\right)\approx \begin{cases} \Phi^{-1}\left(\frac{0.618}{n}-0.00008\right), & i=1 \\ \Phi^{-1}\left(\frac{i-0.347-0.014\ln n}{n-0.028\ln n+0.3060}\right), & i=2,\ldots,n-1 \\ \Phi^{-1}\left(1.00008-\frac{0.618}{n}\right), & i=n \end{cases}$$

(12)

3. Numerical Experiments

In this section, the exact means (M_i) are compared with Blom's models, Filliben's model, simulation results, and our proposed models. The differences between exact values and the approximate values obtained by the different models are expressed as follows:

$\varepsilon_{ij}=M_i-M_{ij},\ i=1,2,\ldots,n,\ \ j=1,2,\ldots,6.$

where M_{i6} represents the mean of 1000000 replications of the order statistics from SN distribution. To evaluate the performances of the models, the plot of ε_{ij} against the exact values M_i, $i=1,2,\ldots,n,\ j=1,2,\ldots,6$ and the following three criteria are used.

i. $MSE=\frac{\sum_{i=1}^{n}\varepsilon_{ij}^2}{n}, \qquad j=1,2,\ldots,6.$

ii. $MAE = \frac{\sum_{i=1}^{n} \left| \varepsilon_{ij} \right|}{n}, \quad j = 1, 2, ..., 6.$

iii. $ME = \underset{1 \le i \le n}{\text{Max}} \left| \varepsilon_{ij} \right|, \quad j = 1, 2, ..., 6.$

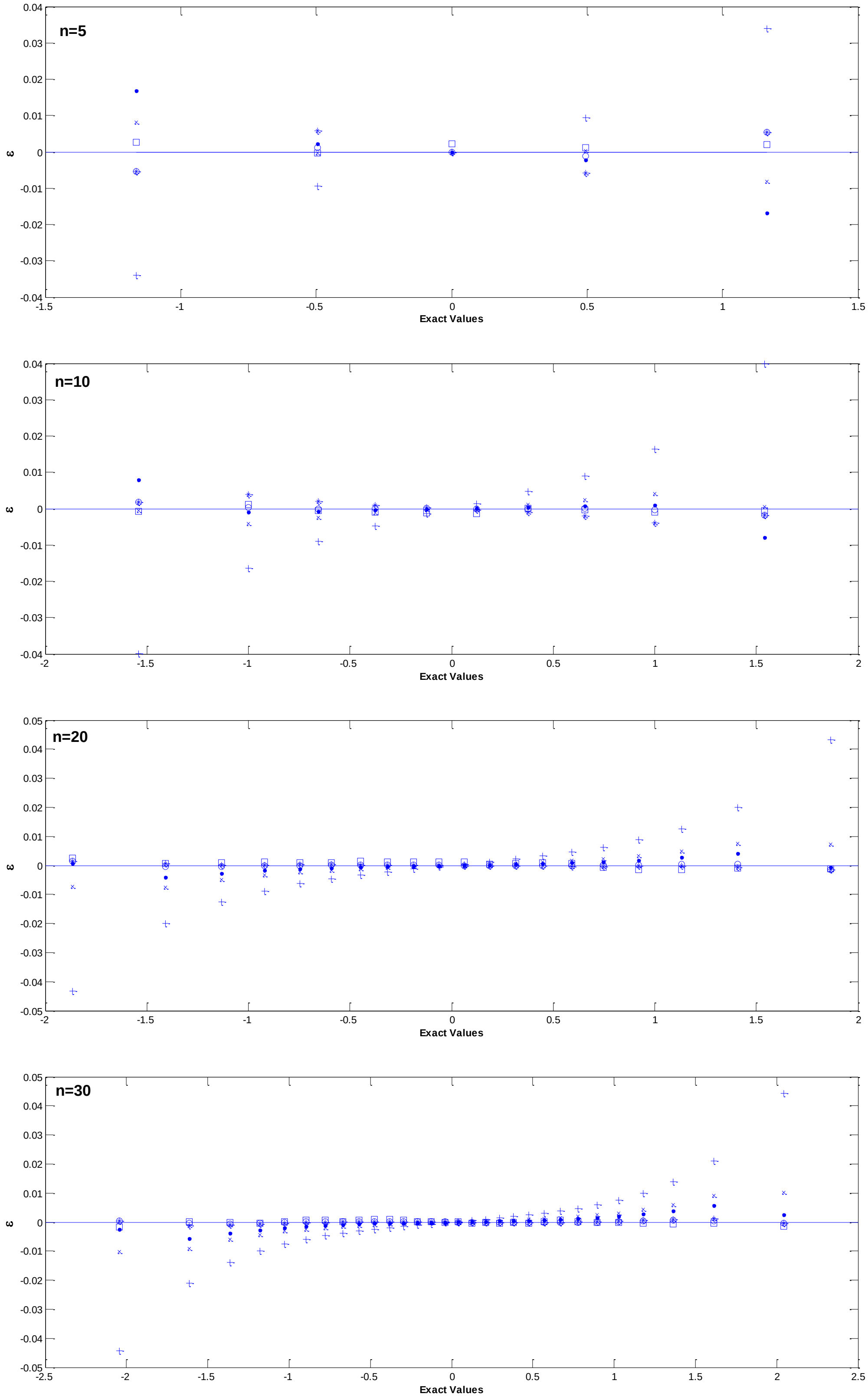
The quantities given in (i), (ii), and (iii) are called the mean square error (MSE), mean absolute error (MAE), and maximum error (ME).

The MSE, MAE, and ME values for all models are given in Table 2.

Table 2 The ME, MAE and RE values for the approximations

n		Blom1 (M_{i1})	Blom2 (M_{i2})	Filliben (M_{i3})	Proposed-1 (M_{i4})	Proposed-2 (M_{i5})	Simulation (M_{i6})
5	MSE	$1.15\text{x}10^{-4}$	$2.64\text{x}10^{-5}$	$4.97\text{x}10^{-4}$	$2.50\text{x}10^{-5}$	$1.23\text{x}10^{-5}$	$1.79\text{x}10^{-5}$
	MAE	0.0076	0.0034	0.0173	0.0045	0.0027	0.0037
	ME	0.0168	0.0081	0.0340	0.0058	0.0054	0.0064
10	MSE	$1.27\text{x}10^{-5}$	$5.01\text{x}10^{-6}$	$3.95\text{x}10^{-4}$	$4.67\text{x}10^{-6}$	$6.97\text{x}10^{-7}$	$1.37\text{x}10^{-6}$
	MAE	0.0020	0.0017	0.0143	0.0018	0.0005	0.0011
	ME	0.0079	0.0042	0.0400	0.0039	0.0017	0.0019
20	MSE	$3.22\text{x}10^{-6}$	$1.52\text{x}10^{-5}$	$2.59\text{x}10^{-4}$	$2.78\text{x}10^{-7}$	$2.08\text{x}10^{-7}$	$1.74\text{x}10^{-6}$
	MAE	0.0013	0.0029	0.0103	0.0004	0.0002	0.0011
	ME	0.0042	0.0075	0.0443	0.0014	0.0014	0.0038
30	MSE	$4.77\text{x}10^{-6}$	$1.74\text{x}10^{-5}$	$1.90\text{x}10^{-4}$	$1.76\text{x}10^{-7}$	$3.74\text{x}10^{-8}$	$3.77\text{x}10^{-7}$
	MAE	*0.0015*	0.0028	0.0081	0.0003	0.0002	0.0005
	ME	0.0058	0.0101	0.0444	0.0011	0.0008	0.0015
50	MSE	$5.99\text{x}10^{-6}$	$1.61\text{x}10^{-5}$	$1.61\text{x}10^{-5}$	$7.55\text{x}10^{-7}$	$6.99\text{x}10^{-8}$	$1.73\text{x}10^{-7}$
	MAE	0.0015	0.0023	0.0058	0.0005	0.0002	0.0003
	ME	0.0074	0.0129	0.0452	0.0029	0.0008	0.0010
100	MSE	$5.37\text{x}10^{-6}$	$1.15\text{x}10^{-5}$	$6.55\text{x}10^{-5}$	$1.12\text{x}10^{-6}$	$7.88\text{x}10^{-8}$	$3.20\text{x}10^{-7}$
	MAE	0.0011	0.0016	0.0036	0.0005	0.0002	0.0004
	ME	0.0090	0.0157	0.0456	0.0048	0.0013	0.0021
150	MSE	$4.55\text{x}10^{-6}$	$8.72\text{x}10^{-6}$	$4.42\text{x}10^{-5}$	$1.05\text{x}10^{-6}$	$4.68\text{x}10^{-8}$	$2.37\text{x}10^{-7}$
	MAE	0.0009	0.0013	0.0026	0.0004	0.0002	0.0004
	ME	0.0104	0.0168	0.0455	0.0056	0.0005	0.0009

 DOI: http://dx.doi.org/10.18187/pjsor.v20i2.4411



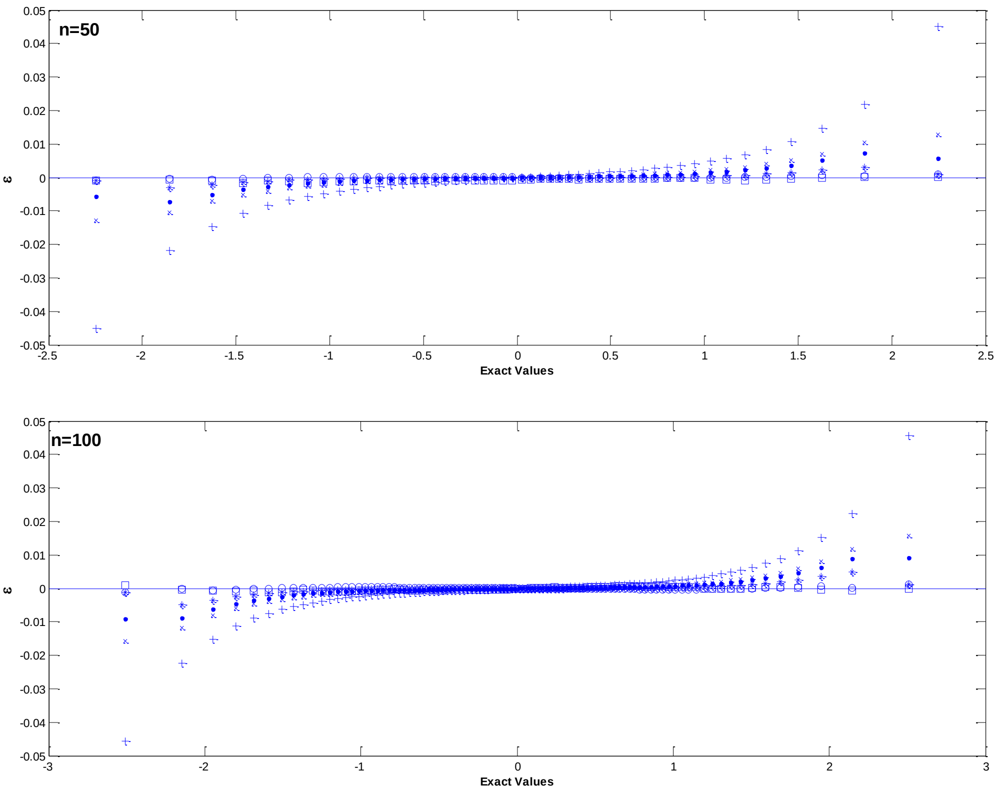


Fig.1. Differences between values M_i and ε_{ij} for different sample sizes.

Fig.1. plotting ε_{ij} against the exact values M_i, $i=1,2,...,n$, $j=1,2,...,6$. Here '.' symbolize Blom1 (M_{i1}) approximation; 'x' symbolize Blom2 (M_{i2}) approximation; '+' symbolize Filliben (M_{i3}) approximation; '*' symbolize proposed-1 (M_{i4}) approximation; 'd ' symbolize proposed-2 approximation (M_{i5}) ; 'W symbolize simulation (M_{i6}).

It is clear from Fig. 1 that the proposed-2 approximation has the smallest deviation among the other approximations and better fits the true values. This approximation is followed by the proposed-1 approximation and the simulation results. The Filliben approximation has the greatest deviation.

The following conclusions are obtained from Table 1.

According to the MSE, MAE, and ME criteria of the approximation methods:

- When the Blom-1 and Blom-2 approximations are compared with each other, the Blom-1 gives better results than the Blom-2 approximation, for $n \le 20$. However, the Blom-2 approximation gives better results, for $n > 20$.
- Similarly, when the Blom approximations and the proposed-1 approximation are compared, the proposed-1 approximation method works slightly better than the Blom approximations, for $n \le 20$. However, the proposed-1 approximation outperforms the Blom approximations, for $n > 20$.
- According to the results obtained here, the Filliben approximation does not perform well in calculating the expected value of the order statistics of the SN distribution.
- The values obtained by the proposed -2 approximation and the simulation study better fit the true values than the other approximations, for all sample sizes. Besides, when the simulation results and the proposed-2 approximation are compared with each other, it is seen that the proposed-2 approximation works better than the simulation study.

The results in Fig.1 are consistent with Table 2. In light of the aforementioned information, we recommend using the proposed-2 approximation for the EV of the order statistics from the SN distribution.

1. Conclusion

In this study, we deal with the problem of finding the expected values of the order statistics of the SN distribution. In this context, two approximate expressions for the expected values of the order statistics of the SN distribution are introduced. We also compare the previously given approximations, simulation results, and our proposed approximations with the exact values. The results show that the proposed-2 approximation we have proposed demonstrates better performance than the other approximations and simulation results.

References

1. Atangana, A., and Gómez-Aguilar, J. F. (2017). A new derivative with normal distribution kernel: Theory, methods and applications. Physica A: Statistical mechanics and its applications, 476, 1-14.
2. Blom, G. (1958). Statistical estimates and transformed beta-variables (Doctoral dissertation, Almqvist & Wiksell).
3. Bose, R. C., and Gupta, S. S. (1959). Moments of order statistics from a normal population. Biometrika, 46(3/4), 433-440.
4. Chen, C. C., and Tyler, C. W. (1999). Accurate approximation to the extreme order statistics of Gaussian samples. Communications in Statistics-Simulation and Computation, 28(1), 177-188.
5. Cho, H. Y. (2021). Normality Test of the Water Quality Monitoring Data in Harbour. Journal of Korean Society of Coastal and Ocean Engineers, 33(2), 53-64.
6. D'Agostino, R. B. (2017). Tests for the normal distribution. In Goodness-of-fit techniques, 367-420.
7. David, H. T. (1963). The sample mean among the extreme normal order statistics. The Annals of Mathematical Statistics, 34(1), 33-55.
8. Du, Y., Fan, B., & Wei, B. (2022). An improved exact sampling algorithm for the standard normal distribution. Computational Statistics, 37(2), 721-737.
9. Evans, M., Hastings, N., and Peacock, B. (2000). Statistical distributions (3rd ed.). New York: Wiley. Balakrishnan, N., & Nevzorov, V. B. (2003). A primer on statistical distributions. New Jersey: Wiley
10. Fard, M. N. P., & Holmquist, B. (2007). First moment approximations for order statistics from the extreme value distribution. Statistical Methodology, 4(2), 196-203.
11. Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. Technometrics, 17(1), 111-117.
12. Fisher, R. A., and Yates, F. (1943). Statistical tables for biological, agricultural and medical research. Oliver and Boyd Ltd, London.
13. Fowlie, A., Handley, W., and Su, L. (2020). Nested sampling cross-checks using order statistics. Monthly Notices of the Royal Astronomical Society, 497(4), 5256-5263.
14. Górecki, T., Horváth, L., & Kokoszka, P. (2020). Tests of normality of functional data. International Statistical Review, 88(3), 677-697.
15. Guan, J., Yuan, P., Hu, X., Qing, L., and Yao, X. (2019). Statistical analysis of concrete fracture using normal distribution pertinent to maximum aggregate size. Theoretical and Applied Fracture Mechanics, 101, 236-253.
16. Harter, H. L. (1961). Expected values of normal order statistics. Biometrika, 48(1/2), 151-165.
17. Harter, H. L., and Balakrishnan, N. (1996). CRC handbook of tables for the use of order statistics in estimation. CRC press.
18. Hastings Jr, C., Mosteller, F., Tukey, J. W., and Winsor, C. P. (1947). Low moments for small samples: a comparative study of order statistics. The Annals of Mathematical Statistics, 18(3), 413-426.
19. Headrick, T. C., and Pant, M. D. (2012). On the order statistics of standard normal-based power method distributions. International Scholarly Research Notices, 2012.
20. Hernandez, H. (2021). Testing for Normality: What is the Best Method. ForsChem Research Reports, 6, 2021-05.
21. Jäntschi, L. (2020). Detecting extreme values with order statistics in samples from continuous distributions. Mathematics, 8(2), 216.
22. Jones, H. L. (1948). Exact lower moments of order statistics in small samples from a normal distribution. The Annals of Mathematical Statistics, 270-273.
23. Kim, N. (2018). On the maximum likelihood estimation for a normal distribution under random censoring. Communications for Statistical Applications and Methods, 25(6), 647-658.

DOI: http://dx.doi.org/10.18187/pjsor.v20i2.4411

24. Lando, T., Arab, I., and Oliveira, P. E. (2021). Second-order stochastic comparisons of order statistics. Statistics, 55(3), 561-579.
25. Liu, Yonghui, et al. (2020). "Diagnostic analytics for an autoregressive model under the skew-normal distribution." Mathematics 8(5), 693.
26. Pahins, C. A., Ferreira, N., & Comba, J. L. (2019). Real-time exploration of large spatiotemporal datasets based on order statistics. IEEE Transactions on Visualization and Computer Graphics, 26(11), 3314-3326.
27. Pakyari, R. (2021). Goodness-of-fit testing based on Gini Index of spacings for progressively Type-II censored data. Communications in Statistics-Simulation and Computation, 1-10.
28. Parrish, R. S. (1992). Computing expected values of normal order statistics. Communications in Statistics-Simulation and computation, 21(1), 57-70.
29. Ruben, H. (1956). On the moments of the range and product moments of extreme order statistics in normal samples. Biometrika, 43(3/4), 458-460.
30. Saha, M., Dey, S., and Wang, L. (2021). Parametric inference of the loss based index C pm for normal distribution. Quality and Reliability Engineering International.
31. Shapiro, S. (2019). Goodness-of-fit tests. In The Exponential Distribution (pp. 205-220).
32. Sharma, R., Pachori, R. B., & Sircar, P. (2020). Seizures classification based on higher order statistics and deep neural network. Biomedical Signal Processing and Control, 59, 101921.
33. Sulewski, P. (2019). Modified Lilliefors goodness-of-fit test for normality. Communications in Statistics-Simulation and Computation, 1-21.
34. Sulewski, P. (2021). Two component modified Lilliefors test for normality. Equilibrium. Quarterly Journal of Economics and Economic Policy, 16(2), 429-455.
35. Triantafyllou, I. S. (2018). Nonparametric control charts based on order statistics: Some advances. Communications in Statistics-Simulation and Computation, 47(9), 2684-2702.
36. Zhao, P., and Balakrishnan, N. (2015). Comparisons of largest order statistics from multiple-outlier gamma models. Methodology and Computing in Applied Probability, 17(3), 617-645.