EUROPEAN ACADEMIC RESEARCH
Vol. VIII, Issue 10/ January 2021

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

Application of Edgeworth Expansions on Some Important Distributions

EDLIRA DONEFSKI
Faculty of Natural and Human Sciences
Department of Mathematics-Informatics-Physics
"Fan S. Noli" University, Korça, Albania
LORENC EKONOMI
Faculty of Natural and Human Sciences
Department of Mathematics-Informatics-Physics
"Fan S. Noli" University, Korça, Albania

Abstract

Edgeworth expansion is used to prove the asymptotic accuracy for the sample distributions and by Hall this expansion is used also to prove the asymptotic accuracy for their bootstrap distributions. Resampling methods have been in the statistical literature for over 50 years. However, it was Efron who proposed the bootstrap as it is today. The bootstrap method is a very general resampling procedure for estimating the distribution of statistics based on independent variables. This technique allows the evaluation of a sample distribution. In this paper we will see application of Edgeworth expansion for some distributions that have special importance in statistics.

We have used the statistical software R and its very useful packages that helped us in our simulations for Edgeworth expansions in both cases: when we use bootstrap and when we don't use it. At the end, we will give some results based on our simulations study.

Keywords: Edgeworth expansion, bootstrap, R programming

INTRODUCTION

Resampling methods have been in the statistical literature for over 50 years. However, it was Efron(1980) who proposed the bootstrap as it is

today. The bootstrap method is a very general resampling procedure for estimating the distribution of statistics based on independent observations. This technique allows the evaluation of a sample distribution.

Edgeworth expansion is an approximation technique that helps us to estimate the distribution function of either the standardized mean, the mean or the sum of independent identical distributed random variables. By Hall (1992) this expansion is used to prove the asymptotic accuracy not only for the sample distributions but also for their bootstrap distributions.

In this paper we will see Edgewoth expansion for three distributions that have special importance in statistics: normal distribution, exponential distribution and lognormal distribution.

We have used the statistical software R and and its very useful packages that helped us in our simulations for Edgeworth expansions in two cases: when we use bootstrap and when we don't use it. At the end, we will give some results based on our simulations study.

MATERIAL AND METHODS

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. By the Central Limit Theorem,

$$S_n = \frac{\sum_{i=1}^{n} X_i / n - \mu}{\sigma / \sqrt{n}}$$

is asymptotically normally distributed with zero mean and unit variance. We are interested in the asymptotic behavior of the difference between the normal distribution $\Phi(x)$ and the distribution function $F_n(x)$ of the S_n. By logarithmic expansion, characteristics functions, using the expansion series of the exponential function and doing the necessary transformations, (Basna, 2010; Butler, 2007; Stuart, 1994; Casella, 2002), we can get the formula

$$P(S_n \le x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + ... + n^{-j/2} p_j(x)\phi(x) + ... \quad (1)$$

called the Edgeworth expansion of the distribution of $P(S_n \le x)$. Here Φ denotes the standard normal distribution function, ϕ denotes the standard normal density. The polynomial p_j have degree of order $3j-1$ and is odd for even j. Hence

$$p_1(x) = -\frac{1}{6}k_3(x^2-1) \quad \text{and} \quad p_2(x) = -x\left\{\frac{1}{24}k_4(x^2-3) + \frac{1}{72}k_3^2(x^4-10x^2+15)\right\}.$$

The third cumulant k_3 refers to skewness, so the term of $n^{-1/2}$ order improves the basic normal approximation of the cumulative distribution function of S_n by performing skewness correction. k_4 refers to kurtosis for the term of order n^{-1} which improves the normal approximation further by adjusting for kurtosis.

Usually (1) exists as an asymptotic series, which means that if the series stop at a specific order the remainder is of smaller order than the last omitted term in the series. It means

$$P(S_n \le x) = \Phi(x) + n^{-1/2}p_1(x)\phi(x) + n^{-1}p_2(x)\phi(x) + \ldots + n^{-j/2}p_j(x)\phi(x) + o(n^{-j/2}) \qquad (2)$$

The restrictions on (2) are

$$E(|X|^{j+2}) < \infty \quad \text{and} \quad \lim_{|t|\to\infty} \sup|\psi(t)| < 1.$$

You can find the proof of this fact in (Hall,1992).

For a symmetric parent distribution $\kappa_3 = \kappa_4 = 0$. Thus the influence of $p_1(x)$ completely disappears and the contribution of the third term in (2) diminishes for large values of *n*. Hence, the resultant distribution is more close to a normal distribution. On the other hand, for a skewed distribution the impact of both second and third terms in (2) will start revealing and results will depart from a normal distribution.

The Concept of Bootstrap and Bootstrap Methodology

The concept of the bootstrap was first introduced in the seminal piece of Efron, 1979, and relies on the consideration of the discrete empirical distribution generated by a random sample of size *n* from an unknown distribution *F*. This empirical distribution assigns equal

probability to each sample item. We will write $\hat{F}_n$ for that distribution. By generating an independent, identically distributed (IID) random sequence (resample) from the distribution $\hat{F}_n$ or its appropriately smoothed version, we can arrive at new estimates of various parameters and nonparametric characteristics of the original distribution *F*. This idea is at the very root of the bootstrap methodology. Singh, 1981 made a further point that the bootstrap estimator of the sampling distribution of a given statistic may be more accurate than the traditional normal approximation. In fact, it turns out that for many commonly used statistics the bootstrap is asymptotically equivalent to the one-term Edgeworth expansion estimator, usually having the same convergence rate, which is faster than the normal approximation. The bootstrap methods can be applied to both parametric and non-parametric models, although most of the published research in the area is concerned with the non-parametric case since that is where the most immediate practical gains might be expected.

Edgeworth Expansion for Bootstrap version

If we recall (1)

$$G_n(x) = P(S_n \le x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + \dots \ , \qquad (3)$$

the bootstrap estimate of *G* admits an analogous expansion (Hall, 1992),

$$\hat{G}_n(x) = P(S_n^* \le x \mid X) = \Phi(x) + n^{-1/2} \hat{p}_1(x)\phi(x) + n^{-1} \hat{p}_2(x)\phi(x) + \dots \ , \quad (4)$$

where S_n^* is the bootstrap version of S_n, computed from a resample X^* instead of the sample *X*, and the polynomial $\hat{p}$ is obtained from *p* on replacing unknows, such as skewness, by bootstrap estimates. The distribution of S_n^* conditional on *X* is called the bootstrap distribution of S_n^*. The estimates in the coefficients of $\hat{p}$ are typically distant $O_p(n^{-1/2})$ from their respective values in *p,* and so $\hat{p} - p = O_p(n^{-1/2})$.

Therefore, subtracting (3) and (4), we conclude that

$$P(S_n^* \le x \mid X) - P(S_n \le x) = O_p(n^{-1}).$$

That is, the bootstrap approximation to G is in error by only n^{-1}. This is a substantial improvement on the Normal approximation, $G \approx \Phi$, which by (3) is in error by $n^{-1/2}$.

R programming and Simulations of Edgeworth Expansions for three important distributions

The bootstrap and related resampling methods are statistical techniques which can be used in place of standard approximations for statistical inference. The basic methods are very easily implemented but for the methods to gain widespread acceptance among users it is necessary that they be implemented in standard statistical packages. R is an integrated suite of an object - oriented programming language and software facilities for data manipulation, calculation, and graphical display. Over the last decade, R has become the statistical environment of choice for academics, and probably is now the most used such software system in the world. The number of specialized packages available in R has increased exponentially, and continues to do so. Perhaps the best thing about R is this: It is completely free to use. The package that we have used here for our simulations is "**EW**" package, (Shao 2003). This package helps us to generate a polynomial function $p_j(x)$ which approximate the error term of order of $o(n^{-j/2})$ in the difference between the normal distribution $\Phi(x)$ and the distribution function $F_n(x)$ of the S_n. When we applicate this package it plots a graph showing the asymptotic manner of polynomial's order. In our simulations we have used the first order of the polynom.

In this paper we have done simulations for these three important distributions: normal distribution, exponential distribution and lognormal distribution.

Let remind shortly the forms of these distributions:

Normal Distribution

The normal distribution is the most important distribution. It describes well the distribution of the random variables that arise in practice, such as the heights or weights of people, the total annual sales of a firm, exam scores et. Also, it is important for the central limit theorem, the approximation of other distributions such as the binomial,etc.

We say that a random variable X follows the normal distribution if the probability density function of X is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty<x<+\infty$$

This is a bell-shaped curve. We write X~ $N(\mu,\sigma)$ and we read: X follows the normal distribution (or X is normally distributed) with mean μ, and standard deviation σ. The normal distribution can be described completely by the two parameters μ and σ. As always, the mean is the center of the distribution and the standard deviation is the measure of the variation around the mean. The special case when μ=0 and σ=1 is known as the standard normal distribution.

Exponential Distribution

The continuous random variable X has exponential distribution if its density has the form

$$f(x)=\begin{cases}\lambda\exp(-\lambda x), & x\geq 0\\ 0, & x<0\end{cases} \text{ for } \lambda>0$$

and its distribution function $F(x)=\begin{cases}1-\exp(-\lambda x), & x\geq 0\\ 0, & x<0\end{cases}$.

Lognormal Distribution

The lognormal distribution is a commonly used distribution in Economics and Finance fields. It has two characteristics: first, the random variable must be non-negative; second, the shape of the density function has a long tail to the right, or is called positively skewed. Let y be a lognormal random variable, i.e., $\ln(y)\sim N(\mu,\sigma^2)$, where $\mu=E(\ln(y))$, and $\sigma^2=Var(\ln(y))$.

Assume that a random sample $y_1,...,y_m$ of size m is available from this distribution.

Let $\hat{\mu} = \frac{\sum_{i=1}^{m}(\ln(y_i))}{m}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{m}(\ln(y_i) - \hat{\mu})^2}{m-1}$.

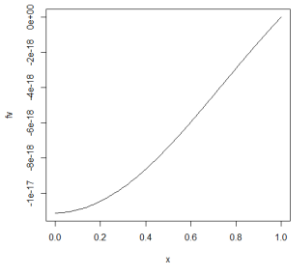
Note that $\hat{\mu}$ and $\hat{\sigma}^2$ are jointly sufficient statistics for μ and σ^2, and they are consistent estimators for large m.

After we have recalled these distribution we will present the results of our simulations in two cases: when we use bootstrap and when we don't use it.

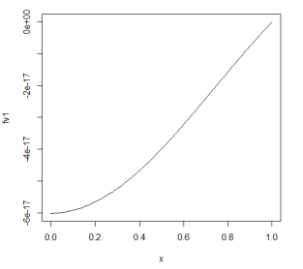
First we have applied the polynomial manner in both cases and after that we have presented the asymptotic manner of polynomial's order in both cases, too. We have done 100 bootstrap replicates and the graphs was all very similar and overlap with each others and so we have presented here only one replication to reduce the complexity of the figures. We conclude that for the three distributions the performance of the bootstrap version and non bootstrap version is very similar. These conclusions are illustrated clearly in the graphics below:

Normal Distribution

polynomial manner

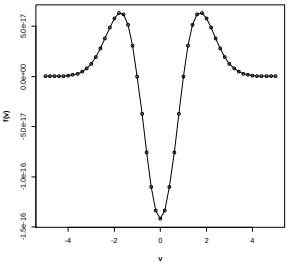


nonbootstrap case

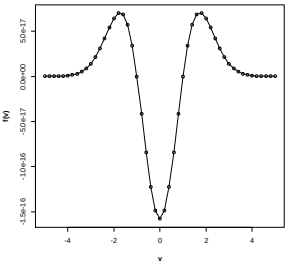


bootstrap case

asymptotic manner of polynomial's order



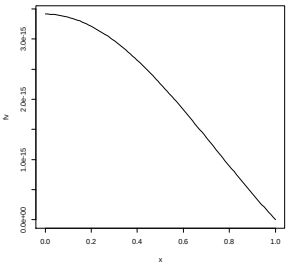
nonbootstrap case



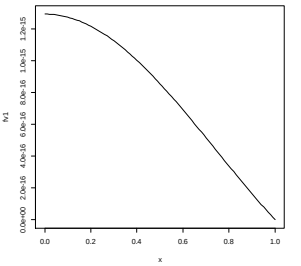
bootstrap case

Exponential Distribution

polynomial manner

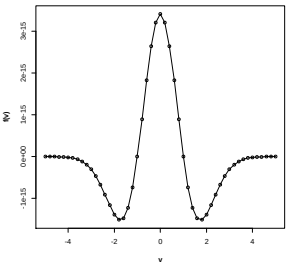


nonbootstrap case

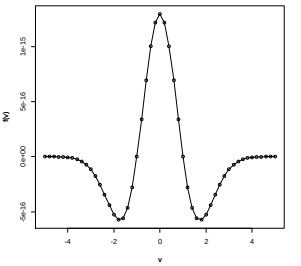


bootstrap case

asymptotic manner of polynomial's order



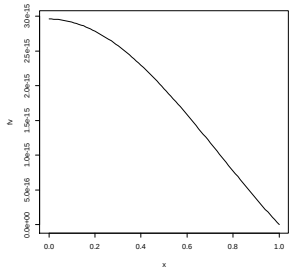
nonbootstrap case



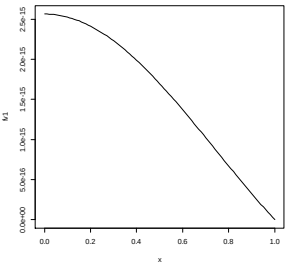
bootstrap case

Lognormal Distribution

polynomial manner

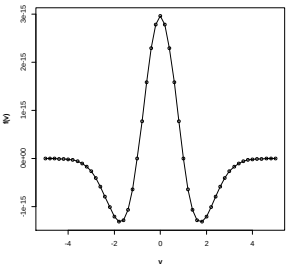


nonbootstrap case

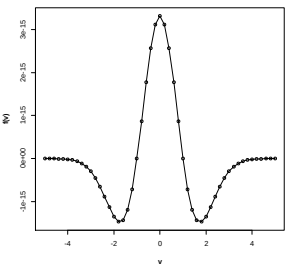


bootstrap case

asymptotic manner of polynomial's order



nonbootstrap case



bootstrap case

CONCLUSIONS

In this paper we have treated Edgeworth expansion for three distributions that have special importance in statistics: normal distribution, exponential distribution and lognormal distribution.
We have used the statistical programming R and a very useful package, "EW" package and the specific commands included in this package that helped us in our simulations for Edgeworth expansions in two cases: when we use bootstrap and when we don't use it.

First we have applied the polynomial manner in both cases and after that we have presented the asymptotic manner of polynomial's order in both cases, too. We conclude that for the three distributions the performance of the polynomial manner and the asymptotic manner of the polynomial's order for bootstrap version and non bootstrap version is very similar.

LITERATURE

1. Basna R. (2010): *Edgeworth Expansion and Saddlepoint Approximation for Descrete Data with Applications in Chance Games.* Linnéuniversitetet.
2. Butler R.W. (2007): *Saddlepoint Approximation with Applications.* Cambridge University Press, New York, USA, 145.
3. Casella G. and Berger R. L. (2002): *Statistical Inference, second edition.* The Wadsworth Group, USA. 83-87.
4. Efron B. (1980): *The Jackknife, the Bootstrap, and Other Resampling Plans.* National Science Foundation Grant, California,105-132.
5. Hall P. (1992): *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York, USA.
6. Shao J. (2003): *Mathematical Statistics*, revised ed, Springer: P70-76, Sec1.5.6
7. Stuart A. and Ord K. (1994): *Distribution theory, vol.1.* Oxford University press Inc., New York, USA, 74-162.