Comparing Edgeworth Expansion and Saddlepoint Approximation in Assessing the Asymptotic Normality Behavior of A Non-Parametric Estimator for Finite Population Total

Jacob Oketch Okungu, George Otieno Orwa, and Romanus Odhiambo Otieno

Abstract — Sample surveys concern themselves with drawing inferences about the population based on sample statistics. We assess the asymptotic normality behavior of a proposed nonparametric estimator for finite a population total based on Edgeworth expansion and Saddlepoint approximation. Three properties; unbiasedness, efficiency and coverage probability of the proposed estimators are compared. Based on the background of the two techniques, we focus on confidence interval and coverage probabilities. Simulations on three theoretical data variables in R, revealed that Saddlepoint approximation performed better than Edgeworth expansion. Saddlepoint approximation resulted into a smaller MSE, tighter confidence interval length and higher coverage probability compared to Edgeworth Expansion. The two techniques should be improved in estimation of parameters in other sampling schemes like cluster sampling.

Keywords — Edgeworth expansion, Nonparametric estimator, auxiliary variables Saddlepoint approximation.

I. Introduction

In the estimation of the finite population total, misspecification of the model can lead to serious errors in an inference especially with regard to the non-sampled part of the population. In the recent past, efforts have been made to explore alternative ways to attenuate the errors. These include the use of nonparametric regression in evolving robust estimators in finite population sampling [1].

Non-parametric estimators have been found to be robust and more precise than their parametric counterparts. It is known, for instance, that a linear regression estimate will produce a large error for every sample size if the true underlying regression function is not linear and cannot be well approximated by linear functions [2].

According to [3], the non-parametric regression estimator of a finite population total is a potent rival to familiar design-based estimators. It has the quality of automaticity associated with design-based estimators, but can better reflect the actual structure of the data, yielding greater efficiency. It can be costly in computer power, and will probably not do as well as a parametric-model based estimator, when the modelling process is done carefully.

Nonparametric regression has its origin in exploration of data. Let $S = \{x_i, y_i\}$, $i = 1, 2, ..., n$ be a data set, then a cloud of points is suggested. It may basically mean drawing a line in the $X - Y$ plane through the cloud of points showing the essential characteristics of the nature of relationship between the variables Y and X. In survey sampling, there are four estimation approaches that can be used in statistical investigations; the design-based approach, model-based approach, model-assisted approach and randomization-assisted approach [4].

II. Background to the Study

The model-based approach has bridged the gap between finite population problems and the rest of statistics. Before the model-based approach, finite population sampling was an eccentric realm where many of the basic concepts and tools of statistics were curiously inapplicable. Statisticians skilled in designing experiments and in applying linear models to make inferences from experimental and observational data found that finite population problems were apparently beyond the scope of their techniques [5]

In the model-based approach, the distribution is a structure that is defined by the population itself and is unknown but can be modelled. The values of the variable Y are believed to be random variables; $Y_1, Y_2, ..., Y_N$ generated by some model. The

Published on January 31, 2023

J. O. Okungu, Meru University of Science and Technology, Kenya.
(corresponding e-mail: jokungu@must.ac.ke)
G. O. Orwa, Jomo Kenyatta University of Agriculture and Technology, Kenya.
(e-mail: gorwa@jkuat.ac.ke)
R. O. Odhiambo, Meru University of Science and Technology, Kenya.
(e-mail: rodhiambo@must.ac.ke).

DOI: http://dx.doi.org/10.24018/ejmath.2023.4.1.167

actual observations for the finite population $y_1, y_2, ..., y_N$ are one realization of the random variables.

The information obtained from the sample is used to predict the information of the non-sampled observations. An appropriate model-based estimator of the finite population total is of the form

$$\hat{T} = \sum_{i=S}^{N} Y_i + \sum_{i \neq S}^{N} \hat{m}(X_i) \tag{1}$$

Where $\hat{m}(X_i) = \sum_i w_i(x) Y_i$.

Reference [6] considered a related nonparametric model-assisted regression estimator, replacing local polynomial smoothing with penalized splines. They extended the local polynomial nonparametric regression estimation to two-stage sampling. In their work, simulation results indicate that the nonparametric estimator dominates standard parametric estimators when the model regression function is incorrectly specified, while being nearly as efficient when the parametric specification is correct.

Reference [1] considered the application of nonparametric regression to the estimation of finite population error variance for a given sample drawn from the population. The error variance obtained by [6] was a function of $\sigma^2(x_j)$ that are unknown. By considering the squared residual $e_j^{\,2} = (y_i - \hat{m}(x_i))^2$ and using some mild assumptions, their study showed that $E(e_j^{\,2} / X_j = x_j) = \sigma^2(x_j) + o(n^{-1})$ implying that $e_j^{\,2}$ is an asymptotic unbiased estimator of $\sigma^2(x_j)$. They obtained an improved estimator of $\sigma^2(x_j)$ by smoothing $e_j^{\,2}$ for $j \in S$ being sample points $(x_j, y_j)'$ close to $(x_i{}', y_i{}')$. Reference [7] used local polynomial regression in the estimation of finite population totals. In his research, he considered the equation $Y_i = m(X_i) + \sigma^2(x_i)e_i$ and applied the technique of using a strip of data around the co-variate in order to fit a line through the set of data (x_j, y_j). The estimator yielded better results in estimating the finite population total. Further, the estimator was found to be asymptotically unbiased, consistent and normally distributed when certain conditions were satisfied.

III. Methods

A. Review of Edgeworth Expansion

Let a random variable Y be the variable of interest and that X is an auxiliary variable associated with Y assumed to be known for all the observable population units. Let *T* be the population total such that $T = \sum_{i=1}^{N} Y_i$. All the sampled units are observed and therefore there exist no error. The task is therefore to determine the characteristics of the non-sampled units $\sum_{i \neq S} Y_j$. T is thus given by the total sampled units and the non-sampled units i.e.

$$T = \sum_{i=S} Y_i + \sum_{i \neq S} Y_j \tag{2}$$

The information obtained from the sample is used to predict the information of the non-sampled observations. In this study, it is assumed that Y is function of X, hence a model of the form

$$Y_i = m(X_i) + e_i \tag{3}$$

It is further assumed that e_i are the error terms which are normally identically and independently distributed wit $E(e_i) = 0$ and $\sigma^2(e_i) = \sigma^2$. Let $X_1, X_2, ...X_N$ be independent and identically distributed (iid) random variables with mean 0 and variance σ^2. Define

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \tag{4}$$

then the characteristics function of S_n is given by

$$\psi_{S_n}(t) = E\left[exp\left\{ \left(\frac{it}{\sqrt{n}} \right) \sum_i X_i \right\} \right] = \left[\psi_X \left(\frac{t}{\sqrt{n}} \right) \right]^n \tag{5}$$

Using the Taylor's series expansion, (4) becomes

$$\psi_X\left(\frac{t}{\sqrt{n}}\right) = E\left\{1 + \frac{itX}{\sqrt{n}} + \frac{(it)^2 X^2}{2n} + \frac{(it)^3 X^3}{6n\sqrt{n}} + \frac{(it)^4 X^4}{24n^2}\right\} + O\left(\frac{1}{n^2}\right) \tag{6}$$

Setting $E(X^4) = \tau$ to take care of kurtosis and $E(X^3) = \gamma$ to take care of skewness, (6) simplifies to

$$\left[\psi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[\left(1 - \frac{t^2}{2n}\right)^n + \left(1 - \frac{t^2}{2n}\right)^{n-1}\left(\frac{(it)^3 \gamma}{6\sqrt{n}} + \frac{(it)^4 X^4}{24n}\right) + \left(1 - \frac{t^2}{2n}\right)^{n-2} \frac{(n-1)(it)^6 \gamma^2}{72n^2}\right] + O\left(\frac{1}{n}\right) \tag{7}$$

Using binomial expansion and from (4) and (6),

$$\psi_{S_n}(t) = e^{-\frac{t^2}{2}}\left[1 + \frac{(it)^3 \gamma}{6\sqrt{n}} + \frac{(it)^4 (\tau - 3)}{24n} + \frac{(it)^6 \gamma^2}{72n^2}\right] + o\left(\frac{1}{n}\right) \tag{8}$$

By defining inversion of a function and incorporating the characteristic function, a function G(x) result into,

$$G(x) = \Phi(x) + \phi(x)\left(\frac{\gamma(x^2 - 1)}{6\sqrt{n}} + \frac{(\tau - 3)(x^3 - 3)}{24n} + \frac{\gamma^2(x^5 - 10x^3 - 15x)}{72n}\right) \tag{9}$$

where $\Phi(x)$ is a standard normal distribution, $\phi(x)$ is a normal distribution and $\left(\frac{\gamma(x^2 - 1)}{6\sqrt{n}} + \frac{(\tau - 3)(x^3 - 3)}{24n} + \frac{\gamma^2(x^5 - 10x^3 - 15x)}{72n}\right)$ is the Edgeworth expansion.

B. Review of Saddlepoint Approximation

Saddle point approximation is remarkable because it usually provides probability approximations whose accuracy is much greater than the current supporting theory would suggest. Saddle point methods are also useful in avoiding much of the simulation requisite when implementing another modern statistical tool, the bootstrap. The most fundamental Saddle point approximation was first introduced by [8] and is essentially a formula for approximating a density mass function from its associated MGF. Assume that the functions are as regular as needed. In other words, when a derivative or an integral is assumed to exist then, the saddle point approximation arises from a natural sequence of approximations that become progressively more local.

Suppose a continuous random variable X has density $f(x)$ defined for all real values of X, then the MGF of density $f(x)$ is defined as the expectation of e^{SX} that is,

$$M(s) = E\left(e^{SX}\right) = \int_{-\infty}^{\infty} e^{SX} f(x)dx \tag{10}$$

over the values of S for which the integral converges. With real values of S, the convergence is always assured at s = 0: In addition, it is presumed that the M(S) converges over an open neighborhood of S designated as say (a, b). Consequently, the CGF of the function is defined as

$$K(s) = ln\{M(s)\} \tag{11}$$

For a continuous random variable X with CGF K and unknown density f, the saddle point density approximation to $f(x)$ is given as

$$\hat{f}(x)=\frac{1}{\sqrt{2\pi K''(\hat{s})}}exp\{K(\hat{s})-\hat{s}x\} \tag{12}$$

Where K'(s̑) is the saddle point equation and s ̂is the saddle point associated with the value x [9]

To approximate the density of the total population total $N\overline{x}$ using saddle point approximation, consider finding the density of $X_i, i=1,2,...,n$ which are iid with CGF K. In this approximation, the Saddle point density is the leading term of the asymptotic expansion as n→∞ of the function, f that is

$$f(N\overline{x})=\hat{f}(N\overline{x})\{1+o(n^{-1})\} \tag{13}$$

where $o(n^{-1})$is the relative error of the asymptotic order indicated such that;

$$f(N\overline{x})=\sqrt{\frac{n}{2\pi K''(\hat{s})}}exp\{nK(\hat{s})-n\hat{s}N\overline{x}\} \qquad \text{(14) [10].}$$

C. The Proposed Estimators

Let T be the population total, defined as the sum of the values of all the population measurements and let the random variable Y be the variable of interest and that X is an auxiliary variable associated with Y assumed to be known for all the observable population units. All the sampled units are observed and the task therefore is to estimate the non-sampled part of the population. The non-sampled part is estimated using the Edgeworth Expansion.

Let S be the sample from the population of N units as in (2). Consider the model $Y_i=m(X_i)+e_i$ where m is an unknown smooth function that depends on the sampled data and is estimated by $\hat{m}(X)$ for the non-sampled data points.

The nonparametric estimator of the finite population total is proposed,

$$\hat{T}_{npe}=\sum_{i\in S}Y_i+\sum_{j\notin S}\hat{m}_e(X_j) \tag{15}$$

Where $\hat{m}_e(X_i)=\frac{1}{n}\sum_{i=1}^{n}S_n\left(\frac{x_0-X_i}{h_n}\right)$ from (4) and

$$\hat{T}_{nps}=\sum_{i\in S}Y_i+\sum_{j\notin S}\hat{m}_s(X_j) \tag{16}$$

Where $\hat{m}_s(X_i)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(\frac{x_0-X_i}{h_n}\right)$ from (12) based on Edgeworth expansion and Saddlepoint approximation respectively.

IV. Results and Discussion

Population of size 1,500 was simulated from three data variables; linear, quadratic and exponential. The linear, quadratic and exponential data variables which have the relations $Y_i=1+2(x_i-0.5)+e_i$, $Y_i=1+2(x_i-0.5)^2+e_i$ and $Y_i=exp(-8x_i)+e_i$ respectively.

The auxiliary variable X was assumed to be uniformly distributed in the interval [0,1] and e_i is assumed to be a standard normal variable defined as $e_i \sim N(0,1)$.

A simple random sample of size 300 was selected randomly from the simulated population index-wise, and replicated 1500 times giving rise to 1500 simple random samples as in [11] and [12]. The proposed estimator was therefore compared to the nonparametric regression estimator due to [5], the design-based Horvitz-Thompson estimator and the Ratio estimator using the amount of bias, MSE and the coverage probabilities.

A. Unconditional Properties

1) Relative Bias of the Estimators

The relative bias of the estimator was obtained using $\left(\left[\frac{\sum_{i=1}^{1500}\hat{T}_i}{1500}\right]-T\right)/T$ where T is the actual population total and $\hat{T}_i$ is the estimator of the population total from the i^{th} sample, for $i=1,2,...,1500$.

TABLE I: Relative Biases

Model	$\hat{T}_{nps}$	$\hat{T}_{npe}$
Linear	-14.566	21.402
Quadratic	20.071	-10.420
Exponential	19.315	-17.352

From Table I, some of the values of the average relative biases are either negative or positive which shows either underestimation or overestimation respectively. For the linear function, the estimator based on Saddlepoint approximation performed better than Edgeworth expansion. For the quadratic and exponential distributions, the Edgeworth expansion performed better than the Saddlepoint approximation.

2) Mean Squared Error (MSE) of the Estimators

The measures for the MSEs were computed for the three data sets, $MSE=\frac{\sum_{i=1}^{1500}(\hat{T}_i-T)^2}{1500}$ and then compared.

TABLE II: Mean Squared Errors

Model	$\hat{T}_{nps}$	$\hat{T}_{npe}$
Linear	0.1318	0.1232
Quadratic	0.1618	0.1644
Exponential	0.4084	0.4010

From Table II, for the linear model, the estimator based on Edgeworth expansion performed better than Saddlepoint approximation. Similarly, Edgeworth expansion performed better in quadratic and exponential distributions.

3) The 95% Confidence Interval Length

The uncertainty in using point estimate is addressed by means of confidence intervals. Confidence intervals provide us with a range of values for the unknown population along with the precision of the method.

The standard error necessitates the construction of the confidence interval. These give the probability to which the range of estimator covers the estimator of the parameter. A 95% confidence interval was therefore constructed such that

$$T=\hat{T}_i\pm Z_{\frac{\alpha}{2}}S.E\left(\hat{T}_i\right) \tag{17}$$

For the extent of coverage of the estimator, the coverage probability was explored more explicitly by approximating

$$P\left(\frac{\hat{T}_i-T}{\sqrt{\text{var}(\hat{T}_i)}}\le 0.95\right)=\Phi(x)+\bigcirc\left(\frac{1}{n}\right)n^{-1} \tag{18}$$

where Φ(x) is the distribution of the estimator which is clearly a function of the variable characteristics and follows the standard normal distribution and O is an order function of the sample size n which is of order n^{-1}.

The empirical results were tabulated in Table III.

TABLE III: 95% Confidence Interval Lengths

Model	$\hat{T}_{nps}$	$\hat{T}_{npe}$
Linear	12.0128	12.953
Quadratic	12.9852	14.498
Exponential	36.2789	36.002

From Table III, for the linear and quadratic functions, the estimator based on Saddlepoint approximation outperformed the proposed estimator based on Edgeworth expansion. For the exponential distribution, Edgeworth expansion performed better than the Saddlepoint approximation method.

4) Coverage probabilities of the Estimator

The coverage probabilities of the proposed estimator were computed using the nominal probabilities; 0.01, 0.05 and 0.10 for the 99%, 95% and 90% confidence levels respectively.

Apart from the 0.01 nominal probability in quadratic and exponential functions and 0.10 nominal probability in the exponential function, the Saddlepoint approximation outperforms the Edgeworth expansion in the estimation. Saddlepoint approximation therefore showed higher coverage rates.

B. Conditional Properties

1) Conditional Biases

Since the estimation is model-based, the 1,500 simple random samples were grouped into groups of 50 so that there were 30 groups. For each group $\bar{\bar{x}} = \frac{1}{30}\sum_{i=1}^{50}\bar{x}_i$ was computed and consequently $\bar{\bar{T}}_{npe} = \frac{1}{30}\sum_{i=1}^{50}\hat{T}_{npe.i}$ and $\bar{\bar{T}}_{nps} = \frac{1}{30}\sum_{i=1}^{50}\hat{T}_{nps.i}$ were also computed. The conditional bias for each group was computed as $\bar{\bar{T}}_{npe} - \bar{Y}$ and $\bar{\bar{T}}_{nps} - \bar{Y}$ where $\bar{Y}$ is the population mean for the survey measurements and $\bar{x}_i$ is the sample mean for the auxiliary variables.

The Fig. 1, 2 and 3 illustrate the behavior of the conditional bias for each estimator when the three mean functions were used. The Fig. 1 shows the conditional bias when linear mean functions was used, Fig. 2 shows the conditional bias when a quadratic mean function was used and Fig. 3 shows the conditional bias when an exponential mean function was used.

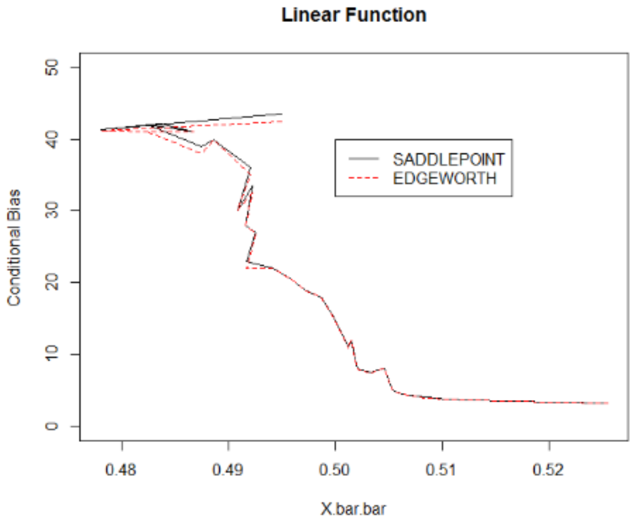

Fig. 1. Conditional biases for the linear function.

From Fig. 1, the proposed estimator due to Edgeworth expansion averagely performs better than the proposed estimator by Saddlepoint approximation.

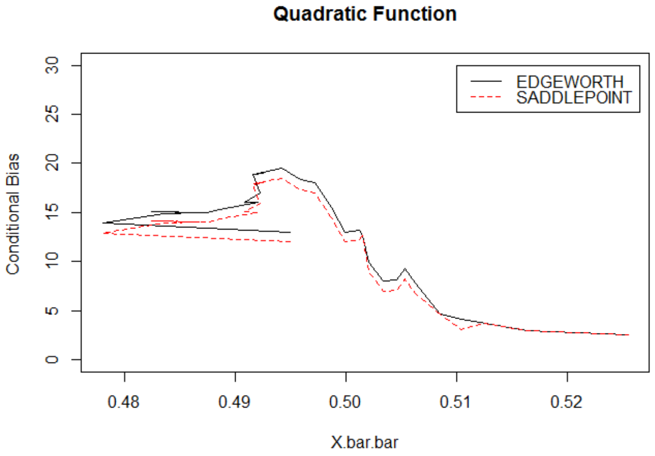

Fig. 2. Conditional biases for Quadratic Function.

From Fig. 2, for the quadratic model used, the proposed estimator by Saddlepoint approximation gives better estimates of the population total compared to the proposed estimator by Edgeworth expansion. This may be attributed to the fact that Saddlepoint approximation has smaller bias compared to Edgeworth expansion.

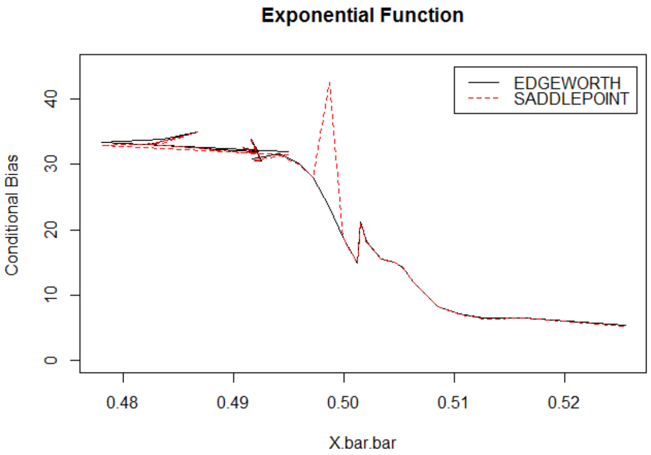

Fig. 3. Conditional Biases for Exponential Function.

From Fig. 3, the proposed estimator due to Saddlepoint approximation gave better estimates of the population total compared to those realized using the Edgeworth expansion.

2) Conditional MSEs

Just like the biases, conditional MSEs were determined in order to establish the robustness of the proposed estimator compared to the designed based, the ratio and the non-parametric Nadaraya-Watson (Dorfman's) estimators.

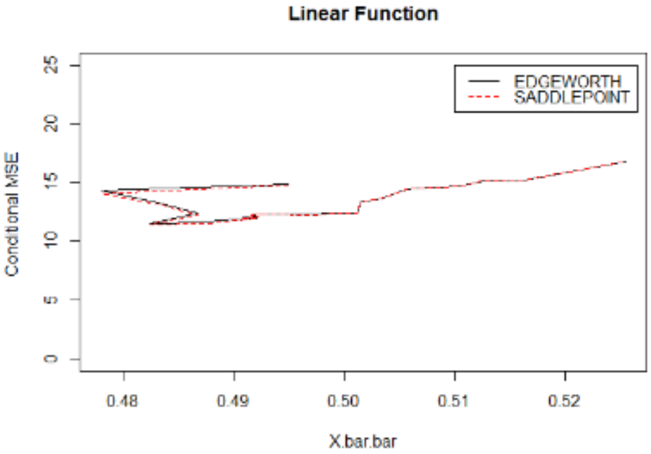

Fig. 4. Conditional MSE for linear model.

From Fig. 4, the estimator based on Edgeworth expansion performs almost the same as Saddlepoint approximation with the Saddlepoint approximation giving smaller MSE for lower sample means.

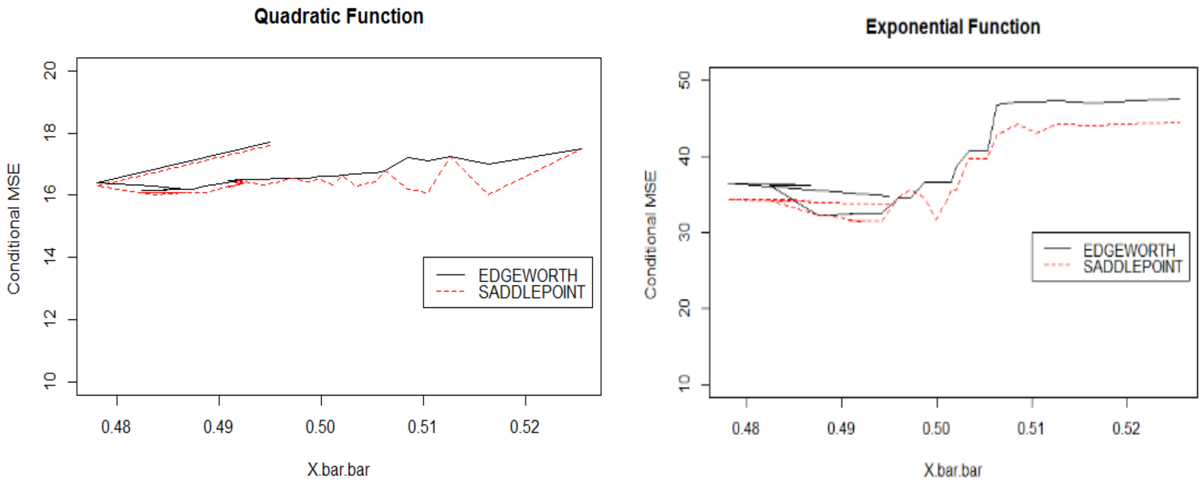

Fig. 5 and 6. Conditional MSE for Quadratic and Exponential models respectively.

From Fig. 5 and 6, the proposed estimator based on Saddlepoint approximation outperformed the proposed estimator based on Edgeworth expansion because in both cases, Saddlepoint approximation has smaller MSE.

3) Conditional coverage properties

Based on the conditional confidence intervals, the coverage probabilities were computed for the 30 samples. The coverage probability was based on the number of observations falling within the confidence interval compared to the total number of observations. From [11] and [12], the proposed estimator based on Saddlepoint approximation outperformed the one due to Edgeworth expansion in all the three theoretical data variables.

V. Conclusion

The proposed estimator based on Saddlepoint approximation performed better than Edgeworth expansion. This is probably due to the fact that Saddlepoint approximation is a compound of the Edgeworth expansion and therefore the average accuracy and precision in its estimation of the finite population total.

Acknowledgment

Much appreciation goes to all those who in a way or the other guided us throughout the duration of this study. We are greatly indebted to the staff of the department of Statistics and Actuarial Science at the Jomo Kenyatta University of Science and Technology and staff in the department of Mathematics at the Meru University of Science and Technology.

Funding

We are greatly indebted to National Research Fund (NRF) for funding the research.

Conflict of Interest

All the authors declare that they do not have any conflict of interest.

References

[1] Odhiambo R, Mwalili S. Nonparametric regression for finite population estimation. *East African Journal of Statistics*. 2000; (Part 2): 107-118.

[2] Laszlo GAKM, Walk H. A Distribution Free-Theory of Nonparametric Regression. Springer-Verlag, New York, 2002.

[3] Dorfman AH, Hall P. Estimators of the finite population distribution function using nonparametric regression. *Journal of the American Statistical Association*. 1992; 21: 1452-1475.

[4] Chambers JM. On Methods of Asymptotic Approximation for Multivariate Distributions. *Biometrica.* 1967.

[5] Dorfman AH. Nonparametric regression for estimating totals in finite population. *Journal of the American Statistical Association.* 1992; 4: 622-625.

[6] Breidt F, Opsomer P. Model-Assisted estimator for Complex Surveys using Penalized Splines. *Biometrica.* 2005; 92(4).

[7] Ombui T. Robust Estimation of Finite Population Total Using Local Polynomial Regression. PhD Thesis, Jomo Kenyatta University of Agriculture and Technology, 2008.

[8] Daniels HE. Saddlepoint approximations. *Annals of Mathematical Statistics*. 1987; 25: 631-650.

[9] Lugannani R, Rice S. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability.* 1980; 12: 475-490.

[10] Easton GS. General Saddlepoint approximation with applications to statistics. *Journal of the American Statistical Association*. 2008; 81: 420-430.

[11] Okungu JO, Orwa GO, Odhiambo RO. Nonparametric Estimator of a Finite Population Total Based on Saddle Point Approximation. *International Journal of Statistics and Applications*. 2020; 10(3): 60-67.

[12] Okungu JO, Orwa GO, Odhiambo RO. Nonparametric Estimator of a Finite Population Total Based on Edgeworth Expansion. *Science Journal of Applied Mathematics and Statistics*. 2020; 8(2): 35-41.