

ESTIMATION OF FINITE POPULATION TOTAL USING BIRTH AND DEATH PROCESS

Lamin Kabareh^{1} Thomas Mageto²*

¹*Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI)*

²*Jomo Kenyatta University of Agriculture and Technology, P.O. Box:62000–
00200, Nairobi, Kenya*

ABSTRACT

Estimation of finite population total using birth-and-death process in the presence of a sample frame is considered. The model based on birth-and-death process model is proposed. Like the existing estimators, this estimation technique deals with initial condition and is based on yearly population totals in order to fit in a model within a given period of time. The model is capable of showing three processes, namely, exponential growth, exponential decay and constant growth. The proposed birth-and-death process model technique has shown to be efficient especially with large data. The empirical study indicated that model is efficient and can estimate properly even in the presence of outliers

Keywords: Birth-and-death, process, model, estimation, finite population, sample frame, exponential growth, decay, constant, empirical study and outliers.

1.INTRODUCTION

Sample surveys are widely used as a cost effective apparatus of data collection and for making valid inference about population parameters. Government bureaus and organizations use such methods to obtain the current information. The foremost aim of a statistician is to obtain

information about the population by deriving reliable estimates of unknown population parameters from a sample.

This study is using estimation techniques to estimate the bounded population and carrying capacity called the Logistic model that do not require any choice of step size as in the case of local polynomial regression estimator or have to be restricted a fix behavior, instead we allow the data to reveal its nature. The logistic model is use for data fitting. The logistic equation was introduced (around 1840) by the Belgian mathematician and demographer P.F. Verhulst as a possible model for human population growth [1].

Under simple random sampling (SRS) without replacement design, [2] proposed an exactly unbiased estimator for ϑ_{yx} . The proposed estimator is given by

$$\hat{\theta}_{HR} = \bar{r}_s + \frac{n(N-1)}{N(n-1)\bar{x}_u} (\bar{y}_s - \bar{r}_s \bar{x}_s) \dots \dots \dots (1.1)$$

where, $\bar{y}_s = \sum_{i \in S} \frac{y_i}{n}$, $\bar{r}_s = \sum_{i \in S} \frac{r_i}{n}$, $r_i = \frac{y_i}{x_i}$, $\bar{x}_s = \sum_{i \in S} \frac{x_i}{n}$, $\bar{x}_u = \frac{t_x}{N}$, the population ratio

$\theta_{yx} = \frac{t_y}{t_x}$, where $t_y = \sum_{i \in U} y_i$ be the population total for the variable Y , $t_x = \sum_{i \in U} x_i$ be the population total for the variable X and U of N units indexed by the set $\{1, 2, \dots, N\}$ a finite population. This estimator can be rewritten under general sampling design $p(\cdot)$. In this case, this estimator is no longer unbiased but still with negligible bias [3].

Under general sampling design, [4] proposed an estimator for estimating the population ratio ϑ_{yx} . This estimator, has negligible relative bias especially for small sample sizes n and approaches zero with increasing n . Under SRS, and based on simulation results, the performance of this estimator is better than that of (1.1). Their estimator is defined by

$$\hat{\theta}_{JM} = \bar{r}_s + \frac{1}{\bar{x}_s} (\bar{y}_s - \bar{r}_s \bar{x}_s) \dots \dots \dots (1.2)$$

Define π_i the first order inclusion probability, by

$$\pi_i = P_r(i^{th} \text{ element } \in s) = \sum_{i,j \in s} P(s) \dots \dots \dots (1.3)$$

For $i \neq j$, the second order inclusion probability is defined by

$$\pi_{ij} = P_r(i^{th} \text{ and } j^{th} \text{ elements } \in s) = \sum_{i,j \in s} P(s) \dots \dots \dots (1.4)$$

The [5] estimator of the population total $t_y = \sum_{i \in U} y_i$ is defined by

$$\hat{t}_{y\pi} = \sum_{i \in U} y_i \frac{I_{\{i \in s\}}}{\pi_i}, \dots \dots \dots (1.5)$$

where $I_{\{i \in s\}}$ is one if $i \in s$ and zero otherwise. Further,

$$\bar{y}_s = \frac{1}{N} \hat{t}_{y\pi} \dots \dots \dots (1.6)$$

can be used to estimate the population mean $\bar{y}_u = \frac{1}{N} t_y$. It can be noted that $\hat{t}_{y\pi}$ and \bar{y}_s are unbiased estimators for t_y and \bar{y}_u respectively. However, $\hat{t}_{y\pi}$ and \bar{y}_s do not use the availability of auxiliary variables in the study. In similar way,

$$\bar{x}_s = \frac{1}{N} \hat{t}_{x\pi}, \text{ and } \bar{r}_s = \frac{1}{N} \hat{t}_{r\pi} \dots \dots \dots (1.7)$$

are unbiased estimators for \bar{x}_u and \bar{r}_u respectively. Where \bar{x}_s is the sample mean of the inclusion probability of the auxiliary variable.

The availability of more than one auxiliary variable is used in literature for estimating the finite population total t_y , or finite population mean y_u .

Under SRS, [6] was the first one who deals with the problem of estimating the population mean using more than one auxiliary variables. His estimator is given by

$$\hat{y}_u = \sum_{i=1}^p w_i \bar{x}_{iu} \hat{\theta}_{yx} \dots \dots \dots (1.8)$$

where p is the number of the auxiliary variables, $\hat{\theta}_{yx} = \frac{\bar{y}_s}{\bar{x}_{is}}$ w_i is the weight of the i th auxiliary variable such that $\sum_{i=1}^p w_i = 1$ \bar{y}_s is the sample mean of Y and \bar{x}_{iu} , \bar{x}_{is} are the population mean and the sample mean of X_i , respectively, for $i = 1, \dots, p$. [7] proposed the following estimator

$$\hat{y}_u = \bar{y}_s \left(w_1 \frac{\bar{x}_{1u}}{\bar{x}_{1s}} + w_2 \frac{\bar{x}_{2u}}{\bar{x}_{2s}} \right) \dots \dots \dots (1.9)$$

for estimating the population mean \bar{y}_u , $w_1 + w_2 = 1$.

[8] studied the general form of (1.9). They proposed two classes of estimators using two auxiliary variables to estimate the population mean for the variable of interest Y.

[9] suggested a new multivariate ratio estimator using the regression estimator instead of \bar{y}_s which used in (1.9). Their estimator is given by

$$\bar{y}_{pr} = \sum_{i=1}^2 w_i \frac{\bar{y}_s + b_i(\bar{x}_{iu} - \bar{x}_{is})}{\bar{x}_{is}} \bar{x}_{iu} \dots \dots \dots (1.20)$$

where b_i , $i = 1, 2$ are the regression coefficients. Based on the mean squares error (MSE), they found that their estimator is more efficient than (1.9) when

$$MSE(\bar{y}_{pr}) < MSE(\bar{y}_u),$$

where $MSE(\bar{y}_{pr})$, and $MSE(\bar{y}_u)$ are defined by Equations (2.4), and (1.2) of Kadilar and Cingi (2004), respectively.

In subsection 2.1 we introduced the proposed birth-and-death process model, while subsection 2.2 talked about the asymptotic properties and section 3.1 talked about the empirical studies. Finally, section 4.0 drew a conclusion on the study. However, our population P(t) will be a continuous estimation to the actual population, which of course changes only by integral increments- that is, by one birth or death at a time.

Suppose that the population changes only by the occurrence of births and deaths- there is no immigration or emigration from outside the country or environment under consideration. It is customary to track the growth or decline of a population in terms of its birth rate and death rate functions defined as follows:

$\alpha(t)$ is the number of births per unit of population per unit of time at time t;

$\beta(t)$ is the number of deaths that occur during the time at time t.

Then the numbers of births and deaths that occur during the time interval $[t, t + \Delta t]$ is given (approximately) by :

$$\text{Births: } \alpha(t) \cdot P(t) \cdot \Delta t,$$

$$\text{Deaths: } \beta(t) \cdot P(t) \cdot \Delta t$$

Hence the change ΔP in the population during the time interval $[t, t + \Delta t]$ of length Δt is

$$\Delta P = \{\text{births}\} - \{\text{deaths}\} \approx \alpha(t) \cdot P(t) \cdot \Delta t - \beta(t) \cdot P(t) \cdot \Delta t \dots \dots \dots (1.21)$$

So

$$\frac{\Delta P}{\Delta t} \approx [\alpha(t) - \beta(t)]P(t) \dots \dots \dots (1.22)$$

The error in this approximation should approach zero as $\Delta t \rightarrow 0$, so- taking the limit – we get the differential equation

$$\frac{dP}{dt} = (\alpha - \beta)P \dots \dots \dots (1.23)$$

in which we write $\alpha = \alpha(t)$, $\beta = \beta(t)$, and $P = P(t)$ for brevity. Equation (1.22) is the general population equation. If α and β are constants, equation (1.22) reduces to the natural growth equation with

$K = \alpha - \beta$. But it also includes the possibility that α and β are variable functions of t . The birth and death rates need not be known in advance; they may well depend on the unknown function $P(t)$ [10].

2. ESTIMATION OF BIRTH AND DEATH PROCESS

This section is purposely considering an estimator, that is the birth and death process model.

2.1 Proposed Birth and Death process model

The birth and death process show a model for changes in the size of populations whose members can die (or dropout). This suggests generalizing the model by permitting transitions from the state E_n not only to the next higher state E_{n+1} , if $n \geq 1$. Suppose at epoch t the system is in state E_n , the probability that between t and $t + h$ the transition $E_n \rightarrow E_{n+1}$ occurs

equals $\alpha_n h + 0$, and the probability of $E_n \rightarrow E_{n-1}$ (if $n \geq 1$) equals $\beta_n h + 0(h)$. The probability that during $(t, t + h)$ more than one changes occurs is $0(h)$ (ie order of magnitude).

However, we can now obtain differential equations for the probabilities $P_n(t)$ of finding the system in state E_n . To calculate $P_n(t + h)$, note that the state E_n at epoch $t + h$ is possible only under one of the following conditions:

- 1) At epoch t the system is in E_n and between t and $t + h$ no change occurs;
- 2) at epoch t the system is in E_{n-1} and a transition to E_n occurs;
- 3) at epoch t the system is in E_{n+1} and a transition to E_n occurs;
- 4) between t and $t + h$ there occur two or more transitions. By assumptions, the probability of the last event is $0(h)$. The first three contingencies are mutually exclusive and their probabilities add. Therefore $P_n(t + h) = P_n(t)\{1 - \alpha_n h - \beta_n h\} + \alpha_{n-1} h P_{n-1}(t) + \beta_{n+1} h P_{n+1}(t) + 0(h) \dots \dots \dots (2.11)$

Transposing the term $P_n(t)$ and dividing the equation by h we get on the left difference ratio of $P_n(t)$, and in the limit as $h \rightarrow 0$;

$$P'_n(t) = \alpha_{n-1} P_{n-1}(t) + \beta_{n+1} P_{n+1}(t) - (\alpha_n + \beta_n) P_n(t) \text{ and } P'_0(t) = \beta_1 P_1(t) \dots \dots \dots (2.12)$$

This equation holds for $n \geq 1$. For $n = 0$ in the same way we have;

$$P'_0(t) = -\alpha_0 P_0(t) + \beta_1 P_1(t) \dots \dots \dots (2.13)$$

If the initial state is E_i , the initial conditions are:

$$P_i(0) = 1, P_n(0) = 0 \text{ for } n \neq i$$

As in many similar cases, the explicit solution of (2.12) is rather complicated, and it is desirable to calculate the mean and the variance of the distribution $\{P_n(t)\}$ directly from the differential equations. We have for the mean

$$M(t) = \sum_{n=1}^{\infty} n P_n(t) \dots \dots \dots (2.14)$$

Multiplying the first equation in (2.12) by n and adding over $n = 1, 2, \dots$, we find that the terms containing n^2 cancel, and we get

$$M'(t) = \alpha \sum (n - 1)P_{n-1}(t) - \beta \sum (n + 1)P_{n+1}(t) = (\alpha - \beta) M(t) \dots \dots \dots (2.15)$$

This is a differential equation for $M(t)$. The initial population size is i , and hence $M(0) = i$. Therefore;

$$M(t) = ie^{(\alpha - \beta)t} \dots \dots \dots (2.16)$$

We see that the mean tends to zero or infinity, according as $\alpha < \beta$ or $\alpha > \beta$. The variance of $\{P_n(t)\}$ can be calculated in a similar way as follows:

$$\frac{d}{dt} [n^2(t)] = 2[n(\alpha_n - \beta_n)] + [\alpha_n + \beta_n] \dots \dots \dots (2.17)$$

$$= 2[\alpha_n - \beta_n](n^2) + [\alpha_n + \beta_n](n) \dots \dots \dots (2.18)$$

$$[n^2(t)] - [n(t)]^2 = var(n) = \sigma^2(t) \dots \dots \dots (2.19)$$

$$\frac{d}{dt} \sigma^2 = 2[\alpha_n - \beta_n](n^2) + [\alpha_n + \beta_n](n) - 2(n)[\alpha_n + \beta_n](n) \dots \dots (2.20)$$

$$= 2[\alpha_n - \beta_n]\sigma^2 + [\alpha_n + \beta_n](n) \dots \dots \dots (2.21)$$

Let $\alpha - \beta = k$. If $\alpha - \beta > 0$, it means the population is growing (exponential growth) implying k is positive. If $\alpha - \beta < 0$, it means the population is decreasing (exponential decay) implying k is negative moving to extinction. If $\alpha - \beta = 0$, it means the birth rate and death rate are equal which leads to the initial population.

Table 1: Census Results

T	YEAR (X_t)	POPULATION TOTAL (P_t)
0	1969	10,942,705
10	1979	15,327,061
20	1989	21,448,774

30	1999	28,686,607
40	2009	38,610,097

The five census years obtained from a sample frame is shown in Table 1 above. However, we aimed at selecting 1969 population as pseudo initial population in order to obtain the population total ten years back before the actual census in 1969. These sample sizes will be used to estimate the population total in 2019 census using the proposed technique.

Here, $P_0 = 10,942,705$ at $t = 0$ (Initial population)

$$P_t = M(t) = 10942705 e^{kt} \dots\dots\dots(2.22)$$

$$P_{10} = 15327061 \text{ at } t = 10 \dots\dots\dots(2.23)$$

$$P_{40} = 38610097 \text{ at } t = 40 \dots\dots\dots(2.24)$$

Substituting equations (2.23) and (2.24) in (2.22) and evaluate simultaneously gives

$$k \approx 0.031$$

Suppose we want to know the population before the actual census begins, we set k as a negative value taking $P_0 = 10,942,705$ as initial population giving us the initial population before census.

$P_0(\text{before actual census}) = 8,025,894$ at $t = 0$ (time before actual census)

resulting to;

$$P_t = 8025894 e^{0.031t} \dots\dots\dots (2.25)$$

2.2. Asymptotic properties

Theorem: Law of large numbers:

Let X_1, X_2, \dots, X_n be iid random variables with common expectation

$\mu = E(X_i)$. Define $A_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then for any $\alpha > 0$, we have

$$P_r[|A_n - \mu| \geq \alpha] \rightarrow 0 \text{ as } n \rightarrow \infty$$

Proof of Theorem:

Let $Var(X_i) = \sigma^2$ be the common variance of the random variables; we assume that σ^2 is finite. With this (relatively mild) assumptions, the Law of Large Numbers (LLN) is an immediate consequence of Chebyshev's inequality. For as we have seen above, $E(A_n) = \mu$ and $Var(A_n) = \frac{\sigma^2}{n}$, so by Chebyshev we have

$$P_r[|A_n - \mu| \geq \alpha] \leq \frac{Var(A_n)}{\alpha^2} = \frac{\sigma^2}{n\alpha^2} \rightarrow 0 \text{ as } n \rightarrow \infty$$

Table 2: Provinces

Provinces	1969	1979	1989	1999	2009
Nairobi	509,286	827,775	1,324,570	2,143,254	3,138,369
C. Province	1,675,647	2,345,833	3,116,703	3,724,159	4,383,743
Coast Province	944,082	1,342,794	1,829,191	2,487,264	3,325,307
E. Province	1,907,301	2,719,851	3,768,677	4,631,779	5,668,123
N.E. Province	245,757	373,787	371,391	962,143	2,310,757
Nyanza	2,122,045	2,643,956	3,507,162	4,392,196	5,442,711
R. Valley	2,210,289	3,240,402	4,981,613	6,987,036	10,006,805
W. Province	1,328,298	1,832,663	2,544,329	3,358,776	4,334,202
Total Population	10,942,705	15,327,061	21,443,636	28,686,607	38,610,097

Table 2 above represents the census population from 1969 to 2009 in the eight provinces in Kenya. Successive sample sizes are selected below to show the law of large numbers.

Here, $N=40$ and $\mu = 2,875,251$

Sample 1: Nairobi (1969 to 2009)

n=5 and $\bar{x}_1 = 1,588,651$

Sample 2: Nairobi and Central

n=10 and $\bar{x}_2 = 2,318,934$

Sample 3: Nairobi, Central and Coast

n=15 and $\bar{x}_3 = 1,915,957$

Sample 4: Nairobi, Central, Coast & Eastern

n=20 and $\bar{x}_4 = 2,371,755$

Sample 5: Nairobi, Central, Coast, Eastern and N/Eastern

n=25 and $\bar{x}_5 = 2,067,957$

Sample 6: Nairobi, Central, Coast, Eastern, N/Eastern and Nyanza

n=30 and $\bar{x}_6 = 2,326,900$

Sample 7: sample 6 and R. Valley

n=35 and $\bar{x}_7 = 2,778,090$

Remark: We can clearly see the sample mean tending to the population mean as we approach the population total N which is in line with the Law of Large Numbers (LLN)

Therefore, $\lim_{n \rightarrow \infty} \bar{x}_n = \mu$

Comment Histec Inunique can track reasonably well throughout up to a sufficiently large number after which, there is a need to shift the initial condition to where the error margin starts increasing in order to maintain precision.

3. MAIN RESULTS:

3.1 Empirical analysis

Figure 3: A plot on birth-and-death process

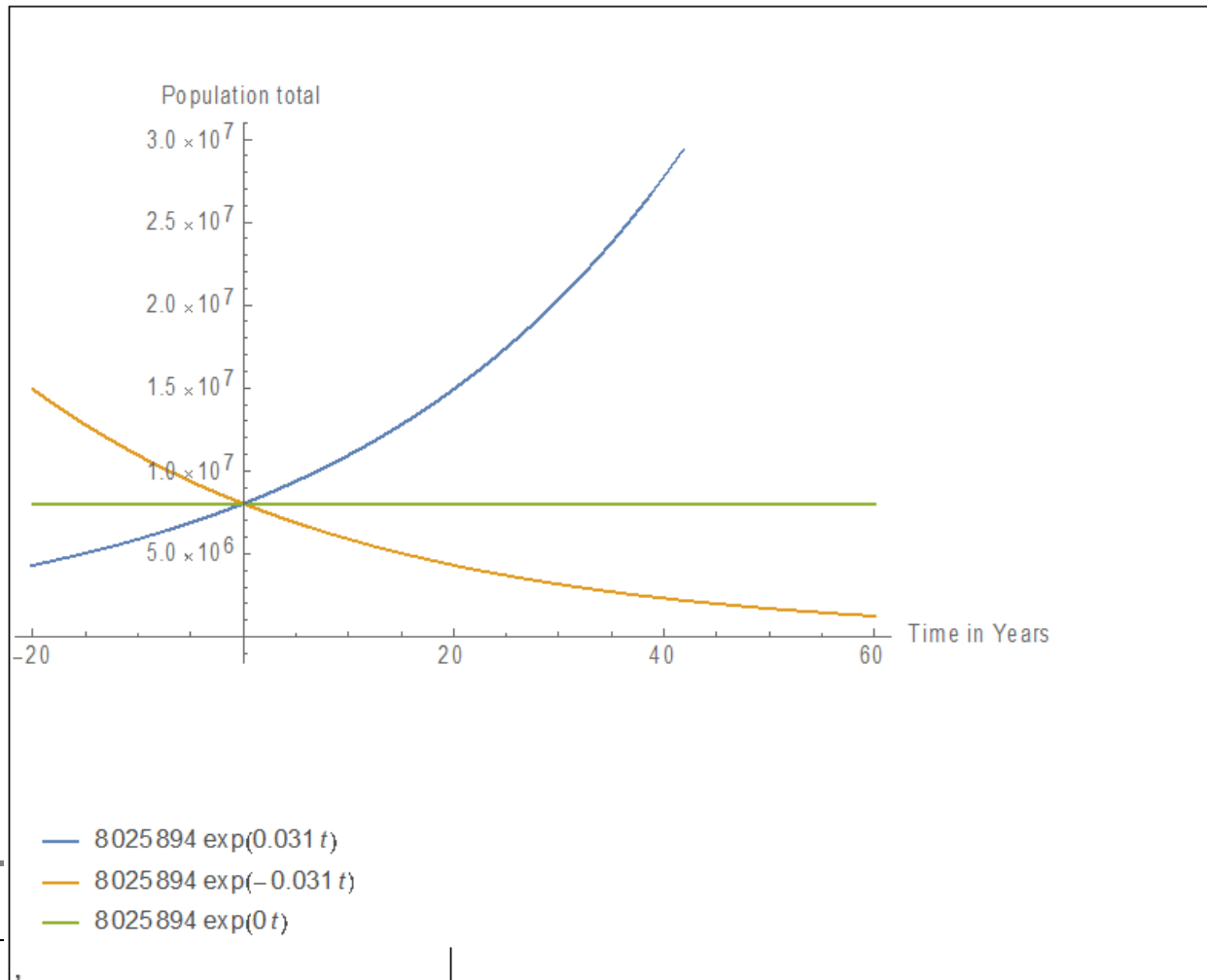


Figure 3 above shows the plot of birth-and-death process. The blue shows a positive growth (exponential growth), purple shows a negative growth (exponential decay) and the yellow shows a constant growth in population. Base on this, the annual population growth and decay figures can be obtained.

Table 3: Estimated population and error calculations

YEAR(t)	ACTUAL POPULATION	ESTIMATED POPULATION	ERROR(t_i)
1969(t=10)	10,942,705	10,942,705	0
1979(t=20)	15,327,061	14,919,559	407,502
1989(t=30)	21,443,636	20,341,702	1,101,934
1999(t=40)	28,686,607	27,734,387	952,220
2009(t=50)	38,610,097	37,813,760	796,337

Table 3 represents the actual population totals, estimated population totals and their corresponding errors from 1969 to 2009.

$$P_{2019}(60) = 51,556,230$$

4.0. CONCLUSION:

In this work, the birth-and-death model is very effective especially in trying to estimate the dynamics of population density in a given area with regard to population growth and decay. It can perform well with a sufficiently large sample size. The birth-and-death model can be more efficient in prediction especially where a regression model is ill conditioned. Which means, in a situation where the coefficients of a regression model are weak to give a better prediction, this technique can perform better in determining population total.

Disclosure of potential conflicts of interest: Authors strongly disclose no conflict of interest with regard to the publication of the paper.

Acknowledgements: We are grateful to God for the grace and mercy rendered to us in seeing us through this work. Special thanks go to the African union for making it possible to pursue this course through scholarship.

References:

- [1] C.Henry Edwards, David E. Penney,2008. Differential equations: Computing and modeling, 4th edition, 79-92.
- [2] Hartley H, Ross A (1954). "Unbiased Ratio Estimates." *Nature*, **174**, 270–271.
- [3] Al-Jararha J (2012). *Unbiased Ratio Estimation for Finite Populations*. LAMBERT Academic Publishing, Germany.
- [4] Al-Jararha J, Al-Haj Ebrahim M (2012). "A Ratio Estimator Under General Sampling Design." *Austrian Journal of Statistics*, **41**, 105–115.
- [5] Horvitz D, Thompson D (1952). "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*, **47**, 663–685.
- [6] Olkin I (1958). "Multivariate Ratio Estimation for the Finite Populations." *Biometrika*, **45**, 154–165.
- [7] Singh D, Chaudhary F (1986). *Theory and Analysis of Sample Survey Design*. New Age Publication, New Delhi, India.
- [8] Abu-Dayyeh W, Ahmad M, Ahmad R, Hassen A (2003). "Some Estimators of a Finite Population Mean Using Auxiliary Information." *Applied Mathematics and Computations*, **139**, 287–298.
- [9] Kadilar C, Cingi H (2004). "Estimator of a Population Mean Using Two Auxiliary Variables in Simple Random Sampling." *International Mathematical Journal*, **5**, 357–367.
- [10] Kabareh, L., & Mageto, T. (2017). Estimation of Bounded Populations and Carrying Capacity with the Logistic Model. *Open Journal of Statistics*, 7(6), 936–943.
<https://doi.org/10.4236/ojs.2017.76065>

