# Estimation Of Population Total Using Model-Based Approach: A Case Of HIV/AIDS In Nakuru Central District, Kenya

Langat Reuben Cheruiyot, Tonui Benard Cheruiyot, Lagat Janet Jepchumba

**Abstract**: In this study we have explored an estimator for finite population total under the famous prediction approach. This approach has been compared with design-based approach using simple random sampling and stratified random sampling techniques. It is shown that the estimators under model based approach give better estimates than the estimators under design based approach both when using simple random sampling (s.r.s) and stratified random sampling. The relative absolute error from both approaches is computed and has been shown to be superior under the super population model than the design based approach. This approach is then applied to predict the total number of people living with HIV/AIDS in Nakuru Central district.

**Index Terms**: Model- based approach, design -based approach, simple random sampling, stratified sampling, HIV/AIDS.

## 1 INTRODUCTION

THE main objective of sample survey is to obtain information about the population. The information that we seek about the population is normally, the total number of units, aggregate values of various characteristics, averages of this characteristics per unit, proportions of units possessing specified attributes etc. The data can be collected by either census or sampling methods. Census is the complete enumeration where data is collected on the survey characteristics from each unit of the population. Sampling methods consist of collection of data on survey characteristics from selected units of the population. A sampling method makes it possible to estimate population totals, averages or proportions while reducing at the same time the size of the survey operations. Over two decades since the first AIDS case was described in Kenya, HIV/AIDS still remains a huge problem for the country in its efforts for social and economic development. Responses to the pandemic have evolved over time as people became aware of this new disease, as they experienced illness and death among family members, and as services have developed to confront this epidemic. Initially many segments of society expressed denial of the disease.


- *University of Kabianga, Mathematics and Computer Science Department, P.O Box 2030-20200 Kericho Kenya. Email: rlangat@kabianga.ac.ke*
- *University of Kabianga, Mathematics and Computer Science Department, P.O Box 2030-20200 Kericho Kenya. Email: tonuib@kabianga.ac.ke*
- *University of Kabianga, Mathematics and Computer Science Department, P.O Box 2030-20200 Kericho Kenya. Email: lagatcjanet@yahoo.com*


While awareness of AIDS has been nearly universal for more than a decade, misconceptions still abound and many still have not dealt with this disease at a personal or community level. In the last five years HIV-related health services have expanded dramatically; they include the widespread availability of testing and counseling, and treatment with antiretroviral drugs, both to prevent mother-to-child transmission and to improve health and prolong life for people with advanced HIV infection and AIDS. While HIV remains an incurable infection, Kenya has now entered an era in which there is new hope in treating and caring for people with AIDS. This hope also offers new effective opportunities for preventing HIV infection as people with HIV infection learn they are infected and learns how to better protect their loved ones. See Kenya AIDS Response Progress Report, (2014) Progress towards Zero for an insight. Many countries have developed surveillance systems for tracking HIV infection and the behaviors that spread HIV. However, countries may lack the capacity to estimate the size of populations of those living with HIV/AIDS so as to enable them plan well. Estimating population totals of people living with HIV is essential because it will enable the government plan well for personnel needed, the drugs and the fund needed to support the infected and affected group such as the orphans. In the past years the government has been running short of the ARV drugs due to poor planning by the agencies required to purchase ARVs. In this study we explore a model-based approach to estimating the population totals of people with HIV/AIDS. Measuring and understanding the impact and magnitude of the human immunodeficiency virus (HIV) epidemic presents many challenges. Yet without accurate measures and estimates of the impact and magnitude of HIV, it is impossible for countries to carry out HIV programme activities, such as advocating for most-at-risk populations, planning and implementing HIV prevention, care, treatment and Evaluation programmes. Establishing the size of populations most-at-risk to HIV allows epidemiologists to develop models which estimate and project HIV prevalence or inform countries of the distribution of HIV incidence within such a country. There are quite a number of researchers in literature who have carried out research on various sampling strategies that can be employed to give desired results. A comprehensive review of the developments in sample surveys can be found in Rao, (2006). For a detailed coverage of the sampling strategies one can see Chambers,

(2003)

## 2 DESIGN-BASED APPROACH

In the design-based framework ***X*** and ***Y*** **(**where ***Y*** is the population characteristic of interest and ***X*** is a known auxiliary variable) are regarded as constants and the only source of randomness is the selection of the sample. A simple random sample can be taken from the population under study or in the case of stratified random sampling, before selecting the sample, the population must be divided into parts which are called sampling units, or units. These units must cover the whole of the population and they must not overlap, in the sense that every element in the population belongs to one and only one unit, Cochran (1977) Let us suppose that the population consists on *N* such units and, of these, *n* are chosen for the sample. The simplest design-based approach is simple random sampling (srs). When the sample is chosen without replacement, there are *N!/n!(N-n)!* ways of selecting the sample and, under simple random sampling, each of these has exactly the same chance of arising. Let the quantity of interest on the $i^{th}$ unit be denoted by $Y_i$**.** The sampling distribution is generated by the randomization process and is formed by the *N!/n!(N-n)!* possible outcomes. The population total is Y=$\sum_{i=1}^{N} Y_i$ and the population mean $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$.

The sample mean $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ Randomization theory works for deriving properties of the sample mean in simple random sampling. As in Cornfield (1944), we let

$$\pi_i = \begin{cases} 1, \text{if a unit } i \text{ is in the sample} \\ 0, \quad \text{otherwise} \end{cases}$$

For a simple random sampling without replacement (*srswor)* each one of the *n* units has the same selection probability, $\pi_i = \frac{n}{N}$, but with more complicated plans each sample unit can have a different value of $\pi_i$, Horvitz and Thompson (1952). The expansion estimator, defined as the product of the population size and the sample mean, $\hat{T}_0$= N$\bar{y}_s$, is an unbiased estimator of the population total $T$, that is $E_\pi\ (\hat{T}_0) = T$.

And $\hat{T}_\pi = \frac{\sum_{i \in s} y_i}{\pi}$

The design variance is given by $Var_\pi\ (\hat{T}) = N^2 \left(1 - \frac{n}{N}\right) . \frac{S^2}{n}$ where $S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1}$

In the model based framework a model is assumed for ***y*** conditioned on ***X*****.** Suppose that the number of units in the finite population is known and that each unit is associated a number $y_i$. The general problem is to choose some of the units as a sample, observe the y's for the same unit ,and then use those observations to estimate the value of some function $(y_1, y_2, \dots, y_N)$ of all the y's in the population. The function $(y_1, y_2, \dots, y_N)$ can be a simple combination of the y's like their total or mean. The prediction approach treats the numbers $(y_1, y_2, \dots, y_N)$ as realized values of random variables $Y_1, Y_2, \dots, Y_N$. After the sample has been observed, estimating $(y_1, y_2, \dots, y_N)$ entails predicting a function of the unobserved Y's. The regression model for prediction is given by

$$y_i = x_i \beta + e_i$$

For the Model-based approach, estimate total population is given by

$\hat{T} = \sum_{i \in s} y_i + \sum_{i \notin s} y_i$ where $\sum_{i \notin s} y_i = \hat{\beta} \sum_{i \notin s} x_i$ and $\hat{\beta} = \frac{\sum y_i}{\sum x_i}$

The sample mean is given by $\bar{y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

## 3 MODEL-ASSISTED APPROACH

Royall and Cumberland (1981), Hansen et al (1983) and Rao (1996), however, demonstrated the model – based approach's poor performance especially in large samples under non-self weighting designs, even if deviations from the model are small. The two groups of practitioners of the two approaches so far highlighted above seem to have taken hard stands. Nonetheless, Brewer (2002) and Särndal et al, (1992) have tried to look at these two approaches from a unified framework of model-assisted approach. In model-assisted approach some auxiliary variables are commonly incorporated in the estimation procedure by using a model, but the inferences are still design based. In this approach, the model is used to increase the efficiency of the estimators, but even when the model is not correct, estimators typically remain design-consistent (Breidt & Opsomer, 2000). Auxiliary information on the finite population is often used to increase the precision of estimators of the population mean, total or the distribution function (Wu & Sitter 2001). As an example, the ratio estimator contains known information (population total) of some auxiliary variable. Under the model-assisted approach the estimate of the total population is given by

$$\hat{T} = \sum_{i \in s} y_i + \sum (y_i - \hat{y})\, \pi_i^{-1}$$

## 4 DESIGN

Our study population consisted of the estimated 80,000 people living with HIV in Nakuru Central District. The study sample comprised of 8 units under simple random sampling and in stratification the sampled units were 7 in stratum 1 and 3 in stratum 2. Strata 1 consisted of hospitals with patients less than one thousand while strata 2 consisted of hospitals with more than 1000 patients as shown in the table 2 and 4. The data collected is given in Table 1 in the appendix.

## 5 RESULTS

We note here that there are hospitals with high number of patients such as Nakuru Provincial General Hospital and Langalanga Health Centre. When such values are included in the sample, then if the size of the sample is small it tends to overestimate the total. The opposite is also true if the sample has centres with small values. One therefore may be tempted to perform purposive sampling, a situation that obviously introduces personal biases. (Sharon, 2009) remarks that in order to ensure that all health facilities are represented; the randomly selected sample should not be adjusted purposively. Taking a stratified sample helps one out under such circumstances.

### 5.1 Design-Based Approach

Under Design based approach, the total population estimate is given by;

$$\hat{T}_1 = \frac{\sum_{i\in s} y_i}{\pi} = 9{,}023 \times \frac{22}{8} = 24{,}813$$

The absolute value of the relative error under this model is given by

$$\frac{(24{,}813 - 82{,}597)}{82{,}597} = 0.699$$

This means the number of patients living with HIV/AIDS is underestimated by 69.9% under this approach.

### 5.2 Model-Based Approach

Under Model-based approach, population total for the year 2013 is given by:

$$\hat{T} = \sum_{i\in s} y_i + \sum_{i\notin s} y_i$$

Where $\sum_{i\in s} y_i$ is the total number patients in the 8 sampled hospitals and $\sum_{i\notin s} y_i$ is the total number of patients in 14 non sampled hospitals. Since $\sum_{i\in s} y_i$ is known, estimating $T$ is equivalent to predicting the number of patients in the year 2013 $(y)$ from the 14 non sampled hospitals.

For the sample of 8 hospitals,

$$\hat{\beta} = \frac{\sum y_i}{\sum x_i} = \frac{82{,}597}{25{,}676} = 3.217$$

Now the number of patients for the year 2013 in the 14 non sampled hospitals are $x_1$=8, $x_2$=2, $x_3$ =24, $x_4$=1346, $x_5$=2627, $x_6$=6, $x_7$=25, $x_8$=3, $x_9$=268, $x_{10}$=15210, $x_{11}$=26, $x_{12} = 602$, $x_{13} = 340$ and $x_{14}$=340 so their predicted number of patients for the year 2013 are

$$\hat{\beta}\sum_{i\notin s} x_i = (3.217)(20{,}837) = 67{,}032.629$$
$$\cong 67{,}033$$

Hence the estimate for the population total become

$$\hat{T}_2 = \sum_{i\in s} y_i + \hat{\beta}\sum_{i\notin s} x_i$$
$$= 9{,}023 + 67{,}033$$
$$= 76{,}056$$

The above value is close to $T = 82{,}597$ from the 22 hospitals. The absolute value of the relative error is given by

$$\frac{(\hat{T} - T)}{T} = (76{,}056 - 82{,}597)/82{,}597$$
$$= 0.079$$

This means that the model based approach using simple random sampling underestimates the population by about 7.9%.

### 5.3 Design-Based Approach after Stratification

We know that $\hat{T} = \frac{\sum_{i\in s} y_i}{\pi}$. Using the data in table 1.2 and 1.3,

$$\hat{T}_{str\ 1} = 1{,}012 \times \frac{17}{7} = 2{,}458$$

And $\hat{T}_{str\ 2} = 70{,}605 \times \frac{5}{3} = 117{,}675$

$$\Longrightarrow \hat{T}_3 = \hat{T}_{str\ 1} + \hat{T}_{str\ 2} = 120{,}133$$

The absolute value of the relative error is given by;

$$\frac{(\hat{T} - T)}{T} = \frac{(120{,}133 - 84{,}597)}{84{,}597} = 0.420$$

### 5.4 Model-Based Approach after Stratification

$$T_{str\ 1} = \frac{\sum y_i}{\sum x_i}(X_1)$$

where $X_1$ = the total in strata 1.
$\sum y_i$ = the total number of patients for the year 2013.
$\sum x_i$ = the total number of patients for the year 2012.

Therefore,

$$T_{str\ 1} = \frac{1{,}012}{379}(1{,}836) = 4{,}902$$

$$T_{str\ 2} = \frac{\sum y_i}{\sum x_i}(X_2) = \frac{70{,}605}{19{,}183}(21{,}213)$$
$$= 78{,}077$$

$$\Longrightarrow \hat{T}_4 = T_{str\ 1} + T_{str\ 2} = 82{,}979$$

The absolute value of the relative error is given by;

$$\frac{(\hat{T} - \mathrm{T})}{\mathrm{T}} = \frac{(82{,}979 - 82{,}597)}{84{,}597} = 0.004$$

## 6. CONCLUSION

From table 6, it can be seen that in both the design and model based approaches the relative errors after stratification are less than the relative errors before stratification. Also the relative error under model based approach is lower than the relative error under design based approach in both sampling techniques (simple random sampling and stratified random sampling). Therefore model based approach is a better approach to use compared to the design based approach while estimating finite population total. It can also be seen that estimated figures of HIV prevalence in Nakuru Central District is about 82,979 which is about 3.7% of the total population. Such estimates help to give the overall picture on the ground and guide in planning purpose.

## 7. FUTURE RESEARCH

In this study we have considered the model -based approach which makes assumptions about the observations not in the sample; that they have the same mean and variance as observations that are in the sample. There is a need study model assisted approach to estimation of finite population totals using simple random sampling and stratified random sampling techniques.

## REFERENCES


[1] Brewer, K (2002). Combined survey sampling inference. London, Arnold a member of the Hodder Headline Group.

[2] Breidt, F. J. and Opsomer, J.D. (2000). Local Polynomial Regression Estimation in Survey Sampling. Annals of Statistics, 28, 1026.

[3] Chambers, R.L. (2003). Which Sample Survey Strategy? A Review of Three Different Approaches. Southampton Statistical Sciences Research Institute. University Of Southampton.

[4] Cochran W.G. (1977). Sampling Techniques (3rd ed.). New York: John Wiley and sons.

[5] Cornfield (1944). On samples from finite populations. Journal of the American Statistics Association.39:236-239.

[6] Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model dependent and probability sampling inference in sample surveys. Journal of the American statistical Associations, 78,776-807

[7] Horvitz, D. G. & Thompson, D. J. (1952). 'A Generalization of Sampling Without Replacement from a Finite Universe', Journal of the American Statistical Association 47, 663–685.

[8] National AIDS Control Council (2014). Kenya AIDS Response Progress Report,2014 Progress towards Zero .

[9] Rao, J.N.K, (2006). Interplay Between Sample Survey Theory and Practice: An Appraisal. Statistics Canada. Business survey Methods. 31. 2,117-138

[10] Rao, J.N.K (1996). Development in sample survey theory. The Canadian journal of statistics, 25, 1-21

[11] Royall R.M, and Cumberland W.G (1981). Prediction models and unequal probability sampling. Journal of the American Statistical Association.

[12] Särndal, C.E Swesson, B and Wretman, J.N (1992). Model- Assisted Survey sampling. New York: Springer Verlag.

[13] Sharon L.L. (2009). Sampling: Design and Analysis 2nd Ed. Richard Stratton,US.

[14] Wu, C, and Sitter, R.R. (2001). A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data. Journal of American Statistical Association.


## APPENDIX: TABLES

**Table1: Data collected**

| | Health Facility Name | 2012 Data ($X$) | 2013 Data ($Y$) |
|---|---|---|---|
| 1 | Afya Medical Clinic | *8* | *6* |
| 2 | AIC Parkview Dispensary | *2* | 16 |
| 3 | Bangii Medical Clinic | 8 | 3 |
| 4 | Bethsaida (AIC) Clinic | *24* | *17* |
| 5 | Coping Centre | 16 | 7 |
| 6 | Family Health Options Kenya (Nakuru) | *1346* | *3269* |
| 7 | Fountain Medical Clinic | 5 | 9 |
| 8 | Fitc | 751 | 1505 |
| 9 | Kapkures Dispensary | *2627* | *2327* |
| 10 | Kiti Dispensary | 28 | 30 |
| 11 | Lalwet dispensary | *6* | *12* |
| 12 | Lanet Health Centre | *25* | *42* |
| 13 | Langalanga Health Centre | 3906 | 7379 |
| 14 | Marie Stopes Nakuru | *3* | *3* |
| 15 | Mother Kelvin Dispensary | *268* | *818* |
| 16 | Nakuru Provincial General Hospital | *15210* | *65009* |
| 17 | Nakuru West (PCEA)Health Care | 121 | 72 |
| 18 | Nakuru West Heath care | *26* | *106* |
| 19 | PCEA Upendo Heath Care | 4 | 18 |
| 20 | Prison Dispensary | *602* | *881* |
| 21 | Sunrise Evans Hospital | *340* | *668* |
| 22 | Valley hospital | *350* | *400* |
| Totals for $N = 22$ | | **25,676** | **82,597** |
| Totals for $n = 8$ | | **4,839** | **9,023** |

Source: *Kenya health information online system*

**Table 2: Stratum 1**

| Unit | Hospital | $X_i$ | $Y_i$ |
|---|---|---|---|
| 1 | Afya Medical Clinic | 8 | 6 |
| 2 | AIC Parkview Dispensary | ***2*** | 16 |
| 3 | Bangii Medical Clinic | 8 | 3 |
| 4 | Bethsaida (AIC) Clinic | 24 | 17 |
| 5 | Coping Centre | 16 | 7 |
| 6 | Fountain Medical Clinic | 5 | 9 |
| 7 | Kiti Dispensary | 28 | 30 |
| 8 | Lalwet dispensary | 6 | 12 |
| 9 | Lanet Health Centre | 25 | 42 |
| 10 | Marie Stopes Nakuru | 3 | 3 |
| 11 | Mother Kelvin Dispensary | 268 | 818 |
| 12 | Nakuru West (PCEA)Health Care | 121 | 72 |
| 13 | Nakuru West Heath care | 26 | 106 |
| 14 | Pcea upendo health care | 4 | 18 |
| 15 | Prison Dispensary | 602 | 881 |
| 16 | Sunrise Evans Hospital | 340 | 668 |
| 17 | Valley hospital | 350 | 400 |
| | **Total** | **1,836** | **3,108** |

**Table 3: Randomly selected Samples in stratum 1**

| Unit | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | 8 | 6 |
| 3 | 8 | 3 |
| 5 | 16 | 7 |
| 7 | 28 | 30 |
| 9 | 25 | 42 |
| 11 | 268 | 818 |
| 13 | 26 | 106 |
| Total | **379** | **1,012** |

**Table 4: Stratum 2**

| | Hospital | $X_i$ | $Y_i$ |
|---|---|---|---|
| 1 | Family Health Options Kenya (Nakuru) | 1346 | 3269 |
| 2 | Fitc | 751 | 1505 |
| 3 | Kapkures Dispensary | 2627 | 2327 |
| 4 | Langalanga Health Centre | 3906 | 7379 |
| 5 | Nakuru Provincial General Hospital | 15210 | 65009 |
| | **Total** | **21,213** | **79,489** |

**Table 5: Randomly selected samples from stratum 2**

| Unit | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | 1346 | 3269 |
| 3 | 2627 | 2327 |
| 5 | 15210 | 65009 |
| Total | **19,183** | **70,605** |

**Table 6: Comparison of Model based and Design based approaches**

| Approach | Sampling Technique | Estimate | Relative Error |
|---|---|---|---|
| *MBA* | $T_2$ | 76,056 | 0.079 |
| | $T_4$ | 82,979 | 0.004 |
| *DBA* | $T_1$ | 24,813 | 0.699 |
| | $T_3$ | 120,133 | 0.420 |

Where $T_1 =$ *DBA* – Design-based Approach under $s.r.s$

$T_2 =$ *MBA* -Model-based Approach under $s.r.s$

$T_3 =$ *DBA* under stratified random sampling

$T_4 =$ *MBA* under stratified random sampling