UNIVERSITY OF ILORIN, ILORIN, NIGERIA

THE ONE HUNDRED AND TWENTY-FIRST (121ST)
INAUGURAL LECTURE

"THE ART OF EVERYDAY SAMPLING"

BY

PROFESSOR ISAAC OLAYIWOLA OSHUNGADE
B.Sc. M.Sc. (Ibadan), Ph.D. (Essex)
PROFESSOR OF STATISTICS
FACULTY OF SCIENCE, UNIVERSITY OF ILORIN

THURSDAY 31ST JANUARY, 2013

This 121st Inaugural Lecture was delivered under the Chairmanship of

The Vice-Chancellor
Professor A.G. Ambali
DVM (Zaria), M.Sc., Ph.D. (Liverpool), MCVSN (Abuja)

January, 2013

Published by:
The Library and Publications Committee,
University of Ilorin, Ilorin, Nigeria.

Printed by:
Unilorin Press

PROFESSOR ISAAC OLAYIWOLA OSHUNGADE
B.Sc. M.Sc. (Ibadan), Ph.D. (Essex)
PROFESSOR OF STATISTICS

Courtesies

The Vice Chancellor,
Deputy Vice Chancellors (Academic, Management services and Research, Training and Innovations)
Registrar and other Principal Officers of the University,
Provost, College of Health Sciences,
Dean of Science and other Deans of Faculties,
Deans of Post-Graduate School and Student Affairs,
Directors of Units,
Professors and other members of Senate,
Head of Department of Statistics and other Heads of Departments,
Members of Academic Staff,
Members of Administrative and Technical Staff of the University,
My Lords Spiritual and Temporal,
Members of my family; Nuclear and Extended,
Distinguished invited Guests,
Gentlemen of the Print and Electronic media,
Great Unilorin Students,
Ladies and Gentlemen.

Introduction

It gives me a great pleasure to stand before you this evening to present the 121st inaugural lecture of this great University. This is the second University Inaugural lecture in 2013 and the 4th from the Department of Statistics.

Sampling has become a good tool for decision making in almost every aspect of our daily life. Government, research institutions, and individuals use it for planning, testing some beliefs (hypothesis) and confirmation of some theories. The topic of the lecture is "**The Art of Everyday Sampling**". This title was chosen in order to;

i. enlighten the whole public on the roles played by sampling methods in practice,
ii. consider the methods of sampling, and
iii. highlight some of the problems in survey sampling and how to solve them.

Statistics is a branch of Mathematics and sampling theory and method depends on it. I have made this lecture as simple as possible in order that the public would understand it.

In our everyday life, our attitudes, knowledge and decisions are based on samples. A **sample** is a group of randomly selected few from the **population**. Samples are used for obtaining reliable, better and timely data when compared with censuses or complete enumeration. No wonder Jesus Christ in His parable of workers paid equally said in the Holy Bible that "Many are called but few are chosen" Matthew 20: 16. In this case, the "Many" are the population and the "few" are the sample.

The main questions when we want to take a sample are always:

- How should the sample be taken?
- How large should be the size of the sample?
- How should the population parameters be estimated? and
- How reliable are these estimates?

The purpose of all these questions is to obtain a representative sample of the population as a whole. A representative sample is one that is unbiased, reliable with least variation and efficient.

Elementary idea of Sampling

Definition: Sampling is the selection of a part or fraction (sample) of a population, observing the selected part with respect to some property of interest e.g. height, cost, population total, and then drawing some conclusions (inference) about the population using procedures based on statistical theory.

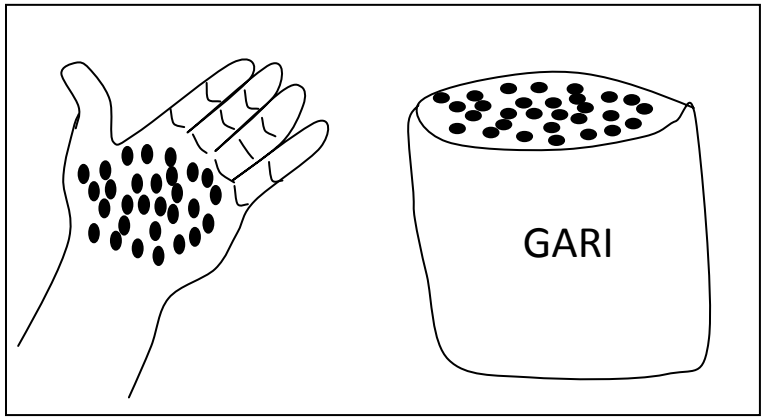
Population is the total number of units or objects of study of interest from which a sample can be taken. The sample must be of the same units as the population. You cannot have in the population of goats, a sample of sheep.

Examples of our everyday sampling are so numerous but simple and common ones include :

i. housewife cooking soup in a pot tastes **a spoonful** of soup to observe the sweetness and salt content of the soup in the whole pot .
ii. medical doctor will examine a patient with **a few** millilitres of blood from the patient after a laboratory test to determine the presence of a disease.
iii. cutting **a piece** of cloth to determine the quality of the whole roll of cloth.
iv. the produce inspector of cocoa beans judges the quality of a sack of cocoa bean by inspecting **a handful** of cocoa beans.
v. **Oshungade** (1989b), used the example of taking **a handful** of *gari* which is a common and popular food in Nigeria to judge the quality of a basin of *gari*.

Some of these concepts like sample, population, random selection, reliability , representativeness (unbiasedness) and sampling distributions were illustrated using this

example . The usefulness of sampling as a basis for decision making on the whole population was also emphasised. Fig.1 below shows the sample and the population



HANDFUL (sample) BASIN of GARI (population)

Fig.1: Sample and Population

The words in bold letters in (i to v) above are the **samples** that are taken from the following **population of:** whole pot, blood, roll of cloth, sack of cocoa, and basin of gari respectively. In the cases (i-v) above, a sample is selected because it is impossible, inconvenient, slow and uneconomical to monitor the entire population.

The problems of survey sampling are how to:

- Design a representative sample.
- Minimise the error of sampling at a minimum cost and a high precision.
- Make correct inference about the sampled population.

The major collector of survey data in Nigeria is the Nigeria Bureau of Statistics (NBS) with her head office in Abuja. Other institutions that collect data using sample survey are universities and research institutions e.g. NISER and IITA at Ibadan.

Statistical Sampling:

Statistical sampling as listed by Dalenious (1989) includes the following five special areas;

i. Experiment.
ii. Monte Carlo sampling (simulation).
iii. Acceptance Sampling (quality control).
iv. Sampling for auditing.
v. Survey Sampling.

The common thing with all these areas is the use of samples.

Survey Sampling

In this lecture I am concerned with the survey sampling which is my area of major interest. The following are the major differences between a survey sampling and the traditional statistical theory or inference:

i. The idea of identifiability of units in survey sampling does not exist in traditional statistical theory.
ii. In survey sampling, we are not confined to independently and identically distributions (iid) as in traditional statistical inference. This is as a result of (i) above.
iii. Inference for the infinite hypothetical population is usually assumed in statistical theory. In this set up, there is usually a sample of n independent

observations $x_1, x_2, \ldots, x_n$ on a random variable X, with the hypothetical density $f(x, \theta)$ and the problem is to estimate the unknown parameter θ only.

iv. Finite real populations are used in survey sampling and the concepts such as parameter, sample, data and estimator are given special meaning in survey sampling. For example, parameters in survey sampling are mean, total, proportion and ratio values.

v. The notion of sampling decision has no direct counterpart in traditional statistical theory, although in the area of experimental design, there is a direct related idea.

A sample survey always brings together three different areas of:

i. survey design (sampling technique and selection of sampling units.)
ii. design of questionnaire.
iii. method of data collection in statistics.

The combinations of the three areas are highly important for good survey design.

Major types of Sampling Techniques

There are two major types of sampling techniques.

i. Non-random sampling
ii. Random sampling which is also called the probability sampling.
    - **A non-random sampling** include judgement or purposive sampling and quota sampling.

Others are use of sample of volunteer subjects and expert choice.

This method limits the sample to units that appear to be representative of the population under consideration. Interviewers are free to choose their respondent satisfying the specified quota and experts pick the typical or representative specimens, units.

The major disadvantages or problems with non-random sampling are:

- with judgement sampling, there will be various differences between the respondents since the interviewers do not have the same method of interviewing and,
- most important sampling error cannot be objectively determined or obtained.

Probability or Random Sampling

Probability sampling is the method recognised in most practical situations to achieve the mirror or miniature of the population i.e. representative sampling. It is a sampling procedure in which every sampling unit belonging to the population has a known and non-zero probability of being selected in the sample. It is free from selection bias or human interference.

An important characteristic of the probability sampling technique is that there is no control over the specific choice or selection of the sampling units. This is achieved through the use of randomisation.

Other important characteristics of probability sampling include:

- the units in the sample are obtained through some mechanical operation of randomisation e.g. use of statistical tables of random numbers .
- there are laid down procedures for any pair of design and estimation procedure and
- the sampling variability or the standard errors of any probability sampling can be estimated.
- appropriate weights to the probabilities of selection are used e.g. stratified sampling.

Consequently, this method demands competent field workers (enumerators) and careful execution of the instructions at all the stages of the survey.

Examples of probability sampling techniques are: **simple random sampling**, **systematic sampling**, **stratified sampling** and **cluster sampling**. Other more complex techniques are **multistage sampling** and **unequal probabilities sampling**.

In summary, probability sampling allows one to have the confidence that the sample is unbiased and one can also estimate how precise the estimates from the data are likely to be. Probability sampling always produce data from a properly chosen sample and are a great improvement over that of a non-random sampling.

Sample Survey:

Sample survey is an organised fact-finding means of collecting data from the sample of units selected from the population. It allows decision to be made over the problems that are meant to be solved.

There are two major categories of sample surveys. These are:

i. **Descriptive survey** it deals with a finite population and involves one subject of investigation and describes the population involved.
ii. **Analytical surveys** involve several subjects of investigation, statistical test are performed. Modelling may also be involved.

In any research based on the use of samples, survey sampling will provide the:

- Sampling theories which include the estimation method and test of hypothesis.
- Methods of data collection through personal interview, mail survey, experimentation, observations and extraction of data from existing published ones.
- Data processing methods.
- Evaluation of the Quality of data by the levels of sampling and non-sampling errors involved.
- Analysis of survey data which would include; editing, processing, and report writing (technical and general reports).

The main goals of survey sampling are to:

- achieve a representative result, in terms of good inference about the target population.
- provide unbiased results through appropriate use of design and its estimator.

Survey sampling can also play the following roles. It can provide:

- estimate of current point parameters like mean, total, proportion and ratio.

- series of data and results over some periods of time for the purpose of looking at the trend.
- a check for the quality of census and administrative statistics.
- small area statistics for constituency and any other very small areas like enumeration areas and villages.

Fields of Application in Sample Survey

Sample survey can be used to collect data on the followings;

- Demography and Medical statistics. e.g. fertility and demographic surveys.
- Housing statistics. e.g. housing condition.
- Industrial statistics.
- Statistics of commerce and services.
- Agricultural statistics.
- Socio-economic studies of households.
- Opinion and attitude surveys.
- Education statistics.
- Other special areas are randomised response techniques, capture-recapture methods and small area statistics.

My Contributions to Statistics and Knowledge

My first employer in statistics was the Federal Office of Statistics (FOS) (Now, National Bureau of Statistics, NBS) at Lagos in 1968. I was employed at an interview as a holder of WASC (O.L) with a minimal knowledge of statistics in Additional mathematics and I

was posted to the Rural Economic Survey unit. This unit deals majorly with Agricultural statistics of Nigeria.

I started with counting of units in the sketched yield plots from the field with a plenimeter and later moved to another sector of the unit dealing with **analysis of variance**. With my experience in this unit, I developed a flair for survey sampling .

I was also taught by an Indian professor, Professor Des Raj, a UNO expert in survey sampling at the University of Ibadan (1971-1973) for a Professional Diploma Course in statistics..

For the degree of B.Sc. (Statistics), my project was "on a sample survey of housing condition at Ilesa" and my project at the M.Sc. (Statistics) was on" the problem of non-response in some sample surveys in Nigeria".

For the Ph.D. Degree of particular interest to me as a survey sampling statistician are;

i. Non-sampling errors (item non-response)
ii. Ratio estimation
iii. Small area statistics and
iv. Statistical Education.

I actually worked briefly on small area statistics see **Oshungade** (1986, 1996a) but I finally settled for the first two in my thesis titled "Non-response and ratio estimation problems in sample survey".

Non-Response

In survey sampling, total survey error is the sum of the sampling and non-sampling errors. Kish (1965) puts the classification of survey errors as follows:

- Sampling error is the error due to the use of any sampling methods or techniques. It can come up through sampling frame biases, consistent sampling bias and constant statistical bias.
- Non sampling errors can be due to (i) errors of non-observation where non-coverage and non-response play major parts and (ii) errors of observation where field data collection and office data processing play the major roles.

In case of sampling errors, we use any of the following to obtain the level or accuracy of the survey methods:

- We can estimate the bias (the difference between the actual true value and estimated value).
- Standard error and mean square error made up of the variance and bias.
- Coefficient of variation to show the quality of your sampling strategy.

Non-response is an important source of non-sampling errors in censuses and survey sampling. According to O' Muircheartaigh et al (1999), no survey can ever attain 100% response. The most damaging is unit non-response where a sampling unit fails to give answer to any of the questionnaire or a battery of questionnaire. It is the failure to obtain information about members of the sample. It may be a failure to obtain any information from the member in the sample due to missing sampling unit or failure to obtain full data from the selected respondents.

The level of non-response in the sample survey can depend on the method of data collection, the subject of survey and the type of population involved. For example, the response rate in mailed questionnaire is always the lowest when compared with the other methods of collection of data such as 'face-to-face' interview. In telephone surveys, this depends on the populations involved. In Nigeria, the mobile phones recently introduced cannot be used for effective large sample surveys.

The major sources of total nonresponse in a sample survey include:

- Outright refusal. i.e. people in the sampling unit not granting interview or not returning the survey.
- Not at home at the time of call.
- Unsuitable for interview due to old age (infirmity) or language barrier.
- Moved away at the time of survey.
- Untraced or unknown address.
- Lost questionnaire.

Effects of Non-Response

The effects of non-response in a sample survey are principally found in the estimation of population parameters such as mean, total, proportion and ratio. Estimates obtained from the respondents alone are often biased because the estimates from non-respondents in many studies have been found to be significantly different from those of respondents.

This bias depends on the proportion of non-response and the difference in the average responses of the responding and non-responding units.

There is also the effect of underestimation of the population variance. If the non-response is not taken into consideration the variance would be less than when the non-response and response group R are considered using a subsample of the non-respondents N_R, hence, a lot of researches have been carried out to find solution to the problem of unit non-response and basically the researches centre on:

- Call back or recalls.
- One call.
- Replacements.
- Subsample of nonresponse and
- Randomized response.

Randomized Response Techniques (RRT) are used to avoid the concealment of information or evasive answers. The first to use RRT was Warner (1965). Others are Fisher, et al (1992) and Jarman (1997).

Usman and **Oshungade** (2012) device a two-way randomized response model (RRM) in stratification and use them to estimate HIV seroprevalence rates in a given population and compare results with the existing seroprevalence rates.

Randomized response techniques (RRT) guarantees the anonymity of respondents in surveys aimed at determining the frequency or proportion of stigmatic, embarrassing or criminal behaviour where direct techniques for data collection may induce respondents to answer or give false response. The motivation was to improve upon the existing RRMs as well as to apply them to estimate HIV seroprevalence rates. Our proposed two-way RRM in stratification for HIV seroprevalence surveys was relatively

more efficient than the Kim and Warde (2005) stratified estimatior for fixed sample size.

The chosen design parameter was 0.7, using the criteria of Quatember (2009) who derived the statistical properties of the standardized estimator for general probability sampling and privacy protection.

Our model was used to estimate the HIV seroprevalence rate in a sample population of adults 3,740 people aged 18 years and above attending a clinic in Kaduna, Nigeria, using a sample size of 550. Our findings revealed HIV seroprevalence rate, as estimated by the model stood at 6.1% with a standard error of 0.0082 and a 95% confidence interval of [4.5%, 7.7%]. These results are consistent with that of Nigeria sentinel survey (2003) conducted by NACA, USAID, and CDC which estimated the HIV seroprevalence in Kaduna state as 6.0%. Hence, our model serves as a new viable methods for HIV seroprevalence.

Notable researchers on recalls include Hansen and Hurwitz (1946), Deming (1953), Stephen and McCarthy (1958), and Dunkelberg and Day (1973). On one call, we have Politz and Simmons (1949-1950) on replacement Kish and Hess (1959), Ericson (1967), Rubin (1977), Singh and Sedransk (1978)..

Oshungade (1989c) conducted a research that summarised the sources and effects of non-response in sample surveys and considered mainly the estimation of mean and variance when non-response is involved in stratified sampling.

A subsample of non-response is assumed to exist in the different strata and the estimates obtained are compared with what would be obtained if one of the suggestions

made by Hansen et al (1946) is used. Optimum allocation to stratum and sample size is also obtained given the cost per stratum and the variance obtained earlier.

The Hansen and Hurwitz data were used to obtain the variance for

i. Only the results from the response stratum $V(\bar{Y}_{st})$ and
ii. The result when the response from subsample strata were used $V(\bar{Y}'_{st})$.

The results obtained are $V(\bar{Y}_{st}) = 0.5377$and $V(\bar{Y}'_{st}) = 0.6189$.

Note that for $V(\bar{Y}'_{st})$, the two strata are involved, hence what was obtained from the subsample is 0.0812. Therefore, it is better to include the subsample in our estimate because an unbiased value of the variance would be obtained.

Overall estimates may be biased if the non-respondents differ significantly from respondents. Our result will give some representation of the strata of non-respondents.

Oshungade (1991) gave an appraisal of non-response in Nigeria industrial survey and examined the problems of non-response in the large scale industrial survey conducted by FOS. The survey collects data on all the manufacturing establishments employing more than ten persons.

It was observed that the response value in the survey was not stable as they vary from year to year and do not follow any obvious pattern.

Oshungade (1989a) examined an adjustment for non-response in sample surveys through Bayesian approach and decision theory. The behaviour of Bayes method as a means of reducing nonresponse bias when predicting

categories for answers to a question that are more than two. This is a means of obtaining a correct answer for a question with more than two answers when such questions are omitted or respondent decided not to respond to it.

Conditions that are necessary for effective prediction are investigated and also looked at how decisions can be made from our predictions.

Item Non-Response

Item non-response occurs when some but not all the required information is collected from a sample unit.
It can occur because of the following reasons:

i. A sample unit may refuse or be unable to answer a particular question e.g. questions on income, drug addiction and crime.
ii. The interviewer may fail to ask the question or to record the answer. This is item non-response due to omission.
iii. A record is inconsistent in the sense that it does not satisfy an editing check based on logical or empirical grounds.

There are four common methods by which item non-response can be tackled after all possible follow ups and call backs of respondents has been done. These are:

i. To ignore all the records with missing values or delete records with missing items if a low number of cases is missing.
ii. Publish the missing item as unknown and put unknown values as one category.
iii. Adjust or reweigh each estimate by ignoring the records with relevant missing items in each case.

iv. Fill in plausible and consistent values for the missing items. This is called imputation and is the most common procedure for dealing with item non-response.

The methods (i) to (iii) above will lead to biased estimates and (i) in particular may lead to unnecessary rejection of available data. Imputation is a tool for combating item non-response when the objectives are to:

(i) Reduce biases in survey estimate arising from missing
(ii) Make analyses easier to conduct and present results
(iii) Give consistent results for the survey.

Values assigned to the missing items allow analyses to be made as if the data were complete. To achieve these objectives, a good imputation procedure must be capable of:

(i) imputing values which are consistent with the edits given that the non-missing data satisfy the edit,
(ii) reducing the non-response bias and preserving the relationship between items as far as possible,
(iii) being set up ahead of time for future use and
(iv) being evaluated in terms of impacts on the bias and precisions of the estimates.

Effects of Item Non-Response

Item non-response without the imputed values gives a biased estimate of the mean, an underestimation of the

variance and a short confidence interval for the estimate. Hertel(1976) has observed that a sample mean solution does not alter the mean for the sample but does reduce the variance for treated items. Sande (1982) noted that:

(i) the usual estimates of variance are inadequate since they do not include the errors due to imputations,
(ii) with hot deck methods , the variance of the estimate in some cases is known to be higher than the variance of the usual simple random sampling.

Sande in the same paper discussed the evaluation of effects of imputation on the Survey estimate and warned that extent of imputation in the variables should be monitored and noted.

Methods of Handling Item Non-Response

There are two approaches to imputation:

(i) To impute the best prediction value for each item non-response or
(ii) To impute by randomly drawing from the respondents values in order to preserve the variability of the observation.

These two approaches lead to three basic methods of imputation, which are

(i) use of class mean called by Hertel (1976) as sample or group means, others call it mean value imputation
(ii) use of regression equations and
(iii) use of hot Deck techniques.

The **class mean** method assigns the average or mean value of the class or group to the missing item. The sample is divided into classes or groups on the basis of the responses to other relevant items. For any particular item for which non-response exists, the mean for the respondent is assigned to all non-respondents in that class.

The **Regression** procedure allows the use of regression equations to predict the missing items by using the responses to other items in the questionnaire as predicting variables, the values imputed are the predicted values from the regression equations. Hawkins (1975) and Ford (1976) have used regression procedures to estimate values for missing items.

Oshungade (1986) demonstrated an example of imputation on continuous data by devising a method known as the Percentage Change method, to estimate the item non-response in a study of the population living in different enumeration areas (EAS). Small area census data for 83 enumeration districts of Colchester are used to estimate the values for the missing data on population total for some of the EAS. The enumeration districts had identical boundaries at the 1971 and 1981 population census.

In this study, we collected four sets of data from the Social Science Research Council (SSRC) Archive based in the University of Essex. These are data on population total P, birth B, school S, and economically archive population E in the two censuses.

P is the variable of interest Y and $B\ is\ X_1, S\ is\ X_2\ and\ E\ is\ X_3$ are the symptomatic or independent variables.

My concern was to estimate the changes in the total population of these identical enumeration areas between the two periods of 1971 and 1981. If we take 1971 as the base year and denote its population as P_{71} and that of 1981 as P_{81} the percentage change is

$$\left(\frac{P_{81} - P_{71}}{P_{71}}\right) 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (1)$$

Similarly we also calculated the percentage changes for B, S and E as in (1)

A sample of 30 EAS out of the 83 is used and the results were later used for all the 83.

A multiple regression is formed from the percentage changes in Y, $X_1, X_2, and\ X_3$. The percentage changes relative to the 1st year 1971 is used in order to put the changes in the independent and dependent variables on a common scale.

We also make use of a sample of the enumeration areas to give a sample regression which has been called the sample-percentage change regression.

Different variations of the percentage change method depends on the use of raw data, regression of the sample and transformation of the data, hence the following variations are obtained. These are;

i. Percentage change of raw data (PCRD).
ii. Sample regression (SR).
iii. Percentage change of proportion (PCP).
iv. Percentage change of square root transformed data (PCSQR).
v. Percentage change of logarithm transformed data (PCLOG).

The two transformation (iv) and (v) are commonly used in practice.

Sample regression (SR) method is a procedure by which the sample survey data is combined with symptomatic information X to obtain the estimate of Y. Essentially, the characteristics of this method are the data used to get the values of the regression coefficients exclusively post censal and current sample data for the variable of interest as the dependent variable, hence the method has avoided the problem of changes in the structural relations. In this study, I have used complete censuses collected in a sample of enumeration areas.

In **Oshungade** (1986), the results for the first four methods, that is, PCRD, SR, PCP, and PCSQR using multiple regression equation are presented in table 1.

To check the accuracy of the different variations of percentage change, we :

i. compared the number of percentage deviation by each method. In this case, the number of enumeration areas with more than 10% deviation or from $\frac{\hat{P}_{h_2}-P_{h_2}}{P_{h_2}}$ determines the direction of the deviation.

ii. calculated the mean of the absolute percentage deviation. i.e.; $P=\frac{1}{83}\sum_{h=1}^{83}\left|\frac{\hat{p}_{h_2}-p_{h_2}}{p_{h_2}}\right|$ and

iii. compared the scatter diagrams of the actual and estimated values and observe the degree of association.

Table 1, gives the summary of the performance of the different variations of percentage change method for estimating value for the missing population values.