Annals of African Medicine
Vol. 5, No. 3; 2006: 158 – 161

OPINION

Statistics in Medical Research: Misuse of Sampling and Sample Size Determination

[1]T. Dahiru, [1]A. Aliyu and [2]T. S. Kene

Department of Community Medicine, Ahmadu Bello University and Ahmadu Bello University Teaching Hospital, Zaria, Nigeria
Reprint requests to: Dr. T. Dahiru, Department of Community Medicine, Ahmadu Bello University, Zaria, Nigeria. E-mail: tukurdahiru@yahoo.com

Abstract
One of the major issues in planning a research is the decision as to how large a sample and the method to be employed to select the estimated sample in order to meet the objective of the research. Sampling is an essential tool for research in medicine. A good number of the medical literature while reporting their sampling method go by stating that the sample was collected by random sampling and no further explanation as how the sample has been drawn as if the word random is generic to all the known sampling methods. The aim of this paper is to sensitise our researchers on the importance of proper sampling and sample size determination. Using a few examples we demonstrated that investigators adhere poorly to the statistical precondition of simple random sampling, have poor understanding of simple random technique, and quite a number of estimated sample sizes were bloated without appreciating the implications of that. Finally, we recommended, among others that investigators should consult biostatisticians at the design stages of their research work and a competent biostatistician should review any article containing even the most elementary statistical procedure.

Key words: Statistics, research, sampling

Résumé
L'un des questions principales d'un recherché pour prendre un décision, comment le grande échantillon et la méthode d'être employé se séléctionner un échantillon éstimé afin d'atteindre le but d'un recherchre. Échantillonnage est un utile essentiale à la recherche en médicine. Le mieux nombre de la litterature de la médicine alors que la reportage de leur échantillonnage méthode noté que l'échantillonnage avait collecté par l'échatillonnage aléatoire et pas explication davantage comment l'échantillon à été attirer si le mot aléatoire est générique aux savoir d'aléatoires échantillonnage. Le but de cette exposé est pour sensibilser notre rechercheur sur l'importance d'échantillonnage proper et détérmin la sauter d'aléatoire échatillonnage. Nous avons utiliser quelque examples prouve que les investigateurs mal obeir en la précondition statistiques d'aléatoire échatillonnage simple, ils ont mal comprend la technique d'aléatoire simple, et un bon nombre d'éstimer les sauters simple était hypertrophié sans appréciation de l'implication. Finallement, nous avons reccomandé que entre autres investigateurs devraient consulter les biostatisciens à l'étage d'ébaucher leur travail et un compétent biostatisticien devrais en revue l'article contienir même le plus procédure élémentaire statistique.

Mot clés: Statistique, recherchre, séléctionner

One of the major issues in planning a research is the decision as to how large a sample and the method to be employed to select the estimated sample in order to meet the objective of the research. Sampling is an essential tool for scientific investigation and research (not only in medical field but in marketing, agriculture, economics, biological sciences etc). Small-scale investigation over small area and population can conveniently be based on simple random sampling. However, for a large widespread population, complex probability sampling might be employed.

The connection between sample size requirements in medical investigations (e.g. cross-sectional, case control, cohort, randomised control trials, field trials) and inference being drawn from results of such

investigations do not seem to be generally appreciated. It is very important and mandatory that sample sizes are determined based on the study design and the objectives of the study. Failure to calculate size of sample with reference to particular study design may lead to incorrect results and conclusions. Of all the statistical procedures dealt with in medical inquiries in this part of the world, perhaps, sampling and sample size determination seemed to be most abused. A good number of the medical literature while reporting their sampling method go by stating that the sample was collected by random sampling and no further explanation as how the sample has been drawn as if the word random is the generic to all the known sampling methods.

If one states that a sample was drawn using simple random technique, one has to explain first the sampling frame and how the sample was drawn. A sampling frame is a complete enumeration of the sampling units in the study population, which may be a list, directory, map, arial configuration, while the sampling unit may be an individual, a household or a school. For example, if it is a study in a village (with a population of say, 500) and the objective is to determine the prevalence of some unusual events or factors among the villagers, the selection unit ideally should be individuals residing in the village. In this case, the list of the names of all inhabitants will be the reference sampling frame. But there are situations where the sampling frame could not be worked out so easily. Taking example of a similar study covering a state, it is almost impossible to draw a list of all inhabitants residing in the state. So here, simple random sampling could not be appropriate; one has to make use of a more simple approach.

This had lead to development of various and more complex techniques of sampling. Here, the appropriate technique is multi-stage sampling i.e. sampling stage-by-stage starting from selection of Local Government Area (LGA), then selection of villages from the selected LGA, and then selection of families / households in the selected villages. Here, the appropriate sampling frame is not the list of individuals but a map of state showing all Local Government Areas in the state, the list of all villages in selected Local Government Areas and finally the list of families / households in the selected villages.

Further, most of the estimates made from so-called random sampling are mentioned in the result without their standard deviation. This is an important statistical oversight. Sampling involves two important processes:

1. Selection (sampling) process, which describes the method as to how some units from the population are included in the sample.
2. Estimation of precision i.e. deviation of sample estimates (means and standard deviations).

Further, to wade off queries from sampling methodology in a medical inquiry, the following points should be mentioned briefly in the report:

1. Description of the nature and content of the population i.e. its individual units, size, time reference.
2. Description of the sampling framework, which the sample is drawn.
3. Consideration of issues regarding the decision of sample size.

The sampling theory affords us the basis for determining the size of sample. The necessary steps involved are:

1. ***Specification of a precision level:*** A decision on the tolerable limits of errors is made, i.e. the researcher makes a statement that it does not matter if his sample estimate does not differ from true population value by a certain amount. For example, suppose a Paediatrician plans a study to estimate the population of malnourished children in a village and suppose that the true proportion of malnourished children is 10%. He is satisfied if his estimate does not differ from true value of 10% by 5% i.e. he is okay with the result of his study if his estimate is within 9.5% to 10.5% (i.e. 10±0.5%).
2. ***Specification of level of confidence:*** This is the degree of uncertainty or probability that a sample value lies outside a stated limits (i.e. 10 ± 0.5) %. Suppose this measure is 5%, the investigator has to accept the unlikely situation of 1 in 20 cases that the sample result falls aside the desired limit; and if it is 1%, then the chance that the sample result falls outside the desired limits in 1 in 400. However, by convention, the mostly used confidence levels are 5% and 1%; but nothing stops the investigator from tolerating 10%, 2.5% etc.

The most important consideration in sampling is the planning of appropriate technique to be used considering the situation on ground; and determination of sample size adequate to ensure confidence on the inference made out of the results of the study within the limitations under which the study (sampling) was conducted.

In order to carry home our message concerning poor understanding of sample estimation and sampling techniques, we reviewed two articles that portray the misuse of sampling and sampling methodology. A caution here is that our medical journals are replete with such misuses and these two articles are only for illustrative purposes.

Review of some published research studies

The following original articles were reviewed and special attention given to the title, objective and sampling techniques (including sample size determined). This list is not meant to be exhaustive; rather it is intended to serve as exposition to several of such abuses of sampling and sampling techniques.

Prevalence of *Trichomonas vaginalis* infection among students of tertiary institutions in Imo State, Nigeria: [1]

The objective of the study was to determine the prevalence of *Trichomonas vaginalis* infection among students of tertiary institutions in Imo State, Nigeria. The estimated sample size was 2419 (510 males and 1909 females) from three tertiary institutions in Imo State. The subjects were sampled as follows: "A random selection of the students was made to ensure

all faculties, residential hostels, and off-campus students were covered in the study". The following questions are pertinent:

1. One will wonder about the usage of "random selection" in the sampling method applied in this study. This "random selection", was it used in statistical sense or in literal sense? If it means random selection in statistical sense i.e. each possible unit (subject) has a known and equal chance of been selected, then the authors failed to tell us about the sampling frame, which is a precondition for random sampling. Further, in order to ensure that all the faculties, residential hostels and off-campus students were covered in the study, the authors must employ a probabilistic sampling technique so as to give every student in his/her location equal opportunity of being selected in the study; and simple random sampling is obviously not appropriate. It is possible to employ simple random sampling and at the end one discovers that all the selected individuals are drawn from one particular location or disproportionately selected from particular location(s). (This is a known flaw in utilising simple random technique). Therefore, one can think of stratified sampling technique.
2. Again one wonders how a sample size of 2419 students was arrived at. Using a random selection and based on the study design (which is a cross-sectional) one can safely assume they used the following formula to calculate the size of sample and have utilised a sampling frame to select their subjects.[2]

$$N = \frac{Z_{\alpha/2} P(1-P)}{d^2}$$

If this is true, then by our calculation, the correct size of sample is 288 students (at prevalence of 24.7%, [1]at 95% confidence level and 5% precision). One can argue that the estimated size of sample is only a minimum number required to make a valid conclusion and generalisation. They enrolled 2491 students, which is about 8 times the required minimum sample. That is true, but the snag here is that the sampling technique employed, which going by the author(s) description was not simple random (and not probability-based) and therefore they cannot make any valid generalisation to any external population of similar characteristics (i.e. students). Further, large sample size can prove anything.

Emergency contraception: a survey of women's knowledge and attitude in a rural setting in Northern Nigeria (Sahel Medical Journal 1999; 2: 73 – 76): The objective of the survey was to assess the women's knowledge and attitude in relation to emergency contraception. The investigators distributed 250 questionnaires out of which only 124 were returned completed. The method of administration of the questionnaires was as follows: "Anonymous self-completed semi-structured questionnaires were administered to randomly selected women resident in semi-urban towns in Gwoza Local Government Area (LGA) of Borno State between January and December 1997". Again the following questions need to be answered:

1. How were the women "randomly selected"? A necessary precondition for random selection is the use of sampling frame, which in this case might be the list of all towns in Gwoza LGA, the list of all households in Gwoza LGA, and list of all women residing in the households. The list of all towns in Gwoza LGA is obtainable but the lists of all households and all women might proved difficult, especially the list of all women. Furthermore, they did not tell us the number of towns selected in the LGA.
2. How was a sample of 250 or 124 respondents arrived at? Using the prevalence of the knowledge, attitude and practice of modern contraception in Nigeria, which they quoted in their introduction as 39% the estimated sample size, is 366 at 95% confidence level and 5% precision level, which is almost three times the 124 respondents employed.

This formula for determining size of sample requires knowledge on population parameters, e.g. mean, variation and proportion. It is therefore necessary that when using this formula an advanced knowledge of the population parameters are required and these can be roughly determined by various methods. The following are few ways, which are practicable and acceptable:

1. ***By pilot surveys:*** This is a small scale preliminary survey arrived at the estimating mean, proportion (or prevalence) etc.
2. ***By use of results of previous surveys:*** Results of previous surveys carried out can be utilised to obtain acceptable population parameters. Most investigators in this part of the world undertaking an enquiry and utilising the formula under consideration use results of previous investigation to determine the size of the sample.
3. ***By intelligent guess:*** An experienced investigator should be able to make a realistic estimate of the population parameter under question. The investigator should be acquainted with population structure and also be conversant with topic of enquiry. For example, a nutritionist should be able from experience to make an intelligent guess about the proportion of children that are malnourished in a certain geographical areas fitting all the factors that have influence on nutritional status under consideration.

The formula under consideration is used to determine the size of sample to be applied in finding the true population proportion P_O of a category or success by simple random sampling. Therefore, if P is the sample estimate of Po, and the entire precision and level of confidence being d and α respectively, then n the minimum size of sample required to give P, n is given by the formula:

$$n = Z^2 \frac{poqo}{d^2}$$

where Z_α is standard normal deviate corresponding to(100% α / 2%) and $q_o = 1 - p_o$ for a population whose size is known (i.e. finite), the

adjusted formula is:

$$n = \frac{Z^2 a/2\ p_o q_o}{Nd2 + Z^2\ p^o qo}$$

where N is the size of the population.

One can make the following conclusions:

1. Majority of investigators in medical sciences do not strictly adhere to the statistical criteria of simple random sampling.
2. For every study design there exist an appropriate technique for sampling and sample size determination. This fact is barely understood and applied appropriately by medical researchers.
3. Sample selection, most at times employ a technique of random sampling and on further scrutiny always reveals confused understanding of the technique used.
4. Most of the estimated sample sizes were well above the minimum required, a few are underestimated. This 'inflated' sample size results from inappropriate use of technique leads only to waste of resources without proportionate increase in the confidence of the inferences made considering the targeted precision level. Remember, large sample size can prove anything and small sample size can prove nothing. Further, it underscores the credibility of the investigators' knowledge of statistics and throws doubts on the originality of the work.

We recommend that the teaching of biostatistics be intensified to students of health sciences and allied courses, both at undergraduate and postgraduate levels. Further, investigators should consult biostatisticians at the design stages of their research work. Any article containing even the most elementary statistical procedure should be reviewed by a competent biostatistician. Finally, wherever possible editorial boards of medical journals should include a biostatistician as an associate editor.

References

Anosike JC. *Trichomonas* among students of higher institutions in Nigeria. Applied Parasitology 1993; 3: 19-25

Lemeshow S, Hosmer DW, Klar J, Lwanga S. Adequacy of sample size in health studies. World Health Organisation, Geneva, 1990; 1