Statistics without tears: Populations and samples

Amitav Banerjee,
Suprakash Chaudhury[1]

Department of Community Medicine, D Y Patil Medical College, Pune, and
[1]Department of Psychiatry, RINPAS, Kanke, Ranchi, India

Address for correspondence:
Dr. Amitav Banerjee, Department of Community Medicine,
D Y Patil Medical College,
Pune – 18, India.
E-mail: amitavb@gmail.com

ABSTRACT

Research studies are usually carried out on sample of subjects rather than whole populations. The most challenging aspect of fieldwork is drawing a random sample from the target population to which the results of the study would be generalized. In actual practice, the task is so difficult that some sampling bias occurs in almost all studies to a lesser or greater degree. In order to assess the degree of this bias, the informed reader of medical literature should have some understanding of the population from which the sample was drawn. The ultimate decision on whether the results of a particular study can be generalized to a larger population depends on this understanding. The subsequent deliberations dwell on sampling strategies for different types of research and also a brief description of different sampling methods.

Keywords: Methods, population, sample

Research workers in the early 19th century endeavored to survey entire populations. This feat was tedious, and the research work suffered accordingly. Current researchers work only with a small portion of the whole population (a sample) from which they draw inferences about the population from which the sample was drawn.

This inferential leap or generalization from samples to population, a feature of inductive or empirical research, can be full of pitfalls. In clinical medicine, it is not sufficient merely to describe a patient without assessing the underlying condition by a detailed history and clinical examination. The signs and symptoms are then interpreted against the total background of the patient's history and clinical examination including mental state examination. Similarly, in inferential statistics, it is not enough to just describe the results in the sample. One has to critically appraise the real worth or representativeness of that particular sample. The following discussion endeavors to explain the inputs required for making a correct inference from a sample to the target population.

Access this article online

Quick Response Code:

Website: www.industrialpsychiatry.org

DOI: 10.4103/0972-6748.77642

TARGET POPULATION

Any inferences from a sample refer only to the defined population from which the sample has been properly selected. We may call this the target population. For example, if in a sample of lawyers from Delhi High Court it is found that 5% are having alcohol dependence syndrome, can we say that 5% of all lawyers all over the world are alcoholics? Obviously not, as the lawyers of Delhi High Court may be an institution by themselves and may not represent the global lawyers' community. The findings of this study, therefore, apply only to Delhi High Court lawyers from which a representative sample was taken. Of course, this finding may nevertheless be interesting, but only as a pointer to further research. The data on lawyers in a particular city tell us nothing about lawyers in other cities or countries.

POPULATIONS IN INFERENTIAL STATISTICS

In statistics, a population is an entire group about which some information is required to be ascertained. A statistical population need not consist only of people. We can have population of heights, weights, BMIs, hemoglobin levels, events, outcomes, so long as the population is well defined with explicit inclusion and exclusion criteria. In selecting a population for study, the research question or purpose of the study will suggest a suitable definition of the population to be studied, in terms of location and restriction to a particular age group, sex or occupation.

The population must be fully defined so that those to be included and excluded are clearly spelt out (inclusion and exclusion criteria). For example, if we say that our study populations are all lawyers in Delhi, we should state whether those lawyers are included who have retired, are working part-time, or non-practicing, or those who have left the city but still registered at Delhi.

Use of the word *population* in epidemiological research does not correspond always with its demographic meaning of an entire group of people living within certain geographic or political boundaries. A population for a research study may comprise groups of people defined in many different ways, for example, coal mine workers in Dhanbad, children exposed to German measles during intrauterine life, or pilgrims traveling to *Kumbh Mela* at Allahabad.

GENERALIZATION (INFERENCES) FROM A POPULATION

When generalizing from observations made on a sample to a larger population, certain issues will dictate judgment. For example, generalizing from observations made on the mental health status of a sample of lawyers in Delhi to the mental health status of all lawyers in Delhi is a formalized procedure, in so far as the errors (sampling or random) which this may hazard can, to some extent, be calculated in advance. However, if we attempt to generalize further, for instance, about the mental statuses of all lawyers in the country as a whole, we hazard further pitfalls which cannot be specified in advance. We do not know to what extent the study sample and population of Delhi is typical of the larger population – that of the whole country – to which it belongs.

The dilemmas in defining populations differ for descriptive and analytic studies.

POPULATION IN DESCRIPTIVE STUDIES

In descriptive studies, it is customary to define a *study population* and then make observations on a sample taken from it. Study populations may be defined by geographic location, age, sex, with additional definitions of attributes and variables such as occupation, religion and ethnic group.[1]

Geographic location

In field studies, it may be desirable to use a population defined by an administrative boundary such as a district or a state. This may facilitate the co-operation of the local administrative authorities and the study participants. Moreover, basic demographic data on the population such as population size, age, gender distribution (needed for calculating age- and sex-specific rates) available from census data or voters' list are easier to obtain from administrative headquarters. However, administrative boundaries do not always consist of homogenous group of people. Since it is desirable that a modest descriptive study does not cover a number of different groups of people, with widely differing ways of life or customs, it may be necessary to restrict the study to a particular ethnic group, and thus ensure better genetic or cultural homogeneity. Alternatively, a population may be defined in relation to a prominent geographic feature, such as a river, or mountain, which imposes a certain uniformity of ways of life, attitudes, and behavior upon the people who live in the vicinity.

If cases of a disease are being ascertained through their attendance at a hospital outpatient department (OPD), rather than by field surveys in the community, it will be necessary to define the population according to the so-called *catchment area* of the hospital OPD. For administrative purposes, a dispensary, health center or hospital is usually considered to serve a population within a defined geographic area. But these catchment areas may only represent in a crude manner with the actual use of medical facilities by the local people. For example, in OPD study of psychiatric illnesses in a particular hospital with a defined catchment area, many people with psychiatric illnesses may not visit the particular OPD and may seek treatment from traditional healers or religious leaders.

Catchment areas depend on the demography of the area and the *accessibility* of the health center or hospital. Accessibility has three dimensions – physical, economic and social.[2] Physical accessibility is the time required to travel to the health center or medical facility. It depends on the topography of the area (e.g. hill and tribal areas with poor roads have problems of physical accessibility). Economic accessibility is the paying capacity of the people for services. Poverty may limit health seeking behavior if the person cannot afford the bus fare to the health center even if the health services may be free of charge. It may also involve absence from work which, for daily wage earners, is a major economic disincentive. Social factors such as caste, culture, language, etc. may adversely affect accessibility to health facility if the treating physician is not conversant with the local language and customs. In such situations, the patient may feel more comfortable with traditional healers.

Ascertainment of a particular disease within a particular area may be incomplete either because some patient may seek treatment elsewhere or some patients do not seek treatment at all. Focus group discussions (qualitative study) with local people, especially those residing away from the health center, may give an indication whether serious underreporting is occurring.

When it is impossible to relate cases of a disease to a population, perhaps because the cases were ascertained through a hospital with an undefined catchment area, *proportional morbidity rates* may be used. These rates have been widely used in cancer epidemiology where the number of cases of one form of cancer is expressed as a proportion of the number of cases of all forms of cancer among patients attending the same hospital during the same period.

POPULATIONS IN ANALYTIC STUDIES

Case control studies

As opposed to descriptive studies where a study population is defined and then observations are made on a representative sample from it, in case control studies observations are made on a group of patients. This is known as the *study group,* which usually is not selected by sampling of a defined larger group. For instance, a study on patients of bipolar disorder may include every patient with this disorder attending the psychiatry OPD during the study period. One should not forget, however, that in this situation also, there is a hypothetical population consisting of all patients with bipolar disorder in the universe (which may be a certain region, a country or globally depending on the extent of the generalization intended from the findings of the study). Case control studies are often carried out in hospital settings because this is more convenient and accessible group than cases in the community at large. However, the two groups of cases may differ in many respects. At the outset of the study, it should be deliberated whether these differences would affect the external validity (generalization) of the study. Usually, analytic studies are not carried out in groups containing atypical cases of the disorder, unless there is a special indication to do so.

Populations in cohort studies

Basically, cohort studies compare two groups of people (cohorts) and demonstrate whether or not there are more cases of the disease among the cohort exposed to the suspected cause than among the cohort not exposed. To determine whether an association exists between positive family history of schizophrenia and subsequent schizophrenia in persons having such a history, two cohorts would be required: first, the exposed group, that is, people with a family history of mental disorders (the suspected cause) and second, the unexposed group, that is, people without a family history of mental disorders. These two cohorts would need to be followed up for a number of years and cases of schizophrenia in either group would be recorded. If a positive family history is associated with development of schizophrenia, then more cases would occur in the first group than in the second group.

The crucial challenges in a cohort study are that it should include participants exposed to a particular cause being investigated and that it should consist of persons who can be followed up for the period of time between exposure (cause) and development of the disorder. It is vital that the follow-up of a cohort should be complete as far as possible. If more than a small proportion of persons in the cohort cannot be traced (loss to follow-up or attrition), the findings will be *biased*, in case these persons differ significantly from those remaining in the study.

Depending on the type of exposure being studied, there may or may not be a range of choice of cohort populations exposed to it who may form a larger population from which one has to select a study sample. For instance, if one is exploring association between occupational hazard such as job stress in health care workers in intensive care units (ICUs) and subsequent development of drug addiction, one has to, by the very nature of the research question, select health care workers working in ICUs. On the other hand, cause effect study for association between head injury and epilepsy offers a much wider range of possible cohorts.

Difficulties in making repeated observations on cohorts depend on the length of time of the study. In correlating maternal factors (pregnancy cohort) with birth weight, the period of observation is limited to 9 months. However, if in a study it is tried to find the association between maternal nutrition during pregnancy and subsequent school performance of the child, the study will extend to years. For such long duration investigations, it is wise to select study cohorts that are firstly, not likely to migrate, cooperative and likely to be so throughout the duration of the study, and most importantly, easily accessible to the investigator so that the expense and efforts are kept within reasonable limits. Occupational groups such as the armed forces, railways, police, and industrial workers are ideal for cohort studies. Future developments facilitating record linkage such as the Unique Identification Number Scheme may give a boost to cohort studies in the wider community.

SAMPLES

A sample is any part of the fully defined population. A syringe full of blood drawn from the vein of a patient is a sample of all the blood in the patient's circulation at the moment. Similarly, 100 patients of schizophrenia in a clinical study is a sample of the population of schizophrenics, provided the sample is properly chosen and the inclusion and exclusion criteria are well defined.

To make accurate inferences, the sample has to be representative. A representative sample is one in which each and every member of the population has an equal and mutually exclusive chance of being selected.

Sample size

Inputs required for sample size calculation have been dealt from a clinical researcher's perspective avoiding the use of intimidating formulae and statistical jargon in an earlier issue of the journal.[1]

Target population, study population and study sample

A population is a complete set of people with a specialized set of characteristics, and a sample is a subset of the population. The usual criteria we use in defining population are geographic, for example, "the population of Uttar Pradesh". In medical research, the criteria for population may be clinical, demographic and time related.

a) Clinical and demographic characteristics define the target population, the large set of people in the world to which the results of the study will be generalized (e.g. all schizophrenics).
b) The study population is the subset of the target population available for study (e.g. schizophrenics in the researcher's town).
c) The study sample is the sample chosen from the study population.

METHODS OF SAMPLING

Purposive (non-random samples)

- Volunteers who agree to participate
- Snowball sample, where one case identifies others of his kind (e.g. intravenous drug users)
- Convenient sample such as captive medical students or other readily available groups
- Quota sampling, at will selection of a fixed number from each group
- Referred cases who may be under pressure to participate
- Haphazard with combination of the above methods

Non-random samples have certain limitations. The larger group (target population) is difficult to identify. This may not be a limitation when generalization of results is not intended. The results would be valid for the sample itself (internal validity). They can, nevertheless, provide important clues for further studies based on random samples. Another limitation of non-random samples is that statistical inferences such as confidence intervals and tests of significance cannot be estimated from non-random samples. However, in some situations, the investigator has to make crucial judgments. One should remember that random samples are the means but representativeness is the goal. When non-random samples are representative (compare the socio-demographic characteristics of the sample subjects with the target population), generalization may be possible.

Random sampling methods

Simple random sampling

A sample may be defined as random if every individual in the population being sampled has an equal likelihood of being included. Random sampling is the basis of all good sampling techniques and disallows any method of selection based on volunteering or the choice of groups of people known to be cooperative.[3]

In order to select a simple random sample from a population, it is first necessary to identify all individuals from whom the selection will be made. This is the sampling frame. In developing countries, listings of all persons living in an area are not usually available. Census may not catch nomadic population groups. Voters' and taxpayers' lists may be incomplete. Whether or not such deficiencies are major barriers in random sampling depends on the particular research question being investigated. To undertake a separate exercise of listing the population for the study may be time consuming and tedious. Two-stage sampling may make the task feasible.

The usual method of selecting a simple random sample from a listing of individuals is to assign a number to each individual and then select certain numbers by reference to random number tables which are published in standard statistical textbooks. Random number can also be generated by statistical software such as EPI INFO developed by WHO and CDC Atlanta.

Systematic sampling

A simple method of random sampling is to select a *systematic sample* in which every n^{th} person is selected from a list or from other ordering. A systematic sample can be drawn from a queue of people or from patients ordered according to the time of their attendance at a clinic. Thus, a sample can be drawn without an initial listing of all the subjects. Because of this feasibility, a systematic sample may have some advantage over a simple random sample.

To fulfill the statistical criteria for a random sample, a systematic sample should be drawn from subjects who are randomly ordered. The starting point for selection should be randomly chosen. If every fifth person from a register is being chosen, then a random procedure must be used to determine whether the first, second, third, fourth, or fifth person should be chosen as the first member of the sample.

Multistage sampling

Sometimes, a strictly random sample may be difficult to obtain and it may be more feasible to draw the required number of subjects in a series of stages. For example, suppose we wish to estimate the number of CATSCAN examinations made of all patients entering a hospital in a given month in the state of Maharashtra. It would be quite tedious to devise a scheme which would allow the total population of patients to be directly sampled. However, it would be easier to list the districts of the state of Maharashtra and randomly draw a sample of these districts. Within this sample of districts, all the hospitals would then be listed by name, and a random sample of these can be drawn. Within each of these hospitals, a sample of the patients entering in the given month could be chosen randomly for observation and recording. Thus, by stages, we draw the required sample. If indicated, we can introduce some element of stratification at some stage (urban/rural, gender, age).

It should be cautioned that multistage sampling should only be resorted to when difficulties in simple random sampling are insurmountable. Those who take a simple random sample of 12 hospitals, and within each of these hospitals select a random sample of 10 patients, may believe they have selected 120 patients randomly from all the 12 hospitals. In statistical sense, they have in fact selected a sample of 12 rather than 120.[4]

Stratified sampling

If a condition is unevenly distributed in a population with respect to age, gender, or some other variable, it may be prudent to choose a stratified random sampling method. For example, to obtain a stratified random sample according to age, the study population can be divided into age groups such as 0–5, 6–10, 11–14, 15–20, 21–25, and so on, depending on the requirement. A different proportion of each group can then be selected as a subsample either by simple random sampling or systematic sampling. If the condition decreases with advancing age, then to include adequate number in the older age groups, one may select more numbers in older subsamples.

Cluster sampling

In many surveys, studies may be carried out on large populations which may be geographically quite dispersed. To obtain the required number of subjects for the study by a simple random sample method will require large costs and will be cumbersome. In such cases, clusters may be identified (e.g. households) and random samples of clusters will be included in the study; then, every member of the cluster will also be part of the study. This introduces two types of variations in the data – between clusters and within clusters – and this will have to be taken into account when analyzing data.

Cluster sampling may produce misleading results when the disease under study itself is distributed in a clustered fashion in an area. For example, suppose we are studying malaria in a population. Malaria incidence may be clustered in villages having stagnant water collections which may serve as a source of mosquito breeding. In villages without such water stagnation, there will be lesser malaria cases. The choice of few villages in cluster sampling may give erroneous results. The selection of villages as a cluster may be quite unrepresentative of the whole population by chance.[5]

Lot quality assurance sampling

Lot quality assurance sampling (LQAS), which originated in the manufacturing industry for quality control purposes, was used in the nineties to assess immunization coverage, estimate disease prevalence, and evaluate control measures and service coverage in different health programs.[6] Using only a small sample size, LQAS can effectively differentiate between areas that have or have not met the performance targets. Thus, this method is used not only to estimate the coverage of quality care but also to identify the exact subdivisions where it is deficient so that appropriate remedial measures can be implemented.

CONCLUSION

The choice of sampling methods is usually dictated by feasibility in terms of time and resources. Field research is quite messy and difficult like actual battle. It may be sometimes difficult to get a sample which is truly random. Most samples therefore tend to get biased. To estimate the magnitude of this bias, the researcher should have some idea about the population from which the sample is drawn. In conclusion, the following quote cited by Bradford Hill[4] elegantly sums up the benefit of random sampling:

> *…The actual practice of medicine is virtually confined to those members of the population who either are ill, or think they are ill, or are thought by somebody to be ill, and these so amply fill up the working day that in the course of time one comes unconsciously to believe that they are typical of the whole. This is not the case. The use of a random sample brings to light the individuals who are ill and know they are ill but have no intention of doing anything about it, as well as those who have never been ill, and probably never will be until their final illness. These would have been inaccessible to any other method of approach but that of the random sample…*
>
> *J. H. Sheldon*

REFERENCES

1. Banerjee A, Chaudhury S, Singh DK, Banerjee I, Mahato AK, Haldar S. Statistics without tears – inputs for sample size calculations. Indian Psychiatr Jr 2007;16:150-2.
2. Barker DJ, Hall A J. Practical epidemiology. 4th ed. ELBS;

1994. p. 30-43.
3. Indrayan A. Basic methods of medical research. India: AITBS Publishers; 2008. p. 116.
4. Hill AB, Hill ID. Principles of medical statistics. 12th ed. New Delhi: B I Publications Pvt Ltd; 1993. p. 12-22.
5. WHO. Health Research Methodology. 2nd ed. WHO Regional office of Western Pacific, Manila; 2001. p. 82
6. WHO. Monitoring immunization services using the lot quality technique. Global program for vaccines and immunization. Vaccine research and development. Geneva: WHO; 1996.

How to cite this article: Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. Ind Psychiatry J 2010;19:60-5.

Source of Support: Nil. **Conflict of Interest:** None declared.

Alcohol: An Autobiography

I am alcohol;

I come in all forms and shapes.
I am the product of wheat, barley and grapes.

I pervade in all aspects of society.
Rum, whiskey, beer, vodka are all my variety.

Young, old, weak and bold, I lure them all.
I am omnipresent, be it shack or mall.

Some like me country way, some like me exotic;
I make anxious comfortable and the composed erotic.

They tell me I have an intoxicating effect ct on all.
I get the party going when the night fall.

I gel with all; soda, coke or water;
Sometime I am good as a tip, some time as barter.

But many don’t know that I have a flipside.
Let me be my own critics as I have nothing to hide.

I wipe people’s tears and drown their sorrow;
Least they realize I scar their tomorrow.

I make the jittery bold and the meek quite impulsive;
Violence, rape etc, their crimes are repulsive.

I make good relations and my company is calming;
But I destroy their relations slowly and they don’t see it coming.

My company gives the youth, new dreams and vision;
But I destroy their reality beyond any reason.

I have made brighter people falter their course;
They went burning their homes for me without any remorse.

I am the means for many people’s cold night slumber;
Slowly I disable them such, which surpasses all number.

I make cultural dehiscence easier in the name of modernization;
Promiscuity, discord, immorality; no more surprises the nation.

My each swallow make them jolly good fellow;
Some continue doing so and I make them yellow.

As time goes by, I become stronger and powerful;
I make their life harder and death very awful.

Some would even vomit out blood but still take me in;
Earlier you had me and now I have you, where is the sin.

And one fine day, my last peg kicks you early to the grave;
My friend, how come you still have been so naïve.

Please don’t curse me; I am not the one to blame;
“Alcohol ruined my life”, these are all excuses lame.

You always knew that I come with a price;
Still you nurtured me and claim yourself as wise.

You tried hiding your inadequacies under my few sips.
It was your own poor coping which brought me to your lips.

I have been destroying families since many a years;
You still embraced me and ironically called cheers.

So, for heaven sake, please take stock of your life now;
I can be shunned, they will tell you, how!

- Dr. Jyoti Prakash