WORLD HEALTH ORGANIZATION
ORGANISATION MONDIALE DE LA SANTÉ

WHO/EPID/SLP/E/5.6

ORIGINAL: ENGLISH

WHO-SPONSORED TRAINING COURSES IN EPIDEMIOLOGY

INDEXED

Self-Learning Package - English Language Course

(prepared by Mr T. K. Sundaresan)

Main Topic: SAMPLING

Sub-Topic: SAMPLING TECHNIQUES

Sampling (Statistics)

Relations to functions of an epidemiologist

Sample surveys are efficient and economical means for obtaining information on matters of epidemiological importance e.g. prevalence of specific diseases. There are several sampling techniques available to suit different practical situations and specific objectives of the investigation. For a given situation one technique can be more efficient and economical than another. A knowledge of these techniques, their advantages and disadvantages will greatly help in choosing the most appropriate technique for that situation.

Educational objective (Ref. No. 5.6)

After studying this document and doing the exercises listed at the end you should be able to suggest the appropriate sampling technique for most epidemiological investigations.

Pre-requisites to understand this section

1. Understanding the concept of random sampling (Educ. Obj. No. 5.2)
2. Understanding the terms "precision" and "accuracy" (Educ. Obj. No. 5.3)
3. Understanding the concept of "Confidence intervals" (Educ. Obj. No. 5.4)

1. Preamble

The epidemiologist, wishing to carry out a sampling investigation, has often to choose one of several techniques of sampling. Some of the considerations that play a role in the choice of the most appropriate technique are: availability of a suitable frame, facility of the field operations, acceptable precision of the estimates, the sub-categories of the population for which separate estimates are required etc. There are many techniques available but in this package the following techniques only will be considered:

1. Simple random sampling
2. Stratified sampling
3. Two-stage sampling, and
4. Cluster sampling

Simple random sampling

Definition: A sample of size n, drawn from a population of size N in such a way that every possible sample of size n has the same chance of being selected, is called a simple random sample.

Example It is desired to determine the prevalence of a certain infectious disease among the 1000 people in a village (N = 1000). The desired accuracy of the estimate dictates that 100 people from the village must be examined for the disease (N = 100). A list of all inhabitants is obtained (the population frame). Then, using a table of random numbers, a sample of 100 persons is selected in such a way that each person in the village (i.e., each entry in the list) has a 10% (n/N) chance of being selected for the examination.

Advantages:

1. Simple design.

2. Requires no knowledge of the distribution of the study variable in the population.

Disadvantages:

1. Requires a population frame which can be expensive or simply not available.

2. Travel expenses could be high if the population is spread over a large area.

Stratified random sampling

Definition: A stratified random sample is one obtained by first separating the population into nonoverlapping groups, called strata, and then selecting a simple random sample from each stratum.

Example Continuing with the example given above for simple random sampling, before selecting the random sample of 100 the entire population is divided into age-groups (stratified on age). Then a simple random sample is taken from within each age-group (stratum). The total number of examinations is still 100, but now the prevalence is available by age-group and the precision of the overall prevalence estimate is improved. This improved precision is due to the fact that sampling is now done within more homogeneous groups (under the assumption that prevalence is correlated with age).

Advantages:

1. Estimates are available for each strata of the population and not just one overall estimate, as in simple random sampling.

2. Some gain in the precision of the overall population estimate, with an appropriate method of stratification.

Disadvantages:

1. A population frame is required.

2. Certain preliminary information is required on the variable(s) to be used for stratification. This may be expensive or simply not available.

3. While stratification on certain variables may improve the precision of the estimate of one parameter (prevalence of a disease, in the above example) it may give rather poor estimates for other characteristics to be studied at the same time (e.g., fertility).

Cluster sampling

Definition: A cluster sample is a simple random sample in which each sampling unit is a collection, or cluster of elements.

Example Continuing with the above example; if a population frame is not available, then the sampling may be done as follows: a list of dwelling units is drawn up, this is much easier than a complete list of individuals. A simple random sample of dwelling units is selected and all individuals in each selected unit is examined.

Advantages:

1. Reduction in travel costs

2. Complete population frame not necessary

Disadvantages:

1. The variability within and between cluster must be considered. If the variability within a cluster is low, then this could lead to unnecessary within cluster precision, using the same number of examinations, while insufficient precision is obtained for the overall estimate.

Two-stage or sub-sampling

Definition: A two-stage sample is obtained by first selecting a sample of groups and then selecting a (sub) sample of elements from within each of the sampled groups.

Example Continuing with the previous example; the investigators decide to use the list of dwelling units, as the population frame is not available, but wish to cover more dwelling units with the same number of examinations. A simple random sample of dwelling units is selected, then, as the second stage, a simple random sample of the individuals within each of the selected dwelling units is chosen.

Advantages:

1. Wider coverage. It is believed that individuals within the same dwelling unit are similar with respect to the quantity measured, then it makes little sense to measure the same thing so many times.

2. Does not require a complete population frame.

3. Reduction of travel costs.

Disadvantages:

1. It sometimes can cause difficulties to include in the sample only a portion of a household or group.

2. Incomplete information concerning the distribution of the quantities measured at the various stages can lead to poor results.

SELF-TESTING EXERCISES*

A. A survey is planned to determine fertility rates in a large geographical area for which no accurate lists of individuals exist; however the villages and towns are well documented.

1. Suggest two possible sampling schemes.

2. State two advantages that each has over the other.

B. It is desired to know the morbidity rate among the population of a city. The results of a recent census, including age and sex, are available.

1. Suggest two possible sampling schemes.

2. State two advantages that each has over the other.

C. It is desired to estimate the prevalence of tuberculosis in a country. The population census is at least 10 years old and it is known that movements of population may have taken place. It is also suspected that the prevalence in urban areas is likely to be different than in rural areas. Make any assumptions, relevant to your country context, on operational factors and suggest a sampling scheme.

---

* Answers are to be found on the following page.

ANSWERS

A.1. Cluster and two stage

2. Cluster over two stage:

- reduced travel cost (must visit fewer villages)

- simplified design

- requires less knowledge of the distribution of the variables measured

Two stage over cluster:

- wider coverage (provides estimates for more towns and villages)

- more flexible, one can alter sampling fractions to attain increased economy if there is some preliminary knowledge about the variability among villages

B.1. Simple random and stratified random

2. Simple over stratified:

- simplified design

- no prior knowledge of the distribution is necessary

Stratified over simple:

- estimates by age and/or sex may be obtained

- more precise overall estimates may be obtained

C. Suppose it is most efficient operationally

1. to examine all members of the household once a household is selected

2. to station the mobile X-ray van and the team in one area for at least one week, and that

3. 50 individuals can be examined in one day, and further

4. at least 20 000 individuals have to be examined to yield an estimate with acceptable precision.

The following can be recommended as one possible procedure.

Because of the operational factors outlined it would be best to adopt a multi-stage cluster sampling, stratified into rural and urban areas. Suppose it is decided to examine 10 000 individuals in the rural and 10 000 in the urban areas. Taking as an example the rural areas: the 10 000 individuals could be considered as the population of 40 clusters consisting of 250 individuals in each cluster (50 individuals per day) for five days a week. If the average household size is 5 this would mean 50 households in each cluster.

Based on the old census, the country can be divided into its major administrative division, and the 40 clusters allocated to these divisions, in proportion to their population. Within each major administrative division minor divisions where the cluster should be located may be selected at random at the second stage. Further stages may be introduced if the divisions are large.

Once the last stage, say a sub-division with a population of around 1000 is reached, a complete census of this sub-division is made and a household selected at random. Fifty consecutive households in a predetermined order will give the required cluster.

* * *