Sampling Methods

Dr. M.M. Mweu,

MBChB Level V Epidemiology,

6 June, 2018

Census vs Sample

Census – every individual in the population is measured

Sample – Only a subset of the population is measured

- Collecting data from sample is convenient to evaluating the entire population & cost-effective
- In a census only source of error is measurement
- In a sample both measurement (variability in outcome due to instrument) & sampling errors (sample-sample variability) present
- Samples drawn to support both descriptive & analytic studies
- Descriptive – describe population attributes (d'se frequency, ave. weight etc.

Sampling cont'd

- Analytic – done to estimate magnitude of associations between outcomes & exposures in the population e.g. is smoking associated with risk of CHD

Hierarchy of populations

(a) Target population – population to which it might be possible to extrapolate results from a study e.g. in a post-surgical mortality study all patients undergoing surgery in Kenyan hospitals will be target population

(b) Source population – population from which study subjects are drawn

Hierarchy of populations

e.g. in the post-surgical mortality study all patients undergoing surgery at KNH may be enlistable

(c) Study sample - consists of the individuals that end up in the study i.e. a sample from the source population

- Internal validity – refers to whether/not the study results are generalisable/applicable to the source population i.e. if probability/random sampling was used
- External validity – relates to how well the results can be generalised to the target/reference population

Hierarchy of populations

- For this, the source population should be representative of the target population i.e. the source population should capture well the attributes of the target population
- Usually it is easier to generalise results of analytic than descriptive studies. Representativeness not necessary to generalise associations
- **Sampling frame** – list of all sampling units in source population
- **Sampling units** – basic elements of the sampled population e.g. households, individuals etc.

Hierarchy of populations

- Objectives of a study influence sampling strategy used – descriptive studies (prevalence/incidence estimation) – random sample required. Analytic – randomness not mandatory

Types of error

- Measurement & sampling error affect study results.
- In analytic studies, 2 types of error recognised:
  - Type I (α) error – you conclude that an association exists when none
  - Type II (β) error – you conclude that no association exists when there is
- *P* is probability of making type I error. ***P*-value** is probability that a difference as large (or larger) than observed could be due to chance if H_o is true

Hierarchy of populations

- **Power** – probability of finding a statistically significant difference when it actually exists and is of certain magnitude $(1 - \beta)$
- Reasons for no associations:
  - Truly no effect of exposure on outcome
  - Inappropriate study design
  - Too small sample size (low power)

Probability & non-probability sampling

Non-probability sampling

- Sample drawn without an explicit method for determining an individual's chance of selection i.e. non-random sample
- Such samples are inappropriate for descriptive but usable in analytic studies
- Types:
  - **Judgement** – judged by investigator as being "representative" of source pop
  - **Convenient** – chosen because it's easy to obtain e.g. households located near the road.
  - **Purposive** – study subjects meet inclusion criteria e.g. have certain attributes of interest e.g. age

Probability & non-probability sampling

Probability sampling

- Every sampling unit in population has a non-zero chance of inclusion in sample
- Random selection of sampling frame necessary

Methods:

a) Simple random sampling

- Every study subject has an equal probability of being included
- Complete list of source population (sampling frame) is required
- Formal random process is used e.g. drawing numbers from hat

Probability sampling

a) Simple random sampling

Use of computer-generated numbers, random-number tables or calculator

- E.g. A list of 40 household heads each having a unique identifier. We want to select 10 households randomly from this list. Using a random-number table we select consecutive 2-digit numbers starting from upper left. If a random number does not match a household number, it is left out. After each number is used it's crossed out so that it's not reused.

Already used, skip

3447	2352	6959	1937	2554	6904	9098	4316
4318	2346	7276	1880	7136	9603	8463	3152
7000	2865	8357	4475	9804	0042	1106	7949

Probability sampling

b) Systematic random sampling

- Complete sampling frame is not required provided an estimate of the total number of units is available. Units are ordered
- Doesn't involve a separate random selection of each unit
- Often used to select large samples from a long list of households
- Steps involved:
  - Calculate sampling interval (no. of households in pop/no. of households in sample
  - Select a random start between start and sampling interval
  - Repeatedly add sampling interval to select subsequent households

Probability sampling

c) Stratified random sampling

- Prior to sampling, population is divided into mutually exclusive strata based on factors likely to affect outcome
- Within each stratum a simple or systematic random sample is chosen
- Simplest form is proportional where no. chosen in stratum is proportional to pop size of stratum. Advantages:
  - Ensures all strata are represented in the sample
  - Precision of overall estimates may be greater than those of SRS
  - Produces estimates of stratum-specific outcomes

Probability sampling

d) Cluster sampling

- A cluster is a natural/convenient collection of study subjects with one/more xteristics e.g. a household is a cluster of people, a clinic is cluster of patients, a classroom is a cluster of students
- In a cluster sample, the primary sampling unit (PSU) is larger than the unit of concern e.g. if you want to estimate the proportion of form 4 students who smoke in Nairobi, you could use a cluster sample in which you randomly select form 4 classes, even though the unit of concern is the student
- In a cluster sample every study subject within the cluster is included in the sample i.e. all students in the selected classes
- Cluster sampling is used because it might be easier to get a list of clusters (e.g. form 4 classes) than it would be to get a list of individuals (e.g. form 4 students) and is often cheaper to sample a small no. of clusters than to collect info from selected individuals within many different clusters
  - In the form 4 smoking survey example, assuming 47 form 4 classes are present in Nairobi, 10 could be randomly selected from a list provided by the County Education Dept. and every student in each of the 10 classes asked to complete a questionnaire

Probability sampling

e) Multi-stage sampling

- Used when it is not possible to obtain a list of all sampling units in the population or not possible to create list, households not ordered
- Initially select clusters – convenient collection of study units e.g. villages, county etc.
- 2-stage sampling common – two ways available:
- a)
  - First select PSUs (larger that unit of concern) with probability proportional to their size e.g. if village size is known, larger villages have higher chance of selection than smaller ones

Probability sampling

e) Multi-stage sampling

- Then select fixed number of SSUs (households) in each of selected villages

- b) If village size is unknown ahead of time:
  - Select PSUs (villages) by simple random sampling
  - Select constant proportion of SSUs (households) in each selected village
- Each of the 2 steps ensures each household has same probability of selection