Quantitative survey methods

COECSA, 18th April 2013, Kampala

Dr. Jefitha Karimurio
University of Nairobi

Outline

1. Introduction
2. Defining the population,
3. Sampling and non-sampling errors,
4. Computation of the minimum sample size,
5. Sampling schemes,
6. Data methods and data collection form,
7. Planning for a survey plan.

Introduction

- A survey is a cross-sectional study: a “snap-shot” of what is happening in the population.
- Data are collected once; no follow-up of participants.
- **Numerical data for quantitative**; textual for qualitative surveys
- Occasionally, data on risk factors are collected.
- Data are analysed to examine the relationships between the risk factors and prevalent health conditions.

Why conduct a survey?

- Surveys provide information required for:
  - **Planning** new projects (baseline surveys)
  - **Evaluation of** existing projects (impact assessment surveys)
  - **Policy formulation, advocacy** and **health promotion**

- To **generate hypotheses** for further studies such as cohort studies and randomised controlled trials

Survey proposal

- A **proposal** and **ethical approval** are required.

- The proposal should provide evidence that the study is:

  - **Feasible** (adequate time, funds, personnel and technology)

  - **Interesting**

  - **Novel** (provides new information)

  - **Relevant** (will advance health policies and clinical practice)

Survey proposal include?

The proposal should include:

- **Clear survey objectives** defined in precise terms for ease of assessing whether they have been achieved.
- **Well-defined population (and study area)** for accurate estimation of prevalence and planning interventions.
- **Accurate procedures (written)** to enhance reliability.
- **Accurate budget** to avoid shortages or wastage of funds

Defining the population

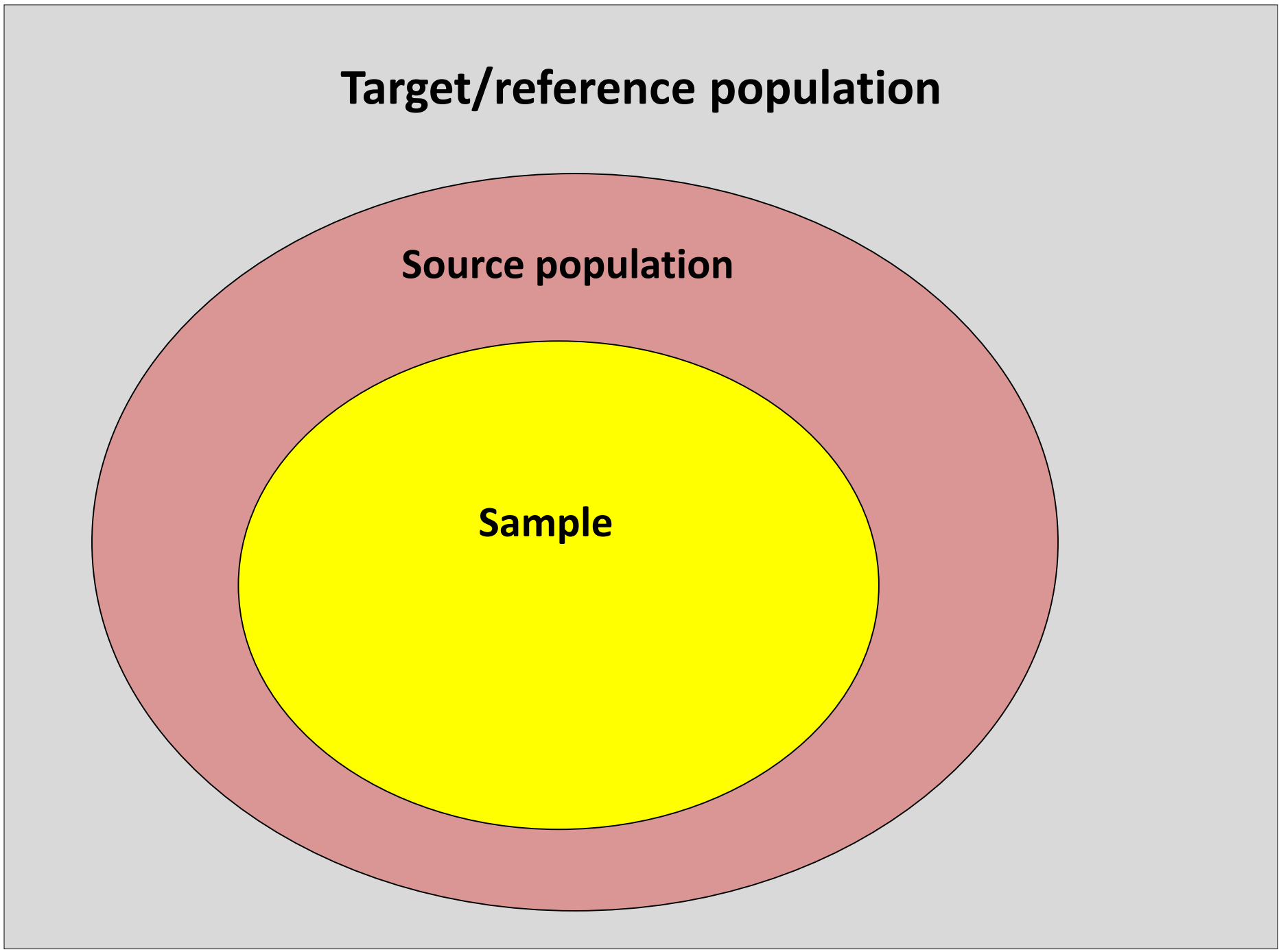
- **Target population:** the total **eligible** population in the study area from which the sample is drawn.

- **Study population (sample):** persons selected and examined.

Other terms:

- **Reference population:** population to which the survey results applies: may be equal to or larger than the target population.

- The term **source population** is used if the sample is drawn from a portion of the target/reference population.


Target/reference population
Source population
Sample

Population parameter and sample statistic (estimator)

- The aim is to estimate true prevalence in the target population.
- Usually, prevalence is measured in a **representative** sample and used as the **estimator** for the true prevalence.
- **Internal validity:** How the inference from the sample approximate the “truth” in the target population.
- **External validity** (generalisability): whether the results can be extrapolated to other settings, times, and so on.

Exercise 1: sampling

- The paper bag contain a many red and few black beans.
- Use the cup provided to draw 20 small sample (1 cup each).
- Estimate the prevalence (%) of black beans in each sample.
- Return the beans into the bag.
- Select 20 large samples (3 cups each) and repeat the exercise.

Results: Small samples

Sample	Prevalence (%)
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Sample	Prevalence (%)
11.	
12.	
13.	
14.	
15.	
16.	
17.	
18.	
19.	
20.	

Results: Large samples

Sample	Prevalence (%)
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Sample	Prevalence (%)
11.	
12.	
13.	
14.	
15.	
16.	
17.	
18.	
19.	
20.	

Discussion

- Are the prevalence estimates of the samples equal? Why?
- Compare the range (highest minus lowest estimate) for the small and large samples?
- Does this affect the sampling distribution?
- From this exercise, can you tell the true prevalence in the bag or in another bag with similar beans?
- What can you do to determine the true prevalence?

Sampling variation

- Multiple samples drawn from the same population rarely provide the same results.
- Sampling variation is the term used to describe this difference. It occurs due to chance.
- Moreover, the estimates from multiple samples are normally distributed around the true prevalence (sampling distribution).
- In a survey **only one sample is studied** and statistical principles used to estimate the true prevalence.

Random Sampling Error (SE)

- Statistical term used to express the **difference between an estimator and the true prevalence**. It is not a “mistake”.
- SE occurs due to chance since the prevalence in the sample and in the rest of the target population may not be equal.
- Minimized by selection of an **adequate/representative** sample.
- The larger the sample the closer the estimator is to the “truth”. No SE/confidence interval if the whole population is examined.
- NB: Standard deviation is for actual data and SE for proportions.

Non-sampling (systematic) errors

- Results from biased survey methods.

- Examples:
  - Biased selection of the sample,
  - Non-response or exceeding the sample,
  - Mistakes in diagnosis and clinical grading,
  - Mistakes in coding, recording and entry of data,
  - Biased analysis and reporting.

Non-sampling errors continued

- Non-sampling errors distort survey results and may lead to ridiculous findings which are not related to the objectives.
- May lead to over or under-estimation of the true prevalence.
- Minimised through meticulous **training and validation** of the enumerators (inter-observer agreement testing) and data clerks.
- Non-sampling errors can neither be corrected nor compensated for using statistical methods/arguments.

Compute of minimum sample size

- Parameters:

  - Expected prevalence (p)

  - Maximum acceptable sampling error (e)

  - Confidence limit (usually 95% or Z score of 1.96)

  - Expected design effect (d)

- Equation 1:

$$\text{Minimum sample size} = d\frac{Z^2 p(1-p)}{e^2}$$

Expected prevalence

- A survey is needed because the prevalence is not known.

- Therefore, you have to estimate or predict the prevalence using:
  - Reports/publications of preceding studies (literature review),
  - World Health Organization estimates,
  - Risk scores from pre-survey risk assessment (e.g. trachoma),

- The **low**er the **prevalence** the **large**r the **sample** (Equation 1).

Minimum acceptable sampling error

- Indicates the desired precision for a specified prevalence estimate. 95% CI is equal to prevalence +/- 1.96 SE.

- Consequently, the **error depends on the prevalence estimate**.

- The standard **relative precision** is 20% of the prevalence estimate,

- Generally, an error of >50% is considered too low for accurate statistical inference.

Examples

- If prevalence = 1%, then 20% of 1% = **absolute precision** of +/- 0.2%; expressed as: 1%(95%CI: 0.8%-1.2%).

- Prevalence = 30%, then 20% of 30% = absolute precision of +/- 6% or 30%(95% CI: 24%-36%).

- If prevalence is 5%, 10%, 15% or 40% then?

- Precision = accuracy of the estimator; confidence level = certainty that the true prevalence is contained in the 95%CI.

Exercise 2: Absolute/relative precision

serial	Assumed prevalence	Precision (maximum error acceptable)		Expected 95% CI	
		Relative	Absolute	Level	Interval
1.	10%	50%	5	95%	5%-15%
2.	10%	30%		95%	7%-13%
3.	10%		4	95%	6%-14%
4.	20%	20%		95%	
5.	35%	20%		95%	
6.	15%		7.5	95%	

Design Effect

- The factor by which the sample for Cluster Random Sampling (CRS) is multiplied, to compensate for increased random SE.
- Penalty for deviating from Simple Random Sampling (SRS).
- **Predicted** using experiences from previous studies. Therefore, 95% CIs for CRS are routinely adjusted for potential clustering.
- **Clustered diseases** like trachoma and **large survey clusters** attract high design effects.

Exercise 3

- Use the hints below to spot the differences in the parameters used to calculate the samples in the next slide.

- Note: the factors that increase the samples size are:
  - Low prevalence,
  - High precision (standard for prevalence surveys = +/-20% of the expected prevalence),
  - High design effect,
  - High confidence level (standard = 95%).

Spot the differences

Variables	Sample 1	Sample 2	Sample 3	Sample 4
Prevalence estimate	35%	20%	10%	5%
Absolute precision (max. acceptable error)	+/-7%	+/-4%	+/-3%	+/-2.5%
Confidence level	95%	95%	95%	95%
Design effect	4.0	4.0	2.0	1.5
Minimum sample	**713**	**1,537**	**768**	**438**

Exercise 4: Computation of sample size

Generate survey sample sizes by feeding different parameters in the:

- Excel sheet with survey sample size computation formula
- Sample size calculation software

Cost of a survey and sample size

- The sample size is computed to inform budgeting/planning.
- The larger the sample the higher the cost: **balance between the sample size and available funds** is critical.
- Usually, it is the precision (minimum acceptable error) of the prevalence estimate which is adjusted to achieve this balance.
- **DO NOT commence a survey without adequate resources**: time, money, manpower/skills, materials and equipment.

Target population and sample size

- If the target population is >5,000 people, it is not used to compute the minimum sample size.
- Further increase has minimal effect on the sample size.
- Sample size calculation formula for population based surveys (Equation 1) assumes population size of >5,000 people.
- Same principle is used in opinion polls: standard sample size used, irrespective of the population of the country.

Selection of sample: Sampling frame

- Equal probability of selection (epsem) methods. Non-probability methods will be discussed in qualitative methods.

- Prepare sampling frame: complete list of all sampling units in the target population from which the sample is selected.

- In the SRS the sampling unit is an individual person while in CRS it is a group of people (cluster).

- A cluster can be a village (rural setting), enumeration areas or a block of houses (urban setting/refugee camps).

What is the frame based on?

- Ideally, a pre-survey census should be conducted. In clinical studies clinic/admission/theatre registers are used.

- Most recent census reports are commonly used because census is too expensive for most eye care projects.

- Where only an old census report is available, project the target population using the average growth rate.

- Cross-border migrations/influx of refugees a challenge. Example: Kenya-Uganda – Southern Sudan border communities.

Selection of the sample: process

- Use a sampling scheme that is easy to implement, economical and does not introduce biased in estimation of the prevalence.
- The sample can be drawn from the sampling frame using: SRS method, one-stage CRS or two-stage CRS.
- The first stage CRS involves selection of clusters and second stage the households.
- Multi-stage sampling: complicated sampling and analysis.

Data collection methods

- Observation: for example a clinical examination or inspection of a household for environment risk factors.

- In-person interview (one-to-one basis). Open ended questions to collect textual data to be covered in qualitative methods.

- Telephone interview.

- Post questionnaires via mail or online: interviewer gives the questionnaires to potential respondents (self-administered).

In-person interviews/examination

- The interviewer asks questions or examines the respondent.
- Does not give the questionnaire to the respondent. Subsequent questions may provide answers/hints.
- Advantages :
  - It reaches people who cannot be reached by telephone, post and internet.
  - It create rapport (respondent can ask for clarifications).

In-person interviews continued

- Disadvantages:
  - Takes a longer period of time than self-administered questionnaire,
  - Expensive due to personnel/travel costs,
  - May not get permission to visit respondents when they are busy at work/school,
  - More prone to bias e.g. due to personal appearance.

Data collection form (questionnaire)

- Define the survey data set and collect the **minimum data set required to meet the objectives** of the survey.

- Define the attributes of each variable for electronic data entry system: name, type, length, code (e.g. Male =2, Female =2) etc.

- Types of data collection forms:
  - fully structured (ordered questions, read word-for-word),
  - semi-structured questionnaire,
  - unstructured interview schedule.

Survey time-table (Gant chart)

ACTIVITIES	TIME (DAYS, WEEKS OR MONTHS)											
	1	2	3	4	5	6	7	8	9	10	11	12
1. Procurement												
2. Set-up survey office												
3. Recruit staff												
4. Prepare data capture tools												
5. Training and validation												
6. Pilot study												
7. Data collection/analysis												
8. Interim report												
9. Final report												
10. Disseminate the findings												

Procurement and recruitment

- List all equipment /supplies required for scheduled activities.
- Procure/transport them to survey sites prior to commencement of data collection.
- List main tasks involved and outline the job descriptions. Recruit adequate staff and allow for possible attrition.
- Include: coordinator, statistician, logistics manager, enumerators, data clerks, drivers, guides and community mobilisation team.

Training, pilot study and validation

- A training includes: background information of the study area, survey methods and logistics.

- A pilot study is conducted during the training workshop to test and improve the data collection tools.

- Inter-observer agreement testing is done to validate the enumerators. Training is repeated agreement is low.

- Also, train and test data entry clerks. A test run electronic data capture tools using data from the pilot study.

Data collection

- Requires **meticulous planning and co-ordination**,
- Support by the local communities is critical. Community and individual **consents** should be taken prior to data collection.
- **Community mobilisation** is required through-out the survey period to ensure high study participation rate,
- Ensure **adequate materials**: time-tables, route maps, directories, data collection tools, consent forms and manuals.
- Arrange for transport , food/accommodation and allowances.

Data management

- Plan for data management and included it in survey budget.
- It involves inspection of the completed forms to ensure they have no mistakes, data entry, cleaning and analysis.
- Send immediate feed-back to the data collection team if mistakes are noted.
- Data analysis to calculate: participation rate, prevalence estimates and correlations between variables.

Reporting and dissemination

- Include reporting of the survey findings in the budget.

- The principle investigator is expected to at least:
  - Prepare and circulate the draft survey report for inputs from peers and partners,
  - Write the final comprehensive report,
  - Brief project partners and the community on key findings.
  - Others: Publish/conferences /WHO meetings etc

Take home message

- A survey is a cross-sectional study.
- Objectives should be clearly and precisely stated because they inform the type data to be collected and activities to plan for.
- The prevalence in a representative sample is used as the estimator for the unknown prevalence in the target population.
- Detailed planning, supervision, training and pilot study are vital.
- Allocate funds for data analysis, reporting and dissemination.

References

- Johnson G, Minassian D, Weale R, West S. The Epidemiology of Eye Disease. London: Imperial College Press; 2012.

- Minassian D. Epidemiology in practice: Sample size calculation for eye surveys: a simple method. J Comm Eye Health 1997;10(23):42-4.

- Rapid Assessment of Avoidable Blindness manual: www.iceh.org.uk/display/WEB/Rapid+assessment+ofavoidable+blindness+%28RAAB%29+CD)