Sampling, Sample Size and Planning for Data Collection

Prof. Andrew Stokes

October 2, 2018

Sampling

Sampling involves selecting units from a population

- why sample?
- why not census?
- how large is the population?

Sampling definitions

- **Sampling unit** Things sampled. Examples: people, clinical episodes, facilities
- **Study population** All sampling units which could possibly be included in the sample
- **Sampling frame** List of all units in the study population. Examples: class lists, e-mail addresses, list of phone numbers from the phone book (less relevant today than 20 yrs ago)

The study population is concept, but sampling frame can be enumerated

What about your study

What is your sampling unit? Study population? Sampling frame?

- Does a sampling frame exist?
- Who could you get it from?
- What will you do if you can't get one?

Probability sampling methods

- Simple random
- Systematic
- Stratified
- Multistage
- Cluster

Simple random sampling

Used where the number of sampling units is small and sampling frame is enumerated. The process involves,

- identifying all units available for sampling
- decide on size of sample
- choose units using lottery

Systematic sampling with equal probability

Systematic sampling is a method of selecting randomly from long list of units

- Make a numbered list of all units
- Sampling interval is the ratio of the total number of units to the desired N.
- For example, if you want to select 6 units from 18, the interval is 18/6 or 3

Stratified sampling

Use stratified sampling When sampling frame contains categories (strata) to be considered separately.

- Examples: SPH and Medical Students
- Male and female students

The first step is to sort list of units by stratum. Then, select units using simple random or systematic random sampling

Multistage sampling

- When design is complex and efficient to select units in stages
- Randomly select primary sampling units at 1st stage. Example: Specific SPH departments or classes
- Within the primary sampling units, randomly select units at 2nd stage. Example: Students
- In very complex sample, additional stages are needed

Multistage sampling example

Let's say the goal is to obtain a representative sample of schoolgoing orphans ages 9-16 in a single district in Zambia

- From 252 total schools, selected 60 schools randomly
- In each school, identified classes with children of appropriate age
- Within each classroom, teachers identified orphans (via form)
- Randomly selected orphan within each class

Why this strategy here? Because you can't come up with a sampling frame of all orphans ages 9-16 in the district. At the smaller level, enumeration is possible.

Multistage versus stratified

Multistage and stratified sampling seem similar, but they have different goals.

- Stratified: we want to represent each group evenly. We have sampling frame by categories
- Multistage: motivation is logistics. First sample higher level, get sampling frame from lower level and sample

Cluster sampling

- When logistically easier to select units in groups or if it makes sense. For example, health research in rural villages or education research in classrooms
- Include entire cluster or randomly select a subsample
- Disadvantage: need large number of clusters for precision (rough rule of thumb: minimum of 30 clusters)

Non-probability sampling: convenience sampling

Study units available at time of data collection selected.

- Sit in village square and sample whoever comes by
- Come to a classroom and whoever is there that day gets sampled
- Set up a table at Chequers Cafe and flag down students as they pass by

Non-probability sampling: quota sampling

Different categories of sample units included until certain number reached in each category

Biases in sampling

- Study volunteers
- Landline owners only
- Non response
- Miss cases of short duration
- Tarmac bias (i.e. more likely to sample those closest to the road)

Sampling for your study

How will you choose your sample?

- Sit out in the medical school lobby and give free cookies to any student that fills out the survey
- Get a list of all student e-mails, randomly select e-mails, and e-mail a survey to each selected student
- Distribute the surveys in a randomly selected class

Sample Size[1]

The issue of sample size is critical when planning a new study. Your goal should be to make sure that you've collected enough data for the results to be useful, without going overboard and collecting more data than are needed. You need to strick a reasonable balance, keeping in mind that,

- If the study is too small, you may miss (fail to detect) important differences between groups. In the case of an intervention, you may not be able to detect its effects.
- If the study is too large, you will be wasting scarce resources.

[1] Source for sample size slides: Smith text, Chapter 5.

Two approaches to sample size calculations

When planning the size of your study, you can focus either on,

- desired precision of outcome measures
- desired power

We'll go over both approaches to calculating sample size

Preliminary notes

- Sample size calculations are approximate. If we knew the input parameters exactly, we wouldn't need to do the study! The goal is to get a rough estimate of how much data need to be collected, balancing power (or precision) on the one hand with resource constraints.
- Sample size calculations are more or less complex depending on the study design. Below we'll focus on data collected at the individual level using simple random sampling. We will assume that in comparing outcomes across groups, there are roughly equal numbers of individuals in each.

Complicating factors

Calculating study size can be more complicated under certain circumstances, for example, if,

- Data are collected at the community-level, using, for example, a cluster sampling design
- Group sizes are unequal
- Interim analyses are planned

Types of error in epidemiological studies

The two types of error in epidemiological studies are,

- Sampling error
- Bias

Bias can persist in the data whether you collect 100 or 1000 observations. Sampling error, in constrast, goes down as you collect more data. Our concern today is with the latter.

Criteria for determining sample size

- **Precision of effect measures**: What is an acceptable confidence interval around your estimate?
- **Power**: Assuming there is a real effect, what's the chance that we will detect it?

More on power

- Power is the probability of obtaining a statistically significant result assuming there is a real difference in outcomes across groups.
- Significance is evaluated with respect to the null hypotheses, which generally states that there is no difference in outcomes across groups or intervention effect.
- A statistically significant result suggests that the data go against the null hypothesis.
- We never know for certain whether we will obtain a significant result in a study, but for an effect of a given magnitude, power calculations allow us to specify the probability of obtaining one.
- A power of 80% to detect a difference of a given size means that if a study were repeated many times, a significant result would be obtained 80% of the time.

Factors that contribute to the power of a study (from Box 5.1 of Smith text)

- The magnitude of the difference between groups in the outcome under study
- Sample size of the study
- The probability level at which a difference is considered statistically significant

Which criteria to use?

What does the prior literature say about the magnitude of the difference between groups? In the case of an intervention, does the prior literature consistently point to an effect?

- If so, focus on the level of precision. If it is already known there is a difference or effect, it is less important to test the null hypothesis
- On the other hand, if little is know about the size of the difference or effect, then use the power criterion.
- This approach helps ensure that you detect an effect conditional on there being one.
- The disadvantage of the appraoch is that it may yield imprecise estimates of the difference or effect.

Multiple outcomes

- If your study has multiple outcomes, decide which one is the outcome of primary interest and use that for sample size calculations.
- Alternatively, if you have several outcomes of equal importance, you would ideally repeat the calculation for each one elect for the largest of the calculated sample sizes.
- In practice, some outcomes may need more data than are feasible to collect (e.g. difference in the child mortality rate between two communities or at two points in time).
- In these cases, try to find a proxy.
- If it's not feasible to design a study of sufficient size for the primary outcome, than the study itself should be modified.

Practical matters

What are some practical constraints to collecting data?

- Time
- Money
- Staff
- Vehicles
- Lab space

What are some of the constraints in the context of this class?

Power curves

- Power curves can be used to study the trade-offs between study size and power.
- Power curves provide a graphical depiction of how power varies with study size for different assumed values of the difference in outcomes or effect.
- For an example, see Figure 5.1 in the Smith text (p. 75).

Study size for adequate precision: proportions

Examples of proportions might include,

- the prevalence of obesity
- the proportion of respondents who develop active TB during follow-up
- what about in your study?

Study size for adequate precision: means

Examples of means might include,

- mean BMI
- mean systolic blood pressure

Precision calculation for a single proportion

$$n = p(1 - p)\left(\frac{Z}{E}\right)^2$$

Precision calculation for a single mean

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

Precision calculation for comparison of proportions

Suppose that the true proportions in groups 1 and 2 are p_1 and p_2 and the relative risk is $R = \frac{p_1}{p_2}$. The approximate 95% CI for R goes from $\frac{R}{f}$ to Rf. The factor f is given by

$$f = \exp\left\{1.96\sqrt{\left[\frac{1-p_1}{np_1} + \frac{1-p_2}{np_2}\right]}\right\}$$

where n is the number of people in each group and f is known as the *error factor*.

Precision calculation for comparison of proportions

The parameters f, p_2 and R are specified by the analyst. The proportion and relative risk estimates are entered as rough approximations and can be drawn from prior studies. Once these values are chosen, the number required in each group n can be calculated as,

$$n = \left(\frac{1.96}{\ln f}\right)^2 + \left\{\left[\frac{R+1}{Rp_2}\right] - 2\right\}$$

Precision calculation for comparison of means

- compare means of values in two groups (for example, mean BMI in women and men or mean BMI in the intervention and control groups)
- Examples from your projects?

Precision calculation for comparison of means

Suppose the true means in groups 1 and 2 are μ_1 and μ_2. We will compare them in terms of the difference in the means, $D = \mu_1 - \mu_2$. The 95% CI for D is given by $D \pm f$, where,

$$f = 1.96\sqrt{\left[\sigma_1^2 + \frac{\sigma_2^2}{n}\right]}$$

where σ_1 and σ_2 are the standard deviations of the outcome variable in the two groups.

Precision calculation for comparison of means

A value of f is chosen by the analyst. Values for σ_1 and σ_2 are often drawn from the literature. For the purposes of this class, we will make the assumption that $\sigma_1 = \sigma_2$. The sample size needed for each group is then calculated as,

$$n = \left(\frac{1.96}{f}\right)^2 (\sigma_1^2 + \sigma_2^2)$$

Estimating sample size needed for adequate power

- What size of difference, D, between the two groups would constitute a finding of public health importance?
- What level of confidence do you wish to have of obtaining a significant result conditional on D being the true difference? Power is entered into the calculations as z_2 (see Table 5.1 in Smith text). A common value is 80% or higher. The corresponding value of z_2 for this level of power is 0.84.

Estimating sample size needed for adequate power

- Choose a significance level for the comparison of the two groups (entered as z_1). A common choice for the p-value is 0.05, which is z_1 of 1.96.
- Approximate estimates of other parameters depending on whether the focus is on proportions, rates, etc.

Power calculations: comparison of proportions

The sample size required in each group to detect a specified difference $D = p_1 - p_2$, with power specified by z_2 and significance level specified by z_1, is given by,

$$n = \frac{[(z_1 + z_2)^2 2p(1-p)]}{(p_1 - p_2)^2}$$

where p is the average of p_1 and p_2

Power calculations: comparison of proportions

To calculate the power of a study of a given size, calculate z_2 and then use Table 5.1 in the Smith text to identify the corresponding power level.

$$z_2 = \left(\sqrt{\frac{n}{2p(1-p)}}\right)(|p_1 - p_2|) - z_1$$

Additional notes on power calculations

- For two groups of unequal size, additional formula is needed
- For sampling designs involving clustering, sample size calculations have to be adjusted
- Sample sizes for class can be done by hand or using software. Most software programs have the ability to estimate sample sizes.
- For more detail on the methods presented in these slides, refer to Chapter 5 of the Smith text.

Planning for data collection

Three stages to the data collection process:

- permission to proceed
- data collection
- data handling

Stage 1: permission

Need consent from relevant authorities at all levels

- Usually an IRB or ethics committee or both
- Permision from participants
- Others?

Stage 1: permission

Formal and informal permissions are both important to consider. It depends on when, where and how you are collecting your data

- In a class
- Will you need space to conduct discussions?
- Will you have a table and ask people to fill out your survey?
- Do you need e-mail addresses?

Informed consent

Basic elements of informed consent for minimal risk research include a description of

- study procedures and what participation involves
- any reasonably foreseeable risks or discomforts to the subject
- any benefits to the subjects or to others which may reasonably be expected from the research

Informed consent also involves a statement concerning confidentiality of records identifying the subject.

Stage 2: data collection

Some logistical questions include,

- who will collect what data?
- how long will it take to collect the data?
- how many surveys will you administer?
- how will you administer the surveys?
- how will you recruit people to participate?

Stage 2: data collection

More logistical questions to consider,

- In what order are the data collected? Qualitative vs. quantitative.
- When should data be collected. Consider accessibility and availibility of population (weather, holidays, etc.)

Stage 2: data collection

- Ensuring quality means thinking early on what might be the potential sources of bias. Consider potential deviations from sampling procedures.
- How can you prevent bias? Fieldwork manual for research team with information on sampling procedures and instruction sheet.
- Pretest research
- Overall, training and good organization will be critical

Stage 3: data handling

- check data for completeness
- number and code questionnaires (this will be automated in KoBo Toolbox)
- Data storage. Who will be responsible for the data? Where will it be stored, backed-up.

Project: next steps

- Sampling plan, sample size calculation, data collection plan

Other upcoming deadlines

- Problem Set 2 (due October 23)

Other items for today

- group work focusing on sampling, sample size calculations and data collection plan

Source material for today's lecture notes

- Peter G. Smith, Richard H. Morrow, David A. Ross. Field Trials of Health Interventions: A Toolbox. Oxford. Chapter 5.