SAMPLING

Introduction

sample = subset of a population

population = all cases that apply to your RQ

element = single member of the population

Sampling Frame = list of all population elements

stratum= sub group of population/sample

unit of analysis = type of case you are collecting data on and to whom you want to generalize

Why Sample?

Can't get data on all population elements or too expensive to do so

goal of sampling = to represent the pop

good sampling designs can produce better data than going after full pop data

Ex: US Census

Why Do We Need A Sampling Design?

1. Different designs are better at producing representative samples for different research goals and populations.

   Representativeness influences generalizability (external validity)

2. Need to proceed in a logical and systematic way. Need to have a plan and keep a record of what you do and decisions you make. So you can defend your sample, and subsequently, what you say about your RQ.

Unit of Analysis and Goals of Analysis

1st step in designing sample is to goal of analysis and identify the unit of analysis

1. Goals of analysis: to understand RQ, to test hypotheses on RQ, to describe population on RQ, to represent population on RQ?

General Qualitative Goal: to understand RQ/process

General Quantitative Goal: to generalize answer to RQ from sample to population

(Read page 97 Qualitative Researching)

2. what is your unit of analysis? Think about what you want to study. Who/what do you want to generalize to?

Is it attitudes about white-collar crime (unit=people) or perhaps policies on white collar crime (unit=state)

Or are you interested in experiences (unit=experience, 1 person could provide multiple cases)

Eligibility for study: adults, English speaking, etc..

3. Your sampling unit and analysis goals determine:

a. who you collect data on

b. how much data you will need to collect

Then the answers to a + b above influence:

- how you will obtain the sample (probability vs. non probability sampling, stratified?)
- the reliability and validity of the data,
- how you will analyze the data,
- what you will be able to say about your RQ and the validity of those statements

(Read page 86-90 of Qualitative Researching for a full discussion of this)

2 types of Sampling Methods: Probability and Non-probability

Probability sampling = each element of the pop has a known probability of selection

*used with surveys

ex. You estimate the # of gay men in the country.

advantages

- can estimate the chances that a sampling design will produce a sample that represents the population. hence, can choose sampling method with best results
- samples usually represent the frame (not always the population b/c of problems with frames)
- knowing the probability of selection allows you to calculate confidence intervals (CIs) around summary stats for your data
    - Example: number of victims of hate crimes

(See pg. 107 Designing Surveys for more examples)

disadvantages

- harder to pull the sample, need more resources, because;
- need a sampling frame

Probability samples require sampling frames.

Need to assess reliability and validity of the frame.

Is it complete? How good is the data? Does it have info needed?

Telephone directories/lists contain
ineligibles: businesses
inaccuracies: fax #'s
missing info: # of residents per household
multiple listings

List of addresses contain
inaccuracies: 2 residences at 1 location
missing info: # of residents per household

Reliability and validity of frame influences what type of probability sample you use.

Exception to using a frame in telephone surveys: RDD Samples.

Select working telephone prefixes, and obtain list of working 1st 2 digits from phone company, phone book, directory. Randomly select the last two digits.

Advantages of RDD:

- Avoids problems with incorrect information on sampling frame (disconnected, changed number, etc.)
- Avoids problems with unlisted numbers
- Don't have to get a list of numbers to serve as a sampling frame

Disadvantages of RDD:

- People without phones are still not included
- Large number of "bad" numbers due to not being in service, a pay phone, a business, etc.
- Households with more than one number are still over represented
- Households with multiple families are still underrepresented

Non-probability sampling = each element of the pop has an unknown probability of selection

*used with qualitative studies and most experiments

advantages: economy and convenience

- often, do not need a sampling frame
- some qualitative projects do use frames:

ex. Study on cultural experiences (get a list of over seas travelers from travel agent)

disadvantage: often results in biased samples (sample doesn't represent population)

There are several types of probability and non-probability samples. Choice of type of sample depends on your desired level of generalizability.

TYPES OF PROBABILITY SAMPLES

*for an overview, see pg. 13-18 Designing Surveys

Steps to Drawing a Probability Sample

1. Define RQ and population (discussed earlier)
2. Find or develop a sampling frame
3. Determine sample size
4. Select Sample

Simple Random Sampling

1. all elements of pop have an = chance of selection

2. randomly chose x # of elements from sampling frame

    to do so, need to number each element on the sampling frame

Results

sample mean = pop mean

Problems

- If you want to look at sub-groups you may not have adequate representation if they are small in #.

- if these sub-groups are related to the RQ, then you have sampling error. Findings will not be accurate.

- Need a sampling frame and one with few errors (ineligible, wrong #/address, multiple listings, people without phones, etc…)

Ex.: RQ = what is the influence of income on smoking, seat belt wearing, amount of saturated fats in diet, etc.. (health behavior). Need to get adequate representation of social classes. If you don't, you could over or under estimate health behaviors

Ex. RQ = how do gender differences influence the effect of income on health behavior. If your sample design generates a sample with disproportionate women, you could over or under estimate health behaviors

Systematic Random Sampling

1. all elements of pop have an = chance of selection

2. chose every x^{th} element from sampling frame

prevents you from having to number each element on the sampling frame

Results

- sample mean = pop mean

Problems

- If you want to look at sub-groups you may not have adequate representation if they are small in #.

Stratified Random Sampling

Random sampling within sub-categories of the population. Sub-groups = stratum

Ex. Influence of Race (whites, black, Asian, Am. Indian, Hispanic) on environmental attitudes and values

Ex. Influence of burglar alarms on victimization rates

*To do this you need data on the strata variable -- need to know someone's race to group them

Can use different data collection methods and sample designs within each stratum:

Data collection methods: telephone, mail, face-to-face surveys

Samples: Random, systematic, proportionate random sampling, disproportionate random sampling

Proportionate Stratified Sampling

take % sample out of each stratum that matches their % in the pop

RQ: does race influence people's access to health resources in NC? If so, how?

ex. Race strata - take 50% sample of whites, 30% black, 5% Asian, 5% Am. Indian, 10% Hispanic

Disproportionate Stratified Sampling

5% Asian American and American Indian will not generate enough cases to statistically analyze data or make sound generalizations to these groups.

So, take higher or lower % sample (to suit needs of research)

ex. Race strata - take 20% sample of whites, 20% black, 20% Asian, 20% American Indian, 20% Hispanic

Ex. National Survey of Families and Households oversampled one-parent households in order to accurately describe these families.

*might disproportionately sample because one stratum is heterogeneous and another

homogeneous. Ex. Race and environmental values (Am. Indian has low variance, Hispanic high)

Why Stratify?

1. To insure that certain key sub-groups will have sufficient sample size for sub-population analyses.

2. To obtain more accurate estimates of incidence/prevalence (sample mean = pop mean)

3. To cut down on costs: can take smaller sample (less sampling error)

ex. Random sample: would need sample of 400 to adequately represent racial/ethnic groups

By creating strata and sampling within you can reduce samples down to n=25 for each stratum. Reducing overall sample (and cost) to 200.

When to Stratify

1. When strata are themselves of primary interest (i.e., they are in your research question, or are key variables in your analysis)

Ex. Race and environmental values: need to represent multiple race groups (whites, black, Asian, Am. Indian, Hispanic)

> *If you didn't stratify you would have to take a very large sample using simple or systematic sampling in order to represent each stratum

Example: influence of masculinity on propensity to commit crime (need to stratify on age cohorts)

2. When variances on key variables differ between stratum.

> If variance on key variables of interest is lower in the subpopulations than in the total population than the standard error of the estimate will go down. (More accurate)

Ex. Race and environmental values (Am. Indian has low variance, Hispanic high)

3. If you need to use different data collection methods for different stratum to obtain most cost efficient data

Strata may require different sampling approaches. Let's say you want to represent the entire U.S. population (goal of the Census). You might create three stratums based on different data collection methods: institutionalized, households, and streets for the homeless.

Households - telephone survey with RDD (medium expense)

Institutionalized - mail survey from resident/member lists (low expense)

Homeless - face-to-face surveys based on area probability survey (high expense)

4. Don't stratify if random sampling from population would likely get needed representation of groups: ex. Gender. Men and women have similar proportions in most populations. If topic is not related to gender, you will likely get equal sample proportions. No need to stratify.

Example: Influence of type of school on whether an adolescent has ever shoplifted (you could argue that gender would have no influence on this relationship. And the population of most cohorts is about 50% M/F. Hence, no reason to stratify on gender)

Cluster and Multi-stage Sampling

Clusters are groups that are more heterogeneous than strata. And you don't need a sampling frame to define them)

Examples: GSS

clustered on regions of country
first stage within clusters is states
second stage is cities
third stage is streets

Advantages: reduces costs (no frames needed for clusters; can use different data collection methods at various stages)

Example: influence of race and class on gun ownership (could create clusters based on neighborhood demographics)

Use when your population is large and heterogeneous on several factors and the

population can be broken down into loose groups

NONPROBABILITY SAMPLING

commonly used with qualitative research and experiments

Qualitative researchers often don't agree with key assumptions of probability sampling:
1. that variables you stratify on have uniform influence on people

Ex. When we stratify on race, we assume race has a similar impact on each stratum

2. Variable analysis: concepts can be turned into standard questions with standard answers that can be coded numerically and analyzed and in turn say something meaningful and honest about the RQ (see page 88 Qualitative Researching)

3. Probability samples most efficiently and effectively produce representative samples: (see page 91 Qualitative Researching)

4. Sampling frames reflect population: (see page 91 Qualitative Researching)

Some qualitative studies do you use sampling frames. Examples: study on culture shock (get list of people who travel globally from travel agent)

Theoretical or Purposive Sampling

1. Chose a group of participants/respondents that will enable you to answer the RQ and that are accessible What are all the situations you need to observe/hear about in order to understand RQ. Rely on theory, lit, experts to answer this.

Ex. RQ = What career plans exist among lobstermen in Belize. Sampling will require researcher to go to Belize to the where there the lobstermen are.

Example: study friendships between gay and straight men

2. Or, chose settings that will enable you to observe topic/behavior in RQ. Rely on theory, lit, experts to identify these. Or identify them once in the field.

Ex. RQ = How do men establish hierarchies in informal interactions? Sampling will require researcher to go to settings including men and hierarchies, such as work settings (fireman, newspaper, car dealerships), friendships, volunteer activities

3. Or, choose groups/settings that will produce desired generalizability.

   - with some people/settings you can go into depth on the RQ
   - some people/settings will give you a range of perspectives on the RQ but no depth into them

4. With any of the above, seek negative cases/instances.

Snowball Sampling

commonly used with qualitative interview studies

start with key informant who identifies eligible people.

Then ask for each interviewee to identify another eligible person. Snowball starts.

Ex. Study on coping mechanisms among HIV+ mothers. Get one, ask her for another name (assumption that she knows others from a support network)

Example: Black and Hispanic gay men's lives

Convenience Sampling

Pick people or settings based on convenience to researcher

Ex. Study on environmental values. Go to a park and to a mall.

Quota Sampling

Figure out the key groups in your study. Then determine the proportion at which exists in the population. Then try to get that proportion in your sample using convenience.

Example: study on students' experiences with drugs. Total sample size is going to be 30. Then go after the correct proportion of freshman, sophomores, juniors, seniors, etc..

ASSESSING CHANGE ACROSS TIME IN SAMPLING

1. **Cross-Section**, ex. North Carolina Employment and Health Survey

2. **Cohort/Trend** data collected from new samples from the same population at different times, ex. General Social Survey; social impact studies

3. **Panel** data collected from the same sample, multiple times. Ex. Study on ethnic diversity among new business owners and success rates

4. **Time Series**: same sample (often a population) at standard time intervals. Examples: economic data; Nurses Survey

DETERMINING NEEDED SAMPLE SIZE (N)

depends on:

- Research design - qualitative interview/observation, quantitative telephone survey or f-to-f survey
- Variability on key variables
- If you are trying to estimate pop values
- If you are testing hypotheses
- If you are comparing sub groups, the size of differences btw. Groups, std error of difference.

Determining Needed Sample Size in Quantitative Research

This is a general review. For more specifics see Advanced Methods Notes

Influences on sample size, in order of effect:

1. Alpha: lower alpha, bigger sample
2. Standard deviation/variance in pop on key variables: Higher variance, bigger sample. (Can sometimes get this from Census data, or data collected in previous studies)

   Sometimes have to estimate from your knowledge of topic or population. Example:

Influence of informed consent on the number of errors that patients make in following post-surgery directions for care. Needed to know about how many patients make lots of errors, some errors, few errors.

3. how big of a difference/relationship do you want to be able to detect? statistical power. Smaller the difference, bigger the sample

    - Example: "driving while black" study. What effect size did they expect between white and black drivers in speeding tickets/pull overs/ etc.

    - Example: informed consent study. How many more errors did they expect to see for patients who received informed consent the day before surgery and those who received informed consent an hour before surgery?

4. Expected RR, Expected or Known Eligibility Rate, and Expected Error Rate on Frame

    Pilot studies help here: Ex. Emergency Medical Services for Children Study

(Se Table on pg. 133 Designing Surveys and read pages 129-130 for examples)

Example: Emergency Medical Services for Children Study, Expected Rates for Pediatricians on Frame

Required Sample Size = 100
Expected RR = 70%
Expected Eligibility = 90%
Expected "Clean" listings = 80%

= 100 / .7 * .9 * .8
=200

Need to go after 200 pediatricians in order to obtain a sample of 100 pediatricians

Determining Needed Sample Size in Qualitative Research

1. Need initial expectations for # of interviews, observations, etc.. Rely on theory, lit, experts to get initial #'s. (See page 99 Qualitative Researching).

    can go through similar math as quantitative if you have a frame:

    what is your desired obtained * expected cooperation rate

    ex. Want 50 interviews. Expect 50% RR. Need to go after 100 interviews.

    Initial quotas change once you learn more about RQ/units of analysis in the field. (Read page 101 Qualitative Researching).

2. While in the field, need to keep track, of what issues you have sufficient data on and which ones you don't. And which data speaks to each issue. Increase sample size as needed.

3. Keep going until you start repeating info/patterns, until you fully understand the RQ

    goal is to understand process, not represent pop

Sampling Bias/Error

*see Designing Surveys ch. 9

1. Sampling frame bias

*ex. Study of depression among widows. You use newspaper to identify local population of widows in recent year. Use these names as your sampling frame. Bias = people who did not appear in newspaper.

Alternative: You get names of widows from a psychiatrist. Biased frame - people on frame are depressed.

2. Response Bias - respondents and non-respondents differ, over or under representation of a group

ex. women and elderly more likely to get mail/answer phone, and respond

ex. households with multiple adults have bigger chance of responding

ex. People with more than 1 phone have unequal probabilities of selection

Ways to avoid:

- Stratify by demographics
- For multiple adults in household: Random sample within sampling unit: Ask for adult with most recent birthday.
- Statistical control: Ask how many phones in household.

3. Design Effects. Bias in obtained samples due to sampling method.

Ex. Cluster samples are less efficient then a simple random sample because each element is no longer selected independently of all others.

Results: Reduction in generalizability.