STA304/1002 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 1: Introduction to Sampling: Basic Concepts & Definitions

Ramya Thinniyam

May 13, 2014

Instructor:

- Dr. Ramya Thinniyam
- Email: ramya.thinniyam@utoronto.ca

  For personal questions
- Phone Number: (416) 978-0673
- Office Location: SS 6025 (Sidney Smith Building)
- Office Hours: Tuesdays and Thursdays 5-6pm (may change/increase before test and project due dates)

Course Outline:

- Textbook: Elementary Survey Sampling by Scheaffer, Mendenhall, Ott (7th edition)

  Covering Chapters 1-9. (Supplemental topics in Chapter 11 if time permits)

- Evaluation:

  • Term Test - 30%
  • Project - 25%
  • Final Exam - 45%

  • Additional homework and data analysis problems will be posted for practice (do NOT have to be handed in).

- Policies on Missed/Late Work:

  • Late projects not accepted.
  • Missed Tests: Declare absence on ROSI on the day of the test and provide documentation to instructor within one week of missed test.
  The missing score will be substituted with your exam.

- See Course Outline for details. You are responsible for understanding and following all policies.

- Check website on Portal/Blackboard for updates/postings/grades/announcements.

Logistics:

- Lectures

  Tuesdays 6-9pm SS 2135
  Thursdays 6-9pm SS 2135

  Notes posted on website before

- Tutorials/Office Hours

  There will be one tutorial on Thur May 15th. Otherwise, no regular tutorials but there will be office hours before assignment due dates and tests. We may have one/two extra tutorials (time permitting).

  TA: Shiva Ashta

  Review concepts, take up homework problems, help with data analysis and projects.
  Introduction to using 'R' in first tutorial.

SURVEY:

- How many of you have taken STA107/257? ECO 227?
- How many of you know what a mean is? Expected value? Variance, etc?
- What discipline are you from? (Sociology, Psychology, Biology, Statistics, Computer Science, etc.)
- What year of study are you in?
- Why are you taking this course? (requirement or elective)
- Are you familiar with using R?

Introduction to Sampling

This course will cover:

- Statistical concepts of taking and analyzing a sample
- Assessing validity of a sample
- Design and analysis of different forms of sample surveys

What is sampling?

Aim of Sampling: We are interested in a population. We take measurements on a sample (subset of the population) to make inferences about the population.

Q: Why do we need to take a sample? Why not use the entire population?

A: Population is not fully obtainable or too expensive/time consuming to take measurements on entire population.

census is not appliable most of time.

Q: What makes a good sample?

A: A sample that is *representative* of the population: characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.

For example: to estimate the GPA of U of T students, would it make sense to take a sample of Biology students only? Males only? 1st year only? Why or why not?

No, not representing the whole students.

Definitions

Observation Unit/ Element: An object on which a measurement is taken. (if it is a human population, element is often called an "individual") each student

Variable: The characteristic that is measured on the element. "GPA"

Target Population: The complete collection of elements we want to make inference about.

Defining the target population is important but sometimes difficult to do. For example: In a political poll, should the target population be all adults eligible to vote? All registered voters? All persons who voted in the last election? The choice depends on what questions you want to answer and will affect the resulting statistics. "U of T students"

☆ Sample: A subset of a population. the target population

This course will discuss what makes a good/bad sample and how to analyze it.

Sampled Population: The collection of all possible observational units that might have been chosen in a sample. ie the population from which the sample was taken. No, but ideally, should be ✓

Sampling Unit: A unit that can be selected for a sample.

For a human population, it does not always have to be the individuals themselves - could be households, schools, etc. Not the unit/element always the same thing

Sampling Frame: A list or specification of all the sampling units. literatually a "list"

Ideally the sampled population = target population. For surveys of people, usually the sampled population is smaller than the target population (because the entire population may not be obtainable).

Example: Absences of Primary School Children

Primary schools in the GTA are being more strict in insisting on school attendance and implementing programs to ensure regular attendance. A study was conducted in the GTA to determine if these programs are working in reducing absences. Households were contacted by telephone and asked to report the number of days the child was absent from school this year. If the household had more than one primary school child, they randomly selected one child. If the household had no primary school children, it was not included. In total, there were 2000 households from which information was collected.

Identify the following:

- Target Population- All Primary school children in GTA
- Element- primary school child (GTA)
- Variable- # of days absent in this academic year.
- Sampling Frame- phone directory / list of house phone #s in GTA
- Sampling Unit- a household
- Sample- 2000 children who were selected

Limitations of the Survey

- What are the possible problems with this study?

1. Undercoverage: not all household are listed on telephone directory (unlisted numbers, those without landlines, etc.)
2. Non-response: certain people may not answer call or the question
3. Include/adjust for confounding factors such as age, race etc.
4. Measurement: people may forget, lie, or misinterpret/count the absences.
5. Make sure numbers selected 'randomly'.

- What is a better way of conducting this study?

Like, go to school, check the attendance sheet for data.
Randomly select 200hundred schools, then randomly select grades, classes...

"cluster sample"

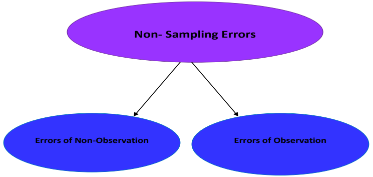
- Identify the above characteristics for your improved sampling method.

Errors

Sampling is not perfect! Errors in estimation can occur:

1. Sampling Errors - Statistical errors, due to randomness. Our sample is not a census (doesn't equal the population) Different samples will yield different results/statistics. Can be estimated with probabilistic statements (margin of errors).
2. Non-sampling Errors - Errors that are not due to randomness, but to do with the way in which sample was selected / data was collected.

"good thing" comparing with the second one.



Errors of Non-Observation: Selection Bias

Selection Bias: occurs when parts of the target population are not included in the sample population or some units are sampled at a different rate than intended.

For example, a survey designed to study household income interview people at a country club (or omit transient persons), etc. This will tend to overestimate the mean/median income.

Sample of Convenience: Units in the sample are selected because they are available and easy to access.

Convenience samples will likely be biased as the units that are easiest to select or most likely to respond are not representative of the harder to select or non-responding units. The units in the sample will often have the same characteristics (not exhaustive of the population's characteristics), hence it is not representative of the population.

Ways in which Selection Bias can Occur:

1. Sampling procedure depends on characteristics associated with the properties of interest (confounding variables)

2. Purposely selecting a representative sample

   Ex. Trying to select the 'average' elements.

   Judgement Sample: Units are selected to be included in the sample by using the investigator's his/her judgement.

3. Misspecifying the target population

   Ex. Election predictions may be wrong if using registered voters from previous years as the target population. The survey would miss new voters and responses from undecided voters.

4. Coverage Errors

- Undercoverage: not including all of the target population in the sampling frame. (more common than overcoverage)

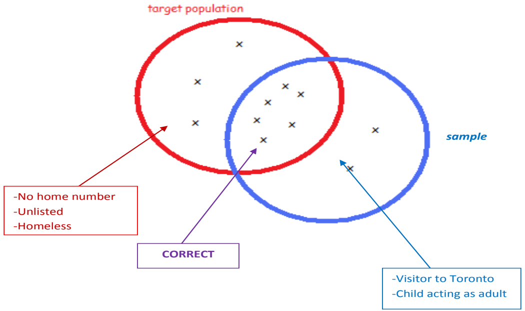
Ex. Telephone survey. How?

- Overcoverage: including units in the sampling frame that are not in the target population. Can occur because of lack of screening or data collector's failure to check sample eligibility

Ex. Telephone survey. How?

Example: Coverage Errors in a Telephone Survey

Target population - all adults in Toronto
Element - adult in Toronto
Sampling frame - telephone book
Sampling unit - telephone number



5. Multiple listing in the sample frame (without adjusting for it in the analysis). Duplicates in the sampling frame will bias results if the same units are sampled more than once.

   Ex. Telephone survey to estimate household size or income. How?

6. Non-response: failure receive response from all in the sample - element refuses to respond, cannot respond, or cannot be reached.

7. Sample of almost all volunteers

   Ex. Call-in polls, online surveys. People with strong opinions tend to volunteer responses.

8. Substituting a convenient member of the population for another who is not available

* Large sample size does NOT mean it is a good sample. Design is far more important! *

Errors of Observation

Need accurate responses in a sample.

Measurement Bias occurs when the measuring instrument tends to differ from the true value in one direction.

Measurement errors can occur due to inaccurate responses from respondents or wrong measurements and/or poor survey design from investigators.

Ex. Wrong measurements - bird counting, agricultural studies, vegetation.

Measurement Error in Surveys of People

Measurement error is a problem for all types of surveys, particularly for surveys involving people because:

- People sometimes lie. Ex. sensitive matters or to obtain certain outcomes
- People do not always understand the question. (due to wording or respondent's understanding)
- People forget

  Telescoping: Respondent's misspecification of incidents. People may not remember exactly when asked about past experiences - durations, dates, frequency of events.
- People may give different answers to different interviewers
- People may answer to please or impress interviewer or to avoid embarrassment

  People tend to agree - studies have shown that the same people agreed to 2 contradictory questions.
  People do not want to admit to taboos, sensitive questions, controversial issues
- Interviewers may misread questions, record incorrectly, antagonize respondent
- Words have different meanings to different people

  Ex. "Do you own a car? " - *you*, *own*, *car* all have different interpretations.....
- Question wording and order have impact on answers. Ex. double negatives, confusing sentence structure, etc.

Reducing Errors

- Reducing Selection Bias:
  - Use Probability Sampling Methods - Chapter 2
- Reducing Non-Response Bias:
  - Callbacks
  - Rewards and Incentives
- Reducing Measurement Error:
  - Careful questionnaire design
  - Testing survey equipment
  - Training interviewers
  - Pretesting survey
  - Check for accuracy in respondent’s data

Example: Abortion Rights Groups Surveying Voters

Source: *Article by Jack Coffman, 1989.*

The U.S. Supreme Court has increased abortion restrictions in some states which appears to open way for action by state legislators. Backers of the abortion rights survey plan to use this survey's results to head off any further anti-abortion legislation. The survey was sponsored by several abortion rights groups and conducted under a consulting firm owned by a state Senator. Volunteers to conduct telephone surveys were not hard to find as many people who are pro-choice and depended on the Supreme Court were worried about a woman's right to choose. More than 7,000 volunteers operated six telephone centers in Minnesota in an effort to contact the families of every registered voter to determine how the voters feel about abortion. So far, about 160,000 families were contacted.

When volunteers phoned registered voters, they asked:
1. *"Do you agree or disagree with the following statement: The decision to terminate a pregnancy is a private matter between a woman, her family, and doctor .... and not a decision to be made by government and politicians?"*

If the person answered 'yes' to 1. , they were then asked :
2. *"In light of the current government threats to safe, legal abortion ... will this influence your opinions of politicians in the future?"*

If the person answered 'no' to 1. , they were then asked :
2. *"Are you opposed to abortion in cases of rape, incest, serious fatal deformity, or to save a woman's life?"*

Example: Abortion Rights Groups Surveying Voters (con'd...)

a) Identify the following:

- Target Population- registered voter
- Sampling Frame- list of residential telephone #s in Minnesota
- Sampling Unit- a household with at least 1 registered voter
- Element- a registered voter

b) Discuss in detail possible sources of sampling and non-sampling errors. Use correct statistical terminology and describe in words.

Example: Abortion Rights Groups Surveying Voters (con'd...)

a) Identify the following:

- Target Population-
- Sampling Frame-
- Sampling Unit-
- Element-

B.

Sampling errors: since sample is not a census and expect to have sample to sample variation

Non sampling errors:

- errors of observation

Measurement bias because:

1). Leading/loaded questions-promoting 'pro-choice' with phrases such as 'terminate pregnancy''private matter''private threats'

Open-ended. Question would be better than agree/disagree.

2). Survey conducted by mostly pro-choice volunteers: without training, they can influence respondents' answers or record incorrectly

- errors of non observation

1). Section bias: undercover age since only household with listed numbers could be contacted/overcoverage since voters' status of the respondent was not cleared.

2). Non-respondence: no indication of now refusals/unanswered calls were dealt with. Can refuse after hearing survey topic.

3). Over-representation of women/elderly people-they are possibly more likely to answer/be at home

b) Discuss in detail possible sources of sampling and non-sampling errors. Use correct statistical terminology and describe in words.

c) Briefly describe how the survey could be improved to obtain reasonable data.