
36-303: Sampling, Surveys and Society

Sampling Frame Review; Questions in Surveys
Brian W. Junker
132E Baker Hall
Brian@stat.cmu.edu

Handouts & Other Things...

- Handouts...
 - These Lecture Notes
- Other Things...
 - Homework solutions are a little slow in coming (this is partly my fault!).
 - I hope to have things caught up by the weekend.
 - Next week we have a midterm test!
 - I will talk a little about the midterm on Thursday, and more next Tuesday.
- Later today or tomorrow I will EMAIL
 - Update on Team Project Assignments
 - Practice problems for calculations with FPC

Outline

- Team Project Progress
- Population, Sampling Frame, Random Sample **[Important for Team Assig. I.4!]**
- Non-response Strategies
- Questions and Answers in Surveys **[Important for Team Assig. I.5!]**

Team Project Progress

- All teams have received feedback on their project I.2 assignments from me.
 - For I.3, EMAIL me a single document (ms word or pdf), that contains complete answers to items A-G for the ONE TOPIC you chose for your 303 project, revised in any way you think is necessary based on my comments and your discussions in your group. *Due this Thursday by midnight.*

- Next Tuesday Feb 15:
 - Team working agreements are due
 - Team Project Assignment I.4 is due (more on this in email I will send later this week)

Population, Sampling Frame, Random Sample

- **Target Population**: The population about which you can make valid inferences from a well-designed survey within your means.
- **Sampling Frame**: A real or theoretical list of all possible individuals in the target population, that you could randomly sample in your survey.
 - Hopefully, differs only in small ways from target pop.
- **Random Sample**: A random, probability based sample (for us, usually SRS without replacement) from the sampling frame.

The only guarantee of a representative sample, that we trust statistical calculations with, is a truly random, probability-based sample from a sampling frame with low coverage error for the target population.

Population and Sampling Frame

- If your population is “all undergraduates at Carnegie Mellon”, then your sampling frame should be one of these (most likely):
 - List of email addresses obtained from Hub or elsewhere; or
 - C-Book student directory
- If your population is “all residents of XYZ part of Pittsburgh” then your sampling frame should be (most likely):
 - A list of addresses that you can visit or send mail to; or
 - A list of phone numbers you can access via random digit dialing

A Good Sampling Frame Simplifies Representative Sampling

- It should be a real or theoretical list that has low coverage error
 - It should contain almost exactly the same individuals as the full target population
- It should be possible to select a real, live random sample from the frame
 - A frame like “all students passing the Fence between 12:00 and 1:00” has obvious problems (coverage error and statistical calculations may not apply).
 - In some extreme cases it is not possible to build a frame and do random sampling (e.g. survey of panhandlers, or survey of people while they are riding the 28X).

Goals for the Random Sample

- For most of the projects, the Random Sample should be an **SRS without replacement** (urn model!) from an explicit sampling frame.
- Some projects may have natural strata (major department of student, fr/so/jr/sr, or 3-4 different college campuses). In that case take an **SRS w/o repl.** from each stratum.
- In some extreme cases it is not possible to build a frame and do random sampling (e.g. survey of panhandlers, survey of riders while they are on 28X).

Non-Response Strategies

- Once you have chosen the random sample, respondent can screw up representativeness by forgetting or refusing to respond to your survey.
- For mail and email surveys, common strategies are
 - Pre-survey announcements
 - Followup reminders
 - Other methods?
- For face to face and telephone surveys, practice:
 - How to pull respondent in, in first 10-30 seconds;
 - How to keep respondent engaged for whole interview
- How many times to re-contact a dead-end in sample?
- IRB: maintain right to refuse, quit early, etc.

Two Fractions: Sampling Fraction vs. Response Rate

- The sampling fraction $f = n/N$ (n = sample size; N = population size)
 - Determines variability of sample;
 - $FPC = \sqrt{1 - f}$ for SE's and similar quantities
- The response rate r/n (r = # who responded; n = number in sample).
 - If the sample is random and $r/n \approx 1$, then the r respondents are probably representative of population
 - If either nonrandom sample, or $r/n \ll 1$, then the respondents are probably not representative of population
- It is much more important for $r/n \approx 1$ than n/N .

Example: Student use/attitudes toward drug/alcohol use (Fictional!)

- Target Population: All currently enrolled undergrads at the Pittsburgh campus of CMU
- Sampling Plan: Advertise on Facebook inviting students to come to www.surveymonkey.com to fill out survey.
 - No sampling frame specified
 - Two sources of coverage errors
 - Not everyone is on Facebook
 - Volunteers are different
 - Nonrandom sample – no way to claim representativeness unless n/N is very close to 1!
 - No plan to identify nonresponders or followup with reminders

Example: Student use/attitudes toward drug/alcohol use (Fictional!)

- Target Population: All currently enrolled undergrads at the Pittsburgh campus of CMU
- Sampling Plan: Take SRS w/o replacement from C-book, email those students to do Surveymonkey survey, email reminders to nonresponders after 1 week.
 - Sampling Frame C-Book is a list of individuals in target pop
 - Low Coverage Error
 - Students who provide wrong or late information for C-Book
 - Random sample – makes sample representative of frame, and because of low coverage error, sample is also representative of target population
 - Followup Nonresponders with email reminder
 - Increase response rate
 - Decrease appearance of confidentiality

Example: Use/attitudes toward PAT bus service (Fictional!)

- **Target Population**: East End Residents (Oakland, Shadyside, Squirrel Hill, Point Breeze)
- **Sampling Plan**: Approach people at bus stop with questionnaire
 - **No sampling frame** specified
 - **Two sources of coverage error**
 - Time/place → who is there (workers, single moms w/kids, ...)
 - Noncoverage of non-riders or infrequent riders
 - **Nonrandom sample** → can't argue “representative”
 - No definition or followup plan for **nonresponders**

Example: Use/attitudes toward PAT bus service (Fictional!)

- **Target Population**: East End Residents (Oakland, Shadyside, Squirrel Hill, Point Breeze)
- **Sampling Plan**: Identify all East End phone prefixes (361, 362, 682, ...), randomly select prefix and 4-digit suffix (ppp-ssss); phone interview starts by verifying residence and then asks about buses.
 - **Sampling Frame** East-end telephone land lines
 - **Lower Coverage Error**
 - Frame + location screen question → subset of target pop
 - Noncoverage of persons w/o landlines (Groves, sect 4.8!)
 - **Random sample** → Easy “representativeness” argument
 - **Nonresponders not mentioned** (but one could call back non-answering numbers [how many times?])

Accessible Sources of Sampling Frames: Face to Face or Mail-Back

■ Residential addresses

- Phone directories are easy, but may have under-coverage problems
- Elaborate schemes based on random selection of block within area, & random selection of house within block

■ Commercial addresses

- Free online services provide addresses of specific business types within x miles of a particular location – under-coverage problems?
- For storefronts, the random block/random storefront scheme can work

■ Man on the Street interviews

- Coverage problems based on respondents' habits
 - Volunteer self-selection problems
-

Accessible Sources of Sampling Frames: Telephone Interviews

- Phone books
 - Easy but can have under-coverage problems
 - University directories can be better
- Random digit dialing
 - Easy to formulate – select a valid prefix at random, then select a random 4-digit suffix (ppp-ssss)
 - Under-coverage (land-lines only)
 - Ineligible numbers (residential vs. business vs. fax, disconnected): Groves 4.8 estimates 6-7 dead-end calls needed for every “good” call
- Advertising a number for respondents to dial in
 - Coverage problems based on where you advertise
 - Volunteer self-selection problems

Accessible Sources of Sampling Frames:

Email/Web Survey (Initial Considerations)

- Can be appropriate for “connected” populations (not yet for “general public”) – University students, online workers, online entertainment users, etc.
 - Sampling frame must be a list of email and/or paper mail addresses
 - University directories (e.g. C-Book)
 - Other email lists from registrar’s office etc.
 - Take random sample from frame, then invite through email and/or paper mail
 - Passive advertisement (Facebook etc.)
 - Under-coverage
 - Volunteer self-selection
 - Facebook can be ok to contact people you have sampled from some other frame, but is not itself a good frame.
-

Accessible Sources of Sampling Frames: Email/Web (Implementation)

- Can embed survey in email and ask respondents to email responses back to you
- Or use an online service, e.g.:
 - ❑ www.surveymonkey.com
 - ❑ www.infopoll.com
 - ❑ www.surveysaid.com
 - ❑ questionpro.com
- Decrease non-response with email and/or paper mail reminders
 - ❑ In order to have non-response followup, you need to know who in your sample has already responded!
 - ❑ Think of ways to assure confidentiality **anyway**.

Non-Framed, Non-Random Samples

■ Non-Framed Samples

□ Challenge: Coverage error

- *Why should anyone believe your sampling method provides good coverage of your target population?*

■ Non-Random Samples

□ Challenge #1: Selection bias

- *What makes people eligible for your survey?*
- *How do you choose among the eligible units?*

□ Challenge #2: Using standard statistical formulae

- *How can we be sure that the sample is large enough to provide good population estimates?*
- *How can we be sure formulae for means, variances, confidence intervals, etc., do not need further modification?*

Non-Framed Samples

- Describe all of the locations, times, and methods of approaching respondents in great detail.
 - ❑ Coverage: Is everyone in your target population accessible at these times and places?
 - ❑ Equally-likely sampling: Is everyone equally likely to be there when you are there?
 - ❑ Eligibility: How will you determine whether to include this respondent in your survey.
 - Target population undergrads, this guy has grey hair and a paunch (are there undergrads like that?)
 - Are there reliable ways to determine eligibility?

Non-Random Samples: Selection Bias

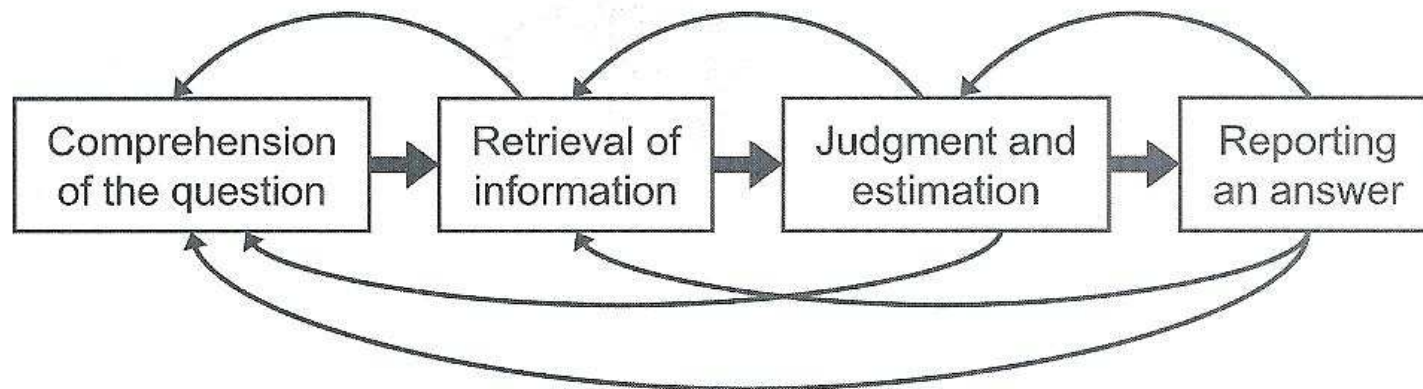
- Volunteer bias (E.g. Facebook or general email invitation to a non-targeted sample)
- Interviewer bias (E.g. you never approach unattractive respondents at the Fence).
- Fix:
 - ❑ Decide on a rule for targeting subjects in advance (select every fifth person who passes by the Fence) and stick with it, no matter what.
 - ❑ Decide how you will follow-up nonrespondents (just like with a targeted random sample!)

Non-Random Samples: Using Standard Statistical Formulae

- Show that your non-random sample “behaves” like a random sample
 - Means match gender, age, college class, income, etc. features of the target population
 - Variances match variances you would expect from a random sample
- As further protection against coverage error, take a larger sample
 - (Sue & Ritter, 2007, *Conducting On-line Surveys*, p. 34):
 - Useful sample sizes are typically 30-500
 - Within that range, sample roughly 10% of total population
 - Sample should be roughly 10 times larger than number of variables being studied
 - Choose the largest sample you can afford

Questions and Answers

- A simple model of the response process (Groves, Ch 7):



- Survey question should be written and refined to
 - Increase validity (reduce errors and misunderstanding)
 - Increase reliability (get the same answer every time)at every stage in the response process!

Comprehension of the Question

- ***Does respondent understand what you intend by the question?***
- Some possible problems:
 - ❑ Not possessing information needed for question
 - ❑ Misunderstanding question wording
 - Grammatical errors or style
 - Too much complexity
 - Question contains false or unproductive assumptions or inferences
 - Vague or unfamiliar concepts, quantifiers, terms

Retrieval of Information

- ***Can respondent recall information from long-term memory?***
- Some possible problems:
 - ❑ Mismatches between terms in question and terms in respondent's memory
 - ❑ Retrieval failures (I forget...)
 - ❑ Distortion or poor reconstruction of remembered events as time goes by
 - Time dilation
 - Rehearsal (or avoidance!) of significant memories

Judgment and Estimation

- ***How does respondent combine, edit, fill in, information needed to answer question?***
- Typical estimation methods for incomplete quantitative memories:
 - **Exact answer:** I just did my taxes, so I know my income is...
 - **Recall-and-count:** Recall events and count them up, add a few in case I forgot some
 - **Rate-based:** Recall the typical rate at which the events occur, and multiply by the time period
 - **Impression-based:** Start with a vague impression (few, some, a lot) and translate into a quantitative estimate
 - Over- & under-reporting – collect validation data!

Judgment and Estimation (continued)

- Typical judgment methods for attitudes:
 - **Deep impressions:** respondent has thought a lot and has deep evidence and reasoning to support attitude
 - **Shallow impressions:** “Gee I was just reading about PAT buses and they don’t sound very reliable.”
 - **Top-down judgement:** “I believe in the free market generally, so I think everyone should pay for their own bus pass.”
 - **Bottom-up judgement:** “I remember being a poor student without access to transportation, so I think people with means should be taxed to make free bus passes available for students.”
 - **Reaction to question wording:**
 - Do you think the United States made a mistake in deciding to defend Korea? [Gallup]
 - Do you think the United States was right or wrong in sending American troops to stop the Communist invasion of South Korea? [NORC]

Reporting the Answer

- ***What does respondent select to respond?***
- Question Format
 - ❑ Open-ended questions (numerical or verbal)
 - ❑ Closed questions with ordered scale (Likert scale)
 - ❑ Closed questions with categorical responses (M/C)
- Failure to Follow Instructions
 - ❑ Comprehension of instructions
- More or less deliberate misreporting & nonresponse
 - ❑ Sensitive questions
 - ❑ Desire to mislead polling organization
 - ❑ Undervalue poll or polling agent (*“it’s just a student project...”*)

Review

- Team Project Progress
 - Population, Sampling Frame, Random Sample
 - Non-response Strategies
 - Questions and Answers in Surveys
-
- I.3 Due in email this Thursday by Midnight
 - I.4 Due next Tues (read revised project schedule carefully – it will be in your email!)
 - Midterm Exam Next Thurs Feb 18