Class 3: Sampling & Data Collection Methods

MAST5953: Creating Your Own Data

Dr. Miriam Sorace

www.miriamsorace.net

23 November 2022

Outline of Today's Class

Collecting Survey Data

Collecting Survey Responses using Google Forms

1. You need a Gmail account
2. You need the list of respondents' email addresses
3. Create your survey
4. Copy the survey link to the invitation email
5. Download the survey responses in .csv format
   - Responses tab - excel icon at the top - File - Download (as .csv)

Google Forms

- Google Forms Demonstration

Coding/Survey Data Entry

Coding

- " the translation of nonnumeric material into numeric data" (Groves et al. 2009: 331)

Coding
Example

Indicate your current status (Check one box):

- ☐ Full-time student
- ☐ Part-time student
- ☐ Applicant, acceptance letter received
- ☐ Applicant, acceptance letter not received

Here, the coding step merely involves assigning a distinct number to each of the four possible answers (e.g., 1 = full time, 2 = part time, 3 = applicant with letter, 4 = applicant without letter). These numbers then become the values in a field of the electronic data file eventually produced.

Coding
Open-Ended Question Example

- Consider this survey item instead:
  - *"What do you think are the most important issues facing [COUNTRY NAME] at the moment?"*
- It can be asked as a (very) long multiple-choice question, but to avoid satisficing and to gain original insights from respondents this is also often asked as an open-ended question - leaving a text field open to respondents.
- Coding here entails (a) retrieving topics via some form of text analysis; (b) assigning codes to the topics.

Survey Data Entry
Best Practices

- Use consistent ID variable coding from one dataset to the next (e.g. individual, country, region)
- Keep the original, un-amended masterfile for record and create a copy that you will use analysis
- Keep any identifying information separate from the masterfile.
- Use standard codes for DKs, refused, and missing answers [typically: 997/998/999 - do not use '0' for this!]
- Create a codebook where all values and their labels are recorded for every variable.
- Run numerical summaries of every variable to check for outlying numbers or any coding error.

In-Class Exercise

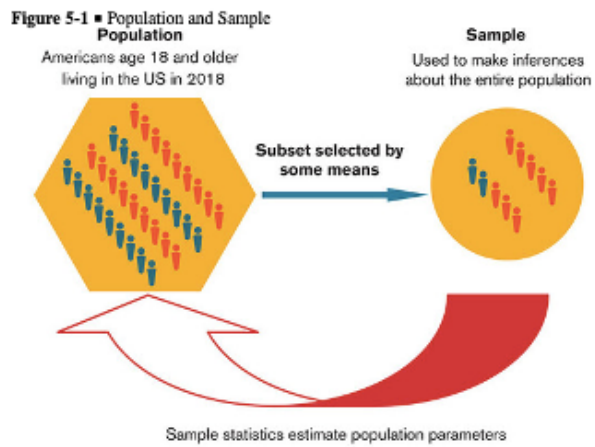
Cognitive Interviewing & Validation

1. See RMarkdown (.Rmd) file in Class 3 folder

Sampling

Population vs. Sampling

1. **Population**: all individuals of a defined group (e.g. all US citizens).
   - Government Census
   - Earliest surveys: e.g. Charles Booth *Life and Labour of the People of London*
   - *Pitfalls*: Impractical/costly
2. **Sample**: a subset of individuals from the population of interest



[1]

[1] Figure taken from Buttolph Johnson, J. et al. (2020) *Political Science Research Methods, 9th Edition*, Sage/CQ Press

How to Sample?
Kirk & the Gallup Poll

Click here for video[2]

[2] Gilmore Girls (2004) *Tippecanoe and Taylor, Too* Season 5, Episode 4

How to Sample?
A Primer

- Using the 'fraction of the population' rule of thumb is wrong - sample size calculations are more complex than that (SurveyMonkey has a nice interactive tool that helps with that: click here to check it out!).
- The larger the sample, the better: but simply going for a very large sample size won't necessarily help you either: you need **probability random sampling**!

Probability Sampling

- Each population unit has an **equal** and **non-zero** probability of selection into the sample.
  1. The researcher procures a representative **sampling frame** (the complete list of elements in the population)
  2. The researcher uses a table of random numbers/a **random number generator** to select sample units from the main sampling frame

Types of Probability Sampling

- **simple**: all units in the sampling frame are sampled at once
- **stratified**: the units in the sampling frame are first split into mutually exclusive *strata*, then random probability sampling within each stratum).
- **clustered**: instead of sampling population *units*, we can consider sampling *areas* or 'clusters' (i.e. households).
- **oversampling/disproportionate sampling**: assigning unequal probabilities to individuals from particular sub-groups of interest to the researcher.

(Non-Probability) Sampling Procedures

Purposive Selection

- Quota sampling
  - Guided by a set of characteristics but units are not representative/selected randomly
- Convenience sampling
  - Guided by 'who you know'/availability
- Snowball sampling
  - Initial recruits offer new names
- They do not adhere to probability sampling and hence cannot benefit from the existence of a theoretical basis to measure estimates variability and bias precisely. No statistical basis to support inference made from non-probability samples.

Reasons **not** to deviate from probability sampling

The Literary Digest Poll: the perils of sampling and non-response error

UPenn Literary Digest Case Study Link

Reasons **not** to deviate from probability sampling

Impossible to calculate standard errors and confidence intervals

- Laws of probability can only apply if samples are drawn at random from the target population - e.g. Fisher (1925), Neyman (1934).
- The sample mean is an unbiased estimator of the true population mean **only if** the sample was randomly selected from the target population.
  - The 'miracle' of probability sampling: the Central Limit Theorem
    - "The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large **random** samples from the population [...], then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n $>$ 30)" (LaMorte 2016).

Ball Pit Exercise

- You have an enormous ball pit and you want to estimate the % of blue units and the % of red units.
- Besides counting each and every ball - which would be unfeasible/too costly - how would you do it?

Ball Pit Exercise

- Use a randomizer to select the sample
- You repeat this sampling exercise several times and record the % of blue units for every sample
- You plot the means (or proportions) of all samples, to construct a *sampling distribution*
- The true population mean will be the sampling distribution mean
- But this is equally time-consuming ... could you do this by sampling *only once?*
  - Yes!

Example: The Ball Pit

Estimating Population Parameters from a Sample

- *Randomly* select 1 sample.
- You calculate that the proportion of blue units is 0.46 (or 46%) the proportion of red units is 1-0.46 = 0.54 (or 54%).
- What is the true population proportion?
- We cannot provide the definitive population mean, but we provide a *range* - i.e. the **confidence interval**.

Example: The Ball Pit

Building the Confidence Interval

- Thanks to the Central Limit Theorem, we **know** that our sample mean comes from a Normally distributed sampling distribution (because the sample was randomly selected!).
  - The property of Normal distributions is that 95% of values lie within + or - 2 standard deviations.
  - Knowing that 95% of sample means will lie within +/- 2 standard errors from the mean of the sampling distribution is useful: it allows us to build a **confidence interval** around our sample estimates & to make inferences about the population.

Example: The Ball Pit

Building the Confidence Interval

- We only therefore need to calculate an extra parameter from our sample: **the standard error**, which is built using the sample standard deviation.
- For proportions, the standard deviation is calculated as p-(1-p) - i.e. 0.46-0.54 in our case.
- We then divide this by the square root of our sample size (e.g. 100) and multiply by 2 (the 2 standard deviations needed for a 95% confidence level) and we get our lower and upper bound estimates!

- 46% sampled are blue
- p=0.46
- $\sigma_p \frac{\sqrt{0.46*0.54}}{\sqrt{100}} = 0.05$
- $\text{ME} = 2*0.05 = 0.1$
- $\text{CI}_{0.95} = [0.45;0.47]$

Example: The Ball Pit

Interpreting the Confidence Interval

- 46% sampled are blue
- p=0.46
- $\sigma_p \frac{\sqrt{0.46*0.54}}{\sqrt{100}} = 0.05$
- ME = 2*0.05 = 0.1
- $CI_{0.95} = [0.45;0.47]$

- 95% of the confidence intervals thus calculated will contain the TRUE population mean.
- there is thus only a 5% chance that the range 0.45 to 0.47 excludes the mean of the population.

So much ado about ... what?

- All this to demonstrate that **random selection** of units into samples is fundamental to derive unbiased estimates of population statistics and allows to estimate CIs
- To determine how large the sample size needs to be, you need more than 'N': you need estimates of variability and error you are willing to tolerate.
  - Better question is: how many units can I *afford*?

What did we learn today?

- How to collect survey responses using Google Forms
- How to code survey responses to make them ready for analysis
- How to validate survey items
- How to sample survey respondents