Class 5: Weighting & Ethics

MAST5953: Creating Your Own Data

Dr. Miriam Sorace

www.miriamsorace.net

14 December 2022

Outline of Today's Class

Total Survey Error - Part 2

Survey Weighting & Multiple Imputation

Ethics

Total Survey Error - Part 2

Sampling Error
The Total Survey Error Framework

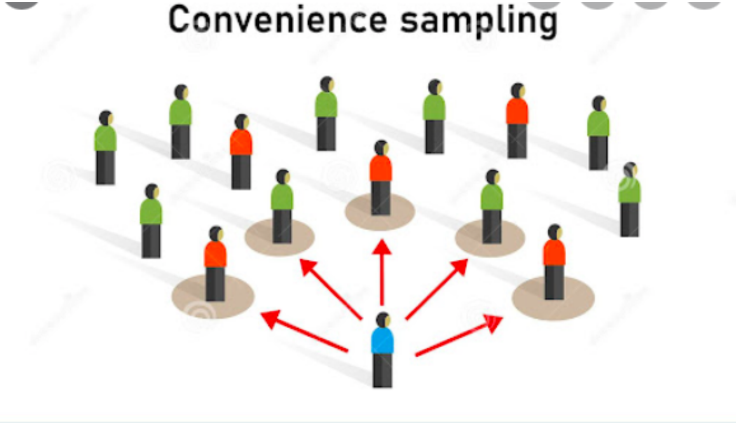


Image from: ResearchArticles.com

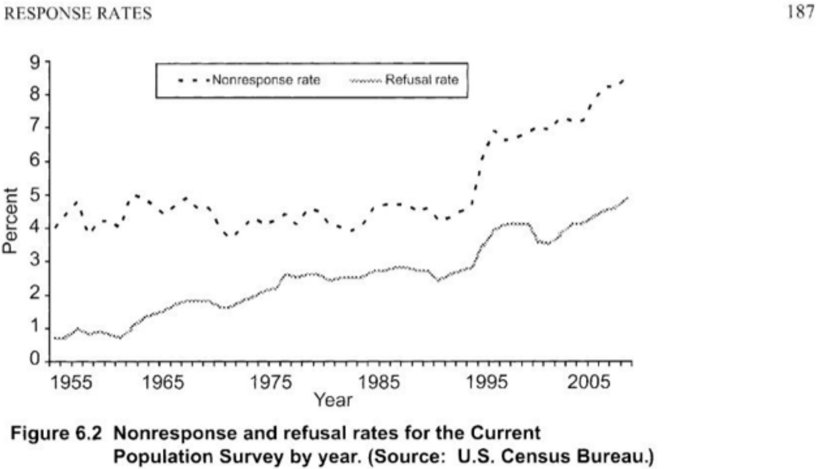
Coverage Error

The Total Survey Error Framework

The Literary Digest
NOVEMBER 14, 1936
Thirty CENTS
IS OUR FACE RED!

Nonresponse Error

The Total Survey Error Framework



RESPONSE RATES 187

Figure 6.2 Nonresponse and refusal rates for the Current Population Survey by year. (Source: U.S. Census Bureau.)

Survey Weighting & Multiple Imputation

Post-Survey Statistical Adjustments

- weighting
- multiple imputation

Weighting

- “ [...] the adjustment of computations of survey statistics to counteract harmful effects of noncoverage, nonresponse, or unequal probabilities of selection into the sample” (Groves et al. 2009: 331)
- “Weighting of survey data is required to ‘map’ the sample back to an unbiased representation of the survey population” (Heeringa et al. 2017: 38)

Differential Selection Probability Weighting

Example

- Need to collect a sample of 125k of the US population (total 199,500,000)
  - The proportion of Latinos in the population is 1/8 (or 12.5%)
- Need to over-sample Latinos (raising their probability of selection to 1/2, or 50%) to carry out some group-specific analysis
  - Instead of 15,625 Latinos in the sample, we will have 62,500
- **Problem**: the aggregate descriptive statistics are not representative of the US population unless respondents are weighted to bring back the Latino probability to 1/8. (any analysis separate by sub-group would be fine though).

Differential Selection Probability Weighting

Example

- What to do?
  - Simply put: Weight each observation by its original proportion in the population.
    - Latinos weight = 0.125
    - Non-Latinos weight = (1-0.125) = 0.875
  - Which more simply means that each Non-Latino observation needs to be multiplied by 7 (i.e. assign Latinos a weight of 1 and Non-Latinos a weight of 7):
    - Non-Latinos weight = 0.875/0.125 = 7

Differential Selection Probability Weighting

Example

- What to do?
  - More technically: weight each observation by the *inverse of its probability of selection*
    - Latinos: 1/(their N in sample / N tot Latino population) = 1/[62500/(199,500,000*0.125)] = 1/(62,500/24,927,500) = 399
    - Non-Latinos: 1/(their N in sample / N tot Non-Latino population) = 1/[62500/(199,500,000*0.875)] = 1/(62,500/174,562,500) = 2,793
  - Which also translates to assigning a weight of 7 to Non-Latinos and a weight of 1 to Latinos (2793/399) = 7

Non-Response Adjustment Weighting

Example

- Imagine that Latinos respond at an 80% rate and Non-Latinos respond at a 99% rate, if one assumes that - within Latinos and Non-Latinos - respondents are a random sample of all sampled persons in the group (missing at random assumption), this 0.80 or 0.99 probabilities can be thought of as *sampling rates*
- What to do?
  - Weight each observation by the *inverse of its sampling rate*:
    - Latinos: $1/0.80 = 1.25$
    - All the rest: $1/0.99 = 1.01$
  - Which then can be cumulated to the original sampling weights via multiplication, to form the below weights:
    - Latinos: 1*1.25 = 1.25
    - Non-Latinos: 7*1.01 = 7.07

Weighting

- "Generally, the *final survey weights* in survey data sets (sometimes referred to as *estimation weights*) are the product of sample selection weights, non-response adjustment factor, and the poststratification factor" (Heeringa et al. 2017: 38)
    1. Apply *design weights* - i.e. adjustments for differential sampling probabilities
    2. Apply *non-response corrections* - i.e. response propensities
    3. Apply *calibration methods*: i.e. raking or poststratification to make sample conform to known auxiliary (extra-survey) population variable distributions
- " The computation of [estimation weights] is therefore an accounting function, requiring only multiplication of the probabilities of selection at each stage of sampling and then taking the reciprocal of the product of the probabilities" (Heeringa et al. 2017: 39)

Exercise: Applied Example

ESS

- Go to the link below: what is the difference among all the weights? What survey errors do they account for?
- ESS Weights

Multiple Imputation

- This method is used for *item non-response*
- Typically analysts/statistical softwares ignore the missing cases, i.e. engage in *casewise deletion* of missing data
- Effectively, casewise deletion is a form of imputation: it simply assumes that each of the missing cases have average values on each covariate
- An alternative is building an explicit multiple imputation model, although many consider this alternative akin to “data fabrication”.

Multiple Imputation

- It entails regression prediction for the missing value.
- It requires that all variables used as predictors in the analytical model are also included on the right-handside as predictors of the missing value.
- It generates series of multiply imputed datasets each with slightly different realizations of the multiple imputation model
- Variation in estimates across the multiple datasets then allows for the estimation of overall variation

Ethics

Ethical Blunders

3 Identical Strangers

- Study by Dr Peter Neubauer, NYU Psychiatric Institute, for a nature versus nurture twin study.

Ethical Blunders

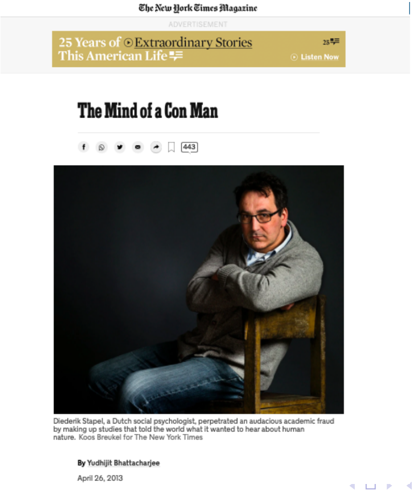
Milgram's Obedience Study

- Study by Dr Stanley Milgram , Yale University, to test the impact of authority on behaviour.

Ethical Blunders

Data Fabrication: The Stapel Case

- NYTimes The Mind of a Con Man



The New York Times Magazine

The Mind of a Con Man

443

Diederik Stapel, a Dutch social psychologist, perpetrated an audacious academic fraud by making up studies that told the world what it wanted to hear about human nature. Koos Breukel for The New York Times

By Yudhijit Bhattacharjee

April 26, 2013

Ethics Principles

- Check out UKRIO's Code of Practice for Research
- From: Groves et al (2009:373):

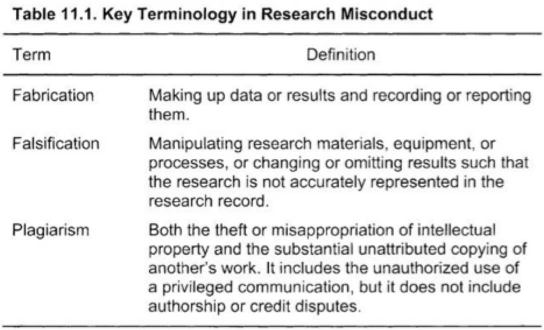


Table 11.1. Key Terminology in Research Misconduct

Term	Definition
Fabrication	Making up data or results and recording or reporting them.
Falsification	Manipulating research materials, equipment, or processes, or changing or omitting results such that the research is not accurately represented in the research record.
Plagiarism	Both the theft or misappropriation of intellectual property and the substantial unattributed copying of another's work. It includes the unauthorized use of a privileged communication, but it does not include authorship or credit disputes.

Key Standards

- Hold client/participant information confidential
- Obtain informed consent
- Be transparent: describe data collection method, survey codebook, survey questions, target population and sampling frames/non-response rates
- Avoid deception
- Consider and mitigate any potential harm to participants
  - Avoid practices that may harm, humiliate or seriously mislead participants.

Core Principles:

- **Beneficence**: minimise possible harm and maximise possible benefits for participants
- **Justice**: do not over-burden some groups in society with research requests
- **Respect**: do not use force, fraud, duress or any other form of coercion, ensure informed and free consent to participate.
  - Avoid to threaten voluntary consent in any way: incentives given to participants must not be coercive or leave them little choice.

Informed Consent

Essential Elements

1. Brief statement on purpose of research and its sponsors/PI
2. Describe duration and the basic procedures involved
3. Describe possible risks/dangers & whether compensation is foreseen (in case of major risks)
    - risks that personal data gets public also need to be considered! Data protection rules.
4. Describe benefits/remuneration to participants
5. Statement on protection of confidentiality
6. Contact information
7. Statement highlighting that participation is entirely voluntary and can be discontinued at any time without penalty

Informed Consent

Some Findings

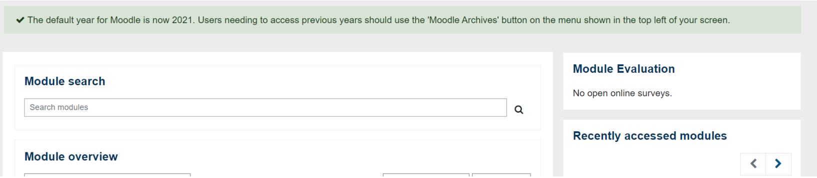
1. Do not include the study's hypotheses not to lead responders
2. Requiring a signature deters participation, so simply clicking a check box is sufficient

The Ethics Checklist & The Consent Form

- Course GitHub Page

Reminder: Survey is on your Moodle Page!

- Check 'Module Evaluation' Box



✓ The default year for Moodle is now 2021. Users needing to access previous years should use the 'Moodle Archives' button on the menu shown in the top left of your screen.

Module search

Module overview

Module Evaluation

No open online surveys.

Recently accessed modules