Survey research in the digital age

Bernhard Clemm von Hohenberg
Department of Computational Social Science
GESIS

Summer Institutes in Computational Social Science
July 28, 2023

Schedule

- 9.00-9.45 Introduction & total error survey framework
- 9.45-10.15 Probability and non-probability sampling
- Coffee break
- 10.30-11.00 Computer-administered interviewing
- 11.00-11.30 Linking surveys to big data
- 11:30-13:00 Intro and begin group exercise
- Lunch (or Eisbach plunge)
- 14:00-15:45 Continue group exercise

Credits

These materials build heavily on Matthew Salganik's 2019 SICSS class as well as Chapter 3 of "Bit by Bit: Social Research in the Digital Age".

Motivating study

Online, Opt-in Surveys: Fast and Cheap, but are they Accurate?

Sharad Goel
Stanford University
scgoel@stanford.edu

Adam Obeng
Columbia University
adam.obeng@columbia.edu

David Rothschild
Microsoft Research
davidmr@microsoft.com

https://5harad.com/papers/dirtysurveys.pdf

Activity

- Design a questionnaire using questions already asked in high-quality, probability-based surveys (i.e., Eurobarometer)

Activity

- Design a questionnaire using questions already asked in high-quality, probability-based surveys (i.e., Eurobarometer)
- Recruit participants from Prolific and have them complete your questionnaire

Activity

- Design a questionnaire using questions already asked in high-quality, probability-based surveys (i.e., Eurobarometer)
- Recruit participants from Prolific and have them complete your questionnaire
- Compare results from your survey to the results from Eurobarometer

Activity

- Design a questionnaire using questions already asked in high-quality, probability-based surveys (i.e., Eurobarometer)
- Recruit participants from Prolific and have them complete your questionnaire
- Compare results from your survey to the results from Eurobarometer
- Try different approaches to weighting and see how the change the estimates

Learnings

This activity will give you practice:

- Designing questionnaires and working with survey software (Google Forms)

Learnings

This activity will give you practice:

- Designing questionnaires and working with survey software (Google Forms)
- Collecting survey data on recruitment platforms (Prolific)

Learnings

This activity will give you practice:

- Designing questionnaires and working with survey software (Google Forms)
- Collecting survey data on recruitment platforms (Prolific)
- Pre-processing/analyzing survey survey data (R)

Learnings

This activity will give you practice:

- Designing questionnaires and working with survey software (Google Forms)
- Collecting survey data on recruitment platforms (Prolific)
- Pre-processing/analyzing survey survey data (R)
- Applying post-stratification and thinking along total survey error framework

Learnings

This activity will give you practice:

- Designing questionnaires and working with survey software (Google Forms)
- Collecting survey data on recruitment platforms (Prolific)
- Pre-processing/analyzing survey survey data (R)
- Applying post-stratification and thinking along total survey error framework

Remember: This is a learning activity so try whatever you want.

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)
- Test questionnaire on your own devices (!)

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)
- Test questionnaire on your own devices (!)
- Publish on Prolific with my help

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)
- Test questionnaire on your own devices (!)
- Publish on Prolific with my help
- Lunch break

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)
- Test questionnaire on your own devices (!)
- Publish on Prolific with my help
- Lunch break
- Download data from Google Forms

Workflow

Recommended work flow (details in `0_exercise-instructions.pdf`):

- Familiarize yourself with the Eurobarometer survey
- Create survey on Google Forms
  - Don't forget to collect the sociodemographics for post-stratification
  - Measure multiple outcomes (estimates are also property of question not just sample)
- Test questionnaire on your own devices (!)
- Publish on Prolific with my help
- Lunch break
- Download data from Google Forms
- Analyze your data and/or apply post-stratification with larger pre-simulated data
  - To post-stratify, we will use Census data from UK Office of National Statistics
  - We compare these post-stratification estimates to the Eurobarometer benchmarks

A quick and dirty tour of post-stratification

Post-stratification

The principle is the following:

1. Chop up the sample into groups

Post-stratification

The principle is the following:

1. Chop up the sample into groups
2. Estimate the mean in each group

Post-stratification

The principle is the following:

1. Chop up the sample into groups
2. Estimate the mean in each group
3. Combine the estimates for each group into an overall estimate

Post-stratification

The principle is the following:

1. Chop up the sample into groups
2. Estimate the mean in each group
3. Combine the estimates for each group into an overall estimate

$$\hat{\bar{y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h$$

where

- N: size of the population
- N_h: size of group h
- $\hat{\bar{y}}_h$: estimated average outcome for group h

Poststratification

Assumptions:

- The realized sample s is partitioned into H groups, $s_1, s_2, \ldots s_H$
- Given s, all elements in s_k are assumed to have the same response probability; different groups can have different response probabilities
- Equivalent to data is missing completely at random (MCAR) within each group
- "Response Homogeneity Group Model" (RHG Model), see Sarndal et al. (1992) Sec 15.6.2 ("A Useful Response Model")

If RHG model holds (and some other minor technical conditions), then the post-stratification estimator is unbiased. See Sarndal et al. (1992) Result 15.6.1

Bias of cell-based post-stratification estimator

If RHG does not hold and if the original sample is simple random sampling without replacement, then (Bethlehem, Cobben, and Schouten 2011, sec. 8.2.1):

$$bias(\hat{\bar{y}}_{post}) = \frac{1}{N} \sum_{h=1}^{H} \frac{cor(\phi_i, y_i)^{(h)} S(\phi_i)^{(h)} S(y_i)^{(h)}}{\bar{\phi}^{(h)}}$$

So, how should we create the H groups?

Bias of cell-based post-stratification estimator

If RHG does not hold and if the original sample is simple random sampling without replacement, then (Bethlehem, Cobben, and Schouten 2011, sec. 8.2.1):

$$bias(\hat{\bar{y}}_{post}) = \frac{1}{N} \sum_{h=1}^{H} \frac{cor(\phi_i, y_i)^{(h)} S(\phi_i)^{(h)} S(y_i)^{(h)}}{\bar{\phi}^{(h)}}$$

So, how should we create the H groups?

- form homogeneous groups where there is little variation in response propensity ($S(\phi_i)^{(h)} \approx 0$) and the outcome ($S(y_i)^{(h)} \approx 0$)

Bias of cell-based post-stratification estimator

If RHG does not hold and if the original sample is simple random sampling without replacement, then (Bethlehem, Cobben, and Schouten 2011, sec. 8.2.1):

$$bias(\hat{\bar{y}}_{post}) = \frac{1}{N}\sum_{h=1}^{H}\frac{cor(\phi_i, y_i)^{(h)}S(\phi_i)^{(h)}S(y_i)^{(h)}}{\bar{\phi}^{(h)}}$$

So, how should we create the H groups?

- form homogeneous groups where there is little variation in response propensity ($S(\phi_i)^{(h)} \approx 0$) and the outcome ($S(y_i)^{(h)} \approx 0$)
- form groups where the people that you see are like the people that you don't see ($cor(\phi_i, y_i)^{(h)} \approx 0$)

In practice this can be difficult because you want to form many groups, but then you have noisy estimates for each group.

Post-stratification

Note:

- Horvitz-Thompson estimation is individual-based weight
- Post-stratification can better be understood as a group-based weight

Exercise plan

Three increasingly sophisticated ways to make group estimate $\hat{\bar{y}}_h$.

- cell-based post-stratification
- model-based post-stratification
- (Extra: multilevel regression post-stratification)

Simple cell-based post-stratification

For our example data, let's form 64 ($2 \times 2 \times 8$) groups:

- sex (2 groups)
- age (4 groups)
- region (8 groups)

$$\hat{\bar{y}}_h = \frac{\sum_{i \in h} y_i}{n_h}$$

h is a group described by a unique combination of gender (2 groups) $\times$ age (4 groups) $\times$ race (5 groups) $\times$ region (4 groups)

Simple cell-based post-stratification

- We can't make an estimate for each group. For example, we don't have any female, 65+ in the East of England

Simple cell-based post-stratification

- We can't make an estimate for each group. For example, we don't have any female, 65+ in the East of England
- This problem can arise if you have too many cell. We have a crude work-around (imputation) in the code.

Model-based post-stratification

$$\hat{\bar{y}}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h$$

where $\hat{\bar{y}}_h$ comes from an individual-level model

$$Pr(y_i = 1) = logit^{-1}(\beta_0 + \beta_{male} \cdot male_i + \beta_{18to29} \cdot 18to29_i + \beta_{30to49} \cdot 30to49_i + \beta_{50to64} \cdot 50to64_i + \beta_{65plus} \cdot 65plus_i + \beta_{NorthernIreland} \cdot NorthernIreland_i + \beta_{Wales} \cdot Wales_i ... + \beta_{London} \cdot London_i)$$

Model-based post-stratification

- Modeling allows you to make more estimates for smaller groups

Model-based post-stratification

- Modeling allows you to make more estimates for smaller groups
- These techniques is widely used by modern pollsters (e.g., YouGov) and political scientists

Model-based post-stratification

- Modeling allows you to make more estimates for smaller groups
- These techniques is widely used by modern pollsters (e.g., YouGov) and political scientists

You're unlikely to get through all of the activity—that's ok!