Introduction to group activity

Matthew J. Salganik
Department of Sociology
Princeton University

Summer Institutes in Computational Social Science
June 20, 2019

The Summer Institutes in Computational Social Science is supported by grants from the Russell Sage Foundation and the Alfred P. Sloan Foundation.

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

ABSTRACT

It is increasingly common for government and industry organizations to conduct online, opt-in surveys, in part because they are typically fast, inexpensive, and convenient. Online polls, however, attract a non-representative set of respondents, and so it is unclear whether results from such surveys generalize to the broader population. These non-representative surveys stand in contrast to probability-based sampling methods, such as random-digit dialing (RDD) of phones, which are a staple of traditional survey research. Here we investigate the accuracy of non-representative data by administering an online, fully opt-in poll of social and political attitudes. Our survey consisted of 49 multiple-choice attitudinal questions drawn from the probability-based, in-person 2012 General Social Survey (GSS) and select RDD phone surveys by the Pew Research Center. To correct for the inherent biases of non-representative data, we statistically adjust estimates via model-based poststratification, a classic statistical tool but one that is only infrequently used for bias correction. Our online survey took less than one-twentieth the time and money of traditional RDD polling, and less than one-hundredth the time and money of GSS polling. After statistical correction, we find the median absolute difference between the non-probability-based online survey and the probability-based GSS and Pew studies is 7 percentage points. This difference is considerably larger than if the surveys were all perfect simple random samples drawn from the same population; the gap, however, is comparable to that between the GSS and Pew estimates themselves. Our results suggest that with proper statistical adjustment, online, non-representative surveys are a valuable tool for practitioners in varied domains.

Activity:

- Design a questionnaire using questions already asked on high quality surveys

Activity:

- Design a questionnaire using questions already asked on high quality surveys
- Recruit participants from Amazon Mechanical Turk and have them complete your questionnaire

Activity:

- Design a questionnaire using questions already asked on high quality surveys
- Recruit participants from Amazon Mechanical Turk and have them complete your questionnaire
- Compare results from your survey to the results from the high-quality survey

Activity:

- Design a questionnaire using questions already asked on high quality surveys
- Recruit participants from Amazon Mechanical Turk and have them complete your questionnaire
- Compare results from your survey to the results from the high-quality survey
- Try different approaches to weighting and see how the change the estimates

Activity:

- Design a questionnaire using questions already asked on high quality surveys
- Recruit participants from Amazon Mechanical Turk and have them complete your questionnaire
- Compare results from your survey to the results from the high-quality survey
- Try different approaches to weighting and see how the change the estimates
- De-identify and "open-source" data by sending to your local organizer (but remember to think about the end at the beginning)

This activity will give you practice:

- Designing questionnaires

This activity will give you practice:

- Designing questionnaires
- Collecting survey data

This activity will give you practice:

- Designing questionnaires
- Collecting survey data
- Analyzing survey data (data wrangling and post-stratification)

This activity will give you practice:

- Designing questionnaires
- Collecting survey data
- Analyzing survey data (data wrangling and post-stratification)
- Working with Amazon Mechanical Turk

This activity will give you practice:

- Designing questionnaires
- Collecting survey data
- Analyzing survey data (data wrangling and post-stratification)
- Working with Amazon Mechanical Turk
- Archiving data for other researchers

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

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms (we have a template)

If you were making up the budget for the federal government this year, would you increase spending, decrease spending, or keep spending the same for the following issues?

	Increase spending	Decrease spending	Keep spending the same
Military Defense	○	○	○
Economic assistance to needy people in the U.S.	○	○	○
Education	○	○	○
Environmental Protection	○	○	○
Economic assistance to needy people around the world	○	○	○
Health care	○	○	○
Rebuilding highways, bridges and roads	○	○	○
Scientific Research	○	○	○
Social Security	○	○	○

or

If you were making up the budget for the federal government this year, would you increase spending, decrease spending, or keep spending the same for: Military defense?

- ○ Increase spending
- ○ Decrease spending
- ○ Keep spending the same

If you were making up the budget for the federal government this year, would you increase spending, decrease spending, or keep spending the same for: Economic assistance to needy people in the U.S.

- ○ Increase spending
- ○ Decrease spending
- ○ Keep spending the same

If you were making up the budget for the federal government this year, would you increase spending, decrease spending, or keep spending the same for: Education?

- ○ Increase spending
- ○ Decrease spending
- ○ Keep spending the same

If you were making up the budget for the federal government this year, would you increase spending, decrease spending, or keep spending the same for: Environmental protection?

- ○ Increase spending
- ○ Decrease spending
- ○ Keep spending the same

Don't forget to collect the information that you will need for post-stratification:

Don't forget to collect the information that you will need for post-stratification:

- gender
- age
- state of residence
- race/ethnicity

How should you ask these questions?

Don't forget to collect the information that you will need for post-stratification:

- gender
- age
- state of residence
- race/ethnicity

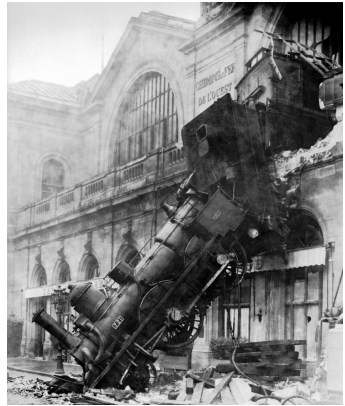
How should you ask these questions? You should copy from the American Community Survey.

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms (we have a template)

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms (we have a template)
- Deploy to MTurk



Last year, every group made at least one error deploying their relatively simple survey.

https://en.wikipedia.org/wiki/Failure#/media/File:Train_wreck_at_Montparnasse_1895.jpg

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms (we have a template)
- Deploy to MTurk

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms (we have a template)
- Deploy to MTurk
- Take a break

Allison Morgan
@alliecmorgan

Following

Just wrapped up the first week of #SICSS2017! On Thursday, we got 50+ online survey responses, all while frolicking in a fountain.

3:24 PM - 24 Jun 2017

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms
- Deploy to MTurk
- Take a break

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms
- Deploy to MTurk
- Take a break
- Validate and pay workers

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you

A quick and dirty tour of the post-stratification methods we will use

Cell-based Poststratification

Let's split the sample into H mutually exclusive and exhaustive groups.

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

where

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

Roughly, we are producing estimates for each group and then putting them together in the right way.

Cell-based 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 poststratification estimator is unbiased. See Sarndal et al. (1992) Result 15.6.1

Bias of cell-based poststratification estimator from non-response

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 poststratification estimator from non-response

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 poststratification estimator from non-response

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.

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

where

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

Use this estimator in three steps:

1. Chop up the sample into groups

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

where

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

Use this estimator in three steps:

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

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

where

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

Use this estimator in three steps:

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

Note:

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

Three increasingly sophisticated ways to make group estimate $\hat{\bar{y}}_h$ (you won't have time to do all of these)

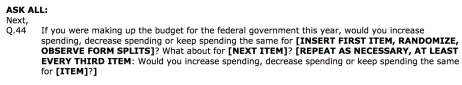
- cell-based poststratification
- model-based poststratification
- multilevel regression postratification (Mr. P)

Data

- Our survey data comes from a survey of 495 people on MTurk collected in less than a week ago
- We will compare to high-quality telephone surveys from the Pew Research Center (links to questionnaires)
- To poststratify our survey data, we will use data from the American Community Survey about the population of the US
- We use multiple questions because estimates are also a property of a question not just a sample.

Data

Some questions have multiple choice answers, but we convert to a series of binary variables for analysis. For example,
As asked:



ASK ALL:
Next,
Q.44 If you were making up the budget for the federal government this year, would you increase spending, decrease spending or keep spending the same for **[INSERT FIRST ITEM, RANDOMIZE, OBSERVE FORM SPLITS]**? What about for **[NEXT ITEM]**? **[REPEAT AS NECESSARY, AT LEAST EVERY THIRD ITEM**: Would you increase spending, decrease spending or keep spending the same for **[ITEM]**?]

As analyzed:

- If you were making up the budget for the federal government this year would you increase funding for scientific research?
- If you were making up the budget for the federal government this year would you decrease funding for scientific research?
- If you were making up the budget for the federal government this year would you keep funding the same funding for scientific research?

This means that each of our estimates are not independent. For real research, you'd want a different way to handle this.

Simple cell-based poststratification

Let's do lots of groups.

- gender (2 groups)
- age (4 groups)
- race (5 groups)
- region (4 groups)
- Makes 160 ($2 \times 4 \times 5 \times 4$) groups

Simple cell-based poststratification

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

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)

Cell-based poststratification

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

Cell-based poststratification

- We can't make an estimate for each group. For example, we don't have any female, 65+, Hispanic living in the South.
- This problem can arise if you have too many cell. We have a crude work-around in the code we provide.

Model-based poststratification

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

where $\hat{\bar{y}}_h$ comes from

$$\begin{aligned} Pr(y_i = 1) = logit^{-1}(&\beta_0 + \\ &\beta_{male} \cdot male_i + \\ &\beta_{30-49} \cdot 30-49_i + \beta_{50-64} \cdot 50-64_i + \beta_{65+} \cdot 65_i + \\ &\beta_{afr-am} \cdot afam_i + \beta_{as-am} \cdot asam_i + \beta_{hispanic} \cdot hisp_i + \beta_{other} \cdot other_i + \\ &\beta_{midwest} \cdot midwest_i + \beta_{south} \cdot south_i + \beta_{west} \cdot west_i) \end{aligned}$$

Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls

David K. Park
Department of Political Science and Applied Statistics,
Washington University, St. Louis, MO 63130
e-mail: dpark@artsci.wustl.edu

Andrew Gelman
Departments of Statistics and Political Science, Columbia University,
New York, NY 10027
e-mail: gelman@stat.columbia.edu

Joseph Bafumi
Department of Political Science, Columbia University, New York, NY 10027

Mr. P

We fit a multilevel logistic regression model for the mean of a binary response variable conditional on poststratification cells. This approach combines the modeling approach often used in small-area estimation with the population information used in poststratification (see Gelman and Little 1997, *Survey Methodology* 23:127–135). To validate the method, we apply it to U.S. preelection polls for 1988 and 1992, poststratified by state, region, and the usual demographic variables. We evaluate the model by comparing it to state-level election outcomes. The multilevel model outperforms more commonly used models in political science. We envision the most important usage of this method to be not forecasting elections but estimating public opinion on a variety of issues at the state level.

https://www.jstor.org/stable/25791784

See also Gelman and Hill (2007), Chapter 14 ("Multilevel logistic regression")

Mr. P.

$\hat{\bar{y}}_h$ comes from

$$
\begin{aligned}
Pr(y_i = 1) = logit^{-1}(&\beta_0 + \\
&\beta_{male} \cdot male_i + \\
&\alpha^{age}_{k[i]} + \\
&\alpha^{race}_{k[i]} + \\
&\alpha^{region}_{k[i]})
\end{aligned}
$$

$$
\begin{aligned}
\alpha^{age}_k &\sim N(0, \sigma^2_{age}) \text{ for } k = 1, \ldots, 4 \\
\alpha^{race}_k &\sim N(0, \sigma^2_{race}) \text{ for } k = 1, \ldots, 5 \\
\alpha^{region}_k &\sim N(0, \sigma^2_{region}) \text{ for } k = 1, \ldots, 4
\end{aligned}
$$

Priors determined by RStanarm (https://cran.r-project.org/web/packages/rstanarm/vignettes/priors.html)

To learn more about Mr. P.

Generally optimistic:

- Park, Gelman, and Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis*.
- Lax and Phillips. 2009. "How should we estimate public opinion in the states?" *American Journal of Political Science*.
- Ghitza and Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science*.
- Warshaw and Rodden. 2012. "How should we measure district-level public opinion on individual issues?" *Journal of Politics*.
- Downs et al. 2018. "Multilevel Regression and Poststratification: A Modelling Approach to Estimating Population Quantities From Highly Selected Survey Samples." *American Journal of Epidemiology*.

Generally cautious:

- Buttice and Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis*.

Notes

- Modeling allows you to make more estimates for smaller groups

Notes

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

Notes

- 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 will not finish this activity. That's OK.

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you
- De-identify and open-source the data that you collected

https://www.youtube.com/watch?v=66oNv_DJuPc

Brief introduction into open-sourcing your data:

- Store your data in a simple format

Brief introduction into open-sourcing your data:

- Store your data in a simple format
- Provide documentation

Brief introduction into open-sourcing your data:

- Store your data in a simple format
- Provide documentation
- Beware of privacy

Store your data in a simple format

Store your data in a simple format

In this case .csv should be good.

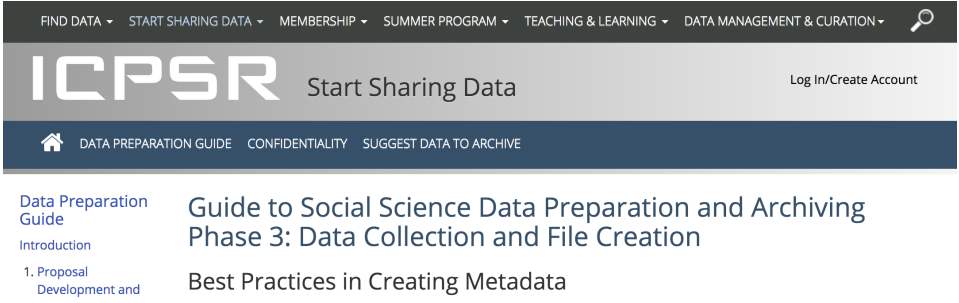
Provide documentation

Provide documentation

What would another researcher want to know?

- How and when was this data collected?
- What do the different variables describe?

Provide documentation (more details)



https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html

Beware of privacy

Beware of privacy

Remove personally identifying information

NIST definition [edit]

The following data, often used for the express purpose of distinguishing individual identity, clearly classify as PII under the definition used by the National Institute of Standards and Technology (described in detail below):[15]

- Full name (if not common)
- Face (sometimes)
- Home address
- Email address (if private from an association/club membership, etc.)
- National identification number (e.g., Social Security number in the U.S.)
- Passport number
- Vehicle registration plate number
- Driver's license number
- Face, fingerprints, or handwriting
- Credit card numbers
- Digital identity
- Date of birth
- Birthplace
- Genetic information
- Telephone number
- Login name, screen name, nickname, or handle

https://en.wikipedia.org/wiki/Personally_identifiable_information

DOI:10.1145/1743546.1743558
Arvind Narayanan and Vitaly Shmatikov

Privacy and Security
Myths and Fallacies of "Personally Identifiable Information"

http://dx.doi.org/10.1145/1743546.1743558

In this case, we recommend:

- Removing PII (name, email address, etc)
- Removing TurkID
- Coarsen age, geography, and race/ethnicity
- Coarsen timestamp
- Anything else?

For more on coarsening, see this code:
https://github.com/compsocialscience/summer-institute/blob/master/2019/materials/day4-surveys/activity/mturk_data_cleaning.Rmd

For more about de-identification, see *Bit by Bit*, Sec 6.6.2 "Understanding and managing informational risk"

The 5Ws of data release:

The 5Ws of data release:

- Who: You

The 5Ws of data release:

- Who: You
- What: make your data available to other researchers in a responsible way

The 5Ws of data release:

- Who: You
- What: make your data available to other researchers in a responsible way
- Where: Dataverse, ICPSR, or an archival data repository

The 5Ws of data release:

- Who: You
- What: make your data available to other researchers in a responsible way
- Where: Dataverse, ICPSR, or an archival data repository (your laptop is not an archival data repository)

The 5Ws of data release:

- Who: You
- What: make your data available to other researchers in a responsible way
- Where: Dataverse, ICPSR, or an archival data repository (your laptop is not an archival data repository)
- When: when you publish your paper

The 5Ws of data release:

- Who: You
- What: make your data available to other researchers in a responsible way
- Where: Dataverse, ICPSR, or an archival data repository (your laptop is not an archival data repository)
- When: when you publish your paper
- Why: It is good for you and it is good for the world

Share your data to your local organizer who can post it here:
https://github.com/compsocialscience/summer-institute/tree/master/2019/materials/day4-surveys/datasets

When you start your projects next week

- plan to release your data
- plan to release your code

Questions

Our recommended work flow:

- Create a write-up that describes what data you will be collecting, why, and how it will be shared with others (for tips, see Meyer)
- Create survey on Google Forms
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you
- De-identify and open-source the data that you collected