# Introduction to group activity on surveys

Matthew J. Salganik
Department of Sociology
Princeton University

# 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.

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

Activity:

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

Activity:

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

Activity:

- Design a questionnaire using questions already asked on a high-quality survey
- Recruit participants from Amazon Mechanical Turk (MTurk) 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 a high-quality survey
- Recruit participants from Amazon Mechanical Turk (MTurk) 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

This activity will give you practice:

- Designing questionnaires

This activity will give you practice:

- Designing questionnaires
- Deploy jobs to MTurk

This activity will give you practice:

- Designing questionnaires
- Deploy jobs to MTurk
- Analyzing survey data (data wrangling and post-stratification)

This activity will give you practice:

- Designing questionnaires
- Deploy jobs to MTurk
- Analyzing survey data (data wrangling and post-stratification)
- Using the total survey error framework to consider and discuss errors in estimates

This activity will give you practice:

- Designing questionnaires
- Deploy jobs to MTurk
- Analyzing survey data (data wrangling and post-stratification)
- Using the total survey error framework to consider and discuss errors in estimates

Remember: This is a learning activity so try whatever you want and don't expect perfection in just a few hours.

- Start with our template and create your questionnaire

The template includes:

- information about who is collecting the data, why, and how it will be used

The template includes:

- information about who is collecting the data, why, and how it will be used
- questions you will use for post-stratification (e.g., sex, age, state of residence, race/ethnicity)

The template includes:

- information about who is collecting the data, why, and how it will be used
- questions you will use for post-stratification (e.g., sex, age, state of residence, race/ethnicity)
- attention check questions

The template includes:

- information about who is collecting the data, why, and how it will be used
- questions you will use for post-stratification (e.g., sex, age, state of residence, race/ethnicity)
- attention check questions
- question to collect the Turk ID so that you can pay your respondents

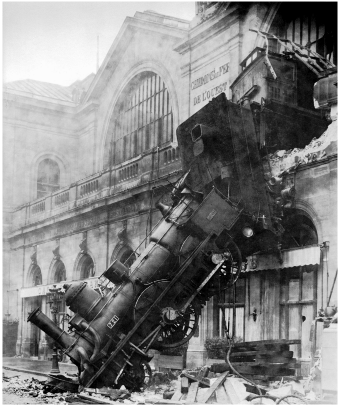
The template is designed to save you time. You will need to add the questions from the Pew surveys

Our recommended work flow:

- Start with our template and create your questionnaire

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire



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:

- Start with our template and create your questionnaire
- Test your questionnaire
- Deploy to MTurk

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire
- 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:

- Start with our template and create your questionnaire
- Test your questionnaire
- Deploy to MTurk
- Take a break
- Validate and pay workers

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire
- 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

Use this estimator in three steps:

1. Chop up the sample into groups

Use this estimator in three steps:

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

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

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

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

where

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

Techniques vary in how they estimate the average outcome for group $h$: $\hat{\bar{y}}_h$

# Cell-based Poststratification

"Response Homogeneity Group Model" (RHG Model), see Sarndal et al. (1992) Sec 15.6.2 ("A Useful Response Model") 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

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.

Note:

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

# Data

- The data that we are providing you for analysis comes from a survey we deployed on MTurk recently

# Data

- The data that we are providing you for analysis comes from a survey we deployed on MTurk recently
- We will compare to high-quality surveys from the Pew Research Center

# Data

- The data that we are providing you for analysis comes from a survey we deployed on MTurk recently
- We will compare to high-quality surveys from the Pew Research Center
- To poststratify our survey data, we will use data from the American Community Survey about the population of the United States

# Data

- The data that we are providing you for analysis comes from a survey we deployed on MTurk recently
- We will compare to high-quality surveys from the Pew Research Center
- To poststratify our survey data, we will use data from the American Community Survey about the population of the United States
- We use multiple questions because estimates are also a property of a question not just a sample.

# 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)

# Simple 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.

## Simple 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.

For the activity, it would be great if you could complete doing simple cell-based poststratifiaction.
At the end of these slides there is more information on these other methods:

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

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you

When you start your projects next week

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

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you, starting with our code scaffolding

Our recommended work flow:

- Start with our template and create your questionnaire
- Test your questionnaire
- Deploy to MTurk
- Take a break
- Validate and pay workers
- Analyze the much larger sample that we have collected for you, starting with our code scaffolding

This activity is our attempt to cover as much of surveys as we can given the constraints of the virtual learning

Let's get started

- Questions?

Let's get started

- Questions?
- Announcements from Lai?

Let's get started

- Questions?
- Announcements from Lai?
- Assign 6 groups with stratified systematic sampling
  - Who has recent experience deploying something to MTurk?

Let's get started

- Questions?
- Announcements from Lai?
- Assign 6 groups with stratified systematic sampling
  - Who has recent experience deploying something to MTurk?
  - Who has recent experience writing a questionnaire?

# Appendix

Appendix with more information about

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

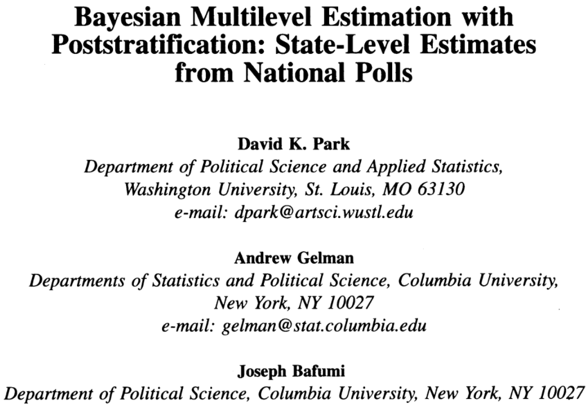
# Model-based poststratification

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

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

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

# Mr. P


**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*

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

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

# 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 an individual-level model

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

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

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*.