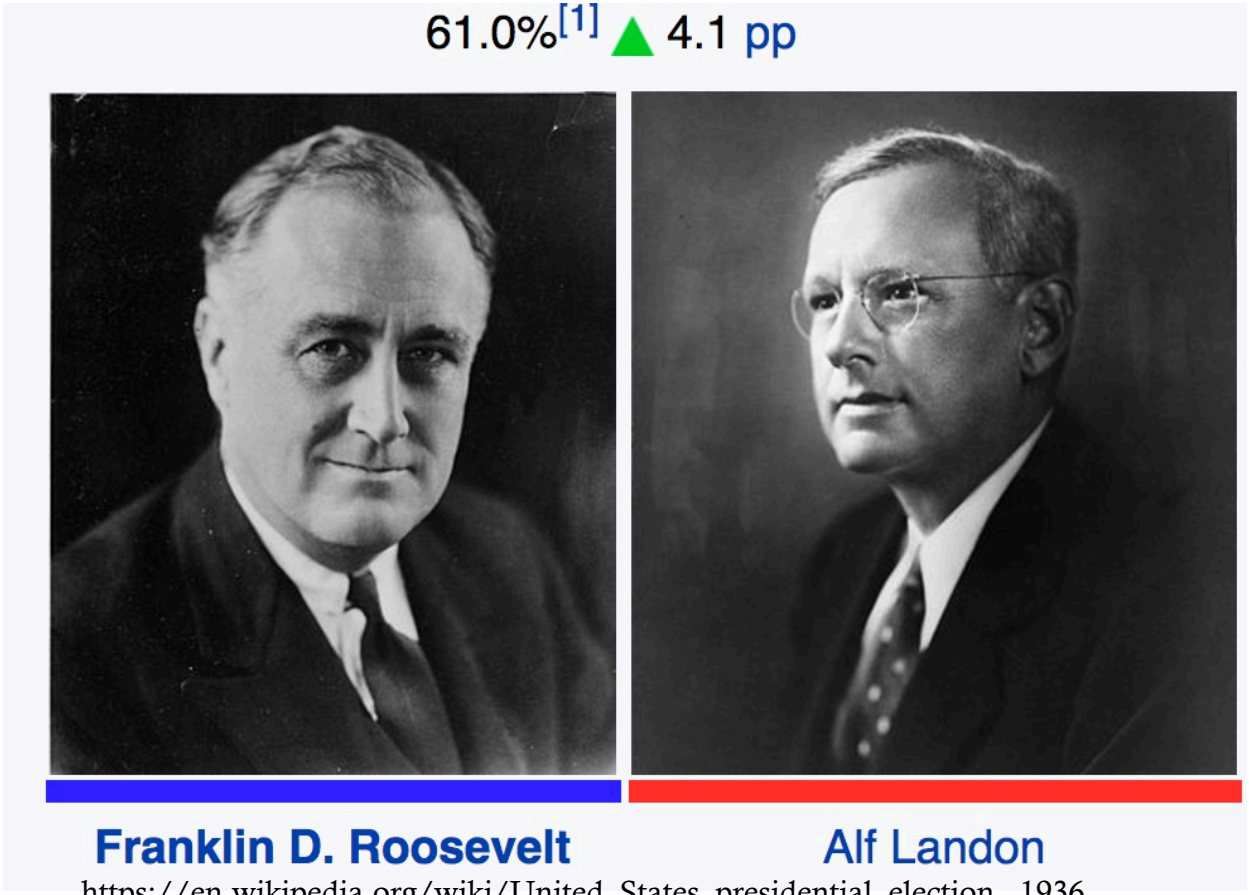
Identify and Adjust for Non-Response Bias

Benjamin Chu

Biomath 204

3/13/2017

Data analysis: Trash in Trash out



https://en.wikipedia.org/wiki/United_States_presidential_election,_1936

Random Sampling and (unit) Non-Response Bias

Random Sampling:

A subset of individuals (sample) chosen from a larget set (population) to be surveyed

(Unit) Non-response bias:

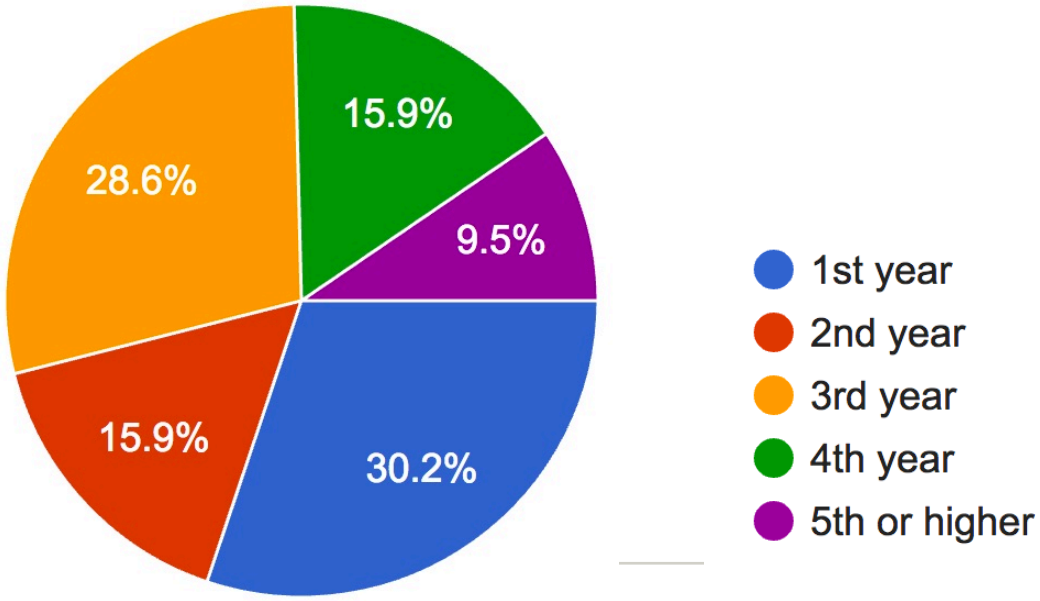
Error due to a proportion of sample population not responding to poll

When present, no amount of data can negate its effect

What is NOT non-response

- "When particular types of respondents are not reached" - trchome
- "Participation is voluntary…" (Wang, 2014)

……correct but ambiguous.



1991 Race and Politics Survey

- Telephone survey with 178 questions
- Up to 4 callbacks (identification)
- **Second survey by mail** (adjustment)

http://astro.berkeley.edu/prospective-students/about-berkeley

Callback: Identify non-response bias

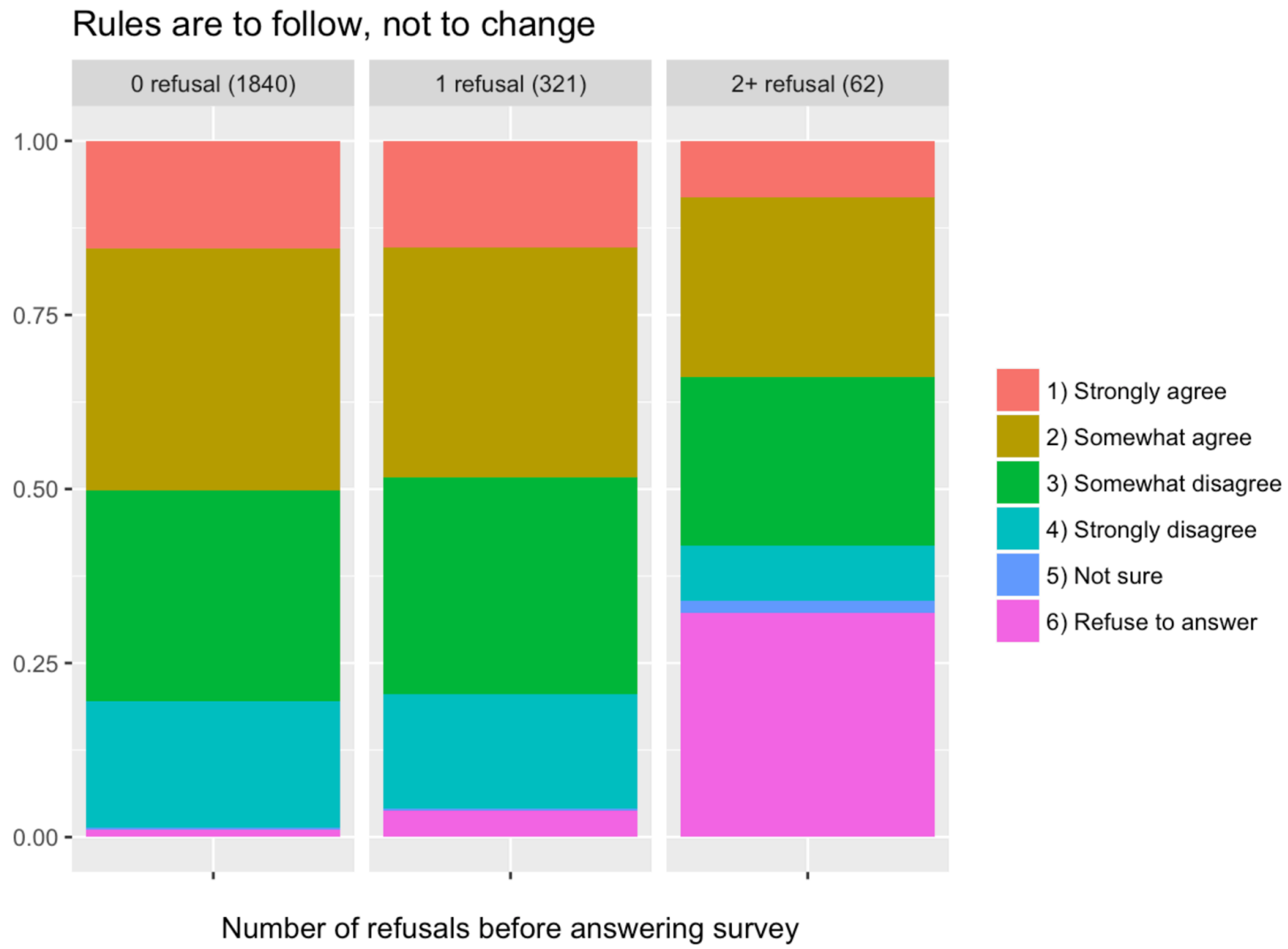
Samples may refuse to take survey

→ **Ask** subset of non-respondents **again**.

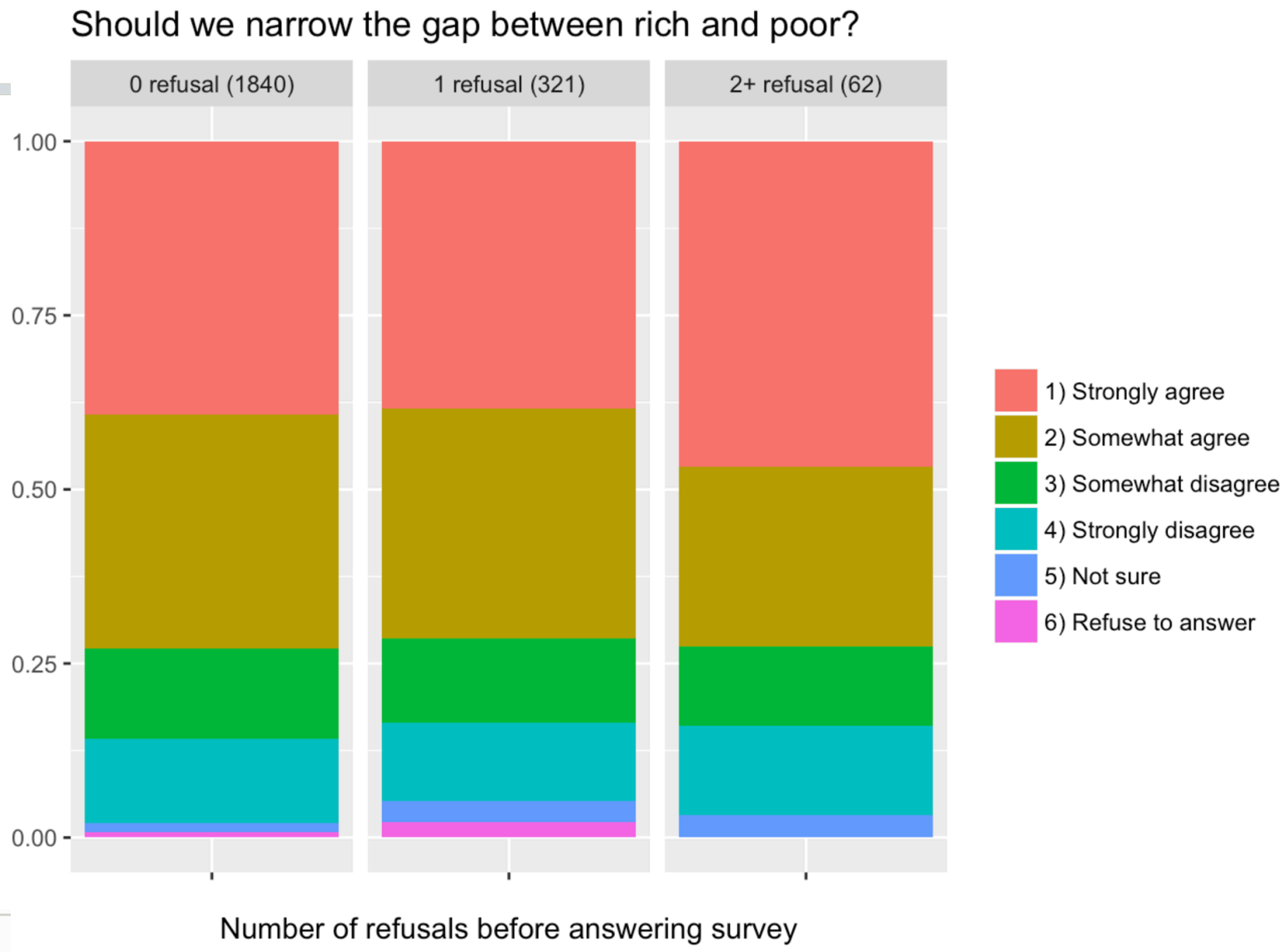
- Assumes "late respondents" ~ non-respondents

Q: How to measure difference between early and late respondents?

Callback – compare initial and late respondents



Another callback example



Estimating (true) general opinion

- Survey goal: to obtain **general opinion** on smth
- Non-response will skew survey result

https://www.buzzfeed.com/matwhitehead/helen-is-all-of-us

Q: how busy?

Weighting

$\beta_i = 1$ if response is strongly agree or somewhat agree.
S_i represents the number of respondents in each cell.
$y_i = \beta_i S_i$ be the variable of interest.
π_i be the probability to be drawn (design weight).
p_i be the response probability .
$w_i = (\pi_i p_i)^{-1}$ the non-response-adjusted weight for observation i.

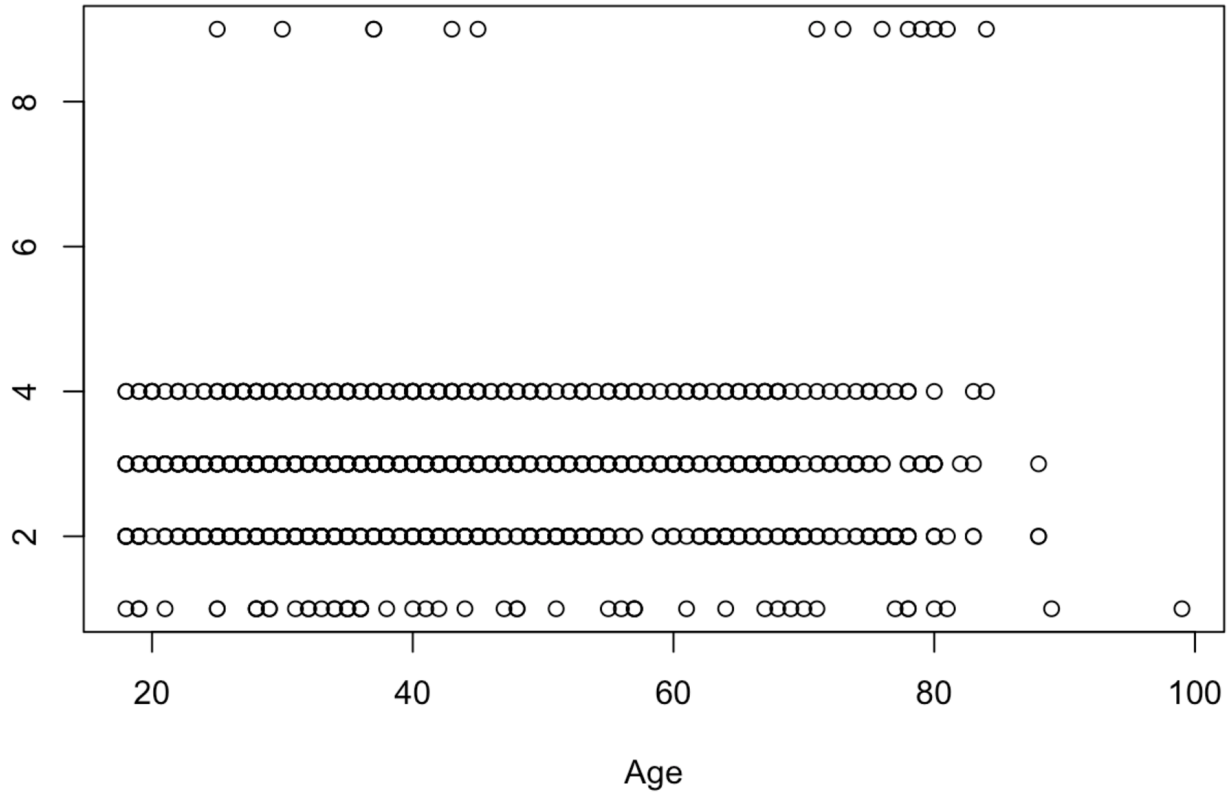
$$\hat{\overline{Y}} = \frac{\sum w_i y_i}{S_i}$$

Idea:

- Divide respondents into cells (require information from all samples)
- Weight individuals in each cell by 1/(response rate)

Descriptive stat:

Most people who don't succeed are lazy



1~4 = Strongly agree ~ Strongly disagree9 = refuse to answer

Pre and Post Weighting

Pre-weighting:

$$\overline{Y} = \sum y_i / \sum S_i = \frac{51 + 341}{51 + 341 + 563 + 229 + 14} = 32.7\%$$

	18~28	29~38	39~48	49~58	59~68	69~78	79~88	89+
Total Sample	444	606	466	259	221	159	52	16
Respondents	213	320	256	148	136	95	28	2
Respondents who agreed	66	103	76	54	36	45	10	2

$$\hat{\overline{Y}} = \frac{\sum w_i y_i}{S_i} = \frac{66\frac{444}{213} + 103\frac{606}{320} + 76\frac{466}{256} + 54\frac{259}{148} + 36\frac{221}{136} + 45\frac{159}{95} + 10\frac{52}{28} + 2\frac{16}{2}}{444 + 606 + 466 + 259 + 221 + 159 + 52 + 16} = 33.0\%$$

Key Takeaway:

- Never trust volunteer sampling (ever)
- All good survey uses **callback**: easiest but tedious
  - used for identification
- Weighting is intuitive, but requires "information" before sampling

References

- http://sda.berkeley.edu/D3/Natlrace/Doc/nrac.htm
- http://sda.berkeley.edu/archive.htm
- 2004 Jan Wang "Non-response in the Norwegian Business Tendency Survey"
- 1984 Roderick Little "Survey Nonresponse bias"
- 2001, Carlson and Williams, "A COMPARISON OF TWO METHODS TO ADJUST WEIGHTS FOR NON-RESPONSE: PROPENSITY MODELING AND WEIGHTING CLASS ADJUSTMENTS"
- Lohr. "Sampling: Design and analysis" 2nd edition
- http://www.trchome.com/research-knowledge/white-paper-library/227-situational-use-of-data-weighting-complete
- http://www.trchome.com/65-market-research-knowledge/white-paper-library/215-non-response-bias-in-survey-sampling-complete
- Jennings lecture notes (7,8,10) for Stat 522 at Purdue