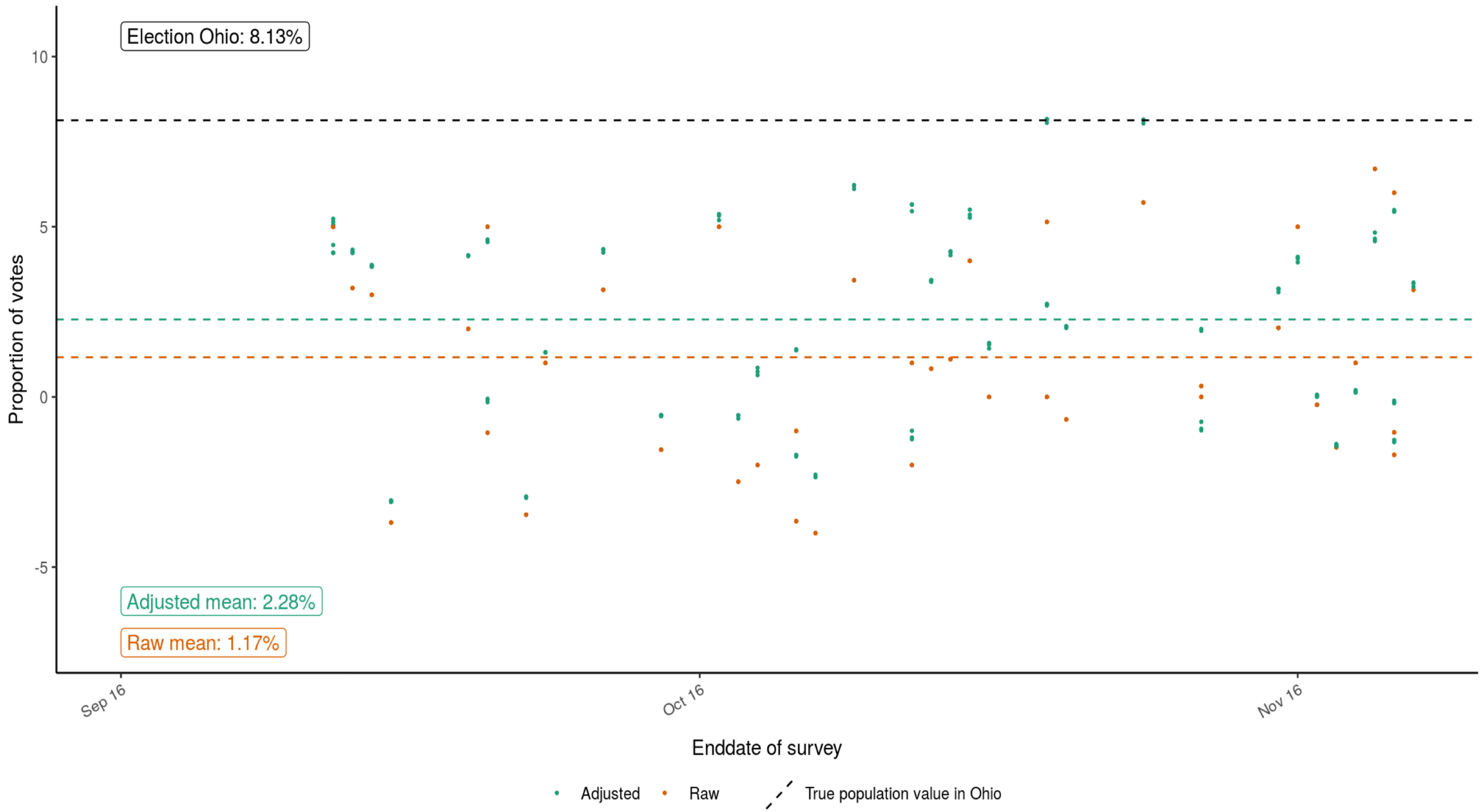
Survey data analysis
Week 49:
“Inference for non-probability samples”

Today

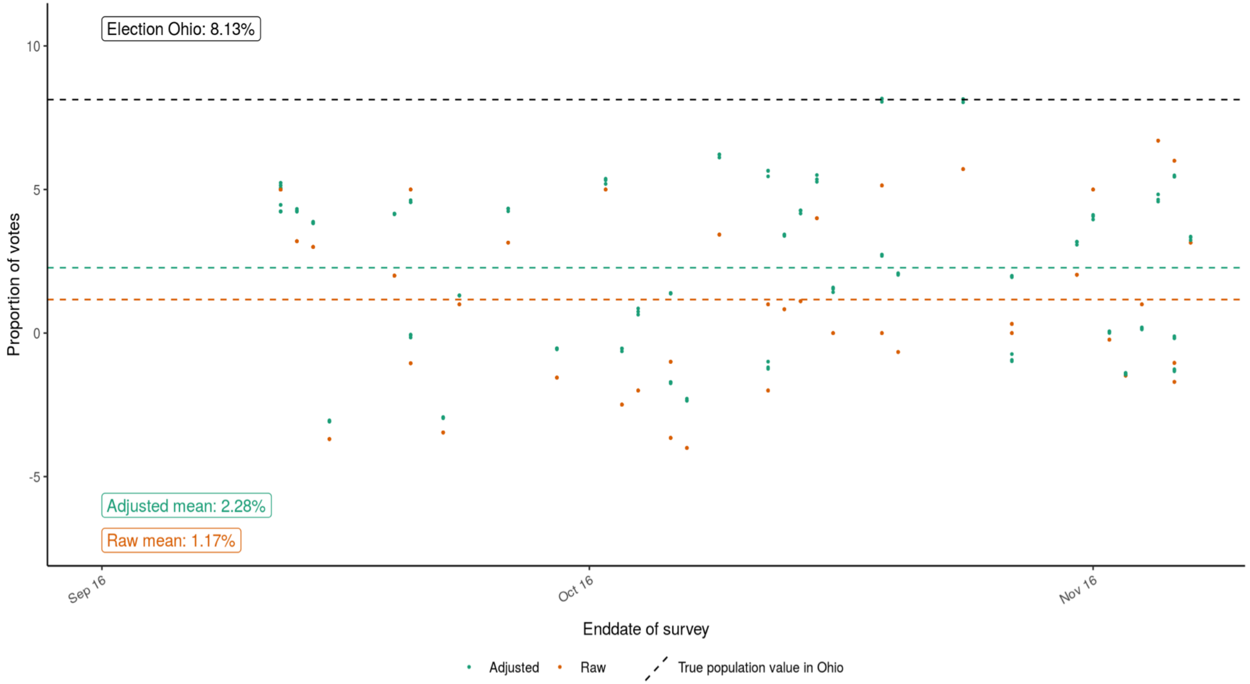
- Lecture
- Inference competition

Back to week 37 (1)



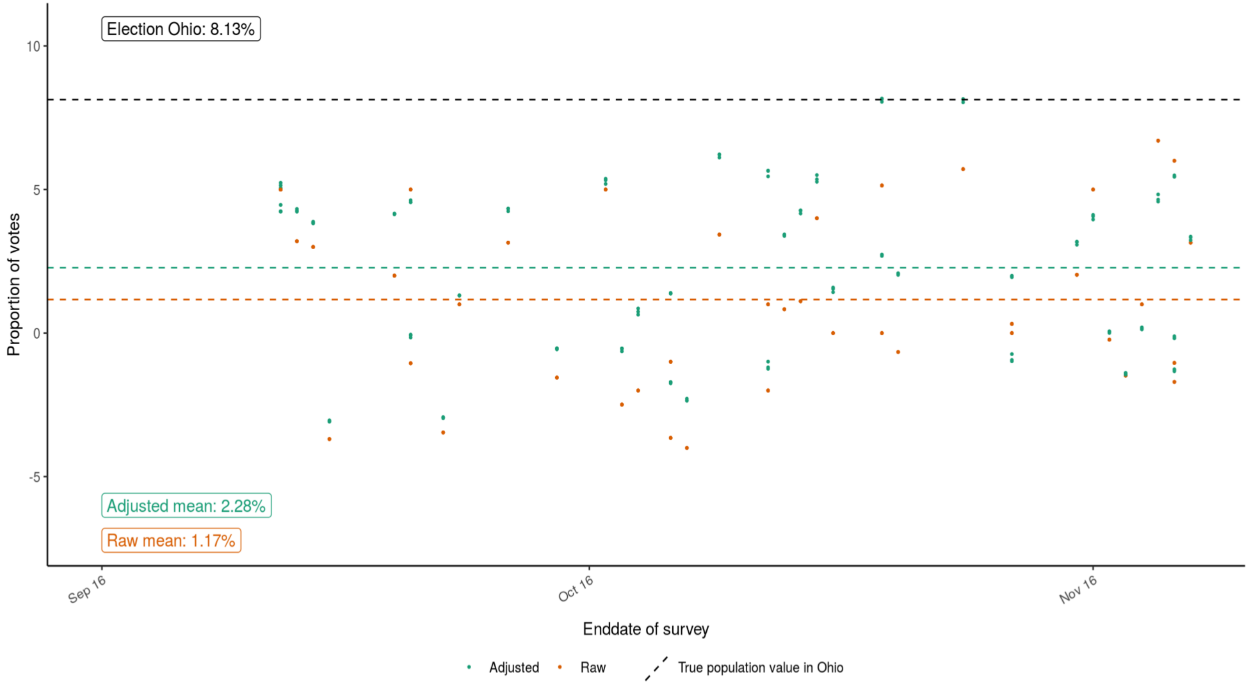
See: https://utrecht-university.shinyapps.io/SDA_shinyelectionbias/

Back to week 37 (1)



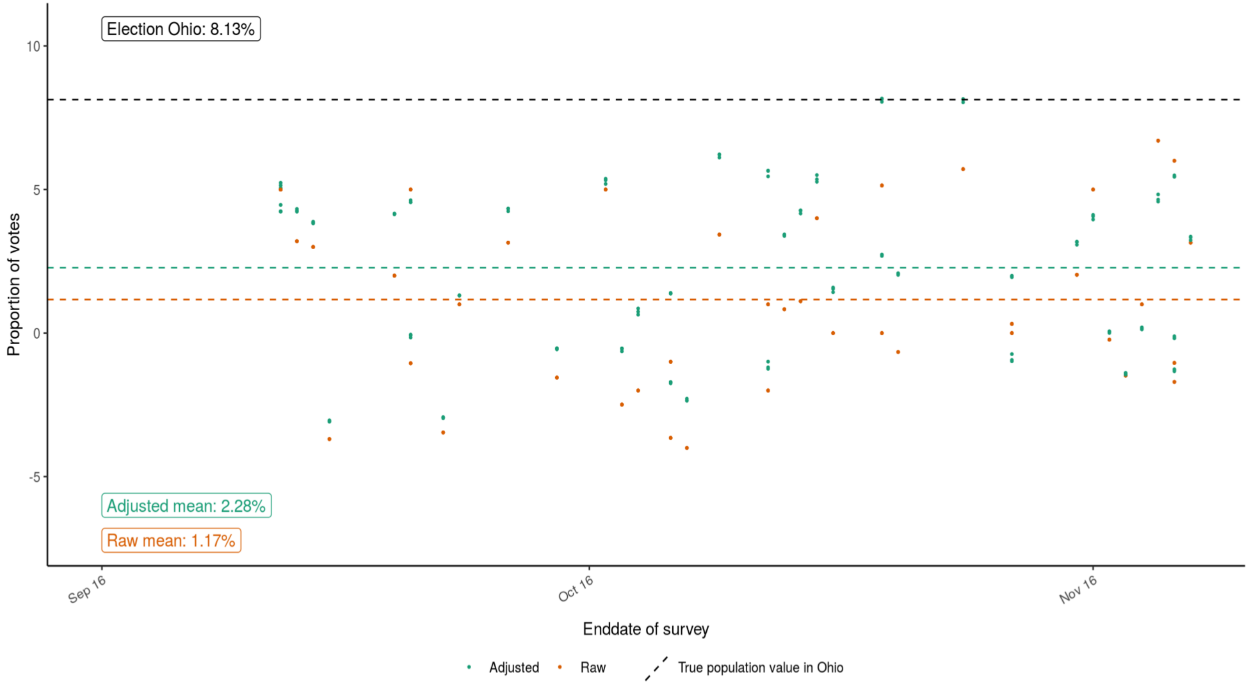
- Adjustments only help a bit on average
- For individual polls they sometimes make matters worse!

Back to week 37 (1)



- Adjustments only help a bit on average
- For individual polls they sometimes make matters worse!
- Grade of pollster/ sample size/ population dont make the difference

We have an inference problem



- Adjustments only help a bit on average
- For individual polls they sometimes make matters worse!
- Grade of pollster/ sample size/ population dont make the difference
- Problems with weighting
- A lot of polls are **not** probability based

Cornesse et al (2020)

- When is a non-probability sample not too bad?
  -
  -
- Global adjustment approaches
  - i.e. Conceptualize as design-based
- Estimate-specific approaches

Cornesse et al (2020)

- When is a non-probability sample not too bad?
  - in change estimates? (perhaps regression?
  - When controls are accurate (quota)
- Global adjustment approaches
  - i.e. Conceptualize as design-based
- Estimate-specific approaches

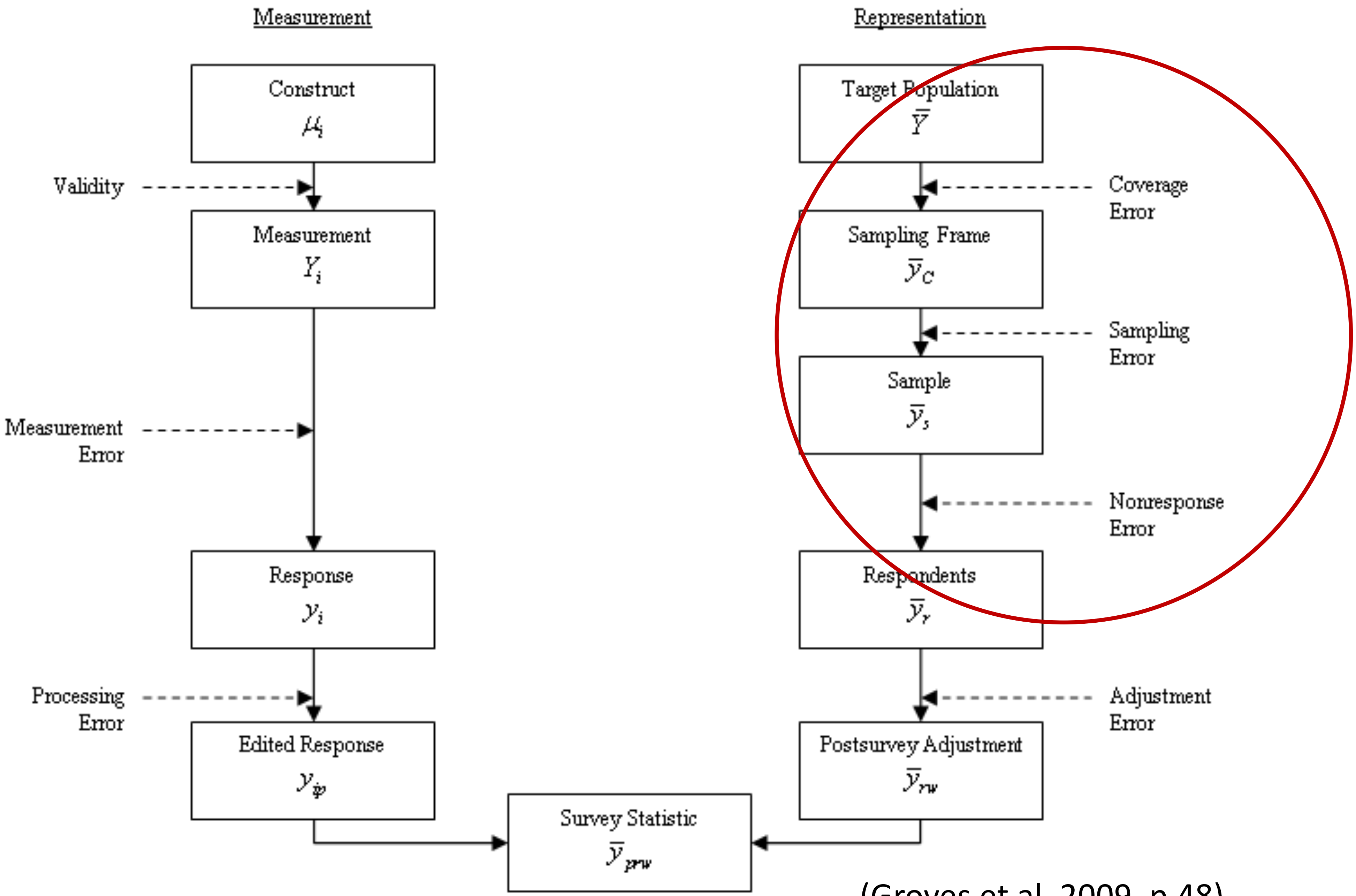
Inference: perspectives from other fields

- Natural sciences
  - Laws of nature: gravity works everywhere
- Social sciences broadly
  - We need descriptives about our population
  - Causal inference with experiments not possible?
- Behavioral sciences (psychology)

Inference: perspectives from other fields

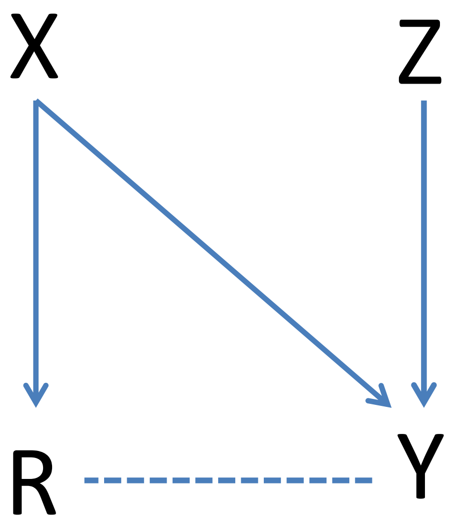
- Natural sciences
  – Laws of nature: gravity works everywhere
- Social sciences broadly
  – We need descriptives about our population
  – Causal inference with experiments not possible?
    - Regression, mediation etc.
  – Focus on external validity
- Behavioral sciences (psychology?)
  - Central role of experiments for causal analysis
  - Focus on internal validity

Selection bias vs. TSE



(Groves et al. 2009, p.48)

Selection bias as missing data problem



Meng 2018 – linking data quality,quantity

- ρ(R,G): correlation between selection bias (R) and variable of interest
- **σ**(G): variation in population of variable of interest
  - E.g. If everyone votes for Clinton, no problem
- Data quantity: $\sqrt{\frac{1-f}{f}}$ (week 39)
  - f=sampling fraction from population.

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}} .$$

  - P. 690 (eq 2.3)

Meng 2018 – linking data quality,quantity

- R mechanism (response)
  - Design based
    - Sampling probabilities are known
    - Nonresponse propensities are modeled.
  - Non-probability: selection probabilities are unknown
- G: estimate of interest (e.g. a mean)
  - Y in missing data literature

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}} .$$

- If correlation [R,G] = 0, no problem with any data
- If R does not vary over elements, no problem

Meng 2018 – final

When can we draw inferences for Big Data (non-probability samples)?

1. Data quality: ρ(R,G): 0
   – design based philosophy
   – Quality and quantity are independent (?)
2. Data quantity: f very large (close to 1)
   – Big data philosophy
   – Quality and quantity negatively correlated?
3. **σ**(G): very small

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}} .$$

Solutions

- 1.Global correction methods
  - Pseudo design based estimation (Elliott & Valliant 2017)
- 2. Estimate-specific methods
  - Calibration (Little, 2004)
    - non-prob -> probability
  - Superpopulation modeling (e.g. Elliott & Valliant, 2020)
  - Mass imputation (Yang and Kim, 2020)
- 3. Sensitivity analyses
  - Meng: for $\rho(R,G)$
  - Pattern mixture models for NMAR (e.g West 2020)

Pseudo design based

non-probability based

gende r	age	educa tion	he alt h	Fav ice
0	34	1	5	vanilla
1	54	2	5	lemon
1	12	3	4	Choc
1	56	3	5	vanilla
0	87	4	2	strawb
1	45	5	3	zabaione
1	67	6	4	lemon
1	23	6	5	straccia
0	16	2	5	vanilla
1	24	4	4	straccia
1	56	2	4	straccia
1	78	3	2	vanilla

Taste	percentage
Vanilla	33%
Lemon	16%
Straccia	25%
Zabaione	8%
Strawberry	8%
Chocolate	8%

Pseudo design based

non-probability based

gender	age	education	health	Fav ice
0	34	1	5	vanilla
1	54	2	5	lemon
1	12	3	4	Choc
1	56	3	5	vanilla
0	87	4	2	strawb
1	45	5	3	zabaione
1	67	6	4	lemon
1	23	6	5	Banana
0	16	2	5	vanilla
1	24	4	4	pear
1	56	2	4	straccia
1	78	3	2	vanilla

Other Probability based survey

gender	age	education	health	P(Response)
0	34	1	5	,24
1	54	2	5	,44
1	12	3	4	.23
1	56	3	5	.56
0	87	4	2	,36
1	45	5	3	.56
1	67	6	4	.44
1	23	6	5	.33
0	16	2	5	,32
1	24	4	4	.43
1	56	2	4	.42
1	78	3	2	.43

Pseudo design based

non-probability based

gende r	age	educa tion	he alt h	P(Respon se)	Fav ice
0	34	1	5	,24	vanilla
1	54	2	5	,44	lemon
1	12	3	4	.23	Choc
1	56	3	5	.56	vanilla
0	87	4	2	,36	strawb
1	45	5	3	.56	zabaione
1	67	6	4	.44	lemon
1	23	6	5	.33	straccia
0	16	2	5	,32	vanilla
1	24	4	4	.43	straccia
1	56	2	4	.42	straccia
1	78	3	2	.43	vanilla

Taste	Raw percentage	Weight (1/p)
Vanilla	33%	1/.39
Lemon	16%	1/.44
Straccia	25%	1/.39
Zabaione	8%	1/.56
Strawberry	8%	1/.36
Chocolate	8%	1/.23

Pseudo design based

non-probability based

gender	age	education	health	P(Response)	Fav ice
0	34	1	5	,24	vanilla
1	54	2	5	,44	lemon
1	12	3	4	.23	Choc
1	56	3	5	.56	vanilla
0	87	4	2	,36	strawb
1	45	5	3	.56	zabaione
1	67	6	4	.44	lemon
1	23	6	5	.33	straccia
0	16	2	5	,32	vanilla
1	24	4	4	.43	straccia
1	56	2	4	.42	straccia
1	78	3	2	.43	vanilla

Taste	Raw percentage	Weight (1/p)	Weighted %
Vanilla	33%	1/.39	33%
Lemon	16%	1/.44	14%
Straccia	25%	1/.39	24%
Zabaione	8%	1/.56	6%
Strawberry	8%	1/.36	9%
Chocolate	8%	1/.23	14%
Ave		.40	

Estimate specific methods

– Calibration (Little, 2004)
  - Bayes

- Superpopulation modeling (e.g. Elliott & Valliant, 2020)

– Mass imputation (Yang and Kim, 2020)

2.1 Calibration

- Conduct a large nonprobability sample
  - Small s.e., large bias(?)
- Conduct a small probability based sample
  - Large s.e., small bias
- Weight non-probabity based -> prob based
  - Small bias (?), small s.e.
  - You can use lots of survey questions, because you conduct 2 surveys
  - Expensive, time consuming

2.1 Little (2004) Calibrated bayes

- Model based (regression) vs. design based
- Solution:
  - Use a model that includes design-based features
    - E.g. A fixed-effects regression model to deal with clustering
  - Bayesian modeling for variance estimation
    - Priors (often uninformative)
    - Posteriors for variance estimation
    - Remember convergence, traceplots,, and how imputations are generated in Mice?

2.2 Superpopulation modeling

- Non-probability based surveys don’t use sample frames
  - We can rake or calibrate to population statistics:

    gender, age, region, ethnicity, income, educatio
- However, what if take an effort to collect more population statistics?
  - Netflix subscription? Voting Behavior, customer of a company, member of organization

2.2 Superpopulation modeling

- However, what if take an effort to collect more population statistics?
  - Netflix subscription? Voting Behavior, customer of a company, member of organization
- i.e. More elaborate weighting

Source:https://www.pewresearch.org/methods/2018/01/26/reducing-bias-on-benchmarks/

Topics and corresponding benchmarks

Topic	Benchmark
Civic engagement	How often talks with neighbors
	Trusts neighbors
	Participated in a school group, neighborhood, or community association
	Volunteered in past year
Family	Marital status
	Presence of children in household
	Household size
Financial	Employment status
	Home ownership
	Family income
	Household member received food stamps
	Health insurance
Personal	Lived in house or apartment one year ago
	Active duty military service
	U.S. citizenship
	Gun ownership
	Smoking
	Food allergies
Political engagement	Voted in 2012
	Voted in 2014
	Contacted or visited a public official in past year
Technology	Tablet or e-reader use
	Texting or instant messaging
	Social networking

Note: See Appendix D for the source of each benchmark, the question text, the response categories, the benchmark estimate, and additional notes.
"For Weighting Online Opt-In Samples, What Matters Most?"
PEW RESEARCH CENTER

2.3 Mass imputation

- We know the population distribution:
  - Gender, age, education, income, region, etc.
- In some cases we have frame data
- Why not impute the whole population?

Mass imputation

- We know the population distribution:
  – Gender, age, education, income, region, etc.
- In some cases we have frame data
- Why not impute the whole population?

gender	age	education	health	Fav ice
0	34	1	5	vanilla
1	54	2	5	lemon
1	12	3	4	Choc
1	56	3	5	vanilla
0	87	4	2	strawb
1	45	5	3	zabaione
1	67	6	4	lemon
1	23	6	5	straccia
0	16	2	5	vanilla
1	24	4	4	straccia
1	56	2	4	straccia
1	78	3	2	???
1	56	4	5	???
….	…	…	..	???
You have X million rows, only X thousand of thesehave Y				

3. Sensitivity analyses

- Cf Meng (2018)
- Pattern Mixture modeling
  - Enter an additional parameter in the model (e.g a selection bias parameter)
  - This parameter can take different forms
    - Covary with Y and all other parameters
  - Simulate
  - Similar to Heckman selection models.

See Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153.

Exercise (class + THE)

- Competition!
  - Three non-probability samples
  - Sample size 30.000
  - June/July 2016
  - You get 15.000 cases
  - And a superpopulation dataset (Mercer, Lau & Kennedy, 2018)
- Goal: adjust your sample:
  - Choose your variables
  - Calibrate, rake, impute?
- Prize: eternal fame and a survey related present

Next week

- Lecture on “designed big data”
- Keep working on your group assignments
- In two weeks -> presentation
  – See schedule on Blackboard
  – Grading rubric

More reading?

- Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153.
- Chen, S., Yang, S., & Kim, J. K. (2020). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*.
- Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.
- Kim, J. K., Park, S., Chen, Y., & Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation. *arXiv preprint arXiv:1812.10694*.
- Rafei, A., Flannagan, C. A., & Elliott, M. R. (2020). Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees. *Journal of Survey Statistics and Methodology*, *8*(1), 148-180.
- Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231-263.