INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Errors in election polls

Raphael Nishimura

November 27, 2020, UFSCar/USP

Outline

- Introduction
- Inferential approaches
- Total Survey Error
- The Literary Digest (1936 U.S. Presidential Election)
- 2016 U.S. Presidential Election Polls

"*All polls are wrong.*
Some polls are useful"
- C. Joy Wilke, 2020

Crisis in Election Polls?

Introduction

- Election polls are a **finite population, descriptive inference** problem
- Well-defined (in space and **time**) finite population *U* of size *N*
  - For example: Votes in the U.S. Presidential Election by April 3, 2020
- Interested in estimating a finite population parameter, say a population total:

$$T_y = \sum_{i=1}^{N} Y_i$$

Introduction

- (Pre-)Election polls are also a two-fold **prediction** problem:

1. With a (responding) sample s of size $n \ll N$, estimate the finite population parameter T_y by predicting the *Y*-values of the $N - n$ unobserved cases:

$$\hat{T}_y = \sum_s y_i + \sum_{U-s} \hat{y}_i$$

2. Predicting the finite population parameter T_y on time *t* using a sample selected on time *t-k*, *k* > 0 (Forecasting modeling)

Inferential approaches

- Design-based inference
  - Inference based on repeated sampling distribution
    - Only applicable for sampling error of probability-based sampled
    - Example: Horwitz-Thompson estimator

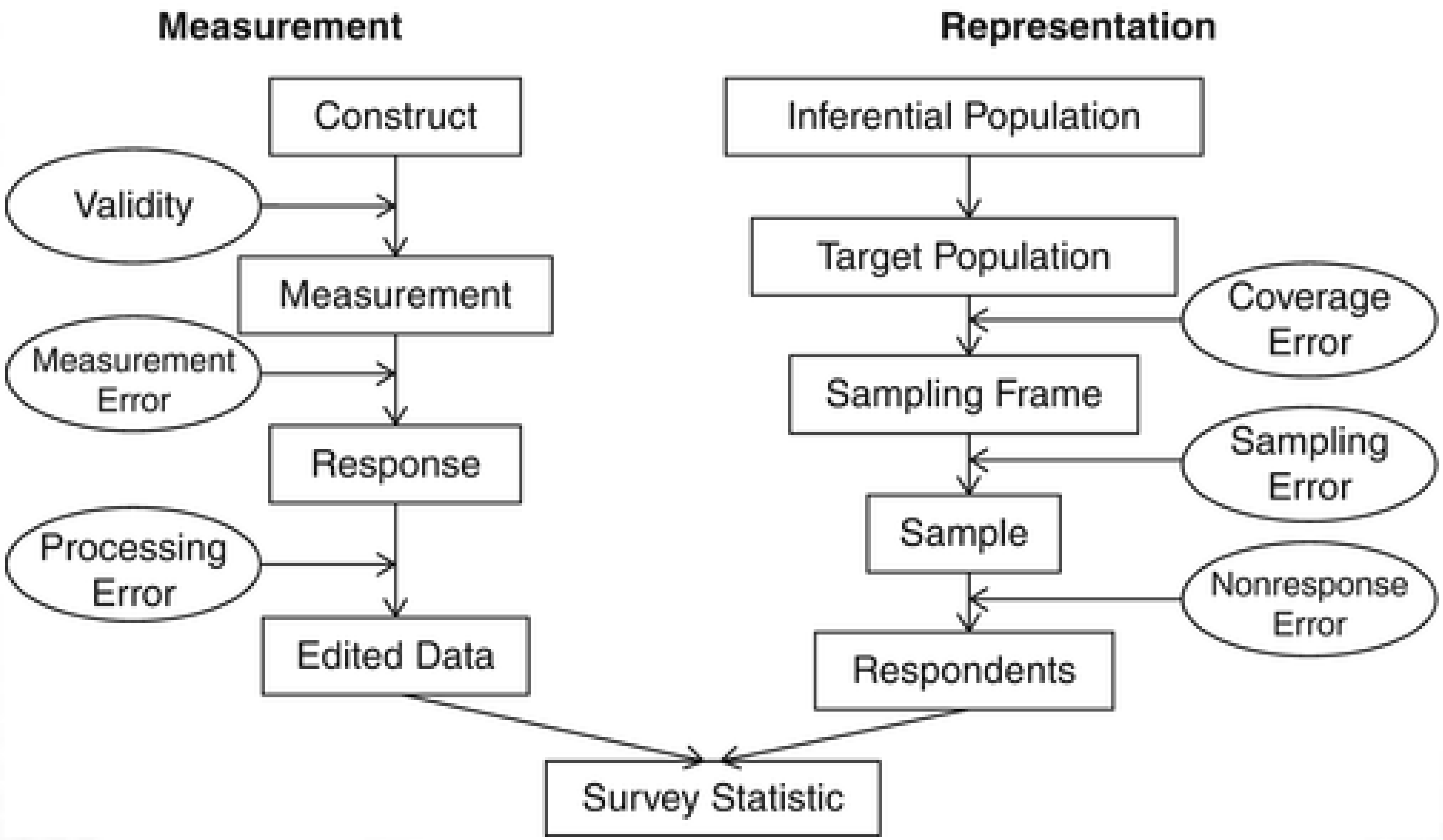
$$\hat{T}_y = \sum_s \frac{y_i}{\pi_i} \qquad E_\pi\left(\hat{T}_y\right) = T_y$$

- Model-based inference
  - Impose a stochastic model to variable *y* and evaluate estimators based with respect to the model: $E_M\left(\hat{\theta}\right) = \theta$
- Model-assisted inference
  - Compromise between design- and model-based
    - Models used to construct estimators
    - Repeated sampling distribution used for inference

Total Survey Error

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = \left[B(\hat{\theta})\right]^2 + V(\hat{\theta})$$

Total Survey Error - Survey cycle



Source: Groves et al (2011)

Total Survey Error

$$MSE(\hat{\theta}) =$$

$$[B_C(\hat{\theta})]^2 + V_C(\hat{\theta}) +$$

$$[B_S(\hat{\theta})]^2 + V_S(\hat{\theta}) +$$

$$[B_R(\hat{\theta})]^2 + V_R(\hat{\theta}) +$$

$$[B_M(\hat{\theta})]^2 + V_M(\hat{\theta}) +$$

...

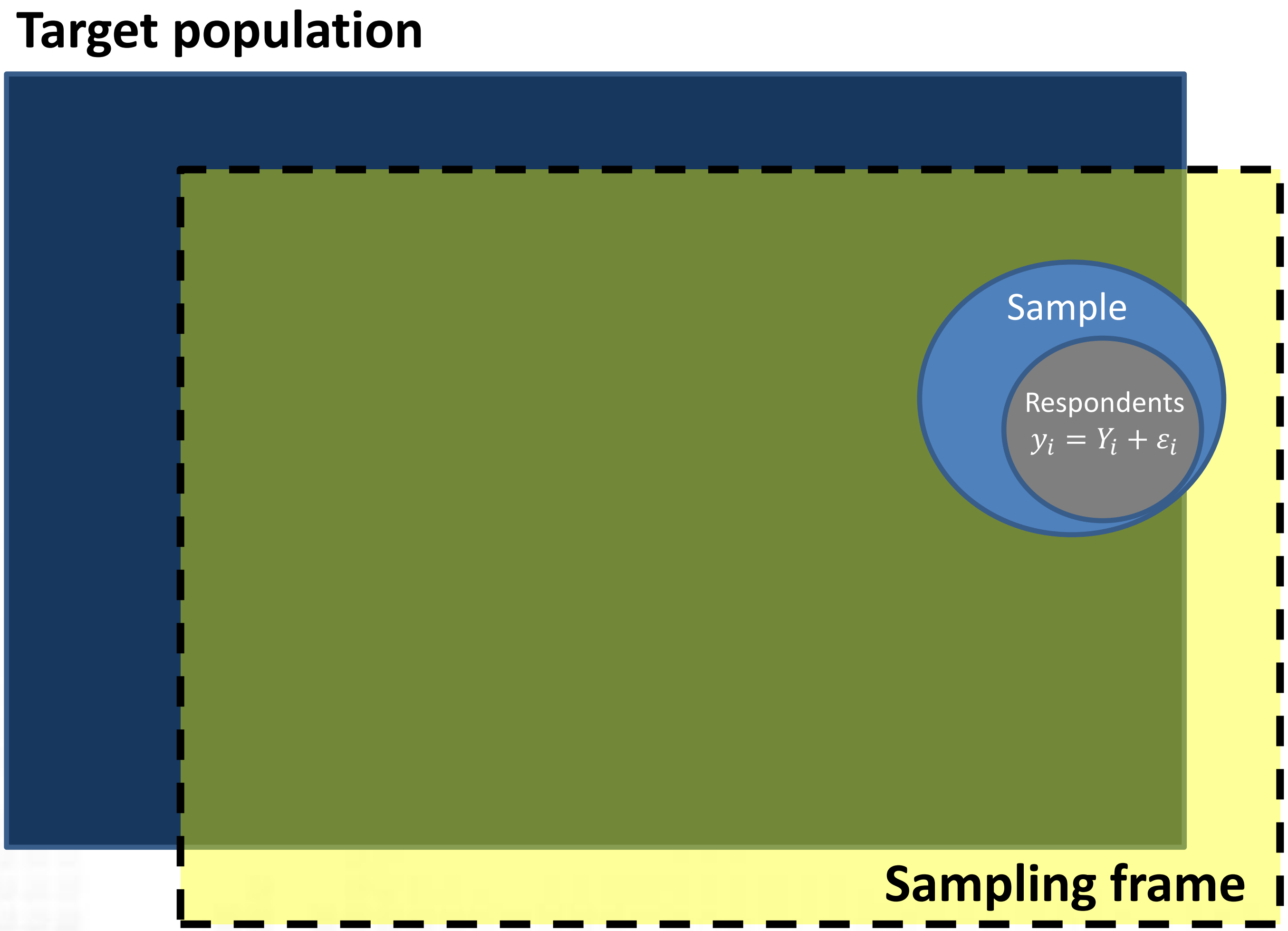
Margin of (sampling) error

$$\left(\pm 1.96\sqrt{V_s(p)}\right)$$

Surveys/Polls: How people think it is…

Target population

Surveys/Polls: How it really is…



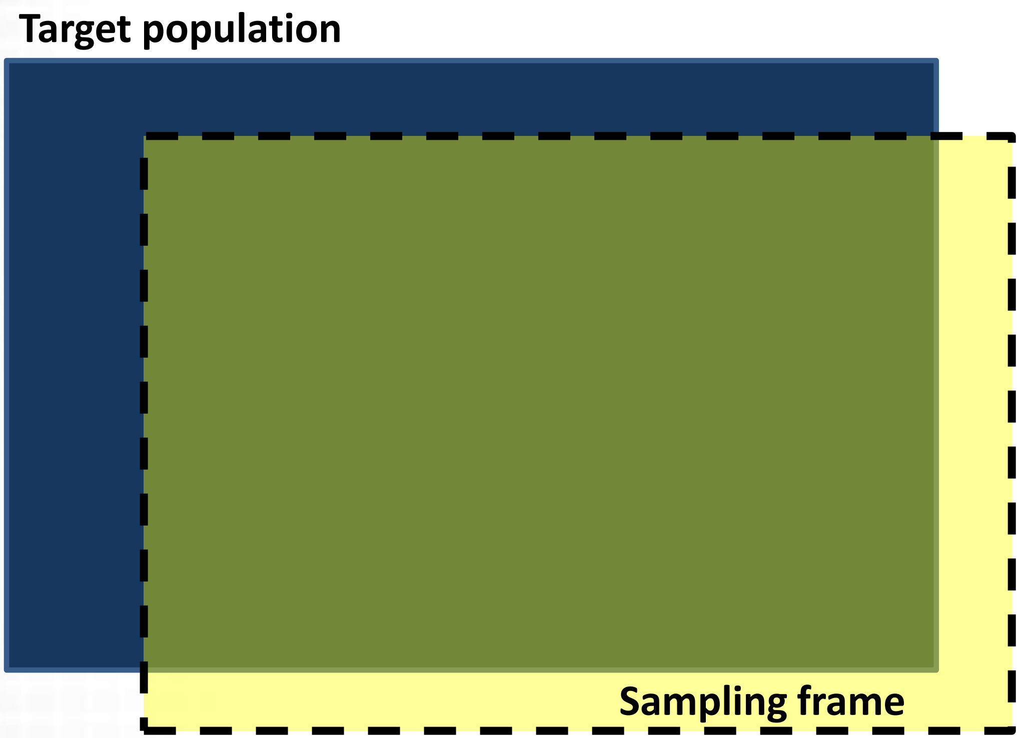
Coverage error

Target population

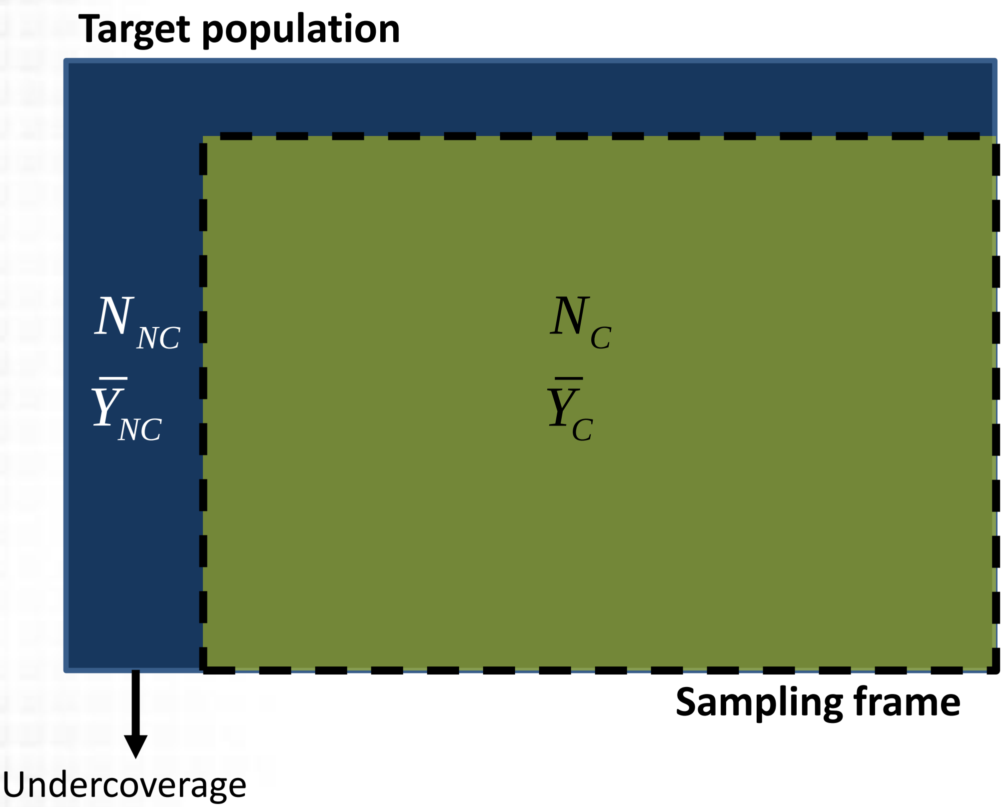
N = Population size

$\bar{Y}$ = Population mean for survey variable Y

Coverage error



Coverage error



(Under)Coverage bias:

$$Bias(\bar{y}) = \underbrace{\frac{N_{NC}}{N}}_{\text{Undercoverage rate } = \left(1 - \frac{N_C}{N}\right)} \left(\bar{Y}_C - \bar{Y}_{NC}\right)$$

N = Overall population size

N_{NC} = Non-covered population size

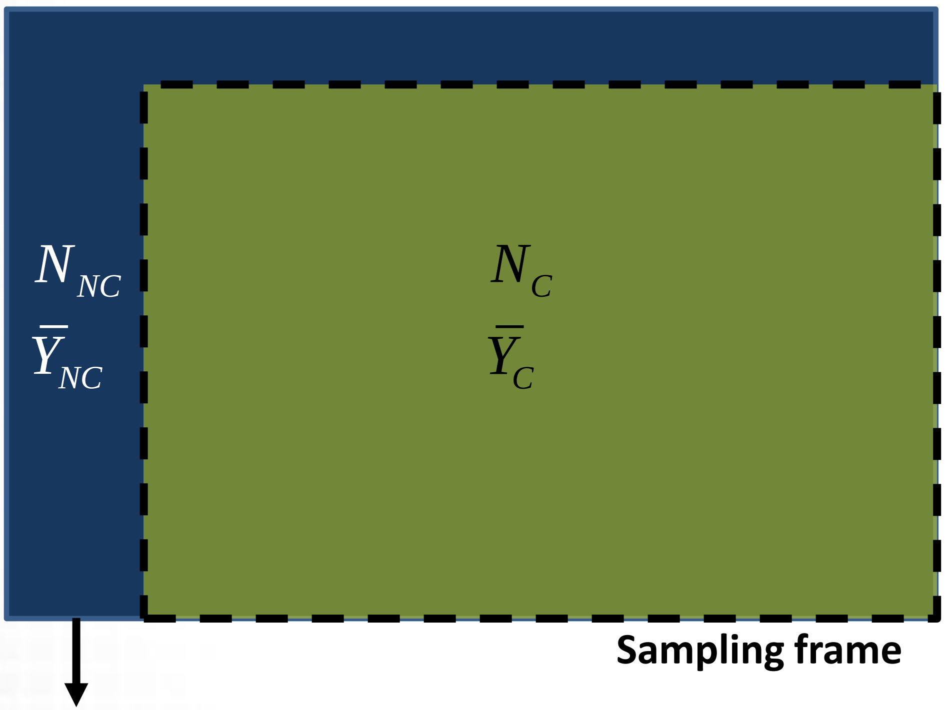
$\bar{Y}_{NC}$ = Non-covered population mean for survey variable Y

N_C = Covered population size

$\bar{Y}_C$ = Covered population mean for survey variable Y

Coverage error

Target population



Sampling frame

Undercoverage

(Under)Coverage bias:

$$Bias(\bar{y}) = \frac{N_{NC}}{N}\underbrace{\left(\bar{Y}_C - \bar{Y}_{NC}\right)}_{\text{Difference between the covered and non-covered populations}}$$

N = Overall population size

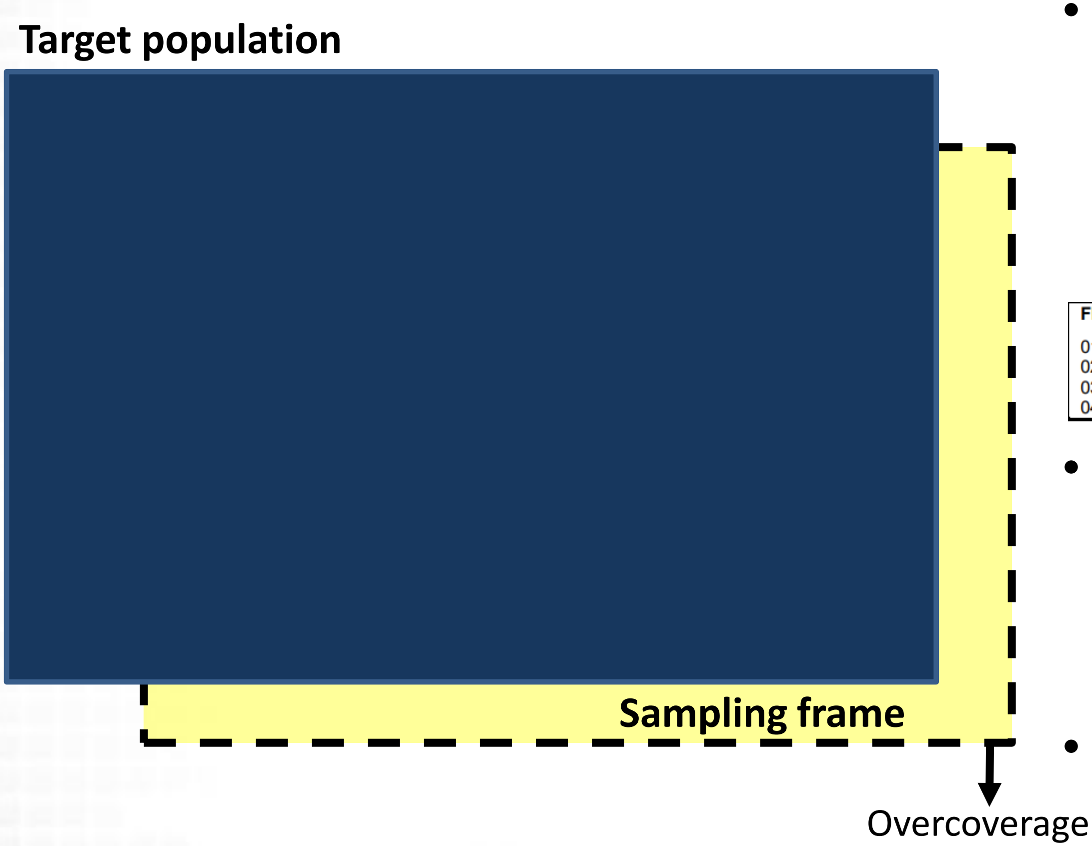
N_{NC} = Non-covered population size

$\bar{Y}_{NC}$ = Non-covered population mean for survey variable Y

N_C = Covered population size

$\bar{Y}_C$ = Covered population mean for survey variable Y

Coverage error



- Typically dealt with screening
  - Example: Municipal election poll → screen-out respondents not registered to vote in the municipality

FILTRO 1) O(a) sr(a) tem título de eleitor? **(CASO SIM)** Vota neste município ou em outro?

01( ) Vota nesse município.............................**APLIQUE FILTRO 2**
02( ) Vota em outro município deste Estado..... **ENCERRE**
03( ) Vota em município de outro Estado...........**ENCERRE**
04( ) Não tem título...**ENCERRE**

- Problems with Pre-election polls:
  - Voting not mandatory
  - Abstention
- Solution: Likely voter models
  - Screening out
  - Turnout score weighting

Nonresponse error

- Unit and item nonresponse

Sample information

Data

001101501	4714221866157137261333346462 .6918657113 .646748168500098992341
001101502	37249 .16114058461581364481306478916646548664656 .38460063
001101503	4554621610682165131631698378841953267865165136065736389 13216
001101504	4726606587193349009319384516167697981832582541937986 61960001
001101505	4119871613194816720067103456467803987198434 6388 .79841316
001101506	6913330687809797109480945097719333494161044060067981949315561
001101507	77135410906781971610654619789741987413102009 0816 .61618165
001103601	8722321378941964549879841513208974009874115121212585555 15115
001103602	651292161139751984165108006815015515151020656751616515511111
001103603	53258 .719675891961651649194865161156789410058405405151251155
001103604	51148298412132021212111210021260681251026546516 .37781335
...	...
001141201	. . . .
001142215	. . . .
002103601	. . . .
...	...

Unit non-response

Item missing values

Nonresponse error

- Unit and item nonresponse
- Missing mechanisms
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)
- Nonresponse bias
  - Deterministic

$$B(\bar{y}_R) = \frac{M}{N}(\bar{Y}_R - \bar{Y}_{NR})$$

  - Stochastic

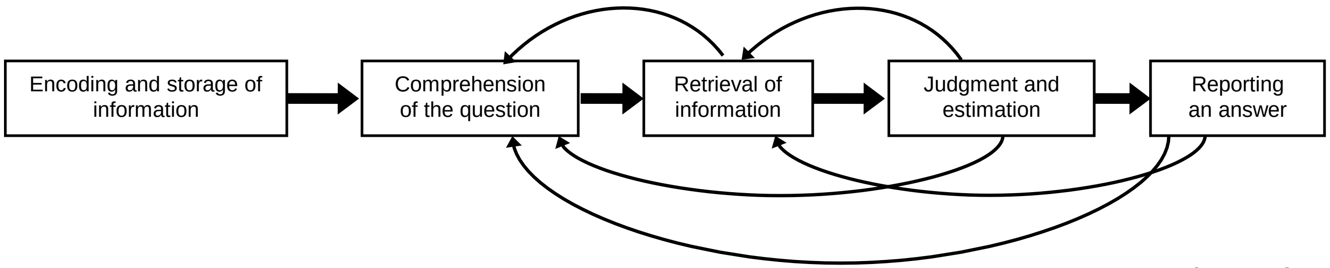
$$B(\bar{y}_R) \approx \frac{1}{\bar{\phi}} \frac{\sum (Y_i - \bar{Y})(\phi_i - \bar{\phi})}{N}$$

Measurement error

- Simple response error model

$$y_i = Y_i + \varepsilon_i$$

- Simple response process model

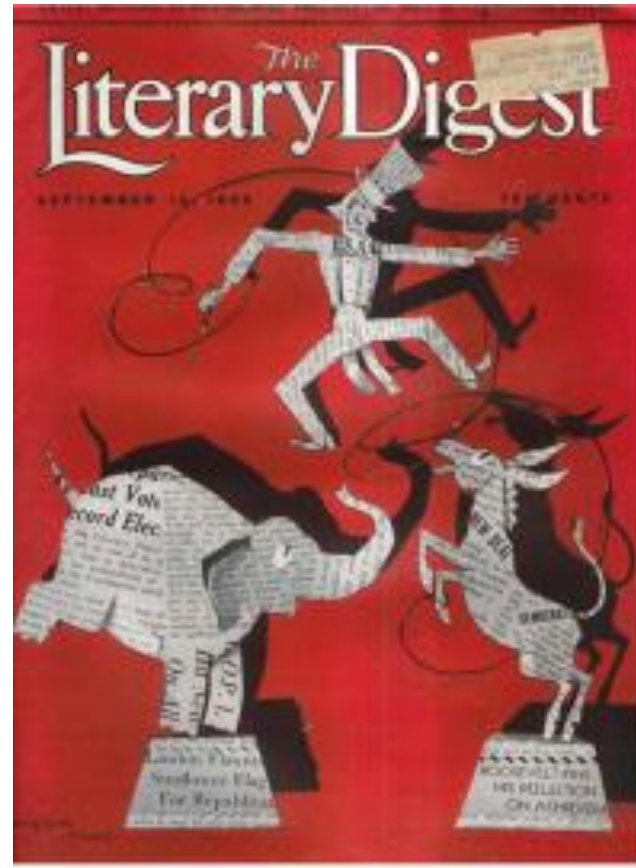


Source: Groves et al (2011)

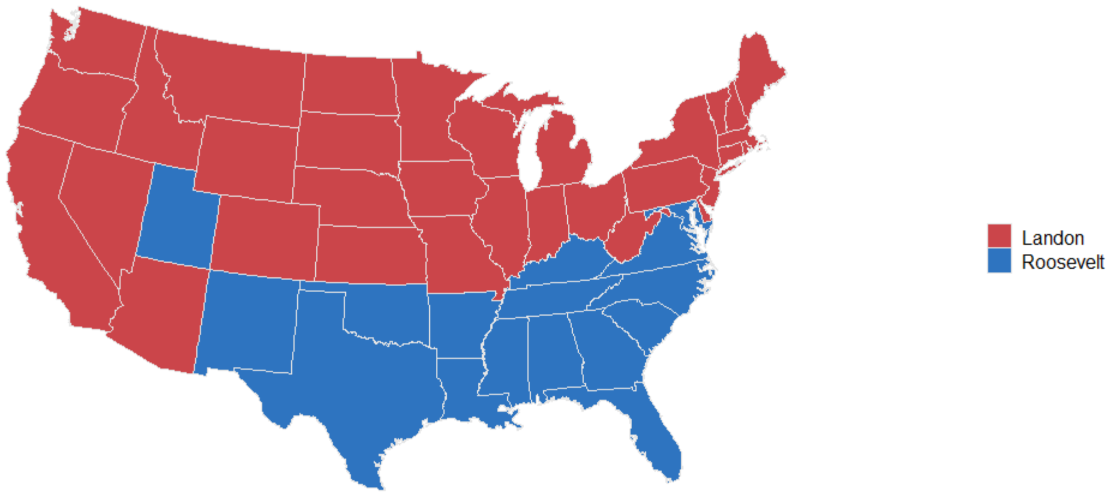
- Questionnaire effects
  - Examples: primacy, recency, order effects
- Interviewer effects
  - Example: Social desirability, interviewer characteristics

The Literary Digest

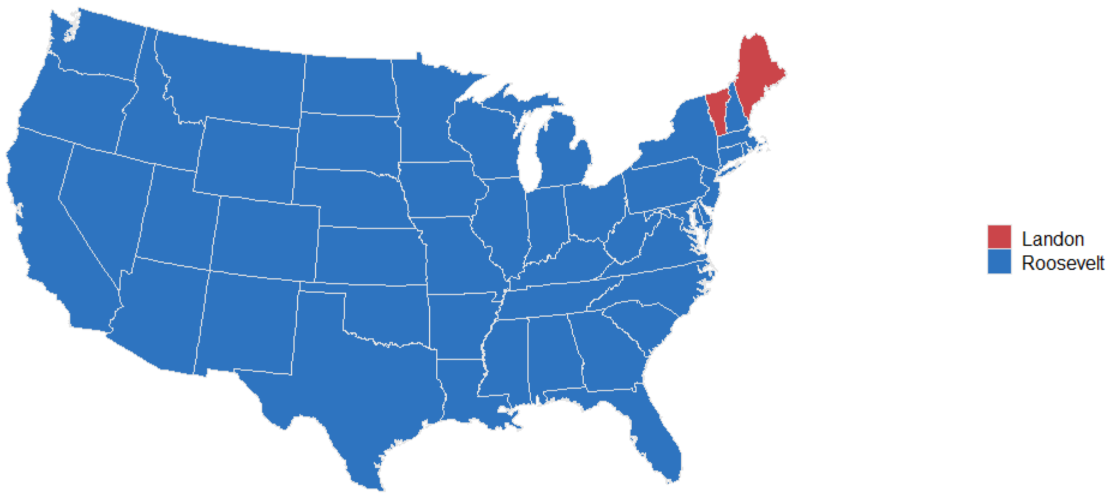
- Accurately predicted 1920, 1924, 1928 & 1932 presidential elections
- 1936 Presidential Poll
  - 10 million ballots sent by mail
  - n ≈ 2,27 million (!!!) respondents (RR=24%)
  - Literary Digest forecast:
    - Landon 57% vs Roosevelt 43%
  - Election results:
    - Landon 39% vs Roosevelt 61%



Literary Digest Election Forecast



1936 Election Results



The Literary Digest: What did go wrong?

- Coverage bias
  - Sampling frame: Magazine subscribers, automobile registration lists and telephone directory
- Nonresponse bias
  - "Low" response rate (24%) and differential nonresponse
- Lohr and Brick (2017)
  - Weighting adjustment by 1932 election vote by state
  - Election results predicts Roosevelt as winner, but estimates are still biased
- See also Meng (2018)
  - *Big Data Paradox*: the more the data, the surer we fool ourselves

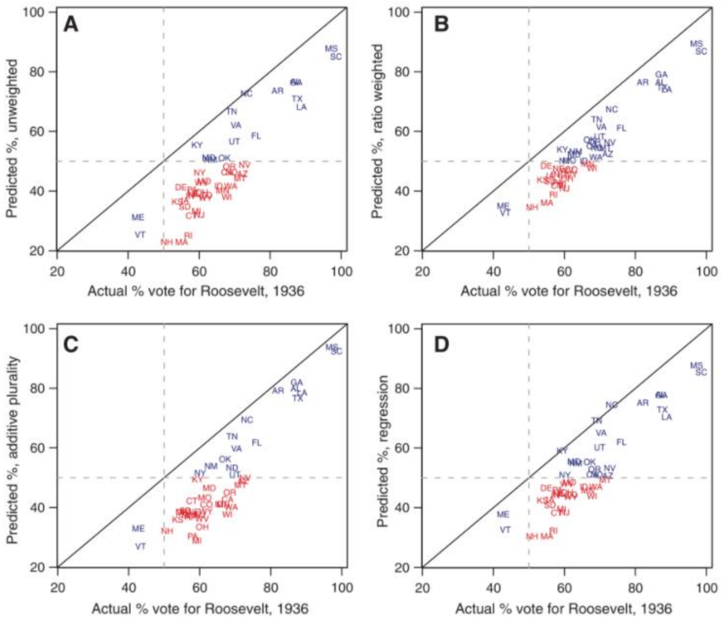


Figure 2: Predicted percentage of vote for Roosevelt in the 1936 election, using (A) unweighted counts from *Literary Digest* poll, (B) ratio adjustment, (C) Robinson's (1932) additive plurality adjustment, and (D) regression prediction.
Note: The states in blue (upper right and lower left quadrants) are those for which the poll predicted the correct winner of the state. The states in red (lower right quadrant) are those for which the poll predicted the wrong candidate would win.

Source: Lohr and Brick (2017)

2016 U.S. Presidential Election Polls

	NYT	538	HuffPost	PW	PEC	DK
Win	85% D	71% D	98% D	89% D	>99% D	92% D

Note: The 538 model shown is its default (polls-only) forecast.

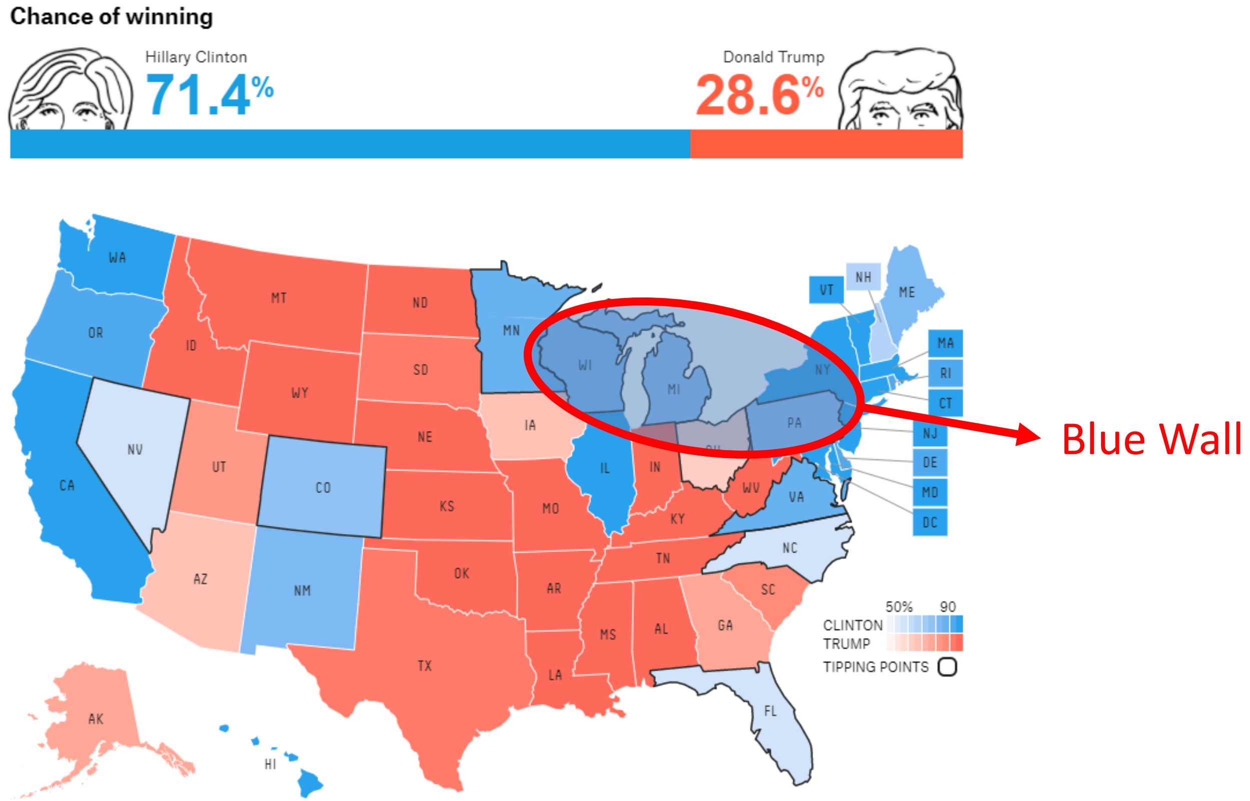
FORECAST

PRESIDENT SENATE

By Natalie Jackson and Adam Hooper
Additional design by Alissa Scheller

PUBLISHED MONDAY, OCT. 3, 2016 12:56 P.M. EDT
UPDATED TUESDAY, NOV. 8, 2016, 12:43 A.M. EST

2016 U.S. Presidential Election Polls



Electoral votes

Hillary Clinton	302.2
Donald Trump	235.0
Evan McMullin	0.8
Gary Johnson	0.0

Popular vote

Hillary Clinton	48.5%
Donald Trump	44.9%
Gary Johnson	5.0%
Other	1.6%

2016 U.S. Presidential Election Polls: Post-mortem

- AAPOR Evaluation of 2016 Election Polls in the U.S. (2017):
  - National polls generally correct and accurate by historical standards

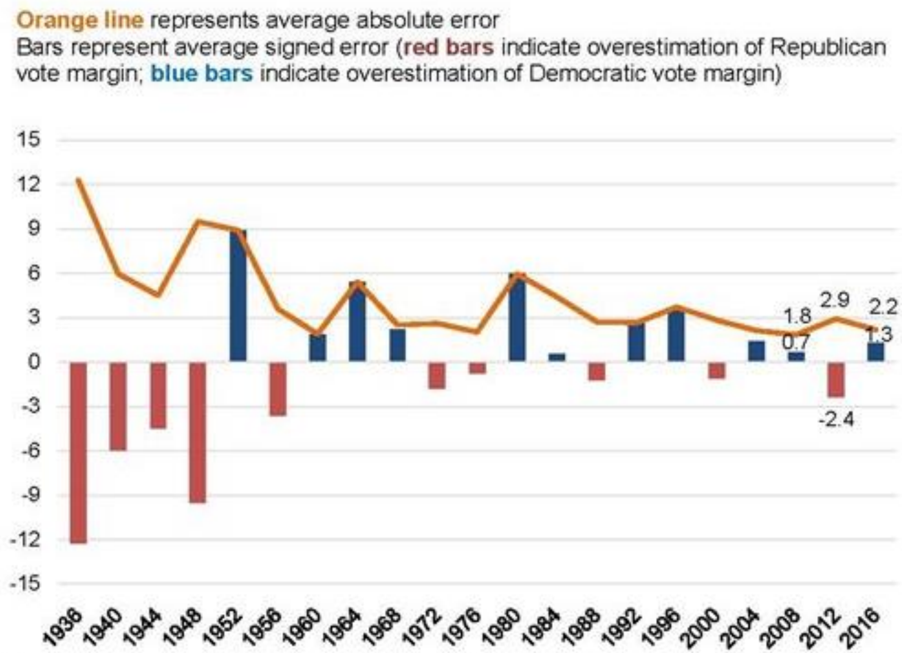


Figure 1. Average Error in Vote Margin in National Presidential Polls, 1936-2016.
Note – The 2016 figures are based on polls completed within 13 days of the election. Figures for prior years are from the National Council for Public Polls analysis of final poll estimates, some occurring before the 13-day period. Figures for 1936 to 1960 are based only on Gallup.

2016 U.S. Presidential Election Polls: Post-mortem

- AAPOR Evaluation of 2016 Election Polls in the U.S. (2017):
  - State-level polls showed a competitive, uncertain contest, but clearly under-estimated Trump's support in the Upper Midwest

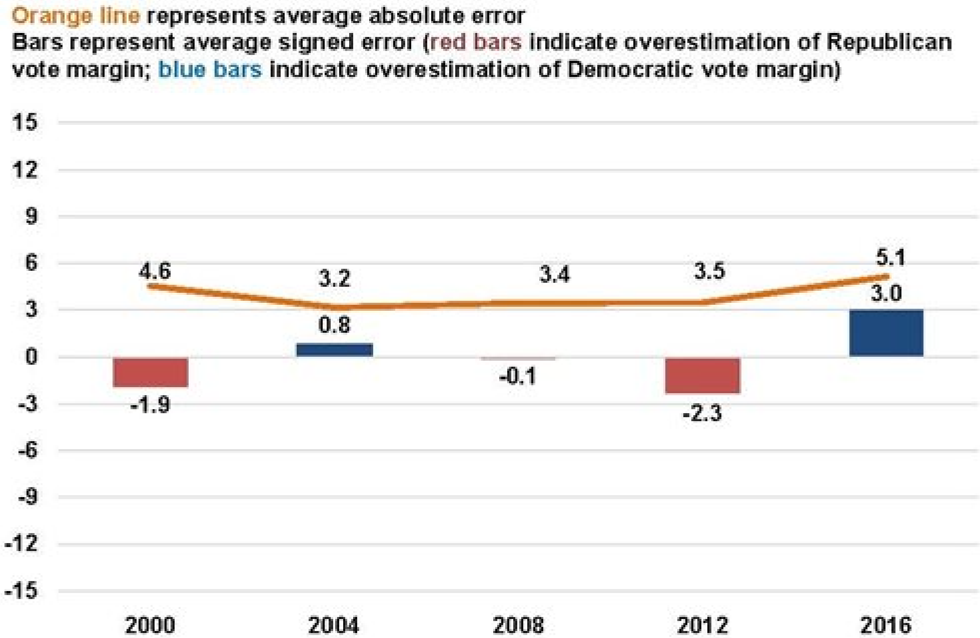
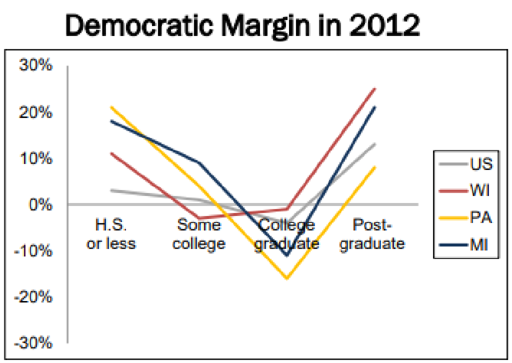


Figure 2. Average Error in Vote Margin in State Presidential Polls, 2000-2016.
Source – Figures for 2000 to 2012 computed from data made public by FiveThirtyEight.com.

2016 U.S. Presidential Election Polls: Post-mortem

- AAPOR Evaluation of 2016 Election Polls in the U.S. (2017):
  - Why polls under-estimated support for Trump?
    - Real change in vote preference during the final week or so of the campaign
    - Unadjusted differential nonresponse bias due to overrepresentation of college graduates, which was correlated with Clinton support



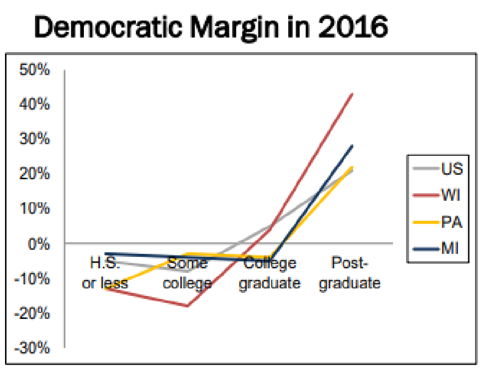


Table 10. Share of Pollsters That Adjusted on Education in Weighting

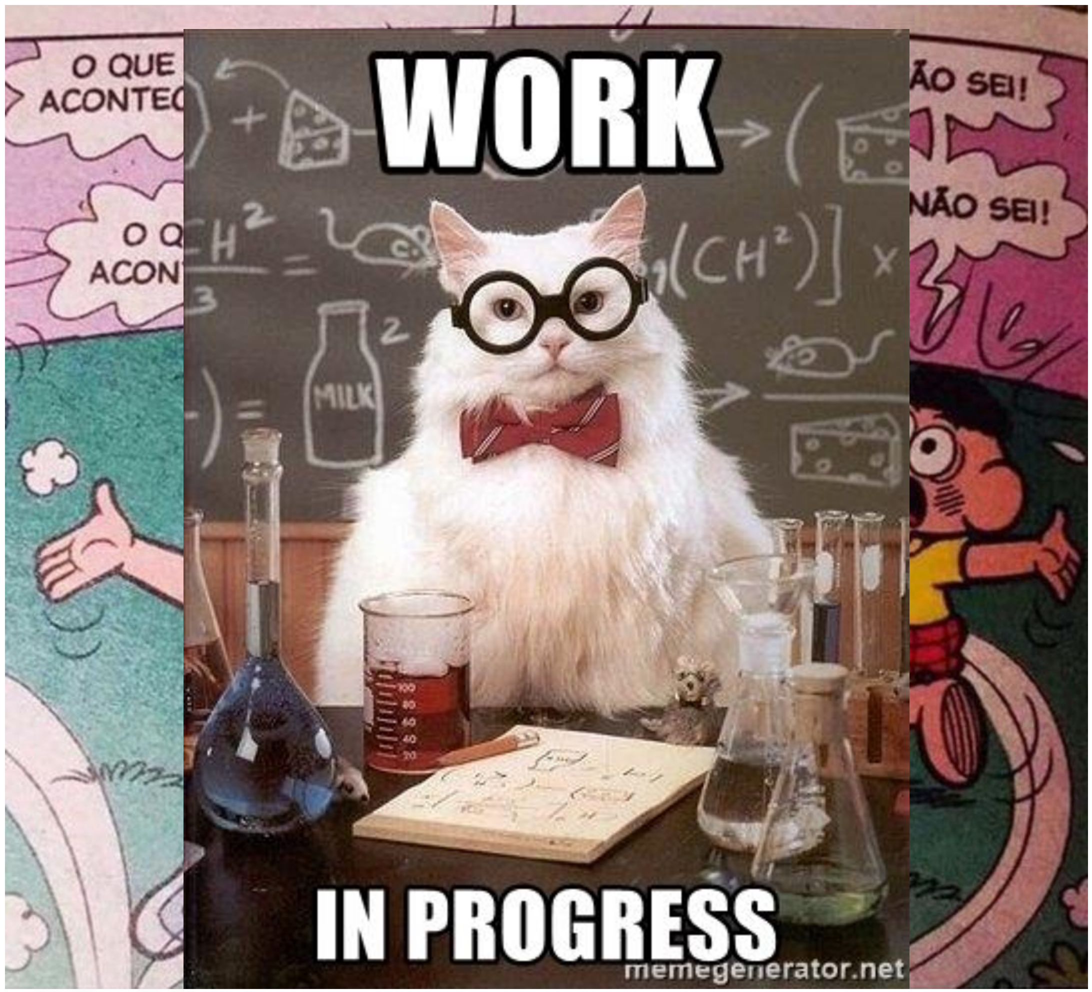
Type of Poll	Share of polls that weighted for education	Number of final polls
Michigan polls	18%	11
Wisconsin polls	27%	11
North Carolina polls	29%	14
Florida polls	31%	16
Pennsylvania polls	33%	18
Ohio polls	36%	11
National polls	52%	21

Note - Figures reflect only polls fielded in the final two weeks and only a given pollster's final poll. The requisite weighting information was missing for 23 polls, which were all imputed as not weighting on education, based on information among similar polls that did disclose their weighting variables.

Source: NEP national Exit Poll 2012, 2016

    - *Shy Trump* effect: Little evidence supporting hypothesis
    - Turnout patterns changed between 2012 and 2016 could have led to mistakes in likely voter models

2020 U.S. Presidential Election Polls?



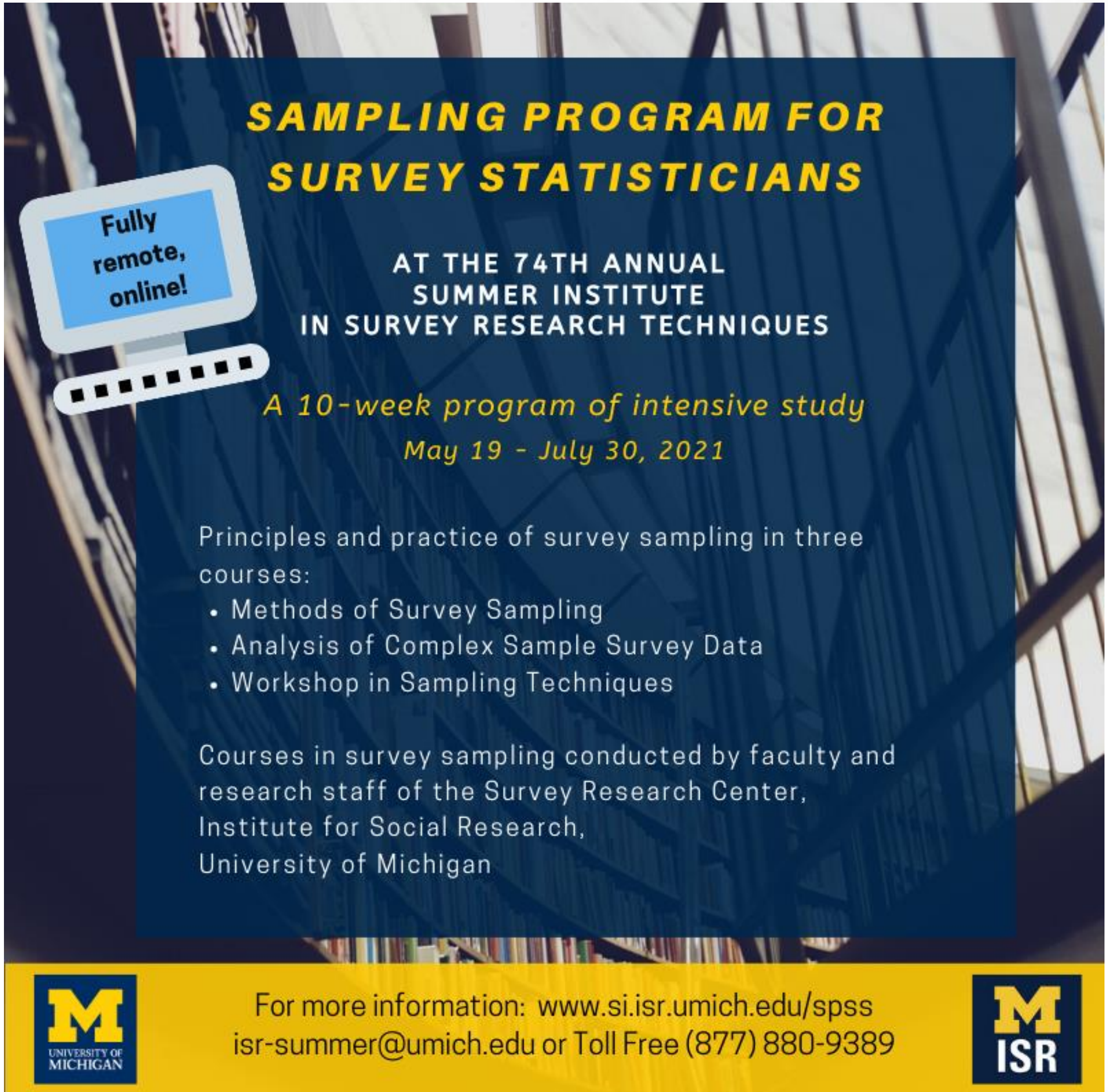
References

- AAPOR An Evaluation of 2016 Election Polls in the U.S. (2017): https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

- Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey methodology (Vol. 561). John Wiley & Sons.

- Lohr, Sharon L., and J. Michael Brick. "Roosevelt Predicted to Win: Revisiting the 1936 Literary Digest Poll." Statistics, Politics and Policy 8, no. 1 (2017): 65-84.

- Meng, Xiao-Li. "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election." The Annals of Applied Statistics 12, no. 2 (2018): 685-726.


SAMPLING PROGRAM FOR SURVEY STATISTICIANS
Fully remote, online!
AT THE 74TH ANNUAL SUMMER INSTITUTE IN SURVEY RESEARCH TECHNIQUES
A 10-week program of intensive study
May 19 - July 30, 2021
Principles and practice of survey sampling in three courses:
• Methods of Survey Sampling
• Analysis of Complex Sample Survey Data
• Workshop in Sampling Techniques
Courses in survey sampling conducted by faculty and research staff of the Survey Research Center, Institute for Social Research, University of Michigan
For more information: www.si.isr.umich.edu/spss
isr-summer@umich.edu or Toll Free (877) 880-9389
UNIVERSITY OF MICHIGAN
ISR

Other resources

- Summer Institute in Survey Research Techniques
  - https://si.isr.umich.edu/
- Michigan Program in Survey Methodology
  - https://psm.isr.umich.edu/
- International Program in Survey and Data Science
  - https://survey-data-science.net/
- 2021 AAPOR Conference
  - https://www.aapor.org/Conference-Events/Annual-Meeting.aspx

INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Thank you!

Raphael Nishimura
raphaeln@umich.edu
@rnishimura