

Survey Research and Design

Political Polling and Election Forecasting

William Marble

October 17, 2023

MINOR IN...

```
from minor import Data
# do I have enough credits for the minor?
def does_minor_require(courses, electives):
    credits = 0
    # ADD up the core courses I've taken
    if "Intro Data Science & Programming" in courses:
        credits += 3
    if "Math & Statistics" in courses:
        credits += 3
    if "Applied Data Science" in courses:
        credits += 3
    # ADD up electives
    credits += len(electives)
    # did I complete the DASA minor?
    if credits == 12:
        print("Success! I have completed the DASA minor!")
    else:
```

Data Science Knowledge

Time as a DASA minor

DATA SCIENCE & ANALYTICS

OPEN TO ALL STUDENTS

DEVELOP YOUR DATA
SCIENCE SKILLS

LEARN APPLICATION
& THEORY

LEARN MORE



QUESTIONS?

EMAIL: JOELLE GROSS JOELLEG@SAS.UPENN.EDU
ASSISTANT DIRECTOR OF DATA SCIENCE & ANALYTICS

Writing Your Survey

- ▶ Continue working on your survey in Qualtrics
- ▶ Keep it polished – do let draft questions hang around
- ▶ Periodically "preview survey" to see how it looks and feels to take
- ▶ Next Tuesday, we will finalize the surveys in class
- ▶ You will export your survey from Qualtrics and send it to us to compile with others
- ▶ Check out question types in Qualtrics!

Asking Many Similar Questions at Once

How much do you approve or disapprove of the job performance of...

	Strongly approve	Somewhat approve	Neither approve nor disapprove	Somewhat disapprove	Strongly disapprove
Democrats in Congress	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Republicans in Congress	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
President Joe Biden	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The Supreme Court	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phillies manager Rob Thomson	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Constant Sum Questions



Please total the choices to 10.

Imagine that you had 10 coins to allocate to the following categories according to how much you identify with the following groups. Place the 10 coins into each category. If you put all 10 into one bucket, it would mean that you identify entirely with that group and not at all with the other groups. If you spread the coins equally across the groups, it would mean that you identify equally with each of the groups.

Americans

1

People living in your city or town

8

People in your occupation

8

People in your ethnic or racial group

0

Total

17

Ranking Questions

Please rank the following political issues from most to least important.

1 Environmental protection

2 Abortion

3 Guns

4 Taxes

5 Trade

6 Labor unions

Goals of Surveys

- ▶ Measure
- ▶ Describe
- ▶ Explain
- ▶ Predict

“Horserace” Polling

- ▶ Who will win in the upcoming elections?
- ▶ Perhaps the most visible type of survey: polls sponsored by and covered by the news media
- ▶ Important in shaping trust in public opinion research
- ▶ But there are unique challenges relative to most public opinion surveys



Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the infinitesimal
poll of ten million voters has been fought
throughout the forty-eight States of the

Lower National Committee purchased Ten
Election Day Boxes?" And all types and sorts

of them, including "Have the Jews purloined

rottened and let the people of the Nation
draw their conclusions as to our security.
Are we safe? Are we free? Will we be safe?
Will we be right in the current P.D.F.? That,
as Mrs. Roosevelt said concerning the P.D.F.,
"We have got to be safe and we have got to be right."
"We never make any claims before election,
but we respectfully refer you to the
conclusion of one of the most trusted sources



ELECTIONS

Battle for Senate majority remains a nail-biter

The 50-50 Senate is still finely balanced a month out from Election Day, the latest surveys show.

Important Caveat

Pre-election polls are at best *predictions*. That said...

- ▶ Widely used for understanding/interpreting political dynamics
- ▶ Campaigns use them for allocating resources
- ▶ They're useful for survey researchers because they present a rare chance for validation

Challenges in Election Polling

- 1 Polls need to identify who will vote in an election that hasn't happened yet.
- 2 People can change their mind before election day.
- 3 Small polling errors can imply incorrect predictions due to electoral rules.

Identifying Voters

Future Voters Are an Unknown (Unknowable?) Population

- ▶ In a typical survey we have a well-defined population and (ideally) a method to identify members of the population
- ▶ E.g. we could – in theory – enumerate every registered voter or every business or every newspaper reporter or
- ▶ But the population of an election survey is the set of people who *will vote* in an election that hasn't yet happened!
- ▶ The task for election polls is both to *predict* who will vote and who they'll vote for

Identifying the Population of “Likely Voters”

- ▶ Pollsters try to identify “likely voters” through questions, demographics, past vote data, etc.

Identifying the Population of “Likely Voters”

- ▶ Pollsters try to identify “likely voters” through questions, demographics, past vote data, etc.
- ▶ Likely voter screens are a form of weighting: try to weight the sample to reflect the future electorate

Identifying the Population of “Likely Voters”

- ▶ Pollsters try to identify “likely voters” through questions, demographics, past vote data, etc.
- ▶ Likely voter screens are a form of weighting: try to weight the sample to reflect the future electorate
- ▶ Two methods:

Identifying the Population of “Likely Voters”

- ▶ Pollsters try to identify “likely voters” through questions, demographics, past vote data, etc.
- ▶ Likely voter screens are a form of weighting: try to weight the sample to reflect the future electorate
- ▶ Two methods:
 - 1 **Deterministic models** that assign every respondent a 0% or 100% chance of turning out

Identifying the Population of “Likely Voters”

- ▶ Pollsters try to identify “likely voters” through questions, demographics, past vote data, etc.
- ▶ Likely voter screens are a form of weighting: try to weight the sample to reflect the future electorate
- ▶ Two methods:
 - 1 **Deterministic models** that assign every respondent a 0% or 100% chance of turning out
 - 2 **Probabilistic models** that estimate the turnout probability for each voter

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?
- ▶ Deterministic or probabilistic likely voter screen?

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?
- ▶ Deterministic or probabilistic likely voter screen?
- ▶ If deterministic: what cutoff to use? How to categorize people as voters or nonvoters?

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?
- ▶ Deterministic or probabilistic likely voter screen?
- ▶ If deterministic: what cutoff to use? How to categorize people as voters or nonvoters?
- ▶ If probabilistic: What modeling strategy? What past data to use to estimate model?

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?
- ▶ Deterministic or probabilistic likely voter screen?
- ▶ If deterministic: what cutoff to use? How to categorize people as voters or nonvoters?
- ▶ If probabilistic: What modeling strategy? What past data to use to estimate model?
- ▶ etc. etc. etc.

Likely Voter Models: Lots of Decisions to Make

Lots of decisions for researchers to make!

- ▶ Which survey questions / variables to include?
- ▶ Deterministic or probabilistic likely voter screen?
- ▶ If deterministic: what cutoff to use? How to categorize people as voters or nonvoters?
- ▶ If probabilistic: What modeling strategy? What past data to use to estimate model?
- ▶ etc. etc. etc.

These are important choices that affect predictions, and there are many defensible choices.

The Perry-Gallup Likely Voter Index

Set of questions gauging likelihood of voting, developed in the 1960s and used through today.

The Perry-Gallup Likely Voter Index

Set of questions gauging likelihood of voting, developed in the 1960s and used through today.

- How much thought have you given to the coming November election? ***Quite a lot, some, only a little, none***
- Have you ever voted in your precinct or election district? ***Yes, no***
- Would you say you follow what's going on in government and public affairs ***most of the time, some of the time, only now and then, hardly at all?***
- How often would you say you vote? ***Always, nearly always, part of the time, seldom***
- How likely are you to vote in the general election this November? ***Definitely will vote, probably will vote, probably will not vote, definitely will not vote***
- In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote? ***Yes, voted; no***
- Please rate your chance of voting in November on a scale of 10 to 1. ***0-8, 9, 10***

Perry-Gallup Index: Deterministic Method

- ▶ Respondents get a turnout score of 0-7
- ▶ Survey firm then comes up with an estimate of overall turnout in the election, based on historical turnout rates, competitiveness, etc.
- ▶ Select the cutoff (from 0-7) that mimics that turnout rate
- ▶ Assume that all respondents above the cutoff will vote, and none below cutoff

Perry-Gallup Index: Deterministic Method

Advantages?

Perry-Gallup Index: Deterministic Method

Advantages?

- ▶ Easy to calculate, simple to interpret
- ▶ Only need survey data and an (external) turnout estimate

Perry-Gallup Index: Deterministic Method

Advantages?

- ▶ Easy to calculate, simple to interpret
- ▶ Only need survey data and an (external) turnout estimate

Disadvantages?

Perry-Gallup Index: Deterministic Method

Advantages?

- ▶ Easy to calculate, simple to interpret
- ▶ Only need survey data and an (external) turnout estimate

Disadvantages?

- ▶ Deterministic method throws away information: some people below cutoff will vote, and their choices may differ from those above cutoff
- ▶ Scale method treats all questions equally, but some questions may be more useful than others
- ▶ Different groups of voters may answer questions differently, even if same likelihood of voting

Performance of Deterministic Perry-Gallup Index, 2014 Elections

The Perry-Gallup index

Score on scale	Share of registered voters	Share of all verified voters	% who are verified voters in each group
7	48%	63%	83%
6	15	15	63
5	10	10	59
4	7	4	22% of verified voters scored between 0 and 5
3	6	4	41
2	6	2	23
1	3	1	13
0	4	1	11
	100	100	

83% of
those who
scored a
7 actually
voted

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

PEW RESEARCH CENTER

Probabilistic Methods

Probabilistic methods combine survey data with validated turnout data from past elections

- ▶ Validated turnout data: match survey respondents to public voting records to see whether they *actually* voted

Probabilistic Methods

Probabilistic methods combine survey data with validated turnout data from past elections

- ▶ Validated turnout data: match survey respondents to public voting records to see whether they *actually* voted

Intuition: Fit a regression model predicting turnout as a function of survey responses (+ maybe demographics). Then, use the model to predict probability of turning out for current sample.

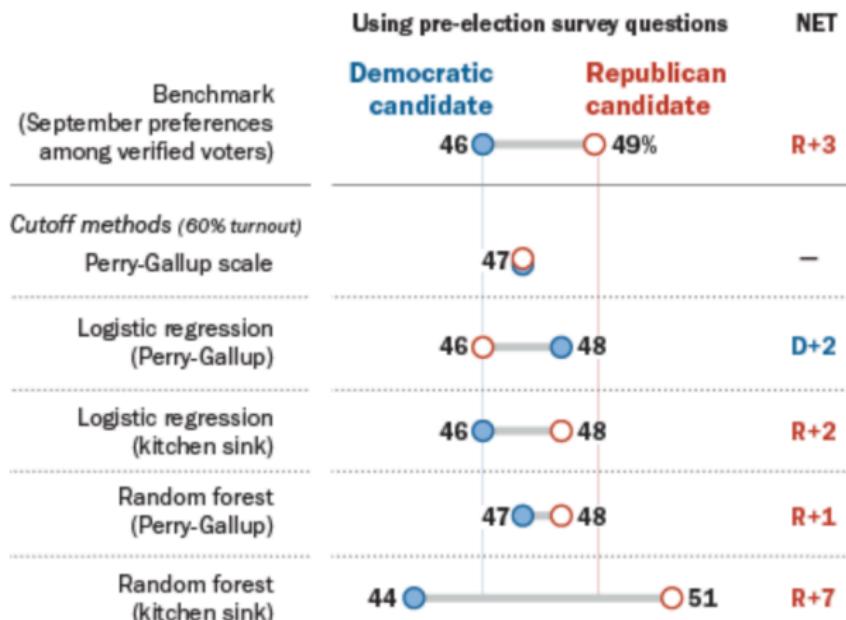
In analysis, there are then two choices:

- 1 impose a cutoff (e.g. $p_i(\text{turnout}) > 0.5 \rightsquigarrow \text{voter}$) and only analyze likely voters
- 2 weight analysis based on $p_i(\text{turnout})$

Pew Comparison of Likely Voter Screens in 2014

Estimates using cutoff methods

September pre-election voter preferences

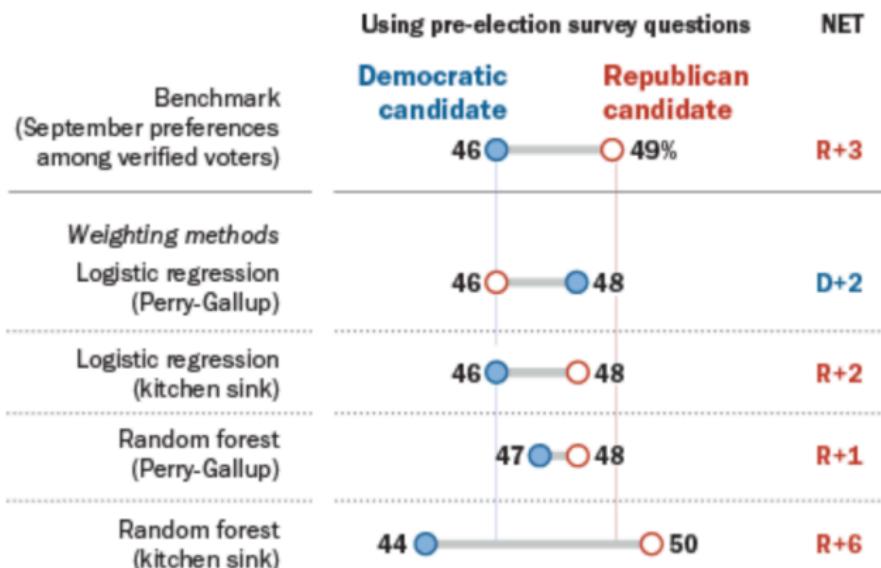


Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

Pew Comparison of Likely Voter Screens in 2014

Estimates using weighting methods

September pre-election voter preferences



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

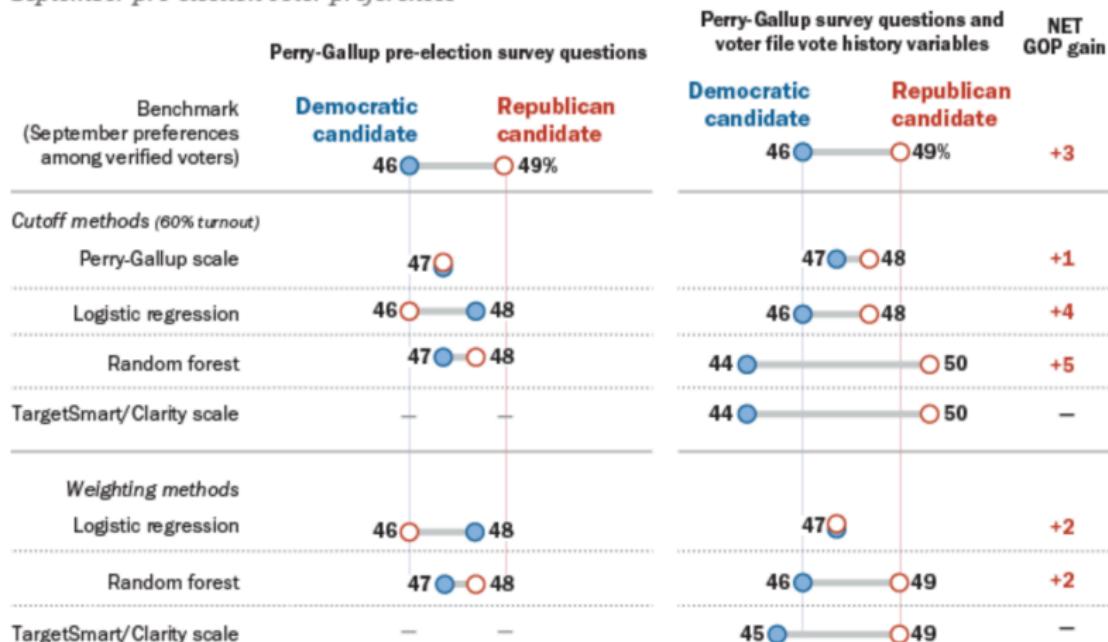
Adding Vote History

- ▶ Might improve turnout predictions if we include (self-reported) past turnout
- ▶ E.g. primary voters more likely to be general election voters

Pew Comparison of Likely Voter Screens in 2014

Voter file vote history improves most estimates

September pre-election voter preferences



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

Other Considerations

- ▶ There are many assumptions in any method:
 - ▶ What cutoff to use \rightsquigarrow assumption about turnout rate
 - ▶ Probabilistic model \rightsquigarrow assumption that correlations between turnout and survey responses are stable across elections
- ▶ Likely voter screens affect sociodemographic composition of predicted electorate

Changing Vote Choice

When Do People Decide Who to Vote For?

- ▶ Pre-election polls really capture a snapshot before the election
- ▶ People may change their mind (or make up their mind) between the time of the poll and the election
- ▶ As always, whether this induces error depends on whether what's captured in the survey represents the population of interest
- ▶ In this case, do late-deciders differ in their vote choice from early-deciders, on average?

Did Late-Deciding Voters Cause 2016 Polling Error?

Time of Decision and Presidential Vote in Key States Won by Trump

State	Decided Final Week	Vote Choice, Late Deciders		Vote Choice, Early Deciders		Trump Gain
		Trump	Clinton	Trump	Clinton	
Florida	11%	55%	38%	48%	49%	2.0%
Michigan	13%	50%	39%	48%	48%	1.4%
Pennsylvania	15%	54%	37%	50%	48%	2.3%
Wisconsin	14%	59%	30%	47%	49%	4.3%
National	13%	45%	42%	46%	49%	0.8%

Analysis from Aaron Blake (2016) using NEP exit poll data, cited in AAPOR report.

Late-Breaking Events in the 2016 Election



Late Campaign Events Increase News Consumption

- ▶ Tyler (2022) uses web-tracking data to study political news consumption in run-up to 2016 election
- ▶ Undecided voters increased their news consumption in the days after Comey letter was released:

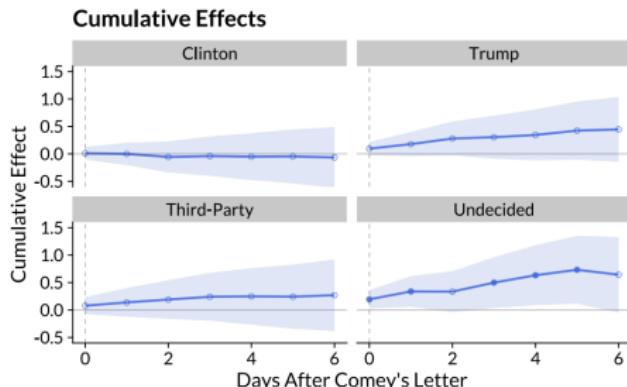


Figure 4: Estimated cumulative effects of Comey's letter (Oct. 28) on the daily share voters in each subgroup who visited a political news page (original outcome between 0 and 1).

Electoral Rules Magnify Small Polling Errors

Plurality Elections

- ▶ In the best-case scenario, polls provide an unbiased estimate of the share of voters who support candidate A over candidate B
- ▶ But electoral rules are discontinuous: A wins if they get even 1 more vote than B
- ▶ Polls are not precise enough to know with certainty who will win in a very tight race
- ▶ For presidential elections, this issue is magnified due to Electoral College

 **TheUpshot**

Alaska Ariz. Colo. Fla. Ga. Ind. Iowa Miss. Mo. N.C. N.H. Nev. Ohio Pa. Wis.

Who Will Be President?

By JOSH KATZ

Estimates on this page are based on pre-election polls. For an estimate including results, The Times is providing live forecasts on election night.

President Senate House

Hillary Clinton has an 85% chance to win.

Last updated Tuesday, November 8 at 10:20 PM ET

CHANCE OF WINNING



85%

Hillary Clinton



15%

Donald J. Trump

ELECTION2016

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

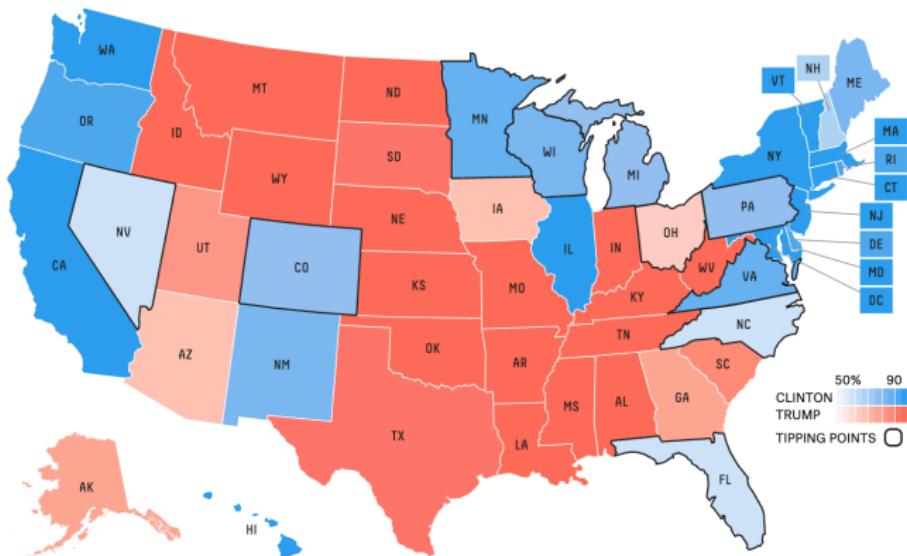


Photos: Getty

Who will win the presidency?



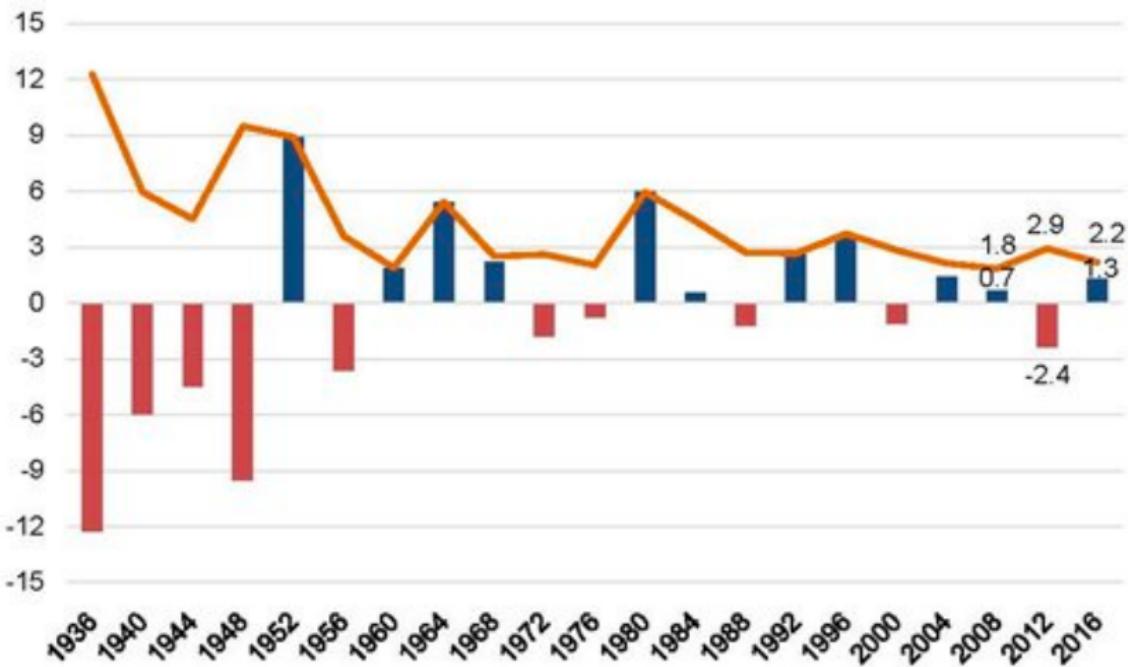
Chance of winning



National-Level Polls in 2016 Were Accurate

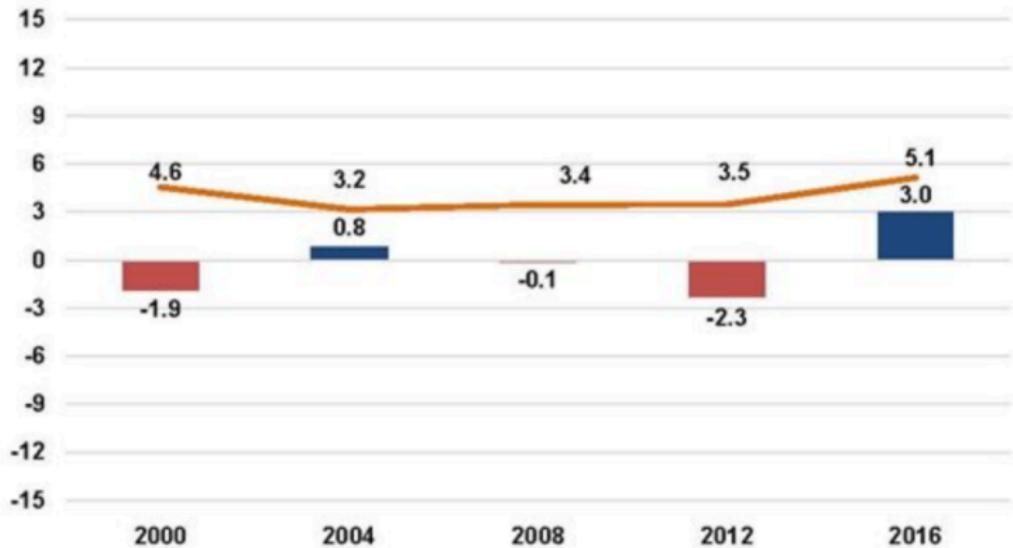
Orange line represents average absolute error

Bars represent average signed error (red bars indicate overestimation of Republican vote margin; blue bars indicate overestimation of Democratic vote margin)



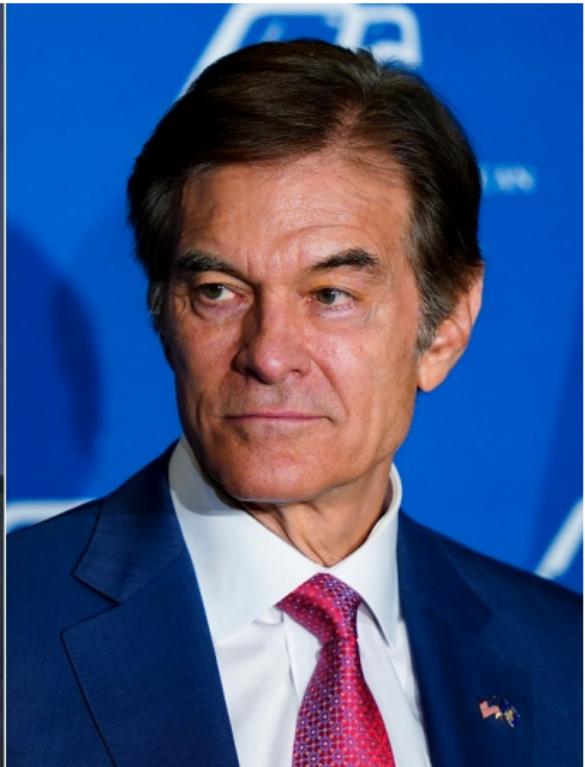
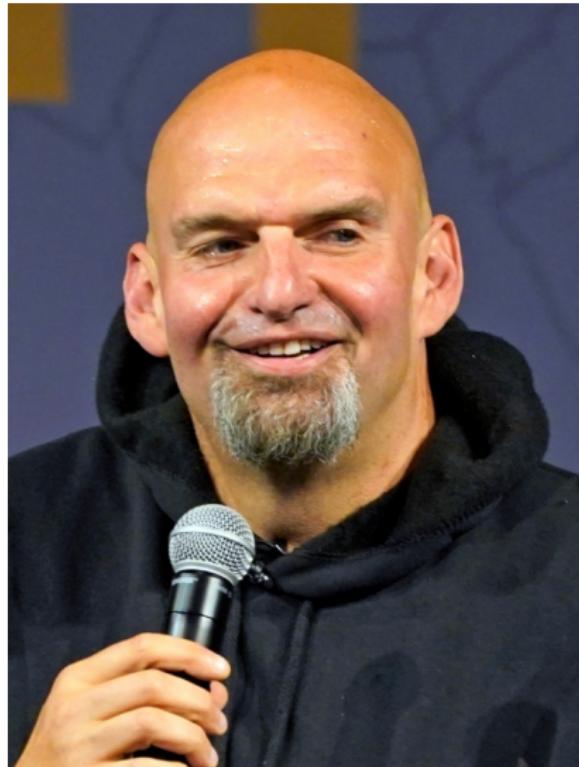
State-Level Polls in 2016 Were Less Accurate

Orange line represents average absolute error
Bars represent average signed error (red bars indicate overestimation of Republican vote margin; blue bars indicate overestimation of Democratic vote margin)



Poll Averaging and Forecasting Models

2022 PA Senate Race



PA Senate Race: One Poll A Month Before the Election

[ABOUT US](#)[ANALYTICS](#)[RACES](#)[THE ALTERNATE ROUTE](#)[Donate](#)

POLL: Fetterman Leads Oz 55% to 36% Among Likely Voters in Pennsylvania's Senate Race

OCTOBER 5, 2022

How Accurate Do You Think This Poll Was?

Ignoring the results, what information would you use to diagnose accuracy?

How Accurate Do You Think This Poll Was?

Ignoring the results, what information would you use to diagnose accuracy?

- ▶ What is the sample size?
- ▶ How was the survey conducted?
- ▶ What weighting targets are used?
- ▶ Who sponsored and conducted the survey?
- ▶ How recently was the survey conducted?

Center Street Poll: Methodology

Methodology: Between Sept. 29-30, 2022, 1,169 adults 18 and older were surveyed, including 971 registered voters and 568 likely voters. Results are now weighted to age/gender of U.S. Census and then by 2020 Presidential vote preference. All results will have 2020 Stated Voter Preference aligned with Actual (50% Biden/49% Trump). Source: Center Street PAC via Momentiv AI.

What Do Other Polls Say?

DATES	SAMPLE	POLLSTER	SPONSOR	RESULT			NET RESULT	
U.S. Senate, Pennsylvania, 2022 ▾ AVG.								
Sep. 29-30	971 RV	◆ Center Street PAC		Fetterman	50%	34%	Oz	Fetterman +16
Sep. 29-30	568 LV	◆ Center Street PAC		Fetterman	55%	36%	Oz	Fetterman +19
Sep. 27-30	500 LV	B+ Suffolk University	USA Today	Fetterman	46%	40%	Oz	More ▼ Fetterman +6
Sep. 23-26	1,000 LV	A- Emerson College	WHTM-TV (Harrisburg, Pa.), The Hill	Fetterman	45%	43%	Oz	Fetterman +2
Sep. 22-26	1,008 RV	A Beacon Research/Shaw & Company Research	Fox News	Fetterman	45%	41%	Oz	More ▼ Fetterman +4
Sep. 19-25	517 RV	B/C Franklin & Marshall College		Fetterman	45%	42%	Oz	Fetterman +3
Sep. 23-24	500 LV	B InsiderAdvantage	WTXF-TV (Philadelphia)	Fetterman	45%	42%	Oz	More ▼ Fetterman +3
Sep. 19-22	1,242 RV	A Marist College		Fetterman	51%	41%	Oz	Fetterman +10
Sep. 16-19	759 RV	Phillips Academy	Abbot Academy Fund	Fetterman	47%	45%	Oz	Fetterman +2
Sep. 13-16	420 LV	B+ Muhlenberg College	The Morning Call	Fetterman	49%	44%	Oz	Fetterman +5
Sep. 13-15	1,078 LV	A- ◆ Trafalgar Group		Fetterman	48%	46%	Oz	More ▼ Fetterman +2
Sep. 6-12	1,188 LV	B+ YouGov	CBS News	Fetterman	52%	47%	Oz	Fetterman +5
Aug. 31-Sep. 7	828 LV	B/C Echelon Insights	NetChoice	Fetterman	57%	36%	Oz	Fetterman +21
Aug. 31-Sep. 3	679 LV	B/C RABA Research		Fetterman	49%	40%	Oz	Fetterman +9

Looking for Outliers

- ▶ All surveys have error, but the sources of error across surveys might be different
- ▶ If one survey has very different results from all the others, seems like there might be something strange about it

Looking for Outliers

- ▶ All surveys have error, but the sources of error across surveys might be different
- ▶ If one survey has very different results from all the others, seems like there might be something strange about it ↗ why?

Looking for Outliers

- ▶ All surveys have error, but the sources of error across surveys might be different
- ▶ If one survey has very different results from all the others, seems like there might be something strange about it ↗ why?
- ▶ Intuition: each poll some error, but similarities across polls should give us more confidence in the results

Poll Averaging

- ▶ Rather than looking at individual surveys, why not combine them together systematically?

Poll Averaging

- ▶ Rather than looking at individual surveys, why not combine them together systematically?
- ▶ Each survey has error, but when averaged together these errors *might* cancel out
- ▶ Simplest case: only sampling error (no bias)
 - Then, combining polls together is like getting a bigger sample size
- ▶ Realistic case:
 - Some errors are idiosyncratic (affect each survey individually), some are systematic (affect all surveys)

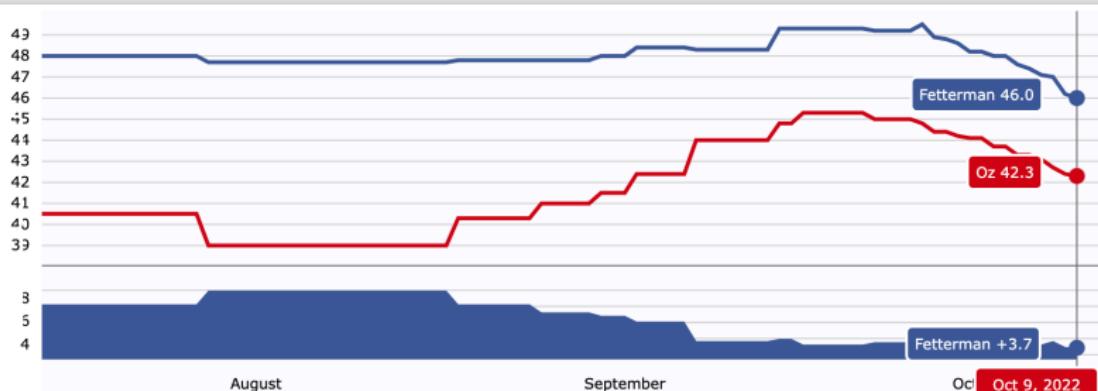


RCP POLL AVERAGE

Pennsylvania Senate - Oz vs. Fetterman

46.0 Fetterman (D) +3.7

42.3 Oz (R)



From: to: Apply

14D 30D MAX Reset

REALCLEARPOLITICS

Are All Polls Equally Informative?

- ▶ Should we count every poll equally?
- ▶ What if I put out a poll of 100 people?
- ▶ What if a poll comes from a campaign or a partisan pollster?

FiveThirtyEight's Adjustments

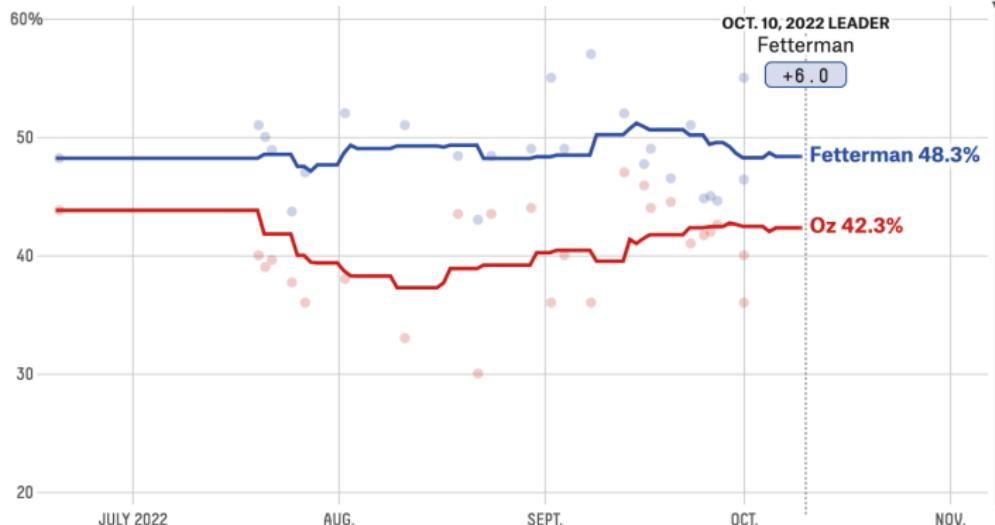
- ▶ Weight polls by recency and sample size
- ▶ Estimate “house effects” —→ does this pollster consistently over/underestimate Democratic vote share?
- ▶ Data-driven likely voter adjustment to make RV and LV polls comparable

FiveThirtyEight Polling Average in PA Senate Race

Who's ahead in the Pennsylvania Senate race?

An updating average, accounting for each poll's quality, recency, sample size and partisan lean.

ELECTION DAY
NOV. 8, 2022



From Polling Average to Election Forecast

What other factors might be useful in predicting which candidate will win?

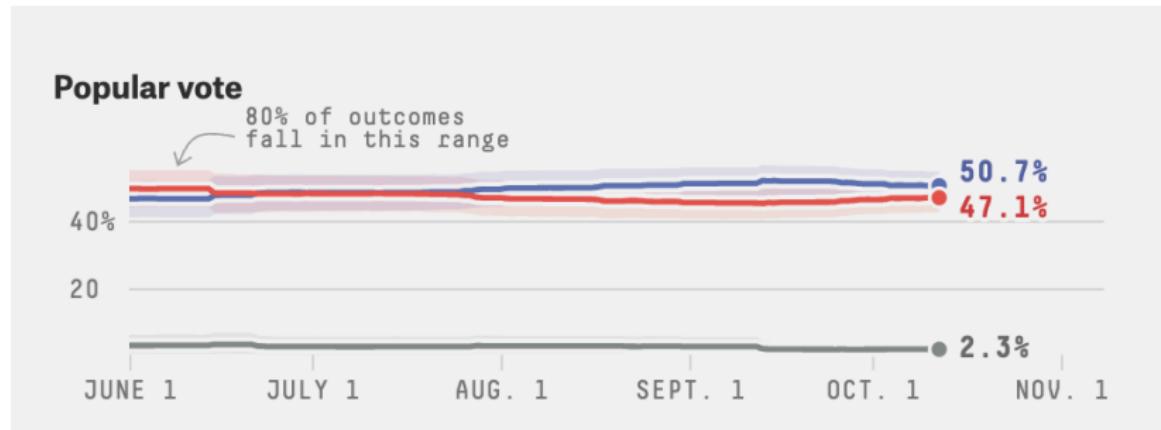
From Polling Average to Election Forecast

What other factors might be useful in predicting which candidate will win?

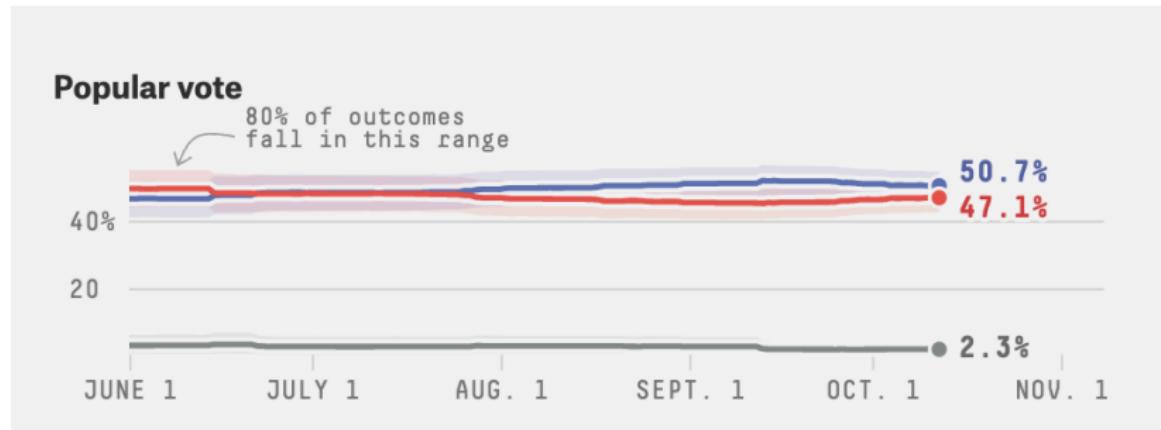
The layers in FiveThirtyEight's House forecast

LAYER	DESCRIPTION	WHICH VERSIONS USE IT?		
		LITE	CLASSIC	DELUXE
1a Polling	District-by-district polling, adjusted for house effects and other factors.	✓	✓	✓
1b CANTOR	A system which infers results for districts with little or no polling from comparable districts that do have polling.	✓	✓	✓
2 Fundamentals	Non-polling factors such as fundraising and past election results that historically help in predicting congressional races.	✓	✓	
3 Expert forecasts	Ratings of each race published by the Cook Political Report, Inside Elections and Sabato's Crystal Ball		✓	

Predicted Vote Share in PA Senate Race (FiveThirtyEight “Deluxe”)



Predicted Vote Share in PA Senate Race (FiveThirtyEight “Deluxe”)



Summing Up

- ▶ Election polling is an important but fraught application of surveys
- ▶ There are lots of decisions to make, beyond the usual survey design considerations
- ▶ Consider: what is the value of horserace polling?
- ▶ Next time: discussion of *Strength in Numbers* and political polling more generally