Brexit, Trump, and Polling: A Form of Statistical Inference

November 16, 2016

Prof. Michael Paul

Prof. William Aspray

Statistics Powerful but can be misleading.

- Ex - newspaper headline:
- People who take short breaks at work are far more likely to die of cancer.
- Sample includes 36,000 workers who leave office to take 10-minute breaks
- Large n!!!
- Finding: workers who take short breaks are 41% more likely to develop cancer over next 5 years than those workers who don’t take short breaks
- How to explain this?

EX 1 cont.

- Leave to go outside to smoke.
- Smoking, not the breaks are correlated to the cancer

- Wheelan, *Naked Statistics*

Ode to the Central Limit Theorem

- We are going to review the CLT in detail next class, but polling is completely dependent on it
- CLT as applied here states:
  - If we take a large representative sample, our sample will look a lot like the population from which it is drawn.
- When polls fail, likely reasons:
  - The sample is too small
  - The sample is not representative (sampling bias, low response rate)
  - Questions phrased inappropriately (don't ask right question; ask confusing questions; bias the answers)
  - Respondents are lying (stigma to admit you don't vote)
  - The polling is so close that the margin of error is significant
  - The timing of the polling is inappropriate (if the population is varying by time)
  - Mistaken notions about the characteristics of the population (e.g. voter turnout)

Brexit and 2016 Presidential Polls

- Polls overwhelmingly reported sentiment for Britain to stay in the EU, but vote 52% - 48% in favor of exit
- Polls report comfortable lead for Clinton over Trump in last month of campaign?
- YES Bloomberg Politics, CBS News, Fox News, Reuters/Ipsos, USA TODAY/Suffolk, Quinnipiac, Monmouth, Economist/YouGov and NBC News/SM, according to RealClearPolitics.
- NO: L.A. Times/USC
- Why?

Why Brexit polls wrong

- Brexit
  - speculate embarrassing to vote for exit, so people would not reveal their voting intentions
  - Telephone polls less accurate than Internet polls
  - Polls undercount people who are hard to reach, especially if the poll has to be concluded quickly and can't try multiple time to reach people
  - Polls overcounted the educated and undercounted the uneducated
  - Turnout models were wrong – who actually turned out to vote
  - Reallocation models for "don't know" responses to leave or remain vote actually made the statistical models less accurate of the real vote (used race and immigrant status; did not use attitudinal similarities)

  [http://www.businessinsider.com/pollsters-know-why-they-were-wrong-about-brexit-2016-7]

Why Presidential Polls Wrong - 1

- Popular vote models were somewhat accurate – Clinton did win the popular vote, not the electoral vote
- State and local polls more recent, more amateurish – they are important to deciding electoral votes
- Much harder to find accurate and random samples of voters since people began using cell phones widely – when could call home lines, less chance of nonresponse bias and know demographic info about the homeowner – and no directories available for cell phone numbers to simplify random selection
- SurveyMonkey Poll had biased sample – people willing to answer SurveyMonkey polls turned out to be more likely to be educated and support Clinton

[http://www.theatlantic.com/politics/archive/2016/11/what-went-wrong-polling-clinton-trump/507188/]

Why Presidential Polls Wrong - 2

- Polls were off by 2 to 3% in typical standard error, but this was enough to make a difference in swing states
- Bad estimates of turnout to vote – voter enthusiasm higher for Trump, which may mean differential turnout
- Trump supporters not admitting their support to pollsters (stigma) – probably only a minor effect
- Third-party collapse: polls running 5% for Libertarian candidate Gary Johnson, but he actually only received 3% of vote and most of these votes went to Trump (people decided on election day they wanted their vote to count)
- Shock related not to the conditional probability of Trump winning given that he had 48% of the two-party support in the polls, but instead to unconditional probability of Trump becoming president given the state of politics two years ago

[http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/]

Use of analytics in Clinton campaign

- Used a custom-built psephological algorithm named Ada to run campaign – more than other presidential campaign in history
- Named after Augusta Ada
- Kept highly secret – on separate server with very little access
- Ran 400,000 simulations every day based on various public and private polls and on voter registration data
- Informed decisions about
  - When and where to send Clinton to speak and her surrogates
  - Where to open campaign offices
  - Where to send Beyonce and Jay Z to give concerts
- Worked well in PA; poorly in MI, WI

[https://www.washingtonpost.com/news/post-politics/wp/2016/11/09/clintons-data-driven-campaign-relied-heavily-on-an-algorithm-named-ada-what-didnt-she-see/#comments]