MATH 3215 Project- Election Polling

Reese Wang

November 28, 2023

1 Public Opinion Polling

Election polling is a type of public opinion polling. It is used to understand what voters are thinking and why they're planning on voting for their choice.

Public opinion was defined by the great political scientist V.O. Key Jr. as "those opinions held by private persons which governments find it prudent to heed."

This can be much simply stated as "what the public wants", or in the context of election polling, election polling can be "who the public wants in its government."

There are other public opinion polls with non-political questions in mind, but I will focus on election polling.

1.1 Polling basics

A **public opinion poll** aims to give everyone in the population an equal voice in current issues. The voice, or public opinion, is derived from asking the population list of curated questions.

One way to ensure that everyone in the population has an equal voice in the poll is to ask the entire population poll questions. However, since the population can be large, pollsters take a **sample** of the population and only ask them questions.

I'll explain how to construct a good public opinion poll in the poll composition section.

1.2 Benefits of public opinion polling

What other ways are there to "measure" public opinion, especially in the context of elections?

1. The election.

   Pros: Votes in elections are weighed equally: everyone who participates has their opinion weighed equally.

   Cons: Election outcomes don't give a lot of information. Election questions tend to be binary (Ex.: Candidate A or Candidate B? For or against

a ballot measure, proposition, or referendum?). Often, pollsters are interested in the nuances behind a candidate or issue (Ex.: Why vote for Candidate A? What is your opinion on Candidate A's policy for policy X? What changes to a referendum would need to be made for you to support it?). I will explore poll question construction in section 2.

2. Public forum

   A **public forum** is a property that is open to public expression and assembly, according to US constitutional law.

   Some examples of public forums include the opinion sections of newspapers, public protests, and online petitions.

   Pros: Public forums are places where citizens can express their thoughts freely, which contains the nuance not found in the yes/no answers of an election.

   Public forums are visible. So we can see the opinions of others

   Cons: Public forums do not represent the population. Citizens who have time to participate skew more educated, affluent, and older. The election will also have voters who participate less in public forum, for a variety of reasons.

2 Poll Composition

A poll is composed of:

1. A set of questions to ask
2. A set of people to interview
3. Mode of interview

2.1 Questionnaire

Poll questions should be answerable and neutral.

There are two general types of questions in a questionnaire:

1. Close-ended questions These questions are either yes/no or multiple choice. Respondents' responses to these questions is easily quantifiable, which makes this type of question preferred for gathering quantitative data.
2. Open-ended questions Questions are open-ended when the pollster does not come up with possible answers beforehand. Each member of the sample responds with their opinion to the question. Open-ended questions create qualitative data, which is harder to analyze.

2.2 Sample

A **sample** is the set of people pollsters interview for their poll, to represent the population. In this section, I explore some techniques used to get a good, representative sample for election polling.

2.2.1 Random Sampling Techniques

Random Sampling is a technique used to achieve the goal of giving everyone in the population an equal voice in current issues. It is a type of **probability sampling**, where the probability of each person being selected to the sample is known.

For political polls, random samples are achieved through **random digit dialing** (RDD) and **registration-based sampling** (RBS).

2.2.2 RDD

RDD is a very popular sampling method for research, known for creating high-quality sample frames. Its **sample frame**, or specific source of respondents, is everyone with access to a phone in the population, which is a large proportion of the voting population.

American phone numbers have 10 numbers: a 3-digit area code, 3-digit exchange number, and a 4-digit number. When creating the sample, pollsters list all working area codes and working exchanges within those area codes.

Since some phone numbers are either invalid or business phone numbers, pollsters must determine whether numbers are likely owned by individuals, without calling every single number. Since phone numbers are typically given out in 100s, one method to determine whether phone numbers are owned by individuals is to separate phone numbers into groups by 100s and randomly choose two per group to call. If both are not owned by individuals, then the group is discarded, as it is likely that most phone numbers in that group are also not owned by individuals.

If the number called is owned by multiple individuals (like a home phone number), pollsters may employ **within household sample selection** to select the voter that they want to talk to. Talking to the person who answers the phone might not be random. For example, if a stay-at-home parent answers the phone during most working hours and the pollster chooses the voter who answers the phone, the poll's sample will include a disproportionate amount of voters who are stay-at-home parents, which is not representative of the voting population. Some techniques employed in within household sample selection include taking a list of all voters in the household and randomly selecting one (the Kish Selection Method), or asking for the individual with the most recent or next birthday (Last-Birthday selection method).

Although RDD has been historically great, its effectiveness has declined as Americans shift away from the telephone and towards the cell phone. Americans are less likely to pick up unsolicited phone calls and more likely to refuse to answer questions, leading to a decrease in **response rate**, or the percentage

of individuals who actually respond to the interview questions. In a study by Pew Research Center in 2016, the response rate for telephone polls is around 9%. Since this is a recent phenomenon, limited research shows that even with the decrease in response rate, there has not been a significant difference in **nonresponse bias**.

2.2.3 RBS

The sample frame for RBS is a list of registered voters. The sample is selected from this list, or voter file. It is advantageous because the phone numbers called are working and used phone numbers, but is disadvantageous for the small amount of people not included in the voter file (people who have recently moved or registered to vote and are not yet on the registered list).

Voter files are used for election surveys at state and local level, but not national level polls. A study from Pew Research Center shows that RBS and RDD yield similar samples, and RBS could be used more on the national level in the future.

2.2.4 Overview of non probability sampling

As discussed in the RDD section, probability sampling is expensive and time-consuming, with a decrease in response rate. There is no method like RDD adapted for modern technology that has the same level of effectiveness in giving everyone an equal chance of being in the sample. For example, surveys on the internet, which most voters are on, is an example of a self-selection or non probability sampling.

However, there is increasing research to see if there are any non probabilistic sampling methods that could approximate or return similar results to a probabilistic sampling method. Probabilistic samples with low response rates use modeling, ranking adjustments, and propensity models to make up for low response rate, all of which are techniques similar to those used in non probabilistic sampling.

I found a couple research papers from the past 10 years on statistical inference with non probability surveys, discussing several of the techniques I have listed above. Many of them are above my understanding level, and I plan to take more time later to read more about them.

2.3 Mode of interview

Just as how technology has transformed sampling methods, the mode of interview has also evolved. Polls were initially conducted in-person, with high engagement. RDD sampling is collected by phone, beginning in the 1980s. Since the internet is popular, poll interviews are also connected online.

What's important to think about with mode of interview is how to get high quality responses. Response rates to polls are low, at around 10%, and some modes of polling create more engagement with people than others.

3 Poll accuracy

No matter how the poll is conducted, people judge polls by their accuracy, or how similar the predicted results are compared to the actual election. Here, I will look at a common parameter to measure poll accuracy, margin of error, and explain how to interpret it.

3.1 Margin of Error

The margin of error quantifies how "off" the poll expects its result to be, compared to the actual outcome. The margin is from sampling, because depending on what specific sample the pollster takes from the sample frame, the result of the poll could be slightly different.

Margin of error can be calculated as follows:

Margin of error$= z \cdot \frac{\sigma}{\sqrt{n}}$,

where z is the z-value of a normal distribution associated with the confidence level, σ is the standard deviation calculated from the sample, and n is the size of the sample.

A margin of error appears in most national election polls. When reading a national election poll, a poll may report that candidate A has 37% support, and candidate B has 44% support, with a margin of error of 5%. This can be interpreted to mean that in the true population of all voters, the true level of support for candidate A is between 32% and 42%, and the level of support for candidate B is 39% to 49%.

The margin of error should also be applied to compare support between two candidates to see who is ahead in the race. To calculate the difference in level of support, take the difference of the support levels and see if the difference is within the **doubled** margin of error. We look at the doubled margin of error because there are two candidates.

In the example above, candidate B is 7% higher, but with the doubled margin of error, candidate B may be behind by 3% or ahead by 10%. Since the margin includes 0, we can conclude that B's lead may be due to sampling, and we cannot conclude that B has more support than A.