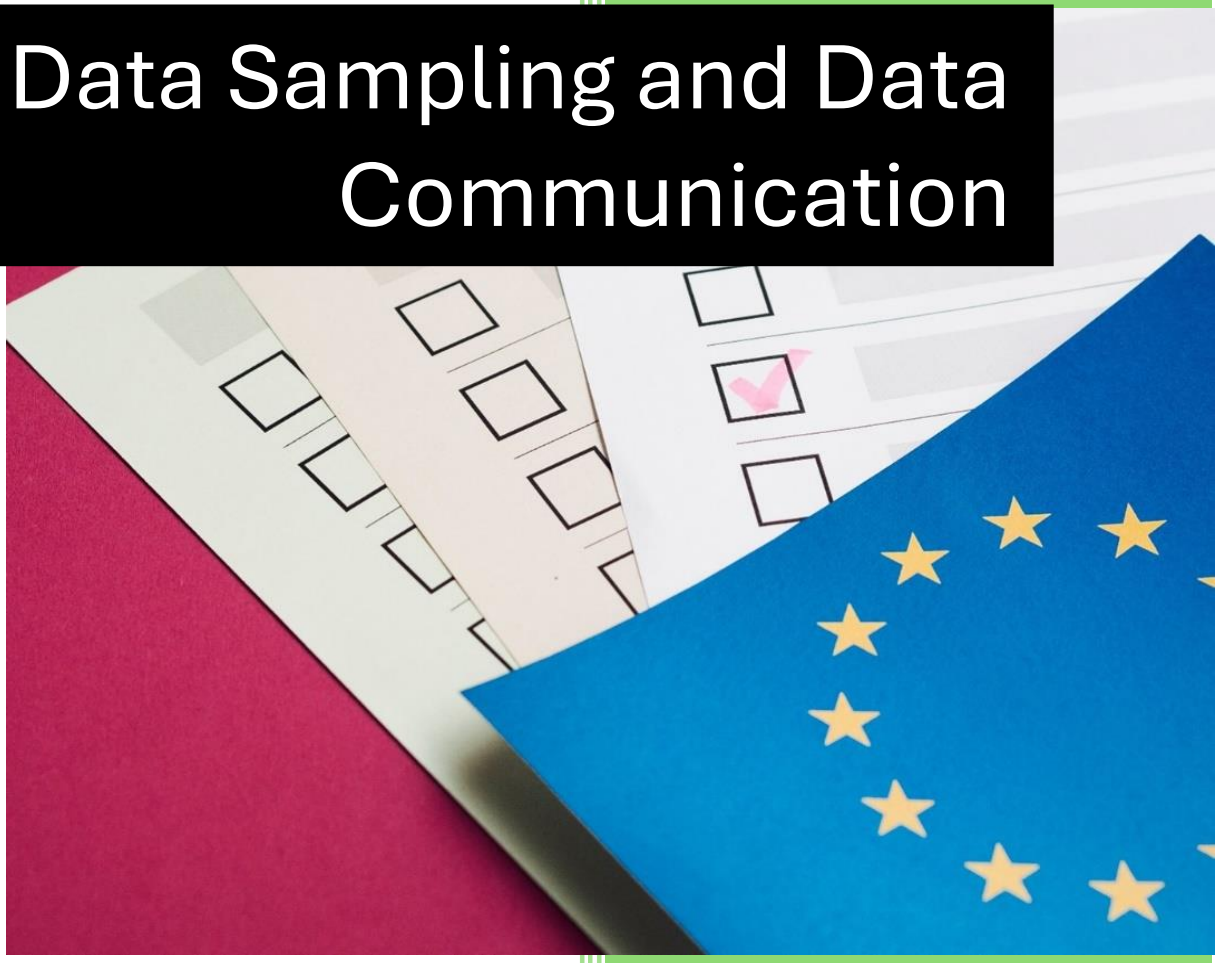
Project

M1 Econometrics, Statistics, Economics & Magistère 2

Data manager

2024

Data Sampling and Data Communication



Written by:

ADJENIA Danélius Dègnon

CHERIF Lamine

SODJINOU Setondji Hippolyte

(Team: Data Manager)

Supervisor:

Quentin LIPPMANN

30/12/2024

amse
école d'économie d'aix-marseille
aix-marseille school of economics

amU
Aix Marseille Université

Table of contents

Introduction

Accurately predicting election outcomes through polling is a critical but complex task. This report explores the challenges and solutions associated with data sampling and communication in electoral polls. By analyzing both ideal and practical polling methods, the report highlights key biases, such as non-response, undercoverage, and measurement errors, that can compromise the reliability of poll results. Through the examination of data from presidential elections in France, Brazil, South Korea, and the United States, this study aims to identify the gaps between poll predictions and actual outcomes while proposing actionable strategies to improve polling accuracy.

Understanding election polls

Ideal conditions to run an election poll

Electoral polls, aimed at predicting election results, can be conducted in various contexts. To ensure the effectiveness of such a survey, we will opt for probability sampling method. This approach, based on random selections, allows for reliable conclusions about the overall population. To achieve a more precise estimation, we will prioritize stratified random sampling.

Indeed, we consider a person registered on the electoral rolls as the unit of observation. The entire group of registered voters, that is, the individuals eligible to vote, constitutes the target population, whether for a specific region or country.

For the sample to be representative of this target population, it must reflect key demographic characteristics such as age, gender, place of residence, level of education, and political preferences. In our study, a sample of 1,000 people will be drawn from a population of 10,000 individuals. The sampling frame used will likely be the list of sampling units, including the names, identifiers, or email addresses of the selected individuals.

The polls will be conducted through various channels: online (secure survey platforms to reach younger generations and tech-savvy individuals), by phone (for elderly people and those with limited internet access), and in person (at public places, community events, or through home visits) to ensure the inclusion of populations that are less accessible by other means.

In summary, an ideal poll would use probability sampling to ensure all population groups are adequately represented. This could include stratified sampling based on income, geographic location, and other relevant demographics. Efforts to increase the response rate, such as follow-ups and incentives for participation, would mitigate nonresponse bias.

How are election polls run in practice?

Electoral surveys often use quota sampling, a non-probability method that allows for the study of populations without a sampling frame. The population is divided into subgroups based on criteria such as gender, age, socio-professional category, type of municipality, and region of residence. Quotas are then set for each subgroup. Researchers fill these quotas using non-random selection methods to obtain a representative sample as quickly as possible.

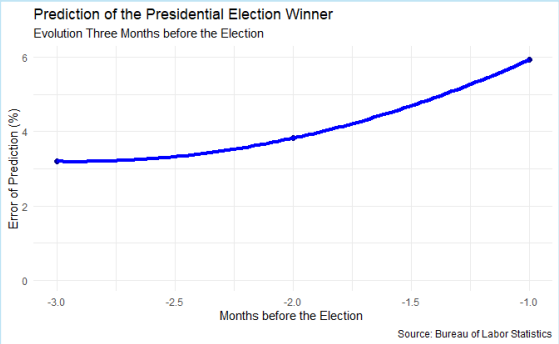
In these polls, the unit of observation consists of individuals registered on the electoral rolls, and the target population remains all registered voters in the region or country. The sample considered is made up of individuals who have completed their surveys, which does not necessarily correspond to those who will vote. In practice, the actual sample is often formed through quota sampling, where certain quotas (such as age, gender, region) are filled based on the availability of respondents, which may lead to a non-fully random selection. The sampling frame used is often less precise, based on phone directories, social media data, or lists of people interested in certain political causes, which may not reflect the

entire target population. These polls are primarily conducted online or by phone, especially in recent years.

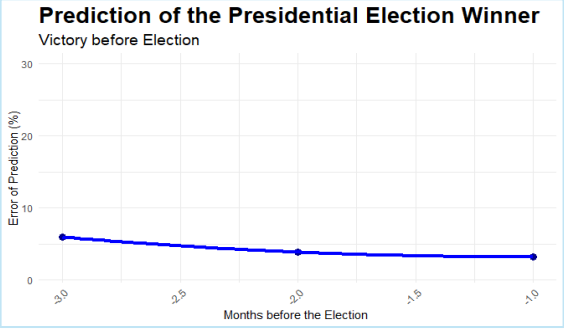
This sampling method introduces selection biases within each quota, making the sample less representative of the population. Additionally, sampling errors are difficult to estimate.

Data collection

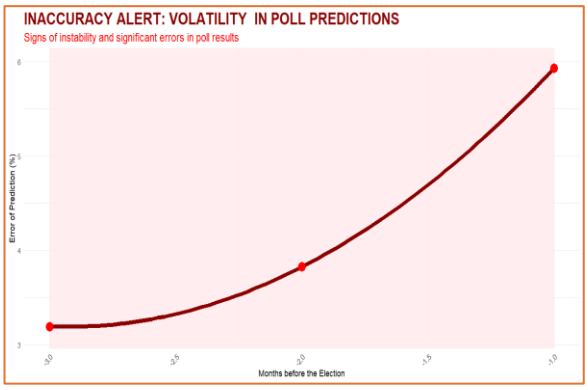
We will analyze the presidential elections, specifically the French presidential election of 2022, the Brazilian presidential election of 2022, the South Korean presidential election of 2022, and the U.S. presidential elections of 2008 and 2012. The data are collected over a month and three months prior to the elections. They include polling results from organizations before the elections on the predicted winner and the actual election results. In total, we have considered 150 observations, with which we will create the following three graphs.



The visualization shows the overall accuracy of polls in predicting election results. We notice an increase in errors as the elections approach. This indicates a bias toward predicting far-right outcomes.



The visualization that convinces polls accuracy.



The visualization that convinces the polls inaccuracy.

The data analysis reveals a gap between the poll predictions and the election results, which can be attributed to selection and measurement biases.

Identifying Problems & Solutions

Selection bias

Election polls often suffer from the underrepresentation of certain population groups. Among these, the most vulnerable groups, such as low-income individuals, the homeless or those living in precarious conditions, as well as the elderly, are the most affected. These groups lack access to online polls, which limits their participation. Additionally, young adults, particularly those aged 18 to 24, are often underrepresented, partly due to their low political engagement or tendency not to respond to surveys. Their mobility, with frequent address changes, also complicates their inclusion in samples based on voter registers.

On the other hand, certain groups are often overrepresented in polls. This includes retirees, politically engaged individuals, and those who regularly participate in elections. Furthermore, highly educated individuals and those living in urban areas, who have better access to technology and greater availability, are more likely to respond to

surveys. These individuals, often more politically active, find the survey process more interesting or rewarding. However, this overrepresentation can skew the results, making the conclusions less representative of the entire population. An example of this can be seen in the US polls of 2016 and 2020, where non-college-educated white voters, a large portion of Trump's base, were underestimated.

Another common bias observed in polls is the non-response bias. On average, nearly 31% of respondents do not answer. These non-respondents often have a lower education level (below high school), a modest income (less than €1500), are renters, and live in rural areas. Non-response bias is unavoidable which makes it the most problematic bias, but it distorts the accuracy of polls and can make the results unreliable. A historical example of this bias is the 1936 Literary Digest poll, which failed to correctly predict the results of the presidential elections, despite sending over 10 million ballots. Less than 25% of the individuals contacted responded, which led to an erroneous estimation of the election results.

Measurement bias

Measurement bias represent systematic errors in the collection or measurement of data. They can be caused by suggestive or poorly worded survey questions. For instance, during the 2008 U.S. presidential election, a question posed was: *"If the 2008 presidential election were being held today and the candidates were Barack Obama and Joe Biden, the Democrats, and John McCain and Sarah Palin, the Republicans, for whom would you vote? If already voted: For whom did you vote?"* This formulation pushes respondents to explicitly choose one of the two parties, limiting nuanced responses. A more neutral rewording could be: *"If the 2008 presidential election were held today, which candidate would you choose? Barack Obama and Joe Biden (Democratic Party), John McCain and Sarah Palin (Republican Party), I don't know / Prefer not to say."*

Additionally, social desirability biases arise when questions implicitly encourage respondents to avoid controversial answers. For example, a question posed during the 2022 Brazilian presidential election was: *"Which candidate do you believe is most capable of governing the country and promoting national unity?"* This wording leads respondents to choose a candidate perceived as more socially acceptable rather than the one they truly prefer. A more balanced and open formulation could enhance the quality of the responses obtained.

Regardless of the methods used, electoral polls often struggle to account for rare or unexpected events. These are frequently overlooked, which can bias the results. To address this, techniques such as stratified sampling with disproportional allocation, two-phase sampling, multiple frame surveys, or network/snowball sampling can be used, as well as increasing the sample size. Lastly, events occurring just before elections (whether economic, political, social, or international) can significantly influence public opinion and affect the quality of the data collected. This highlights the complexity and limitations of polls in dynamic contexts.

Far-right challenges

Predicting the voting shares of far-right parties presents several challenges, including social desirability bias, which may lead respondents to underreport their preferences. Additionally, traditional sampling methods often fail to capture rural or marginal voters, who are typically more inclined to support far-right parties. Biases related to media coverage and the stigmatization of far-right ideologies further complicate the analysis.

To mitigate these biases, indirect questioning techniques can be employed, such as the randomized response technique or the item count technique. In the latter, the control group receives a list of non-sensitive items, while the treatment group gets the same list with the sensitive item included. Respondents only report the total number of "yes" responses, thus preserving anonymity while collecting reliable data.

Moreover, question phrasing can be improved using forgiving wording that assume the behaviour being studied, or by framing questions within a specific temporal context. Finally, self-administration and audio computer-assisted self-interviewing (ACASI) methods may be more effective for improving the accuracy of surveys, as they provide respondents with greater confidentiality. These methodological adjustments are essential for obtaining more representative and reliable results.

Recommendations

Current polling methods face three primary challenges: non-response bias, undercoverage of specific demographics, and measurement bias. Non-response is unavoidable, but its rate can be reduced through a systematic approach. Before data collection, surveys should be carefully designed with clear communication and incentives to encourage participation. During data collection, follow-up procedures, interviewer training, and mixed-mode collection methods can increase response rates. After data collection, adjustments such as weighting the sample to match the population and imputing missing answers help address non-response bias effectively.

To mitigate undercoverage, mixed approaches like combining stratified sampling with quota sampling can ensure broader representation, particularly for rural and low-income populations. Reducing measurement bias requires better survey design and neutral question phrasing. Implementing these solutions involves allocating resources to reach underrepresented groups, regularly validating polling methods against actual outcomes, and training pollsters to avoid leading or biased questions. These efforts collectively enhance the accuracy and reliability of polling results.

Conclusion

This analysis underscores the limitations and biases inherent in current polling methods, including non-response, undercoverage, and measurement errors. While these issues present significant challenges, implementing improved survey designs, employing advanced sampling techniques, and ensuring broader demographic representation can enhance the reliability of poll results. By applying these strategies and continuously validating polling methods against actual election outcomes, researchers can build more accurate and representative polling systems. Ultimately, addressing these challenges is essential for fostering trust in electoral predictions and the democratic process.

Resources

• List of Wikipedia pages for election polls.

https://en.wikipedia.org/wiki/Opinion_polling_for_the_2002_French_presidential_election

https://en.wikipedia.org/wiki/Opinion_polling_for_the_2022_Brazilian_presidential_election

https://en.wikipedia.org/wiki/Opinion_polling_for_the_2022_South_Korean_presidential_election

https://en.wikipedia.org/wiki/Nationwide_opinion_polling_for_the_2008_United_States_presidential_election

https://en.wikipedia.org/wiki/Nationwide_opinion_polling_for_the_2012_United_States_presidential_election

• Access to data visualization tools (e.g., R, Python)

• Course materials on sampling methods and data visualization techniques