Lab 10 – MATH 240 – Computational Statistics

Sankalp Ojha

Colgate University

Mathematics

sojha@colgate.edu

04/10/2025

Abstract

This document provides a basic template for the 2-page labs we will complete each week. Here, briefly summarize what you did and why it might be helpful. Provide all the top-line conclusions, but avoid providing *all* the details. Results should be limited to "we show X, Y, and Z."

Keywords: Margin Of Error (MOE).

1 Introduction

Gallup polls published a document called *"How Are Polls Conducted?"* that describes how Gallup selects people to include in its poll and other details. Gallup states towards the end of the document:

"For example, with a sample size of 1,000 national adults (derived using careful random selection procedures), the results are highly likely to be accurate within a margin of error of ±4 percentage points."

"If Gallup increases a poll sample size to 2,000, the results would then be accurate within ±2of the underlying population value, a gain of two percentage points in terms of accuracy, but with a 100% increase in the cost of conducting the survey."

In a study from February 3-16, 2025 poll of 1,004 adults aged 18+ living in all 50 U.S. states and the District of Columbia showed that only 39% of Americans were satisfied with the United States's position in the world, whereas 59% of people were dissatisfied. **Note:** 2% had no opinion on the United States's position. Gallup reported the ±4% margin of error.

Through this lab, we aim to study how larger sample sizes lead to lowering the margin of error. Margin of error is a measure which aims to describe the variation in our using the sample proportion $\hat{p}$ as an estimate of p. The measure aims to tell us how much the sample proportion to differ from the proportion for the entire population.

Statisticians and quantitative researchers aim to report their studies with at least 95% confidence. Having a 95% confidence indicates that the researcher is 95% confident that if we conduct the study repeatedly, the interval would contain the actual proportion 95% of the time.

For this lab, we first ran simulations of 10K random polls for a sample size of 1004 individuals and again with a sample size of 2008 individuals to analyze whether the statement of Gallup claiming doubling the sample size will decrease the MOE. Both simulations were conducted assuming we had a known probability, p, of 0.39. We then used resampling to obtain the MOE and sampling distribution for $\hat{p}$. We finally ran simulations where we varied both the sample size, n, and probability, p, to see how varying combinations of n and p affect the estimated MOE and Wilson MOE.

2 Simulations

To conduct our basic and resampling simulations, we need to follow the process of setting a sample size, declaring an amount of polls to randomly conduct, and calculating the MOE for the given data. The Simulation over n and p and Actual Margin of Error Calculation sections differed as they ran simulations varying both the n, sample size, and p, probability, parameters. To summarize the data from our simulations, we used the `ggplot2` (Wickham, 2016) and `patchwork` (Pedersen, 2024) packages to make and combine the plots. We used the `xtable` (Dahl et al., 2019) package to create tables to tabulate our data.

2.1 Basic Simulation

For the basic simulation, we ran a simulation with 10K random polls. We assumed the probability of surveying a satisfied US citizen was 0.39, as giving by the Gallup study. The sample size for the first simulation was $n = 1004$ and $n = 2008$ individuals for the second simulation. We used `ribom()` to generate our random samples within the constraints.

2.2 Resampling

In the previous section, we conducted samples based on an assumed value of p, 0.39. Resampling will allow us to forgo the assumption made about the sample proportion in the previous section and simulate our own sample distribution of $\hat{p}$ for a Gallup survey. **Note:** The data which can be used for resampling is constrained to the sample size of $n = 1004$. We ran 1000 resamples to find the sample distribution of $\hat{p}$.

2.3 Simulation over n and p

So far the simulations were conducted under a constrained sample size and proportion. The next simulation aims to represent how the MOE varies as the sample size and proportion change for a sample. Using nested `for` loops with varying sample sizes from 10 to 3000 in increments of 10, proportions from 0.01 to 0.99 in increments of 0.01, and a set poll size of 10K, as in the basic simulation, simulations were ran for each condition and saved to a `tibble`. The MOE was summarized using a `geom_raster()` plot.

2.4 Actual Margin of Error Calculation

3 Results

Tie together the Introduction – where you introduce the problem at hand – and the methods – what you propose to do to answer the question. Present your data, the results of your analyses, and how each reported aspect contributes to answering the question. This section should include table(s), statistic(s), and graphical displays. Make sure to put the results in a sensible order and that each result contributes a logical and developed solution. It should not just be a list. Avoid being repetitive.

3.1 Results Subsection

Subsections can be helpful for the Results section, too. This can be particularly helpful if you have different questions to answer.

4 Discussion

You should objectively evaluate the evidence you found in the data. Do not embellish or wish-terpet (my made-up phase for making an interpretation you, or the researcher, wants to be true without the data *actually* supporting it). Connect your findings to the existing information you provided in the Introduction.

Finally, provide some concluding remarks that tie together the entire paper. Think of the last part of the results as abstract-like. Tell the reader what they just consumed – what's the takeaway message?

Bibliography: Note that when you add citations to your bib.bib file *and* you cite them in your document, the bibliography section will automatically populate here.

References

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.

Pedersen, T. L. (2024). *patchwork: The Composer of Plots*. R package version 1.3.0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

5   Appendix

If you have anything extra, you can add it here in the appendix. This can include images or tables that don't work well in the two-page setup, code snippets you might want to share, etc.