Lab 10 – MATH 240 – Computational Statistics

Caroline Devine
Colgate University
Mathematics Department
cdevine@colgate.edu

04/08/2025

Abstract

This lab explores how sample size and population proportion affect the margin of error in polling results. By using four approaches, we analyzed the accuracy of each method: basic simulation, resampling, simulations, and analytical evaluation. Our results highlight that large sample sizes reduce margin of error, but that the margin of error also depends on the population proportion. Additionally, we found that the Wilson margin of error formula is the most efficient and corresponds to the simulations over variations of n and p.

Keywords: Satisfaction with Polls, Basic Simulation, Margin of Error Estimates, Wilson's Estimate

1 Introduction

Gallup published a document called *"How are Polls Conducted?"* (Gallup, Inc., 2020) which outlines how Gallup selects people to include in its poll. Focusing on two excerpts describing sample sizes of 1,004 and then 2,008, Gallup highlights how the results are accurate within a certain margin of error ($\pm 4\%$ for $n = 1000$ and $\pm 2\%$ for $n = 2000$). This margin of error tells researchers how much they can expect the sample proportion to differentiate from the proportion for the entire population. The goal of statisticians and quantitative researchers is to report a margin of error that provides 95% confidence meaning they are 95% confident that if they were to conduct the poll repeatably, they would contain the actual proportion 95% of the time on the interval. The goal of this lab is to explore how sample size affects the margin of error as well as how to estimate this accurately.

2 Methods

We will explore different simulations of different sample sizes and look at the differences in margins of errors through various mediums. To do this, we will look at basic simulation, resampling, simulations over different sample sizes and proportions, and calculate the actual margin of error.

2.1 Part 1: Basic Simulations

Initially, we conducted a basic simulation study assuming a true probability, $p = 0.39$, that someone is satisfied with the position of United States in the world today to use the binomal distribution which measures the number of successes in a fixed number of trials. This was done using `rbinom()` function in `R`. We started with a sample size of 1,004 to replicate Gallup's example and doubled it for the second simulation completed. The sample proportion of the satisfied participants for each simulated poll was computed and plotted to visualize the distribution for both simulations. We also estimated the margin of error through calculation of the range of middle 95% for comparison to Gallup's results.

2.2 Part 2: Resampling

We assumed a true population proportion p in part 1 which allowed us to see how the polls look like under that assumption, but in reality, we do not always know the true population proportion. To remove this assumption, we preform resampling to approximate the sampling distribution of $\hat{p}$ using the Gallup survey data. This data reported that 39% of respondents were satisfied with the position of the US in the world today, 59% were dissatisfied, and 2% had no opinion. These were categorized into "satisfied", "unsatisfied", and "no opinion" categories for analysis, with the focus on estimating the proportion of respondents who were "satisfied". Using the same sample size ($n = 1004$), we preformed 1,000 resamples and plotted them on a histogram with a superimposed density curve for visualization purposes. Margin of error was also calculated for comparison.

2.3 Part 3: Simulation over n and p

To further explore how sample size(n) and the true population proportion(p) affects the margin of error, we simulated 10,000 simulations of a combination of increasing sample sizes and proportions. This is similar to the simulations in part 1, but at a larger scale. For each case, we calculated the margin of error using half of the range between the 2.5th and 97.5th percentiles of the sampling distribution. Using `geom_raster()` function in `R` which produces a heatmap, we plotted the margin of error estimates across different values of n and p to visually see how population proportion and sample size affects margin of error.

2.4 Part 4: Actual Margin of Error Calculation

Estimating the margin of error from simulations is strong, but we can also find the actual margin of error through derivations of the Central Limit Theorem to find the Wilson Estimate analytically. The Wilson Estimate is a weighted average between $\hat{P} = \frac{X}{n}$ and $\frac{1}{2}$ where the weights are n and Z^2 respectively. This provides an approximation of the margin of error called the Wilson margin of error formula:

$$z_{1-\alpha/2} * \frac{\sqrt{n\hat{p}(1-\hat{p}) + \frac{z^2_{1-\alpha/2}}{4}}}{n + z^2_{1-\alpha/2}}$$

We computed the Wilson margin of error using the same values from part 3 for n and p and calculated z to be 1.96 using the standard normal distribution for a 95% confidence interval. The table of errors were plotted them using a heatmap for comparison purposes. We did not need to use simulations for this step.

3 Results

3.1 Basic Simulations

The basic simulation study's goal was to examine how different sample sizes affects the sampling distribution and thus, the margin of error. Figure 1 shows the visual representation of the sampling distrbutions and their density curves. Both appear to be approximately symmetric and bell-shaped. For `n = 1004`, the distribution is a relatively normal distribution is centered around 0.39. For `n = 2008`, the distribution has less variability because the spread of the sample proportions is more concentrated. Table 1 corresponds to the graphical conclusions showing the margin of error decreasing as sample size increasings, aligning with Gallup's reported margins of error for each sample size.

Sample Size	Range of middle 95%	Margin of Error
1004.00	0.06	0.03
2008.00	0.04	0.02

Table 1: Margin of Error by Sample Size

3.2 Resampling

The 1,000 resamples of the same sample sizes using Gallup's reported survey data to approximate the sampling distribution of the sample propotion resulted in a margin of error around ($\pm$2.93%. The range of the middle 95% can be seen in Table 2. The margin of error and the shape of the distribution (Figure 2) approximately matches the basic simulation. This reinforces the reliability of Gallup's estimate, but shows there are slight differences. Due to the limited data from Gallup, we were only able to resample with one sample size, meaning we could not double the sample size as we did in the basic simulation. Ultiamtely, this shows that resampling is an efficient tool to estimate sampling distributions and variability when the true population proportion is unknown, which in reality, is typically accurate.

Sample Size	Range of middle 95%	Margin of Error
1004.00	0.06	0.03

Table 2: Margin of Error with Resampling

3.3 Simulations and Wilson Estimate

As we increased the sample size (n) and the true population proportion(p), the results of the simulated trials showed that larger sample sizes do reduce the margin of error as seen in the left plot of Figure 3. The darker colors represent low margins of errors and the lighter ones reflect higher margins of errors. Although, the margin of error does not soley depend on the sample size; it also depends on p. Figure 3 shows that when p is close to 0 or 1, the margin of error is smaller due to the limitations of the parameter space. This reflects higher variability when the population proportions are close to 50% and sample sizes are small. This adds a missed refinement to Gallup's story. The actual margin of error calculation, shown in the right plot of Figure 3, reflects similar results across all n and p combinations. The Wilson formula, used to generate the actual margin of error calculation, allows for an analytical check to the simulation approach. The results of both the simulation and analyitcal approach align strongly, with small variations in the simulation approach which can be accounted by sampling error. These results highlight how both simulations and analyziation approaches are accurate, with the Wilson formula method providng more efficient results.

4 Discussion

By analyzing how different approaches through basic simulation, resampling, simulation, and analytical evalutation, we are able to see how different sample sizes and population proportions affect the margin of error. This is extremmely important for statisticans and quantitative researchers for ensuring that the information they are reporting is accurate. The overall takeaway shows that Gallup is correct that larger sample sizes reduce the margin of error, but missed that the margin of error is also sensitive to the value of the population proportion. This can become an issue if it is overlooked when reporting information about satisfactions with polls in the United States. It is important to report accurate and full information, especailly in today's political climate.

References

Gallup, Inc. (2020). *How are Polls Conducted?* Accessed April 8, 2025.

5 Appendix

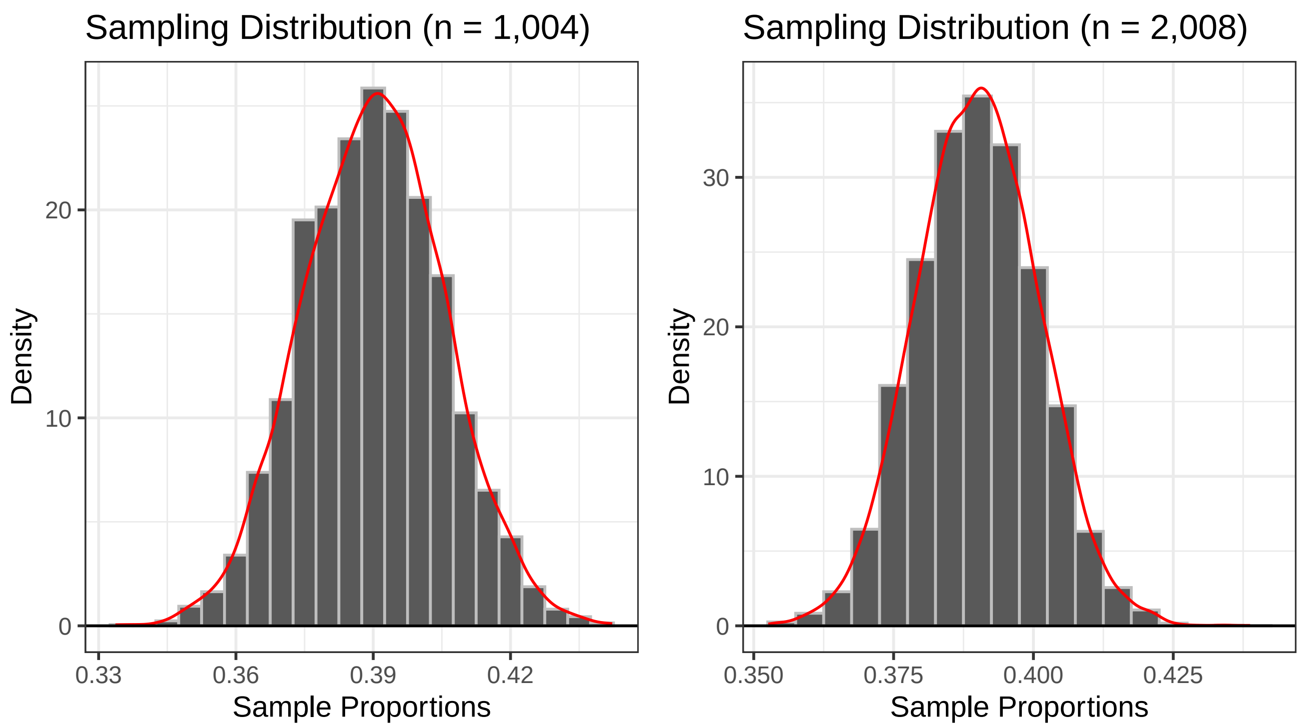

Figure 1: Sampling Distributions

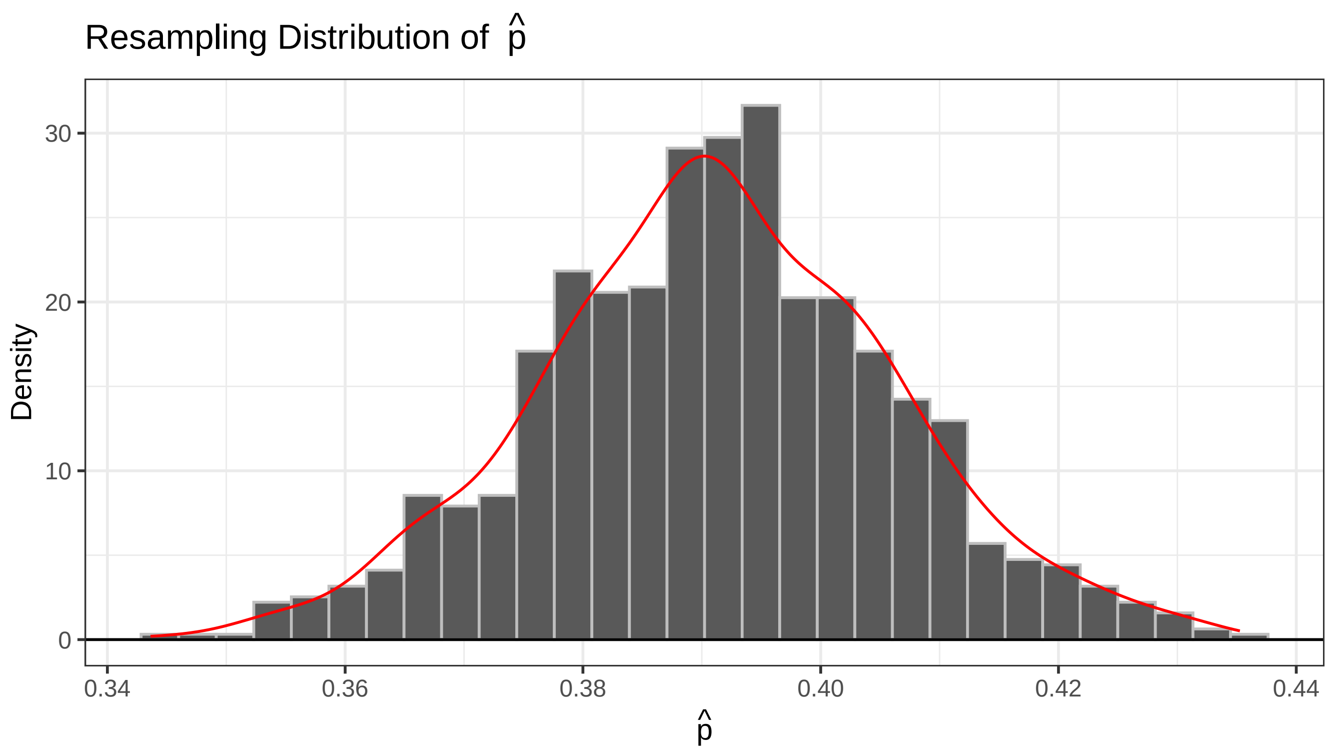

Figure 2: Resampling Distributions

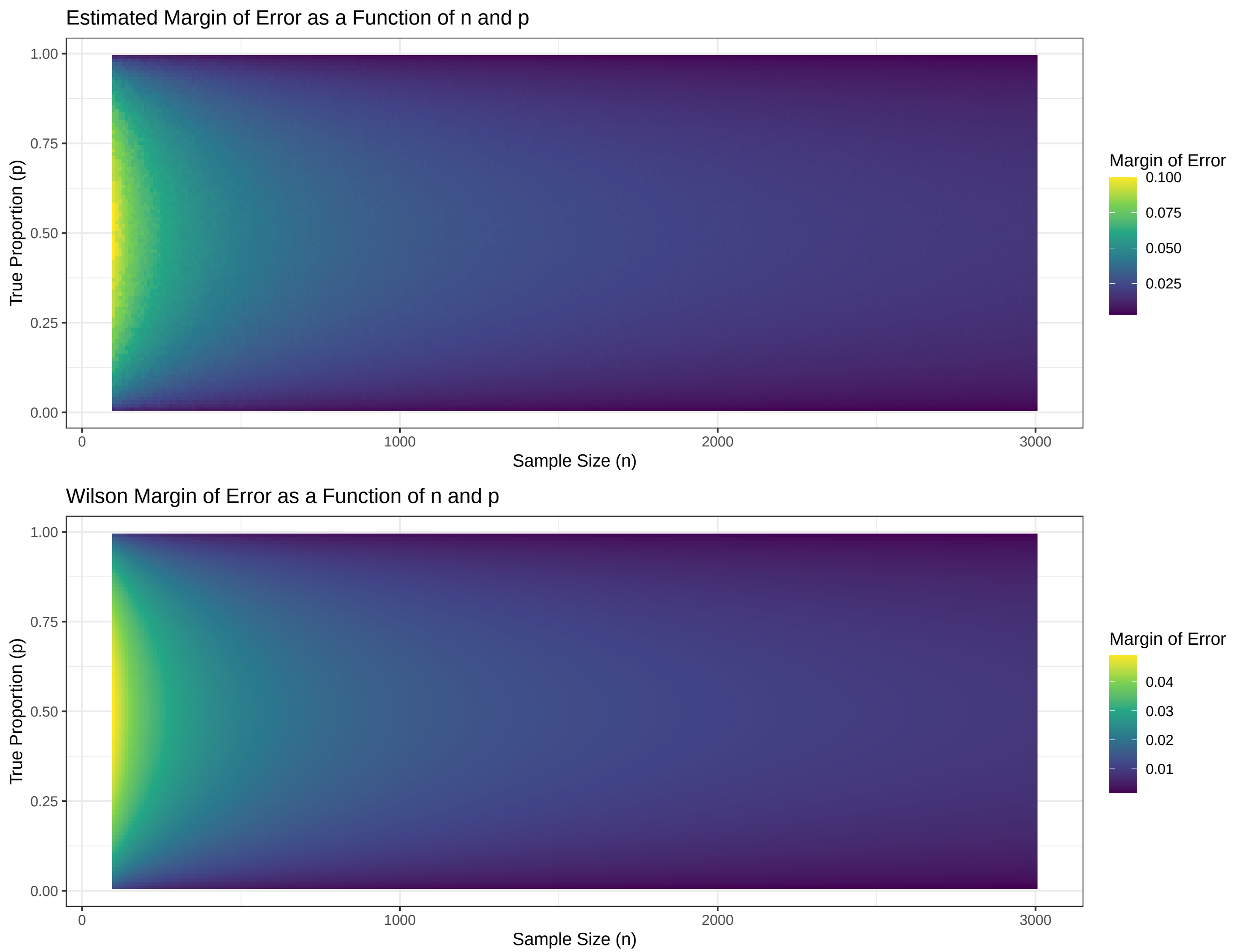

Figure 3: Heatmap: Margin of Error Comparisons