Lab 10 – MATH 240 – Computational Statistics

Caroline Devine
Colgate University
Mathematics Department
cdevine@colgate.edu

04/08/2025

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: What topics does the lab cover concerning class? List 3-4 key terms here, separated by semicolons.

1 Introduction

Gallup published a document called *"How are Polls Conducted?"*. It outlines how Gallup selects people to include in its poll as well as other details. Focusing on two excerpts describing sample sizes of 1,000 and then 2,000, 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 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:

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

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

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

Figure 1

3.2 resample

```
## Error in parse(text = input):  <text>:5:0:
unexpected end of input
## 3:       include.rownames = FALSE,
## 4:       size = "small",
##   ^
```

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.

3.3 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.
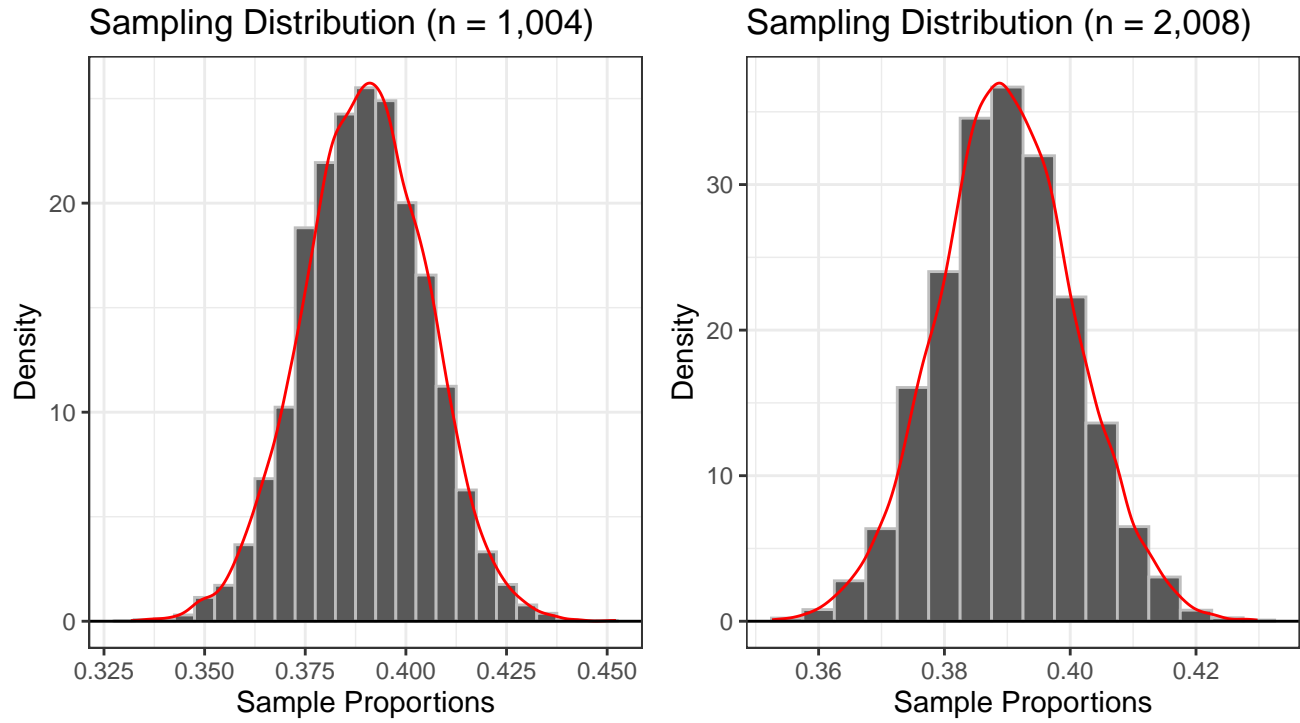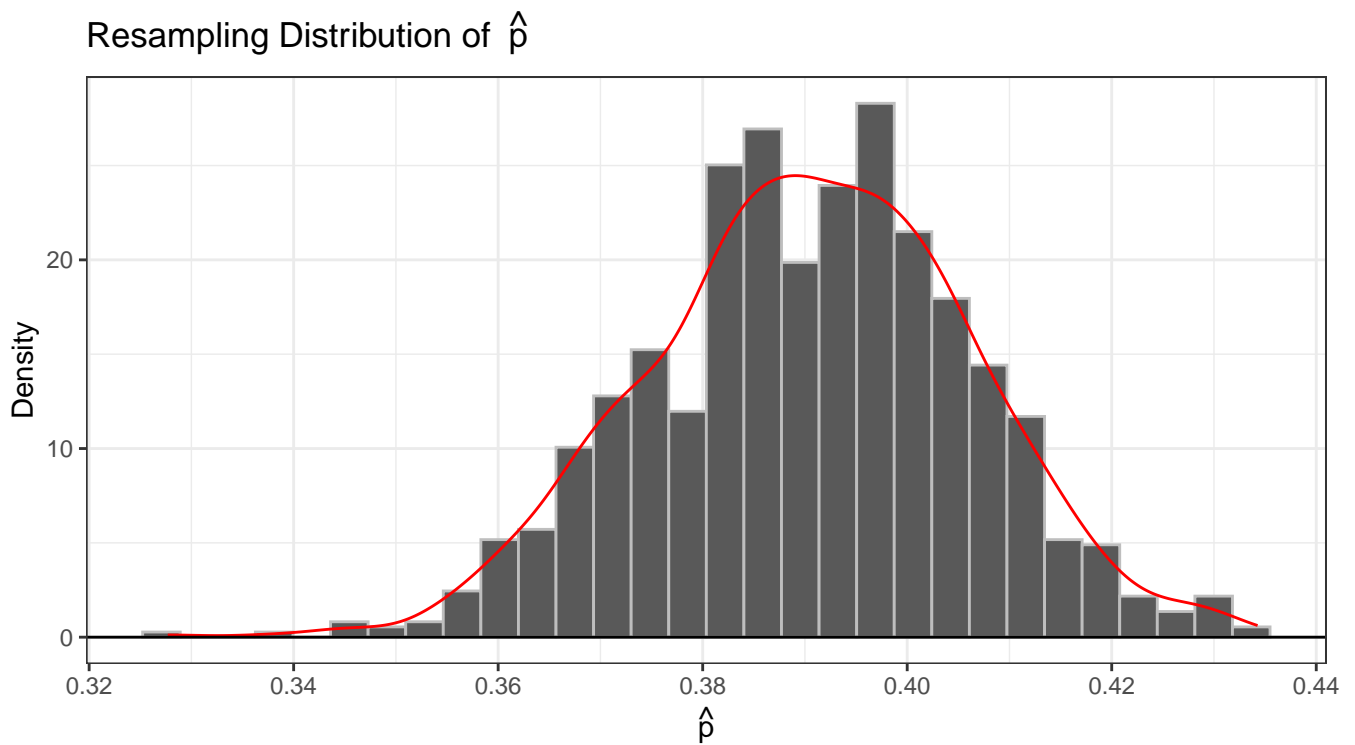
5   Appendix

Figure 1: Sampling Distributions

Figure 2: Resampling Distributions