Lab 10 – MATH 240 – Computational Statistics

Camilo Granada Cossio
Colgate University
Department of Mathematics
cgranadacossio@colgate.edu

Abstract

In this lab, we explored how sample size and population proportion affect the margin of error in survey sampling. We conducted simulation studies and apply the Wilson margin of error formula to compute and visualize margins of error. Our findings demonstrate that while larger sample sizes generally reduce margin of error, the true proportion play a crucial role, especially at the extremes. This analysis provide better context for interpreting polling results like those reported by Gallup.

Keywords: Sampling distributions; Margin of error; Wilson margin of error; resampling

1 Introduction

Public opinion polls often report a fixed or rounded margin of error, such as ±4%, but this simplification hides important nuances. The margin of error depends not only on the sample size but also on the estimated proportion of success. In this lab, we investigate how both factors contribute to uncertainty in polling estimates.

We begin by using simulated polling data to visualize the sampling distribution of the sample proportion under assumed population parameters. Then, we use resampling techniques to examine variation. Finally, we compute margins of error using the Wilson margin of error formula across a range of values to generalize our findings.

1.1 Intro Subsection

Gallup states a single margin of error for their polls. However, the sampling distribution's spread varies depending on the true proportion being estimated. When the true proportion is close to 0 or 1, variability is constrained, and the margin of error is smaller. This lab aims to make that relationship clear through simulations and mathematical reasoning.

2 Methods

We conducted three main analyses:

1. Simulation study: We assumed a true proportion of 0.39 and generated 10000 samples for various sample sizes using `rbinom()`. We visualized the sampling distributions and calculated the middle 95% range of simulated sample proportions.

2. Resampling: We simulated an original sample of 1004 individuals with 39% satisfaction and resampled 10000 times with replacement to estimate sampling variability.

3. Wilson margin of error formula: We computed the margin of error using the Wilson formula across values of sample size (100 to 2000) and true proportion (0.01 to 0.99). This allowed us to create a heatmap of margin of error as a function of both variables using `geom_raster()`.

`R` packages used include `tidyverse` (Wickham et al., 2019), `ggplot2` (Wickham, 2016), and `patchwork` (Pedersen, 2024) for data manipulation and visualization.

2.1 Mathematical Derivation

We used the Wilson margin of error formula:

$$\text{MOE}_{\text{Wilson}} = z_{1-\alpha/2} \times \frac{\sqrt{n\hat{p}(1-\hat{p}) + \frac{{z_{1-\alpha/2}}^2}{4}}}{n + {z_{1-\alpha/2}}^2}$$

where $z^2_{1-\alpha/2}$ is the critical value from the standard normal distribution (1.96 for 95% confidence).

3 Results

We began with a simulation assuming p = 0.39 and n = 1000, and found that the middle 95% range of sample proportions was approximately [0.361, 0.419], giving a margin of error of about 2.9%. When we doubled the sample size to 2000, the margin of error shrank to about 2.1%, matching our theoretical expectations.

In the resampling section, we created a sample of 390 satisfied and 610 unsatisfied responses, and conducted 10000 resamples. The histogram of sample proportions mirrored the previous simulation, reinforcing the expected behavior of sampling variability.

Using the Wilson margin of error formula, we computed the margin of error across a grid of n and p values. A `geom_raster()` plot revealed that the margin of error is highest around p = 0.5 and decreases toward the boundaries at 0 and 1. It also decreases as sample size increases, though with diminishing returns.

4 Discussion

Our simulations and mathematical calculations show that a fixed margin of error (like ±4%) is overly simplistic. While larger samples do reduce variability, the proportion of interest strongly influences the precision of estimates. The Wilson margin of error provides a more nuanced and accurate measure of uncertainty, especially for small or extreme values of p.

For poll readers, this means interpreting results with caution. A statement like 39% ±4% is only valid under specific assumptions, and real-world data may have more or less uncertainty.

Future polls should report margins of error as a function of both sample size and observed proportion, or at least clarify the assumptions behind a fixed error range.

References

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.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

5 Appendix

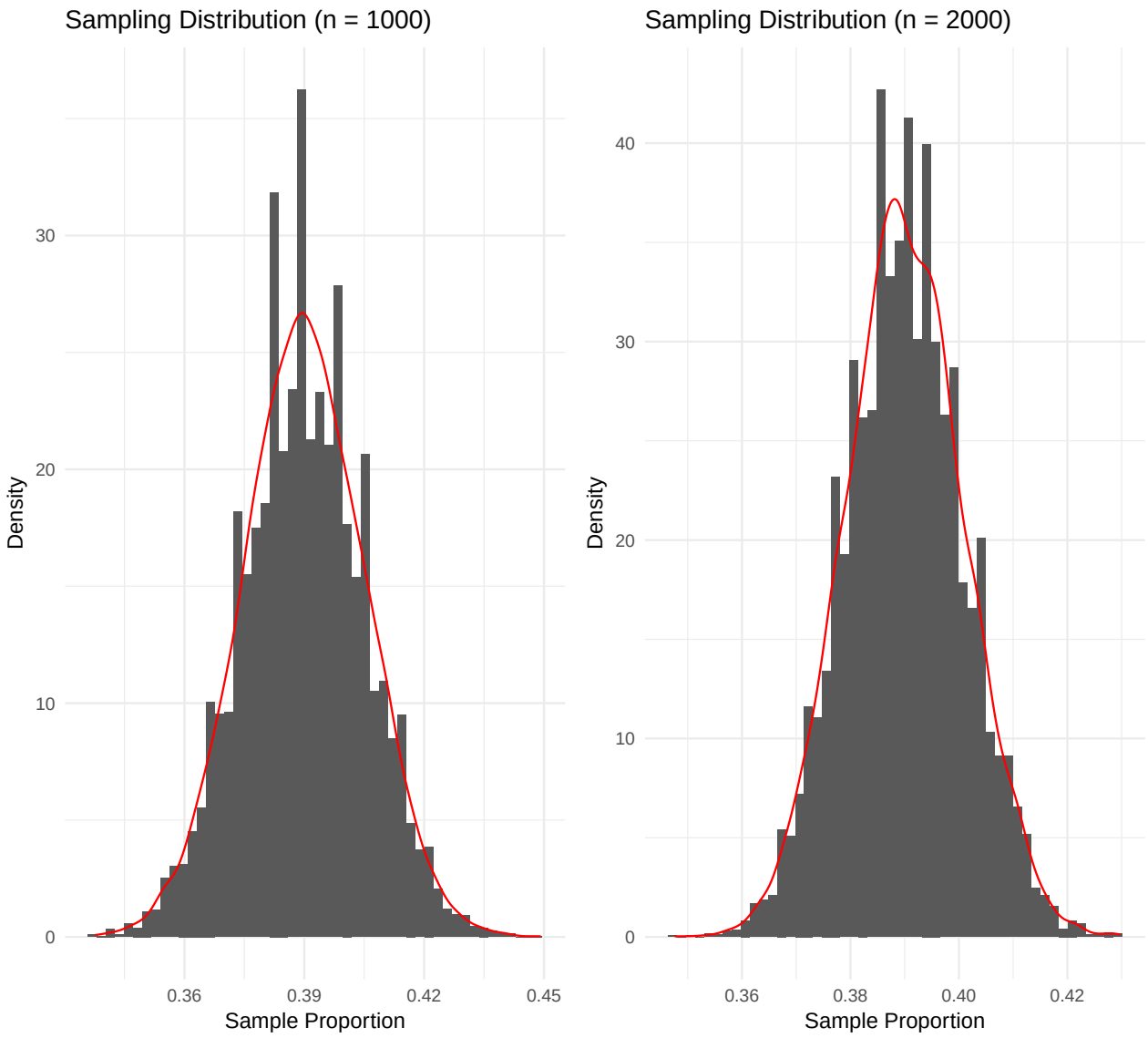

Figure 1: Sampling Distributions for n = 1000 and n = 2000

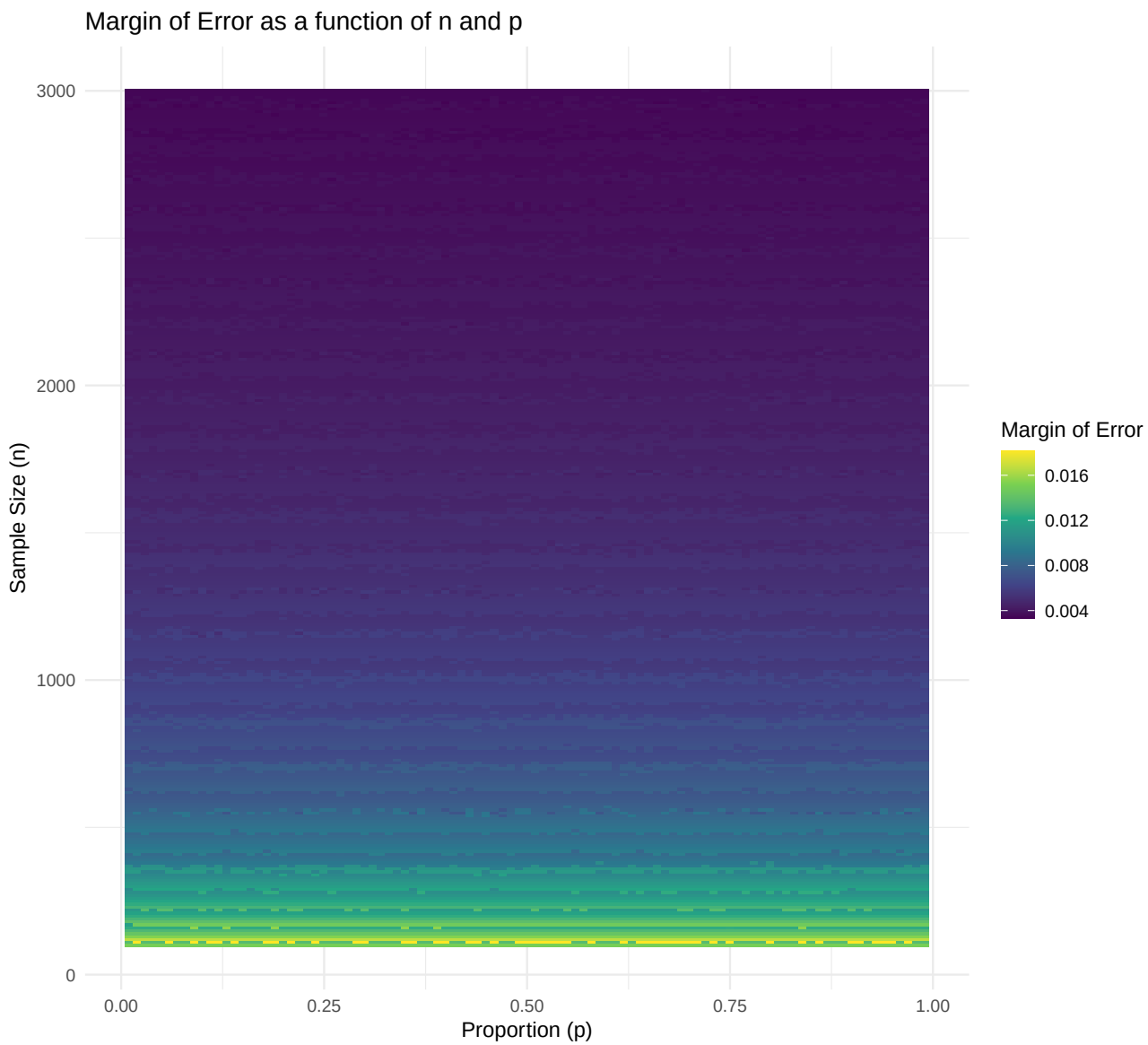

Figure 2: Margin of Error Heatmap for Resampling of Gallup Data

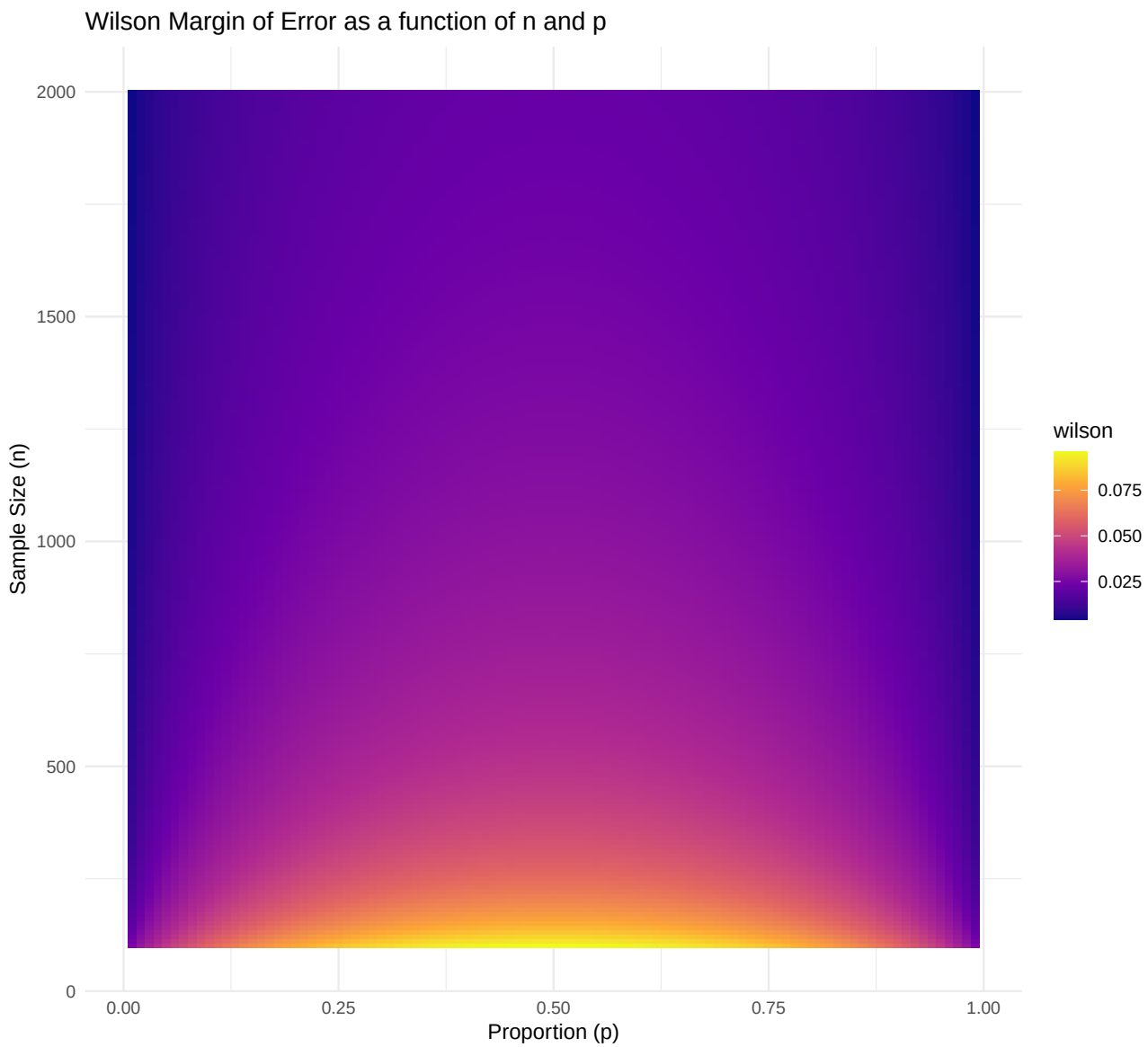

Figure 3: Wilson Margin of Error Formula Heatmap