Lab 10 – MATH 240 – Computational Statistics

Andrew Li
Colgate University
Math department
ali@colgate.edu

04/08/2025

Abstract

In this lab, we conducted a simulation study and explored resampling to learn how they can be used to model real world data. Then, we explored what the margin of error of a value can mean and what variables have an impact on it.

Keywords: Margin of error, resampling, sample size

1 Introduction

Gallup polls described their polling process in a document called "*How are Polls Conducted*" and made several claims about their margin of error, one being doubling sampling size from 1004 to 2000 halves their MOE from 4% to 2%. The margin of error helps us understand the variation of our sample proportion as an estimate of the population proportion, and we can estimate the MOE by assuming the population proportion to create a simulated study. However, there are times when the population proportion is unknown, and resampling makes it possible to gather information about the population by using the original sample.

To see what other factors may affect the margin on error besides sampling size, we calculated the margin of error that provides 95% confidence interval over different sample sizes and proportions. Then, we found the Wilson MOE formula to get more information.

species	Mean	Median	SD	IQR
Adelie	3700.66	3700.00	458.57	650.00
Chinstrap	3733.09	3700.00	384.34	462.50
Gentoo	5076.02	5000.00	504.12	800.00

Table 1: This is a table.

2 Methods

First, we tested the claim if increasing sample sizes from 1000 to 2000 would reduce the MOE. With the `rbinom()` function from the `stats` (R Core Team, 2024) package, we created 10000 polls of size 1004 with an assumed population proportion of 0.39 and plotted a histogram of the sampling distribution using `tidyverse` (Wickham et al., 2019) package. Then, we repeated the process with polls of size 2008 to measure the difference in moe.

Using the Gallup poll from February 2025, we have a sample size of 1004 from people across the 50 states and have a sample proportion of 39% for the people who were satisfied with global standing of the US. Knowing this, we represented that sample with a numeric vector of size 1004. 39% of the entries were 1's to stand for the people who were satisified and the rest were 0's for those who were dissatisified or neutral. Those two groups can be put together as the focus on is on the 39%.

By resampling the data 10000 times with the basic `R` function `sample()`, we can approximate the sampling distribution for the sample proportion without knowing the actual population proportion.

3 Results

The histograms of the polls are shown below. They both follow a relatively normal distribution with little skewness, but the sampling distribution for the smaller sample size in Figure **??** has a larger variability compared to the Figure 2. This matches with the calculated margin of errors as the polls with sample size of 1004 had a MOE of 0.0309 (approximately 3.1%), and the polls with sample size of 2008 had a MOE of 0.0211 (approximately 2.1 %). The MOE for sample size 1004 is within the projected 4% and while the 2.1% is slightly more than the Gallup polls statement of within 2% MOE, it is still very close for the polls of sample size 2008.

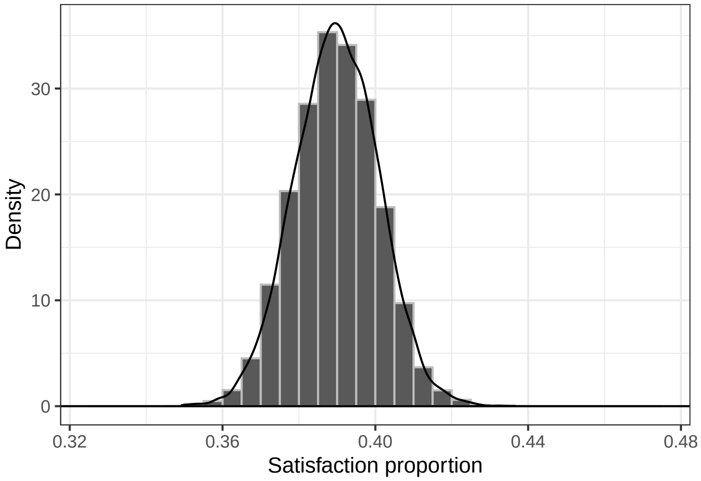

Figure 2: 10000 polls with p= 0.39 and n = 2008

The histogram of the resampled polls is shown in Figure 3. It follows a relatively normal distribution and looks closer shape to the Figure ??. The MOE is 0.0303 (approximately 3%), which matches the Gallup's statement that their February 2025 poll was within 4% MOE.

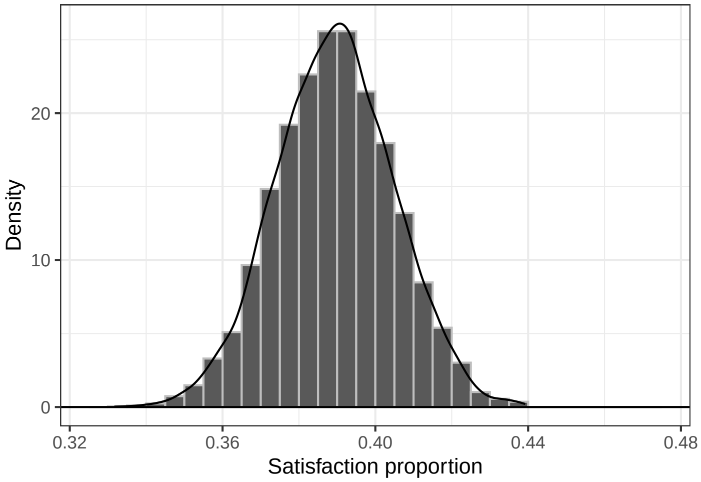

Figure 3: 10000 polls from resampling orginal sample

We finally look at the effects of the sample size and proportions on MOEs. Figure 4 is the estimated MOEs and Figure 5 is the Wilson MOEs. As sample size increases, the MOEs decreases. However, the MOE also decreases as proportion gets closer to 1 or 0 because the formula for MOE uses p(1-p), which becomes incredibly small when dealing with sample proportions close to the maximum or mimimum values.

Figure 4: MOE when comparing n vs p

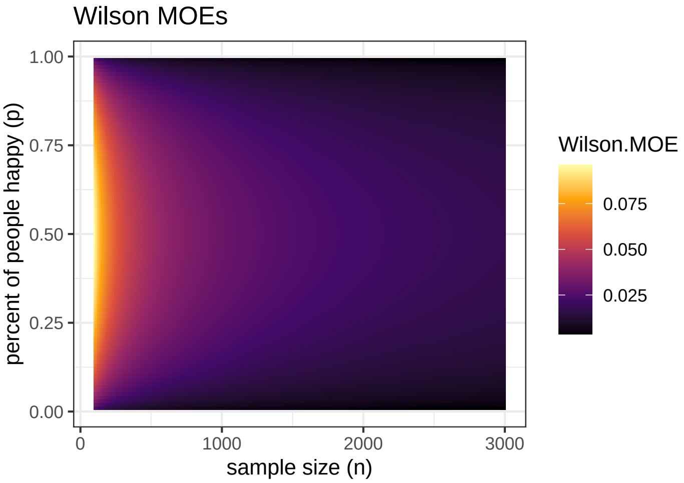

Figure 5: Wilson MOE when comparing n vs p

4 Discussion

After looking at the data, there is a decrease for increasing sample sizes, but the margin of error only gets reduced by 1% for doubling the sample size from 1000 to 2000. Depending how long it takes to conduct these polls or the cost of them, it might not be worthwhile to upscale them. Bigger sample sizes does reduce the margin of error, but so does sample proportion close to 1 and 0.

References

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.