

# Lab 10 – MATH 240 – Computational Statistics

Brendan Mariano  
Colgate University  
Mathematics  
bmariano@colgate.edu

## Abstract

This document provides a breakdown of why Gallup's assertion that doubling his sample size would halve his margin of error. We additionally look at the margin of error over different sample size and probability values.

**Keywords:** Resampling; Simulations; Margin of Error; Wilson's Estimate

## 1 Introduction

Recently, an author named Gallup published a document where he analyzed the proportion of people who were currently satisfied with the position of the United States. Within it, he made a claim stating that with 1,000 people in his sample he had a margin of error of 4% and that with 2,000 it is 2%. Our goal in this lab was to assess his claim and to generalize our findings to other sample sizes from 0 to 3,000 and other probabilities. We went about solving the first part of our problem by simulating his experiment using the same chance of satisfaction and sample size; to verify our findings, we also analyzed the margin of error using resampling. We address the second part of our question by calculating the margin of error from a simulation with various  $n$  and  $p$  values and by supporting our findings using the actual margin of error calculation (the Wilson estimate).

	Data	Margin of Error	Size
1	Sample	0.03	1004.00
2	Sample	0.02	2008.00
3	Resample	0.03	1004.00
4	Gallup	4.00	1004.00
5	Gallup	2.00	2008.00

Table 1: Comparing Margin of Error

## 2 Methods

Our overall goal was assess how valid Gallup's margin of error estimations were. Our first method was to test Gallup's claim about having a 4% margin of error by using a basic simulation which had 10,000 polls, the same probability (.39) and an equivalent sample size (1004). Using this data, we produced a graph using ggplot (Wickham, 2016) to view the distribution of the proportion satisfied. Since Gallup also claimed that the margin of error for 2004 would be 2%, we ran the same simulation using a sample size of 2008 and plotted that data.

We later tested Gallup's theory using resampling instead of a simulation since it is possible that it may only be true for one method and plotted the data for this. Another possibility is that Gallup's theory may only be true for certain probabilities or sample sizes. As a result, we calculated the margin of error using a simulation with 10,000 polls from sample sizes 100 to 3000 and with probabilities from .01 to .99. We plotted the margin of error in a raster plot to show its value for all of the given parameters. To be safe, we also calculated the margin of error with the same range of parameters using the actual formula (Wilson's formula) and put it into a raster plot. Using these two raster plots, we were able to assess the truth of Gallup's statements.

### 2.1 Methods Subsection

Much like the Introduction, subsections can be helpful for the Methods section. For example, you might describe data collection and the statistical analyses of the collected data in different subsections. Or, you may have different questions that require distinct methods.

## 3 Results

We found in the first part of the lab that the margin of errors we calculated (table 1) were inconsistent with the numbers proposed by Gallup. For a size of 1004, we found that the margin of error for our simulation and resampling were the same at 3%, but they were both less than Gallup's value. When we used the simulation with a size of 2008, we found that the margin of error was around 2%.

In terms of our simulation using different  $n$  and  $p$  values (figure ), we found that the margin of error was moderate to dark blue at any place where  $n$  is greater than 1,000 and that the difference in color at any probability from  $n = 1,000$  to  $n = 2,000$  was negligible. The raster plot for our Wilson estimate was identical and showed the same pattern. 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 phrase 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?

## References

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.