

Understanding Political Surveys and Their Implications

Prof. Jonathan Cervas

2024-08-29

```
# Remove all objects just to be safe.  
rm(list=ls(all=TRUE)) # Remove all objects from R history  
options(scipen=999) # Turn off Scientific Notation  
set.seed(666) # Setting a seed for reproducibility
```

Introduction

This assignment will guide you through the principles of political surveys, including the importance of sample size, the calculation of margin of error, and the impact of nonresponse bias.

We know from our reading and logic that the margin of error in a poll goes down as sample size goes up. We can show this in R using the following:

The margin of error for a proportion p is typically calculated using the formula:

$$\text{Margin of Error} = Z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

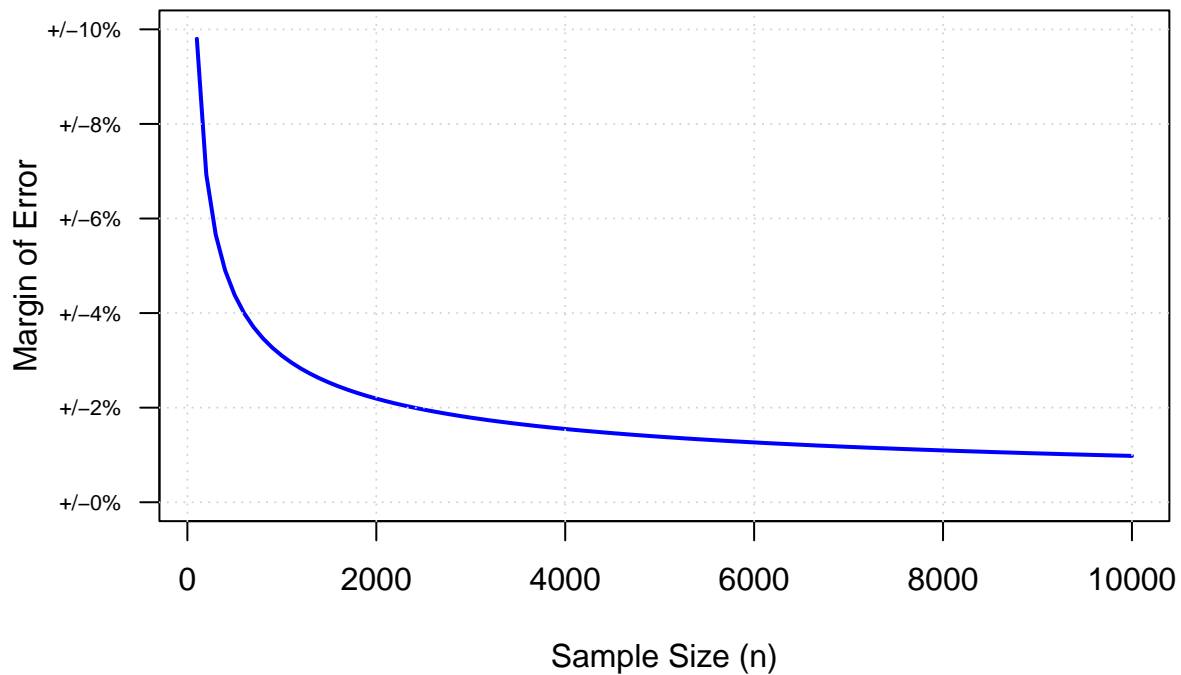
z is the z-score corresponding to the desired confidence level (e.g., 1.96 for a 95% confidence level),

p is the proportion (e.g., 0.5 for the maximum margin of error),

n is the sample size.

```
# Set parameters  
z <- 1.96 # z-score for 95% confidence level  
p <- 0.5 # worst-case scenario for proportion  
n <- seq(100, 10000, by=100) # sample sizes from 100 to 10,000  
  
# Calculate margin of error for each sample size  
margin_of_error <- z * sqrt(p * (1 - p) / n)  
  
# Plot  
plot(n, margin_of_error, type="l", col="blue", lwd=2,  
      xlab="Sample Size (n)", ylab="Margin of Error",  
      main="Margin of Error vs. Sample Size",  
      ylim = c(0,0.1),  
      yaxt = "n")  
  
y_values <- seq(0,0.1,0.02)  
axis(side=2, las=2, at=y_values, label=paste0("+/-", y_values * 100, "%"), cex.axis = 0.65)  
  
grid()
```

Margin of Error vs. Sample Size



Part 1: Simulating a Political Survey

Activity 1: Population Simulation

In the 2020 election, the candidates received these number of votes:

This implies that Trump won what share of the two-party vote?

We will explore how random sampling can produce relatively accurate polls, assuming no errors beyond sampling error.

Simulated 2020 Elections

Let's simulate a population of 100,000 voters based on this proportion:

In order to better understand how polls are conducted and how margin of error is calculated, we can create a simulated world reflecting the US in 2020.

Let's first create a vector of the 'population' of voters, and in that assign Trump voters the value 1 and Biden voters a value 0

We can simulate this in R using `rbinom()`, which will create a vector of 0 and 1s to some size in proportion to the value we give. Let's start with a smaller electorate (100000) but with the same underlying vote choices. We can see if it is right.

Now let us visualize this using a barplot

Bonus, let's try to "map" the voters for fun.

Sampling from the population

We can simulate to see how accurate our poll is. First, let us draw **ten** voters, completely randomly, from our population.

We can take the `mean()` to see what percentage voted for Trump, since our "population" is made up of 0 and 1s.

How accurate was it? We can actually find the deviation, since we know the true value (**47.7%**).

```
# Our sample missed the true value by
```

Now let's poll **ten** random people **1,000** times and see how accurate each is, both individually and in the aggregate.

How accurate were these polls?

Plot Trump's vote share in each simulation (hint: Histogram is a good way to visualize this kind of data)

Now plot the accuracy of the poll compared to Trump's actual vote share.

Ideally, we sample **one** subset of the population and produce error this low on the first try!

Sample a larger number of people

Let's see if our error goes down as we sample a larger number of people.

In R, we can first write a **function** to do a simulation. In the function, allow for a parameter to vary the number of people sampled, and for changing the number of simulations for greatest flexibility.

Now run your function to sample for 1,000 people.

Let's see what the 95% confidence intervals are:

We can see a significant reduction in uncertainty about Trump's vote share when we ask 1,000 people compared to when we only ask 10 people.

Plot the new histogram

Law of large numbers

If you are running a campaign and only have so many resources, you may not want to poll more people than necessary to understand the current state of the election. So how many people should you poll?

Simulate polls with sample sizes ranging from 10 to 5000

Create a box and whisker plot to demonstrate the reduction in error as the sample size increases

It is also useful to see the reduction in error as a function of sample size. Create a line plot showing the difference between the 2.5 percentile and the 97.5 percentile.

```
# 95 percentile for each number of persons sampled
```

Alternative Way to do this, calculating the margin of error through simulation

Write a function to calculate the margin of error and mean of each simulation:

Plot the mean of each sample, with the sample size on the x-axis and the mean Trump support on the y-axis.

```
# Plotting the Means
```

Plot the margin of error as the sample size increases

```
# Plotting the Margin of Error
```

Non-response

We also learned about non-response and how it might bias our sample. How might it matter if 40% of those who support Trump fail to respond to the pollster?