Predicting the Outcome of the 2016 USA Presidential Election

In 2012, data scientists, including Nate Silver, accurately predicted the U.S. presidential election outcomes by aggregating data from multiple polls. By combining poll results, they provided more precise estimates than a single poll could achieve.

In this exercise, we aim to predict the result of the 2016 U.S. presidential election by analyzing polling data and aggregating results.

The data for this exercise is in a CSV file named `2016-general-election-trump-vs-clinton.csv`. Note that some rows may represent subgroups (e.g., voters affiliated with specific parties) and contain NaN values in the "Number of Observations" column. Exclude such rows from your calculations to avoid errors.

Now do the following tasks:

1. Let X_i be a random variable where:
   - $X_i = 1$ if the i-th voter supports the Democratic candidate.
   - $X_i = 0$ if the i-th voter supports the Republican candidate.

   With $i = 1, 2, \ldots, N$, the Central Limit Theorem (CLT) states that if N is large:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \hat{p} \approx N\left(p, \frac{\hat{p}(1-\hat{p})}{N}\right)$$

   where p is the true proportion of voters supporting the Democratic candidate. Based on the CLT result, derive and compute the 95% confidence interval (CI) for p.

2. Suppose the true population proportion $p = 0.47$. Perform a Monte Carlo simulation with $N = 30$ and 10^5 iterations to show that the CI derived in Question 1 captures the true proportion p approximately 95% of the time.
3. Load the data from the dataset into your coding workspace, and then make a data frame containing only the columns `Trump`, `Clinton`, `Pollster`, `Start Date`, `Number of Observations`, and `Mode`. Exclude any rows where the `Number of Observations` is missing.

4. Create a time-series plot of poll results showing support percentages for Trump and Clinton, using different colors for each candidate. Include a smooth trend line to visualize support trends over time.
5. Calculate the total number of voters observed by summing all poll observations in the dataset.
6. Calculate the estimated proportion of voters favoring Trump and Clinton. Display these estimates in a table.
7. Using the aggregated data, compute the 95% confidence intervals for Trump and Clinton support proportions.
8. For illustrative purposes, assume there are only two parties, and let p denote the proportion of voters supporting Clinton. Consequently, $1 - p$ represents the proportion supporting Trump. We define the spread as the difference in support between Clinton and Trump:

$$d = p - (1 - p) = 2p - 1$$

Using the aggregated poll data, we estimate p as $\hat{p}$. Therefore, the estimated spread d can be approximated as:

$$d \approx 2\hat{p} - 1$$

This also implies that the standard error for the spread is twice as large as the standard error for $\hat{p}$. So, our confidence interval for the spread d is:

$$CI\ for\ d = (2\hat{p} - 1) \pm 1.96 \times (2 \times SE_{\hat{p}})$$

where $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ is the standard error of $\hat{p}$.

a) Calculate the 95% confidence interval for the spread d, using the formula provided above.

b) Conduct a hypothesis test to determine if the spread d is significantly different from zero by testing $H_0: d = 0$ vs. $H_a: d \neq 0$. Provide the test statistic and p-value.