ReadmeFile

This codes begins by loading the datset from the '*dslabs*' package in R Programming Language (Irizarry 2018). The data contains the Polls results from US 2016 presidential elections which are collected from 'HuffPost' Pollster and 'RealClearPolitics. An initial observation of dataset provided (polls_us-election_2016) suggest there are n = 4208 observations in total with 15 columns. Six columns are identified with no new or supportive information for the analysis. The columns regarding raw polls of candidates such as Gary Johnson and Evan McMullin are removed from dataset as they contains no predictive value. Subsequently, the underlying assumptions of sample sizes are investigated to determine whether they are normally distributed. Then the research question is formulated for this study, and parametric and non-parametric t-tests are applied to test the null hypothesis. Thereafter, a simulation fucntion is written to generate a random dataset that follows a normal distribution using '*rnorm* fucntion'. Simulation is conducted for both two sample t-test and Wilcoxon t-test to measure power and size of the null hypothesis under different sample sizes. Simulations repeatedly generate random data then analyze each data set and count the proportion of results that are significant. That proportion is the calculated power for sample sizes.

The simulation fucntion uses '*grid_search()',* to run simulation repeatedly and collate the results and graphically show them using '*ggplot*'. . The '*grid_search()*' function allows to run the function iteratively with varying parameters. In similar vein, the simulation is conducted to generate a dataset that follows chis-square distribution. The estimated result of power are then detailed in a table and showed graphically. At the end, this coding attempts to show a type II error graphically.