Problem Set 2

Directions: Complete each of the following prompts. Your answers must be professionally organized, typed, and all Stata code needed to reproduce your results must be appended to the end of your assignment. For any tables or figures you produce, make them as readable as possible. This means cleaning up axis labels, tick marks, using readable legends/color schemes, and so forth. Do not copy and paste directly from your Stata or R console. Your work is due via email by 11:59:59 pm on September 17, 2021.

1. Using the dataset you built for Problem Set 1, do the following with the assistance of Stata or R:
    a. Create a table of descriptive statistics that include mean, standard deviation, range, and a brief description of the following variables:[1]
        i. Total population of a state
        ii. Percent of a state that voted for Biden
        iii. Dummy variable for whether Biden won a state's 2020 Electoral College votes
        iv. Percent of a state that is urban
        v. Dummy variable for whether a state flipped its vote between 2016 and 2020
    b. Using the results from the previous table, give a brief, textual analysis of these descriptive findings.
    c. Generate the following figures via individual plots or via one lattice of plots:
        i. A boxplot of the percent urban variable, over the dummy variable for which candidate won the state
        ii. A histogram of state population size
        iii. A bar graph of the mean population of a state, over whether it flipped between 2016 and 2020
        iv. A scatterplot with Biden's percent of a state vote on the y-axis and the percent of a state that's urban on the x-axis.
        v. A histogram of the percent of a state that is urban
        vi. A pie chart showing the percentage of states via their status as having either flipped or not between 2016 and 2020.
    d. Using the results from the previous graphs, give a brief, textual analysis of what you see descriptively from each figure.
2. By hand, and showing how you arrived at your results, perform the following:
    a. Suppose you gather a sample of 5 individuals and their heights in inches. You observe the following: $H = \{62, 60, 68, 65, 72\}$. Using an α-threshold of 0.05 (one-tailed), construct a confidence interval to test the hypothesis that the average height of the entire population from which the sample was drawn is taller than 60 inches.
    b. Suppose you want to examine the fertility of pileated vs. yellowhammer woodpeckers. You go and observe the clutch size of four nests from each species of woodpecker and find $P = \{2, 4, 1, 3\}$ and $Y = \{4, 5, 3, 6\}$. Construct a null and alternative hypothesis for the relationship between clutch sizes in the two populations of interest. Using an α-threshold of 0.05 (two-tailed), use a z-table to test your alternative hypothesis.
    c. Using the data you collected for Problem Set 1, test the hypothesis that Biden-won states are more urban than Trump-won states in the 2020 presidential elections (feel free to have the software report to you sample means and so forth, but do the hypothesis-test by hand).
3. Answer the following:
    a. What is the null hypothesis, and what role does it play in hypothesis-testing?
    b. What is the difference between a Type I and Type II Error, and how does our choice of α affect the likelihood we commit either?

[1] If you're unsure what this table should look like, you can consult this paper's 19th page for a rough approximation.

c. Explain the logic of the central limit theorem and how we use the sampling distribution along with its dispersion parameter, the standard error, to engage in hypothesis-testing.
d. Explain the difference between probability and non-probability sampling along with their relative (de)merits. Give examples if needed.
e. Discuss some of the most common ways by which we measure central tendency and dispersion in our variables and under what conditions some measurement strategies might be more preferable to others.
f. Explain the concept of statistical significance. What is it; how do we assess it; and why might the concept of a binary significant/insignificant relationship be problematic for inference?
g. Explain the difference between one and two-tailed hypothesis testing and the types of alternative hypotheses for which either is appropriately implemented.