Assignment 15: Impact of Sample Size on ANOVA Results

110 Points scaled to 20 Points

Introduction

In this assignment, you will conduct an experiment to explore how the size of the sample drawn from a population impacts the results of an ANOVA test. Specifically, you will assess for differences in the percent of households in a county that do not have internet access ("per_no_internet") within different subregions of the country ("SUB_REGIONS"). You will make use of the "us_counties_data.csv" dataset. Make sure to read the associated description of the data ("us_counties_data_ DESCRIPTION.pdf") prior to undertaking the assignment. Note that not all of the assumptions of ANOVA may be met. For example, there is likely some spatial autocorrelation in the data, which may invalidate the assumption of independent samples. However, we will not worry about testing assumptions and assume that the results are valid.

Objectives

- *Conduct an Analysis of Variance (ANOVA) test and interpret the results*
- *Create stratified random samples and use loops to perform multiple iterations of a test using different random subsets of a population*
- *Explain the impact of sample size on the central tendency and variability of statistical results*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and answers to the questions embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and be answered within Markdown cells.*

Background Questions

Question 1. State the null hypothesis for the ANOVA test for this specific research question. (4 Points)

Question 2. State the alternative hypothesis for the ANOVA test for this specific research question. (4 Points)

Question 3. Why do we need to use an ANOVA test to investigate this research question as opposed to a T-Test? (4 Points)

Question 4. Explain how the F-Value associated with an ANOVA test is calculated. What defines the numerator and denominator? What defines the degrees of freedom? (8 Points)

Question 5. List and explain the assumptions of ANOVA. (8 Points)

Experiment and Questions

This experiment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following tasks and answer the associated questions.

Task 1. Read in the data table. (4 Points)

Task 2. Create a grouped boxplot that shows the distribution of percent of households without internet access grouped by subregion. (6 Points)

Task 3. Perform 50 iterations of an ANOVA test using different random subset of the available counties. Perform 50 replicate tests for each of the following sample sizes: 5, 10, 20, 30, 40, 50. (20 Points)

Task 4. Extract the obtained p-value for each test and save them to a table. (10 Points)

Task 5. Calculate the mean and standard deviation of the obtained p-values for each sample size. (10 Points)

Question 6: Based on the grouped boxplot, what do you expect to be the results of the ANOVA test? Explain your reasoning. (8 Points)

Question 7: Compare the mean p-values obtained relative to the different sample sizes. Does sample size impact the p-value obtained when using different samples from the same population? (8 Points)

Question 8: Compare the standard deviation of the p-values relative to the different sample sizes. Does sample size impact the variability of the p-value obtained when using different samples from the same population? (8 Points)

Question 9: Summarize how the size of the sample drawn from the population may impact the results of an ANOVA test and how this may impact your interpretation and use of the results. (8 Points)