Project
Comparison of Multiple Distributions

Mothers and their babies data is a dataset of data related to newborns and individual mothers. It contains 1236 samples and 23 independent variables. The independent variables include infant survival, birth weight, date of birth, sex, mother's ethnicity, age, education level, height, weight, and smoking status. This dataset can be found at https://www.stat.berkeley.edu/users/statlabs/labs.html. A clean version is uploaded on Moodle.

In this project, we are interested in the relationship between maternal smoking and babies weight, and whether different conditions of smoking lead to changes in the weight of different groups of neonates. The variable wt contains babies birth weight in ounces (999 = unknown) and smoke contains mothers smoke history (0 = never, 1 = smokes now, 2 = until current pregnancy, 3 = once did, not now, 9 =unknown).

Tasks:

1. Please use descriptive statistics to briefly describe the distribution of these two variables. For discrete variables, consider the count of the data, for continuous variables, consider the distribution of the data and the central tendency of the different groups.

2. Do the babies birth weights differ between the categories? Conduct a global test.

3. Are there pairwise differences between the resulting birth weights? Consider all pairs of categories and conduct two-sample tests. Adjust the test results with the Bonferroni correction and the Tukey's Honest Significant Difference (HSD). Please also calculate the Tukey's confidence interval.

4. Please compare the results of these two correction methods with the non-adjusted test and give a reasonable explanation.

Remember to check that all assumptions hold before applying the respective tests. You can assume all significance levels are defined as $\alpha = 0.05$. Please also state clearly the null hypotheses and the alternative hypotheses.