NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp Fall 2015

Foundations of Statistics

Question #1: Body Temperature

The `01temp.txt` dataset records the body temperatures, gender, and heart rate of 130 individuals; you may assume that the observations are independent of one another.

1. Read the data into R and provide basic numerical EDA to describe the data.
    a. Provide univariate EDA for all variables, and bivariate EDA where appropriate.
2. Provide basic graphical EDA to describe the data.
    a. Provide univariate EDA for all continuous variables, and bivariate EDA where appropriate.
3. You have heard that the average human body temperature is 98.6 degrees Fahrenheit. Does this data support this claim? (Perform a hypothesis test to determine whether the true population mean body temperature is 98.6 degrees Fahrenheit.)
    a. What is a 95% confidence interval for the average human body temperature?
    b. Interpret your results in context of the problem.
4. Is there a significant difference in body temperature between males and females? If so, quantify this difference.
    a. What is a 95% confidence interval for the average difference in human body temperature between males and females?
    b. Interpret your results in context of the problem.
5. You believe the variances of male heart rate and female heart rate are different; specifically, you claim that females have a lower heart rate. Test this claim.
    a. What is a 95% confidence interval for the ratio between female and male heart rates?
    b. Interpret your results in context of the problem.

Question #2: Plant Growth

Load the `PlantGrowth` dataset located in the `datasets` library; this dataset contains dried plant weight measurements for the same species of plant under three different conditions (two separate growth treatments, and a control group where no treatment was applied). You may assume that the observations are independent of one another.

1. Create side-by-side boxplots of the plant weights segmented by the type of applied treatment. Describe the features of the graph.
2. Calculate the standard deviations of each conditional distribution of plant weight based on the applied treatment. Do these differ significantly?
    a. **NB:** To avoid increasing our chance of encountering a "false positive," we must avoid applying three separate F-tests (treatment #1 vs treatment #2; treatment #1 vs control; treatment #2 vs control). As an alternative, Bartlett's Test of Homogeneity of Variances allows us to simultaneously test for the similarity of a group of variances, rather than just a pair. Implement this test using the `bartlett.test()` function and report your results.
3. Is there a significant difference in the weight of plants based on the growth treatments they were given? Conduct a hypothesis test and report your results in context of the problem.
    a. Given the results of the Bartlett test, is the result of your hypothesis test valid?

Question #3: Gender, Hair, & Eye Color

Load the `HairEyeColor` dataset located in the `datasets` library; this is a three dimensional dataset that records the hair color, eye color, and gender of 592 different statistics students.

1. Visualize the entire dataset using a mosaic plot. Describe what you see.
    a. Which category combinations receive more observations than expected? Fewer observations than expected?

2. Reduce your dataset to focus on just females with brown and blue eyes (but still include all hair colors). Create another mosaic plot and describe what you see.
    a. Conduct a hypothesis test to see if hair and eye color are independent of one another for this reduced dataset. Report your results in context of the problem.
3. For the reduced dataset, which category combination contributed most to any statistical deviations? Which category contributed the least? By how much for each?