Homework 1

Statistical Inference

1- A teacher wants to determine which of the following methods can be effective for students to enhance their learning experience in class:

   a. study with instrumental music
   b. study with vocal music
   c. study without music

   Design an experimental study to help the teacher answer his research question.

2- Answer the following questions and explain your reasons.

   a. The sales team of a sunglasses company is trying to test their sunglasses' effectiveness by examining the annual sales of sunglasses.
   In 2017, the company sold 3,000 sunglasses.
   In 2018, they sold 2,500 sunglasses.
   They assume that their sunglasses are less effective, which is why sales have decreased.

   Does this experiment have a potential confounding variable? If the answer is yes, find at least one confounding factor.

   b. Down syndrome occurs when an individual is born with three copies of chromosome 21 instead of the normal two copies. One can demonstrate that the frequency of Down syndrome increases with birth order, with a frequency of about 57 per 100,000 live births in 1st born children rising to about 164 per 100,000 in 5th born children. However, the frequency of Down syndrome also increases with maternal age, starting at about 40 per 100,000 live births in mothers under the age of 20 and rising slowly at first until age 35-39 when the frequency is about 270 per 100,000 live births, and then jumping to about 855 per 100,000 live births in mothers 40 years of age or older. It is certainly not surprising that the increase in birth order correlates with an increase in maternal age. Mothers giving birth to their firstborn will have a younger age distribution compared to those giving birth to their fifth child.

   Considering the figure below, explain which of them is more important. Birth order or maternal age? Is the association between birth order and Down syndrome confounded by maternal age? Or is the association between Down syndrome and maternal age confounded by birth order?

Homework 1

Statistical Inference

c. A teacher believes that using a new software can help students to get higher scores on their exams. To test this hypothesis, she encourages students to use the software on school computers for one hour after their class finishes. The teacher concludes that the software is actually effective because students who use them obtain higher scores in comparison to their peers.

What is the confounding variable in this study? How the confounding variables shouldbe controlled?

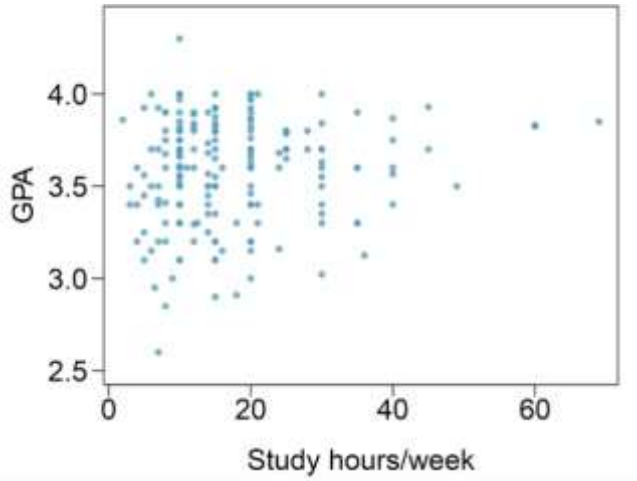
d. In an observational study, more than 50 people who previously worked for an international organization applied for a position and were asked to get a letter of recommendation from their previous employers. It was observed that letters of recommendation containing more detail were much more persuasive and acceptable than those containing less detail.

Is there a confounding variable in this study? If the answer is positive, explain.

3- Which type of sampling is used for each of the following instances? Why?

a. Surveying every passenger in five flights on a particularly selected day at an airline company.
b. Putting names of all employees in a bowl and choosing one of the names without looking into the bowl.
c. Surveying 100 students by getting random samples of 25 seniors, 25 sophomores, 25 juniors, and 25 freshmen.
d. Taking the last phone number of each page in the registration book of a competition to announce the winners of the competition.

4- A survey was conducted on 193 Yale University undergraduate students who took Statistical Inference course in 2016. This survey asked the students about their GPA, which can range between 0 and 4 points, and number of hours they studied per week. The relationship between these two variables is shown in scatterplot below:

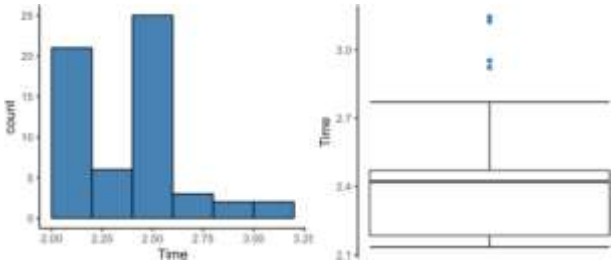


Homework 1

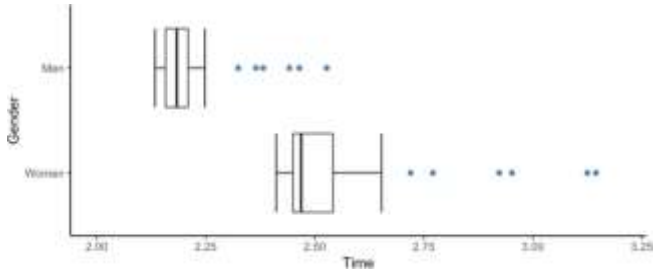
Statistical Inference

a. What is the explanatory variable and what is the response variable?
b. Describe the relationship between these two variables. Make sure to discuss unusual observations, if any.
c. Is this an experimental or an observational study?
d. Can we conclude that studying longer hours leads to higher GPAs?

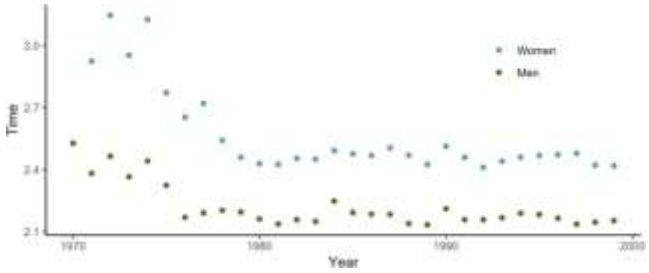
5- The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



a. What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?
b. What may be the reason for the bimodal distribution? Explain.
c. Compare the distribution of marathon times for men and women based on the box plot shown below.



d. The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



Homework 1

Statistical Inference

6- (R) Test scores for a college statistics class held during the day are:

99,56,78,55.5,32,90,80,81,56,59,45,77,84.5,84,70,72,68,32,79,90

Test scores for a college statistics class held during the evening are:

98,78,68,83,81,89,88,76,65,45,98,90,80,84.5,85,79,78,98,90,79,81,25.5

a. Find the smallest and largest values, the median, and the first and third quartile for the day class and night class.
b. Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.
c. Should we always remove outliers from datasets?
d. Create a box plot for each set of data. Use one number line for both box plots(in a single plot).
    i. For each set, based on the plot, would you expect the mean of the values to besmaller or larger than the median? Explain your reasoning.
    ii. Which box plot has the widest spread for the middle 50% of the data(the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

7- (R) In this Question, you should analyze the built-in diamond dataset. Your plots must have a proper title, x-label and y-label. Also, **do not** use any non-built-in R packages, e.g., ggplot2, etc. Follow the instructions:

a. Identify the variables and their types.
b. Use an appropriate diagram to visualize the number of diamonds in each clarity groups.
c. Plot a histogram of the distribution of 'price' and discuss its skewness.
d. Use side-by-side boxplots to display the distribution price along with 'clarity'. Then identify outliers for each group.
e. Plot a pie chart that visualizes the frequency of each color of diamonds. Each category must have a percentage and should have a unique color. Draw a legend for your pie chart.
f. Use a scatter plot to determine the relationship between 'depth' and 'price'. Interpret your plot.