Homework 1

Statistical Inference, Fall 2021

1- Answer the following questions for each section:

- Explain the type of study (Observational or Experimental).
- Identify explanatory variables.
- Identify response variable.
- Comment on whether or not the findings of the study can be used to establish causal relationships.

a. In a study published in the July issue of Psychosomatic Medicine, Joseph Boscarino examined the prevalence of heart disease and PTSD in more than 4,000 Vietnam veterans. The more severe the PTSD diagnosis, the greater the likelihood of death from heart disease, the study showed.

b. In order to assess the effectiveness of 3 different types of diets on weight loss, in a 2-year trial, researchers randomly assigned 322 moderately obese subjects to one of three diets: low-fat, Mediterranean, low-carbohydrate. Results show that the Mediterranean and the low-carbohydrate diets may be effective alternatives to low-fat diets.

c. Educational psychologists investigate the impact of different types of instruction on learning. In one study, researchers taught a math lesson to 9th-graders using the "Inventing to Prepare for Learning (IPL)" instructional cycle. The second group of students received traditional "tell and practice" instruction. After the lessons, both groups studied a worked example of a math problem on their own. Then they took a test that included problems like the worked example. The journal Cognition and Instruction published the results in 2004.

2- In each of the following results, is there a potential confounding variable? If the answer is yes, find the confounding factor and explain.

a. It is known that murder rates and ice cream sales are highly positively correlated throughout the year, That is, as murder rates rise, so does the sale of ice cream.

b. A study claims that living next to high-voltage transmission lines causes cancer.

3- Define the sampling method and explain your response:

a. A hospital wants to survey religious participants in their city about what they seek from a hospital chaplain, so they randomly select 5 religious meetings in the city and survey every participant in those meetings.

b. A factory manager takes an alphabetized list of workers' names and picks a random starting point. Every 20th worker is selected to take a survey.

c. A research team is seeking opinions about the Covid-19 vaccine amongst various age groups. Instead of collecting feedback from 326,044,985 U.S citizens, random samples of around 10000 can be selected for research. These 10000 citizens can be divided into age groups of 18-29, 30-39, 40-49, 50-59, and 60 and above.

d. A school principal tends to conduct a survey concerning personal hygiene. He has the list of all students in a table in Excel. He generates random numbers using Excel and asks corresponding students to answer the survey.

4- Many people believe that the Telegram Messenger is superior to the WhatsApp because it is much more user-friendly. Experimenters decided to test this claim. They take a sample of 52 social media users, each claiming to have experience (each person has a different level of expertise) with both applications. They give them a series of tasks to do on each application and record the amount of time it takes to complete the tasks in total.
   a. Describe how you might design an experiment for this purpose.
   b. Does your experimental design use blocking? Explain why you did or did not include blocking in your design.

5- For each of the following parts, explain (use more than three words) the most concerning the potential source of bias (If any).
   a. Rasul wants to use Mahan Airlines to travel around the world, but he is worried about his life, so he decides to interview a passengers who have used Mahan in last month about the safety of Mahan planes. According to another survey, 90% of Mahan customers have a cell phone and 91% of them answer the unknown phone number. So, Rasul finds their cell phone number and asks them whether their travel with Mahan was safe or not. Finally, he finds that all the answers are 'Yes, my travel was safe!'.

   b. A psychology professor wants to study the popularity of meat-based foods amongst undergraduate students at her university. She sends out a survey to everyone enrolled in Introduction to Psychology courses at her university. They all complete it in exchange for course credits.

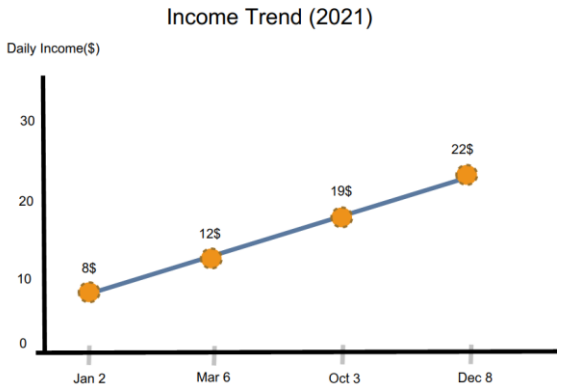
   c. Coca-Cola company wants to make an advertisement video based on a real survey. So, the company employs a marketing person and asks him to do a survey about Coca-Cola's popularity. The marketing person prepares two groups of cola cups. One is filled with Coca-Cola and the other with Pepsi-Cola. He labels Coca-Cola and Pepsi-Cola cups with 'A' and 'B', respectively. Then he takes a random sample of people and gives them a cola cup with the label 'A' and a cola cup with the label 'B' and asks them which cola cup they prefer.

d. A social scientist selects a random sample of girls walking in Naser-Khosro street in the morning, asking their name, last name and if they have experienced sexual abuse. Of the 20 samples surveyed, 2 say ‘yes’ and the others say ‘no’. The scientist uses data and concludes that 90% of girls walking in Naser-Khosro in the morning have not experienced sexual abuse.

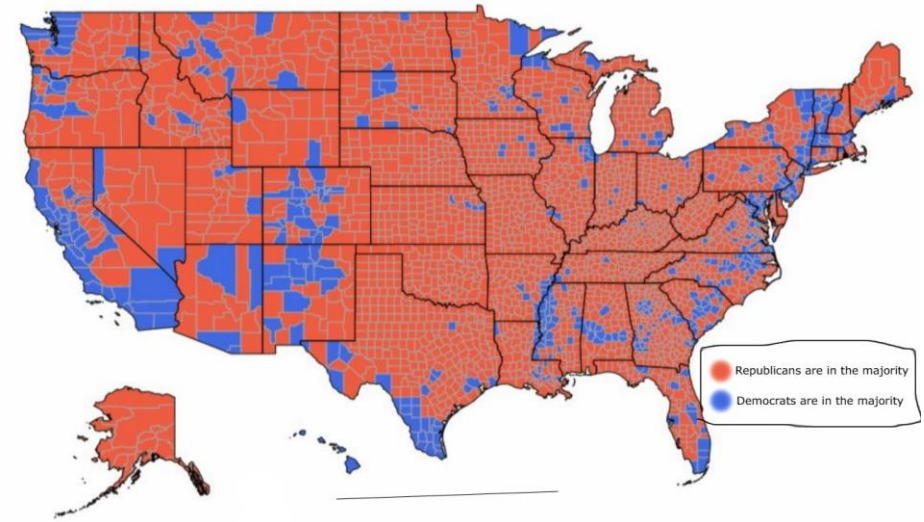
e. According to StatCounter, more than 72% of social media traffic in Iran is dedicated to the Instagram app. Snapp company wants to gauge its popularity amongst Iranian people. So, the company decides to send a survey as a direct message to every Iranian Instagram user. The company also gifts participants a free travel. Reports show that 90% of participants are satisfied with Snapp.

6- Determine if the following statements are true or false based on the corresponding diagram. If false, explain your reasoning.

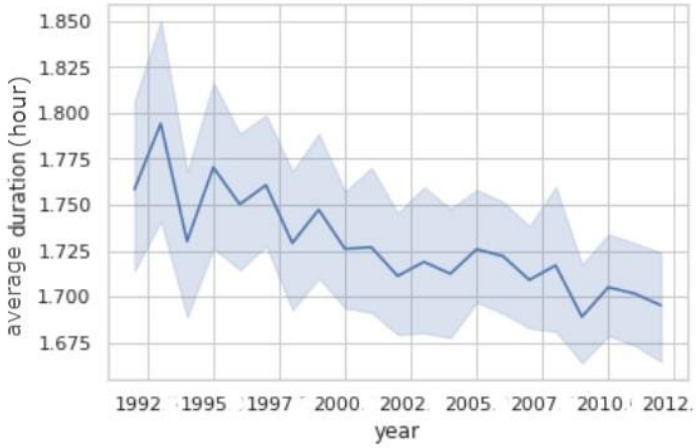
a. An employee’s daily income increases linearly over time in 2021 (from Jun 2 to Dec 8).



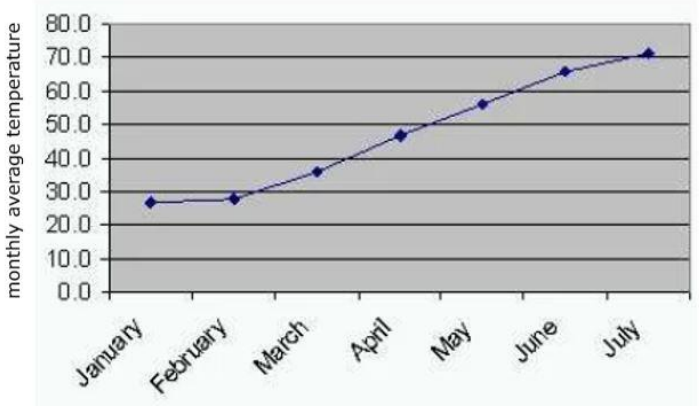
b. The following map is the electoral vote map of America. According to this, most of the Americans vote for republicans.



c. The average duration of IMDB movies decreases strongly over time.



d. According to the monthly average temperature trend of the U.S., we must take the global warming problem seriously.

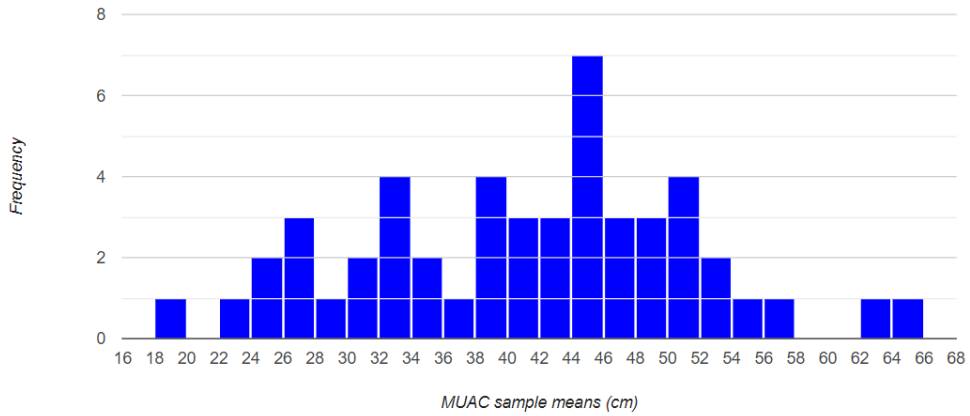


7- Researchers claim that the average mid-upper arm circumference (MUAC) of people who smoke cigarettes is 24cm. Rasul (that all of his friends smoke cigarettes) thinks researchers are wrong because his friends have much bigger arms. So, he selected a random sample of 15 persons from his friends and gauged their MUACs. The mean MUAC of his randomly selected friends was 37cm. He decided to perform a simulation to validate his findings. For this purpose, he repeatedly took a random sample of size 15 from smokers and gauged their MUACs. He did this 50 times and recorded the mean of the MUACs in each sample and prepared the following histogram:

a. Explain the appropriate null hypothesis and alternative hypothesis for his significance test in terms of words.
b. Based on the simulation results, what is the approximate p-value of the test?
c. Interpret the p-value?

Homework 1

Statistical Inference, Fall 2021



8- **(R)** Below are the final exam scores of twenty introductory statistics students.

57, 66, 72, 78, 79, 79, 81, 81, 82, 83, 84, 87, 88, 88, 89, 90, 91, 92, 94, 95

a. Create a vector of scores.
b. Calculate **median**, **mode**, **variance** and **standard deviation** of scores.
c. Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.
d. Plot the boxplot.
e. Plot the histogram and the density of scores in a single plot.
    i. Based on the plots, discuss the skewness of scores.
    ii. Based on the plots, would you expect the mean of this dataset to be smaller or larger than the median? Explain your reasoning.
    iii. What is the best measurement of the center for the scores? Why?

9- **(R)** In this part, you are going to study and analyze the IMDB movies 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 movies produced yearly.
c. Plot a histogram of the distribution of 'USA_gross_income' and discuss its skewness.
d. Use side-by-side boxplots to display the distribution of movie durations along with 'tomatometer_status'. Then identify outliers for each group.
e. Categorize all the movies based on their durations into 4 groups: "very long" (>200), "long" (>150), "standard" (>100), and "short" (<=80). Plot a pie chart that visualizes the frequency of these five categories. 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 'USA_gross_income' and 'worldwide_gross_income'. Interpret your plot.