Homework 1

Statistical Inference, Spring 1400

1- Answer the following questions for each section:

I. Define the type of the study.

II. Identify explanatory variables.

III. Identify response variable.

IV. Can the result of the study be used to establish causal relationships?

a. According to a study published in the journal *Psychological Medicine*, ayahuasca—a drink made of Amazonian plants—can help people with depression. The study involved 29 participants who all had depression. Doctors gave each of the participants either ayahuasca or a placebo beverage that looked and tasted similar. Those who drank ayahuasca showed significantly lower levels of depression compared to those whom drank the placebo beverage.

b. According to a study published in the journal *Social Psychological and Personality Science*, people with religious affiliations live longer, on average, than those without religious affiliations. The study analyzed more than 1000 obituaries and accounted for other variables like sex and marital status.

c. A recent study showed that viral medications might be useful for fighting certain bacterial infections. Participants with gastrointestinal issues were assigned to either a placebo or viral medication for the first four weeks of the study, followed by an additional four weeks on the other treatment. Their symptoms showed significantly more improvement from the viral medication than the placebo.

d. A study published in the journal *Environmental Health* found that flight attendants get cancer more often than the general population. The 5366 flight attendants in the study had higher incidence rates for every type of cancer examined, even though they had relatively lower rates of smoking and obesity.

e. Employees at a video streaming website suspect that users on desktop computers spend more time on their websites than users on mobile devices. They plan on taking a large random sample of users to see if their suspicions are correct.

f. Maryam has developed an exercise app where users log their workouts. She wonders if sending users a daily notification that reminds them to use the app will result in users logging more workouts. Subsequently, she released an update to her app so that half of the users will have daily notifications and the other half won't have daily notifications (randomly determined for each user). After a few weeks, she would like to see if one group logs more workouts on average than the other group.

Homework 1

Statistical Inference, Spring 1400

2- Define the sampling method and explain your response:

a. A manager associated each employee's name with a number on one ball in a container, then drew balls without looking to select a sample of 5 employees.

b. A truck manufacturer selects 33 trucks at random from each of 66 models for safety testing.

c. Security workers at an airport randomly choose one of the first 50 people to pass through a checkpoint for extra security screening. After that person, they choose the 50th person who passes through for extra screening as well.

d. A principal orders t-shirts and wants to check some of them to make sure they were printed properly. She randomly selects 2 out of the 10 boxes and checks every shirt in those 2 boxes.

e. Each student at a school has a student identification number. Counselors have a computer generate 50 random identification numbers, and the students associated with those numbers are asked to take a survey.

f. A student council surveys 100 students by taking random samples of 25 freshmen, 25 sophomores, 25 juniors, and 25 seniors.

g. A large bakery mass produces cakes on an assembly line. Each shift, a quality control expert randomly selects one of the first ten finished cakes, and every tenth cake thereafter. Employees weigh those cakes and give the cakes a detailed visual check.

h. A school chooses 3 randomly selected athletes from each of its sports teams to participate in a survey about athletics at the school.

3- Identifying the population and sample. Is there any problem with choosing the sample this way?

a. Administrators at Riverview High School surveyed 10 volunteers of their seniors to see how seniors at the school felt about the lunch offering at the school's cafeteria.

b. A city council member wanted to know how her constituents felt about a planned rezoning. She randomly selected 75 names from the city phone directory and conducted a phone survey. About 40% of these names answer the phone.

c. The state Department of Transportation wants to know about out-of-state vehicles that pass over a toll bridge with several lanes. A camera installed over one lane of the bridge photographs the first tenth vehicle's license plate that passes through that lane.

d. Ali wants to know whether the food he serves in his restaurant is within a safe range of temperatures. He randomly selects 70 entrees and measures their temperatures just before he serves them to his customers.

Homework 1

Statistical Inference, Spring 1400

4- Answer the following questions about the source of bias:

a. A polling firm wants to contact a random sample of people likely to vote in an upcoming nation-wide election. They will use a random digit dialer to generate and call phone numbers at random, so the poll will include people with landlines, unlisted numbers, and mobile phones. The random digit dialer skips invalid phone numbers. If a person doesn't answer a call, the dialer will try one more time, and then skip that number. When a person does answer, a pollster asks a set of initial questions—such as the person's age and whether they voted in previous elections—to see if they are likely to vote in the upcoming election. If they are eligible and likely to vote, the pollster will ask a series of questions about the election. **Which of the following is *not* a potential source of bias in their poll? Tell the type of bias for each item.**

I. Some people might refuse to share their personal information over the phone.
II. People who are likely to vote but don't have a telephone are excluded from the pool.
III. Some people will not answer calls from an unfamiliar caller.
IV. The pool excluded people too young to vote.
V. Some people might say they are registered and suggest they are likely to vote when they aren't.

b. A high school has a policy that students' phones must be kept away during class. A principal used the school roster for polling a random sample of 50 students, and only 10% said that they ever had their phone out during class. The next day, the principal observed classrooms and noticed that approximately 25% of students had their phones out at some point during the class. **Explain the most concerning the potential source of bias in the principal's poll?**

c. An airline wants to survey customers about their overall satisfaction. They take a random sample of 1000 customers who have flown in the past month and email them a survey. The email also offers those who complete the survey a 25$ gift card that can be used almost anywhere. **Explain why nonresponse bias may occur?**

5- Answer the following questions and explain your reasons.

a. Research shows that the amount of coffee drunk is directly related to college students' grades. In other words, on average, the more coffee a student drink, the higher his or her score in the exam. **Does this experiment have a potential confounding variable? If the answer is yes, find the confounding factor and explain.**

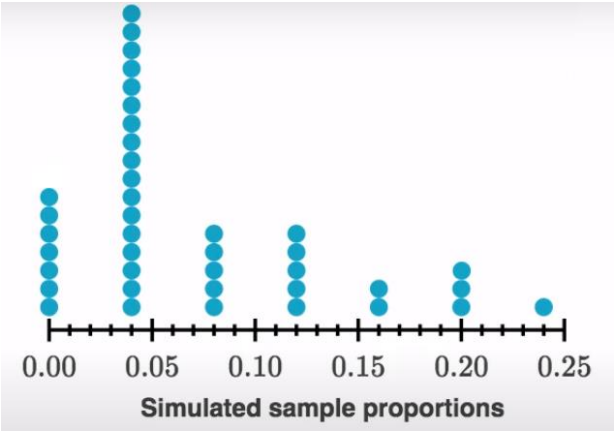
b. The sales team of a sunglasses company is trying to test their sunglasses' effectiveness by examining the annual sales of sunglasses.
In 2017, the company sold 3,000 sunglasses.
In 2018, they sold 2,500 sunglasses.

They assume that their sunglasses are less effective, which is why sales decreased. **Does this experiment have a potential confounding variable? if the answer is yes, find at least one confounding factor and explain.**

c. A study was done to compare the lung capacity of coal miners with the lung capacity of farmworkers. The researcher studied 400 workers of each type. Other factors that might affect lung capacity are smoking habits and exercise habits. The smoking habits of the two types of workers are similar, but the coal miners generally exercise less than the farmworkers. **What is the confounding variable in this study?**

d. Medical device manufacturers want to evaluate whether the new blood pressure device they produced performs better than the previous devices. For doing research, manufacturers assign one group of users to the new medical device completely random, and also keep the group members hydrated very well, and assign the other group to the old device. **How should confounding variables be controlled?**

6- A researcher read an article that said 6% of people exercise daily. But he thinks it's higher for teenagers. To test his theory, he took a random sample of 25 teenagers, and 20% of them were exercising daily. To see how likely a sample like this was to happen by random chance alone, the researcher performed a simulation. He simulated 40 samples of n=25 teenagers from a large population where 6% of the teenagers were exercising daily. He recorded the proportion of Those who was a daily exercise in each sample. Here are the sample proportions from his 40 samples:



a. For the dataset represented by the dot plot, find mean, median, mode, range, and the interquartile range.
b. Explain the appropriate null hypothesis and alternative hypothesis for their significance test in terms of words.
c. Based on these simulated results, what is the approximate p-value of the test?
d. What does this p-value say?

Homework 1

Statistical Inference, Spring 1400

7- Answer the questions in each of the following parts.

a. Medical researchers found that there is a correlation of 0.95 between smoking and lung cancer. **Determine if the following statements are true or false, and explain why.**

I. There is a positive linear association between the two variables.
II. There is a strong correlation between the two variables.
III. Smoking causes the increase of lung cancer.

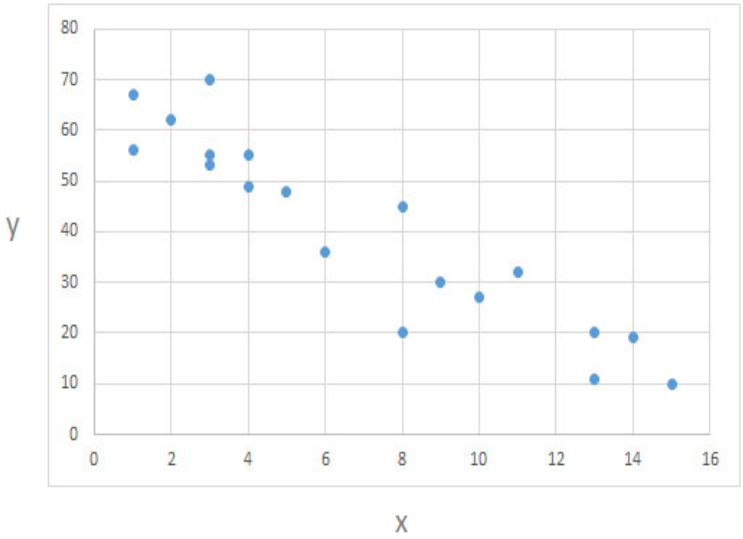
b. A study on the smartphone usage done by national researchers, and they found the following correlations:

- The correlation between the number of texts sent each day and a person's average credit card debt is 0.45.
- The correlation between the number of texts sent each day and the number of books read each month is −0.30**.**

Determine if the following statements are true or false, and explain why.

I. As the number of texts sent each day increases, average credit card debt increases.
II. Sending more texts causes people to read less.
III. A person's average credit card debt is related more strongly to the number of texts sent each day than the number of books read each month is related to the number of texts sent each day.

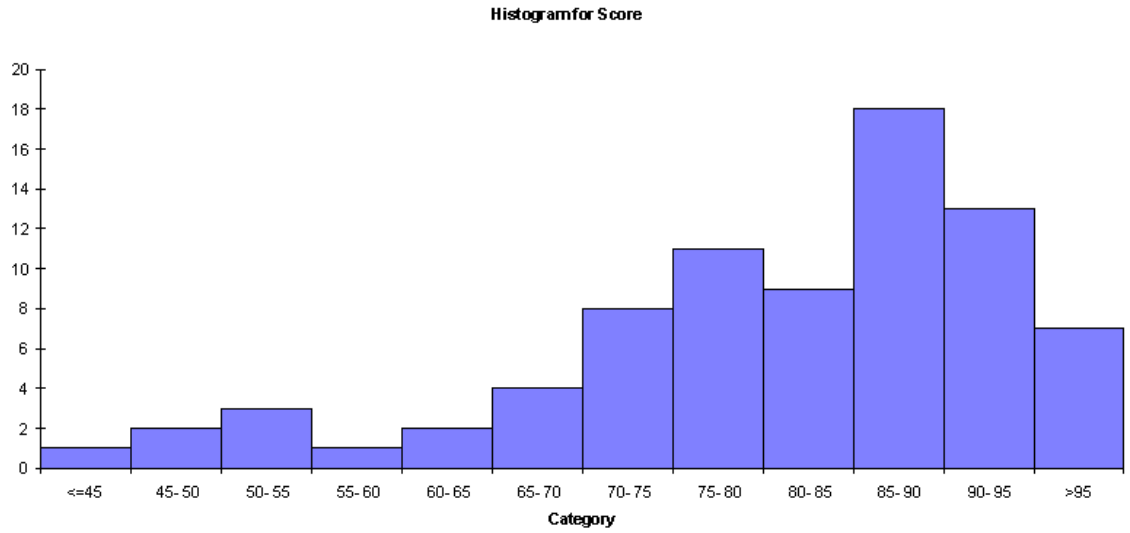
c. The data collected in a research contains data points with two parameters X and Y. A scatterplot of Y versus X is shown below.



i. What type of an association is apparent between X and Y?

ii. What type of an association would you expect to see if the axes of the plot were reversed, i.e., if we plot X versus Y?
iii. Are X and Y independent? Explain your reasoning.

8- A dataset is shown below with a histogram.



a. What is the best measurement of the center for this dataset? Why?

b. What is the best measurement of spread? Explain your reason.

9- **(R)** There exist 5 types of positions in a software company: "UI Developers", "Back-end Developers", "management", "HR" and "HSE". Each type has 8, 12, 4, 3, 3 employees in order. Answer the following questions about this company:

a. Create two vectors containing the position types and their corresponding population.

b. Plot a bar chart of distributions of employees in different position types. Note that the plot must have a proper title in green color, and the x-label and the y-label should be in blue.

c. Consider each position type has below salaries. Visualize a plot with 5 boxplots that shows salaries of different groups of position types.

UI Developers → 75000, 25000, 48000, 42000, 35200, 45000, 23000, 45500

Back-end Developers→ 20000, 80000, 36000, 46300, 41000, 43000, 22000, 37000, 39000, 43500, 69000, 5000

Management→ 80000, 67000, 56000, 82000

HR→ 45000, 39000, 30000

HSE→ 12000, 25000, 31500

d. According to part c, find the exact quartile of the salary of each group and calculate IQR. Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.

e. Discuss the skewness of distributions in each group salary. Then plot histogram and density plot of each group in 5 plots (both histogram and density plot must be in a single plot).

f. Categorize all employees based on their salary into 5 groups: “very high” (>50000), “high” (>40000), “middle” (>30000), “low” (>20000), and “very low” (<=20000). 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.

g. For back-end developers group calculate mean, median, variance, and standard derivations.