Homework 1 Solution

Statistical Inference, Spring 1400

1- Question 1

a. Type→ experimental study
Explanatory→the type of drink
response→depresion level
causal relationships→Yes

b. Type→ observational study
Explanatory→ religious affiliations
response→lifespan
causal relationships→No

c. Type→ experimental study
Explanatory→the type of treatment (viral or placebo)
response→symptom improvement
causal relationships→Yes

d. Type→ observational study
Explanatory→ the type of person(flight attendant or not)
response→incidence rate of cancer
causal relationships→No

e. Type→ observational study
Explanatory→the type of device
response→time spent on the website
causal relationships→No

f. Type→ experimental study
Explanatory→notification presence
response→the number of workouts
causal relationships→Yes

2-

a. Simple random sample

b. Stratified random sample
c. Systematic random sample
d. Cluster random sample
e. Simple random sample
f. Stratified random sample
g. Systematic random sample
h. Stratified random sample

3-

a. The population is all seniors at Riverview High;
 the sample is the 100 seniors surveyed.
 Voluntary bias

b. The population is everyone listed in the city phone directory;
 the sample is the 75 people selected.
 Non-response bias

c. The population is all of the vehicles that pass through the lane with the camera;
 the sample is the group of every tenth vehicle that passes through the lane.
 Convenient bias

d. The population is all of the entrees Lucio serves;
 the sample is the 70 selected entrees.
 Convenient bias or Anecdotal Evidence

4-

a.

I. True → this sort **of nonresponse** could potentially lead to biased results

II. True→ this poll excludes people who don’t have access to telephones(an example of undercoverage) so this a source of bias **non-response** and can be **convenient bias**

III. True → this sort of **nonresponse** could potentially lead to biased results

IV. False→ since the goal of the poll is to reach potential voters there’s no reason to include people who are too young to vote

V. True→False answers like these might lead to biased results(an example of **Response Bias or any bias about giving false answers**)

b. Response Bias occurs when people systematically give wrong answers. in this context, students who have their phone out during class might not want to admit that to the principal since they are breaking school policy

c. Satisfied customers might be less likely to complete the survey than dissatisfied customers. This would be nonresponse, which is when people of interest can’t be reached or refuse to participate. If satisfied customers opt out, then the survey results might suggest that customers are less satisfied overall than they really are.

5-

a. Yes. Amount of sleep a student gets each night
The only confounding variable in this experiment is the amount of sleep that each student gets. A confounding variable is one that has an impact on both the dependent and independent variables. It is possible that the amount of sleep a student gets is related to caffeine intake, which in turn affects the grade a student receives on a test or assignment.

b.

Yes. each of the following three can be a confounding variable:

1)An increase in prices may have led to decreased sales.
2) If the same people live in the city during 2017 and 2018, people may already have sunglasses in 2017 and might not need to buy them.
3) There may have been less sunny days in 2017 than in 2018, therefore decreasing the need for sunglasses.
Any of these answers could explain why sunglasses sales dropped. You cannot assume any specific cause explains a change in data like this. further experimentation should be done rather than assuming cause and effect.

c. Exercise is a confounding variable in this study Because exercise affects lung capacity and more farmworker's lung capacity may be affected by exercise.

d. The group receiving the old device should also be required to stay hydrated.
When comparing the effectiveness of a treatment, one should try to ensure that only the treatment varies across groups. In this case, the new device is compared to an old device. However, the new device also requires that users stay well hydrated. If we observe any positive effects from the new device, we won't know whether the new device is effective, or if merely staying well hydrated is actually what is effective. To rule out this confounding variable, we should also ask the group using the old machine condition to stay hydrated as well.

6-

Mean = (0.0 *7 + 0.04*17 + 0.08*5 + 0.12*5 + 0.16*2 + 0.2*3 + 0.24) / 40 = 0.071
Mode = 0.04
Median location : n+1/2= 41/2=20.5
Median = 20.5th data point = (20th data point + 21th data point)/2 = (0.04 + 0.04) /2 = 0.04
Range = Max – min = 0.24 – 0 = 0.24
interquartile range:
Q1 location : (20+1)/2 = 10.5
Q1=10.5th data point = (0.04+0.04)/2=0.04
Q3 location :20+ (20+1)/2 = 30.5
Q3=10.5th data point = 30th data point + 31th data point = (0.12+0.12)/2 = 0.12
interquartile range = 0.12 - 0.04 = 0.08

b. H_0: $p_{\text{teenagers}} = 6\%$ H_a: $p_{\text{teenagers}} > 6\%$

there is a null hypothesis that the proportion of people that are teenager and are exercising daily, would be as the same as the proportion of people who exercise daily as a whole, that would be 6%.

So his alternative hypothesis would be the proportion, the true population parameter for teenagers, is greater than 6%.

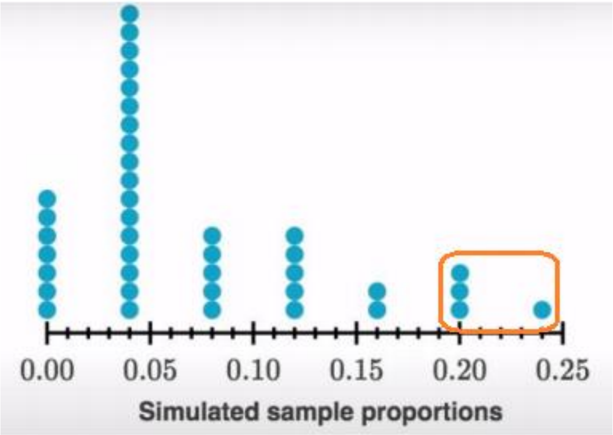
c. p-value ≈ 0.1

The researcher found that 20% of her sample of 40 students were vegetarians.

Since his alternative hypothesis is Ha:p>6%, we can find the approximate p-value of this result by looking at how often a sample result as high or higher than 20% occurred in the simulation.

The simulation produced a sample proportion at or above 20% in 4 out of 40 samples.

So p-value= 4/40 = 0.1



d. This p-value says that when we take a sample of 25 teenagers from a large population where 6% of them are exercising daily, there is about a 10% chance that 20% or more of the teenagers in the sample are exercising daily.

(so to see whether or not could reject the null hypothesis he took a random sample of 25 teenagers and you calculate the sample proportion and then you figure out what is the probability of getting a sample proportion this high or greater and if it's lower than a threshold then you will reject your null hypothesis and that probability we call the p-value the p-value is equal to the probability that your sample proportion is going to be greater than or equal to 20% if you assumed that your null hypothesis was true.)

7-

a. i) False

ii) True

iii) False

Since the correlation is positive, there must be a positive association between the two variables but we don’t know anything about the type of this association and therefore statement i is incorrect. Statement ii is correct since a correlation of 0.8 to 1.0 on an absolute value scale of 0 to 1.0 is considered to be a strong correlation. Statement iii is incorrect since correlation does not mean causation.

b. i) True

ii) False

iii) True

i is correct because there is a positive correlation between the number of texts sent each day and average credit card debt.

ii is incorrect because the word "cause" was used in the statement. Correlation does not mean causation. There is a relationship between the number of texts sent each day and the number of books that a person reads each month. However, the number of texts sent each day does not cause a person to read a certain number of books each month.

iii is correct because the absolute values of the correlations indicate which correlation is stronger. 0.45 is a stronger correlation than -0.3.

c. i) negative linear relationship

The data points follow an overall linear trend, as opposed to being randomly distributed. Though there are a few outliers, there is a general relationship between the two variables. A line could accurately predict the trend of the data points, suggesting there is a linear correlation. Since the y-values decrease as the x-values increase, the correlation must be negative. We can see that a line connecting the upper-most and lower-most points would have a negative slope. An exponential relationship would be curved, rather than straight

ii) Association would still be negative.

iii) No, they are not independent. As explained in part (i) they are negatively associated.

8-

a. **Median**. The median is the value in the center of the data. Half of the values are less than the median and half of the values are more than the median. It is probably the best measure of center to use in a skewed distribution.

b. **Interquartile Range**. For skewed distributions, the outliers can greatly affect the value of the mean and the standard deviation. Therefore, the IQR would be a better measure of spread because it is not greatly influenced by outlier values.

Choosing a measure for the center and spread, depends on what kind of distribution it is, whether it's symmetric or if it's skewed with outliers. If the distribution is symmetric then the mean can be used to find the center. When it is skewed right or left with high or low outliers then the median is better to use to find the center. The best measure of spread when the median is the center is the IQR. As for when the center is the mean, then standard deviation should be used since it measures the distance between a data point and the mean