Homework 1 (Solutions)

Statistical Inference, Fall 1401

1)(10)

1. Randomly select 30 students from the class.
2. Randomly assign 10 students to each group (no music, music without lyrics, music with lyrics)
3. Give all 30 students an exam to assess their baseline knowledge.
4. Tell the 30 students to study the material on the exam for 2 hours. Provide the same study materials to all students, and provide all students listening to music the same music within each group. Make sure the environments all the students are studying in are as similar as possible.
5. After the 2 hours, give all students another exam on the same material.
6. Compare the score differences between all three groups.

2)(15 - each 3.75)

a.

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 fewer 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.

b.

The figure shows that at any given maternal age, the birth order has little, if any, effect; the frequency of Down syndrome is low in young moms regardless of birth order, and the frequency of Down syndrome is high in older moms regardless of birth order. In other words, if we control for maternal age, birth order is *not* associated with increased prevalence of Down syndrome; it is not an independent risk factor. However, within each stratum of birth order, prevalence increases with maternal age, meaning that, controlling for birth order, the strong association with maternal age persists. Given our definition of confounding, the effect of birth order was confounded by maternal age, since maternal age made it appear that there was an association with birth order. However, when stratified by both birth order and maternal age, we can see that birth order did not have an independent effect. The apparent association with birth order was totally the result of confounding and overestimation caused by maternal age.

c.

In this experiment, the fact that the students are willingly staying an extra hour at school is not taken into account, and this is the confounding variable

Confounding can invalidate conclusions of the study. Because it's probable that students who used the software, had higher scores on the exams even without using it. To eliminate the effect of the confounding variable we can change the experiment. We can group students by their scores. For example we can group them in "Excellent", "Good" and "Weak" groups. Then we can ask half of each group to use the software and half of them not to use it. Now the scores of the exam can show us the effect of using software on the scores of the students.

d.

The explanatory variable is details of letter of recommendation and the response variable is its persuasiveness. The confounding variable can be number of words. The more number of more of are used, the more letter of motivation contains details. Also, if few number of words is used, the letter becomes less valued and it is considered as a prepared template, and it will not be as persuasive as needed. We can alter the study to examine the relationship between numbers of words used in the study and level of the letter's persuasiveness. Also, a study found that "women are significantly less likely to receive excellent recommendation letters than their male counterparts at a critical juncture in their career and the company may want to hire men for the position. Other correct answers are accepted as well.

3)(10 - each 2.5)

a. Cluster random sampling; The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random

b. Simple Random sampling; every member and set of members has an equal chance of being included in the sample.

c. Stratified random sampling; the population is first split into groups. The overall sample consists of some members from every group. The members from each group are chosen randomly.

d. Convenience sampling / Not Random

Homework 1 (Solutions)

Statistical Inference, Fall 1401

4)(10 – each 2.5)

a. The explanatory variable is study hours per week and the response variable is the GPA.
b. There seems to be a positive association, that is, that the more study hours, the higher the GPA. Some unusual observations are: there is a person with a GPA above 4, but 4 should be the limit. There are two people that claimed to study more than 60 hours per week.
c. It is an observational study.
d. We can't because it is an observational study, we can just point to an association (or correlation).

5)(10 – each 2.5)

a. In the histogram, the two modes are more visible. In the boxplot, the outliers and median value are more apparent.
b. Men and women on average have very different finishing times. One mode is for men while the other is for women.
c. The women's finishing times are higher and more variable than the men's finishing times.
d. The women's time is always higher than the men's time, and they are both decreasing over time. They decrease rapidly at first, then at a slower pace over time.

6)(10 – each 3.75)

a. Whether the patients were cured or not and gender appear to be dependent, as the proportion of women who were cured is higher than the proportion of men who were cured.
b. **Null hypothesis**: gender and being cured by the treatment are independent, and observed difference in number of men and women who are cured is due to chance.
c. **Alternative hypothesis**: gender and being cured by the treatment are dependent, and observed difference in number of men and women who are cured is not due to chance.
d. We write *male* on 35 cards and *female* on 25 card. Then we shuffle these cards and split them into two groups: *cured* group of size 28 representing the patients who were cured after being under treatment and *not cured* group of size 32 representing the patients who weren't cured after being under treatment. We calculate the difference between proportion of men who were cured and the proportion of women who were cured and record this value. We repeat this procedure 100 times to build a distribution centered at zero. Finally, we calculate the fractions of simulation in which the calculated differences are ≥ 0.2971 or ≤ -0.2971. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.
e. rejected in favor of the alternative Observed difference between proportions (female-male) = 0.2971. so we have to count the simulations in which the difference between proportions are equal or greater than 0.24. As we can see in the figure, three points are placed after 0.2971 or

before - 0.2971, so p-value is 0.01 and we reject the null hypothesis in favor of the alternative hypothesis.

7)(15)

```
outlier <- function(scores){
  Q1 = summary(scores)[["1st Qu."]]
  Q3 = summary(scores)[["3rd Qu."]]
  IQR = Q3 - Q1
  upper_whisker = Q3 + 1.5* IQR
  lower_whisker = Q1 - 1.5* IQR
  outliers <- scores [scores > upper_whisker | scores < lower_whisker]
  if (length(outliers) == 0)
    return ("none")
  return (outliers)
}
day = c(99,56,78,55.5,32,90,80,81,56,59,45,77,84.5,84,70,72,68,32,79,90)
night = c(98,78,68,83,81,89,88,76,65,45,98,90,80,84.5,85,79,78,98,90,79,81,25.5)
print("day")
print(summary(day))
print("night")
print(summary(night))
boxplot(day, night,
        main = "Multiple boxplots for comparision",
        names = c("Day", "Night"),
        xlab = "Score",
        col = c("orange","red"),
        border = "brown",
        horizontal = TRUE,
        notch = FALSE
)
print("day")
print(outlier(day))

print("night")
print(outlier(night))

segments(x0 = mean(day), y0 = 0.61,
         x1 = mean(day), y1 = 1.4,
         col = "blue", lwd = 3,lty=2)

segments(x0 = mean(night), y0 = 1.61,
         x1 = mean(night), y1 = 2.4,
         col = "blue", lwd = 3,lty=2)
```

a.(2)

```
print(summary(day))
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
32.00   56.00   74.50   69.40   81.75   99.00

print(summary(night))
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
25.50   78.00   81.00   79.05   88.75   98.00
```

b.(5)

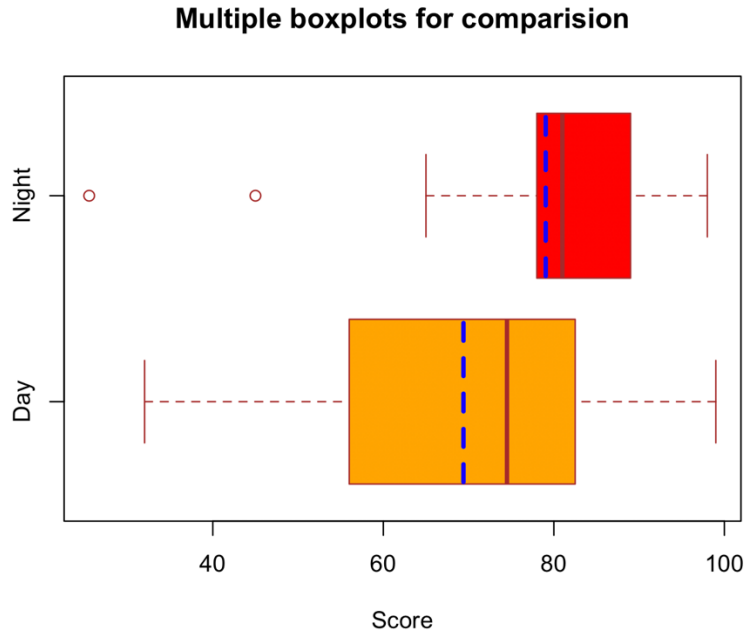
```
> print(outlier(day))
[1] "none"
> print(outlier(night))
[1] 45.0 25.5
```

The calculations are shown in the outlier function

c.(2)

no, they may have important information about data that can lead us to interesting results. Therefore, erasing them may cause the loss of valuable information.

d.(6 – plot 2,di 2, dii 2)



i. left skew(data has outliers,if you answer(correctly) the question with removing outlier no marks have been reduced) => mean < median

ii. The first data set has the wider spread for the middle 50% of the data. The IQR for the first data set is greater than the IQR for the second set. This means that there is more variability in the middle 50% of the first data set.