Mohammad Javad Ranjbar

810101173

Homework 1

Statistical Inference, Fall 2022

Question #1:

First off, the teacher needs to pick random samples. In this case, we use the Stratified sampling technique. We divide the population into groups based on the student's records(grades) and, we sample from within each stratum.

It is suspected that females are better at multitasking, which might affect the result, so we block for gender.

In this experiment, there are 2 explanatory variables: music and lyrics, 1 blocking variable: gender, and 1 response variable: exam performance.

The teacher needs to divide students into three groups. It should be mentioned that female and male students should be assigned equally to each group. The groups are as followed:

1- Control group: Students that study without music (Half female and half male).
2- Experimental (treatment) group: which consists of two groups:
    a. Students that study music without lyrics (Half female and half male).
    b. Students that study music with lyrics (Half female and half male).

Finally, the teacher analyzed each group's performance on the same exam.

Question #2:

a) There are several confounding variables for this exterminate such as:
    a. These glasses were accessible to a Limited number of people and, they bought their glasses in 2017. Therefore, they do not need to buy new glasses in 2018. Hence, if this company does not increase its number of branches around the country or globe, annual sales will decrease each year.
    b. The number of sunny days in each year varies. In 2018, there might have been more cloudy days. Therefore, people might feel less need to buy sunglass.
    c. There might be a new company that represents better deals for glasses and people got more inclined to buy glasses from them.

    All the reasons above can be considered confounding variables.

b) Based on the provided figure it is obvious that for each maternal age, the order of birth has almost no effect. In other words, the risk of having a child with down syndrome is low for parents at a young age regardless of birth order. If we analyze each stratum of the birth order, it is shown that the association for maternal age persists. Hence, we realize that the effect of birth order was cofounded 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 confounding caused by maternal age.
c) Several factors can be cofounding viable of this experiment

a. Students that study at school, normally work in groups. Therefore, each student helps their groupmates in any aspect that they lack. This kind of teamwork might help students to obtain higher scores regardless of using the software. Hence, the association with higher scores with the software was confounded by students studying together. To control this variable, the teacher needs to ask students to don't work together and work alone.
b. Practicing the topics immediately after learning that topic helps students to learn and remember those topics better. This could be a confounding variable. To eliminate this, the teacher needs to ask students to study subjects that are not related to today's curriculum.
c. In the question, it is not clear whether the teacher is present during that session or not. If the teacher is there to help them, this might accelerate students to learn any subject regardless of the software. To eliminate this confounding variable (Teacher helping them), the teacher needs to refuse to answer their questions.

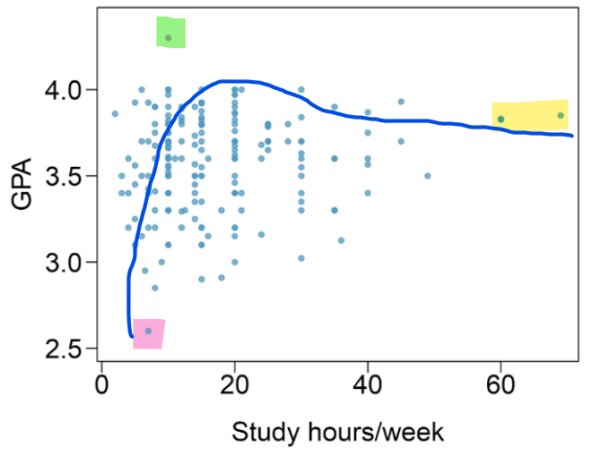
d) employees with higher abilities in their jobs, make their employers happier and more satisfied of their work. As a result, they will write a more detailed and better letter of recommendation. In addition, when these employees have lots of expertise, there are more things to explain in their letters of recommendation. This means, these employees were accepted because of their skills and the detailed recommendation letter was written because they were experts. Hence, their skillfulness might be the confounding variable in this experiment.

Question #3:

a) Multistage sampling: We divide the population based on the days of the week, then we choose one day of the week. For that day, we divide the population based on different flights(clusters), choose five flights, and survey all the passengers.
b) Simple Random Sampling (SRS): Each employee has the same chance of being selected.
c) Stratified sampling: The population is divided into homogenous strata (seniors, sophomores, juniors, and freshmen), then randomly sample from within each stratum.
d) Systematic sample: In this instance, elements of a sample are chosen at regular intervals of population. We choose the last phone number on the first page and with an interval of one page, we choose the next number.

Question #4:

a) In this experiment, there is 1 explanatory variable: study hours per week, and 1 response variable: GPA.
b) There is a positive, curved, and weak relationship between study hours/week and GPA. The study hours and GPA do not have a strong correlation. However, it seems like with 25 hours of study you get the best result.



Several students study way above average (Yellow dots) and don't get the best result. This case might be caused by getting mentally tired after long hours of study. Therefore, they don't perform well in exams. Some students have worked a little less than average and got a very bad GPA (Pink dot) this might be due to having work or problems outside the university (family problems, etc.) which hinders them to focus well on their studies and adds to their stress which leads to a bad performance in exams. Based on the figure, studying around 25 hours get students the best result. There is a student, (Green dot) that has a GPA over 4 which should be a mistake during data gathering.

c) This is an observational study; because we collect data in a way that does not directly interfere with how the data arise.
d) No, we can't conclude this statement. Based on the figure, people with average study hours have higher GPAs, and the relationship between study hours and GPA in longer hours gets negative. This means, longer hours lead to a worse GPA.

Question #5:

In the histogram we can perceive this information:

a. Boxplot does not show distribution.
b. Two modes in the histogram are shown. In other words, the histogram is bimodal.
c. Number of all the contestants from 1990 to 1999.
d. Exact number in each timespan(bins).

However, In the box plot, some information is provided that in the histogram it is not explicitly apparent such as outliers, median, IQR, min, max, and 25th and 75th percentile.

b) We have two peaks of data, which usually indicates you've got two different groups. This might be caused by the difference between women finishing time and men's finishing time. One mode is for men while the other is for women.
c) Men generally have shorter finishing times and their distribution is close to symmetric. However, Women's finishing times are higher and the distribution is right skewed.
d) Men's finishing times are all lower than women finishing times. The finishing time for both groups is decreasing each year. Differences between fishing times in the early years (1970-1980) with the finishing time after 1980 are pretty huge. In other words, the slope of decreasing in that period was rapid but became slower over time.

Question #6:

a) Based on the plot, it seems like gender has an impact on the treatment result. Because Women have a considerably better cure rate ($\frac{16}{25} = 0.64$) than men ($\frac{12}{35} = 0.34$). Therefore, it is not independent.

b) Null hypothesis (H_0):There is nothing going on. In other words, getting cured and gender are independent, observed difference in treatment is simply due to chance.

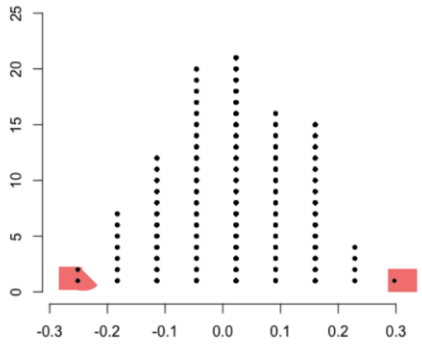
Alternative Hypothesis (H_A): There is something going on. In other words, getting cured and gender are dependent, the observed difference in treatment is not due to chance.

c) We consider H_0 true and conduct this simulation many times. We evaluate the probability of observing an outcome at least as extreme as the one observed in the original data and if this probability is low, reject the null hypothesis in favor of the alternative. In other words, If the results from the simulations look like the data. We can conclude the difference between the proportions of cured patients between males and females was not due to chance and therefore, treatment and gender are independent. Otherwise, it is not due to chance and they are dependent.

Gender	Cured	Not Cured	Total
Male	12	23	35
Female	16	9	25
total	28	32	60

$$Differnece = \frac{16}{25} - \frac{12}{35} = \frac{52}{175} \approx 0.3$$

d) Based on the below figure. The probability of observing a 0.3 difference is very low. Therefore, we reject the null thesis and interfere that treatment and gender are dependent.



Question #7:

a)

```
> dtest_score <- c(99,56,78,55.5,32,90,80,81,56,59,45,77,84.5,84,70,72,68,32,79,90)
> summary(dtest_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   56.00   74.50   69.40   81.75   99.00
> ntest_score <- 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)
> summary(ntest_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.50   78.00   81.00   79.05   88.75   98.00
```

b) Yes. The evening class has two outliers as followed. A potential outlier is defined as values outside of the range of threshold values for outlier detection. We use Interquartile Range to calculate this threshold. For the day class the calculation is as followed:

$IQR = \mathrm{Q}_3 - \mathrm{Q}_1 = 81.75 - 56 = 25.75$

$\mathrm{T}_{\min} = \mathrm{Q}_1 - (1.5 * \mathrm{IQR}) = 56 - 1.5 * 25.75 = 17.35$

$\mathrm{T}_{\max} = \mathrm{Q}_3 + (1.5 * \mathrm{IQR}) = 81.75 + 1.5 * 25.75 = 120.375$

$Rang_{\mathrm{upper}} = min(120.375, 99) = 99$

$Rang_{\mathrm{lower}} = max(17.35, 32) = 32$

Now, values that are outside of the range of $X < Rang_{\mathrm{lower}} | X > Rang_{\mathrm{upper}}$ are the outliers for the day class we do not have any outliers.

```
> IQR <- IQR(dtest_score)
> Tmin <- quantile(dtest_score, 0.25) -(1.5*IQR)
> Tmax <- quantile(dtest_score, 0.75)+(1.5*IQR)
> dtest_score[which(dtest_score < max(Tmin,min(dtest_score)) | dtest_score > min(Tmax,max(dtest_score)))]
numeric(0)
```

For the evening class the calculation is as followed:

$IQR = \mathrm{Q}_3 - \mathrm{Q}_1 = 88.75 - 78 = 10.75$

$\mathrm{T}_{\min} = \mathrm{Q}_1 - (1.5 * \mathrm{IQR}) = 78 - 1.5 * 10.75 = 61.875$

$\mathrm{T}_{\max} = \mathrm{Q}_3 + (1.5 * \mathrm{IQR}) = 88.75 + 1.5 * 10.75 = 104.875$

$Rang_{\mathrm{upper}} = min(104.875, 98) = 98$

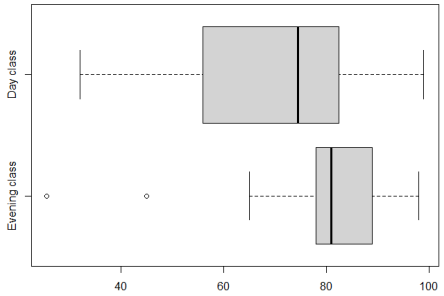
$Rang_{\mathrm{lower}} = max(61.875, 25) = 61.875$

Now, values that are outside of the range of $X < Rang_{\mathrm{lower}} | X > Rang_{\mathrm{upper}}$ are the outliers for the evening class we have two outliers:

```
> eIQR <- IQR(ntest_score)
> TNmin <- quantile(ntest_score, 0.25) -(1.5*eIQR)
> TNmax <- quantile(ntest_score, 0.75)+(1.5*eIQR)
> ntest_score[which(ntest_score < max(TNmin,min(ntest_score)) | ntest_score > min(TNmax,max(ntest_score)))]
[1] 45.0 25.5
```

c) No. We should not remove the outliers. Outliers serve many useful purposes, including: Identifying strong skew in the distribution. Identifying data collection or entry errors. Providing insight into interesting properties of the data. We could remove them if they are critically wrong. For instance, in this dataset. If the scores are negative, they are wrong and they should be removed.

d)
```
> boxplot(ntest_score, dtest_score,names = c("Evening class","Day class"),horizontal = TRUE)
```



i. Both of the boxplots are skewed to the right. In this case, the means should be smaller than the median in each class.

ii. $IQR_{\mathrm{Day}} > IQR_{\mathrm{Evening}} \rightarrow 25.75 > 10.75$

The day class has the widest spread in the middle. The larger the value, the data is more spread out and less consistent. The smaller the value the data is less spread out and more consistent. Therefore, in the day class, the scores are more variable than evening class.

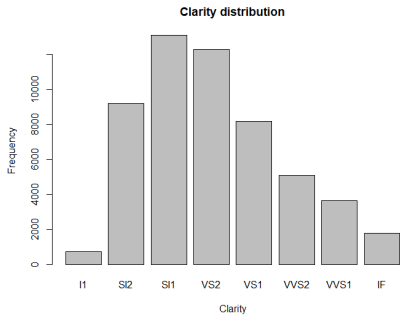
Question #8:

a)

```
> names(diamonds)
 [1] "carat"   "cut"     "color"   "clarity" "depth"   "table"   "price"   "x"       "y"       "z"
> sapply(diamonds,class)
$carat
[1] "numeric"

$cut
[1] "ordered" "factor"

$color
[1] "ordered" "factor"

$clarity
[1] "ordered" "factor"

$depth
[1] "numeric"

$table
[1] "numeric"

$price
[1] "integer"

$x
[1] "numeric"

$y
[1] "numeric"

$z
[1] "numeric"
```

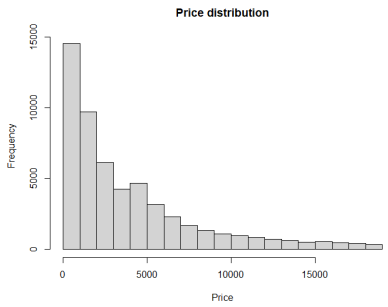
b)

```
> plot(diamonds$clarity, main = "Clarity distribution",xlab = "Clarity", ylab="Frequency")
```



c) The Histogram is right skewed. Which makes sense for this dataset. It is obvious that expensive diamonds are generally rare, and cheap diamonds are more common.

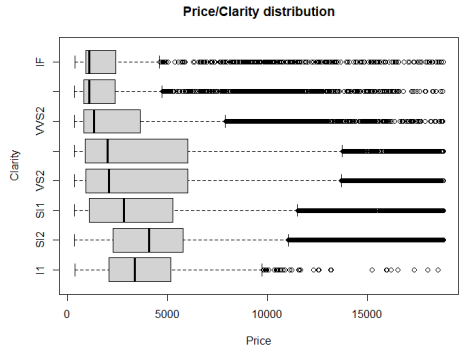
```
> hist(diamonds$price, main = "Price distribution",xlab = "Price", ylab="Frequency")
```



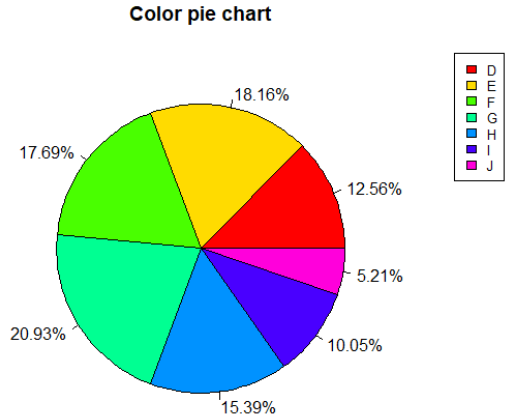
d)

- The outliers in “IF” clarity groups are diamonds with prices over 4596 and under 369.
- The outliers in “VVS2” clarity groups are diamonds with prices over 4717 and under 336.
- The outliers in “VS2” clarity groups are diamonds with prices over 7900 and under 336.
- The outliers in “S1” clarity groups are diamonds with price over 11489 and under 326.
- The outliers in “S2” clarity groups are diamonds with price over 11043 and under 326.
- The outliers in “I1” clarity groups are diamonds with price over 9727 and under 345.
-

```
> PC_plot<-boxplot(diamonds$price ~ diamonds$clarity ,main = "Price/Clarity distribution",xlab = "Price", ylab="Clarity",horizontal = TRUE)
> PC_plot$stats
     [,1]  [,2]  [,3]  [,4]  [,5] [,6] [,7] [,8]
[1,]  345   326   326   334   327  336  336  369
[2,] 2080  2264  1089   900   876  794  816  895
[3,] 3344  4072  2822  2054  2005 1311 1093 1080
[4,] 5161  5778  5250  6024  6023 3640 2379 2389
[5,] 9727 11043 11489 13710 13736 7900 4717 4596
```



e)
```
> pie( table(diamonds$color), labels = paste0(round(100 * table(diamonds$color)/sum(table(diamonds$color)), 2), "%")
+      , main = "Color pie chart", col=rainbow(length(levels(diamonds$color))))
>
> legend("topright", levels(diamonds$color), cex = 0.8,fill = rainbow(length(levels(diamonds$color))))
```



f) Based on the figure, it is perceivable that in each price range diamonds can have any given depths and there is no real relationship between 'depth' and 'price'.

```
> plot(diamonds$depth , diamonds$price ,main = "Depth & Price ",xlab = "Price", ylab="Depth")
```