



Answer:

**Table** represents data from survey of 1691 respondent, which captures the “Sex” “Age” “Marital” Status of the respondent. Data collected has information about number of Cigarettes smoke over weekdays column “amtWeekdays”, and weekends in column “amtWeekends”. Smoking habit is recorded in “smoke” column with “Yes” and “No”.

**sex**: is categorical variable, only possible values of “Male” and “Female”, we can also say it *categorical variable Ordinal (Ordered Variable)*

**age :** Age is discreate variable, since it would take number with jumps 1,2,..and store the age of the respondent in the column field.

**marital:** Marital status is ordered categorical variable, Ordered because it can have marital status “Single” “Married”.

**grossIncome:**It lookscategorical variable, as we have some factors which this data fields can take , like “Under , above certain number”

**smoke:** It is a categorical variable and can have “YES” and “NO” only

**smtWeekends:** This would be discreate variable as we can’t smoke 1.5 cigarette a day .

**amtWeekdays:** This would be discreate variable as it can be whole numbers only (0 to n).



Answer:

*By Checking the count of total observations, we can say that total 1691 participants were included in survey.*

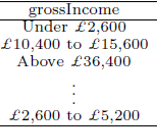


**sex**: is **categorical variable**, only possible values of “Male” and “Female”, we can also say it *categorical variable Ordinal (****Ordered Variable****)*

**age :** Age is **discreate variable**, since it would take number with jumps 1,2,..and store the age of the respondent

**marital:** Marital status is again **ordered categorical variable**, it can have only marital status “Single” “Married”.

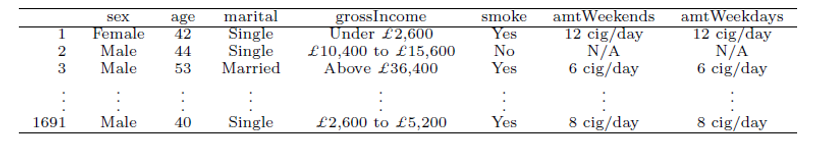
**grossIncome:**It lookscategorical variable, as we have some factors which this data fields can take but order is not clear here so it is an **unordered categorical variable .**

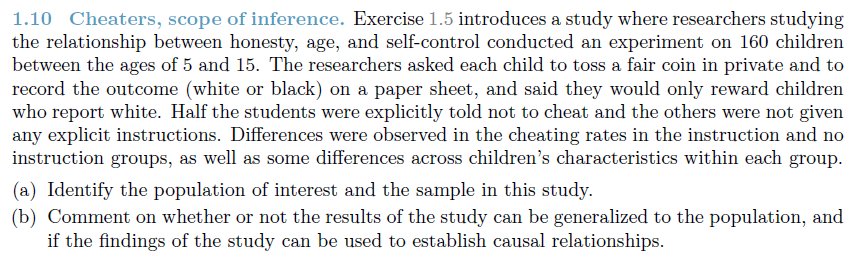


**smoke:** It is a **ordered categorical variable,** since itcan only have “YES” and “NO” only

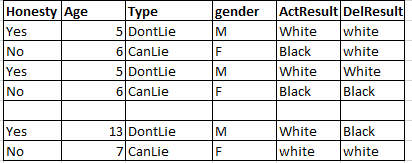
**smtWeekends:** This would be **discreate variable** as we can’t smoke 1.5 cigarette a day .

**amtWeekdays:** This would be **discreate variable** as it can be whole numbers only (0 to n).







According to the data recorded sample of study looks like below, 

Population of Interest: Data was recorded for 160 children between age 5 – 15 with. Group of children are divided into two group with one group has direction of not to lie and other has no direction (they may lie).



Given the sample size of 160 children, it seems very less also we are not considering the education level, and influence circle of the kids   
(family type “Join” “Nuclear” “Single Parents” etc. )

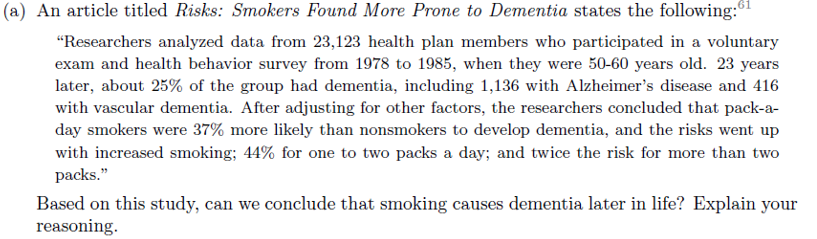
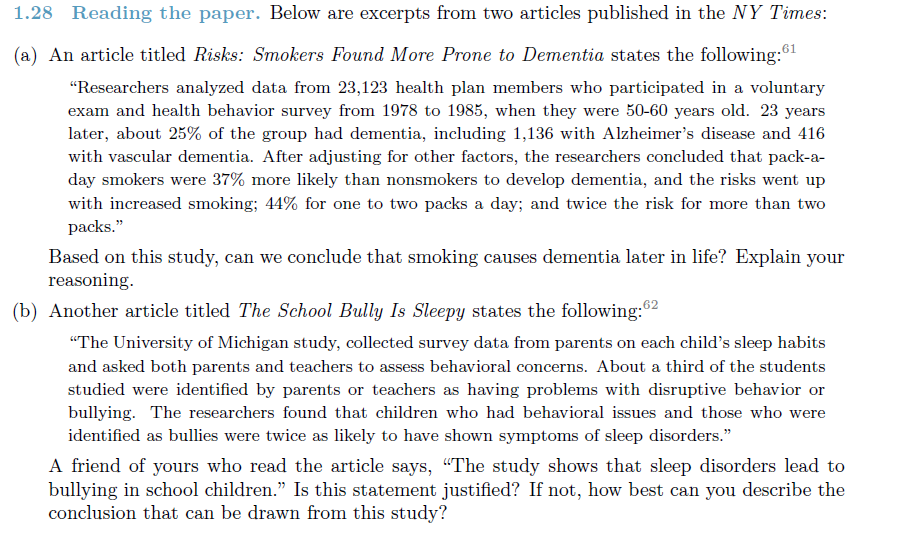
Its not clear from the sample that how these children were selected.

Result from this survey can give us some hint about the behavior which should be Experimented further based on the results from this survey.

Possible Casual relationship:

a. We might have relation between Age, and Honesty level of the them.

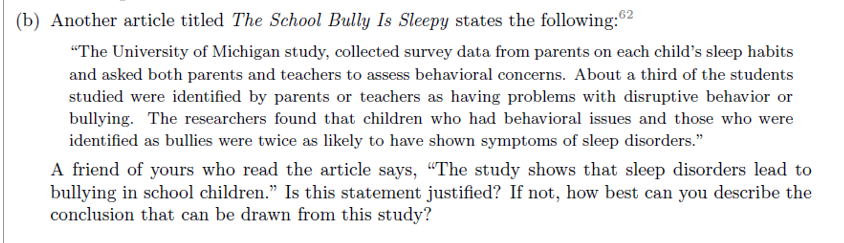
b. Effect of direction (one group can lie or one group can’t lie ) to the group and Honesty.



Answer:

Based on texts : “After adjusting for other factors, the researchers concluded that pack-aday smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs.”

We can say that smoking increased dementia, more important to note that risk increases if number of packs increases more than two pack as day.

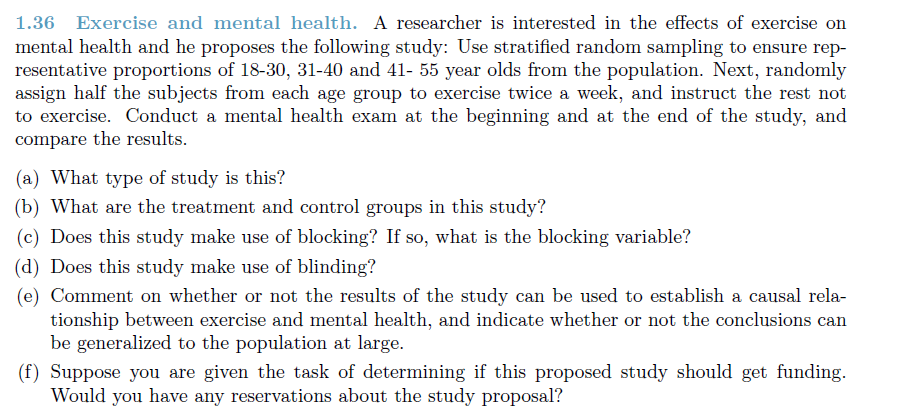


Answer :

Researchers suggest that children with behavioral issue and bullying are more at risk of showing Sleep Disorder. As per data 1/3rd of data support this conclusion. Other side of the coin says that only if children are suffering from both problem then they are at the double risk but its not the same risk if you have just one of the problems. Sleep disorder doesn’t always result in bullying.

Data suggest that 1/3rd of datapoints are supporting the statement and   
conclusion shared by friends. But its not clear how only sleep disorders can result in Bullying.

It may be good to do further controlled study and see if controlling the sleeping disorder would result in no bullying or change in behavior.





Sampling of data is done with “multistage sample” technique. This type of study is called Blocking study, since people are divided in each group of age , and one group is advised to exercised and other is advised not to .

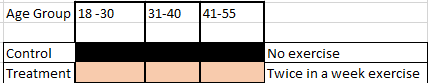


Treatment groups: This group of the people are who has been asked to do exercise twice a week.

Control Group: This group of the people are advised not to exercise.



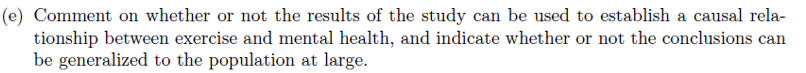
Simple depicting data group:



Study make use of blocking, the categorical variable indicating who can exercise and who can’t is the blocking variable .



Blinding is the case where both respondent and survey recorder (executors ) are not informed about the control and treatment group. Here people are asked to not to do exercise and are grouped , respondent are aware of the constraint hence its not making use of blinding.

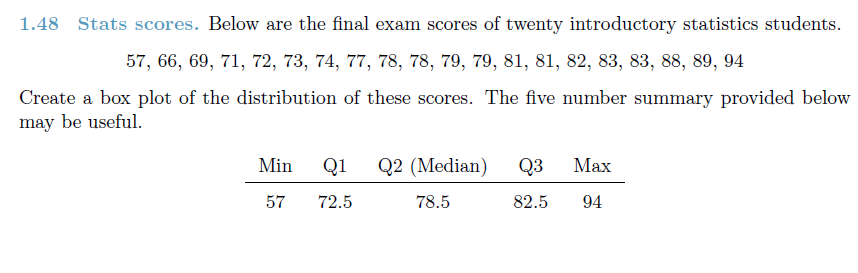


Since its blocking study and people part of the control group with different age group may not be feeling good after skipping the exercise. So its very much interesting to see how mental health shows vs Exercise.

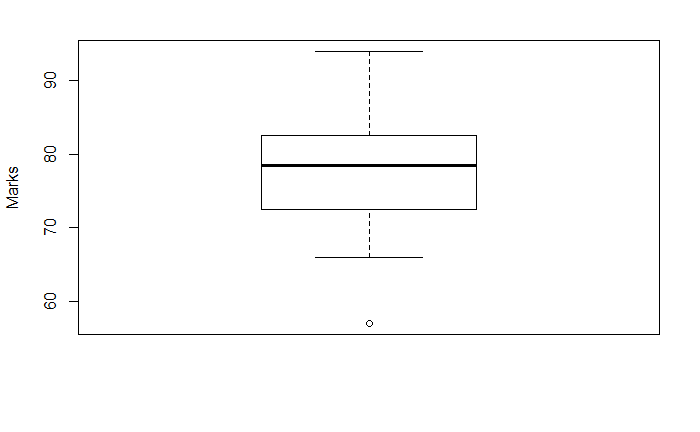
This conclusion may not be generalized 100% as we are not able to control the emotional variable of people who were asked not to exercise. May be a blind study on such group would give more better result.

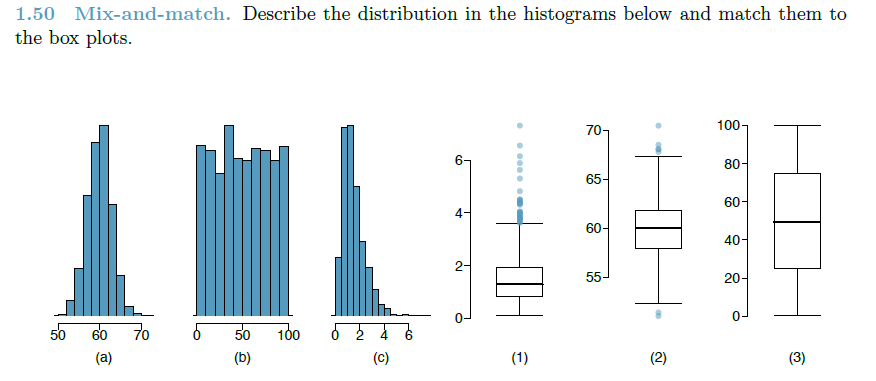


Since we are checking mental health against exercise pattern, it important that we don’t add external emotional variance to the data, it’s not 100% possible to control all variable but we can minimize it with some technique. Keeping group blind about control and treatment or keeping them separate might have better result.



stmark<- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)

boxplot(stmark, ylab= "Marks" )  
  




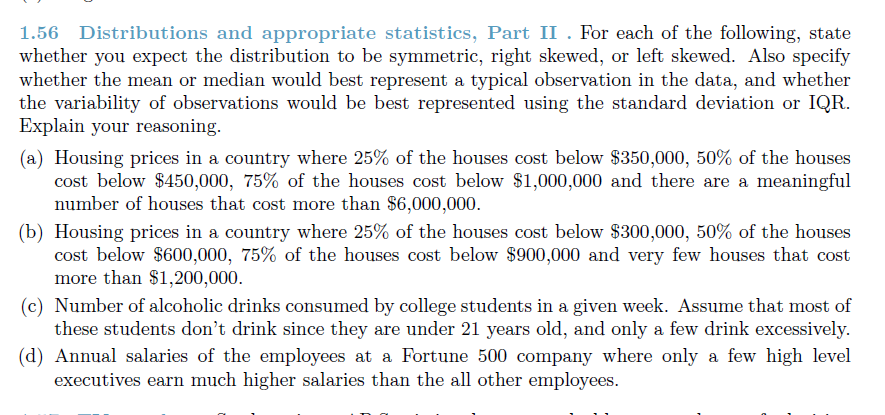
Bar chart of (a) is symmetrical distribution,with data ranging from 50 to 70, with median close to 60.

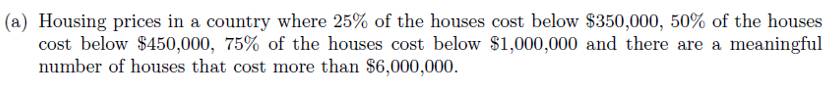
BOX Plot (2), shows the same pattern of median close to 60 and data in the range of 50 to 70. There are few outliers present at outer and lower whisker .

Bar char (b) is multimodal distribution, most of the data points are between 0 to 100. Median of the data is close to center of datapoint i.e. 50.

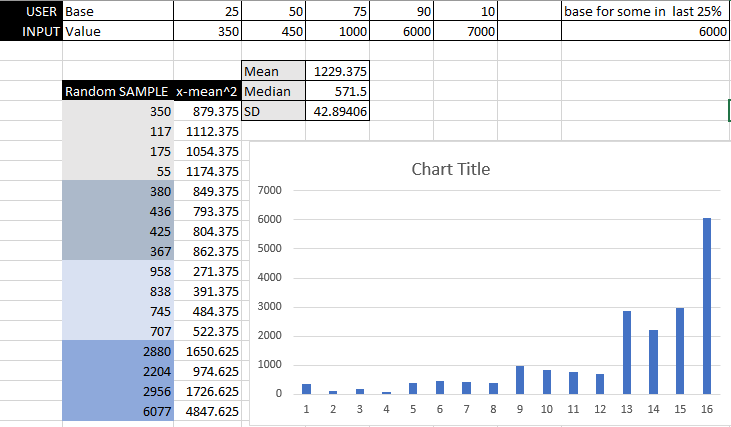
Box plot (3) shows bigger IQR due to the data points, ranging from 0 to 100.

Bar chart (c) is right skewed graph ,with data ranging from 0 to 6. Box plot (1) shows median of data somewhere close to 1 more than 50% of data points are below 2. Sine there are more outliers even it explains due to data being right skewed.





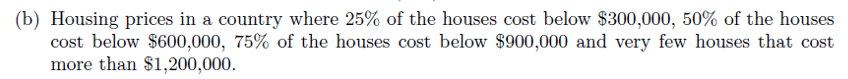
I created a sample data set based on given view:

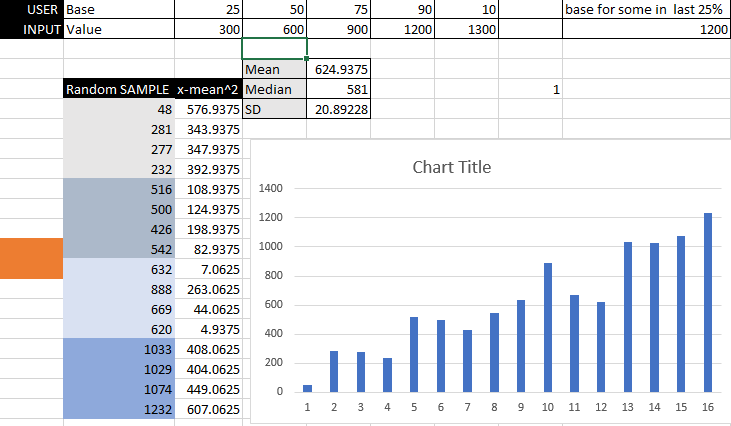


It would be left skewed distribution, as price is going up if data points are sorted by price.

Median of the house price would be good starting point to check the housing value in the area. Mean would get impacted by 25% of 6000k and 600K+ price of houses and may not give right picture.

With box plot showing datapoints in IQR would give us the hint where the 50% of the data lies between 1st and 3rd quartile. IQR range would give little better info about pricing.



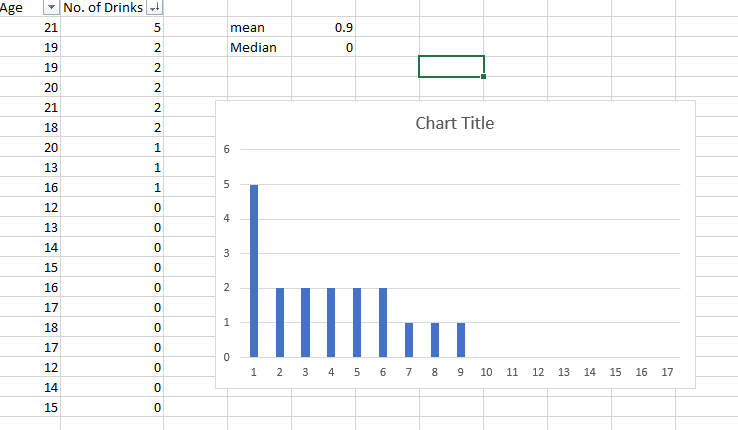


Based on sample dataset its clear data is left skewed, here based on data points median of the data would make more sense.

Observations can be best represented using IQR as we can see that what frequency of data falls within IQR and understand outliers, which could be due to high price point for some house (1200+).

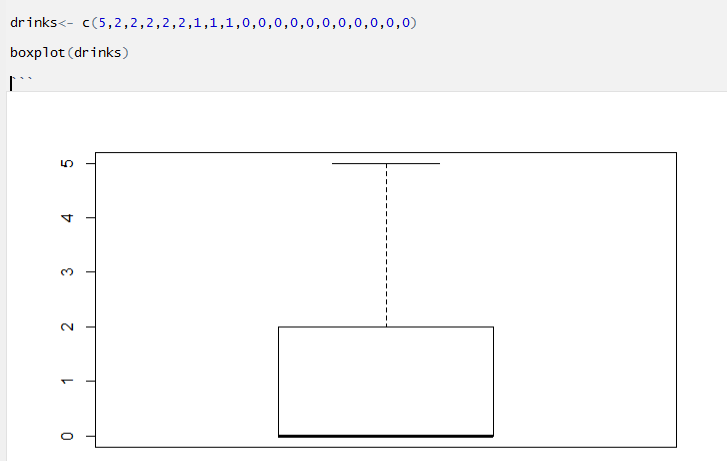


Depending on data points if data is sorted in decreasing order drinks , then we get right skewed distribution.



Mean may give wrong info indicating average 0.9 is being consumed by everyone in the age group. Median gives at least a right number 0 which seems a good sign given the age group of the students.

Below boxplot does start from Zero and IQR shows that most of the students are under 2 drinks per week and max going up to 5 and no outlier. Since there is no lower whisker it mean that at least 50% of class doesn’t consume such drinks.

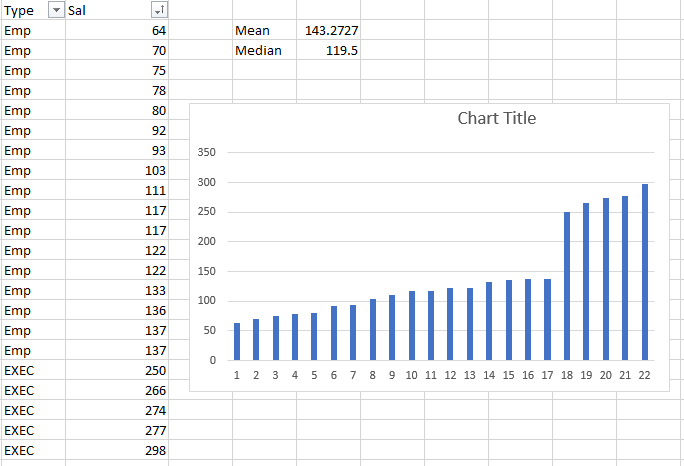




Based on datapoints if data is sorted in increasing order of salaries then we see that graph is left skewed distribution.

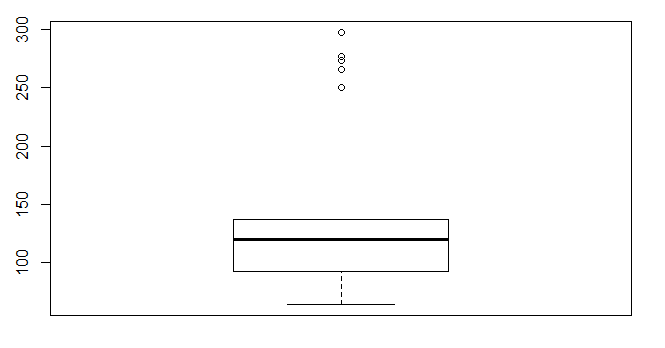
Since few data points are very high and not in natural distribution mean would not give right info about salary.

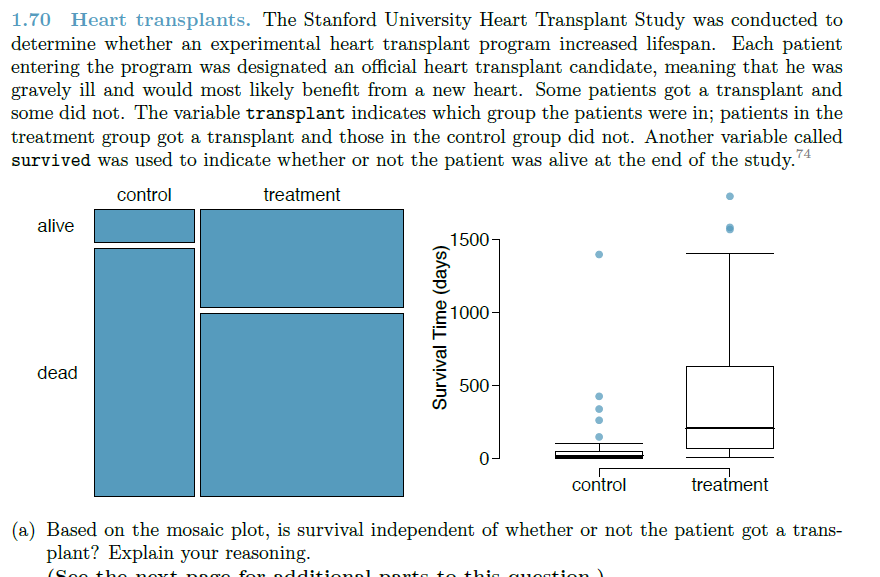
On the other side Median would be right indicator to check for some data points in the set to understand how the salaries are given. Since it would say at least 50% from here are getting more than this number and below it is getting less than median.



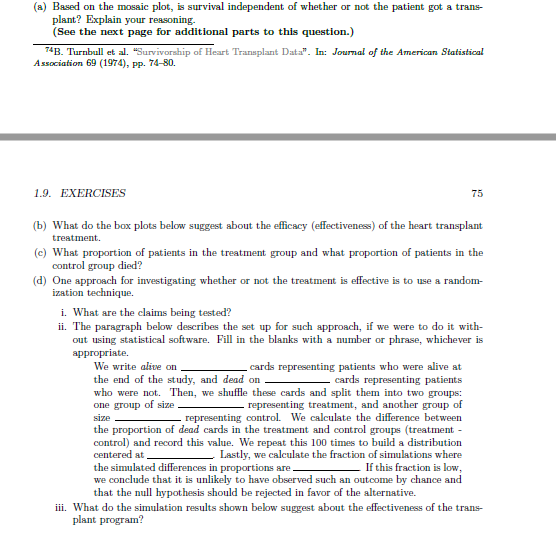
Boxplot below very clearly indicate that median and group of people falling in between 1st and 3rd Quartile. We can also see the executive staying as an outliers and not even close to outer whisker.

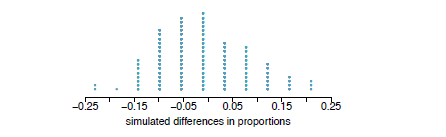
I feel IQR would give good glimpse of salary distribution.



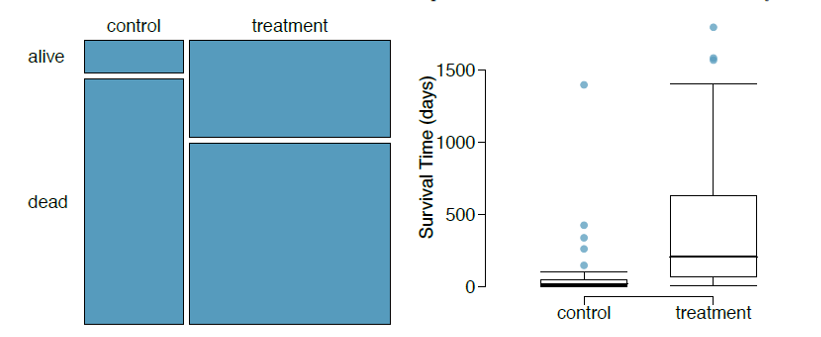


Based on sample of data all the patients were very ill, so if we can see more than 50% death even after transplant in the treatment side of the chart, which looks on mosaic plot, it can be said that treatment might have saved some percentage of the people but not everyone, as we can see close to 85% died in control group.









Boxplot for treatment group shows more that 75% of the population survived below 600 days, also at least 50% lived form 200 – 600 days (between median and 3rd quartile )

Control group boxplot is very depressing, even the outer whisker falls below the median of the TREATMENT Group.

It can be said the treatment did add big number of days for survivors and was a great success.



Based on R code :

#For treatment group

round(prop.table(table(heartTr[which(heartTr$transplant=="treatment"),4])),2)

alive dead

0.35 0.65

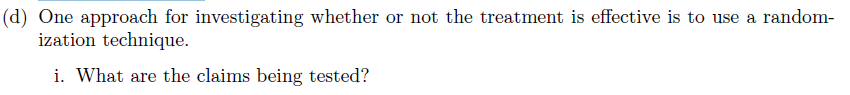
#For Control group

round(prop.table(table(heartTr[which(heartTr$transplant=="control"),4])),2)

alive dead

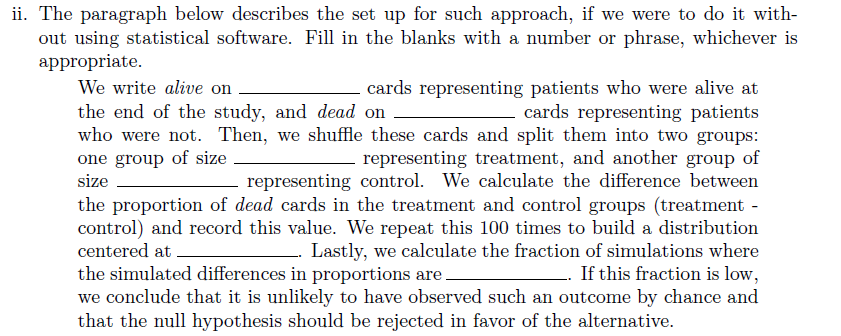
0.12 0.88

In the treatment group, 0.65 died. In the control group, 0.88 died. The difference is about -0.23.



**H0- Independent Model**: The transplant treatment has no effect on survival. The observed higher survival rate was due to chance.

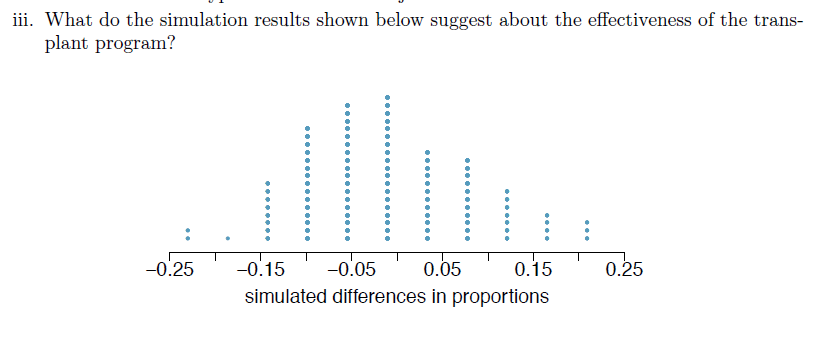
**HA- Alternative model**: The transplant treatment has effect on patient survival. The observed higher survival rate was not due to chance.



By R:

|  |  |
| --- | --- |
|  |  |

We write “alive” on **28** cards representing patients who were alive at the end of study, and “dead” on **75** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size **69** representing treatment, and another group of size **34** representing control. We calculate the difference between the proportion of “dead” cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at **0**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **less than -0.23**. If this fraction is low, we conclude that it is unlikely to have observed an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.



The simulation shows that the different in proportion being less than 0.23 (23%) due to chance is highly unlikely. Therefore, the proportion different is not due to chance and the evidence is sufficiently strong to reject H0, and the alternative is accepted - that the transplant is effective.