Professor Moody

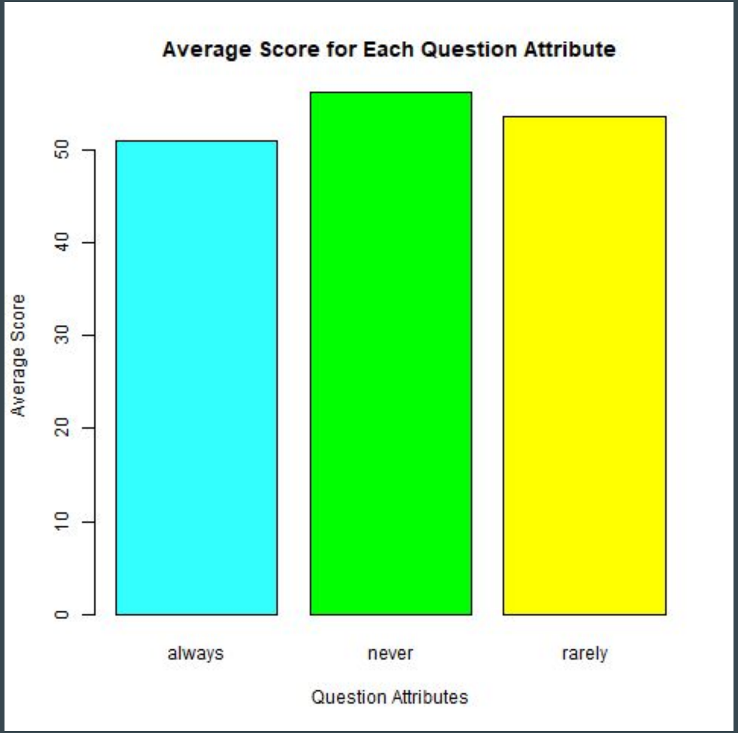
•••

By: Rohit Manjunath

My Hypothesis

- As you can see in the barplot on the right, never asking questions in class has the highest average score. Rarely asking questions has the second highest average score.

- Based on this information my hypothesis is, the average score of students who never ask questions is higher than the average score of students who rarely ask questions.



barplot(tapply(moody$score, moody$questions, mean), main = "Average Score for Each Question Attribute", col = c("#33FFFF", "green", "yellow", "orange", "red"), xlab = "Question Attributes", ylab = "Average Score")

The Process

- To test our hypothesis we normalize the data and calculate the p-value.

- First we can start by splitting the main dataset into two subsets. One that will contain data of students that never ask questions and another that will contain data of students that rarely ask questions.

- We need the following data to calculate the p-value:
  - Standard Deviation of both the datasets.
  - The mean of both the datasets.
  - The difference in standard deviation of the datasets.
  - Finally, Z-score of the datasets.

Data Preparation

- Let us start by subsetting our data into students that never and rarely ask questions.

```r
neverAskQuestions <- subset(moody, moody$questions == "never")#Subset of students who never ask questions.
rarelyAskQuestions <- subset(moody, moody$questions == "rarely")#Subset of students who rarely ask questions.

neverScore <- neverAskQuestions$score#Stores only the score attribute.
rarelyScore <- rarelyAskQuestions$score#Stores only the score attribute.
```

- The above code gives us two important sets of data. “`neverScore`” which stores the scores of all students that never ask questions. Similarly, “`rarelyScore`” stores the scores of all students that rarely ask questions.

Calculating the Standard Deviation, Mean, and Length.

- Now we can get the standard deviation, mean, and length of both the subsets by using the functions "`sd()`", "`mean()`", and "`length()`".

```
neverSD <- sd(neverScore)
rarelySD <- sd(rarelyScore)

neverMean <- mean(neverScore)
rarelyMean <- mean(rarelyScore)

neverLength <- length(neverScore)
rarelyLength <- length(rarelyScore)
```

```
Output:
 ●   neverSD : 25.68241
 ●   rarelySD : 25.36096
 ●   neverMean : 56.32474
 ●   rarelyMean : 53.69217
 ●   neverLength : 287
 ●   rarelyLength : 253
```

- As you can see the standard deviation of both datasets are really close. The mean of both datasets have a difference of somewhere around 2.7(rounded). Is that a random difference or does that actually mean anything? The p-value will tell us.

Calculating the Difference in Standard Deviation and Z-value

- Now we can calculate the difference in standard deviation and Z-value using the known formulas.

```
bothSD <- sqrt((neverSD^2) / neverLength + (rarelySD^2) / rarelyLength)

Zscore <- (neverMean - rarelyMean) / bothSD

Output:
  ●   bothSD : 2.200094
  ●   Zscore : 1.196569
```

- Now we have everything we need to calculate the p-value.

P-value

- We can use the function “pnorm()” to assist us in calculating the p-value.

```
p = (1 - pnorm(Zscore)) / 2

Output:
 ●  p : 0.05786865
```

- As we can see the p-value is 0.05786865. Since the p-value is just greater than 0.05, we fail to reject the null hypothesis.

Thank you!