Exam 1

Mitchell Meier

October 9, 2020

Problem 1

Introduction

For this report, we are going to analyze a set of data describing the past data scores for a class, with some information about each student included for context. Including these variables about the student, in addition to their exam scores, will hopefully help us in determing if there is an coorelation between a student's exam scores and their other traits

For the columns in our data set, we have:

- *Section*, which is a categorical nominal data type that lists the section number the student was in (either Section 101 or Section 102)
- The next three columns, *Exam 1, 2, and 3*, are all quanitive discrete data types. These columns are the primary focus of our data set
- After the exam scores we have the *Attendance* column, which is a categorical ordinal data type with four possible values (listed highest to lowest): frequently, occasionally, rarely, never
- The next column, *College level*, is also a categorical ordinal data type. The values for this colum are: senior, junior, sophmore, and freshman
- Our last column is *Major*, which is a categorical nominal data type

Our assigned objectives for this study are to find if a student's:

1. attendance has any impact on their exam scores
2. exam 1 score impacts their exam 2 and exam 3 scores
3. attendance is dependent on their college level
4. college level has any impact on their exam scores
5. major has any impact on their exam scores
6. section has any impact on their exam scores
7. section has any impact on their attendance

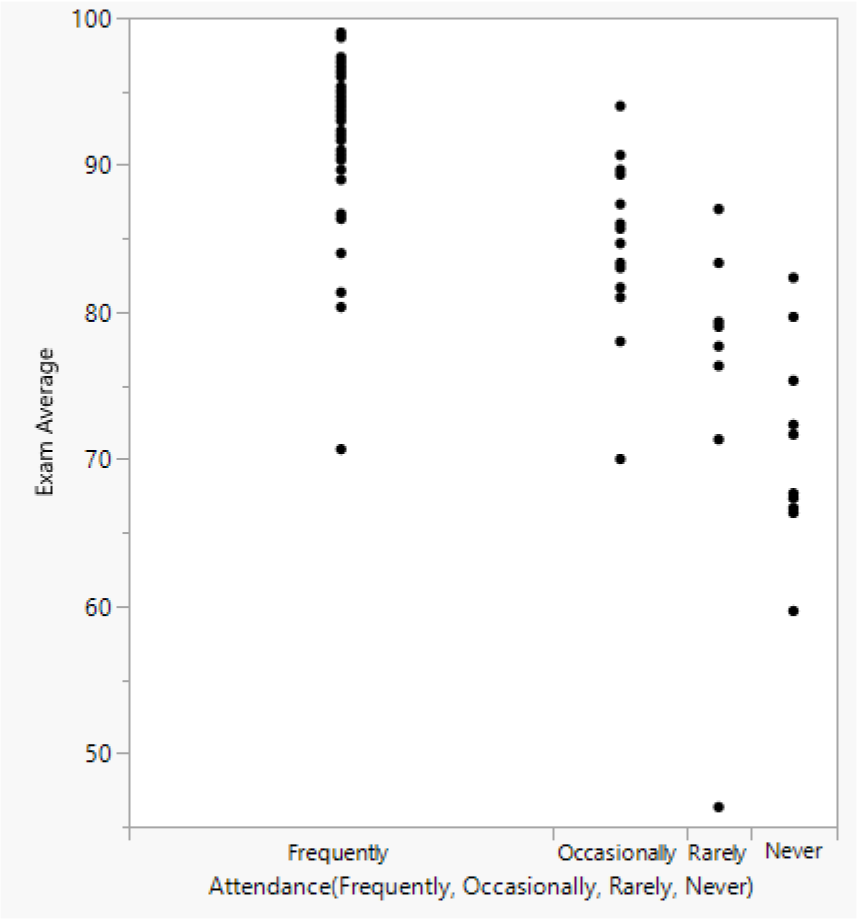
Our main goal heading into our analysis is to see how our columns influence a student's exam scores. Hopefully we will be able to analyze this data to a point where we can predict how a student will score on their exams based on their other traits. We may also discover by looking at the individual exam data that the exam number may also factor into a student's success

Analysis

For the analysis section, we'll be splitting up our analysis between each objective, and use graphs and statistical summaries to make inferences on the data. To help with our analysis, I decided to add one additional column to our data, called *Exam Average*, which contains the average score for each student out of their three exams. I believe this column should still fall under the requirement of ”provided data”, since it is entirely derived from provided data. This column will make it easier in comparing exam scores to other attributes

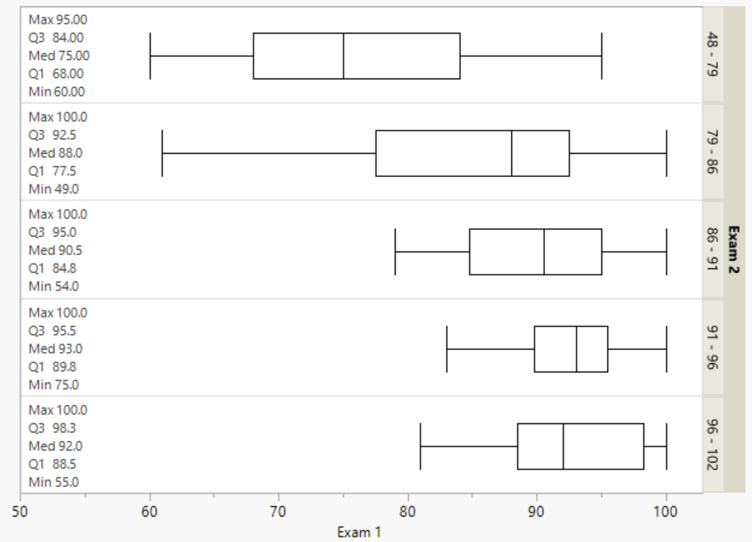
1. Attendance seems to have one of the most distinct relationships with average exam score. As we can see in our coorelation graph, a large majority of the highest scores were from student that attened class frequently. In fact, out of the 44 highest averaging student, 42 of them visted class frequently, and the other two visited occassionally. The average (of exam averages) for students who visited class frequently and occassionally were high as well, with only one data point in each that could be considered an outlier. I believe it is safe to say that students who attend STAT 3113 more frequently will be more likely to do better on their exams.

Student's attendance and average Exam Score

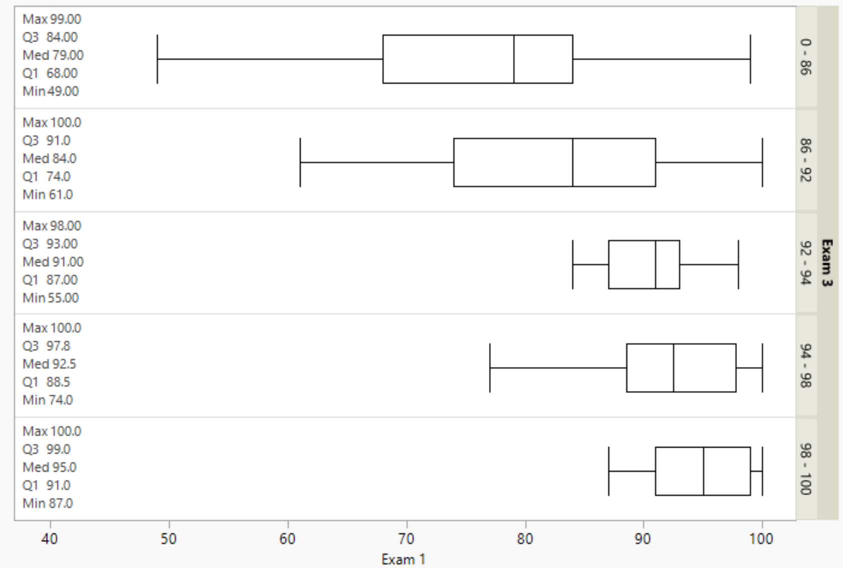


2. Students in STAT 3113 seem to be consistent between tests on average. There is a steady trend showing that the score a student gets on exam one is similar to the scores they get on exam two and three. There do seem to be cases where this does not hold true, however, each exam has some outliers where a couple students have drasticly different score between exams. But for the majority of students, this trend does in fact hold true

Student's Exam 1 score and Exam 2 score

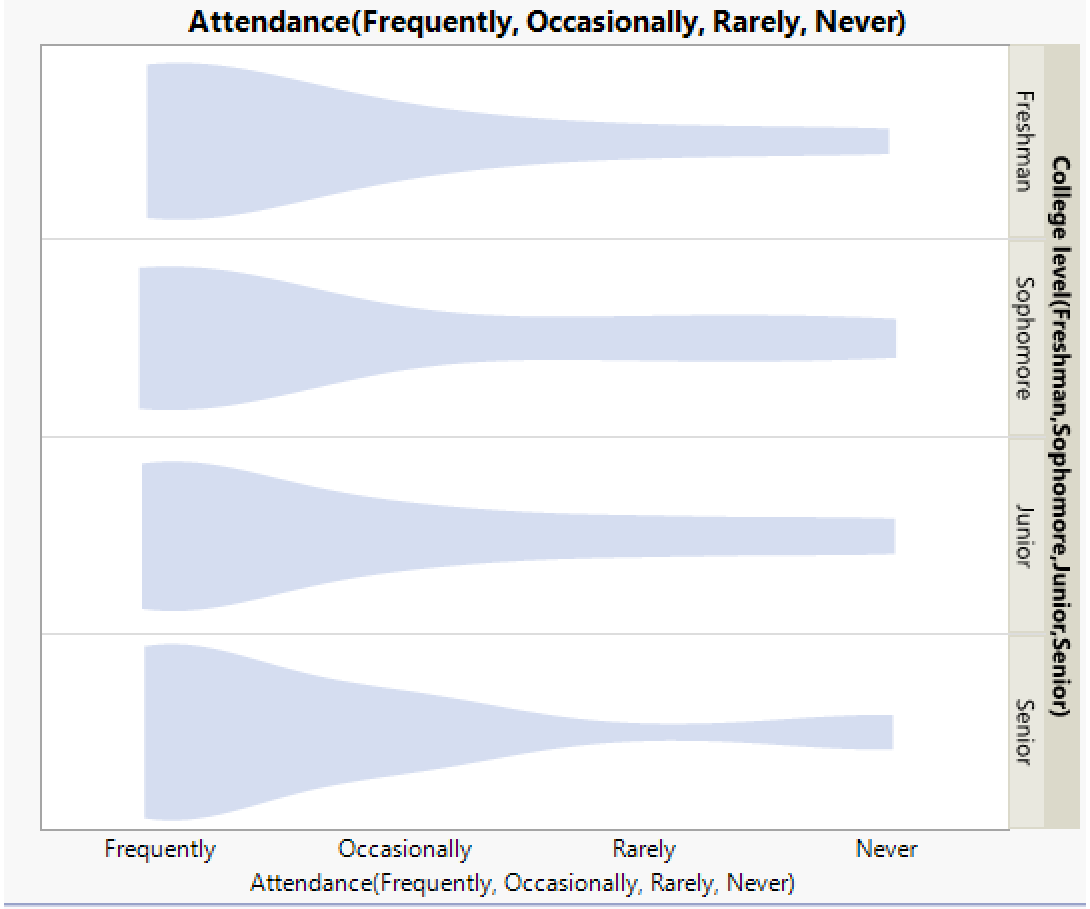


Student's Exam 1 score and Exam 3 score



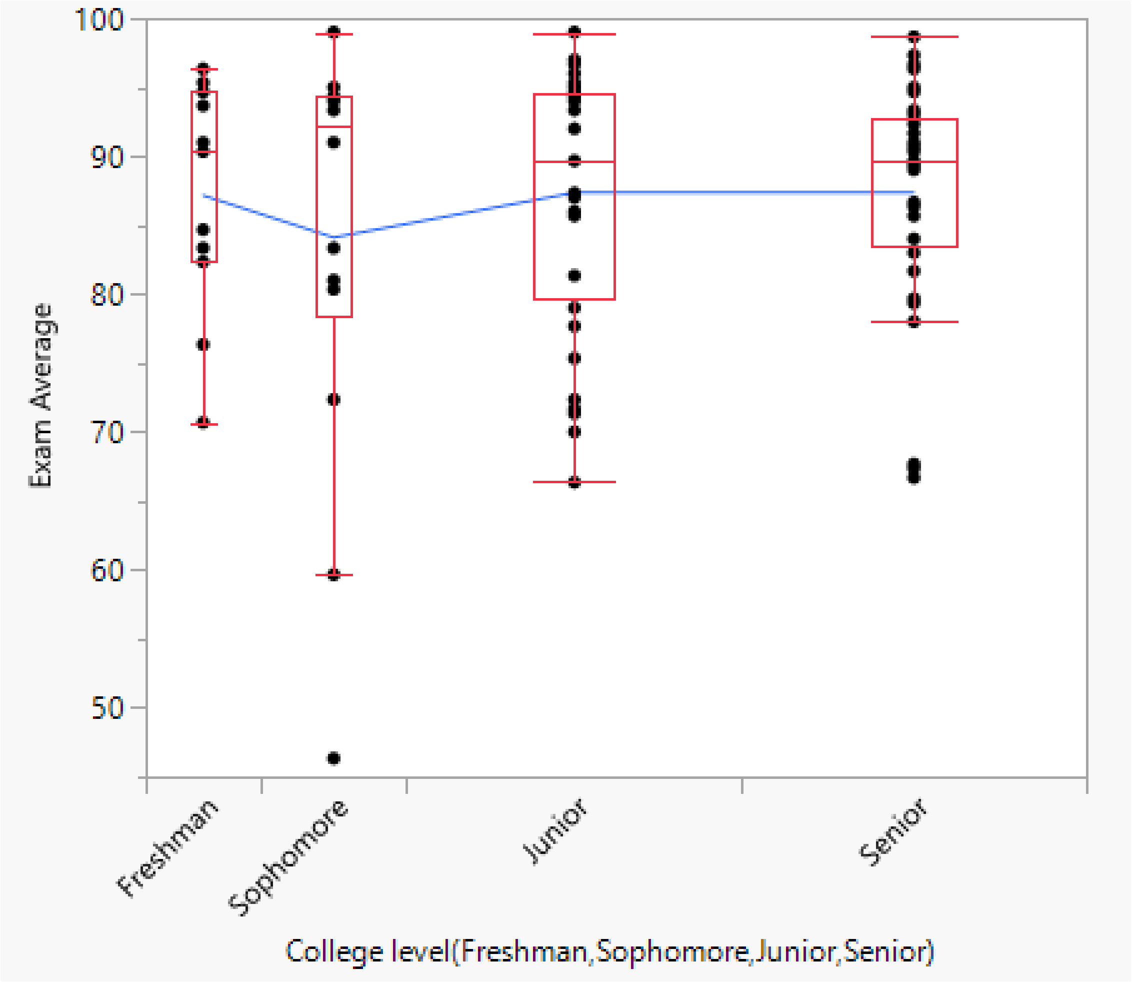
3. Since this study had a disproportiante amount of upper class men to lower class men, I decided the best way to evaluate the coorelation between attendance and college level was a contour graph. The contour graph will give more of an emphasis on the proportion of students (per class level) that atteneded class frequently, rather than just the number. From looking at the contour graph, there seems to be negligable difference between class levels when it comes to attendance. All classes had a large chunk of their students attending frequently, with a slightly smaller percentage of students attend occasionally. Only a slightly larger percentage of sophmores and juniors attened class rarely or never compared to freshman are seniors.

Student's Attendance by College Level



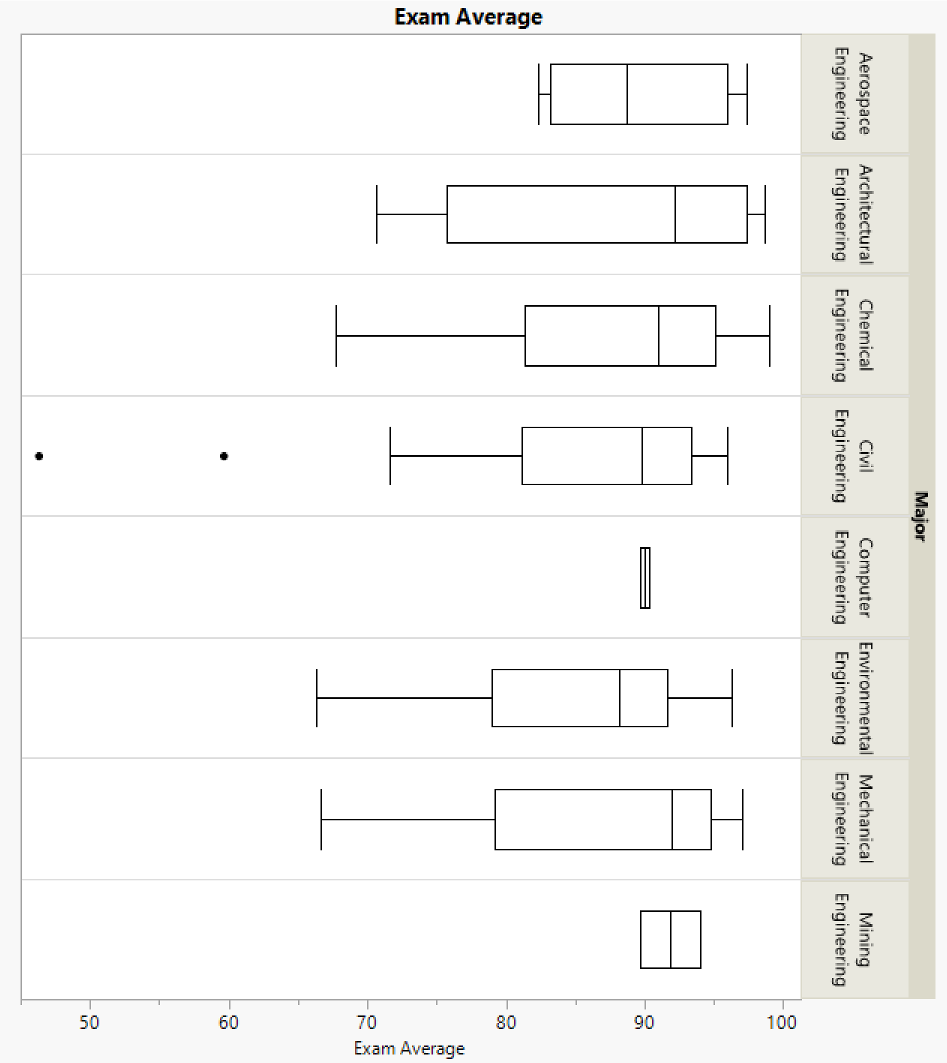
4. The relationship between class level and exam scores seems to be similar to our last relationship. Our boxplots show that the interquartile ranges for each class level are very close to each other, as well as the means. The sophmore class is the only class level with a slightly lower mean compared to the other class levels, but this could simply be attributed to the sophmore class having the lowest outlier by far.

Student's average Exam Score by College Level



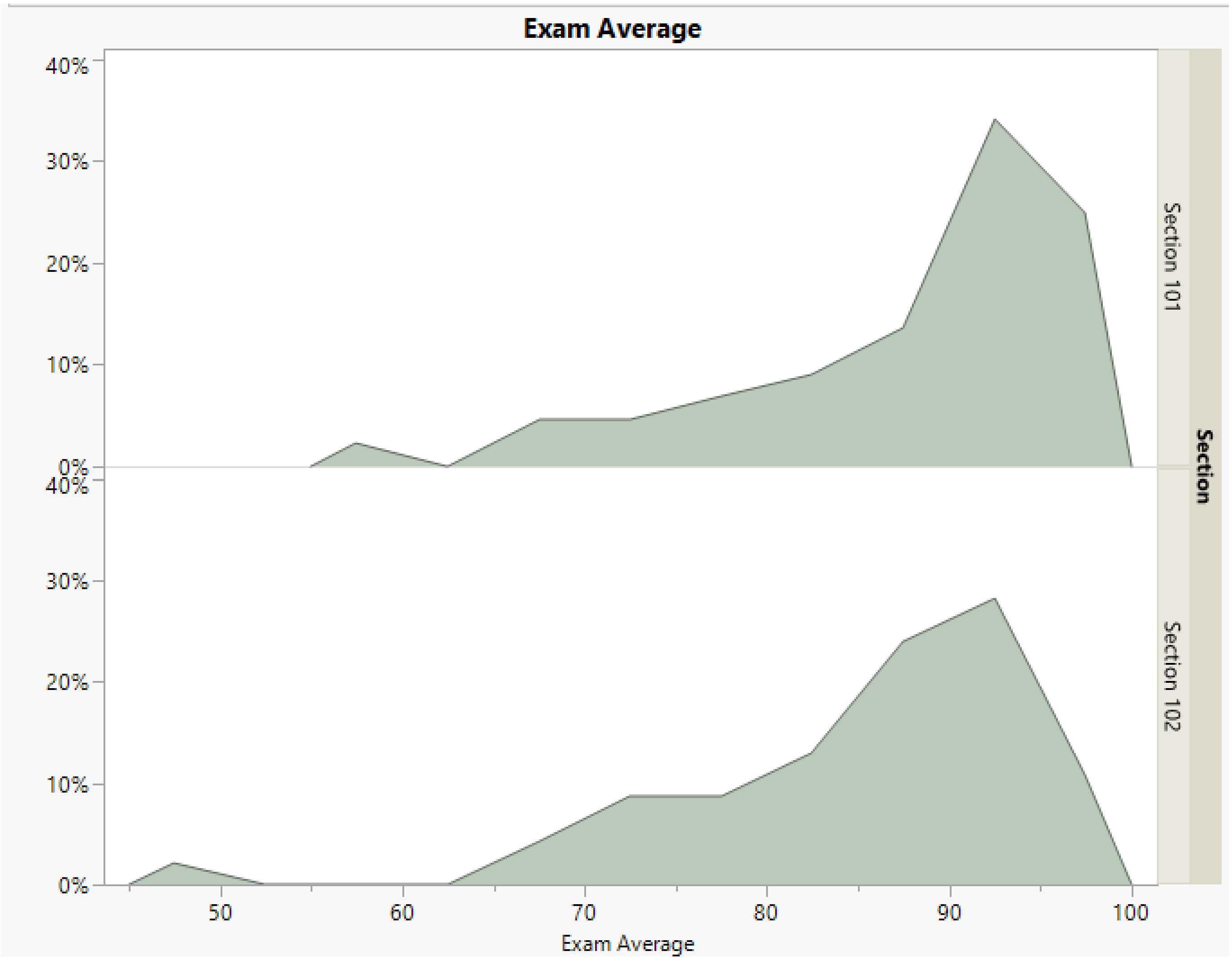
5. According to our data, the top 3 performing majors, based on average exam scores, are Architectrual Engineering, Mechanical Engineering, and Mining Engineering. However, when we compare the reliability of this data to the rest of our data, it is much lower. This is because we have a wide variety of majors that have taken STAT 3113, so our data for most of these majors consists of a comparibly smaller sample size. With that being said, I believe there is insufficent data here to draw a strong conclusion on which major's have a better chance of obtaining higher exam scores.

Student's average Exam Score by College Level



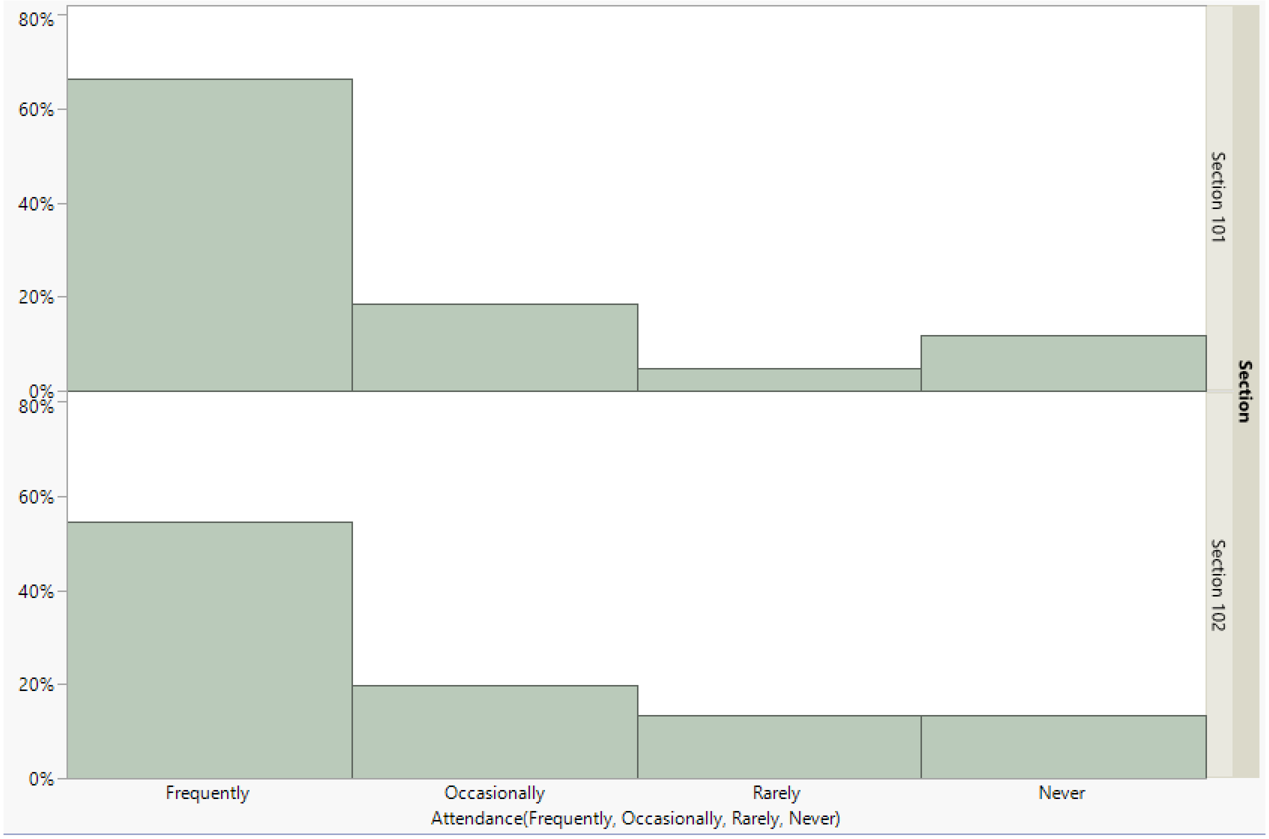
6. Using a histogram to analyze our section data, it seems that there is not a big difference between taking section 101 and section 102. Both sections have similar data distributions. If we were to point out one difference, however, section 101 seems to be slightly more left skewed, though it is barely noticeable. Odds are that section number will not have an impact on a student's exam score.

Student's average Exam Score by College Level



7. **Additional Objective:** For my additional objective, I wanted to see if students' attendance was coorelated to section number at all. If it was, there could be a good amount of factors not listed in our data that, if listed, could provide a good deal more insight into why student's grades are high or low. Perhaps a certain section is too early in the morning for students to attend, or they don't attend because they find their paticular section's lectures uninteresting (very unlikely for our class :)). Looking at our data, it seems that there is a slight different between the attendance in our two sections. Less students are attending frequently in section 102 than 101, though not by a wide margin (less than 10 percent).

Student's Attendace by Section Number



Conclusion

To summarize the conclusion's by objective:

1. Attendace heavily impacts exam scores. The data infers that students who attended class more often are more likely to score higher on exams

2. Students who were the higher scores on exam one tended to also be higher scores on exams 2 and 3. The data did not guarantee that students would keep their high scores, but in the majority of cases they would.

3. The data suggested that students attended class slightly more than juniors and sophmores, but not by a wide margin. The most likely assumption to be true here is that a student's college level will have little impact on their attendance

4. Similar to how college year coorelated with attendance, it also does not seem to influence exam score that much. The best reccomendation to make to a student regarding when to take STAT 3113 is to not worry about what year is generally best to take it. It is more likely to be dependent on the student's course load by year

5. Major may or may not impact exam score. Going off the data in our study alone, a case could be made for certain major's excelling more in this class. However, not only was there not too much difference between majors, but the sample size for each major is quite low for this experiment. So to conclude on this objective, I believe it can be hypothesised that major does not impact exam score, but I am not confident enough with our sample size to say that with certainty

6. The data shows that the students from section 101 and 102 have similar exam scores. So going off this realtionship alone, it would be easy to conclude that no one section has an inherent advantage. However, more factors are usually involved when it comes to differences between sections, so I decided to explore section relations more with my additional objective

7. In our addditional objective, the data shows that there is a slight coorelation between attendance and section number. Students in section 101 seem more inclined to attend class for whatever reason compared to 102. This could be due to a variety of different factors within the section, or it could just be due to the kinds of students in each section. More coorelation analysist, and possibly even new data points, would be needed to find the reason for this difference in attendance

Problem 2

I. We require a geometric distribution model in order to determine the number if driver files that are randomly drawn until the company finds one that has not exceeded their deductible

Our parameter is $p = 0.88$ so our expected value is

$$E(X) = \frac{1}{p} = \frac{1}{0.88} = 1.14$$

We know this problem represents a geometric distribution because it is an experiment that has independent trials with the same constant probability p of success, and it also ask how many trials will be conducted until a success is obtained

II. We need to find the mean of the binomial distribution of X in this problem to determine the number of defective transistors in a batch of fifty produced

Our parameters are $p = 0.02$ and $n = 50$, so our expected value is

$$E(X) = n * p = 0.02 * 50 = 1$$

We know this problem represents a binomial distribution because each trial has a boolean outcome (success or failure), all trials are independent, and the probabilty p is the same for each trial

III. We require a geometric distribution model in order to determine the number of transistors examined before finding the first defective transistor

Our parameter is $p = 0.02$ so our expected value is

$$E(X) = \frac{1}{p} = \frac{1}{0.02} = 50$$

We know this problem represents a geometric distribution because it is an experiment that has independent trials with the same constant probability p of success, and it also ask how many trials will be conducted until a success is obtained

IV. We need to find the mean of the poisson distribution of X in this problem to determine the number of volcanic eruptions in japain within one year

Unfortanutley, we are not given any λ, only the time period that the λ would reside in. This means we have insufficent data to find the expected value $E(X)$

V. We need to find the mean of the binomial distribution of X in this problem to determine the number of tours that are unsuccessful

Our parameters $p = 0.2$ and $n = 10$, so our expected value is

$$E(X) = n * p = 0.2 * 10 = 2$$

We know this problem represents a binomial distribution because each trial has a boolean outcome (success or failure), all trials are independent, and the probabilty p is the same for each trial

VI. We need to find the mean of the poisson distribution of X in this problem to determine the number of power failures in Planet007 per three weeks

Our parameter $\lambda = 9$ because we know if our λ for a week is 3, we know our lambda in a time period three times that is $3 * \lambda = 9$, so our expected value is

$$E(X) = \lambda = 9$$

We know this problem represents a poisson distribution because it is a distribution of discrete occurences over an interva, where those occurences can range from 0 to ∞

Problem 3

To show the geometric distribution is a valid probability distribtion, lets make up an example experiment that falls under the category of a geometric distribution

Let's say a baseball team has a 25 percent chance of winning a game for each game they play. We want to show that, given enough games, the probability that this baseball team has won at least one game is approximatley equal to 1.

The value of our random variable X for a given x numbered game is

$$f(x) = (1-p)^{x-1}p = 0.75^{x-1} * 0.25$$

Let's give this baseball team an infinite number of games to play in order to win at least one. With an infinite number of games, our new probability function would be

$$f(x) = \sum_{x=1}^{\infty} 0.75^{x-1} * 0.25$$

Calculating each value of x for this summation, we can start to see a trend

$$x = 1, f(x) = 0.25$$
$$x = 2, f(x) = 0.1875$$
$$x = 3, f(x) = 0.140625$$

So we see that our probability of the baseball team winning its first game is going down by the game. This is because for each new game, we must take the probability that the baseball team would win the game times the probability that the baseball team has not won a game to this point

If we continute, x gets smaller at a rate that the total summation of probabilites will never be greater than 1, but it will get very close, close enough that you could determine that

$$f(x) = \sum_{x=1}^{\infty} 0.75^{x-1} * 0.25 \approx 1$$

Proving that this geometric distrubution is a valid probability distribution