INVESTIGATING UNDERGRADUATE STUDENT UNDERSTANDING OF GRAPHICAL DISPLAYS OF QUANTITATIVE DATA THROUGH MACHINE LEARNING ALGORITHMS

by

ALEXANDER JAMES LYFORD

(Under the Direction of Jennifer J. Kaplan)

Abstract

The purpose of this research is to provide insight into student understanding about graphs of quantitative, univariate data. Specifically, students' understanding of the variability of data displayed graphically is investigated. This research also utilizes an ensemble of machine learning algorithms to further investigate this knowledge and expedite the investigation process for large data sets.

A total of nine constructed-response items are disseminated to students through an online homework platform. The responses to these questions are examined to determine the prevalence of particular misconceptions about variability in graphs and to investigate the relationships between these misconceptions. In addition to the nine online homework items, this research includes face-to-face, task-based interviews with 19 students from the same introductory statistics course at the University of Georgia. Students are asked a series of questions that are isomorphic to their online counterparts. The differences in completeness and correctness are analyzed between responses given online and those given during face-to-face interviews.

A rubric is constructed for each of the nine constructed-response items and used to categorize student responses. Each rubric has multiple bins, and student responses may be assigned to one or more of the rubric bins. Multiple statistics PhD students read a small subset of student responses for each item and categorize these responses into the appropriate bin. Next, a series of eight machine learning algorithms is constructed using the previously categorized responses as training data. These algorithms are then tuned to make accurate predictions about the uncategorized responses.

Finally, an ensemble of the eight machine learning algorithms is constructed to combine the votes of each of the algorithms. The results of these ensemble categorizations show that students struggle to compare the variability between two graphs, even when students have a correct understanding of the statistical definition of variability. Students often assess the variability in a graph by the variation in the heights of the bars or dots. This research provides valuable information about the different ways students view variability in graphs and demonstrates a method in which constructed responses can be categorized in an automated fashion.

INDEX WORDS: Statistics education, Machine learning, Constructed-response items, Question development

INVESTIGATING UNDERGRADUATE STUDENT UNDERSTANDING OF GRAPHICAL DISPLAYS OF QUANTITATIVE DATA THROUGH MACHINE LEARNING ALGORITHMS

by

ALEXANDER JAMES LYFORD

B.S., The University of Georgia, 2011

M.S., The University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

INVESTIGATING UNDERGRADUATE STUDENT UNDERSTANDING OF GRAPHICAL DISPLAYS OF QUANTITATIVE DATA THROUGH MACHINE LEARNING ALGORITHMS

by

ALEXANDER JAMES LYFORD

Major Professor: Jennifer J. Kaplan

Committee: Nicole Lazar

Jaxk Reeves

Laine Bradshaw

April Galyardt

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

May 2017

Acknowledgements

I want to thank my friends and family for providing me support throughout this arduous process. I want to thank my mother, Connie-Sue, for providing me a safe haven every day where I can sleep in a warm bed, be with the family I love, and eat something other than pasta. I want to thank my advisor (and second mother), Jennifer Kaplan, for her companionship, unending wisdom, and a kindness so great that she read my dissertation in its entirety at least twice in the span of a week. I want to thank my father, Dan, for teaching me how to be a good son, a good husband, and a good person. I'm not quite there yet, but I wouldn't be close if it weren't for my father.

I want to thank my step-father, Harvey, for adopting me into his family and providing happiness in my life during times when I need a friend. I want to thank my step-mother, Sabrina, for keeping me grounded and reminding me that life isn't about me, but about kindness and compassion for others. I'm still working on that, too. I want to thank my sister, Jillian, for reminding me that siblings don't have to hate each other— I'm going to miss our 1 a.m. trips to Waffle House. I want to thank my grandfather, Harvey, for teaching me about the world and how to make a name for myself (in a good way). I want to thank my grandmother, Fran, for reminding me that family is forever and that compassion should know no bounds. Lastly, I want to thank my wife, Shannon, for taking this leap of faith in marrying me and putting up with me throughout my academic career.

There were times when writing this dissertation was incredibly rewarding. There were also times when I wanted to give up altogether. I wouldn't have made it this far if it weren't for the tremendous support of my friends and family.

Additionally, this material in this dissertation is based upon work supported by the National Science Foundation under Grant Number 1322962. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Contents

1 Introduction

In the world today, data are ubiquitous— it is nearly impossible to read an article, make a purchase, watch a show, or do almost anything without confronting some sort of graphical display of data. Because of this, there is an increasing need for basic understanding of graphical displays amongst all undergraduate students. One of the consistently more difficult tasks for students in introductory statistics courses of any level is interpreting the variability displayed in univariate graphs such as histograms and stem-and-leaf plots (Cooper and Shore, 2008). For non-statistics majors, these interpretations are useful for comparing the variability across different graphs, such as the distributions of ratings for two different products. For statistics majors, understanding univariate displays of data is necessary for understanding sampling distributions and discrete distributions like the Binomial or Poisson. Franklin et al. (2007) claims that understanding the omnipresence of variability in data is vital for anyone approaching or solving a statistical problem. In either case, it is desirable that undergraduate students in an introductory statistics course develop a solid understanding of the statistical definition of variability and its applications to real-world graphs.

For both statistics and non-statistics majors, it is important to track student progress through the development of their knowledge about graph, and immediately identify any related systematic misconceptions or deficiencies. Cooper and Shore (2008) suggest that instructors should consistently provide examples of different types of graphs with varying types of data to help students get a better understanding about how variability is represented in graphs. Most introductory

statistics courses, however, have hundreds (and sometimes thousands) of students (Blair et al., 2013). Thus, identifying the misconceptions held by these students is a daunting task. This dissertation proposes a construction of an ensemble of eight different machine learning algorithms to automate the categorization of student responses to open-ended questions about statistics. The models are initially trained by hand-scoring a subset of student responses to questions about graphical displays of quantitative, univariate data. These responses are then used to train the models in the ensemble to make predictions about future responses. The benefit of this approach is that the instructors of large-format, introductory statistics courses can give open-ended questions to their students and receive immediate feedback about what their students do or do not know about a multitude of topics. Tracking the quality of these responses over the course of the semester can demonstrate students' growth of knowledge. These models may then be employed on responses to questions administered to students in an introductory statistics course. The effectiveness of these models is both assessed at categorizing student responses and at providing useful feedback about what students do (or do not) understand about certain ideas related to variability as displayed in graphs.

This research aims both to create ensemble models and to use them to learn more about students' ideas about variability when reading graphs of univariate, quantitative data. This research takes a mixed-methods approach in an attempt to answer the following research questions:

1) **Dot Plot vs. Histogram:** Do students interpret graphs differently, including the graph's variability, when data are shown as a histogram versus a dot plot?

2) **Medium:** Does the proportion or prevalence of various misconceptions related to variability and graphs of univariate data change based on the medium

through which students communicate, including interviews and online assessments?

3) **Misconception:** What types of misconceptions do students have about variability in univariate data when the data are presented in graphs?

4) **Ensemble:** Can an ensemble of several machine learning algorithms accurately mimic hand scoring of student responses to questions about variability?

The quantitative analysis involves analyzing student responses to a variety of multiple choice and free response questions. The free response portions are analyzed using numerous machine learning algorithms together in an ensemble to make categorizations about student responses. The qualitative analysis is based on a series of task-based interviews with undergraduate students in an introductory statistics course at the University of Georgia. These interviews are used to gain deeper insight into student thinking about graphs of univariate data and to supplement the findings of the quantitative analysis.

1.1 Misconceptions about Variability in Histograms

This section contains a description of what the current literature says about how students read and interpret various graphical displays, including histograms, case-value bar charts, frequency bar charts, and dot plots. Some of the major misconceptions students hold with regards to these graphs are described in this section. Additionally, some of the correct and incorrect responses students give to questions about variability in graphs are identified. In the context of this study, the presence of a misconception is henceforth defined to be a response that gives an incorrect interpretation of data displayed in a graph of any form.

Elementary statistics, first encountered by most students in high school (although increasingly more so in middle school), may be taught by an enthusiastic mathematics teacher with no formal statistical training (Garfield and Ben-Zvi, 2004). It is often taught, however, by a less-than-enthusiastic general math teacher begrudgingly completing a short statistics unit situated toward the end of the course (Garfield and Ben-Zvi, 2004). In both circumstances, statistics as a discipline may be portrayed in a narrow, limited fashion in which seemingly every statistical phenomenon can be depicted through a bar chart (Ben-Zvi and Amir, 2005). This depiction may stem from the minimal statistical knowledge of the teachers and the limited curriculum used for instruction (Bright and Friel, 1998). One of the drawbacks of this curriculum is that most charts, such as bar charts, histograms, and ribbon plots, are referred to by instructors as *bar charts* (Konold and Higgins, 2002). Cooper and Shore (2010) found that 64% of K-12 teachers identified a histogram as a bar chart. For students in courses taught by such instructors, this approach to teaching statistics could systematically reinforce a set of incorrect notions about the presentation of univariate data in any two-dimensional form, leading to a number of common misconceptions about the variability in such data.

The first of these misconceptions, the **Bar Chart Misconception**, is that the depiction of any two-dimensional graphical display of univariate data containing bars must be a case-value bar chart (or case-value plot) and thus its bars must be displaying individual, non-aggregated data (Konold and Higgins, 2002). This implies that students often only see bars as a representation of an individual data point and not as an aggregation of a number of data points (delMas et al., 2005). When data are displayed in a histogram, this fundamental error in reasoning might inhibit the student from correctly ascertaining many important characteristics of the distribution as visualized in the graph, especially its variability. This error in

reasoning may occur because students, thinking that a histogram is equivalent to a bar chart, interpret the heights of each bar as individual data values and not a frequency of many data values (delMas et al., 2005). In this manner, many students would identify the frequency value of a particular bar and believe that it represents a value of the quantitative variable of interest from a single person. This inability to separate depictions of individual data points (bar chart) and aggregated data (histogram) has been shown to be reversible through deliberate instruction (Ben-Zvi and Arcavi, 2001).

A second common misconception held by elementary statistics students, the **Bar Height Misconception**, is that the variability in a given histogram is depicted through the variation (or lack thereof) in the heights of its bars (Cooper and Shore, 2010; Chance et al., 1999). This misconception typically, though not necessarily, follows from **Bar Chart Misconception**. That is, students who think that the heights of the bars represent individual data values will in turn assess the variability of the data in the graph by looking at the variability of the heights of the bars (Garfield and delMas, 1990). A student without **Bar Chart Misconception**, however, may still possess this second misconception; even students with a correct understanding of the difference between the axes in a bar chart and a histogram will more often than not choose a bumpier histogram as being more variable than a flatter one (Meletiou-Mavrotheris and Lee, 2010).

A third misconception that students often have regarding the identification of variability in histograms, the **Axis Order Misconception**, is the idea that the ordering of the x-axis is arbitrary (delMas et al., 2005). Since bar charts depict categorical variables, the bars and associated categories do not have an order, in contrast to histograms, which depict a quantitative variable. The presence of this misconception may lead to erroneous interpretations of the overall characteristics of

the histogram, including its variability. Much like with **Bar Height Misconception**, believing that the x-axis is arbitrarily ordered leads to an inability to correctly identify the approximate location of the mean, therefore making it difficult to assess the level of variability in the data presented in the histogram.

A fourth and final misconception regarding the assessment of variability in histograms, the **Range Misconception**, is that students treat the range of the data set as the only determining factor when deciding which of two data sets are more variable. This misconception implies that a data set containing a larger range must necessarily be more variable than one with a smaller range (Meletiou-Mavrotheris and Lee, 2010). Though this misconception is not unreasonable, it may directly interfere when defining variability as a measure of deviation of a value from some measure of center. Specifically, assume one histogram contains data that have almost no variability but a few outliers that cause the range of the data to be relatively large. Assume a second histogram depicts data that are relatively uniformly distributed over a reasonable range and thus highly variable. Students with this misconception often claim that the first graph displays data that are more variable than the second solely because the range is larger, despite clear clustering of data in the first graph. This misconception is typically only manifested if students do not have any of the first three misconceptions, as it requires a correct understanding of the purpose of the x- and y-axes of a histogram (Garfield, delMas, and Chance, 1999).

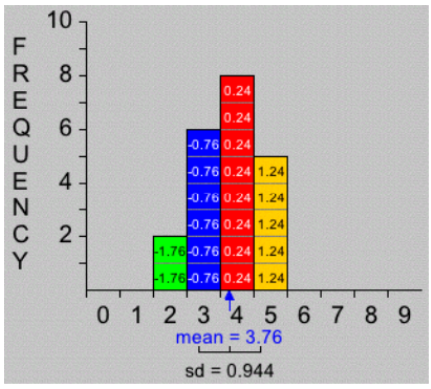
One potential solution to correct these misconceptions is through carefully designed instruction in the first instance in which bar charts and histograms are encountered, as it has been shown to increase student understanding with regard to interpreting variability in a histogram (Meletiou-Mavrotheris and Lee, 2010). While it may be impossible to fully correct these misconceptions without such instruction (or even with it), presenting identical data in slightly different fashion (say, a dot

plot), may decrease the propensity to commit many of the aforementioned errors by fundamentally changing the manner in which students approach a two-dimensional display of univariate data. Bakker (2004) demonstrates that the type of graph used to display data (e.g., dot plot, case-value plot, bar chart) is related to a student's ability to correctly perceive the shape of the distribution. Bakker (2004) also finds that dot plots seem to be the best starting point for students when trying to learn about the shape of a distribution. These same influences may hold true for a student's ability to correctly perceive variability.

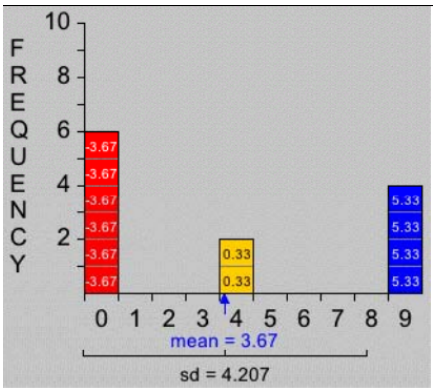
1.2 Histograms vs. Dot Plots

There are some advantages of displaying univariate data in a dot plot versus a histogram. First, it highlights individual data values stacked on top of each other as part of a whole, instead of one solid bar like in a histogram. Segmenting bars (or separating the bar into individual boxes like dots in a dot plot) has been shown to increase student understanding of how far individual values are away from the mean (delMas and Liu, 2003). This segmentation is instrumental in understanding the amount of variability present in the data, as having lots of individual data values far away from the mean implies a much larger variability. An example of such a segmentation can be seen in Figure 1.1 (from delMas and Liu, 2003). With the segmentation, students are able to visualize the individual data points that make up each bar in the histogram (equivalent to the representation of a dot plot) and the corresponding deviation from the mean contributed by each individual data point. When students compared the variability present in subfigure 1.1(a) with subfigure 1.1(b) after segmenting and labelling the histogram (now essentially a dot plot) in this manner, they were able to determine almost unanimously that the data in subfigure 1.1(b) contained more variability than those displayed in subfigure 1.1(a).

Figure 1.1: Interview Tasks from (delMas and Liu, 2003)



(a) A demonstration of segmented bars, nearly isomorphic to a dot plot.



(b) A second graph with segmented bars displaying data with more variability.

This same depiction paired with the one in Figure 1.2 helped students correct the fourth misconception, that the range of the data is not equivalent to its variability. When given the segmented data and labeling, the deviation in the same way, students were able to visualize why Figure 1.2 displays more variable data than either plot in Figure 1.1.

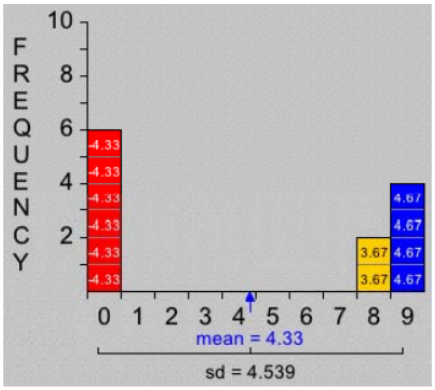


Figure 1.2: A third graph with segmented bars displaying data with more variability than the previous two.

While it is unclear how much of this newfound understanding was a product of the segmentation of the bars or the labeling of the deviation, it was clear

that both had some impact on improving understanding of variability in the data (delMas and Liu, 2003). Additionally, it has been shown that asking students to write data values on sticky notes in order to create a line plot leads to a greater understanding of the purpose of the x-axis (**Axis Order Misconception**) and a better idea of the aggregated data (**Bar Chart Misconception**) (Bright and Friel, 1996). This representation, isomorphic to a dot plot, helped students understand both the numerical ordering of the x-axis, but also the meaning of the frequency in the y-axis.

There does, however, exist some precedent that dot plots may not be a perfect substitution for histograms in all cases. In a Fright and Briel (1995) study, middle school students who did not display appropriate aptitude for understanding of the x- and y-axes for bar charts also misinterpreted the meaning of the dots in a dot plot with only one axis (meaning no y-axis for frequencies).

1.3 Interplay between Shape, Center, and Variability

Many research studies have explored the connection between a student's understanding of the shape, center, and variability of a distribution displayed graphically. Cobb (1999) finds that students need to understand shape before being able to learn about the variability in a graph. Following the **Bar Height Misconception**, students sometimes equate certain shapes of graphs with certain amounts of variability (Kaplan et al., 2014). For example, students may look at the probability distribution function of a uniformly distributed variable and claim that it has significantly more variability than a bell-shaped distribution of variable simply because it is flatter (Kaplan et al., 2014).

Because the fundamental normative description for variability, a numerical summary of the magnitude of the distance of points from the center of a distribution, is so inherently connected with center, it is difficult for students to assess the magnitude of the variability in a given graph without understanding how to determine the center of the graph (delMas and Liu, 2005; Cobb et al., 2003). Students tend to develop understandings about the center of distributions displayed graphically at a very young age (Ben-Zvi and Amir, 2005). Students, however, may hyper-focus on the central values while ignoring many other important features of graphs such as their variability or shape (Kaplan et al., 2014). Because of this, it is important to highlight the natural connection between the center of a graph and the variability in its data.

There are studies (delMas and Liu, 2003; Bakker 2004) that demonstrate that horizontal segmenting in frequency graphs helps students identify which of two graphs contain data that are more variable. In both of these studies, students were given specific instruction about variability and graphs prior to being asked questions. In the delMas and Liu (2003) study, students used computer software to play a game that taught them how to identify graphs with larger standard deviations. In the Bakker (2004) study, students were also given several lessons about how to identify variability in different graphs. Although it is clear that the segmented graphs led to understanding of variability when data are presented graphically, it is unclear whether this was a result of the direct instruction or because of the segmentation itself. This research is conducted with students who have received no additional instruction about variability and graphs outside of the standard introductory statistics curriculum. While there is some literature about student misconceptions about histograms (Kaplan et al., 2014), there is sparse literature discussing the misconceptions associated with other univariate displays of data, such as dot plots.

The work presented here explores the connections between dot plots, histograms, and abilities of students to correctly compare measures of variability across data sets depicted graphically. Student responses are collected from large numbers of students in an introductory statistics course at the University of Georgia. These responses are analyzed using an ensemble of machine learning algorithms designed to categorize student responses into a number of bins. Chapter 2 and Chapter 3 present the statistical theory behind the ensemble model used in this study and provides details of the mixed methodology employed to answer the four research questions posed by this work. Chapter 4 contains the results of the mixed methods study, and Chapter 5 provides a discussion of the results. In addition, limitations and potential for future directions are discussed.

2 Ensemble and Machine Learning Literature Review

To investigate the **Ensemble** research question, a series of machine learning algorithms are constructed in tandem to make classifications about uncategorized student responses to statistics content questions, specifically about variability. A machine learning algorithm is a class of statistical algorithms that can learn from and make predictions about data (Kohavi, 1998). These classifications are used to identify possible misconceptions held by students (**Misconception** research question) as well as the differences in responses to questions about dot plots versus histograms (**Dot Plot vs. Histogram** research question). In essence, a series of proven classification techniques are used to make accurate, overarching classifications about new data. To achieve this, a technique known as ensemble learning is used to combine the classification abilities of multiple machine learning algorithms. Although machine learning algorithms can be used for regression, here they are used solely for classification. As such, the terms algorithm and classifier are used interchangeably throughout this dissertation.

2.1 Ensemble Modeling

Ensemble learning is a technique that employs multiple machine learning algorithms, because the combined knowledge of multiple algorithms often produce more accurate results than any one algorithm alone (Dzeroski and Zenko, 2004). This ensemble of algorithms is more accurate when the number of algorithms in the ensemble is very

large (Hansen and Salamon, 1990). In this context, each algorithm in the ensemble makes a single classification prediction for any given observation (i.e. student response). Under uniform vote weighting, the vote of each algorithm is counted, and the classification that receives the majority vote is the classification selected by the ensemble. There are, however, many ways to weight the votes of each of the classifiers in the ensemble. Regardless of vote weighting, it has been shown that ensemble methods often produce a lower classification error rate than the error rate of any of the individual classifiers in the ensemble, including the best one (Dzeroski and Zenko, 2004).

Ensemble methods for classification can only be more accurate than any of their individual classifiers if the individual classifiers in the ensemble are both 'diverse' and 'accurate' (Hansen and Salamon, 1990). There is, however, no formally accepted definition for ensemble diversity (Kuncheva and Whitaker, 2003). Thus, measuring diversity is not straightforward. Some define diversity as having each of the errors in classification made by individual machine learning algorithms in a given ensemble be uncorrelated with one another (Hansen and Salamon, 1990). Others define diversity using a single metric, like Yule's Q statistic, a special case of Kruskal's gamma (Kuncheva and Whitaker, 2003).

This research utilizes Cohen's Kappa (Cohen, 1960) and Fleiss' Kappa (Fleiss, 1971), a generalization of Cohen's Kappa, to ensure diversity by determining if the errors made by each of the classifiers in the ensemble are uncorrelated. Cohen's Kappa is used for analytic rubrics (defined in Section 7.3) and for quantifying the coding agreement between two independent raters. Cohen's Kappa values larger than .8 are said to demonstrate strong agreement (Cohen, 1960), and this threshold is used as a baseline for determining if raters are agreeing sufficiently often. Fleiss' Kappa is used for holistic rubrics and any instances where there are

multiple raters. For this purpose, multiple raters could either be three or more experts coding a particular response, or three or more machine learning algorithms coding a particular response.

To demonstrate the calculation of Fleiss' Kappa, assume the following. Let N be the total number of documents in the document-term matrix (see Section 7.3 for a definition of a document and document-term matrix). Let n represent the number of individual classifying algorithms in the ensemble. Let K represent the number of mutually exclusive rubric bins, or alternatively the number of potential categorizations for the given problem. Documents are indexed $i = 1, ...N$, categories are indexed $j = 1, ..., K$, and n_{ij} represents the number of algorithms that classified the i-th document into the j-th category. First, calculate:

$$\begin{aligned} \overline{P} = & \quad \tfrac{1}{N} \textstyle\sum_{i=1}^{N} P_i \\ \overline{P} = & \quad \tfrac{1}{N \cdot n(n-1)} \big(\textstyle\sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2 - n \big) \end{aligned} \tag{1}$$

Where P_i represents the proportion of pairs of algorithms that agreed on a classification for document i. That is, $P_i = 1$ if each of the algorithms agreed on the same classification (regardless of its correctness) and $P_i = 0$ if each of the algorithms chose a different classification than the rest. Thus, $\overline{P}$ calculates the average of all of these agreements. Next, calculate:

$$\overline{P}_e = \sum_{j=1}^{K} p_j^2 \tag{2}$$

Where p_j represents the proportion of all classifications made by the algorithms belonging to classification j. Thus $\overline{P}_e$ represents the sum of the square of the proportions of classifications made for each category. That is, if all documents were classified as category 1, then $p_1 = 1$ and $p_2 = p_3 = ... = p_K = 0$. If half of the documents were classified as category 1 and half as category 2, then $p_1 = p_2 = 0.5$ and $p_3 = p_4 = ... = p_K = 0$, etc.

Finally, the formula for Fleiss' Kappa is:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$$

Where $1 - \overline{P}_e$ represents the degree of agreement that is attainable beyond random chance, and $\overline{P} - \overline{P}_e$ represents the degree of agreement that is actually achieved by the ensemble beyond random chance. Equation 1 and Equation 2 are used to calculate $\overline{P}$ and $\overline{P}_e$, respectively.

Diversity, specifically when measured as a function of the correlation of the errors each classifier makes, is important in an ensemble. Diverse classifiers are important so that mistakes in a certain direction made by one classifier (say the propensity of a certain classifier to over-classify responses into bin 1) are not made in the same direction by all classifiers in the ensemble (as they would if the classifiers were not diverse). In other words, it desirable to not have all of the classifiers consistently make errors in the same direction, or more ideally to make errors randomly in any direction instead of systematically in one direction (Banfield et al., 2005). To demonstrate the importance of diversity, consider the extreme case where every classifier in the model is identical. That is, say there are k identical classifiers in an ensemble, each with correct classification probability equal to 0.8.

Because the classification of classifiers $1, 2, ..., n$ is all exactly the same, the true probability that the ensemble makes the correct decision is not depicted by Figure 2.1, but is instead equal to 0.8. This is because the errors that the classifiers make are correlated (in this extreme example, they are perfectly positively correlated). Thus, it is necessary to have uncorrelated (or at least minimally correlated) classifiers to avoid this issue.

Hansen and Salamon (1990) state that an ensemble is 'accurate' if the probability that each individual classifier correctly categorizes any given item must be greater than 50% in the long run. Take as an example the case in which a researcher wishes to categorize responses into one of two bins, 1 and 2. Suppose the probability that any given response belongs to one of these bins is equal (thus, $p(Bin1) = p(Bin2) = .5$). A classification algorithm that employed purely random guessing would correctly categorize any given response with probability 0.5. If the correct classification rate for each classifier in the ensemble is exactly 50% (equal to random guessing), then regardless of the number of classifiers in the ensemble, the overall probability that the majority vote in the ensemble is in favor of the correct classification is exactly 50%. This is true because each of the algorithms has only a 50% chance of correctly classifying a given response. Thus, any arbitrary combination of these algorithms would also produce only a 50% chance of correctly classifying the response, regardless of how each algorithm's vote is weighted. To construct an ensemble to successfully categorize items into one of these two bins, the accuracy of each individual algorithm must therefore be better than 50%.

Assuming each algorithm in the ensemble is equally accurate and its classifications are independent, the probability that a given ensemble correctly categorizes a particular item is equal to:

$$P(\text{Correct Classification}) = \sum_{k=\lceil \frac{n}{2}+.5 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

where n represents the total number of algorithms in the ensemble, and p represents the probability that any one algorithm correctly classifies a given response (i.e. its accuracy). Thus, the probability that the ensemble makes a correct classification is the probability that the majority of the algorithms in the ensemble (at least $\lceil \frac{n}{2}+0.5 \rceil$ of them) makes the correct classification.

Take for example the following 4 separate instances. In each of these instances, the probabilities of making a correct classification for each of the classifiers in the ensemble are fixed to be 0.5, 0.55, 0.6, and 0.8 respectively. The y-axis represents the expected probability that the majority vote of the ensemble is correct. Note that ensembles with 95% long-run accuracy whose individual algorithms have 55%, 60%, and 80% accuracy require approximately 200, 100, and 5 total algorithms. Thus, increasing the accuracy of individual algorithms greatly reduces the number of algorithms required in an ensemble to achieve high overall accuracy.

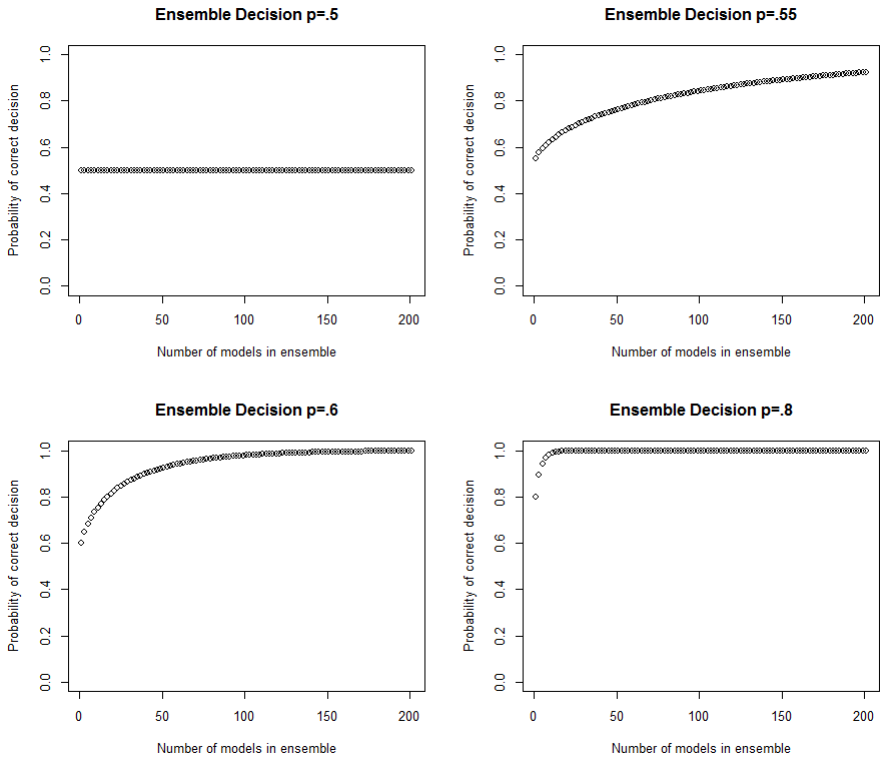


Figure 2.1: The expected probability of a correct ensemble decision for various classifiers with fixed probabilities of correct classification.

If there are n number of bins into which items can be classified, then each individual classifier must still be able to correctly classify any given item with probability greater than 0.5. Unlike in the aforementioned example where $n = 2$, this is not equivalent to saying that each classifier must decide, with probability greater than 0.5, that a given item belongs in the correct classification bin. For example, suppose there are $n = 4$ bins. Classifier A might calculate the following probabilities that Item Z belongs in each of the 4 bins:

$$P(\text{Item Z belongs in Bin 1}) = \quad 0.4$$
$$P(\text{Item Z belongs in Bin 2}) = \quad 0.3$$

$$P(\text{Item Z belongs in Bin 3}) = \ 0.2$$
$$P(\text{Item Z belongs in Bin 4}) = \ 0.1$$

Suppose that Item Z belongs in Bin 1. Even though Classifier A only assigned a 40% probability that Item Z belongs to Bin 1, Classifier A would place Item Z in Bin 1 because it had the largest probability of all available bins. For a classifier to be 'accurate' enough to be placed into an ensemble, it must correctly categorize, on average, at least half of the responses in any given training set of data. This is true regardless of the individual probabilities that any item belongs to any of the n bins.

Both the conditions of perfect accuracy (an algorithm that predicts the correct categorization with probability 1) and perfect diversity (an ensemble of completely uncorrelated algorithms) are rarely met in practical cases. For example, sometimes one or two classifiers individually perform poorly or are highly correlated with each other. Typically, steps are taken in each ensemble technique to ensure that they are met with as high a probability as possible (Banfield et al., 2005). Some classification methods, such as Bagging and Boosting help reduce classifier correlation through bootstrap sampling. Random Forests is a technique that employs random feature bagging to reduce classifier correlation. Therefore, except under extreme circumstances that guarantee individual poor performing or highly correlated classifiers, for example extremely small sample sizes, it is always advantageous to use ensemble modeling instead of an individual classifier (Banfield et al., 2005).

The aforementioned calculations assume that each vote in the ensemble is weighted equally. There are, however, many different ways in which the votes of each algorithm in an ensemble can be weighted. A voting scheme in which each algorithm's vote receives equal weight in the classification chosen by the ensemble is

often simple and effective (Dietterich, 2000). This is known as uniform weighting. There are many other weighting schemes that calculate the weight of an algorithm's vote based on various criteria (Muhlbaier et al., 2009). Fung et al. (2006) proposes a method of calculating the weight of an algorithms vote based upon the location of the response in n-space. That is, algorithms that predict particularly well for certain areas are given higher weight than those that do not predict as well in the same area. Other, cross-validation-based vote weighting schemes prove effective when particular algorithms have a high rate of false positives or false negatives (Dietterich, 2000). A variety of ensemble vote weighting functions are utilized in this research.

2.2 Ensemble Weighting Functions

Each algorithm in an ensemble votes on the classification of a particular response. There are numerous different vote-weighting schemes that can be used to dictate how votes from each algorithm are counted in the ensemble. The general form of an ensemble vote-weighting function is:

$$\tilde{y}(x;\alpha) = \sum_{j=1}^{p} \alpha_j y_j(x)$$

where $\tilde{y}(x;\alpha)$ represents the ensemble classification for response x using vote-weighting function α, p represents the number of unique classification algorithms in the ensemble, $y_j(x)$ represents the predicted classification of response x by algorithm j (where $j = 1, ..., p$), and α_j represents the weight applied to algorithm j's classification. Although there are no restrictions about the values of α_j, typically $0 < \alpha_j < 1$ for all j. That is, suppose the vote from algorithm 1 (with weight α_1) is valued three times

more than the vote from algorithm 2 (with weight α_2). Any scalar value for α_1 could be valid so long as $\frac{\alpha_1}{\alpha_2} = 3$. Thus, it is acceptable that $\alpha_1 = 3$ and $\alpha_2 = 1$. These weights, α_j, are typically (though not necessarily) standardized such that their sum is equal to 1. In the previous example, a standardization in weights would yield $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = \frac{1}{4}$ so that $\frac{\alpha_1}{\alpha_2} = 3$ and $\sum_{j=1}^{2} \alpha_j = 1$.

Uniform vote weighting, also known as a majority voting rule, states that the vote of each algorithm is weighted equally. Thus, under uniform vote weighting, $\alpha_j = 1$ for all j and the classification receiving the most votes (regardless of which algorithms produced the votes) is the classification chosen by the ensemble.

Each algorithm reports a measure of certainty that the algorithm has made a correct prediction. This measure of certainty is known as the predicted probability that is has made a correct prediction. Probability-based voting weights the vote of each algorithm by this measure of certainty. Thus, the weights in predicted-probability based voting are $\alpha_j = P_j$, where P_j represents this measure of certainty (Muhlbaier et al., 2009). The calculation involved in the predicted probability of a correct classification is different for each of the eight algorithms. For example, a classification tree calculates this predicted probability of success by identifying the proportion of responses at a particular terminal leaf with the given classification. That is, if X out of the N total training data responses belonging to a terminal leaf are coded as category A, then the classification tree algorithm outputs a classification of category A with predicted probability of success to be $P_j = \frac{X}{N}$. Although other algorithms calculate this measure of certainty in a slightly different manner, most algorithms assess their predicted-probabilities of success as a measure of the proportion of responses with the target classification in the given subspace (Muhlbaier et al., 2009).

A dynamic ensemble weighting function weights the votes of the algorithms depending on the location of the response in n-space. There are an infinite number of dynamic vote-weighting functions, but the one used in this research is called Dynamic Classifiers Weighting and was developed by Fung et al. (2006). In two dimensions (i.e. if there were only two features in the feature space), a particular algorithm may more correctly categorize responses located in the first quadrant than in the second quadrant. The corresponding α_j value would be higher (i.e. the algorithm's vote's weight would be higher) when making predictions about responses located in the first quadrant than responses located in the second quadrant. Significant amounts of training data are used to determine the efficacy of the algorithms in different regions, and this is often computationally expensive (Fung et al., 2006).

Cross-validation-based (or CV-based) vote weighting schemes weight the votes of the algorithms in an ensemble by various performance metrics calculated during the cross-validation step of the model-training process. Although there are an infinite number of CV-based vote weighting schemes, the one chosen in this study weights the vote of each algorithm based on its propensity of false-positives (if the predicted categorization is positive) or false-negatives (if the predicted categorization is negative). For this implementation, $\alpha_j = 1 - FPP$ or $\alpha_j = 1 - FNP$, where FPP represents the probability, based on cross-validation testing, that the algorithm make a false positive prediction and FNP represents the probability, based on cross-validation testing, that the algorithm will make a false negative prediction. The former equation is only used when the algorithm makes a positive categorization, and the latter equation is only used when the algorithm makes a negative categorization. In the instances where there are greater than two categories, each category is collapsed into a binary form (i.e. is the response in the particular category or not) for the purposes of these calculations.

2.3 Machine Learning Algorithms in the Ensemble

Machine learning algorithms involve inferring a classification function from a set of labeled training data. The general form of machine learning for classification involves training the classifiers on a set of training data, and then testing their accuracy on a separate set of testing data (Adeli and Hung, 1994). Both sets of data must already contain labels for machine learning to occur. Subsequent sections discuss each individual machine learning algorithm to be used in the classification ensemble.

2.3.1 Classification Trees

Classification Trees, also known as Decision Trees or Survival Trees, are a machine learning technique that creates a binary decision making structure to classify new data into one of potentially many mutually exclusive categories. The basic structure of a classification tree begins with a central node (known as a leaf) with at least two attached branches (Rokach and Maimon, 2008). For the purposes of this research in classifying student responses to questions about statistics, the central leaf might say, *Did the student use the phrase 'Central Limit Theorem?'* This leaf would then have two branches- one labeled *yes* and one labeled *no*. Each of those branches would lead to another leaf that might ask a similar question such as, *Did the student use the word 'Expectation?'* This leaf would have two more branches, and the process would repeat until a set of specific stopping criteria were achieved.

There are many different ways to construct classification trees. One of the more common ways is through maximizing the homogeneity of classifications at any given leaf. Homogeneity is a measure of the proportion of responses at a given leaf that have the same categorization. If all of the responses at a particular leaf have the same categorization, then this leaf is said to be perfectly homogenous.

The classification tree constructed in this manner is derived by first constructing attribute-value pairs for each of the text responses in the data set (Gelfand et al., 1989). An attribute is defined as any of the words appearing in the training data set, and its corresponding value is either a 1 or 0 if the word were used or not used in the selected text response, respectively. The attribute-value pairing that separates the training data into homogenous categories is chosen to be the central leaf (Srivastava et al., 1999). If the presence or absence of a certain attribute always identifies each response in the training data as belonging to one category or the other, then this attribute is selected to be the central leaf. This rarely occurs for all but the smallest data sets (Gelfand et al., 1989).

Assuming there does not exist an attribute that splits the training data into homogenously categorized groups, the classification tree algorithm cycles through all remaining attributes until it identifies one that promotes the highest degree of homogeneity possible. This is known as maximizing homogeneity. For example, assume there are 100 observations with two classifications, correct and incorrect, and two remaining attributes A and B. Next assume that the use of attribute A would split the remaining sample into two groups, a *yes* group and a *no* group, where the *yes* group and *no* group each have 25 *correct* classifications and 25 *incorrect* classifications. Now assume that the use of attribute B would also split the remaining sample into a *yes* group and a *no* group, but there would be 45 *incorrect* classifications and 5 *correct* classifications in the *yes* group, and conversely 5 *incorrect* classifications and 45 *correct* classifications in the *no* group. In this example, attribute B would be chosen as the next leaf because it maximizes the homogeneity within the groups. Table 1 and Table 2 give a summary of Attributes A and B.

Once the attribute for the central leaf is identified in this manner, it now has two branches extending from it: one indicating the presence of the attribute in

a given text response and one indicating its absence. A new leaf is formed at each of these branches, and the process of selecting the leaf that maximizes the homogeneity repeats in an identical manner to the selection of the first leaf. (Quinlan, 1986). This process continues until a set of stopping criteria are reached. In general, the stopping criteria include a hard stopping criteria, if a node splits the data into perfectly homogeneous categorizations, and soft stopping criteria, if there are too many branches in the tree (Quinlan, 1986). The specifics of the stopping criteria used here can be found in Section 3.7.1.

Table 1: Attribute A summary

		Classification	
		Correct	Incorrect
Branch	Yes	25	25
	No	25	25

Table 2: Attribute B summary

		Classification	
		Correct	Incorrect
Branch	Yes	5	45
	No	45	5

Once a classification tree is formed in this manner, it can be used to classify new, uncategorized responses. To categorize new responses, the algorithm begins at the central leaf and determines if the response contains the given attribute. Then, it follows the branch corresponding to whether or not the given attribute is present or absent in the response. It then arrives at a new leaf containing an attribute. The algorithm continues this process until the bottom of the tree is reached. If the bottom-most leaf contains homogenously categorized responses in the training data, then the new response receives this same categorization. If the training data were

not homogenously split at this final leaf, then the split of larger size is chosen to be the categorization of the new response (Magerman, 1995).

2.3.2 Bagging Classification Trees

Bagging is an ensemble learning technique which utilizes multiple resamplings of training data to construct several different classification algorithms. Bagging hypothesizes that a series of votes from these separate classifiers derived from the sets of re-sampled data makes more accurate classifications than any one classifier, including one derived from the original data (Dzerosky and Zenko, 2004). In other words, in an attempt to reduce both the overall variability of classifications and classification error rate, Bagging partitions the N data points into a training and a test set, say T_{train} and T_{test} respectively. Then the N_{train} data points in T_{train} are each given equal probability of being selected ($p(n) = \frac{1}{N_{train}}$) and a sample of size N_{train} is drawn with replacement and a classifier is created. This process is repeated a desired number of times (say 100) and each of the classifiers is used in an ensemble to vote on the classification. The classification that received a majority vote is selected to be the most likely classification (Breiman, 1996).

One way to develop a classifier is to bag classification trees. In this manner, a set number (typically 100) of bootstrapped samples are drawn from the population and a classification tree is constructed for each of them. The classification trees are created in such a way as to maximize efficiency, so that numerous trees can be constructed in a reasonable amount of computing time (Dzerosky and Zenko, 2004). This increase in efficiency may lead to small decreases in individual classifier accuracy, especially when classifying outlier points. Bagging seeks to overcome this by having each classification algorithm vote in an ensemble (Breiman, 1996). It is expected that with 100 bagged trees to use in an ensemble to make a classification,

the number of trees that were created 'too quickly' and missed constructing specific leaves necessary for intricate classifications are only a fraction of the overall number of trees in the ensemble. Each of the trees in the bagging ensemble gets one vote, and that bagging ensemble's vote is worth one vote in the overall ensemble.

2.3.3 Boosting Decision Stumps

Boosting is a machine learning algorithm that turns weak learning algorithms that perform slightly better than random guessing into a single ensemble of algorithms with arbitrarily high accuracy (Freund and Schapire, 1995). Weak learning algorithms are those that are quick to construct, but are often not as accurate as algorithms whose run-times are significantly longer (Freund and Schapire, 1999). Boosting, much like bagging, utilizes the re-sampling of data to generate new classifiers. While Boosting and Bagging both work to combine many classifiers into a single, improved classifier, Boosting modifies the re-sampling probability distribution using a set function (Freund and Schapire, 1996). There are many functions that can be used to re-weight the training data using to train the individual classifiers. Some of these re-weighting functions include: AdaBoost (Freund and Schapire, 1999), LP-Boost (Demiriz et al., 2002), and LogitBoost (Friedman et al., 2000). Each of these algorithms performs well under certain circumstances (Friedman et al., 2001). The AdaBoost re-weighting function tends to perform better than other known schemas when decisions stumps, a one-leaf decision tree with two branches, are used as the weak learners in an ensemble (Friedman et al., 2000).

Bagging is essentially a variance-reduction technique and is useful for highly variable structures like decision trees (Breiman 1998). Decision stumps, however, have very little variance and often high bias. Thus, bagging performs poorly with decision stumps, whereas re-weighting algorithms such as AdaBoost perform

very well by minimizing both the inherent bias in the creation of each stump (Friedman et al., 2000).

The essence of Boosting involves varying the probabilities of being selected for the bootstrapped sample for each of the responses in the training set of data (Breiman 1998). Unlike bagging, in which each response has an equal probability of being selected at each iteration of training, in boosting successive iterations increase the probability of the selection of misclassified responses for the new training set by some function (Freund and Schapire, 1999). In its simplest form, this increase in probability involves multiplying the probability of being selected for each of the misclassified points, then re-normalizing these probabilities to sum to 1 . The advantage of this re-weighting of probabilities is that in most cases it leads to significantly reduced prediction error when compared to Bagging, because it systematically lowers the probability of many easy-to-classify cases being selected in the re-sampled data sets, as these cases ultimately do not help improve the accuracy of each classifier (Breiman 1998). The disadvantage of this method is, for some outlier cases, the classifiers often continuously misclassify the outlier cases, causing them to appear in every re-sampled data set (since they have such a high probability of being sampled in each successive bootstrapped sample since it keeps getting misclassified) (Freund, 1999). This leads to an overall decrease in final Boosting ensemble prediction accuracy (Breiman 1998).

The most prominent utilization of the boosting of decision stumps uses the AdaBoost re-weighting algorithm (Freund and Schapire, 1999). This re-weighting algorithm constructs hundreds of decision stumps based on the boosted data set. Unlike in the boosting described by Breiman (1998), AdaBoost applies a more complex function to the reweighting scheme. Given a finite number of classifications, AdaBoost first constructs an error function $E(f(x), y, i) = e^{-(y_i f(x_i))}$ where x_i rep-

resents the ith resampling of the data, $f(x_i)$ represents the classifications by the decision stump created by the ith resampling, and y_i represents the vector of correct classifications for sample i (Freund and Schapire, 1999). The problem then becomes a minimization problem where ones tries to find the decision stump (essentially the best single attribute) that minimizes the error function as a function of the weight applied to each data point. Thus, ones tries to find the decision stump $h(x)$ that minimizes $\sum_{i=1}^{n} w_i E(h(x), y, i)$ where w_i is the vector of weights assigned to each of the i data points. One then minimizes this function at each of the boosting iterations until one has reached a designated level of convergence or stopping criteria (Freund and Schapire, 1999). The weights of individual data points themselves are determined by the function $w_{i,t+1} = w_{i,t} e^{-y_i \alpha_i h_t(x_i)}$, where $w_{i,t}$ represents the vector of weights of sample i at iteration t, and α is a function of the error of the previous set of decision stumps (Freund and Schapire, 1999).

An effective visualization explaining the process and utility of boosting comes from Meir and Ratsch (2003), and is pictured on the following page. The color of the observations indicates their correct classification, and the diameter of the point indicates that observation's proportional weight in the sample. The dashed lines represent the boundaries formed by individual classifiers, and the solid green line represents the decision boundary formed by the combined classifier. Figure 2.2 shows that a strong learner (first iteration) splits the data reasonably well. In the second, third, and fifth iterations, the individual classifiers (pink dashed lines) do not perform particularly well. Specifically, the classifier with a circular-shaped boundary splits the data extreme poorly. The figure also shows the results after a hundred iterations of constructing several weak learners about data that are hard to classify. In this case, 'hard to classify' is synonymous with being close to the boundary. The overall boosted ensemble of weak learners classifies much better than the original

strong learner due to the sections of easy to classify points, (in this example in the top right and bottom left of the data), being removed from the training data as they give no real predictive power after the first iteration. The difficult to classify points, however, remain in the data set and often appear multiple times. This allows for a more refined classification line to be drawn.

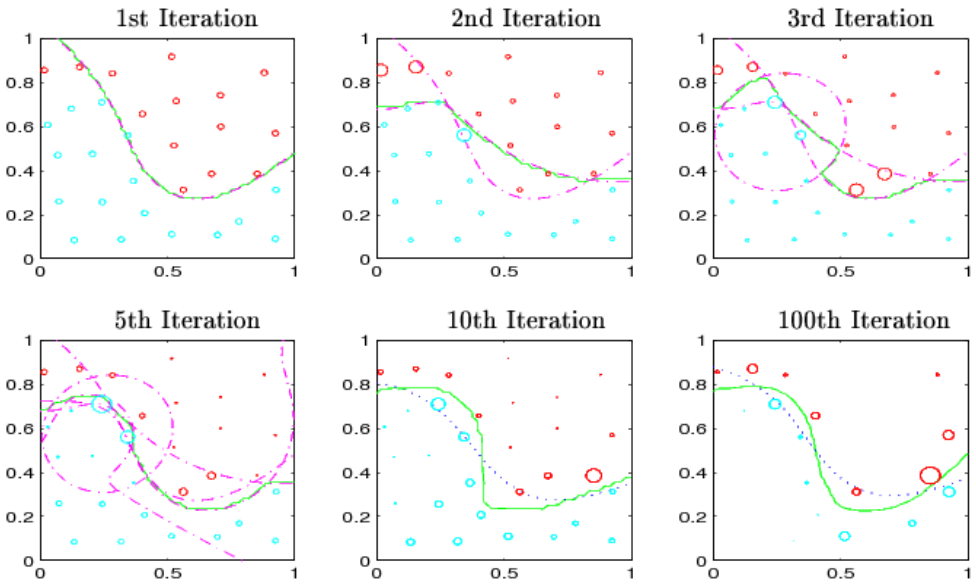


Figure 2.2: This figure details the boosting process. All data points have an equal chance of being selected in the first iteration of the resampling of the training data. Hard-to-classify data points (i.e. those near the boundary) are given more weight and appear more often in future resamplings.

2.3.4 Random Forests

Random forests describe the general ensemble modeling technique of constructing many classification trees (as described in Section 2.3.1) and outputting the majority classification of the forest of classification trees (Liaw and Weiner, 20012). Each individual classification tree is a classifier in the ensemble, and the overall ensemble is known as the random forest. This method utilizes the idea of bagging (as described in Section 2.3.2), but instead of bagging the individual classification trees,

random forests utilize the bagging of features. It is difficult to select useful attributes when the number of attributes is significantly larger than the number of responses (Ho, 1998).

The random selection of features to be used is calculated via the random subspace method. Let $g = 1, ..., F$ be the total number of features and n_{tree} be the total number of decision trees constructed by the random forest ensemble. This method selects a subsample of the F features, t_g, to be used for each of the n_{tree} classification trees. Each of the non-selected $F - t_g$ features are essentially set to 0 as all of the samples are projected onto the subspace containing only the t_g selected features. In order to classify any new subject, the subject is projected onto the same subspace as its corresponding decision tree and a classified accordingly. Unlike in typical decision tree construction, each of the n_{tree} decision trees constructed in each of the subspaces in this manner is fully-split. This means that they are split until each final leaf is completely homogeneous. Thus, each decision tree predicts with perfect accuracy on training data, since it is split until all leaves contain no mixtures of classifications.

One of the main advantages of this method is that each of the decision trees in the random forest is constructed independently. Unlike boosting, for which the weights of the data in each subsequent sample depends on the weights and classification error of the previous sample, the random forest algorithm can be implemented in a parallel environment for faster computing. Also unlike many typical decision tree construction methods, there is no danger of being trapped in local optima since there is no 'hill-climbing' (Ho, 1998). This means that the resulting forest is less likely to over-fit the training data and thus perform poorly on new testing data. Finally, the main criticism of boosting decision trees is that the individual trees in subsequent samples become highly correlated with each other, since many of the

same data points get used over and over. With random forest construction, none of the individual classifiers are highly correlated since features to be used in each tree are selected in a random fashion. This lack of correlation is important because classifiers in an ensemble must remain relatively uncorrelated, especially when making errors, for the entire ensemble classification error to asymptotically approach zero (Hansen and Salamon, 1990).

2.3.5 Elastic-Net Regularized Generalized Linear Models

A generalized linear model (GLM) is a generalization of linear regression that predicts a particular response variable (e.g., the categorization of a student response) as a function of numerous predictor variables. Through the use of a link function, the errors of the response variable for a GLM need not be normally distributed as is the case in simple linear regression (McCullagh and Nelder, 1989). In this instance, the logit link function is used. The elastic-net is an approach to model selection via penalized maximum likelihood and it is especially effective for sparse input matrices (Tibshirani et al., 2012). The goal of the elastic-net implementation used in this research is to solve Equation 3 over a large number of values λ, where $l(y, n)$ is the negative log-likelihood for observation i. (Friedman et al., 2012).

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \Big[\frac{(1-\alpha)||\beta||_2^2}{2} + \alpha ||\beta||_1 \Big] \tag{3}$$

If $\alpha = 0$, then the elastic-net implementation in Equation 3 is equivalent to ridge regression and reduces to Equation 4. The second term in Equation 4 represents the L^2 norm (i.e. euclidian distance) of the coefficient vector that is used

for penalization. The euclidian distance is calculated by $\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$, where p and q are euclidian vectors.

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \Big[\frac{||\beta||_2^2}{2} \Big] \tag{4}$$

If $\alpha = 1$, then the elastic-net implementation in Equation 3 is equivalent to lasso regression and reduces to Equation 5. The second term in Equation 5 represents the L^1 norm (i.e. manhattan distance) of the coefficient vector that is used for penalization. The manhattan distance is calculated by $\sum_{i=1}^{n} |p_i - q_i|$, where p and q are vectors in the plane.

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \Big[||\beta||_1 \Big] \tag{5}$$

The advantage of the elastic net is that a variety of values of α are tested to determine which best improves model prediction (Zou and Hastie, 2005). Hastie and Qian (2014) note that the ridge penalty tends to shrink coefficients of correlated predictors toward each other, while lasso tends to overvalue a single predictor. A mixture of these approaches often leads to the benefits of both with neither of the drawbacks (Hastie and Qian, 2014). The variable selection for the implementation used here (from the *glment* package in R) itself functions similarly to lasso regression in that it selects important covariates for the regression model first through a forward variable selection algorithm by finding the covariate most correlated with the response variable of interest. Instead of simply adding that variable to the model and moving to the next-highly correlated variable as in forward selection, the technique

increases the coefficient of the corresponding initially chosen covariate continuously in the direction of the sign of its correlation with the response variable of interest until another variable becomes more correlated with the response than the initial covariate. This process is then repeated until all covariates are in the model (Tibshirani et al., 2012). Once the model is created in this fashion, it is used to predict the bin to which each item belongs to.

2.3.6 Maximum Entropy Modeling

Maximum entropy modeling utilizes a multinomial logistic regression model to make classifications about unknown responses (Jurka and Tsuruoka, 2015). Maximum entropy models function similarly to the Naive Bayes classifier. Unlike the Naive Bayes classifier, however, maximum entropy modeling does not assume that each of the features are independent of one another (Berger, 1996). Whereas multinomial logistic regression models attempt to classify data (i.e. responses) into one of many mutually exclusive categories, the maximum entropy model takes a unique approach to optimizing the variable selection process for the logistic model. First, a maximum entropy model implemented via the *maxent* package in R begins (as described by Vryniotis, 2013) by calculating the empirical probability distribution of the feature-classification pairings in the data:

$$\tilde{p}(x, j) = \frac{1}{N_{train}} \cdot I(x, j)$$

Where N_{train} is the size of the training set, x is a particular feature (typically an n-gram), and j is a particular classification. I(x,j) represents the number of times that a document with feature x is classified as classification j. Now, define:

$$f_i(x,j) = \begin{cases} 1, \text{if document } i \text{ is classified as } j \text{ and } containsfeaturex \\ 0 \text{ otherwise} \end{cases}$$

Thus, $f_i(x,j)$ is an indicator function which indicates if a particular document i contains feature x and was classified as j. As per Vryniotis (2013), one then calculates the expected value of any given feature function f_i by:

$$\tilde{p}(f_j) = \sum_{x,j} \tilde{p}(x,j) f_i(x,j)$$

Based on the principle of maximum entropy, one should select the logistic model $p*$ that is as close as possible to a uniform distribution of classifications. That is, without external knowledge about how certain features may or may not indicate the probability of a particular classification, one should prefer distributions (in this instance, specifically the conditional distribution of the classification variable given a particular document (McCallum et al., 1999))that are as close to uniform as possible. Thus, one must calculate the logistic model $p*$ which maximizes:

$$p* = \arg\max_{p \epsilon C} \left(- \sum_{x,j} \tilde{p}(x) p(x|j) log(p(j|x)) \right)$$

This model intended to be used when there are numerous predictor variables, sometimes hundreds of thousands (Jurka and Tsuruoka, 2015). For the case of natural language processing, it is often important to identify which words (either their presence or absence) are significantly correlated with a particular classification. It is equally important to identify potential interactions between words (i.e.

the presence of two words individually is uncorrelated with a particular categorization, but their presence in a response together is highly correlated with a certain categorization). The following example shows why assessing the potential statistical significance of interaction terms in a logistic regression model is a key step in the variable selection process.

Assume a professor poses the following question:

What is another common name for the Gaussian distribution?

The correct answer to this question is 'normal distribution.' A student using both the words *normal* and *distribution* would likely, but not necessarily, indicate the student has achieved the correct answer. Consider the following hypothetical responses:

1) This is definitely not a *normal* question to ask. I have never heard of the word Gaussian.
2) Another common name is the Poisson *distribution*.
3) Another name for this is the *normal distribution*.

In response 1, the student used the term *normal* but answered the question incorrectly. In response 2, the student used the term *distribution* but answered the question incorrectly. In response 3, the student used the 2-gram *normal distribution* and answered the question correctly. This is an example where the 1-grams *normal* and *distribution* may be moderately positively correlated with the correct answer, but the interaction term between *normal* and *distribution* will be highly significant

and have a positive coefficient. This implies that the 2-gram of *normal distribution* may be a better indicator of a correct response.

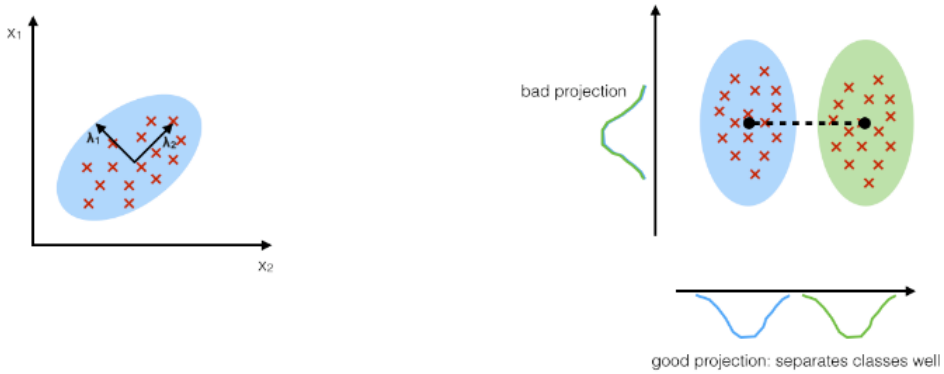
With few predictors variables, investigating all possible predictor variable combinations and their interactions is computationally trivial for most data sets (Malouf, 2002). Investigating these combinations for data sets with many predictor variables exponentially increases the run-time required (Zou and Hastie, 2005). In order to identify important predictor variables and potential corresponding interactions in a reasonable amount of computational time, maximum entropy modeling uses an iterative reweighting procedure, in this case maximum a posteriori (MAP) estimation. This is essentially equivalent to using MLE using regularization of the weights. The solution to this MAP estimation, in this case, is found through the iteratively reweighted least squares (IRLS) algorithm found in (Bishop, 2006).

2.3.7 Scaled Linear Discriminant Analysis

Linear Discriminant Analysis is a dimension reduction technique often used in natural language processing and other scenarios where the number of predictors is significantly larger than the number of data points, $p >> n$ (Raschka, 2014). Scaled Linear Discriminant Analysis (SLDA) is a generalization of Fisher's linear discriminant, which attempt to utilize a linear combination of features to separate classes. SLDA is most similar to principal component analysis (PCA), except that SLDA attempts to model the distinct differences between each of the different classifications of data. When the number of data points (i.e. responses) per classification bin is relatively large, SLDA tends to outperform PCA significantly (Martinez et al., 2001). The basic structure of SLDA is as follows, as described by (Raschka, 2014).

First, compute the mean vectors for each of the bins in the data set. Each of these vectors should be of dimension F, where F is the number of features of interest in the data. Second, calculate the associated variance-covariace matrices, known as scatter matrices. Third, computer the eigenvalues and eigenvectors for the scatter matrices. Fourth, construct a matrix of a fixed number of the eigenvectors with the largest eigenvalues. Finally, use this matrix to project the samples onto this new subspace.

The Figure 2.3b from (Raschka, 2014) demonstrates the difference between traditional PCA versus LDA. Figure 2.3a shows the PCA process of locating the two component axes which maximize the variance that is accounted for. Figure 2.3b demonstrates the LDA process of locating the axes on which to project for maximal class separation (i.e. the black dotted line).



(a) This figure shows the PCA process of locating the axes which account for the most variation in the data.

(b) This figure shows the LDA process of locating the axes of maximal class separation.

Figure 2.3: This figure details the difference between PCA (left) and LDA (right).

2.3.8 Support Vector Machines

Support Vector Machines (SVM) are a classification technique that construct a division between classes with a margin as large in magnitude as possible (Joachims 1998; Hearst et al., 1998). In two dimensions and if the classes are linearly separable, then this division is a vector. To find this maximally separating vector, SVM constructs two parallel lines—one containing two data points from one class, and the other containing two data points from the other class. SVM continues to construct all possible parallel lines in this manner, and then selects the pair of parallel lines with the largest distance between them (Sebastiani, 2002). These maximally distant, class separating parallel lines are known as support vectors, and the midparallel line formed perfectly in the middle of these lines is the class separating vector.

In F dimensions with separable classes, a series of $(F-1)$-hyperplanes are constructed until a separating hyperplane is found that separates the training data as maximally distant as possible. Support hyperplanes are formed instead of vectors (Hearst et al., 1998). A visualization of this technique in two dimensions can be seen in Figure 2.4. If the data are perfectly separable in n-space, then the hyperplane that separates the data and categorizes new data is called the maximum margin hyperplane.

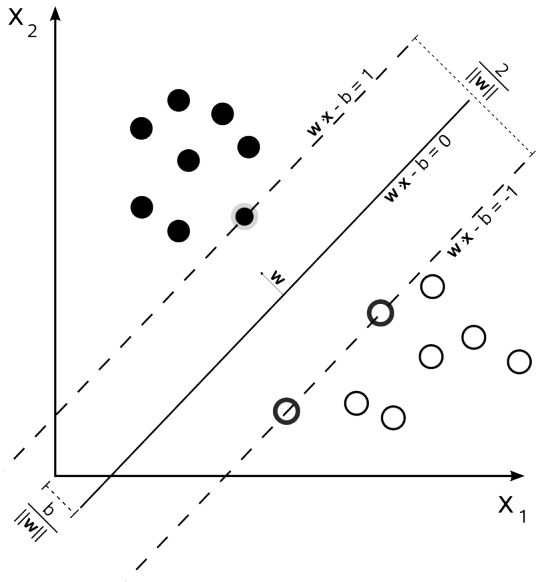


Figure 2.4: The image above shows 3 lines that each separate the black dots from the white dots. While both lines function perfectly as separators for these data, the line in the middle, or the line which maximizes the distance from each point to the line, is the maximizing linear classifier chosen by SVM.

If the data are not perfectly linearly separable, then the $(n-1)$-hyperplane is drawn in a similar fashion by maximizing the distance from the points to the plane (Yang and Liu, 1999). In general, the higher the dimension of the data and the more potential categorizations one has, the more difficult the task of creating a separating hyperplane may become. It has been shown that with thousands of samples, SVM still performs reasonably well even with tens of thousands of dimensions (Jin and Wang, 2012).

Joachims (1998) argues that many, if not most text categorization problems are linearly separable. With sparse document-term matrices and linearly separable classes, SVM should perform very well for text categorization, especially in high dimensional feature spaces. One of the more appealing aspects of SVM is its inherent lack of tuning parameters. This makes it easy to use and allows multiple in-

stances of SVM to be run consecutively without adjusting various tuning parameters (Joachims, 1998).

Additionally, SVM avoids feature selection completely; it inherently takes into account all features of the data. Many other machine learning algorithms attempt to select only the important features to use for classification, while in many text analysis circumstances there are few (if any) irrelevant features (Joachims, 1998). A classifier using only the 'worst' features outperforms a classifier that classifies at random, and thus feature selection (any reduction in the number of features of the data) results in data loss.

3 Methods

Overall, this research took a mixed-methods approach to answer the four proposed research questions. The general structure of this mixed-methods design is described in this section. The subsequent subsections describe all aspects of this methodology. This includes the development of constructed-response instruments, their integration into online student homework assignments, the creation of a rubric of categorizations for each question, the process of hand-coding training responses for these questions, and the development of an ensemble of machine learning algorithms to make categorizations for new responses based upon hand-coded training data.

This research followed an explanatory sequential design, described in Creswell (2015) as a design whose intent is first to use quantitative methods and then use qualitative methods to help explain in more depth the results found through the quantitative methods. The general manner in which this design was implemented is as follows.

From a quantitative perspective, a series of 9 online free-response instruments were constructed (as described in Section 3.3) to help answer the four proposed research questions. Through the process of hand-coding these responses, insight was gained to answer the **Misconception** and **Dot Plot vs. Histogram** research questions. Then, an ensemble of machine learning algorithms was constructed to categorize the online student responses into categories of interest. The efficacy of these models was then evaluated using data hand-coded by experts, ultimately determining if an ensemble of machine learning algorithms could be trained to accurately

categorize student responses to a variety of question archetypes to answer the **Ensemble** research question.

A series of 13 interview tasks were constructed for the qualitative portion of this study. These task-based interviews were used to investigate the **Misconception** research question involving student misconceptions about variability in graphs. In addition, the responses to these tasks were compared to the responses from their isomorphic online counterparts. This comparison helped answer the **Medium** research question. To better answer the **Medium** research question, each of the interviews were transcribed and analyzed, and close attention was paid to the completeness and correctness of responses to each of the tasks performed by the participants. A positivistic approach to these interviews was taken. That is, this research took a scientific approach to find the true prevalence of many previously described misconceptions about variability in graphs. This involved categorizing whether or not students appear to have any of the misconceptions detailed in Section 1.

Finally, the results of the interviews were combined with the quantitative data to make conclusions about students misconceptions about variability in graphs. This combination culminated in answering the **Medium** and **Misconception** research questions.

3.1 Quantitative Research Setting and Participants

All of the data in this study came from students in an undergraduate introductory statistics course at a large research institution in the Southeastern United States. In this course, students met for three hours a week in lecture with an additional hour-long session in a computer lab. This course served both statistics major and non-majors, and it included most of the standard topics taught in an introductory

statistics course including Chi-square tests and linear regression, but not ANOVA. The course textbook was *Statistics: The Art and Science of Learning from Data 3rd Edition* by Agresti and Franklin (Agresti and Franklin, 2013). There were approximately 1200 students enrolled in any given spring or fall semester, and typically fewer than 5% of the students enrolled were statistics majors. There were approximately 200 students enrolled in this course during any given summer semester.

Students in this course used WebAssign (`www.webassign.net`) for all lab, homework, and test submissions. WebAssign is an online instructional platform through which students can submit typed responses to instructor-created questions. Questions could take the form of multiple choice or free response. Students were given homework credit for their responses to each question used in this research project regardless of the correctness of their responses.

The data for each of the 9 free-response items were collected from students taking the introductory statistics course across five semesters from 2014 to 2015. These data consisted of over 25,000 total responses to free-response items. Some items also had an additional multiple-choice selection related to the constructed-response prompt. The textual responses to the 9 free-response questions are referred to as 'the data' for the remainder of this study. The data for each student were downloaded from WebAssign, and then immediately de-identified using a one-way encryption macro in Excel. This macro converted a student's name to a 40 character alphanumeric string. While the string could not be traced back to any individual student, a student's name always yielded the same 40 character string. This allowed for responses to be linked to the same student across multiple items.

3.2 Qualitative Research Setting and Participants

This section describes the qualitative methods, which consist of 19 one-on-one, task-based interviews with introductory statistics students from a large research institution in the Southeastern United States. These interviews were conducted over a period of three weeks in spring 2016. The 19 interviews were used to answer the **Misconception**, **Dot Plot vs. Histogram**, and **Medium** research questions.

In February of 2016, approval was given to conduct a series of interviews from consenting participants. A recruitment script was posted on eLC, an online learning and class management system, in order to recruit 20 participants. This script can be seen in Figure 7.26 in the Appendix. There was an initial surge of responses, so after a few hours, 20 respondents were randomly selected for participation. Each selected student received an email with further details about the interview study, and a time was set up to meet individually with each of the 20 students. Before conducting each interview, students were provided with a consent form, seen in Figure 7.26 in the Appendix. The consent form required that each participant be a at least 18 years of age and a currently-enrolled introductory statistics student. After accepting the interview invitation, one participant revealed herself to be under the age of 18 and thus was excluded from the study. This ultimately left 19 interviews ranging from twenty minutes to an hour in length.

Each of the 13 interview tasks was printed on separate sheets of paper. Beginning after the third interview, students were read the prompt aloud prior to being handed the sheet of paper with the corresponding task. This was done in response to the first two interviews, in which students would often begin by looking at the graph in the task instead of paying attention to the written prompt. This change appeared to alleviate this issue. Each of the 19 interviews was recorded

using S Voice, an Android application that records sound through a cell phone. This application was chosen due to its free nature and because it recorded sounds at a reasonably high quality, making for a simple transcription process. Each of the 19 interviews were then transcribed and relevant portions were coded for later use.

3.3 Item Construction

There were 22 free-response items constructed to answer the four proposed research questions. A total of 9 of these items were given to students in an online format over five semesters, and 13 of these items were given to students in face-to-face interviews. Online and interview items are categorized as either *Describe* or *Compare* items. *Describe* items ask students to describe the distribution shown in a given graph (histogram or dot plot) and potentially ask students to answer a question about their description of the graph in context. *Compare* items ask students to compare the variability of the data displayed in two side-by-side graphs. Both rubrics are found in Section 3.4.

In each of the items used in this research, students were asked questions about variability in data. When students were asked to describe the variability in a set of data or to compare the variability in two sets of data, they were instructed to consider standard deviation as the measure under consideration.

3.3.1 Describe Items

Describe items asked students to describe the distribution shown in a given graph. This section lists each of the *Describe* items and gives a brief description of their construction and purpose. Item 1 (Figure 3.1) provides an example of a typical *Describe* question.

Prompt- The histogram below shows the distribution of yearly income in dollars for a random sample of 356 adults living in Atlanta, GA. Describe as completely as possible the distribution shown in the histogram, being sure to explain what the graph tells you about yearly income for adults in Atlanta.

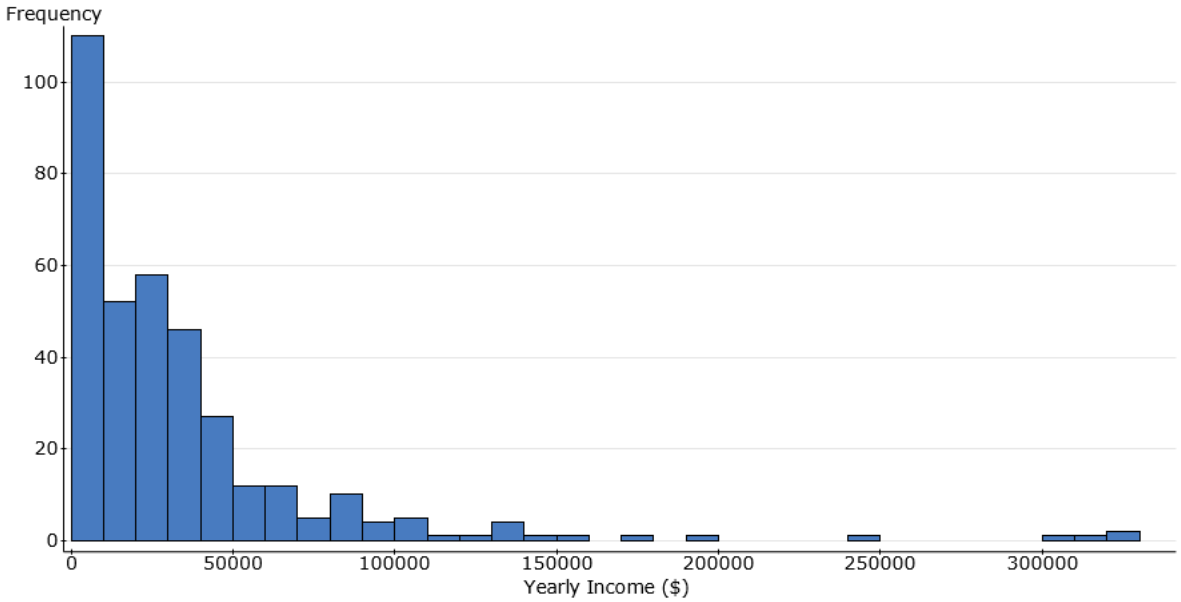


Figure 3.1: Item 1- Atlanta Income

The *Describe* rubric can be seen in Table 3. The *Describe* rubric was an analytic rubric that categorized responses as discussing the shape, center, variability, and/or context of the given graph. This rubric was created mainly to answer the **Dot Plot vs. Histogram** and **Medium** research questions by determining if students described histograms and dot plots in the same manner and by determining if students described graphs in a similar manner across different mediums. Each of the responses was categorized as including or not including a discussion of shape, center, variability, and context. The decision for each of these categorizations was made independently. Thus, a student's categorization for one particular category had no effect on the likelihood of the same student's categorization in another category.

Therefore, students could be categorized into any number of the four bins. This rubric was used for items asking students to describe a given graph, specifically in Item 1, Item 3, Item 5, Item 8, and Item 9. The rubric for Item 9 did not contain a context category due to the nature of the question prompt.

Table 3: This table shows the condensed *Describe* rubric for a general question.

Category	**Requirements**
Shape	Students must correctly discuss the shape of the graph by using appropriate terminology (e.g., asymmetric, unimodal, skewed).
Center	Students must give a valid measure of center (e.g. mean, median, mode, average) and correctly state its location.
Variability	Students must discuss either the range of the data, highlight potential outliers, locate the maximum and minimum values, or give an approximation of the standard deviation directly.
Context	Students must answer the question within the context of the problem by using appropriate units with their answer (e.g., 10,000 dollars) and identifying the subject of each unit (e.g., Atlanta adults).

Item 1- Atlanta Income

Item 1- Atlanta Income (Figure 7.1) was given to 1155 students in Spring 2014. It was developed at the onset of the research in this dissertation in order to determine what students would say when asked to describe a histogram. It was ultimately the catalyst for the remainder of the items described in this dissertation. All responses to this item were categorized using the *Describe* rubric.

Item 3- Student Sleep V1

Item 3- Student Sleep V1 (Figure 7.3) was given to 1188 students in Fall 2014. The histogram in this item was symmetric and unimodal. It was developed

both to determine if students possessed the **Bar Chart Misconception** and to help answer the **Medium** research question, and it functioned in an equivalent manner to that of Item 1. All responses to this item were categorized using the *Describe* rubric.

Item 5- Student Sleep V2

Item 5- Student Sleep V2 (Figure 7.5) was given to 1188 students in Fall 2014. The histogram in this item is identical to the graph in Item 3— symmetric and unimodal. The prompt is essentially identical to the prompt used in Item 3. This item was created to answer the **Dot Plot vs. Histogram** research question through its comparison with its isomorphic dot plot counterparts. It was also used to answer the **Medium** research question in its comparison with its interview task counterpart. Finally, this item was used to determine if student responses to identical items varied significantly in completeness and correctness over different semesters. All responses to this prompt were categorized using the *Describe* rubric.

Item 8- Coffee Consumption

Item 8- Coffee Consumption (Figure 7.8) was given to 1176 students in Fall 2015. This bimodal histogram was created for three reasons. First, it was created to determine how students would describe a bimodal distribution (namely, a distribution students had not been directly exposed to in class such as a uniform, skewed, or bell-shaped distribution). Second, it was created to test the **Medium** research question and is identical to Task 5, its interview task counterpart. Finally, it was created to test the **Dot Plot vs. Histogram** research question, and its isomorphic dot plot is given in Task 7 (using hours of TV watched instead of ounces of coffee consumed). All of the responses to Item 8 were categorized using the *Describe* rubric.

3.3.2 Compare Items

Compare items asked students to compare the variability of data shown in two side-by-side graphs. This section lists each of the *Compare* items and gives a brief description of their construction and purpose. Task 8 (Figure 3.2) exemplifies a typical *Compare* question.

Prompt: The two histograms below show test scores of two different classes on the same test. Which of the two classes had test scores that were more variable (i.e., have the higher standard deviation)?

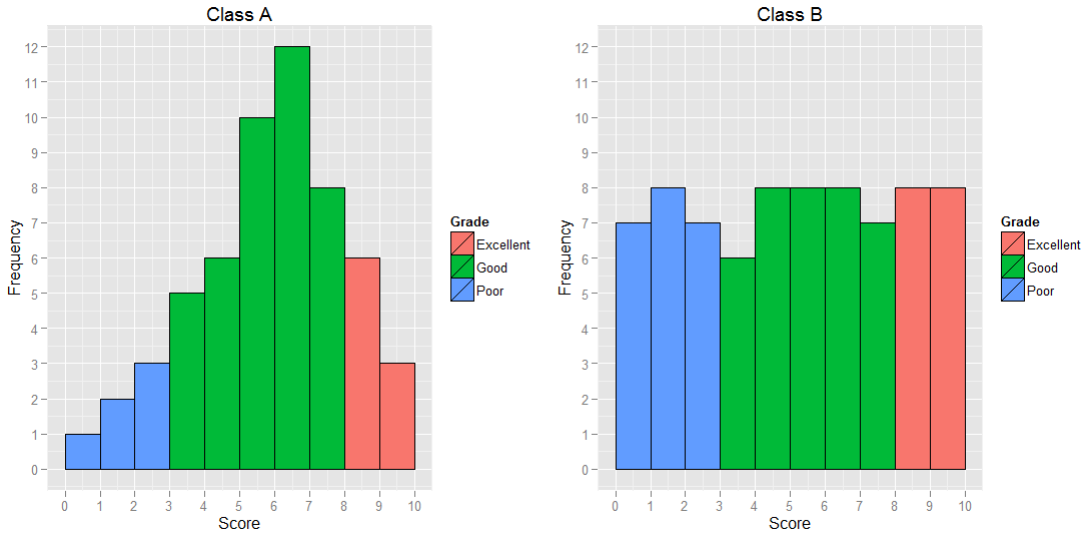


Figure 3.2: Item 9- Test Scores for Large-Test Histogram

The *Compare* rubric can be seen in Table 4. The *Compare* rubric was a holistic rubric used for items that asked students to compare the variability between two graphs. This rubric was used to categorize the manner in which a student justified their answer. This rubric contained five categories for student justifications: Height, Normal, Spread, Correct, and Other. The Height category was created in an attempt to capture students with the **Bar Height Misconception**. The Spread

category was created in an attempt to capture students with the **Range Misconception**. The other categories were created based upon the types of justifications given by students and were used to address the **Misconception** research question as a whole.

Unlike the *Describe* rubric, the bins presented here were mutually exclusive—students could only be placed into one of these bins. The reasons for utilizing a holistic rubric over an analytic rubric were twofold. First, students had to give *some* justification for their answer. Thus, students necessarily belonged to one of the five bins. Second, utilizing a holistic rubric scheme improved ensemble accuracy by necessitating the categorization of each response into only one bin. That is, some categories (e.g., Height) proved particularly challenging to categorize when treated as a binary category (i.e., a student either used the height justification or they did not), but when the ensemble was forced to place each response into one of the five categories, responses belonging to these difficult categories were more likely to be categorized correctly.

Each *Compare* question utilized a two-part question stem in which students were asked to determine which of two graphs had more variability (multiple choice) and then asked to explain their answer (free-response). Initially, the *Compare* rubric was used to categorize both responses from students who made the correct multiple choice selection and responses from students who made the incorrect multiple choice selection. During the process of hand-coding, it became apparent that students selecting the correct answer either gave a completely correct justification (coded as 'Correct') or in rare cases a nonsensical justification (coded as 'Other,' e.g., "I chose this answer because I knew it was right"). While these correct responses were still classified according to the *Compare* rubric, only responses from students making an incorrect multiple choice selection are analyzed in Chapter 4. Students

making the incorrect multiple choice selection sometimes included a correct justification to their incorrect multiple choice answer, and thus the 'Correct' category is maintained in the rubric. The items that used this rubric were: Item 2, Item 4, Item 6, Item 7, and Item 9.

Table 4: This table shows the condensed *Compare* rubric for a general question.

Category	**Requirements**
Height	Students specifically refer to the variation in the heights of the bars or dots. These students often give a correct definition of variability, but read the histogram or dot plot as if it were a case-value bar chart.
Normal	Students claimed they selected the histogram or dot plot because it was approximately normally distributed.
Spread	Students refer to their selection as having data that are more spread out or having a larger range. In some cases this statement may be true, however in no cases did a larger range imply more variability.
Correct	Students give a correct interpretation of the histogram or dot plot, but then select the incorrect answer.
Other	Students either give no justification for their answer, or give a nonsensical justification.

Figure 3.3 gives an example of five student responses that were categorized into each of the five categories in the *Compare* rubric.

Height: The frequencies are not very consistent as seen in Graph B. In Graph A there a lot of peaks and valleys in the graph. The frequency numbers in Graph A are further spread out and varied.
Normal: I know it is more variable because it has a normal distribution curve.
Spread: Class A are more variable because the graph is spread further out.
Correct: Class B has lots of dots everywhere, but class A has most of the scores in the middle.
Other: The dispersion of the scores was a dead give away. There was a clear difference in which question the students did better on.

Figure 3.3: Sample student responses for a *Compare* question.

Item 2- Test Score Variability Histogram V1

Item 2- Test Score Variability Histogram V1 (Figure 7.2) was given to 1155 students in Spring 2014. Students were shown two histograms, each depicting the distribution of test scores on a 10 question test for two different classes and asked to determine which of the two graphs showed data that were more variable. This item was constructed specifically to address the **Dot Plot vs. Histogram** research question. Each of the graphs in this item are histograms, and the graphs in Item 4 are isomorphic dot plots. Each student randomly received one of these six pairings seen in Figure 7.2 in the Appendix.

Item 4- Test Score Variability Dot Plot V1

Item 4- Test Score Variability Dot Plot V1 (Figure 7.4) was given to 1188 students in Fall 2014. Students were shown two dot plots, each depicting the distribution of test scores on a 10 question test for two different classes, and asked to determine which of the two graphs showed data that were more variable. This item was constructed specifically to address the **Dot Plot vs. Histogram** research question. Each of the graphs in this item are dot plots, and the graphs in Item 2 are isomorphic histograms. Each student randomly received one of these six pairings seen in Figure 7.4 in the Appendix.

The purpose of this item was twofold— are students able to compare the variability of data displayed in two side-by-side dot plots, and is there a difference in the proportion of students who can accurately compare the variability when the two displayed graphs are histograms versus when they are dot plots (**Histogram vs. Dot Plot** research question)? There were two parts to this question. The first part was multiple choice with two possible selections and asked students which of

the two histograms had more variable test scores. The second part asked students to explain how they knew that their selection had more variable scores. Responses to the second part of this item were coded using the *Compare* rubric.

Item 6- Test Score Variability Combination

Item 6- Test Score Variability Combination (Figure 7.6) was given to 1073 students in Spring 2015, 1176 students in Fall 2015, and 1086 students in Spring 2016 for a total of 3335 students. Responses in this item were categorized according to the *Compare* rubric. This item was created after reviewing the responses to Item 2 and Item 4. There were three specific misconceptions that appeared in the responses to these items. An isomorphic histogram and dot plot were created to investigate each of these three misconceptions for a total of 6 pairings. The first of these misconceptions was related to the **Bar Height Misconception** and is discussed in further detail in Chapter 4. The first pairing in this item was created to address this. The second misconception involved students' struggles with interpreting a histogram with an inverted shape. To determine if these struggles were due to the shape of the graph or the fact that it was originally paired with a uniform graph, this inverted shape was then paired with the Bumpy graph. Finally, the last pairing was designed to investigate the **Range Misconception**— that a data set with a larger range necessarily has data that are more variable. The full prompt is identical to that of Item 2 and Item 4 and can be seen in Figure 7.6 in the Appendix. All responses were categorized using the *Compare* rubric. Each student randomly received one of these six pairings seen in Figure 7.6 in the Appendix.

Item 7- Colored Test Score Variability Combination

Item 7- Colored Test Score Variability Combination (Figure 7.7) was given to 1176 students in Fall 2015 and 990 students in Spring 2016. Responses to this item were categorized according to the *Compare* rubric. This item is isomorphic to Item 6, and was created to determine if adding a coloring scheme and categorizing test scores as poor (0-3), good (4-7), or excellent (8-10) had any effect on students' interpretations of the graphs. A legend is provided to students containing this information. This was initially done in an attempt to improve readability and target students with the **Bar Chart Misconception** in hopes that the coloring and categorization scheme would help students understand that the bars/dots in the given graphs were aggregated data and not a single data value. The results of this coloring and categorization are found in Chapter 4. Each student randomly received one of these six pairings seen in Figure 7.7 in the Appendix.

Item 9- Test Scores for Large-Test Histogram

Item 9- Test Scores for Large-Test Histogram (Figure 7.9) was given to 1176 students in Fall 2015. Students randomly received one of three prompts asking each student to describe either the differences in center, in shape, or in variability between the two histograms. It was designed to evaluate the difference in the descriptions of shape, center, and variability in student responses when asked to generally compare two histograms (e.g., Item 2) versus being specifically asked to compare either the shape, center, or variability of two histograms. It was also constructed to answer the **Medium** research question and is identical to Task 9. Responses to this item were coded using the modified *Compare* rubric. This rubric did not contain a context category, as none of the prompts specifically asked students to discuss the context of the histograms.

3.3.3 General Quantitative Item Construction

A summary of each of the 9 constructed-response items is seen in Table 5. Each item was constructed to assess whether students had at least one of the four misconceptions about graphs: the **Bar Height Misconception**, the **Bar Chart Misconception**, the **Range Misconception**, and the **Axis Order Misconception**. The graphs used in each item were either dot plots, histograms, or both. In items where both graph types were used, students only viewed either a set of histograms or a set of dot plots. That is, any given student only compared the variability between two histograms or two dot plots, never between a histogram and a dot plot.

Table 5: A summary of each of the nine quantitative items used in this study.

Item	Name	Type	Graph	Versions	N
Item 1	Atlanta Income	Describe	Histogram	1	1155
Item 2	Test Score Histogram V1	Compare	Histogram	1	1155
Item 3	Student Sleep V1	Describe	Histogram	1	1188
Item 4	Test Score Dot Plot V1	Compare	Dot Plot	6	1188
Item 5	Student Sleep V2	Describe	Histogram	1	1188
Item 6	Test Score Combo	Compare	Both	6	1073
Item 7	Colored Test Score Combo	Compare	Both	6	1176
Item 8	Coffee Consumption	Describe	Histogram	1	1157
Item 9	Large-Test Histogram	Compare	Histogram	1	1087

There was only one version of graphs for Item 1, Item 2, Item 3, Item 5, Item 8, and Item 9. There were six versions of graphs for Item 4, Item 6, and Item 7. Students responding to these items were randomly given one of the six possible versions for each item. Each of the specific versions can be seen in Figure 7.10 and Figure 7.11. There were eight total graph types used across each of these versions. These eight graph types can be seen in Figure 3.4. They are henceforth named Tall Center, Inverted, Bumpy, Bell-Shaped, Uniform, Very Peaked, Spaced Uniform, and Short Uniform. Each of the pairings in Item 4, Item 6, and Item 7 consisted of some

isomorphic pairing of exactly two of these graphs. Some graphs were given as dot plots and others as histograms, as indicated in Table 5.

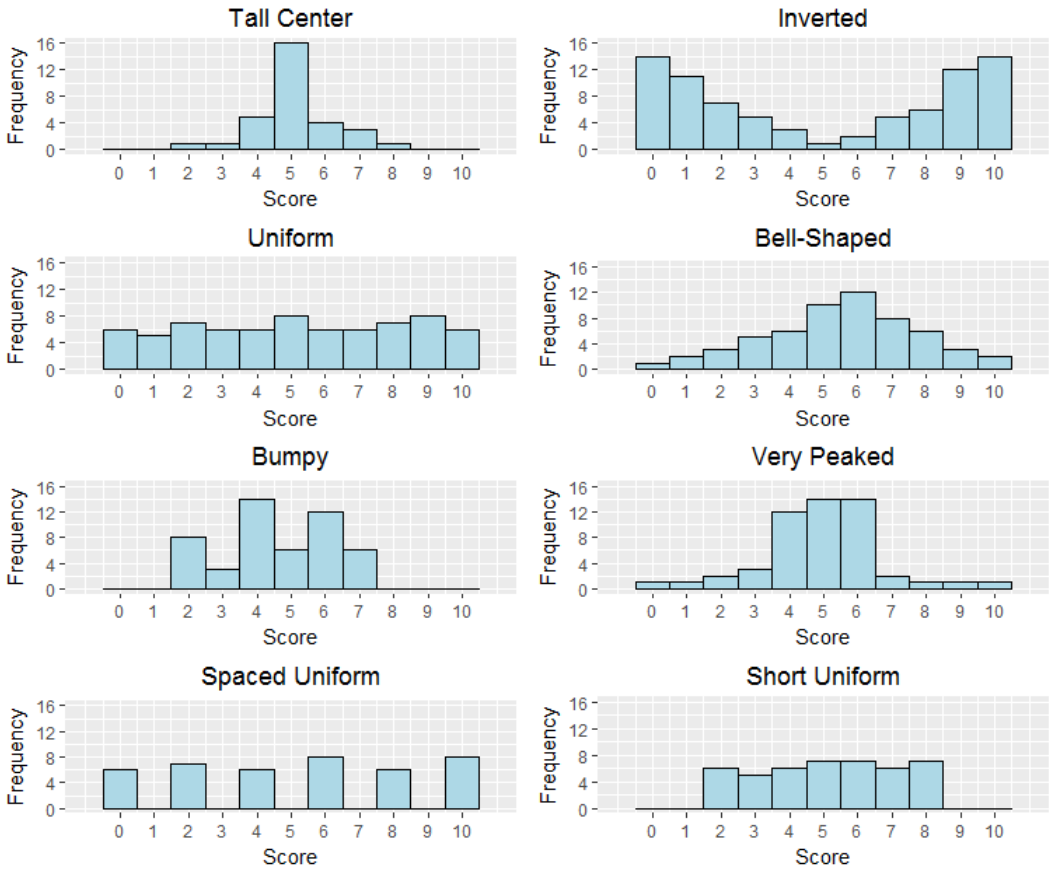


Figure 3.4: This figure shows the 8 graphs for the different pairings in Item 4, Item 6, and Item 7.

Pairings of graphs shown in Figure 3.4 were used to assess particular student misconceptions about graphs. While some pairings target specific misconceptions, others are used more generally to answer the **Medium Misconception** and **Bar Chart Misconception**. For example, the Bumpy and Uniform graphs are paired together to determine if students have the **Bar Height Misconception**—the heights of bars vary in the Bumpy graph but not the Uniform graph, however, the variability in the data is larger in the Uniform graph. The Very Peaked and

Short Uniform graphs were paired together to determine if students held the **Range Misconception**— the range in the Very Peaked graph is larger than that of the Short Uniform graph, however, the variability in the data is larger in the Short Uniform graph.

3.3.4 General Qualitative Task Construction

There were ultimately 13 interview tasks used in this study, although the process of developing the tasks for the interviewees to complete was iterative in nature. Two tasks were removed between the second and third interviews after it became clear that they were not functioning as had been intended. Conversely, task prompts were refined in order to better instruct students to answer as completely as possible.

Table 6 gives a summary of these 13 interview tasks. A more detailed description of each of the tasks can be seen in Section 7.2.

Table 6: A summary of each of the thirteen interview tasks used in this study.

Task	Name	Type	Graph	N
Task 1	Raw Data V1	Compare	Raw	19
Task 2	Raw Data V2	Compare	Raw	19
Task 3	Raw Data V3	Compare	Raw	19
Task 4	Water Histogram	Describe	Histogram	19
Task 5	Coffee Histogram	Describe	Histogram	19
Task 6	Exam Dot Plot	Compare	Dot Plot	19
Task 7	TV Dot Plot	Describe	Dot Plot	19
Task 8	Colored Two Test	Compare	Histogram	19
Task 9	Blue Two Test	Compare	Histogram	19
Task 10	Olympics	Compare	Histogram	18
Task 11	Advertising Histograms	Compare	Histogram	17
Task 12	Colored Two Test V2	Compare	Dot Plot	19
Task 13	City/Country Route	Compare	Histogram	18

3.3.5 Isomorphic Items

Fourteen of the twenty-two constructed items and tasks were isomorphic so that responses about variability in graphs could be compared across different graph types and mediums to answer the **Dot Plot vs. Histogram** and **Medium** research questions, respectively. Items that were isomorphic across mediums can be used to determine if student responses are of equal completeness and correctness in both online and face-to-face mediums. Items that were given as isomorphic dot plots and histograms can be used to determine if there are differences in the proportionality of certain misconceptions between dot plots and histograms. Table 7 shows the isomorphic groups, where items in a given column are each isomorphisms of one another.

Table 7: This table shows the groupings of online items and interview tasks by their isomorphisms. Items and tasks in the same column are isomorphic to one another.

Description	Compare Variability V1	Symmetric and Unimodal	Compare Variability V2	Bimodal	Unimodal vs. Skewed
Type	Compare	Describe	Compare	Describe	Compare
Isomorphisms	Item 2 Item 4	Item 3 Item 5 Task 6	Item 6 Item 7 Task 8 Task 12	Item 8 Task 5 Task 7	Item 9 Task 9

3.4 Coding Process

In this research, ensembles of machine learning algorithms were constructed to analyze the responses to the online items and to investigate the prevalence of the misconceptions listed in Section 1, as well as discover potentially new misconceptions that students have about variability. To accomplish this, a subset of student responses to the 9 online items were hand-coded by utilizing either the *Describe*

rubric or the *Compare* rubric. These hand-coded responses were used as training data for each of the machine learning algorithms in the ensemble. The following outline details the step-by-step process of hand coding responses for a given item.

Step 1: Briefly read over a subset of the student responses to ensure the item of interest was interpreted by students in the intended manner.

Step 2: Determine possible alterations to the *Describe* or *Compare* rubric, described below, for any given item.

Step 3: Two experts, typically two statistics Ph.D. students, hand code the first 100 responses according to the rubric.

Step 4: The two experts then convene and discuss disagreements in first 100 responses. This includes potential refinement of the rubric for clarity or the presence of new, important categories.

Step 5: Both experts hand-code the next 500 responses.

Step 6: The experts convene and discuss disagreements, ultimately agreeing upon a final training set for the ensemble.

Step 7: If there are categories of interest with few responses in the first 600 responses (or if the raters feel that more responses are needed to get a full sense of the data), the remaining responses are categorized.

Table 8 shows the inter-rater reliability results for items utilizing the *Describe* rubric archetype, where κ represents the value of Cohen's Kappa and A represents the proportion of responses that were independently, identically categorized by the raters. Across all items, there was strong inter-rater reliability for responses in the shape, center, and context categories. There was acceptable agreement between raters for the variability category.

Table 8: This table shows the inter-rater reliability statistics for each of the analytic categories in the *Describe* rubric archetype.

	Item 1		Item 3		Item 5		Item 8		Item 9	
Date Created	Spring 2014		Fall 2014		Fall 2014		Fall 2015		Fall 2015	
Number Hand-Coded	1155		1188		1188		600		600	
	κ	A	κ	A	κ	A	κ	A	κ	A
Shape	.941	.977	.960	.985	.952	.979	.917	.929	.931	.917
Center	.907	.938	.946	.976	.954	.978	.934	.951	.906	.924
Variability	.803	.925	.811	.922	.901	.935	.863	.908	.817	.898
Context	.969	.980	.934	.972	.927	.965	.959	.980	-	-

Table 9 shows the inter-rater reliability results for items utilizing the *Compare* rubric archetype, where κ represents the value of Cohen's Kappa and A represents the proportion of responses that independently, identically categorized by the raters. Across all items, there was strong inter-rater reliability for responses in the height, normal, and spread categories. There was acceptable agreement between raters for the correct category. The results for the 'other' category were passable, and the raters sometimes struggled to agree about which responses belonged in this category— it was often difficult to determine if a student were discussing the heights of bars/dots in the incorrect justifications if the word 'height' or other synonym was not used. After convening and discussing disagreements for the other category for Item 2 and Item 4, slight rubric refinements and extra practice allowed for better rater agreement and higher kappa values for this category in Item 6 and Item 7.

Table 9: This table shows the inter-rater reliability statistics for each of the holistic categories in the second rubric archetype.

	Item 2		Item 4		Item 6		Item 7	
Date Created	Spring 2014		Fall 2014		Fall 2015		Fall 2015	
Number Hand-Coded	1155		600		600		600	
	κ	A	κ	A	κ	A	κ	A
Height	.927	.953	.885	.932	.911	.925	.898	.947
Normal	.981	.990	.971	.986	.963	.985	.936	.958
Spread	.872	.905	.802	.891	.890	.910	.879	.901
Correct	.835	.892	.794	.856	.875	.899	.813	.899
Other	.746	.812	.717	.813	.774	.871	.805	.876

3.5 Data Cleaning

The first step in the ensemble categorization process for a given item is to clean the data set of student text responses. The following are brief descriptions of each of the cleaning measures used in this dissertation and the common circumstances under which they are employed. The default setting used in this work are given in bold. The extent to which any of the following techniques are used, however, varies from item to item. The specific data cleaning settings for each particular item are displayed in Section 4.6.

1) **Set minimum word length:** Words below a designated length are removed from the data set. If the difference between a correct answer and an incorrect answer is 'on the sphere' versus 'in the sphere,' then setting a minimum word length too high (i.e. above 2) will significantly decrease the overall ensemble accuracy, since identifying the use of 'in' versus 'on' is critical to a correct categorization. Conversely, setting a minimum word length too low allows words like 'a,' 'an,' 'it,' and others to flood the machine learning algorithms

with words not useful for making categorizations. In many instances, setting a minimum word length of 3 is a good balance between these two. **The default setting for data cleaning in this dissertation is to set the minimum word length at 3.**

2) **Set n-gram length:** An n-gram is the number of consecutive words that are taken together as a single entity when used for algorithm construction and evaluation. Converting n-grams into a numerical quantity representing the number of times an n-gram was used in a given response is known as feature extraction. A 1-gram means that each word is used by itself to make categorizations. A 2-gram implies that every two adjacent words are used for prediction. For example, in the sentence 'The normal distribution is bell-shaped,' there are four 2-grams: *The normal*, *normal distribution*, *distribution is*, and *is bell-shaped*. The use of 2-grams (or n-grams where $n > 2$) is critical in situations where the words 'normal' and 'distribution' may not indicate a correct answer, but where 'normal distribution' together as a 2-gram might indicate a correct answer. **The default setting for data cleaning in this dissertation is to set the n-gram length at 1.**

3) **Remove numbers:** This removes all stand-alone numbers from the data set. Removing numbers from a data set may be useful if students can provide a variety of correct answers using a different sets of numbers. For example, if students choose to describe the modal clump of a unimodal, symmetric histogram, they may arbitrarily decide where the modal clump begins and ends. Conversely, removing numbers from responses in which the categorization is based upon a correct, finite set of numerical values will decrease categorization accuracy. **The default setting for data cleaning in this dissertation is to remove numbers.**

4) **Remove punctuation:** This removes all symbols and punctuation from the data set, including but not limited to (- . ? ! ; : < > =). This removal, however, collapses responses such as '7-9 hours' into '79 hours.' This may have a negative effect on model accuracy. **The default setting for data cleaning in this dissertation is to remove punctuation.**

5) **Remove stopwords:** Stopwords are commonly used English words that tend to provide no contextual meaning in a particular sentence. Some examples of common stopwords are: and, the, in, on, of, a, and to. The removal of these words generally has a positive effect on algorithm accuracy, as these stopwords tend to appear in all possible categorizations of a response, often leading to confusions about categorizations in the algorithms. In cases where one or more of these stopwords are critical to a specific categorization, (e.g., the 'sphere' example in item 1) of this list), the removal of stopwords will have a negative effect on algorithm accuracy. **The default setting for data cleaning in this dissertation is to remove stopwords.**

6) **Stem words:** Stemming words is the process of removing prefixes and suffixes from words in an attempt to group words of similar semantics, despite these words having different prefixes and suffixes (known as stems). For example, the stemming process would make the 2-grams 'normal distribution' and 'normally distributed' into the same 2-gram 'normal distribute.' In this example, distribut-ion becomes distribute, and normal-ly becomes normal. Thus, both become 'normal distribute.' There are many cases where both 'normal distribution' and 'normally distributed' have the same semantics, and thus treating them as identical 2-grams will lead to significantly increased model accuracy. Conversely, there are situations where these individual 2-grams have significantly different meanings, where the prior refers to a specific continuous

distribution and the latter refers to the distribution of a random variable. **The default setting for data cleaning in this dissertation is to stem words.**

Once the data are cleaned as described above, the responses are ready to be used in the ensemble classification process. Let $\mathbf{M}$ be an NxM matrix, where N represents the number of student responses (also known as documents), and M represents the total number of unique n-grams (also known as terms) used in any of the responses. This matrix $\mathbf{M}$, henceforth referred to as the document-term matrix, serves as the input training matrix for each of the eight machine learning algorithms. The final step in the data cleaning process is to adjust the sparsity of $\mathbf{M}$.

7) **Adjust matrix sparsity:** The document-term matrix $\mathbf{M}$ is often extremely sparse. In most cases, fewer than 0.1% of the $n \cdot m$ cells are nonzero. Thus, to improve model accuracy and decrease runtime, it is often advantageous to reduce this sparse matrix into one that is more densely populated. The simplest manner in which to reduce the sparsity of $\mathbf{M}$ is by setting a minimum number of word uses for inclusion into the column space of $\mathbf{M}$. For example, one might say that a word must be used in at least 5 separate documents in order to warrant its own column in $\mathbf{M}$. This can significantly reduce the dimensionality of $\mathbf{M}$ by removing words that were only used in a few responses and likely do not have a significant influence in predicting the categorization of a particular response. **The default setting for data cleaning in this dissertation is to set the matrix sparsity threshold to be** 0.99**, such that the bottom 99% of words (when ordered by number of occurrences) are removed from M.**

3.6 Algorithm Evaluation

The goal for each of the items described in Section 3.3.1 was to develop an ensemble of machine learning algorithms that most accurately predicts the correct categorization of new, uncoded responses to the same question. The measures of accuracy used in this study are precision and recall. In the general case, both measures are treated equally.

The precision of an algorithm for a specific category j is the proportion of student responses that the algorithm placed into category j that actually belong to category j.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Thus, an algorithm's precision for a specific category decreases as it erroneously adds responses that do not truly belong to the selected category. In essence, low precision can be analogous to a high rate of false positives. Its calculation purpose is to identify an algorithm's propensity to correctly identify responses that belong to a given category.

An algorithm's recall for a specific category j is defined as the proportion of student responses that truly belong to category j that are correctly classified as category j.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Thus, in the trivial case, an algorithm that assigns all responses to category j has perfect recall for category j (since all responses from category j are necessarily assigned to category j). This algorithm, however, will have very low precision. These two measures together help identify struggling algorithms and address their possible inclusion or exclusion from the ensemble for a given item.

For each of the items used in this dissertation, at least two raters coded student responses according to the corresponding coding rubric. In order to test the categorization efficacy of each algorithm, leave-one-out cross-validation was utilized. Once a prediction was made for each of the responses by each of the eight algorithms, the accuracy of each algorithm at categorizing responses into each of the potential rubric categories was analyzed for a given item. For each rubric category, this involved examining the precision and recall of each algorithm. This process was repeated for variations in the tuning parameters and for different ensemble weighting functions. Ultimately, the combination of tuning parameters and machine learning algorithms that led to the highest precision and recall for the coded data was selected. That is, the combination that led to the highest average of the values recall and precision was considered the best combination and was used for classification.

3.7 Algorithm Tuning

At most eight different machine learning algorithms were used to make categorizations for each of the 9 items listed in Section 3.3.1. While the individual tuning parameters for each of the eight algorithms change from item to item, the following subsections describe the default tuning parameters used in this dissertation for each of the eight algorithms. Some algorithms have many tuning parameters, including

specific construction algorithms— others have relatively few. The specific computation for each of the 8 algorithms in this research can be found in the following **R** packages in parenthesis: support vector machines (*e1071*), elastic-net regularized generalized linear models (*glmnet*), maximum entropy (*maxent*), scaled linear discriminant analysis (*lda*), bagging classification trees (*ipred*), boosting decision stumps (*caTools*), random forests (*randomForest*), and classification trees (*tree*). This section describes default tuning parameters used in this dissertation to classify responses. The default values used here differ slightly from the default tuning parameters in each algorithm's corresponding **R** package.

3.7.1 Classification Trees

The default search algorithm used for the construction of classification trees is the Iterative Dichotomiser (ID3) Information Gain algorithm (Quinlan, J. R. 1986). The ID3 algorithm is considered a greedy algorithm, and constructs new leaves in a classification tree iteratively as follows:

1) Determine if the presence or absence of any word guarantees a specific classification in the training data.

1a) If this occurs, this word becomes the new node.

2) Determine which word partitions the data in a manner that minimizes entropy and thereby maximizes information gain.

2a) The word which partitions the training data in this manner becomes the new node.

3) Repeat recursively until stopping criteria are achieved.

Let Ent represent the entropy of a collection of responses S with K potential classifications (e.g., correct or incorrect). Then:

$$Ent(S) = -\sum_{i=1}^{K} P_S(i) \cdot log_2 P_S(i) \tag{6}$$

where $P_S(i)$ represents the proportion of responses in S belonging to classification i. This equation, therefore, is minimized when all members of a collection S have the same categorization, and it is maximized when there are equal numbers of members from all K classifications in S. The entropy of a feature, in this case an n-gram, is a measure of how well the presence or absence of a particular n-gram separates the responses into distinct categorizations.

The information gain for classification i a collection of responses S is defined by:

$$Gain(S, j) = Ent(S) - \sum_{j=1}^{K} P_S(j) Ent(S_j) \tag{7}$$

where $P_S(j)$ represents the proportion of responses in S belonging to classification j. Thus the information gain is calculated for each of the K potential classifications, and measures how well a given n-gram separates the training data into one of many distinct classifications.

The ID3 algorithm attempts to maximize the information gain at the creation of each leaf until one of three specific stopping criteria are achieved. The

stopping criteria used here to determine when new leaves should cease forming are threefold:

- **Stop forming branches when every response at a node falls into the same branch.** As the tree progresses, if a certain decision at a leaf causes all responses to either belong to the *yes* branch or the *no* branch, then stop generating new branches at that leaf.
- **Stop forming branches when the addition of any new branch adds approximately zero value to the overall classification tree.** This most often occurs when a tree becomes many leaves deep in a certain area. Because each leaf contains at least two branches, every time a leaf is traversed the data is necessarily split into smaller subsets. Every branch represents a characteristic that each subset of data does or does not contain. Therefore, after traversing several branches, the remaining responses in this subset of data become very specific and often very few in number. This threshold value is set to be 5, meaning that a minimum number of 5 observations are needed for a new leaf to be constructed.
- **Stop forming branches if the tree becomes more than 32 leaves deep in any path.** While this is a fairly rare occurrence, occasionally many subsequent leaves will generate branches that only eliminate one or two responses. For example, if the item prompt were, '*Describe the Central Limit Theorem*', a leaf that asked, '*Did the student say the word Central*' followed by a leaf that asked, '*Did the student say the word Limit*', etc. would lead to a very elongated tree that offers almost no real classification value.

Once constructed, new responses follow the series of decisions at each leaf as given by the classification tree. Once a terminal leaf is reached, the new response is

given a categorization equal to the most frequent categorization of training responses at that same terminal leaf.

3.7.2 Bagging Classification Trees

Bagging is a technique that uses bootstrapping of data to produce more diverse trees and reduce the likelihood of over-fitting the training data set. For the bagging of classification trees used in the ensemble, the classification trees are also formed using the ID3 algorithm. These trees are computed, however, without soft stopping criteria. Thus, the only stopping criteria for a classification tree constructed in this section is that new branches cease forming after the tree becomes more than 32 leaves deep in any path. While the removal of this soft stopping criteria would likely lead to severe over-fitting of the tree to the training data, the bagging process of constructing several trees works to alleviate this issue. The specific tuning parameters for the bagging of classification trees in the ensemble are:

- **The number of data points in any one bootstrapped sample of training data:** This parameter is set to be equal in size to the number of data points in the training data set. Since each resampling is done with replacement, this does not imply that the same sample of data is used to construct each tree.
- **The number of classification trees to construct:** This parameter is set equal to 50. Increasing this parameter significantly increases the run-time of the classification ensemble. Decreasing this parameter significantly impacts the accuracy of the bagging classification tree algorithm, as the construction of fewer trees is more likely to overfit random features of a particular resampling of the training data set.

3.7.3 Boosting Decision Stumps

Boosting is a technique that uses bootstrapping with a non-uniform weighting function for resampling to produce more diverse trees and reduce the likelihood of overfitting the training data set. A decision stump is a one-leaf classification tree. There are three tuning parameters to be set for the construction of numerous decision stumps through boosting:

- **The number of data points in any one bootstrapped sample of training data:** This parameter is set to be equal in size to the number of data points in the training data set. Since each resampling is done with replacement with varying weights for the probability that each point is chosen in the resampled data, this does not imply that the same sample of data is used to construct each tree.

- **The number of decision stumps to construct:** This parameter is set equal to 100. Increasing this parameter significantly increases the run-time of the classification ensemble. Decreasing this parameter significantly impacts the accuracy of the boosting algorithm— the more stumps are constructed, the higher probability that 'problem' points are identified and used to create more robust distinctions between classifications.

- **The reweighting function used for resampling training data:** With each new decision stump creation, a new bootstrapped sample is drawn from the training data. As points are misclassified with previous decision stumps, their corresponding probabilities for being selected in the new resampling of data (i.e. the weight of each misclassified point) is increased. The function that performs this is the AdaBoost reweighting algorithm.

3.7.4 Random Forests

A random forest is a collection of several classification trees, where each tree is formed using a different set of attributes (i.e. n-grams). There are two tuning parameters to be set for the construction of random forests in the ensemble:

- **The number of attributes (n-grams) to be used for each forest construction:** The default value set here is $\frac{N}{10}$, where N represents the total number of n-grams in the document-term matrix. Increasing this number exponentially increases the run-time of the classification ensemble. Decreasing this number leads to reduced accuracy.
- **The number of forests to construct in the ensemble:** This parameter is set equal to 50. Increasing this parameter significantly increases the run-time of the classification ensemble. Decreasing this parameter may significantly impact the accuracy of the random forest ensemble.

The formation of a random forest proceeds identically to that of a standard classification tree, except that only a randomized subset of features are allowed to be used by any one tree in the random forest. Each of the trees constructed in this manner are used together in an ensemble to make one overall classification.

3.7.5 Elastic-Net Regularized Generalized Linear Models

There are numerous tuning parameters for the generalized linear model portion of this algorithm. These parameters include those in standard variable selection techniques, such as the alpha parameter by which to include variables in forward selection. A detailed description of the defaults for these parameters can be found in the *glmnet* package in **R**. There is only one specific parameter with which the elastic

net penalization function uses, and it can be seen in Equation 8. Since the elastic net functions as a hybrid of both the L1 norm penalization from lasso regression and the L2 norm penalization from ridge regression, the elastic net penalty can weight this hybridity with the parameter α. If $\alpha = 1$, then this penalty function is equivalent to the lasso penalty. If $\alpha = 0$, then this function is equivalent to the ridge penalty:

$$\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1 \tag{8}$$

where β represents the vector of covariate coefficients. By default, $\alpha = 0.5$ is used. Often a range of α values between 0 and 1 is investigated determine which value of α best improves algorithm performance.

3.7.6 Maximum Entropy Modeling

Since maximum entropy modeling is being used as a tool for variable selection in a multinomial logistic regression model that also utilizes interaction terms between the covariates, there are no tuning parameters of interest other than those in the standard variable selection process. The details of this variable selection can be found in Section 2.3.6 as well as the *maxent* package in **R**.

3.7.7 Scaled Linear Discriminant Analysis

The tuning parameters for the scaled linear discriminant analysis algorithm remain unchanged throughout the model construction process. The specific details about the functionality of the SLDA used here can be found in the *lda* package in **R**.

3.7.8 Support Vector Machines

The tuning parameters for the support vector machine algorithm remain unchanged throughout the model construction process. The specific details about the functionality of the SVM used here can be found in the *e1071* package in **R**.

3.8 Ensemble Construction

After the algorithms were optimally tuned for categorizing responses for a particular rubric, they were combined into an ensemble containing all eight machine learning algorithms. Although each individual algorithm in the ensemble produces one classification vote for each response, the manner in which these votes are combined depends on the weighting function used. A detailed description of ensemble weighting functions can be found in Section 4.4, and the specific ensemble weighting function used for each category of each item is shown in Section 4.6.

4 Results

The results of the research in this dissertation are presented in this section. When relevant, results are split into quantitative and qualitative subsections. In many cases, the results are discussed in tandem as the qualitative results provide deeper insight into many of the quantitative analyses.

4.1 Dot Plot vs. Histogram

The **Dot Plot vs. Histogram** research question seeks to determine if students interpret graphs differently, specifically in terms of the graphs' inherent variability, when identical data are presented in a histogram or in a dot plot. This research question was addressed by the items in which students were asked to compare the variability between two data sets (*Compare* items), where the data were given in the form of a dot plot or a histogram, and then asked to explain their choice. This research question was also addressed by items in which students were asked to describe a given dot plot or histogram (*Describe* items). This section details the similarities and differences regarding student responses to these items. Section 4.1.1 provides results from student responses to online items, while Section 4.1.2 provides results from student responses to interview tasks. Some items were asked exclusively in online or in-person format, while others were asked in both settings. A comparison of student responses across online and in-person mediums can be found in Section 4.2.

4.1.1 Online Items

When answering online questions about the comparison of variability between two graphs, students responded differently to isomorphic histograms and dot plots. Each pairing of graphs contained a multiple choice prompt in addition to a constructed-response prompt. This section begins by analyzing the differences in multiple choice results (see the prompt for Figure 7.4 for an example of the prompt for each of these pairings). The graphics for each of the following 12 pairings can be seen in Figure 7.10, Figure 7.11, and Figure 7.12.

For each pairing, a two-proportion test was performed to determine if there were significant difference between the proportion of students giving a correct answer for the histogram version and the proportion of students giving a correct answer for the dot plot version. A Bonferroni correction was used to deal with the issue of conducting twelve separate hypothesis tests (thus using $\alpha = \frac{0.05}{12} = 0.00417$). Pairings with significant differences in the proportion of correct responses after the Bonferonni correction are marked with asterisks seen in Table 10.

Table 10: This table shows the percentage of students giving a correct answer for each of the pairings of dot plots and histograms. Significant differences (after Bonferroni correction) are marked with two asterisks.

Pairing	Graphs	% Correct Histogram	N_{Hist}	% Correct Dot Plot	N_{Dot}	Chi-Square Statistic	p-value
1	Very/Peaked Uniform	69.2	227	**75.1**	209	1.599	0.206
2	Bumpy/Very Peaked	65.9	229	**74.9**	243	4.172	0.041
3	Unif/Bumpy	69.3	212	**69.5**	236	0.000	1.000
4	Spaced Uniform/Very Peaked	75.3	182	**75.4**	192	0.000	1.000
5	Spaced Uniform/Bumpy	**90.0****	179	69.5	167	**21.515**	**0.000**
6	Inverted/Bumpy	68.5	159	**85.0****	140	**10.292**	**0.001**
7.1	Bell-Shaped/Uniform	36.9	179	**47.0**	215	3.676	0.055
7.2	Bell-Shaped/Uniform	31.7	224	**43.5****	186	**10.113**	**0.001**
8.1	Inverted/Bumpy	81.9	215	**85.6**	181	0.730	0.393
8.2	Inverted/Bumpy	93.0	172	**94.5**	201	0.147	0.701
9.1	Very Peaked/Uniform	24.1	145	**36.2****	138	**7.841**	**0.005**
9.2	Very Peaked/Uniform	39.2	148	**47.9**	144	1.908	0.167

In 11 of the 12 pairings, students were more likely to correctly identify the graph that contained more variable data when presented with a dot plot instead of a histogram (and significantly so in 3 of these 11 cases). Although the proportion of students making the correct selection for dot plots is higher in almost all cases than the proportion of students making the correct selection for histograms, the relatively small sample size leads to insufficient evidence that this difference is significant in many cases.

In pairing 5, students were significantly more likely to answer correctly when viewing histograms versus viewing dot plots. Although it is unclear exactly why this difference was so pronounced, it may have been due students viewing an in-class example similar to the graphs shown in pairing 5. In this in-class example, students were shown a pair of histograms similar in shape to the graphs of pairing 5, and the answer to this example was given to students a few weeks before completing the homework assignment containing pairing 5. Since the in-class example used histograms instead of dot plots, it is possible that students recalled this similar in-class example and chose the correct answer for pairing 5 accordingly.

Student justifications for correct responses (i.e. the constructed-response portions to correct multiple choice answers) were similar across all 12 pairings and both graph types. That is, students who provided a correct multiple choice answer for any of the 12 pairings tended to give a similar justification for the correct answer. Students provided one of two typical justifications:

> **Online Response to Pairing 1:** I chose [the flat graph] because the results were much more spread out than [the peaked graph], so it definitely has more variability.

Online Response to Pairing 9.1: Students made scores all over the place in [the flat graph], but most of the students made about the same grade in [the peaked graph].

In the first archetype response, students identified the spread of the data as the key to it having more variability. In the second archetype, students gave a more informal justification noting that scores were 'all over the place.'

A small proportion ($< 1\%$) of students presumably guessed the correct answer as their multiple choice selection was correct, however, their justifications were nonsensical and their responses were categorized into the 'Other' category. Otherwise, there were few differences in the types of correct responses for both graph types. Table 11 shows the combined categorizations of incorrect responses for the 12 total pairings. Across all items, incorrect responses were most likely to be categorized as 'Height.' That is, it was most common for students to incorrectly identify that the variation in the heights of the bars (or stacks of dots) was the determining factor about which graph contained more variable data. When viewing a dot plot versus a histogram, students were more likely to provide a correct justification (the 'Correct' category) but make an incorrect multiple choice selection. That is, when viewing a dot plot, many students gave a reasonable justification for why one particular graph contained data that were more variable. These same students, however, then made the incorrect multiple choice selection. This phenomenon did not occur as often in responses to histograms.

Table 11: Categorization of incorrect responses to all histogram and dot plot pairings. Relative proportions are indicated in parentheses.

Graph	Height	Normal	Spread	Correct	Other	Total
Histogram	503 (59.7%)	81 (9.6%)	149 (17.7%)	41 (4.9%)	69 (8.2%)	843 (100%)
Dot Plot	351 (48.5%)	58 (8.0%)	132 (18.3%)	102 (14.1%)	80 (11.1%)	723 (100%)

A Chi-squared test based on the data in Table 11 showed significant evidence of a difference in the distributions of categorizations of responses between histograms and dot plots for introductory statistics students (d.f. = 4, $\chi^2 = 49.818$, $p < 0.0001$). The two main contributors (in terms of their Chi-squared contributions) were the significantly smaller proportion of students in the 'Height' category for dot plots and the significantly smaller proportion of students in the 'Correct' category for histograms.

This trend for dot plots— students less frequently giving incorrect justifications referring to the heights of bars/dots and more frequently giving 'correct' justifications for incorrect responses— was seen throughout each of the 12 parings. Since the response patterns were similar across all pairings, pairing 7.1 and 7.2 are discussed in detail as examples of the patterns seen in the data. Table 12 shows the distribution of incorrect response justifications for both graph types for pairing 7.1, found in Figure 7.11. Pairing 7.1, the Bell-Shaped/Uniform pairing, was one of the potential pairings from Online Item 6 in Figure 7.6. Each of these 227 students made the incorrect multiple choice selection, and incorrect responses were categorized into one of the five aforementioned mutually exclusive categories. The ensemble categorization of these responses, (whose analytics are found in Table 21), can be seen in Table 12.

A Chi-squared test based on the data in Table 12 showed significant evidence of a difference in the distributions of categorizations of responses between the histograms and dot plot for pairing 7.1 for introductory statistics students (d.f. = 4, $\chi^2 = 23.244$, $p = 0.0001$). Although the Normal, Spread, and Other categories have nearly identical representation, there was a significantly larger proportion of students in the 'Correct' category for the dot plot version than the histogram version across both pairings. Additionally, there were significantly fewer students in the Height

category in the dot plot version when compared to the histogram version. Most pairings displayed a similar distribution to the one shown in Table 12— students referenced the variability in the heights of the bars or dots more frequently when describing the histogram versus the dot plot. Likewise, there were significantly more students in the Correct category for dot plots than for histograms.

Table 12: Categorization of incorrect responses to pairing 7.1.

Graph	Height	Normal	Spread	Correct	Other	Total
Histogram	71 (62.8%)	21 (18.6%)	9 (8.0%)	0 (0%)	12 (10.6%)	113
Dot Plot	52 (45.6%)	22 (19.3%)	7 (6.1%)	20 (17.5%)	13 (11.4%)	114

Student responses to pairing 7.2 (seen in Figure 7.11, and also in its original item in Online Item 7 in Figure 7.7), an isomorphic but colored version of pairing 7.1, were nearly identical proportionally to those of pairing 7.1. The ensemble categorization of these responses, (whose analytics are found in Table 21), can be seen in Table 13. A similar Chi-squared test based on the data in Table 13 also showed similar evidence of differences for the distribution of categorizations of responses to pairing 7.2 (d.f. $= 4$, $\chi^2 = 14.025$, $p = 0.0072$). Similar to pairing 7.1, there were more students giving correct interpretations with the dot plot version than with the histogram version. Thus, more students correctly interpreted the dot plot but chose the graph with less variability despite giving this correct interpretation. This only occurred twice in the histogram version.

Table 13: Categorization of incorrect responses to pairing 7.2.

Graph	Height	Normal	Spread	Correct	Other	Total
Histogram	107 (70.0%)	19 (12.4%)	10 (6.5%)	2 (1.3%)	15 (9.8%)	153
Dot Plot	65 (61.9%)	9 (8.6%)	10 (9.5%)	12 (11.4%)	9 (8.6%)	105

4.1.2 Interview Tasks

Throughout the interview process, respondents often gave differing responses to isomorphic items depending on whether the data were presented as a histogram or a dot plot. Table 14 shows the distribution of students' graph selections for interview tasks 8 and 12, which are identical to the histogram (see Figure 7.20) and dot plot (see Figure 7.24) from pairing 7.2 in Table 10. The left-most graph was bell-shaped, and the right-most graph was relatively flat. Students were asked which of the two graphs contained data that were more variable, although two students elected to respond that both graphs contained data that were equally variable. When asked to explain, both of these students indicated that the ranges for each of the graphs were from 0 to 10 (in both the histogram and the dot plot), and thus their variability had to be equal. Table 14 shows that two students made the incorrect selection when viewing the histogram, but then made the correct selection in the isomorphic dot plot in a subsequent task. There were no students who changed from the correct answer when viewing the histogram to the incorrect answer when viewing the dot plot.

Table 14: Detailed results for pairing 7.2 in the interview portion of this study indicating how each student answered both the histogram and dot plot isomorphisms.

		Histogram			
		Correct	Incorrect	Same	Total
Dot Plot	Correct	9	2	0	11
	Incorrect	0	6	0	6
	Same	0	0	2	2
	Total	9	8	2	19

In addition to asking each student to determine which of the two graphs (for both the histogram and isomorphic dot plot) contained data that were more variable, each student was also asked why he or she chose said graph. Students' rationales differed greatly depending on the graph type. Students who determined

that the peaked histogram contained more variable data often highlighted the variability in its frequency values or the variation in the heights of the bars. Interviewee 8 indicated this directly.

> **Interviewee 8:** The...the frequencies of the bars looks more different. And variability means a difference of frequency.

Interviewee 11 also noted the variability in the frequency values in A and the lack thereof in B.

> **Interviewee 11:** A [the left histogram] has more of a distinction in the frequency versus class B [the right histogram] who have about the same frequency for each type of grade.

Interviewee 12 used hand motions and a 'low to high' description to indicate that the frequencies are changing, and thus the graph is highly variable.

> **Interviewee 12:** There's just more... all of these [the right histogram's] bars are in the same area for frequency. This one [the left histogram] is just more variable. It starts off low, and then it goes high, and then it goes low again.

Students who selected the incorrect choice for the dot plot version occasionally mentioned the frequency in their explanations. These students, however, tended to refer to specifics dots or individual subjects displayed in the graph. Interviewee 6 pointed to outlier points specifically in his description of the dot plot.

> **Interviewee 6:** Even though it [the left dot plot] would look like it's approximately normal, it seems like there's kind of... outliers— subjects that are not anywhere near the mean and on the ends of the normal

curve on both sides. That's definitely showing that it's more spread out overall, and so I'd say more variable.

Interviewee 18 referred to the dots directly and noted that there are dots 'all over the place,' implying larger variability.

Interviewee 18: There's like an equal distribution of scores in B [the right dot plot]. So it's not like everyone is getting between a 5 and a 7. The dots are evenly distributed between 0 and 10. People are making scores all over the place.

Two students, interviewees 5 and 16, selected the left (incorrect) histogram in Task 8 of the interviews, but also selected the right (correct) dot plot in Task 12 as containing data that were more variable. Both students used similar reasoning to justify their incorrect decision regarding the histograms and their correct decision regarding the dot plots. Interviewee 16 noted the even bars in the right histogram and incorrectly assessed it as having little variability. A few minutes later, Interviewee 16 gave a correct description of the dot plots including identifying that the right dot plot had more variable data.

Interviewee 16: [Regarding the histogram pairing] I'd say class A [the left histogram] is more variable, because class B [the right histogram] has got about, like, the same amount of even bars, where as in class A a lot of people did good, but fewer people did poor and excellent.

Interviewee 16: [Regarding the dot plot pairing] Class B [the right dot plot] because each their scores are a little more evenly spread out. In Class A [the left dot plot], the majority are doing good, and then like very few are in excellent or poor. In B, it's got a little more variability...because it's got an even frequency throughout.

Interviewee 5 incorrectly responded to the histogram pairing by claiming that the differing heights of bars in the left histogram implied high variability. Interviewee 5 then correctly answered the isomorphic dot plot pairing by identifying that 'the dots are everywhere' and thus the scores were more variable.

> **Interviewee 5:** [Regarding the histogram pairing] Class A [the left histogram] because the bars are different heights but in B [the right histogram] they're all the same.
>
> **Interviewee 5:** [Regarding the dot plot pairing] Class B [the right dot plot] has more variability because the dots are everywhere and people made lots of different scores unlike A [the left dot plot].

Interview Task 4 (Figure 7.16) asked students to describe a unimodal, symmetric histogram about undergraduate students' water-drinking habits. Interview Task 6 (Figure 7.18) asked students to describe the same unimodal and symmetric graph, but in dot plot form instead of histogram form. The dot plot in Task 6 depicted exam scores on a 10 question exam.

Most of the students who could correctly interpret the meaning of graphs of univariate, quantitative data described both the histogram in Task 4 and the dot plot in Task 6 in a nearly identical fashion. When asked to describe the corresponding histogram or dot plot, these students targeted three main areas:

- **Center:** Students often began by identifying the center of the distribution. This sometimes included identifying the mode or perceived mean directly, but other times students identified the modal clump (e.g., most of the data are between 5 and 7).
- **Variability:** Many students identified the range of the data set and/or commented about the data set's overall variation (i.e. standard deviation).

- **Shape:** Students often concluded by mentioning that the graph was bell-shaped or approximately normal.

The following excerpts are three examples of interviewees' responses to Task 4 and Task 6. In these cases where students could correctly interpret the given graphs, there were few differences between their descriptions of histograms and of dot plots. Interviewees 11, 14, and 19 each identify the center of the graph, briefly describe its shape, and finally give an rough approximation of the graph's variability in context.

> **Interviewee 11:** [Regarding the histogram in Task 4] In the 58, there were 12 students who drank 6 cups of water, and that was the highest frequency. The most was 10 cups an the least was 0 cups of water per day.
>
> **Interviewee 11:** [Regarding the dot plot in Task 6] The average was around 4 to 8, so, it was probably an average medium [difficulty] test. Very few students did really well and really poorly.
>
> **Interviewee 14:** [Regarding the histogram in Task 4] Most of them are around...the mean is around 5 to 7— that's where the biggest percentage of the students are. It's pretty much bell [shaped]. Other than that, it's got a good amount of variability with some students drinking 0 cups per day or 10 cups per day. A normal looking graph.
>
> **Interviewee 14:** [Regarding the dot plot in Task 6] The majority of students only scored between 5 and 6, and there seems to be a great range of students who scored a lot less and ones who did well. It seems like it's got that bell-shaped curve that most professors would look for.

Interviewee 19: [Regarding the histogram in Task 4] Most of them drank 6 [cups of water] with a frequency of 12, and then it's a normal distribution. There doesn't really seem to be any outliers because both sides are pretty even.

Interviewee 19: [Regarding the dot plot in Task 6] The test must have been pretty difficult because the highest frequency was a 6. So, that's like failing. Only about 5 [students] made A's, and someone actually made a 0. So it seems pretty hard.

Much like in Tasks 4 and 6, students described the right-skewed histogram in Task 5 (Figure 7.17) and the isomorphic dot plot in Task 7 (Figure 7.19) in a similar manner. The following is an excerpt from Interviewee 6, who accurately described the distributions in both tasks. This person, like most of the interviewees who could correctly interpret graphs, gave a similar description of each— first by identifying the first mode, then by identifying the second mode and the overall shape, and finally by identifying the highly variable nature of the data.

Interviewee 6: [Regarding the histogram in Task 5] It seems that most people who do consume coffee are drinking between 0 and 10 ounces...or between 0 and 30 is the majority of the people, meaning that the majority of people who consume coffee consume less. But then there are outliers that are a little spread out, making it a little more variable, because there are some people who drank 100 ounces per day.

Interviewee 6: [Regarding the dot plot in Task 7] So this pretty much seems to be what people call bimodal, in the fact that there's kind of two main curves, there seems to be like two high points where people either seem to watch 0 hours of TV or they watch upwards of 10— is this hours

per day? That's a lot of TV. They watch upwards of 7 hours of TV. So you have this two-pronged kind of grouping where people seem to watch a lot of TV or none at all. There's definitely a lot of variability going on in this data.

There were a few interviewees who struggled to correctly interpret the histogram. The following is an excerpt from an exchange with Interviewee 13 regarding Task 5. This student gave a terse description of the histogram in Task 5, and an incorrect definition of its associated variability.

Interviewer: What does this graph tell you about UGA undergrad's coffee drinking habits?

Interviewee 13: A lot of people drink 0 cups of coffee, and then it varies between 0 and 100.

Interviewer: What do you mean by 'it varies?'

Interviewee 13: Um... There is scattered data. So the bars start high then they go low at the end, so lots of variability.

Conversely, Interviewee 13 and the interviewer had a nearly identical interaction a few minutes later with the isomorphic Task 7. This student again gave a succinct description of the given dot plot and again mentioned that 'it varies a lot.' When asked to explain, this time the student correctly defined the variability exhibited in the dot plot.

Interviewer: What does this graph tell you about UGA undergrad's TV watching habits?

Interviewee 13: Well it varies a lot because it goes from 0 to 10 hours and a lot of people watch 0 and a few people watch 10.

Interviewer: Well, so what do you mean by 'it varies?'

Interviewee 13: Well people watch lots of different numbers of hours of TV— they're all over the place.

4.2 Medium

This section details the similarities and differences amongst student responses to online items and their isomorphic face-to-face interview task counterparts. It seeks to shed light on the **Medium** research question, which aims to determine if there are significant differences in the correctness and completeness of student responses to online items versus responses given in face-to-face interviews.

4.2.1 Correctness

This section examine the similarities and difference in the correctness of student responses to online items and face-to-face interview tasks. It does so by examining isomorphic questions asked in both mediums.

Table 15 shows students' results for Pairing 7.2 (Figure 7.11). In interview form, this encompassed Interview Tasks 8 (Figure 7.20) and 12 (Figure 7.24). Students saw the pair of histograms in Task 8 and the pair of dot plots in Task 12. In its online form, this encompassed two of the six possible pairings from Item 7 (Figure 7.7). That is, students were randomly assigned one of six pairings of data (three histograms and three dot plots). Only students who received the histograms and dot plots in Pairing 7.2 ($n = 224$ and $n = 186$, respectively) are included in Table 15. In this item, students were asked which of two graphs (either a pair of dot plots or a pair of histograms) contained data that were more variable. A

two-sample test for equal proportions showed no evidence of a significant difference across mediums in the percentage of students correctly answering either the histograms ($\chi^2 \approx 0, \text{d.f.} = 1, p \approx 1$) or the dot plots ($\chi^2 \approx 0, \text{d.f.} = 1, p \approx 1$).

Table 15: This table shows the overall percentage of correct responses for pairing 7.2, a subset of Online Item 7 and identical to Interview Tasks 8 and 12.

Medium	% Correct Histogram	% Correct Dot Plot
Online	31.7 ($n = 224$)	43.5 ($n = 186$)
Interview	31.6 ($n = 19$)	42.1 ($n = 19$)

A similar trend as the one previously shown was observed in each of the overlapping online items and interview tasks. Table 16 shows the proportion of correct responses for the remainder of overlapping items (or subset pairings of items) for which there existed a correct response (i.e. there does not exist a correct response for the 'describe the graph' items, but there does exist a correct response for the 'choose the correct graph' items). Three separate two-sample tests for equal proportions showed no evidence of a significant difference across mediums in the percentage of students correctly answering any of the three remaining questions ($\chi^2 = 0.005, \text{d.f.} = 1, p \approx 1$ for Test Score Dot V2; $\chi^2 \approx 0, \text{d.f.} = 1, p \approx 1$ for Test Score Hist V3; $\chi^2 \approx 0, \text{d.f.} = 1, p \approx 1$ for Country/City Variability).

Table 16: A comparison of the percentage of correct responses to each of the remaining overlapping isomorphic online items and interview tasks.

	Test Score Dot V2	Test Score Hist V3	Country/City Variability
Online	59.4% (n=215)	51.2% (n=178)	42.7% (n=1175)
Interview	63.1% (n=19)	52.6% (n=19)	42.1% (n=19)

4.2.2 Completeness

Students choosing the correct graph in both Task 8 and Task 12 gave nearly identical responses to those from the isomorphic Item 7. That is, student responses to the interview tasks justified their responses using nearly identical justification as those from Item 7. The primary difference between these responses was the average word length— student responses to online items averaged 15 ($s = 9.1$) words in length, whereas student responses to interview tasks averaged 75 ($s = 45.7$) words in length.

In Online Item 8 (Figure 7.8) and Interview Task 5 (Figure 7.17), students were asked to describe what the depicted histogram told them about how many ounces of coffee undergraduate students drink per day. In order to compare the completeness of responses to this question across mediums, the presence of the four main characteristics of a histogram is examined— its Shape, Center, Variability, and Context. Table 17 shows the comparison of the discussion (or lack thereof) of these four categories in online responses versus face-to-face responses. Definitions of a description of Shape, Center, Variability, or Context can be seen in Table 3. Ensemble analytics for Item 8 ensemble can be seen in Table 21.

Table 17: A comparison of histogram descriptions for online versus face-to-face responses

Question	Medium	% Shape	% Center	% Variability	% Context	N
Item 8	Online	71.9	27.9	15.6	44.6	1156
Task 5	Face-to-Face	73.7	26.3	15.8	68.4	19

A Chi-squared test showed no evidence of a significant difference in the distributions of responses for students discussing each of the four elements of a histogram across mediums ($\chi^2 = 1.4815, \text{d.f.} = 4, p = 0.6865$). Although the percentage of students discussing the shape, center, and variability remained nearly identical

across mediums, there appeared to be a slight difference in the percentage of students discussing the context of the histogram. Although this difference was not statistically significant, it is likely in part due to the average length of response across the different mediums. Students responding to the online version of the histogram often used few words to convey their responses.

> **Response to Online Item 8:** Skewed right. Mean is greater than median and range is very large.

> **Response to Online Item 8:** Bimodal and skewed right, mean is about 40 with very large variability.

> **Response to Online Item 8:** Skew right. They drink between 0 and 30 and range is 100.

The aforementioned responses were each coded positive for shape, center, and variability, but were all done using fewer than 12 words. The following responses came from the face-to-face version of the same question, and were also coded positive for shape, center, and variability. These responses were typically much longer, but still conveyed the same basic information about the shape, center, and variability of the histogram. In many cases, these responses also included a discussion of the graph's context, or their descriptions of the shape, center, and/or variability were given in context.

> **Interviewee 4:** A lot of people don't drink coffee. About half maybe? So about half of UGA doesn't drink coffee at all. But some people drink a lot of coffee. Just a few though, so it has two main groups of people. The data is pretty spread out too.

> **Interviewee 16:** It's not as high as I would expect honestly. This graph looks more skewed right than I would expect. But there is a little bit of

a second bump near the right end between 60 and 90. That's probably during exam week. But pretty much overall it's skewed right, with a mean maybe around 20 or 30. It's hard to tell because the data is so spread out.

Interviewee 18: Most students do not drink very much coffee. It's skewed right, which means that more people are drinking amounts that are on the left side of the graph. Between 0 and 10 ounces, about 100 people said they drink that much. After that, it tapers down pretty substantially, and as it goes down, the value get much smaller other than the small subset of students between 60 and 90. It kind of looks bimodal.

4.3 Misconceptions

This section utilizes both ensembles of machine learning algorithms and interview tasks to quantify the proportions of students with each of the four established misconceptions about quantitative displays of univariate data described in Section 1.1. Table 18 quantifies the percentage of responses to several different questions that contained a particular misconception, shown in the left-split column. In the right-split column, the table shows the percentage of incorrect responses that contained the misconception of interest. Ensemble analytics for each of these predictions can be found in Table 21.

Table 18: This table shows, for several different items and tasks, the overall proportion of students exhibiting a given misconception, as well as the proportion of incorrect answers exhibiting the same misconception.

		Bar Chart Misconception		Bar Height Misconception		Axis Order Misconception		Range Misconception	
Item	Name	% of Total	% of Incorrect	% of Total	% of Incorrect	% of Total	% of Incorrect	% of Total	% of Incorrect
Item 6	Score Variability V2	32.5 (n=3335)	50.6 (n=1319)	28.1 (n=3335)	42.4 (n=1319)	0.1 (n=3335)	0.3 (n=1319)	12.3 (n=1112)	31.2 (n=440)
Item 7	Score Variability V3	31.8 (n=1125)	50.4 (n=709)	41.4 (n=1125)	71.6 (n=709)	0.8 (n=1125)	1.1 (n=709)	8.1 (n=185)	21.9 (n=91)
Item 8	Coffee Histogram	7.5 (n=1327)	N/A	25.1 (n=1327)	N/A	0.6 (n=1327)	N/A	N/A	N/A
Task 3	Raw Data	N/A	N/A	N/A	N/A	N/A	N/A	63.2 (n=19)	92.3 (n=13)
Task 5	Coffee Histogram	10.5 (n=19)	N/A	15.8 (n=19)	N/A	5.3 (n=19)	N/A	N/A	N/A
Task 8	Score Variability Histogram	31.6 (n=19)	83.3 (n=6)	36.8 (n=19)	71.4 (n=19)	0 (n=19)	0 (n=19)	N/A	N/A

4.3.1 Bar Chart Misconception

The **Bar Chart Misconception** is the belief that any two-dimensional graph with bars must be a case-value bar chart, and its bars (or stacks of dots) must therefore represent individual, non-aggregated data. As such, this section quantifies the prevalence of this misconception through the examination of student responses to Item 6 (Figure 7.6), Item 7 (Figure 7.7), Item 8 (Figure 7.8), and their corresponding interview tasks— Task 8 (Figure 7.20) and Task 5 (Figure 7.17).

The propensity of a student to believe that, when viewing a histogram, the student were instead viewing a case-value bar chart varied significantly across different items. Specifically, students were more likely to misunderstand a histogram as a bar chart when being asked to describe the graph's variability (or compare the variability across multiple histograms) than when being asked general questions about the same histogram. Table 18 shows that 32.5% and 31.8% of students exhibited the bar chart misconception for Item 6 and Item 7, respectively. These items

asked students to compare the variability between two histograms or dot plots, and students often compared the variability as if they were viewing case-value bar charts instead of histograms or dot plots. The sample responses exemplify students with the bar chart misconception. These responses claim that the bars (or dots) represent a value for a single subject rather than an aggregated value for many subjects. This type of response comprised approximately 50% of all incorrect answers to Item 6 and Item 7.

> **Response to Online Item 6:** Class A is more variable because out of the 10 students in class B, most of them got the same grade but students in A did all over the place.
>
> **Response to Online Item 6:** Most students made about the same grade in B but students made lots of different grades in A.

4.3.2 Bar Height Misconception

The **Bar Height Misconception** is the misconception that the variability in a given histogram or dot plot is depicted through the variation (or lack thereof) in the heights of its bars or dots. There was a strong propensity for students with the bar chart misconception to also exhibit the bar height misconception. This relationship is discussed in greater detail in Section 5.2.3. When comparing the variability in two graphs, as per Item 6, Item 7, and Task 8, 32.5%, 31.8%, and 31.6% of students exhibited the bar height misconception, respectively. This proportion dropped significantly in the two questions that required students to describe a histogram, Item 8 and Task 5. Student responses to these questions exhibited the bar height misconception 7.5% and 10.5% of the time.

Students with the bar height misconception sometimes understood the premise of variability as a function of deviation from a measure of center, but they applied this definition in the wrong way, as per the first response below. Many other students identified the variation in the heights of bars or dots as the source of variability in the data, as per the second response.

> **Response to Online Item 7:** The first one has more variability because the bar heights are way different from the mean bar. The second one has almost no variability because all of the bars heights are right around the center bar.
>
> **Response to Online Item 7:** The bar heights vary a lot more because they go up and down in the first one, so the first class has more variability.

4.3.3 Axis Order Misconception

The **Axis Order Misconception** is the idea that the ordering of the x-axis is arbitrary. Despite being a well-established misconception in the statistics education literature, very few students across both mediums directly exhibited the axis order misconception. Since each of the histograms and dot plots in the online items and interview tasks had a well-labeled x-axis, (that is, each bin in the histogram or dot plot was labelled directly), it was unlikely that students would believe the x-axis to be arbitrarily ordered. Across both interview tasks, 5.3% and 0% of students exhibited this misconception (only 1 student total). With regards to online Item 6, Item 7, and Item 8, only 0.1%, 0.8%, and 0.6% of students directly exhibited this misconception. The following is a sample response from Item 8 that exhibits the axis order misconception.

Response to Online Item 8: The first couple students drink a lot of coffee, but then other students don't drink as much coffee, although there are some students at the end that drink a medium amount.

4.3.4 Range Misconception

Students with the **Range Misconception** believe that the range of a data set defines its variability— thus, a data set with a larger range is necessarily more variable than one with a smaller range. Although range is one informal measure of variability, each of the items and tasks presented in this research asked students to assess variability as a function of the standard deviation of the data. Students with this misconception believe that finding the range is equivalent to finding the variability (i.e. standard deviation).

There were several items and tasks designed to assess the prevalence of the range misconception. Table 18 shows the proportions of students with this misconception across Item 6, Item 7, and Task 3. Although the proportions of students exhibiting this particular misconception varied across the three related questions, student responses to online items and in-person tasks were similar. When asked to define variability at the beginning of each of the face-to-face interviews, students gave a wide range of answers. Many students, like Interviewee 6, included range in their description of variability.

Interviewer: In your own words, give me a statistical definition for the word 'variability.'

Interviewee 6: I would say it means the spread of data. How much the data differs from a measure of center, mean, median, and how much it's spread out and how big of a range it has.

Interviewee 6 mentioned the range when asked to describe variability in general. When asked to determine which of two data sets have more variability as a function of standard deviation, she began by looking at the range but quickly corrected herself and correctly identified the data set with larger variability. In the aforementioned task, (Task 3, Figure 7.15), students were shown two data sets—one with a smaller range but more variable data (list E) and one with a larger range and less variable data (list F).

Interviewee 6: [In response to Task 3] I'm going to say list F. I can see that list F has a larger range...wait... it just has two outliers. The data from list E are actually more variable than F because they are farther from the mean than [the data in list F].

Other students, however, gave similar definitions of variability which included, either in part or in whole, a mention of range. Interviewee 5 appeared to equate range and variability in her definition.

Interviewer: In your own words, please give a statistical definition for the word 'variability.'

Interviewee 4: Variability means a bigger range... so if things are really spread out then they are really variable.

Unlike Interviewee 6, Interviewee 4 did not appear to associate range as a measure of variability, but rather as a synonym. This is evidenced by her response to Task 3.

> **Interviewee 4:** [In response to Task 3] List F has more variable data because... 44 [the subtraction of 45 (the maximum) and 1 (the minimum)] is a lot bigger than 36 [the subtraction of 41 (the maximum) and 5 (the minimum). So definitely list F because the range is a lot bigger.

Interviewee 11 seemed conflicted about which list to choose—she ultimately chose the wrong one. She described not being able to remember how to calculate variability or how to find it in a data set.

> **Interviewee 11:** [In response to Task 3] Now I can't remember if the variability depends on all of them together or just the lowest and highest numbers... I guess it's list F [the one with a larger range and less variability] because it's got more spread between the highes and lowest number.

Many students—12 out of the 19 interviewed—gave nearly identical responses to this task. Each claimed that the list with a larger range must be more variable. Task 3 was designed to be intentionally difficult, and the proportion of students exhibiting the range misconception was expectedly higher than that of the online items. Item 6 and Item 7, each questions about histogram or dot plot variability, saw only 12.3% and 8.1% of students exhibit the range misconception compared to the 63.2% of Task 3. The range misconception comprised 31.2% and 21.9% of incorrect responses to Item 6 and Item 7, whereas the range misconception was seen in 92.3% of incorrect student responses in face-to-face interviews. The following three student excerpts show three students' justifications of their incorrect answer. Much like the interview responses, students' online responses exhibiting this misconception often described the variability only in terms of the range.

Response to Online Item 6: Class A goes from 1-10 and B is only from 2-9, so class A is more variable.

Response to Online Item 6: The range is 9 in A so it is more variable, since the range is only 7 in class B.

Response to Online Item 7: The range in A is slightly more than B, so it is more variable.

4.4 Ensemble Weighting

An ensemble weighting algorithm is a function that provides weights for the votes of each algorithm in the ensemble. For a more detailed discussion of ensemble vote weighting, see Section 2.2. To determine which of the ensemble vote weighting algorithms described in Section 2.2 might provide more accurate classifications, each vote weighting scheme was tested for three selected data sets.

The three selected data sets were students' responses to Item 1, Item 3, and Item 4. Two categories were selected from the rubrics used to categorize responses to each of these 3 items— Shape and Variability for Item 1, Center and Variability for Item 3, and Bar Height and Bar Chart for Item 4. For each of the three items, the four potential vote-weighting algorithms were utilized (under the default data cleaning scheme, described in Table 20) to categorize responses into the two selected categories. Table 19 shows the results of the different vote-weighting algorithms on each item and category combination. Since Item 1 and Item 3 utilized an analytic rubric and Item 4 utilized a holistic rubric, the results in Table 19 show Cohen's Kappa for Item 1 and Item 3 and Fleiss' Kappa for Item 4. The specifics of each data set, including rubric definitions, can be seen in Section 3.4.

Table 19: A comparison of the effects of various vote weighting techniques on Cohen's Kappa.

	Uniform	Probability-Based	Dynamic	CV
Item 1- Shape	0.891	**0.942**	0.931	0.935
Item 1- Variability	0.867	0.908	**0.918**	0.915
Item 3- Center	0.911	**0.975**	0.973	0.970
Item 3- Variability	0.817	**0.922**	0.912	0.891
Item 4- Bar Height	0.719	0.884	0.904	**0.947**
Item 4- Bar Chart	0.784	**0.916**	0.883	0.901

Uniform vote weighting was inferior to the other three vote-weighting schemes in all six test cases. There was, however, no overall most effective vote-weighting scheme for all of the data sets. Probability-based vote weighting outperformed the other three techniques in four of the six test cases, and dynamic vote weighting and cross-validation-based vote weighting each were top performers on a particular category. Based on these results, the difference in vote-weighting technique appeared to be minimal at best. Since the efficacy of each vote-weighting scheme appeared different for different data sets, each of the vote-weighting schemes were tested on all categorizations.

Table 21 shows the results of the best vote-weighting scheme for each of the 45 ensembles used in this dissertation. Of the four vote-weighting schemes, only three were utilized in the final form of any ensembles— uniform vote weighting was unused. In each case, at least one of probability-based, dynamic, or CV vote weighting outperformed uniform vote weighting. Probability-based vote weighting was used more often than dynamic or CV vote weighting. Probability-based vote weighting tended to perform better for generic categories and when specific words almost assuredly placed a response into the given category. For example, students in the Context category for Item 1- Atlanta Income were almost assuredly responding in context if their response contained the words 'Income,' 'Dollars,' or 'Atlanta

adults.' Although these types of categories were easier to predict than other, more convoluted categories, the probability-based weighting function performed the best.

CV-based vote weighting tended to perform better in categories where a small, specific subset of words was present in both correct and incorrect responses (such as the 'Height' categories representing the presence of the Bar Height Misconception). For example, students in the 'Height' category believed that the variability in the data is represented by the variation in bar heights. Students in this category would likely use words like 'bar,' 'height,' or other words describing the height of a specific bar in the graph. Students without this misconception (and thus not belonging to this category) would also likely use similar words to correctly describe the graph. These students often correctly identified the center of the graph by describing the location of the mode using similar terminology as those with the bar height misconception. CV-based vote weighting helped reduce the number of false positive categorizations by reducing the weights of votes from algorithms prone to false positives.

Dynamic vote weighting rarely outperformed the other algorithms, however, it consistently did so in the 'Correct' and 'Axis Order' categories. Due to the obfuscating nature of the calculations performed by this method as described in Fung et al. (2006), it is unclear why this vote-weighting scheme outperformed the other three for these particular categories. The use of dynamic vote weighting for these two categories (across all items) lead to a higher recall (i.e. missed fewer responses that truly belonged in either the 'Correct' or 'Axis Order' categories) than the other vote-weighting schemes.

4.5 Ensemble Training

There were five total data cleaning schemes used across all ensembles in this dissertation. Table 20 shows the summary characteristics of the data cleaning schemes. Each cleaning scheme contains accompanying algorithm tuning settings if algorithm tuning differed from the default scheme. The default data cleaning scheme is described in Section 3.5, and the default algorithm tuning scheme is described in Section 3.7. Each individual data cleaning scheme is described in greater detail later in this section.

Table 20: This table gives a summary of each of the five data cleaning schemes.

Cleaning Settings/Scheme	Default	2	3	4	5
Minimum word length	3	1	1	3	3
N-gram length	1	2	2	5	1
Numbers	Removed	Removed	Removed	Removed	Not Removed
Punctuation	Removed	Removed	Removed	Removed	Not Removed
Stopwords	Removed	Removed	Removed	Removed	Not removed
Matrix sparsity	0.99	0.99	0.99	0.99	0.99
Algorithm Tuning Changes	-	Tree-based methods altered to grow larger	Elastic-net uses Lasso penalization	C4.5 algorithm used for tree construction	-

Cleaning scheme 2 was exclusively used for algorithm predictions in the 'Shape' category. Many responses to items containing this category were shorter in length than a typical response. Typical responses belonging in this category were: 'Bell-shaped,' 'It looks normal,' or 'Skewed right.' To accommodate these atypically terse responses, the standard scheme was altered to accommodate bigrams and words of any length. The use of bigrams and words of any length helped identify features (such as the bigram 'skewed right') that were useful in differentiating between short, incorrect responses and short, correct responses. The tree-based algorithms in this scheme (Classification Trees, Bagging Classification Trees, Boosting Decision Stumps, and Random Forests) were tuned to utilize the large influx of features, since the addition of bigrams changed the number of features from tens of thousands to

hundreds of thousands. Trees were pruned in a way that allowed them to grow very large (i.e. removing the standard stopping criteria that ceases forming branches after a tree becomes 32 leaves deep). The boosting of decision stumps involves the creation of several (default 100) one-leaf trees. With only 100 stumps, it became likely that no features indicative of the 'Shape' category (e.g., the bigrams 'right skew,' 'left skew,' or 'normal distribution') were ever chosen to be the central leaf of a decision stump. In scheme 2, this default was changed to 1000 to better accommodate the larger number of potential features to be used in the creation of new decision stumps.

Cleaning scheme 3 was used exclusively for algorithm predictions in the Variability category across all items. This scheme was essentially identical to scheme 2 with an additional change to the Elastic-Net regularization of the logistic model. For the Variability category, using the L1 norm penalization function (used in lasso regression) vastly outperformed both the L2 norm penalization (ridge regression) and any weighted combination of these using the Elastic Net (described in Section 3.7.5). Thus α was set to 1 for the calculation of the parameter penalty. The rest of the scheme was identical to that of scheme 2.

Cleaning scheme 4 was used sparingly. This scheme was primarily used for the Height category, and it was used in one instance for the Correct category. This scheme was similar to the default scheme with two exceptions. Feature extraction was done in a unique was to avoid removing any potentially useful features. In this manner, all n-grams up to 5-grams (i.e. 1-grams, ... , 5-grams) were considered features. This exponentially increased the size of the document-term matrix, and so a more efficient tree-building algorithm was required to construct trees in a reasonable computation time. The C4.5 algorithm (Quinlan, 1993) was used in place of the standard ID3 algorithm to construct trees more efficiently. This design allowed

more complex sentence structures to be captured in a single feature— a necessary step in determining which students had the bar height misconception and which did not.

Cleaning scheme 5 was used for each of the Bar Chart categories and for two Correct categories. This scheme specifically did not remove punctuation, numbers, or stopwords during the cleaning process. These features proved particularly useful in determining which responses belonged in the Bar Chart category, as students often referred to the heights of the bars by location on the x-axis (i.e. a number) and by its height on the y-axis (i.e. a number). Students understanding that the data were aggregated (i.e. not having the bar chart misconception) tended to give responses of grouped numbers using a dash. For example, many students without the bar chart misconception would respond 'Most students slept 6-8 hours,' or 'The mean of this graph is 6-8.' Removing the dash or numbers from these responses obfuscated the accurate categorization of responses into the Bar Chart category.

4.6 Ensemble Accuracy

In this section, the ensemble analytics are provided for each of the ensembles used throughout this research to make categorizations of student responses. The training of each ensemble involved data cleaning and algorithm tuning, and the three primary measures of ensemble efficacy used in this dissertation— recall, precision, and Cohen's/Fleiss' Kappa—are shown here. Item 6, Item 7, and Item 8 use holistic rubrics, and thus the Kappa metric used is Fleiss' Kappa. The remaining items utilized analytic rubrics, and thus Cohen's Kappa is used. A more detailed discussion of these metrics can be found in Section 3.6. Table 21 shows the resulting three efficacy metrics for each ensemble used in this dissertation.

Table 20 shows the ensemble vote-weighting function and data cleaning schemes used to weight the votes and tune the algorithms in the corresponding ensemble. In addition, it describes any changes made to tuning parameters in the eight machine learning algorithms. Ensemble vote-weighting functions are described in detail in Section 2.2

Table 21: This table shows the three ensemble efficacy metrics, the weight function, and data cleaning scheme used for each of the categories predicted by the ensembles.

Ensemble Question	Category	Recall	Precision	Kappa	Weight Function	Cleaning Scheme
Item 1- Atlanta Income	Shape	0.99	0.99	0.98	Probability	2
	Center	0.86	0.96	0.83	Probability	Default
	Variability	0.82	0.96	0.76	Probability	3
	Context	0.87	0.96	0.83	Probability	Default
Item 2- Test Score Variability Histogram V1	Height	0.91	0.94	0.88	CV	4
	Normal	0.99	0.99	0.97	Probability	Default
	Spread	0.95	0.99	0.90	Probability	Default
	Correct	0.79	0.84	0.72	Dynamic	5
Item 3- Student Sleep V1	Shape	0.96	0.99	0.90	Probability	2
	Center	0.86	0.92	0.86	Probability	Default
	Variability	0.70	0.92	0.71	Probability	3
	Context	0.96	0.98	0.94	Probability	Default
Item 4- Test Score Variability Dot Plot V1	Height	0.89	0.94	0.88	CV	4
	Normal	0.99	0.99	0.95	Probability	Default
	Spread	0.97	0.99	0.96	Probability	Default
	Correct	0.81	0.88	0.76	Dynamic	5
Item 5- Student Sleep V2	Shape	0.98	0.99	0.95	CV	2
	Center	0.83	0.91	0.82	Probability	Default
	Variability	0.73	0.85	0.70	Probability	3
	Context	0.97	0.99	0.96	Probability	Default
Item 6- Test Score Variability Combination	Height	0.91	0.94	0.87	CV	4
	Normal	0.96	0.99	0.91	Probability	Default
	Spread	0.93	0.95	0.87	Probability	Default
	Correct	0.76	0.72	0.71	Dynamic	4
	Bar Chart	0.80	0.83	0.75	CV	5
	Axis Order	0.76	0.72	0.70	Dynamic	Default
	Range	0.93	0.95	0.93	Probability	Default
Item 7- Colored Test Score Variability Combination	Height	0.95	0.96	0.88	CV	4
	Normal	0.99	0.99	0.99	Probability	Default
	Spread	0.94	0.96	0.87	Probability	Default
	Correct	0.71	0.90	0.83	Dynamic	5
	Bar Chart	0.84	0.81	0.72	CV	5
	Axis Order	0.78	0.79	0.70	Dynamic	Default
	Range	0.96	0.97	0.91	Probability	Default
Item 8- Coffee Consumption	Shape	0.88	0.84	0.76	Probability	2
	Center	0.92	0.90	0.85	Probability	Default
	Variability	0.74	0.81	0.76	Probability	3
	Context	0.96	0.99	0.95	Probability	Default
	Bar Chart	0.83	0.87	0.72	CV	5
	Axis Order	0.79	0.78	0.70	Dynamic	Default
	Range	0.92	0.98	0.96	Probability	Default
Item 9- Test Scores for Large-Test	Shape	0.91	0.91	0.86	Probability	2
	Center	0.92	0.98	0.89	Probability	Default
	Variability	0.80	0.83	0.72	Probability	3

Each of the ensembles in Table 21 performed reasonably well. All Kappa values were larger than 0.7, and the recall and precision of each of the ensembles were typically larger than 0.85, although they were always larger than 0.7. Regardless of item, the corresponding ensemble predicted better for certain categories when compared to predictions for other categories. Across all items, the ensembles performed particularly well when categorizing responses into the Shape or Context categories. Predictions made about responses in the Center category predicted with high accuracy, especially those in Item 8. For other items, this accuracy was acceptable but noticeably less than that of the responses in Item 8. Responses belonging to the Variability category were particularly difficult to categorize across all items, and the biggest struggle was with the ensembles' precision (i.e. its ability to correctly identify responses with a discussion of variability) and not its ability to correctly identify responses lacking this discussion of variability.

Many of the holistic rubrics (here for Item 6, Item 7, and Item 8) had one or two categories that were particularly troublesome. Specifically, the Axis Order and Correct categories were the hardest to categorize across all items. Item 6 and Item 7 were overall particularly challenging for the ensembles to correctly classify due to the large variation in student responses belonging to these categories relative to those in other categories. For these items, there were six possible pairings of graphs, and each pairing contained displays of data that were drastically different than the other pairings. This lead to an increase in the number of different words students could use to correctly (or incorrectly) describe each of the aspects of the corresponding histograms or dot plots.

5 Discussion

This section begins with a brief summary of the scope and design of this dissertation. It then contains a discussion of the results of Section 4 and each of its corresponding subsections. This section concludes with a discussion of the limitations of this study and of potential future directions for research.

5.1 Study Summary

Students' understanding of variability in data plays a key role in solving statistical problems (Franklin et al., 2007). Cooper and Shore (2008) stated that one of the consistently more difficult tasks for students in introductory statistics courses of any level is interpreting the variability displayed in univariate graphs such as histograms and dot plots. This study sought to quantify the proportion of students with misconceptions about these graphs, specifically related to the variability in the data they display. In addition, this study sought to develop a manner in which instructors could receive meaningful, instantaneous feedback about their students' knowledge through the use of an ensemble of machine learning algorithms. This dissertation began by developing nine constructed-response questions to be given to thousands of undergraduate students at a large research institution in the Southeastern United States. Students' responses were recorded and used as training and testing data for eight machine learning algorithms. These machine algorithms were trained to categorize responses for each of the nine items. The predictions of individual algorithms were combined into an ensemble using one of four vote-weighting functions. The

accuracy of each ensemble was optimized by changing tuning parameters for each algorithm and modifying the overall data cleaning process. When peak ensemble performance was achieved, the ensemble was used to categorize student responses into many categories of interest.

Finally, face-to-face, task-based interviews were performed with 19 undergraduate students enrolled in the same introductory statistics course at the same institution. These interviews were used to determine if students' responses to online homework questions were of a similar completeness and correctness as those responses given to isomorphic interview questions. From the results of these interviews, it was ultimately concluded that students' responses to isomorphic online questions and in-person interview tasks were of a similar completeness and correctness. Due to this continuity of responses, it seems reasonable to conclude that student responses to online homework assignments were an accurate representation of the students' knowledge about variability in graphs. Thus, an ensemble of machine learning algorithms could be used to evaluate the efficacy of a particular lecture or determine the pre-requisite knowledge (or lack thereof) of a class of undergraduate statistics students simply by categorizing student responses to constructed-response items.

5.2 Discussion of Results

This section gives a concise summary and discussion of the results contained in this dissertation. The answers to each of these four research questions: **Dot Plot vs. Histogram**, **Medium**, **Misconceptions**, and **Ensemble Accuracy** are discussed in the subsequent subsections.

5.2.1 Dot Plot vs. Histogram

In both online items and interview tasks, students typically performed better on questions related to variability in graphs involving dot plots than their isomorphic histogram counterparts. Students performed better when viewing dot plots over histograms in 11 of the 12 isomorphic *Compare* questions in which both dot plots and histograms were utilized. Essentially, students more often correctly identified which of two graphs contained more variable data when viewing a dot plot than when viewing a histogram. One of the most common misconceptions that seemed to inhibit students from making the correct selection in these types of items was the **Bar Height Misconception**. Since this misconception was approximately 10% less prevalent in student responses to dot plots, it is plausible that relatively more students understood that a dot plot was showing aggregated data. As further evidence of this, the difference in the proportion of correct responses for dot plots and histograms for each question approximately corresponded to the difference in prevalence of the **Bar Height Misconception** between dot plots and histograms (also approximately 10%).

Regardless of whether the student was viewing a pair of histograms or dot plots, students giving a correct multiple choice answer to a *Compare* question used nearly identical words in the constructed-response portion to justify their multiple choice answers (aside from the standard differences in descriptive words between the two graphs). That is, the only differences in word usage between correct responses to dot plots and correct responses to histograms involved words related to dots and bars. Students selecting the incorrect multiple choice gave a variety of incorrect justifications in the constructed-response portion of the *Compare* question, and these responses differed depending on graph type. Averaged across all items, 60% of students giving an incorrect answer focused on the bar height when comparing the

variability between two histograms. This percentage was only 48% for dot plots. Unlike responses to histograms, many incorrect responses to dot plots contained partially correct justifications (approximately 14% for dot plots compared to 5% for histograms). For example, students often gave a correct synopsis of the dot plot (e.g., 'There are dots spread out everywhere'), but then chose the other graph with *less* variability. It is unclear why such a large proportion of students made this decision across each of the online items. Further research is needed to investigate why this might have occurred, as there were relatively few responses to histograms that contained a reasonably correct justification but an incorrect graph selection. Since these incorrect selections occurred over many semesters across many different questions, it is unlikely that these incorrect selections are due to students accidentally selecting the wrong multiple choice answer.

When asked *Describe* questions during interviews, students were much more likely to point at individual dots (often outlying points) and make observations about their value or judgments about why the dots were there in the context of the problem. There was likewise evidence of this in the online responses where students would more often identify outlier points in the questions involving dot plots versus those involving histograms. There was little evidence of this occurrence in both online items and interview tasks. When discussing general features about the distribution, there were very few differences between student responses to dot plots and histograms. In both instances, students frequently addressed the center of the graph and then made some mention of its overall shape. Students were slightly more likely to discuss the variability for dot plots versus histograms. Most of the discussion of variability involved identifying outlying points in the dot plots. The identification of these outlier points appeared slightly more frequently in responses describing dot plots than those describing histograms.

It is unclear whether or not viewing a dot plot instead of a histogram has any long-term effect on student knowledge. That is, there was significant evidence that students more often interpret correctly the variability displayed in a dot plot than the variability displayed in a histogram. There was no evidence, however, that this increase in interpretability leads students to overcome the many misconceptions present in univariate graphs. A more longitudinal research would be required to determine if such evidence exists.

5.2.2 Medium

The primary difference between student responses to online items and their isomorphic interview task counterparts was the length of the response. For example, the average response length for a response to Item 8 was 15 words ($s = 7.8$). For its isomorphic interview task counterpart, Task 5, responses averaged 58 ($s = 35.3$) words. This ratio—around 1:4 in terms of word count—was consistent throughout each of the online and in-person responses. Despite the differences in response length, students gave nearly identically constructed responses to isomorphic questions across both mediums. Students' responses to both online items and interview tasks were both the same levels of correctness and completeness. That is, the same students who gave correct responses to the online items also gave correct responses to the isomorphic interview tasks, and the same students who gave incorrect responses to the online items also gave incorrect responses to the isomorphic interview tasks. Additionally, student responses contained the same level of completeness across both mediums. Although responses to online items typically contained fewer words than their interview counterparts, the same justifications to both incorrect answers and correct answers were used throughout each of the responses.

5.2.3 Misconceptions

The prevalence of the four established misconceptions about graphs of univariate data remained consistent across mediums. The prevalence of these misconceptions about graphs, however, varied significantly across items. First, significantly fewer students exhibited the **Bar Chart Misconception** and **Bar Height Misconception** when viewing a dot plot (22% and 69%, respectively) than when viewing a histogram (41% and 60%, respectively). This may be due to the fact that dots, moreso than bars, help students better understand that the bars and stacks of dots represent aggregated data and not data from a single individual. Thus, larger variations in bar heights do not necessarily equate to a large variability in the data. There were no other significant differences in the prevalence of the four established misconceptions in responses between isomorphic histograms and dot plots.

Some questions were significantly more likely to evoke responses that contained particular misconceptions. Interview Task 3 (Figure 7.15) and Pairing 9.1/9.2 from Item 7 (Figure 7.12) were specifically designed so that the data set with the larger range had less variability. As such, there were significantly more responses to these questions categorized as having the **Range Misconception** (63% and 48%, respectively, as compared to 8% from all other items). Results showed that Task 3 was as equally challenging as Pairing 9.1/9.2 from Item 7, however the percentage of responses containing the **Range Misconception** varied significantly. Although each question was answered correctly approximately 40% of the time, responses to Item 7 exhibited a variety of incorrect justifications, whereas each of the incorrect responses to Task 3 appeared to be due specifically to the **Range Misconception**. The raw data form of Task 3 was only given to students in interview form. For future research, it would be useful to give the same question as an online version so that a larger sample of responses could be obtained.

The **Axis Order Misconception** occurred rarely across all online items and interview tasks. This misconception, written about in delMas et al. (2005), appears near-inseparable from the **Bar Chart Misconception**. That is, students who believe that histograms or dot plots are actually case-value bar charts often also believe that the x-axis is arbitrarily ordered. Very few students (typically around 1% per question) provided evidence in their responses that they understood that they are viewing a histogram or dot plot *and* believe the x-axis to be arbitrarily ordered.

Overall, it appears that determining whether or not a particular student has a given misconception with very high accuracy (>99%) may require more than one item. This is due in part to the error rates of the ensembles at making predictions about the four primary misconceptions. While the error rates for the ensembles in this dissertation are not incredibly high (in all cases fewer than 5%), asking several (3-5) questions about a similar topic would greatly improve the chances of the ensembles making a correct overall prediction about whether or not the student has a particular misconception. In addition, students that exhibited a particular misconception on one item did not necessarily do so for another item. From this study, it is impossible to determine whether this was due to the student developing the particular misconception over the course of the semester, having that particular misconception targeted by an instructor, or due to some other cause. Regardless, there is evidence that a single question is insufficient at determining whether a student may or may not have a particular misconception. Even at the aggregate level, multiple questions may be required to accurately determine what proportion of the class has a particular misconception.

5.2.4 Ensemble Training and Accuracy

Section 4.6 detailed the efficacy of the 45 different ensemble models used in this dissertation to make categorizations about student responses. Overall, the models in this dissertation performed particularly well— 27 of the 45 models had Precision, Recall, and Kappa values larger than 0.8. Some of these ensembles, however, took tens of iterations before performing at an acceptable level. Although some categories were particularly easy for the ensemble to correctly categorize (e.g., categories about context, categories about the basic shape of the graph), others proved more challenging and required particular feature extractions to prove effective. This was particularly the case in hard-to-classify categories such as 'Height' from the *Compare* rubric and 'Variability' from the *Describe* rubric. Aside from algorithm and ensemble tuning, rubrics had to be revised several times to ensure that the given codes aligned with the goals of the categorization scheme developed by statistics experts. Most of the categories that were too difficult to accurately classify (and thus omitted from this dissertation) were due to the relatively low prevalence of student responses belonging to that category. In order for the algorithms to classify appropriately, results from this dissertation showed that a minimum of approximately 10-15 responses were required in each rubric category for the training data set. In addition, one should not underestimate the amount of time required to train an accurate ensemble for some of these categories (typically tens of hours), and this has proved to be a potential drawback of this technique.

Table 21 in Section 4.6 also gave a potentially misleading look into the efficacy of ensemble models for general categorization. There were at least ten categories for the 9 items used in this dissertation that had to be removed or completely reworked due to the inability of an ensemble of machine learning algorithms to correctly identify responses belonging to these particular categories, even after nu-

merous revisions. Many of these categories were removed due to the low prevalence of responses in this category, but others were removed or altered due to an overlap between two mutually exclusive categories (e.g., many categories like 'Fully correct' and 'Partially correct' were removed due to the difficulty of differentiating between border cases). The prompts of some items were worded in such a way that no meaningful classifications could be made by any ensemble, and these items had to be removed from the study. In short, the particular ensembles used in this dissertation performed well. This, however, may not be the case for ensembles making classifications about student responses for all categories. Ensembles used for classification may require a substantial time investment in refining both the item and prompt, as well as the rubric classification bins.

5.3 Study Limitations

Although the results in this dissertation illuminated the propensity of students to have particular misconceptions about variability in graphs, they did so only regarding students at a single research institution. That is, the results in this study are only generalizable for students at the single institution, and the percentages of students with particular misconceptions may differ significantly across institutions. Additionally, all supervised learning algorithms, machine learning algorithms which utilize a training set of data and a test set for accuracy testing, assume that new, uncategorized data behave in an identical fashion to that of the training data. Thus, if new responses to constructed-response questions use a different vernacular or are otherwise constructed in a substantially different manner than those responses in the training data set, the ensembles constructed here may perform significantly worse than what is shown in Section 4.6 (Ikonomakis et al., 2005).

This study only utilized nine online items to make judgments about students' misconceptions. Additionally, misconceptions in the *Compare* rubric were assumed to be mutually exclusive (that is, a student could only have one of the given misconceptions). This was done to improve model accuracy and appeared to be a reasonable assumption, but it is almost assuredly not the case that each student only had at most one misconception. Ideally, a larger number of items would be given to the same group of students. Due to limitations on the number of questions any given student could receive, only a small subset of the nine online questions were given to students in any one semester. Moreover, the results of this study are contingent upon each student giving their best effort when responding to online constructed-response questions. Although students' responses to online items appeared to match their responses to in-person interview tasks, this was only true for $n = 19$ students. It is also possible, albeit unlikely, that students did not respond to questions in either medium in a manner that would demonstrate each student's true knowledge. An interview study with more students might further illuminate the validity of these issues.

There were several online items for which students randomly received one of six possible pairings of either a histogram or a dot plot. Post-hoc analysis showed significant evidence that certain pairings appeared more often for students than other pairings. Due to the nature of the online platform used to distribute homework questions to students, WebAssign, it was impossible to determine how or why this may have occurred. Despite this strong evidence indicating that certain pairings were more likely to be shown to students than others, there was no reason to believe this ultimately had any effect on the results of the study (besides differing sample sizes). Each pairing was given to at least 150 students in all cases, and after controlling for other covariates, there was no evidence of a relationship between the rate at which

students received a specific pairing and the likelihood that the given student gave a correct response.

5.4 Future Directions

This section describes the future directions for this research. It begins with a pathway for expanding the number of questions with a trained ensemble of machine learning algorithms. It then discusses the further work necessary for better optimization of the algorithms in order to categorize with improved efficacy.

5.4.1 Expanding the Study

This study gave a description of common misconceptions and their corresponding prevalence held by students at a large research institution in the Southeastern United States. In the future, it would be useful to carry out an identical study at other institutions to compare the results across multiple institutions, particularly institutions from different geographical regions or institutions with different student demographics. A wider array of student responses would also augment the ensemble accuracy by diversifying the training data. In particular, data from other institutions would shed light on the generalizability of the work in this dissertation— if students from other institutions use a similar vernacular as those at the institution used in this study, then ensembles built from data gathered from this institution could be used to make prediction about new responses from other institutions. This would imply that data from a single, representative institution could be collected and used when categorizing responses from any other institution. A quantification of any amount of decrease in model efficacy should also be considered.

5.4.2 Expanding the Questions

Many items in this study, both online and in-person questions, could be revised to more clearly target and identify the presence of particular misconceptions. When originally written, it was unclear how prevalent particular misconceptions would be and how to best construct an item to identify if a particular student had or did not have the target misconception. Certain questions, such as Task 5 and Item 8, were revised multiple times after students struggled to understand what the histogram in each question was displaying. Other questions—such as Task 3 or several of the pairings from Item 6—proved particularly difficult for students to answer correctly. Further investigation is required to determine exactly why these questions (among others in this study) were significantly more difficult.

In addition to refining the current set of questions, one could expand the scope of the items and tasks to cover a wider range of statistical topics. For example, one could develop a series of questions about any particular statistical topic (say, the Law of Large Numbers or the Central Limit Theorem). These questions could be asked to students and their responses used as training data for all future responses to this particular question. Then, instructors would have a larger database of trained ensembles with which to categorize students' responses across a variety of statistical topics. This variety in question type would allow instructors to choose question with a trained ensemble, ask the question to their students, and receive immediate feedback about how their students are doing. In the future, it would be largely advantageous to have a set of questions with trained ensembles for every introductory statistics topic.

5.4.3 Enhancing the Models

This study mainly utilized eight machine learning algorithms. There are, however, other machine learning algorithms that could be used to make predictions in a similar manner. At the onset of this study, Neural Networks were used to categorize student responses. Although Neural Networks have a history of making accurate classifications for complex data sets (Haykin, 1998), it became evident that Neural Networks required responses with more words than the observed in this study. Responses in this study typically contained fewer words than was required for an accurate neural network categorization. If future questions demanded responses of a greater length (i.e., three or more sentences), machine algorithms such as neural networks could potentially be used to make classifications. Half of the machine learning algorithms in this study are tree-based: Classification Trees, Bagging Classification Tress, Boosting Decision Stumps, and Random Forests. Although the ensembles in this study typically contained four non-tree-based algorithms, it would be advantageous to try implementing a wider variety of algorithms, such as Naive Bayes or Relevance Vector Machines, to reduce potential correlation between the models in the ensemble.

With regard to ensemble classification accuracy, the use of a wider range of algorithm vote-weighting mechanisms to improve accuracy could be investigated. There was significant evidence that a simple, uniform vote-weighting scheme was outperformed by nearly any other vote-weighting scheme. Across all items and for all categories used in this study, at least one of the three other vote-weighting schemes (probability-based, dynamic, and CV-based) outperformed simple uniform vote-weighting. Each of these schemes was used on data from questions about statistics. Thus, testing these vote-weighting schemes using responses to questions from different disciplines would help strengthen this claim. In addition, alternate vote-

weighting schemes could be tested to determine if there exist vote-weighting functions that were more efficient than the ones used in this study. One such method involves a regression-based approach. In this method, each individual algorithm is a predictor variable in a cumulative logistic regression model that predicts the overall classification. There also exist numerous dynamic ensemble vote-weighting functions that were untested in this dissertation. Given the lack of an observed pattern regarding which combinations of vote-weighting schemes and rubric categories led to higher model accuracy, these alternate vote-weighting schemes may ultimately yield ensembles with higher precision and recall for certain data sets or when a particular data cleaning scheme is used.

6 References

Adeli, H., & Hung, S. (1995). *Machine learning : neural networks, genetic algorithms, and fuzzy systems.* New York, NY : Wiley.

Agresti, A. & Franklin, C. (2013). *Statistics : The art and science of learning from data (3rd ed.).* Upper Saddle River, NJ: Pearson.

Alpaydin, E. (2010). *Introduction to Machine Learning.* Cambridge, MA: The MIT Press.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, 3(2), 64—83.

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1), 49—62.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, 45(1-3), 35—65.

Ben-Zvi, D., & Sharett-Amir, Y. (2005). How do primary school students begin to reason about distributions. In *Reasoning about distribution: A collection of current research studies. Proceedings of the fourth international research forum on statistical reasoning, thinking, and literacy (SRTL-4)*, University of Auckland, New Zealand, 2—7.

Berger, A., Pietra, V., & Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1), 39—71.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* New York, NY: Springer. 206–209.

Blair, R., Kirkman, E.E. & Maxwell, J.W. (2013). *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States.* Providence, RI: American Mathematical Society.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, 26(3), 801—849.

Bright, G. W., & Friel, S. N. (1998). Graphical representations: Helping students interpret data. In S. P. Lajoie, S. P. Lajoie (Eds.), *Reflections on statistics: Learning, teaching, and assessment in Grades K12*, 63—88. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9—13.

Cohen, J.(1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20 (1): 37—46.

Cooper, L., & Shore, F. (2008). Students misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, 16(2), 1—13.

Cooper, L. L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education*, 18(2), 1—16.

Creswell, J. (2015). *A concise introduction to mixed methods research.* Thousand Oaks, CA: SAGE.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1—15. Heidelberg, Germany: Springer.

delMas, R. C., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students statistical reasoning. *Journal of Statistics Education*, 7(3).

DelMas, R., & Garfield, J. (1990). The use of multiple items to identify misconceptions in probabilistic reasoning. In *Research Papers from the Third International Conference on Teaching Statistics*. University of Otago, Dunedin, New Zealand.

delMas, R. C., & Liu, Y. (2003). Exploring Students Understanding of Statistical Variation. In C. Lee (Ed), *Reasoning about Variability: A Collection of Current Research Studies* [On CD]. Dordrecht, the Netherlands: Kluwer Academic Publisher.

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students difficulty reading and interpreting graphical representations of distributions. In *Fourth Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-4)*. Auckland, New Zealand.

Dzeroski, S., & Zenko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. *Machine Learning*, 54(3), 255—273.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. Alexandria, VA: American Statistical Association.

Freund, Y., & Schapire, R. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal Of Computer And System Sciences*, 55(1), 119—139.

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. *ICML*, 148—156.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Japanese Society For Artificial Intelligence*, 771—780.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337—407.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (Vol. 1)*. New York, NY: Springer series in statistics.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189—1232.

Fung, G., Yu, J., Haixun, W., Cheung, D., & Huan, L. (2006). A Balanced Ensemble Approach to Weighting Classifiers for Text Classification. *Sixth International Conference On Data Mining*, 869—873. Hong Kong, HK.

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In *The challenge of developing statistical literacy, reasoning and thinking*, 397—409. New York, NY: Springer.

Gelfand, S. B., Ravishankar, C. S., & Delp, E. J. (1989). An iterative growing and pruning algorithm for classification tree design. In *Systems, Man and Cybernetics, IEEE International Conference*, 818—823. Ulm, Germany.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12, 993—1001.

Hastie, T., & Qian, J. (2014). *Glmnet vignette* [PDF Document]. Retrieved from `http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf`.

Haykin, S. S. (1998). *Neural Networks : A Comprehensive Foundation*. Upper Saddle River, N.J. : Prentice Hall.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4), 18—28.

Ho, T. (1998). The Random Subspace Method for Constructing Decision Forests (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832—844.

Iba, W., & Langley, P. (1992). Induction of one-level decision trees. In *Proceedings of the Ninth International Conference on Machine Learning*. Aberdeen, Scotland. 233—240.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8), 966—974.

Jin, C., & Wang, L. (2012). Dimensionality dependent PAC-Bayes margin bound. *Advances in Neural Information Processing Systems*, 1034—1042.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* 137—142. Heidelberg, Germany: Springer.

Kaplan, J., Gabrosek, J., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, 22(2), 17.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(2-3), 271—274.

Konold, C., & Higgins, T. (2002). Highlights of related research. *Developing mathematical ideas: Working with data*, 165—201.

Kuncheva, L., & Whitaker, C. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2), 181—207.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18—22.

Magerman, D. M. (1995). *Statistical decision-tree models for parsing* [PDF Document]. Retrieved from `http://acl-arc.comp.nus.edu.sg/archives/acl-arc-090501d4/data/pdf/anthology-PDF/P/P95/P95-1037.pdf`

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation (PDF). *Sixth Conf. on Natural Language Learning (CoNLL)*, 49–55. Baton Rouge, LA.

Martinez, A. M., & Kak, A. C. (2001). Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2), 228—233.

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (1999). A Machine Learning Approach to Building Domain-Specific Search Engines. *International Joint Conference On Artificial Intelligence*, 16(2), 662—667. Chicago, IL.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. New York, NY: Chapman and Hall.

Meir, R., & Ratsch, G. (2003). An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning*, 118—183. Canberra, Australia.

Meletiou-Mavrotheris, M., & Lee, C. (2010). Investigating college-level introductory statistics students prior knowledge of graphing. *Canadian Journal of Science, Mathematics and Technology Education*, 10(4), 339-355.

Muhlbaier, M. D., Topalis, A., & Polikar, R. (2009). Learn C^{++}. NC: Combining Ensemble of Classifiers With Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes. *IEEE transactions on neural networks*, 20(1), 152—168.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning Vol 1.* 81—106.

Raschka, S. (2014). *Linear Discriminant Analysis bit by bit* [PDF Document]. Retrieved from `http://sebastianraschka.com/Articles/2014_python_lda.html`.

Rokach, Lior, & Maimon, O. (2008). *Data mining with decision trees: theory and applications.* Toh Tuck Link, Singapore: World Scientific.

Meir, R., & Ratsch, G. (2003). An introduction to boosting and leveraging. In *Advanced lectures on machine learning*, 18—183. Heidelberg, Germany: Springer.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3), 297—336.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

Srivastava, A., Han, E. H., Kumar, V., & Singh, V. (1999). Parallel formulations of decision-tree classification algorithms. In *High Performance Data Mining.* 237—261. New York, NY: Springer.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2), 245—266.

Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 42—49.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301—320.

7 Appendix

7.1 WebAssign Items

7.1.1 Item 1- Atlanta Income

Prompt- The histogram below shows the distribution of yearly income in dollars for a random sample of 356 adults living in Atlanta, GA. Describe as completely as possible the distribution shown in the histogram, being sure to explain what the graph tells you about yearly income for adults in Atlanta.

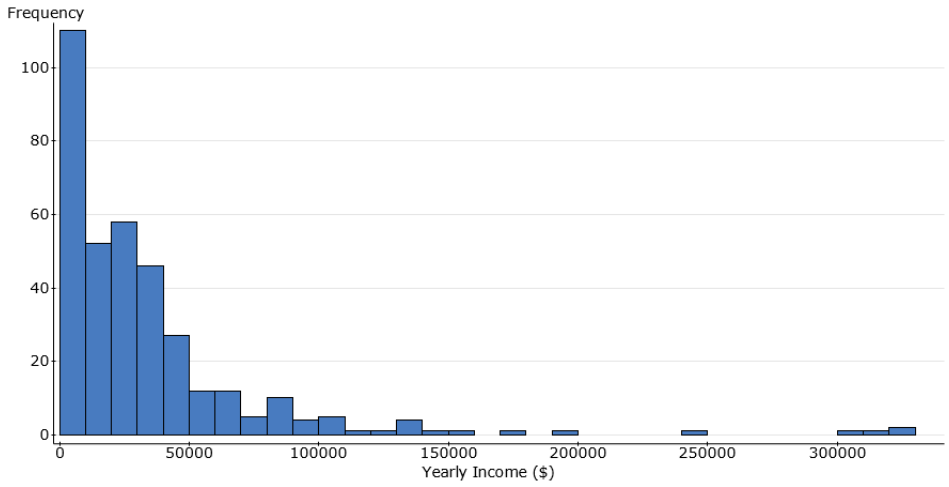


Figure 7.1: Item 1- Atlanta Income

7.1.2 Item 2- Test Score Variability Histogram V1

Prompt- The histograms below show the distribution of scores on a 10 item test for two classes.

a. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?

A) Class A has more variable scores

B) Class B has more variable scores.

b. Explain how you know from the graphs that the scores in the class you chose are more variable.

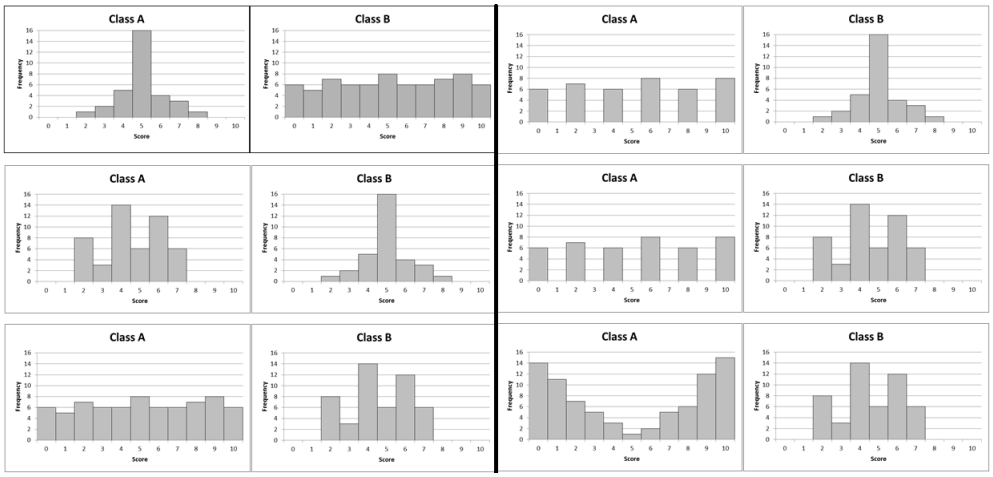


Figure 7.2: Item 2- Test Score Variability Histogram V1

7.1.3 Item 3- Student Sleep V1

Prompt The histogram below shows the distribution of the number of hours a random sample of 471 high school students in Georgia slept on the previous school night.

Distribution Prompt- Describe as completely as possible the distribution shown in the histogram.

Variable Prompt- Describe as completely as possible what the graph tells you about the number of hours high school students in Georgia sleep on school nights.

Both Prompt- Describe as completely as possible the distribution shown in the histogram, being sure to explain what the graph tells you about the number of hours high school students in Georgia sleep on school nights.

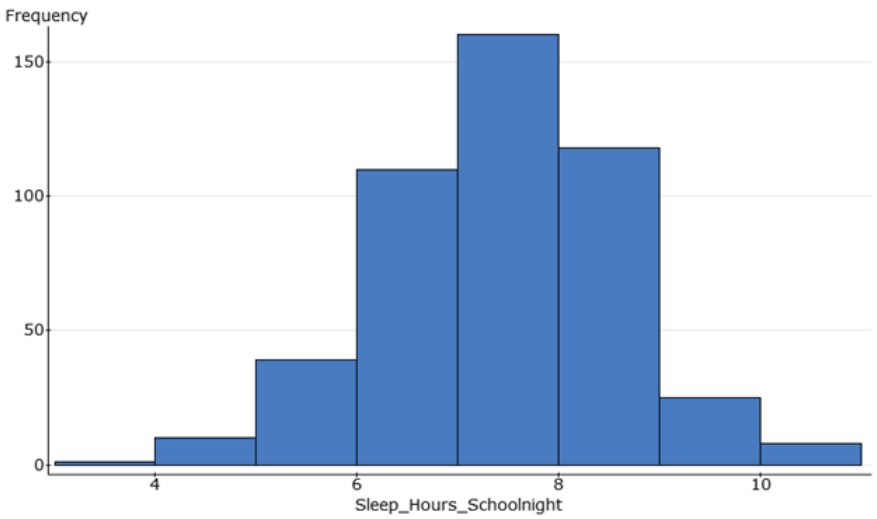


Figure 7.3: Item 3- Student Sleep V1

7.1.4 Item 4-Test Score Variability Dot Plot V1

Prompt- The dot plots below show the distribution of scores on a 10 item test for two classes.

a. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?

A) Class A has more variable scores

B) Class B has more variable scores.

b. Explain how you know from the graphs that the scores in the class you chose are more variable.

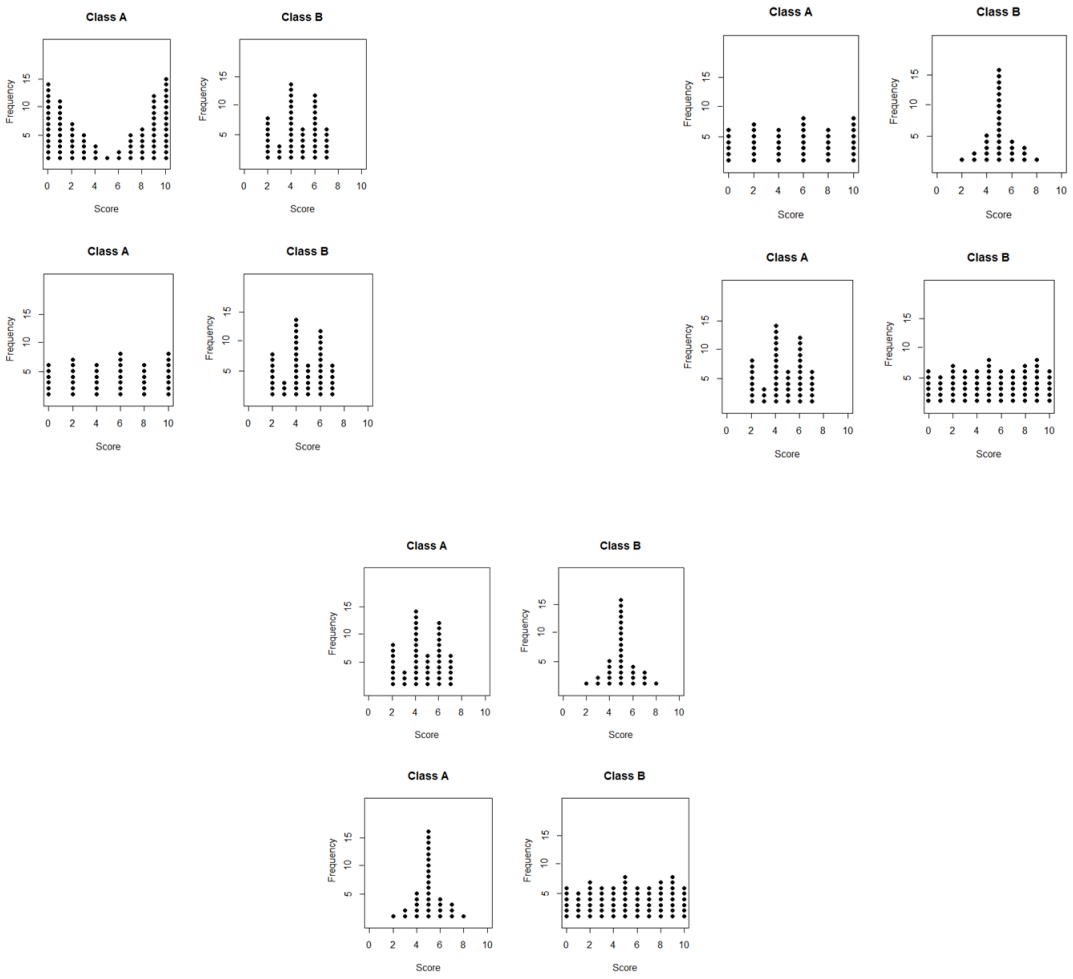


Figure 7.4: Item 4- Test Score Variability Dot Plot V1

7.1.5 Item 5- Student Sleep V2

Prompt The histogram below shows the distribution of the number of hours a random sample of 471 high school students in Georgia slept on the previous school night. Describe the distribution of the number of hours high school students in Georgia sleep, as shown in the histogram.

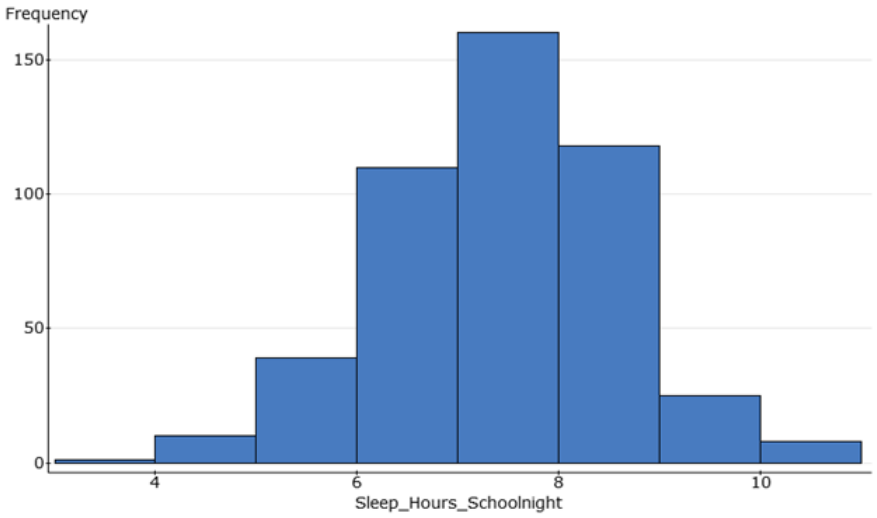


Figure 7.5: Item 5- Student Sleep V2

7.1.6 Item 6- Test Score Variability Combination

Prompt- The dot plots (histograms) below show the distribution of scores on a 10 item test for two classes.

a. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?

A) Class A has more variable scores

B) Class B has more variable scores.

b. Explain how you know from the graphs that the scores in the class you chose are more variable.

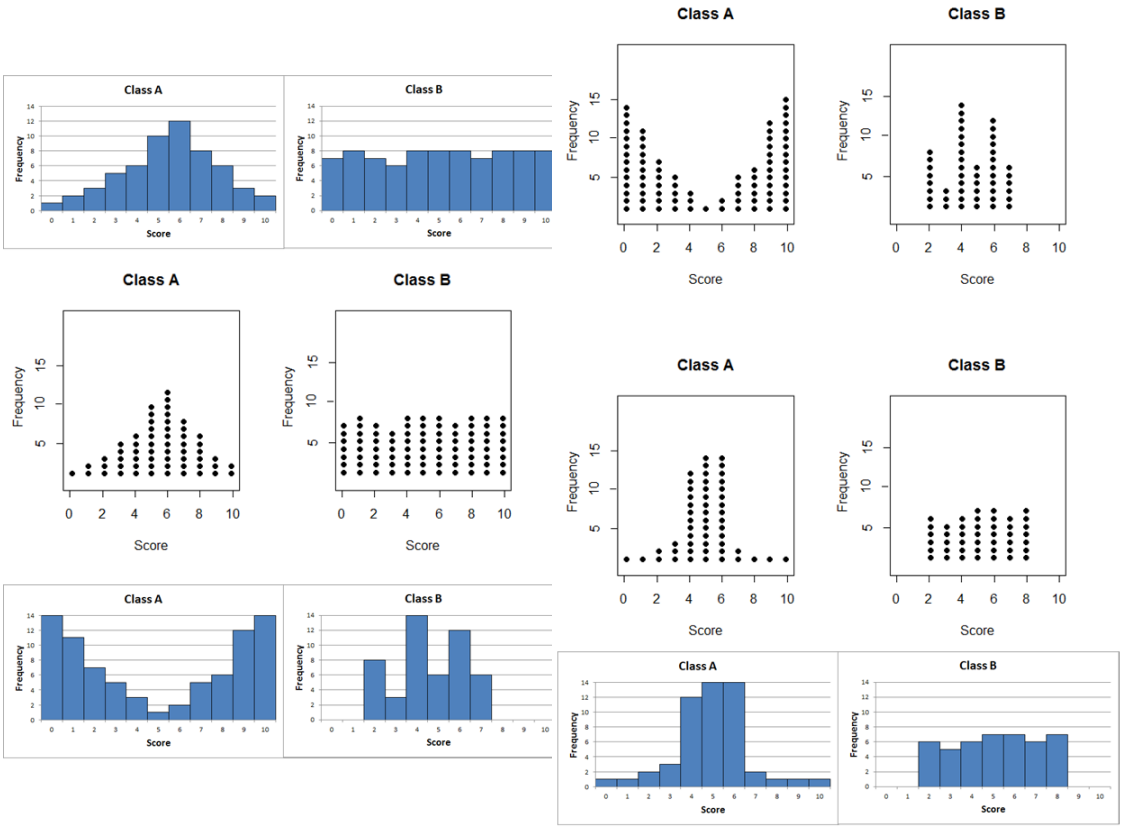


Figure 7.6: Item 6- Test Score Variability Combination

7.1.7 Item 7- Colored Test Score Variability Combination

Prompt- The dot plots (histograms) below show the distribution of scores on a 10 item test for two classes. Note that scores on each test have been classified as Excellent, Good, or Poor.

a. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?

A) Class A has more variable scores

B) Class B has more variable scores.

b. Explain how you know from the graphs that the scores in the class you chose are more variable.

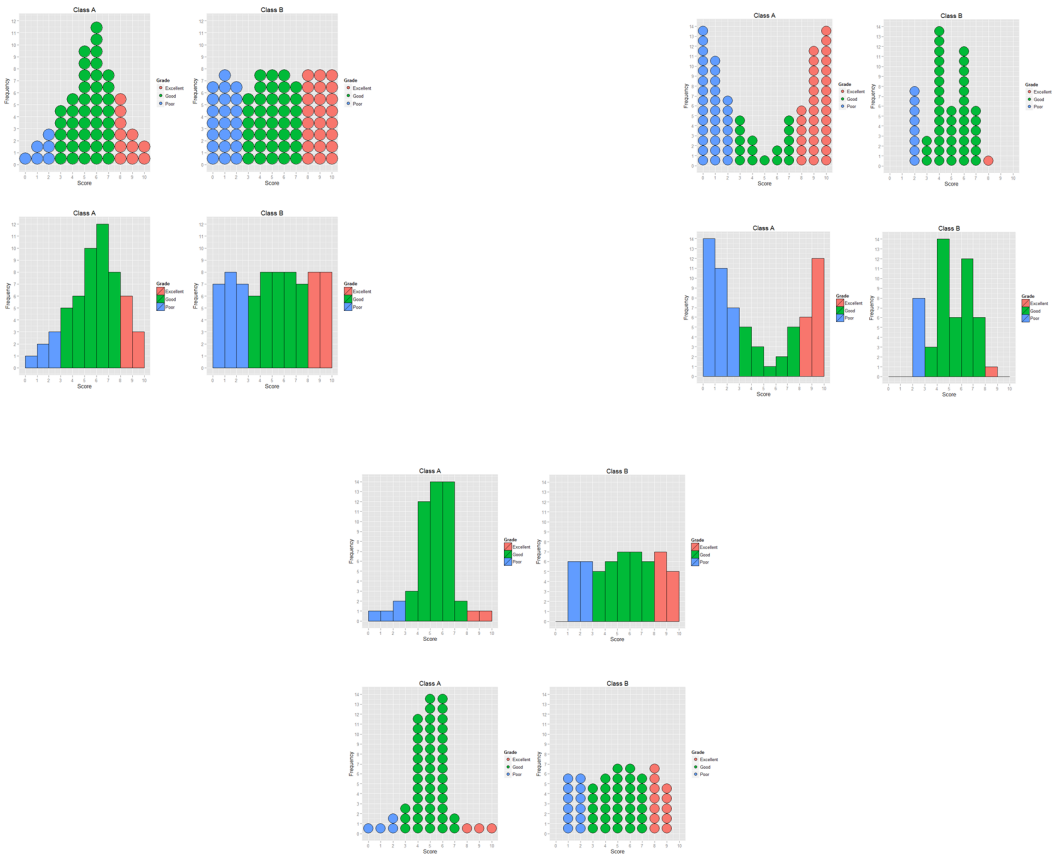


Figure 7.7: Item 7- Colored Test Score Variability Combination

7.1.8 Item 8- Coffee Consumption

Prompt- The histogram below shows the distribution of the number of ounces of coffee a random sample of 237 college students drank the previous day.

Describe the distribution of the number of ounces of coffee college students drink as shown in the histogram.

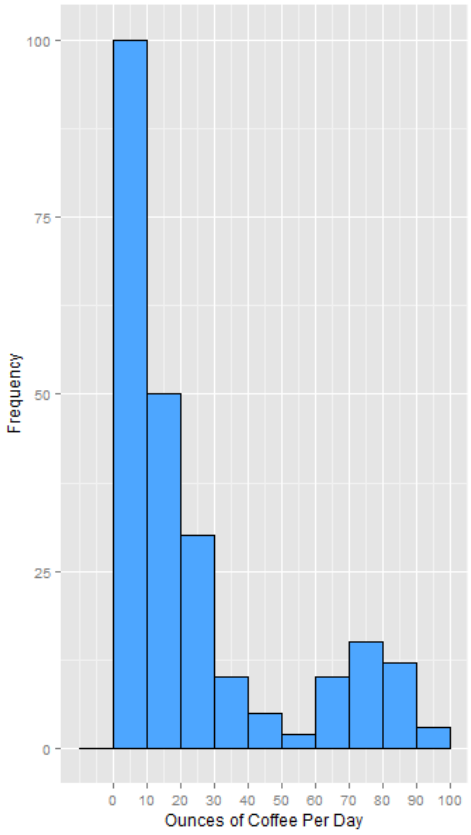


Figure 7.8: Item 8- Coffee Consumption

7.1.9 Item 9- Test Scores for Large-Test Histogram

Prompt- The histogram on the left shows the distribution of Class A's test scores for a mathematics test. The histogram on the right shows the distribution of Class B's test scores on the same test.

- **Center** Compare the centers of the distributions of test scores for Class A (Left Histogram) and Class B (Right Histogram).
- **Shape** Compare the shapes of the distributions of test scores for Class A (Left Histogram) and Class B (Right Histogram).
- **Variability** Compare the variability of the distributions of test scores for Class A (Left Histogram) and Class B (Right Histogram).

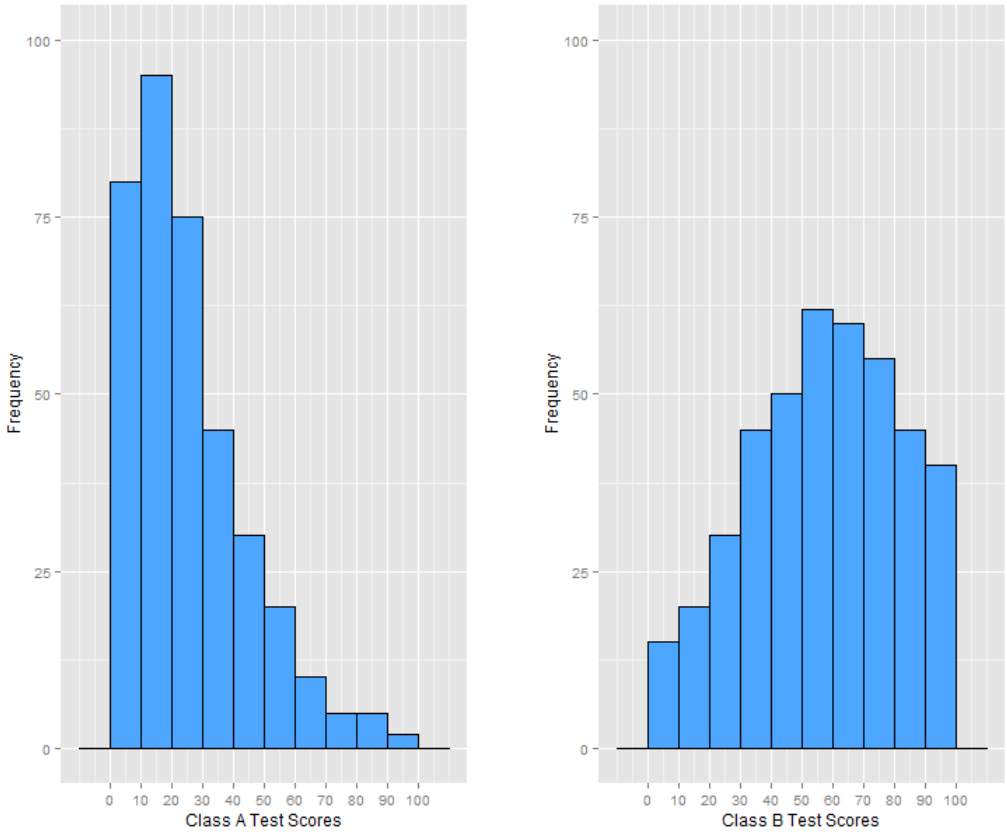


Figure 7.9: Item 9- Test Scores for Large-Test Histogram

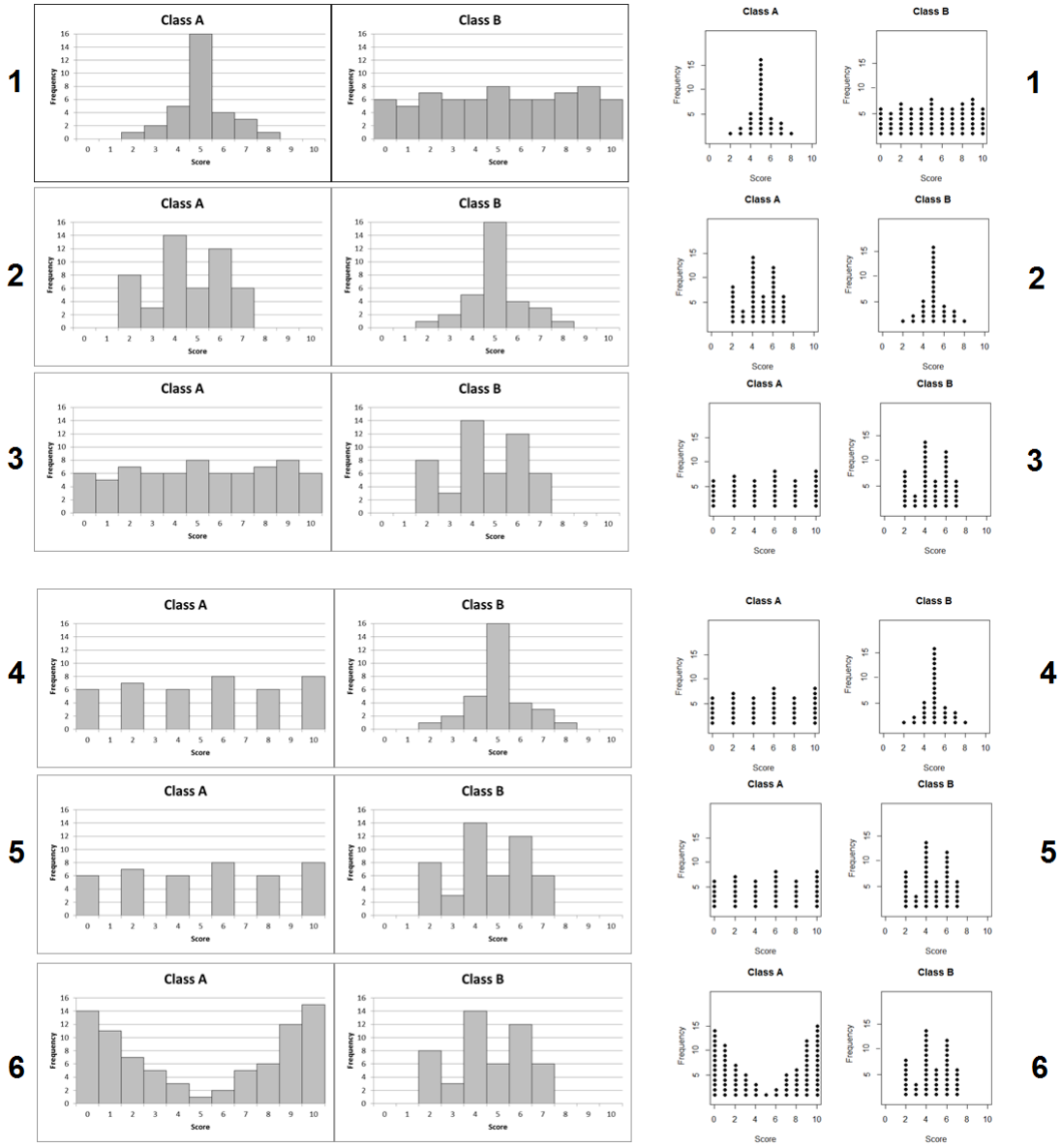


Figure 7.10: Pairings 1 through 6. Graphics are from Item 2 (histograms, left) and Item 4 (dot plots, right).

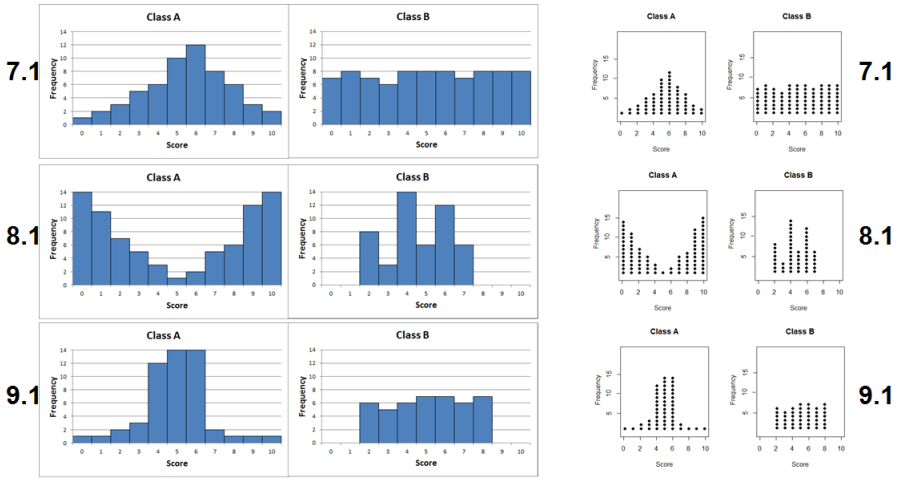


Figure 7.11: Pairings 7.1 through 9.1. Graphics are from Item 6.

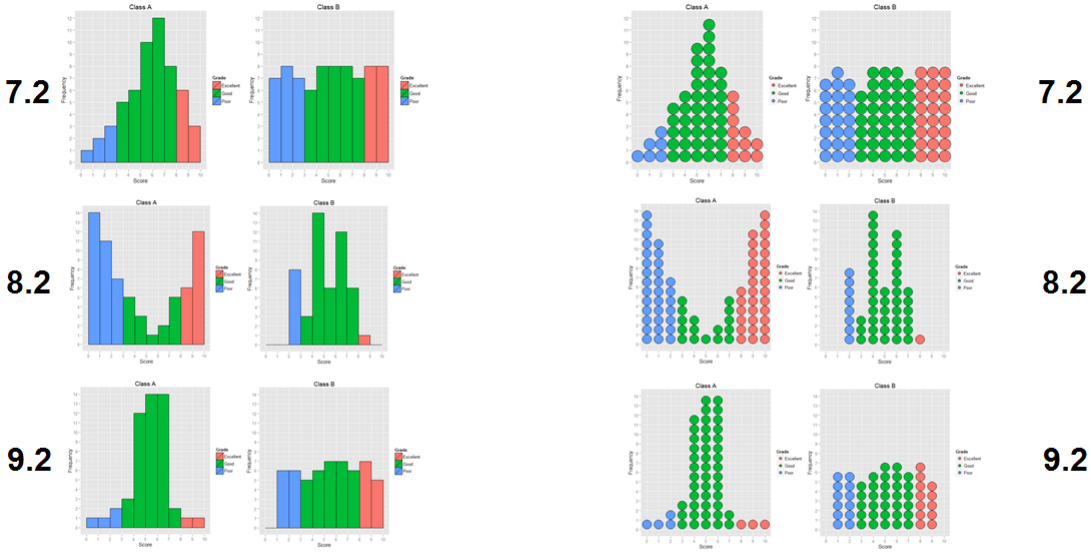


Figure 7.12: Pairings 7.2 through 9.2. Graphics are from Item 7.

7.2 Interview Tasks

This section details the final form of the 13 tasks that I used for the 19 interviews.

7.2.1 Task 1- Raw Data V1

Determine whether list A or list B contains data that are more variable (i.e. have more variability), or if both lists have data that are approximately equally variable. Describe why one list is more variable than the other or why they're both approximately equally variable.

A) $1, 3, 5, 7, 9, 11, 13$
B) $5, 6, 7, 8, 9, 10, 11$

Figure 7.13: Task 1

7.2.2 Task 2- Raw Data V2

Determine whether list C or list D contains data that are more variable (i.e. have more variability), or if both lists have data that are approximately equally variable. Describe why one list is more variable than the other or why they're both approximately equally variable.

C) $10, 10, 10, 20, 20, 20, 50, 50, 50$
D) $16, 18, 20, 22, 24, 26, 28$

Figure 7.14: Task 2

7.2.3 Task 3- Raw Data V3

Determine whether list E or list F contains data that are more variable (i.e. have more variability), or if both lists have data that are approximately equally variable. Describe why one list is more variable than the other or why they're both approximately equally variable.

E) $5, 12, 15, 21, 22, 23, 31, 36, 41$
F) $1, 20, 21, 22, 23, 24, 25, 26, 45$

Figure 7.15: Task 3

7.2.4 Task 4- Water Histogram

The histogram shows the number of cups of water 100 randomly surveyed respondents drink each day. Describe what the histogram tells you about how much water the respondents drink per day.

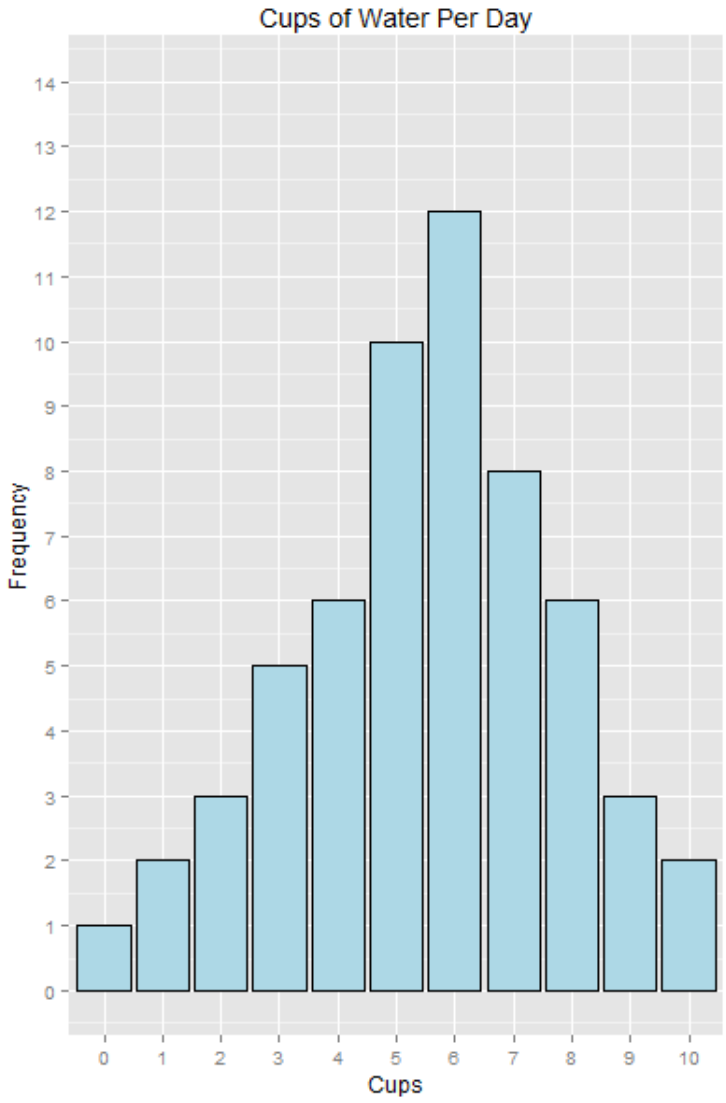


Figure 7.16: Task 4

7.2.5 Task 5- Coffee Histogram

The histogram below shows the number of ounces of coffee per day consumed by a randomly selected sample of 250 UGA college students. Describe what the histogram tells you about how much coffee UGA college students drink.

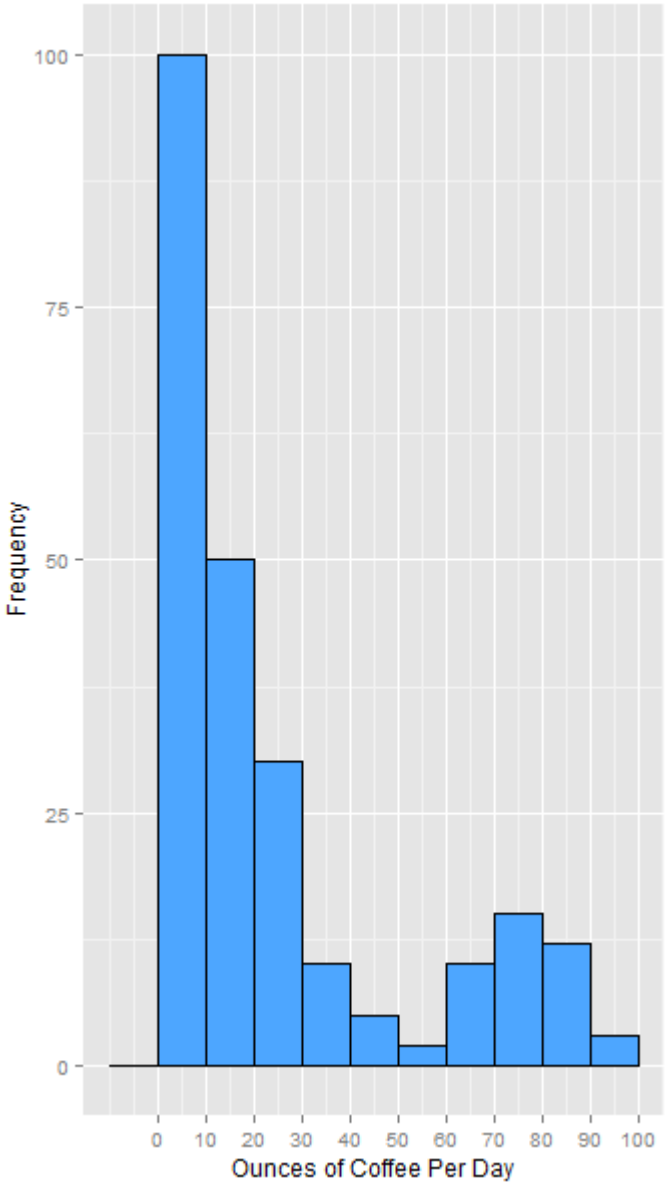


Figure 7.17: Task 5

7.2.6 Task 6- Exam Dot Plot

The dot plot below shows the exam scores for 58 students on a 10 question test. Describe what the dot plot tells you about how well students in this class did on the exam. (Interpret meaning of dot plot)

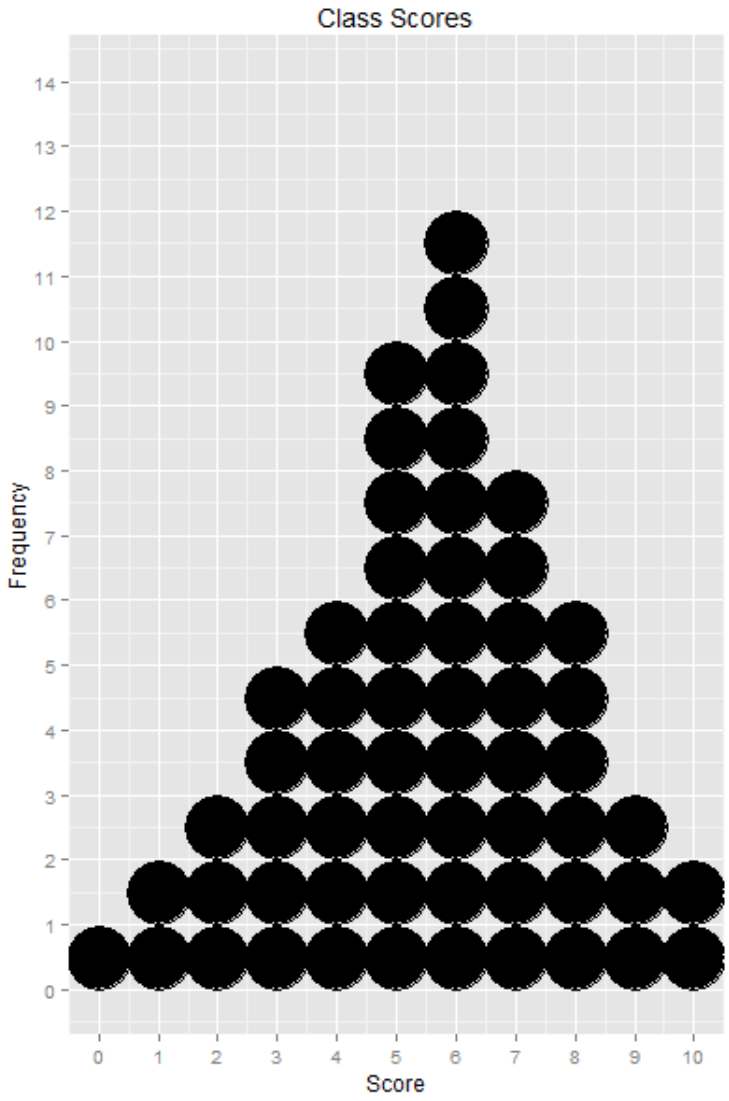


Figure 7.18: Task 6

7.2.7 Task 7- TV Dot Plot

The dot plot below shows the number of hours of TV watched by a randomly selected sample of UGA undergraduate students. Describe what the dot plot tells you about how much TV UGA undergraduate students watch.

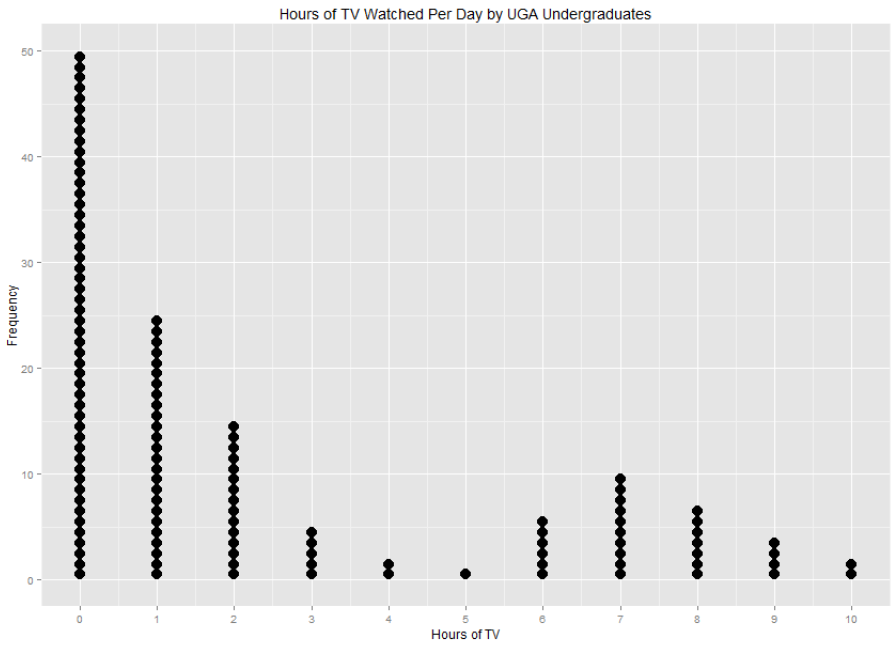


Figure 7.19: Task 7

7.2.8 Task 8- Colored Two Test Histogram

The two histograms below show test scores of two different classes on the same test. Which of the two classes had test scores that were more variable? How did you know they were more variable? Which class, A or B, would you rather take a test in? (Space out questions and allow time for full response to each)

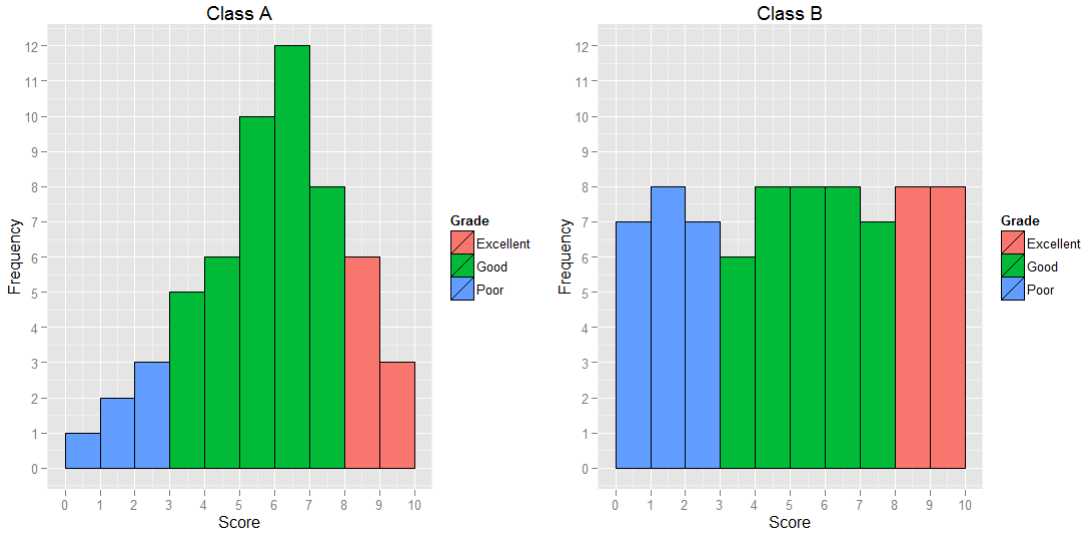


Figure 7.20: Task 8

7.2.9 Task 9- Blue Two Test Histogram

The two histograms below show test scores of two different classes on the same test. Which of the two classes had test scores that were more variable? How did you know they were more variable? Which class, A or B, would you rather take a test in? (Space out questions and allow time for full response to each)

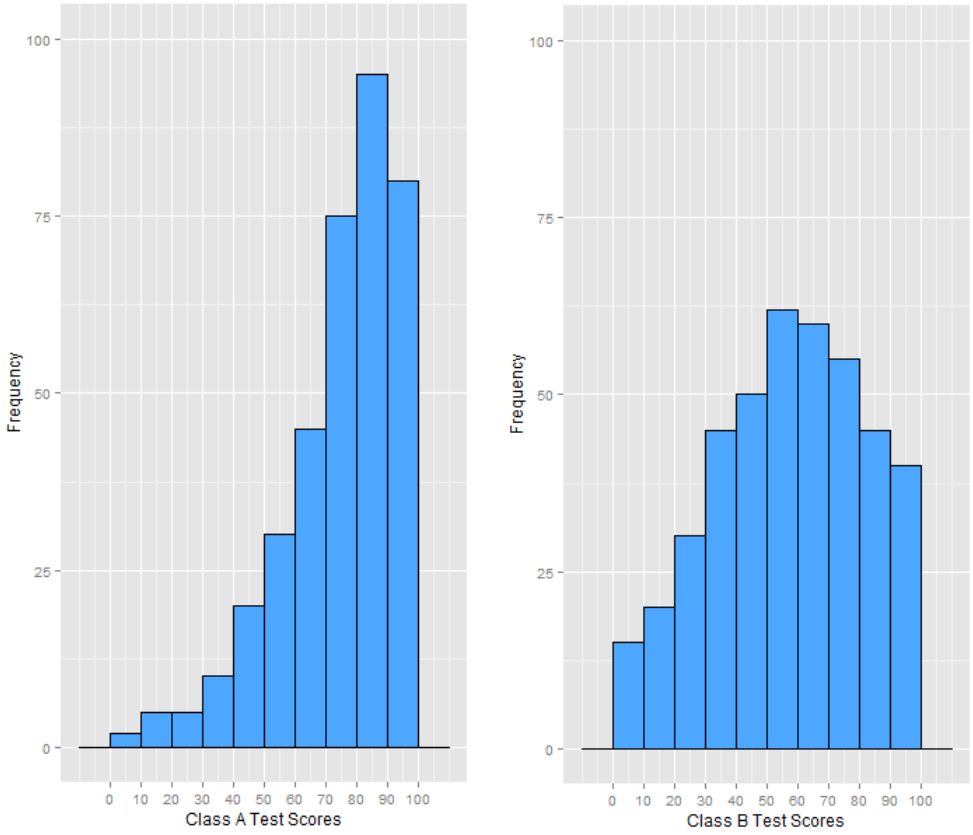


Figure 7.21: Task 9

7.2.10 Task 10- Olympics

The overlayed histograms below show Judges scores on a value event from 1960 and 2012. Describe what happened to judges scores between 1960 and 2012. (Filler question to determine how students with no exposure to overlayed histograms might interpret its display.)

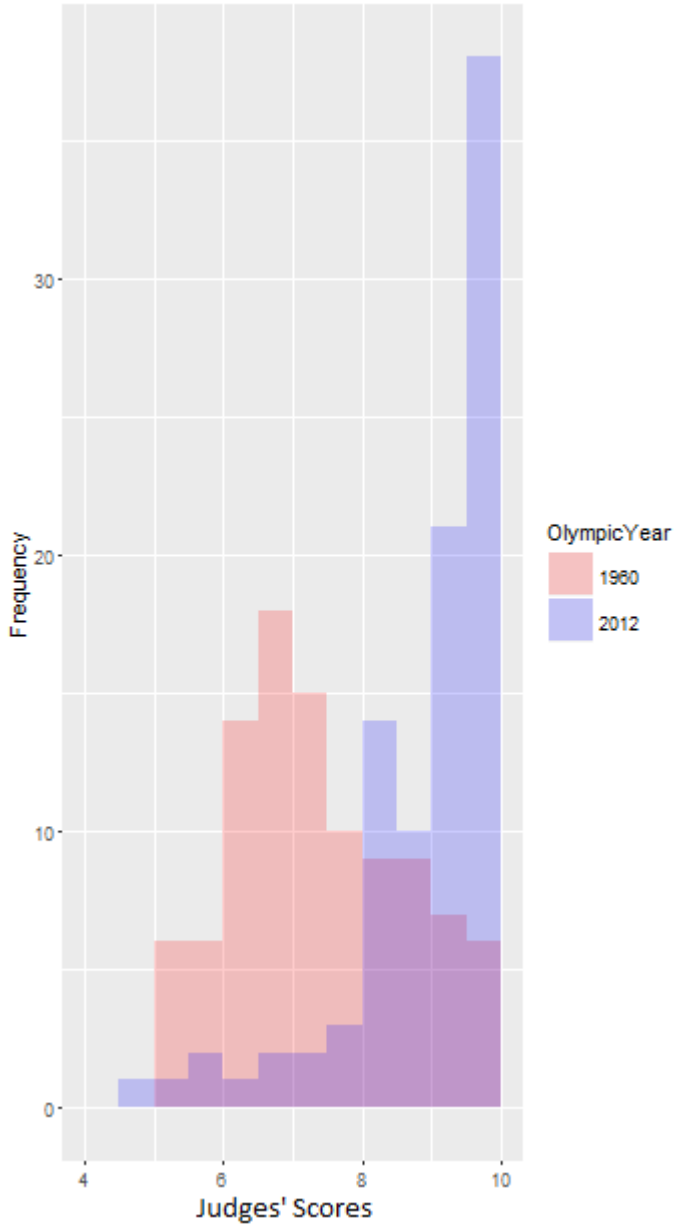


Figure 7.22: Task 10

7.2.11 Task 11- Advertising Histograms

The two histograms show how much money advertising companies A and B earned each of their many clients. Assuming they cost the same amount to hire, which of these two companies would you rather hire to advertise for your business? (Why?)

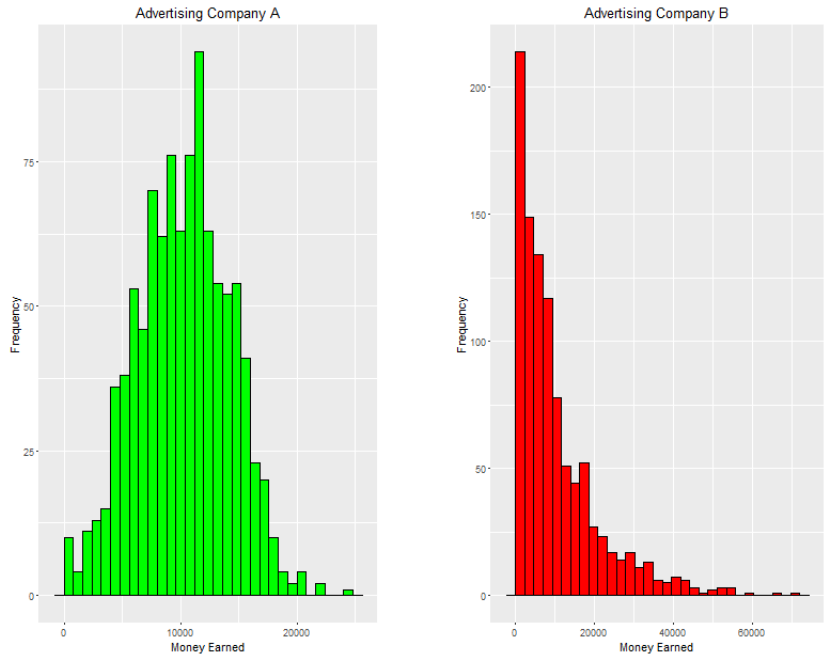


Figure 7.23: Task 11

7.2.12 Task 12- Colored Two Test Dot Plot

The two dot plots below show test scores of two different classes on the same test. Which of the two classes had test scores that were more variable? How did you know they were more variable? Which class, A or B, would you rather take a test in? (Space out questions and allow time for full response to each)

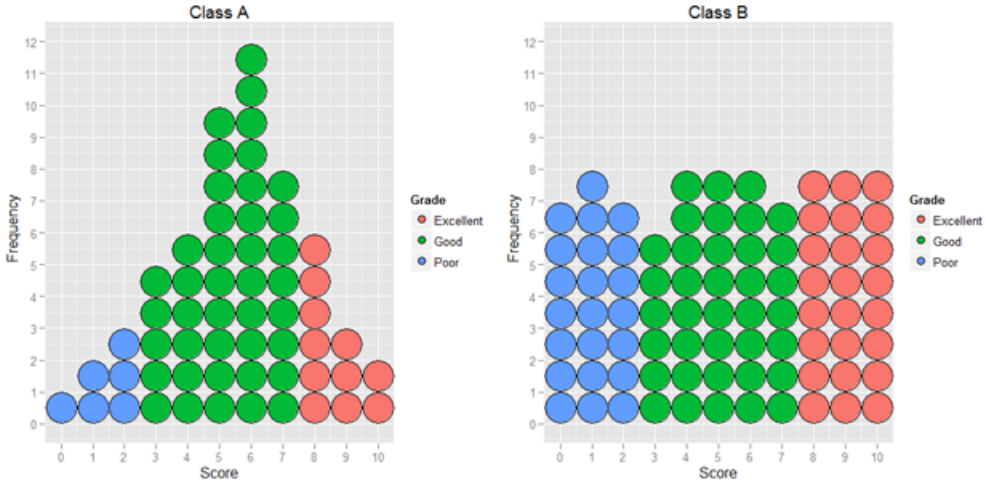


Figure 7.24: Task 12

7.2.13 Task 13- City/Country Route

A commuter records his travel time on two separate routes into Atlanta, the Highway Route and the Back Road Route. The histograms show the results of his trips. Just based on the data, which of these two routes would you rather use to take to work? Assume you only had 25 minutes to get to work and you could NOT be late. Which of these two routes would you rather take?

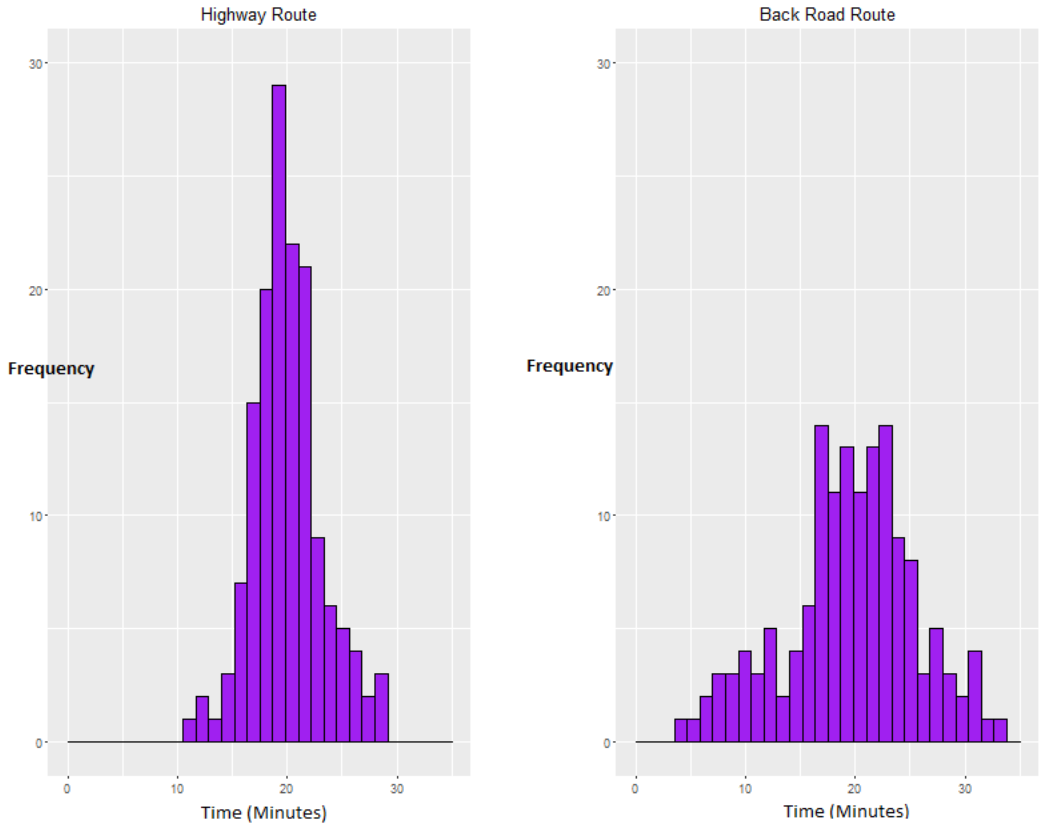


Figure 7.25: Task 13

Figure 7.26: Recruitment script for interviews

Recruitment Script

Dear STAT 2000 students:

I am a graduate student under the direction of Dr. Jennifer J. Kaplan in the Department Statistics at The University of Georgia. I invite you to participate in a research study entitled "Undergraduate Student Understanding of Variability in Graphical Representations of Univariate Data". The purpose of this study is to learn how undergraduate students answer questions about variability in various graphs such as histograms, dot plots, and bar charts. We obtained your contact information from your STAT 2000 course professor and/or computer lab teaching assistant.

In order to be eligible for this research, you must both be:

1) Currently enrolled in STAT 2000 at the University of Georgia and
2) 18 years or older

Your participation will involve being interviewed for up to 1 hour. Your involvement in the study is voluntary, and you may choose not to participate or to stop at any time without penalty or loss of benefits to which you are otherwise entitled. If you decide to withdraw from the study, the information that can be identified as yours will be kept as part of the study and may continue to be analyzed, unless you make a written request to remove, return, or destroy the information. There is no expected risk to you during this study; however your interview will be audio-recorded. The study may benefit you by allowing you to think more critically about the way you perceive variability in histograms, dot plots, and bar charts. Your participation in the study will help Statistics educators understand how students both correctly and incorrectly reason through statistical questions about variability in graphs, and which types of graphs make answering those questions easier or harder.

You will receive an incentive of 20 dollars at the conclusion of your interview.

If you are interested in participating in this study or have any questions about this research project, please feel free to call me at (678) 438-4375 or send an e-mail to ajlyford@uga.edu.

Thank you for your consideration!

Sincerely,

Alex Lyford

Figure 7.27: Consent form for interviews

Consent Letter

Dear STAT 2000 student:

I am a graduate student in the Department of Statistics at The University of Georgia working under the direction of Dr. Jennifer J. Kaplan . I invite you to participate in a research study entitled "Undergraduate Student Understanding of Variability in Graphical Representations of Univariate Data." The purpose of this study is to learn how undergraduate students answer questions about variability in various graphs such as histograms, dot plots, and bar charts. .

We are asking you to take part in a research study for which you must be at least 18 years old to participate. If you are younger than 18, please let us know.

Your participation will involve being interviewed for up to 1 hour. The interview will consist of several task-based statistics questions. The interview will be audio-recorded and transcribed using a pseudonym. Your involvement throughout the study is voluntary, and you may choose not to participate or to stop at any time without penalty. If you decide to withdraw from the study, the information that can be identified as yours will be kept as part of the study and may continue to be analyzed, unless you make a written request to remove, return, or destroy the information.

In order to maintain anonymity, you will be assigned a code and pseudonym. All publications about this project will use pseudonyms for the subjects. The assigned code will be generated from your name using a non-reversible algorithm. All data collected from you will be stored using this code. The only person who will know whether you gave consent to have your data used for this project is the person to whom you are giving consent.

There is no expected risk to you during this study, but your interview will be audio-recorded. In order to be used for the data analysis, the recordings will be later transcribed. I plan to keep the recordings for up to five years, and any written work and transcripts indefinitely. These data will be stored in a password-protected personal computer and on a portable hard drive.

The study may benefit you by allowing you to think more critically about the way you perceive variability in histograms, dot plots, and bar charts. Your participation in the study will help Statistics educators understand how students both correctly and incorrectly reason through statistical questions about variability in graphs, and which types of graphs make answering those questions easier or harder. You will receive an incentive of 20 dollars at the conclusion of your interview.

If you have any questions about this research project, please feel free to call me at (678) 438-4375 or send an e-mail to ajlyford@uga.edu. After I graduate in May 2017, please direct any questions or concerns to Dr. Jennifer J. Kaplan in the Department of Statistics (jkaplan@uga.edu). Questions or concerns about your rights as a research participant should be directed to The Chairperson, University of Georgia Institutional Review Board, 609 Boyd GSRC, Athens, Georgia 30602; telephone (706) 542-3199; email address irb@uga.edu.

Thank you for your consideration! Please keep this letter for your records.

Sincerely,

Alex Lyford

7.3 Definitions

This section defines a number of natural language processing and machine learning terms used frequently throughout this dissertation. More succinct definitions may also be found in the instance where each term first appears in its corresponding subsection. Some of these terms may have multiple meanings, but the definition provided defines the interpretation to be used henceforth.

- **Attribute**— Any of the n-grams appearing in the document-term matrix. Equivalent to feature.
- **Classifier**— A statistical algorithm which classifies an unlabeled student response into one of potentially many mutually exclusive categories.
- **Document**— Any individual student response to a single item.
- **Document-term matrix**— An nxm matrix, $\mathbf{M}$, where n represents the number of documents, and m represents the total number of unique n-grams used in any of the responses.
- **Feature**— Any of the n-grams appearing in the document-term matrix. Equivalent to attribute.
- **Feature Extraction**— Conversion of features (n-grams) into a quantitative measure of the number of instances of a given feature (Alpaydin, 2010).
- **Machine learning algorithm**— A class of statistical algorithms that can learn from and make predictions about data.
- **n-gram**— A phrase of n contiguous words. These are used as predictor variables in the machine learning algorithms.

- **Precision**— An algorithm evaluation metric. The proportion of documents that a particular machine learning algorithm placed into a given category that actually belong to that category.
- **Recall**— An algorithm evaluation metric. The proportion of documents that truly belong to a given category that are correctly classified by a particular machine learning algorithm to be in that category.
- **Rubric**— The document containing specific definitions and delineations for each category. An *analytic* rubric contains multiple binary categories that are not mutually exclusive (i.e. Did the student address Shape? Did the student address Variability?). A *holistic* rubric contains one set of several, mutually exclusive categories.