Donald Lee Beadle
MA 345
Statistics Final

Statistics is the collection and analysis of data. This data can be relevant to any field and help provide a better understanding of results. Statistics can be used to measure data that is only numerical based. Or, statistics can be used to measure data that is in categories or classes. This versatility allows it to be applicable from mathematics to marketing. Also, because statistics contains such a broad spectrum of applications, it is divided into descriptive and inferential statistics.

Descriptive statistics includes forming a means to represent some data in a clear manner. This can be used in the case of having some data and being able to describe the nature of it. Conversely, inferential statistics focusing on performing analysis on some sample of data to make a prediction about a larger population. The goals of both types of statistics can vary based on the necessary application. However, each type has its own associated goals. For example, a representation of data to show marketing growth over a number of years could be considered descriptive statistics. Meanwhile, making a hypothesis, designing an experiment, and analyzing the results is inferential statistics. Statistical experiments include observed units, variables, populations, and samples. Based on data, inferences are then made to draw possible conclusions.

Probability distributions are statistical representations that show the probability of some event occurring. There are many types of probability distributions that each have a specific purpose. The normal distribution is a symmetric distribution. This is also called the bell curve due to its appearance. In this type, the mean equals the median. Also, because of its shape, this type is readily identifiable. The next type of distribution is the binomial distribution. Binomial distributions are used with data that has “yes” or “no” or “positive” or “false” answers. Another distribution is the exponential distribution. In this type, the mean equals the standard deviation. This distribution can be associated with a rate which can be expressed through the steepness of a curve. However, it should be noted that there are many more types of distributions in statistics.

The Central Limit Theorem is essential in cases where a random sample size is large. From this theorem, the sampling distribution will approximate a normal distribution. Through repeated sampling, we observe that the mean of the samples will be approximate to the population mean.

Confidence intervals are a means of establishing estimations for some aspects on the data. In these cases, there is some unknown parameter that needs to be discovered. Through confidence intervals, we can estimate a value for that parameter. These parameters can be population mean or variance. There are two main types of confidence intervals to perform based on the sample size. If the sample is large, sample size is greater than 30, then a z statistic is used. If the sample size is relatively small, smaller than 30, a t statistic is used. Confidence intervals are valuable to inferential statistics for establishing a degree of certainty for values. The percent of error can range from different levels of confidence.

Hypothesis testing is another key aspect of inferential statistics. This testing allows for two hypotheses to be tested with a degree of certainty. There are two general hypotheses. The Null hypothesis and the alternative hypothesis. The Null hypothesis represents what currently exists. The alternative hypothesis represents what the experiment wants to prove. Hypothesis testing, like confidence intervals, is dependent on the size of samples with z statistics and t

statistics. Hypothesis testing can focus on single-sample testing or multi-sample testing. By performing hypothesis tests, one can reject or fail to reject certain hypotheses.

Another important component of statistics is ANOVA. ANOVA, or analysis of variance, focuses on designed experiments. This analysis helps determine certain changes in variables. These designed study variables include the response variable, factors, factor levels, treatments, and the experimental unit. ANOVA provides a means of generalizing testing to observe differences. The last component covered is linear regression. Regression provides a means to model values based on the equation of line. This line and the corresponding points around it form a relationship for analysis. The correlation component expresses how strongly the variables of the regression relate to one another. The deterministic component expresses a proportion.

Overall, this was a very informative and effective course. I am not usually a big fan of group work assignments but I enjoyed my discussion group. My main feedback for this course is related to RStudio. I really enjoyed learning how to work in the RStudio environment and perform the necessary computations in a software that has relevant real-world applications. During the course of the semester, I used RStudio to create PDFs for other classes as well as perform computations. In fact, I used RStudio to create a statistical report for my analysis of scheduling algorithms for my CS 420 Operating Systems class. We needed to analyze the means and variances (predictability) of each algorithm to determine an order of efficiency. I imported the text files into RStudio and was able to create a clean, informative report in the form of a PDF. My only recommendation to the course would be to have more R assignments early on in the semester. This is so that students can become more quickly immersed in the environment. I only add this because I know a couple of people in my group had trouble using it early on because they were not used to a terminal/console style environment.