POMONA COLLEGE

In Partial Fulfillment of the Requirement for a

Bachelor of Arts in Mathematics

Towards a More Conceptual Way of Understanding and Implementing Inferential Rules

Author:
Thalia Rodriguez

Advisor:
Dr. Jo Hardin

April 7, 2014

Contents

1 Introduction 5

2 Background: Using Computer Simulation Methods vs Traditional Methods for making inferences 9

2.1 Computer Simulation Methods 9

2.2 Traditional Methods 13

2.3 T-test 17

2.3.1 The Wilcoxon Rank-Sum Test 20

2.4 Summary 20

3 Using intuition to make inferences 21

3.1 Distribution of Samples 21

3.2 Making the call 26

4 Methods 29

4.1 Power 29

4.2 Normal and T Distribution 30

4.2.1 Normal Distribution 30

4.2.2 T-Distribution 31

5 Results 33

5.1 Normal Data 33

5.2 T-Distributed Data 34

5.3 Normal data with uniform outliers 35

5.4 Size 36

6 Discussion 39

Acknowledgements

I would like to express the deepest appreciation to my thesis advisor, Professor Jo Hardin, who is such an amazing person: she continually conveyed a spirit of adventure in regard to research and an excitement in regard to teaching. Without her guidance and persistent help, this thesis would not have been possible.

In addition, a thank you to my math professors, Amy Radunskaya, Vin de Silva, and Shahriar Shahriari, all of whom became the mentors that helped me achieve my academic endeavor of becoming a Mathematics major. I thank Kathy Sheldon who is responsible for getting everything (literally everything) done and for always being so friendly. I would also like to thank Pomona College and the entire Mathematics Department for such a wonderful experience.

Last but not least, I would like to thank my family: my parents Carolina Rodriguez and Ignacio Rodriguez for being such an inspiration in everything I do.

Chapter 1

Introduction

There is a growing movement, based on research in statistics education, indicating that students should begin working with precursor forms of statistical inference earlier than in standard curricula. In particular, proponents recommend using computer simulation methods (CSMs) to teach statistical concepts that tend to be very difficult or abstract. Proponents argue that this will give students more direct ways of answering real world questions. They claim that traditional methods of teaching introductory statistics are ineffective because they fail to establish a clear link between statistics and its applications in the real world. Another advantage to integrating computers in the classroom is that it allows students practice accomplishing computational tasks quickly and efficiently.

Basic concepts in statistics that are important to teach at an elementary level include data collection, data-analysis, and inference [Albert and Beliner, 1994]. Ideas of sample design and the design of experiments should be addressed in data collection. According to Albert and Beliner, the data-analysis portion of introductory statistics should include methods of summarizing data, describing relationships between variables, and graphics [Albert and Beliner, 1994]. In inference, the emphasis should be on statistical reasoning instead of technical details. Statistical reasoning includes distinctions between populations and samples, the variation of statistics from sample to sample, the correct interpretation of confidence intervals and hypothesis tests, and the role of the sample size in inference.

David S. Moore claims, "Statistics teaching should get students actively involved in working with data to discover principles for themselves" [Peterson, 1991]. Moore is a statistics professor at Purdue University and is also the author of a number of widely used statistics textbooks. His claim agrees with the underlying idea that learning statistics can be elucidated and made more engaging through computer-based techniques such as resampling. Resampling is the drawing of repeated samples from the given data. Different methods such as bootstrap and jackknife have been popular resampling methods. Peter C. Bruce and Julian Simon state that students who learn resampling are able to solve problems better than conventionally taught students [Lock et al., 2013].

The following is a review and critical examination of the literature related to the teaching and learning of statistics using computer simulation methods implemented in introductory statistics courses in the post-secondary arena. A discussion of the type of simulation method used, evidence of a theory of learning, and other comments specific to the article are also discussed.

Opponents of computer simulation methods are skeptical about the ways in which resampling is used to make inferences. Albert and Berliner claim that the key is that resampling results do produce estimates that are correct under the hypothesized chance model. However, if the hypothesized model is untenable, the resulting inferences are uninteresting and subject to misuse [Albert and Beliner, 1994]. Other general concerns about resampling and its role in inference arise from the claim that one can make inferences about a populations by reusing the sample many times.

Nicholas Horton and colleagues have created a structure that is parallel to hypothesis testing. In paticular, Horton et al. propose some specific and highly visual proposals for making inferences. They propose some specific milestones to guide the student's on 'making the call' that one set of values tend to be larger than another set of values by looking at some practical simulations, in this case, box plots [Horton et al., 2011]. Many considerations that students encounter when reasoning from the comparison of box plots include sample size, random variation in a sample, and interpretation of differences in spread. Horton and his colleagues also highlight four main ideas: (1) working with minimal set of the biggest ideas of statistical inference, (2) the mechanics of the inferential step are not demanding, (3) inferences should be able to be performed by looking at graphs, (4) methods should have connections to more formal methods to be used later [Horton et al., 2011]. The first idea addresses the belief that learners cannot successfully address too many issues simultaneously. This involves using simple ideas and representations that are already well understood by the vast majority of students in oder to attempt to make few conceptual connections that matter most. Thus, the basic approach of Horton et al. is too create a big picture so that over the years students can refine these intuitive rules in order to understand the more traditional ways of forming statistical inference.

Horton and his colleagues propose a convincing case for the role of computer-generated animations in teaching. However, no one has really questioned how compatible these recommendations are in comparison to more traditional calculus inference methods. Their driving question is: Can we actually make the differentiate two populations simply by viewing parallel boxplots?

Although there are many articles in the education literature that recommend using a computer basis to teach statistics better, there is very little research that has been published comparing methods, computer based and conventional, as to their accuracy for making inferential claims. In 2011, Horton and colleagues introduced new ways of considering sampling distributions by investigating distributions of boxplots (i.e. the distribution of the sample) instead of distributions of sample statistics. I will investigate the rules of Horton et al. associated with inferring that population A tends to have larger values than population B. This will be accomplished by comparing their rules to standard tests of center t-tests.

The following section will include the background on past literature which will cover the benefits of computational methods. It will be followed by some background on the traditional way of making inference by using sampling distributions of statistics. The method in which we will be investigating Horton et al's rules will be introduced in the subsequent section. Succeeding the methods will be the results, followed by a discussion. The conclusion will sum up the paper.

Chapter 2

Background: Using Computer Simulation Methods vs Traditional Methods for making inferences

2.1 Computer Simulation Methods

Technology has transformed both the ways in which statistics courses are taught and the way in which statistics is practiced. Previously, many assumptions were made in order to simplify statistical models and many problems that were hard to deal with and interpret analytically now have approximate solutions [Chance et al., 2007]. Simulation is a use of technology that can be a very powerful tool in helping students learn statistics [Chance et al., 2007]. Nonetheless, Hawkins [1990] cautioned, "Technology can enhance the processes of teaching and learning statistics. However, not all technology is fit for this purpose, and the use of this technology is not always appropriate." Technology-based teaching may be less than optimal because the software may be inadequate, the use of the technology may be inappropriate or the students may not experience what many predict they do. Although computer-based technology has increased the range of possible classroom activities, it can become a challenge for teachers who attempt to understand what each method has to offer and how their students will best understand [Hawkins, 1990]

There is a growing movement to introduce concepts of statistics and probability into the elementary and secondary school curriculum. For example the implementations of the Quantitative Literacy Project (QLP) of the American Statistical Association (ASA) is one indication of interest in this movement [Scheaffer, 1988]. Teaching and learning of statistics over the past few decades has changed by integrating the use of computers, especially in statistics postsecondary classrooms. As a result of computer development, there is more accessibility for students of user-friendly statistics packages such as SAS, SPSS, Excel, and MINITAB [Scheaffer, 1988]. Proponents of computer simulations argue that it allows students to accomplish computational tasks more quickly and efficiently so that they may focus more on statistical reasoning. This technology offers an end to tedious computations in data analysis, but it presents the possibility of a total lack of understanding for what is being done in the analysis, and the assumption that if the computer or calculator has done it then it must be right [Marasinghe et al., 1996]. Therefore, the use of technology must be meticulous in the ways in which it serves to enhance student learning and understanding in statistics.

Computer simulations methods (CSMs) allow students to experiment with random samples from a population with known parameters for the purpose of clarifying abstract and difficult concepts and theorems of statistics. For example, students are able to generate 50 random samples of size 30 from a non-normal distribution and compute the mean for each random sample. A histogram of the sample means can show the student that sampling distributions of the sample mean is the normal distribution. Indeed, computer simulations are valuable because abstract concepts can be easily illustrated.

Resampling methods emphasize intuitive reasoning and involve using a computer to perform experiments on the available data in order to get meaningful answers to statistical problems. As a result, students avoid having to deal with complicated formulas, cryptic tables and other forms of mathematics-sometimes black boxes to them- to get their results. A resampling experiment uses a mechanism such as a coin, a die, a deck of cards, or a computer's stock of random numbers to generate sets of data that represent samples taken from some population. Different methods such as bootstrap and jackknife have been popular resampling methods. The term jackknife was used to describe a technique that could be used to estimate bias and to obtain approximate confidence intervals [Duckworth and Stephenson, 2003]. Efron [1979] introduced the bootstrap as a general method for estimating the sampling distribution of a statistic based on the observed data. The Bootstrap Method's main idea is to use the empirical cumulative distribution function as a proxy for the true cumulative distribution function of the data. Critics of this method believe that you are trying to get something for nothing. However, they fail to realize that standard methods of inference also use only the one data set to make cliams about the distribution, including standard error of a statistic of interest.

The following is a review and critical examination of the literature related to the teaching and learning of statistics using computer simulations methods implemented in introductory statistics courses in the post-secondary arena. A discussion of the type of simulation method used, evidence of a theory of learning, and other comments specific to the article are also discussed.

Applets on the World Wide Web (WWW) are the latest Internet resources many educators are now using to illustrate statistics concepts. Ng and Wong [1999] reported using simulation experiments on the Internet to illustrate Central Limit theorem (CLT) concepts. The CLT can be demonstrated graphically online at a specific website. The program allows the user to choose a distribution from which the data are to be generated, a sample size for each sample, and the number of samples to be drawn [Ng and Wong, 1999]. The user can observe how fast the probability histogram approaches the normal curve as the sample size increases. It also allows the user to compare sampling distributions of other statistics including the median and the standard deviations.

The following introduces some of the first published incarnations of sampling in the classroom in which the technology is computationally light. Dallal [1990] proposed using simulation methods to illustrate the idea of inference and sampling error using MINITAB. Students can be given a population for which they can calculate the mean and standard deviation. From this population, each student generates a sample to find a 95 percent confidence interval for the population mean based on their sample mean. After, the class can determine how many of the sample confidence intervals actually contain the true population mean.

Simon and Bruce are responsible for some of the first software programs for resampling in the late 1980's (Resampling Stats Website). They developed a software package called Resampling Stats. The program allows the user to obtain numerical answers to statistical problems by writing short programs in a simple command language. For more information see the user guide Resampling: Probability and Statistics as Radically Different Way by Peter C. Bruce and a textbook to accompany Resampling Stats by Julian Simons [Simon and Bruce, 1991]. Simon and Bruce's view on the traditional teaching of probability and statistics revolves around the idea that conventional methods are not successful because they are taught through highly mathematical and difficult paradigms. This makes it frustrating for students to learn, and they become uninterested. The students hardly get insight, and they are unable to deal with practical problems. In addition, they choose particular statistical tests without understanding the underlying physical process of how the data were generated. Ultimately, they are not forced to think hard about the problem at hand. By using resampling statistics, students are able to fully understand what they are doing by learning from the step-by-step logical way of thinking about hypothesis testing. “People who use conventional methods can avoid this hard thinking by simply grabbing the formula for some test without understanding why they chose that test. But resampling pushes you to do this thinking explicitly” [Boomsma, 1990].

Although the overall consensus about computer simulations is that it appears to facilitate student understanding of difficult or abstract concepts, there is still some skepticism and criticism. Albert and Beliner argue that Simon and Bruce present a narrow aspect of statistics because they focus on probability and the inferential aspects of statistics [Albert and Beliner, 1994]. They believe that not enough is said about data analysis even when data-analysis techniques such as frequency tables and histograms are used in summarizing simulation reports. Other more general concerns arise from the claim that one can make inferences about a population by reusing the sample many times. Albert and Beliner make a harsh remark that a student who completed Simon’s book they would come to the conclusion “Statistics is easy: I can collect the data any way I like, conjure any resampling analysis I like (if I do not get the desired answer, try another statistic), and finally, I might as well use tiny samples because enough resampling allows inference to whatever population I decide I want to infer to” [Simon and Bruce, 1991]. This remark is very critical about Simon’s book but it is their honest reaction on what they believe it is accomplishing. Ultimately, Albert and Beliner do not recommend that book for a statistics course and claim that the approach to teaching statistics is not complete or valid. However, there has now been research to suggest when resampling techniques are appropriate which is not when sample sizes are too small. Resampling is indeed better understood. For example, modern textbooks have started using bootstrap confidence intervals as part of the cannon of introductory statistics material [Lock et al., 2013]

Despite the development of software programs, very little research has evaluated the effectiveness of simulation activities to improve student’s understanding about statistics [Chance et al., 2004]. In order to investigate the effectiveness of of simulation software on student’s understanding of sampling distribution, Chance et al. [2004] performed a study to document student learning on sampling distributions, while also providing feedback for further development and improvement of the software and learning activity. Four questions guided the investigation “how the simulations could be utilized more effectively, how to best integrate the technology into instruction, why particular techniques appeared to be more effective, and how students understanding of sampling distributions was affected by use of the program” [Chance et al., 2004]. Five sequential research studies were conducted in order to consecutively work on reconstructing the previous study. The series of studies revealed that several ways of using the software was not sufficient to affect meaningful change in students' misconceptions of sampling distributions [Chance et al., 2004]. The problem usually came when students were asked to distinguish between the distribution of one sample of data and the distribution of several sample means. Researchers found that having students make predictions about distributions of sample means drawn from different populations under different conditions such as sample size, and then asking them to use simulations to determine the accuracy of their predictions, improves the impact of technology on the student's reasoning [Chance et al., 2004].

Using many of the innovative ideas described above may be beneficial to students academically, but these methods must be evaluated more deeply, documented empirically, and rest on the foundation of a specific learning theory, especially if claims are made that student achievement is increased. Although proponents encouraged readers to use computer simulations and graphics to enhance students' understanding, it is not readily apparent that these methods offer a better instructional method than a more traditional approach. More empirically and theoretically grounded research must be done in order to determine whether these intuitive methods are compatible with the traditional methods.

2.2 Traditional Methods

The objective of statistics is to extract information from some data in order to be able to say something about a larger population. Traditionally we base inference on the sampling distribution of a statistic. Statistical inference is defined as making generalizations about a large population by using data from a sample. This is important because when you have some research data you want to be able to interpret what your result mean in terms of statistical significance. Making statistical inferences is usually done with p-values, hypothesis tests, and confidence intervals based on a sampling distribution. A sampling distribution is a theoretical distribution of the values that a specified statistic of a sample takes on in all of the possible samples of a specific size that can be made from a given population. A sampling distribution describes probabilities associated with a statistic when a random sample is drawn from a population. It is the probability distribution or probability density function of the statistic. Derivation of the sampling distribution is the first step in calculating a confidence interval or carrying out a hypothesis test for a parameter. A statistic is a quantity that is calculated from a sample of data. It is considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform a hypothesis test. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviors that would distinguish the null from the alternative hypothesis where such an alternative is prescribed, or that would characterize the null hypothesis if there is no explicitly stated alternative hypothesis. An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows p-values to be calculated. A statistic is important because it is used to give information about unknown values in the corresponding population. For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn. Hypothetically consider drawing more than one sample from the same population and the value of a statistic will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal.

Central Limit Theorem

The central limit theorem (CLT) states that given a population with a finite mean and a finite non-zero variance the sampling distribution of the mean approaches a normal distribution with a mean and a variance as N, the sample size, increases. The expressions for the mean and variance of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the sample mean approaches a normal distribution as the sample size increases.

The CLT describes the distribution of sample means of any population, no matter what shape, mean or standard deviation the population has. In many settings, the distribution of sample means approaches normality very rapidly. By the time the sample reaches 30 observations, the sampling distribution is often almost perfectly normal.

The shape of the sampling distributions tends to be normal. It will be normal if the population from which the samples are drawn are normal. The sample size (n) in each sample is relatively large, around 30 or more. The expected value of the statistic is the name given to the mean of the distribution of sample means (or of any other sample statistic).The expected value is equal to the population mean. The standard error is the name given to the standard deviation of the distribution of sample means. Typically, the smaller the population standard deviation, the smaller the standard error. The standard error is equal to the standard deviation of the population divided by the square root of the sample size. Typically, the larger the sample size, the smaller the standard error.

Significance Tests

Once sample data has been gathered through and observational study, statistical inference allows analysts to assess evidence in favor or so some claim about the population from which the sample has been drawn. The methods of inferences used to support or reject claims based on sample data are known as *tests of significance*

Every test of significance begins with a null hypothesis. The null hypothesis represents a theory that has been put forward either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the new drug.

The alternative hypothesis is a statement of what a statistical test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect on average, compared to that of the current drug.

The final conclusion once the test has been carried out is always give in terms of null hypothesis. We either “reject H_0 in favor of H_A” or “do not reject H_0”; we never conclude “accept H_0.”If we conclude ”do not reject the null hypothesis”, this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against the null in favor of the alternative; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Rejecting the *null* hypothesis when it is in fact true is called a *Type I error*. Since there are too many sources of error to be controlled, for example, sampling error, researcher bias, problems with reliability and validity, researchers can only quantify Type I errors. Many people decide on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha)

Statistical significance means that there is a good chance that we are right in finding that a relationship exists between two variables. But statistical significance is not the same as practical significance. We can have a statistically significant finding, but the implications of that finding may have no practical application. The researcher must always examine both the statistical and the practical significance of any research finding.

For example, a large clinical trial may be carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistical significant difference in lifespan when using the new treatment compared to the old one. But it increases less than 24 hours, and with poor quality of life during the period of extended life. Most people would not consider the improvement practically significant.

Steps in Testing for Statistical Significance

1. ***State the Research Hypothesis:*** A research hypothesis states the expected relationship between two variables

2. ***State the Null Hypothesis:*** A null hypothesis usually states that there is no relationship between the two variables. It also states the opposite relationship from the research hypothesis. Researchers use a null hypothesis in research because it is easier to disprove a null hypothesis than it is to prove a research hypothesis. The null hypothesis is the researcher's "straw man." That is, it is easier to show that something is false once than to show that something is always true. It is important to note that we never accept the null hypothesis.

3. ***Type I and Type II Errors***

   Even in the best research project, there is always a possibility that the researcher will make a mistake regarding the relationship between the two variables. There are two possible mistakes or errors. The first is called a Type I error. This occurs when the researcher claims that a relationship exists when in fact there is no relationships in the populations.

   In a *type I error,* the researcher should not reject the null hypothesis but the researcher ends up rejecting the null hypothesis. The probability of committing a Type I error is called α.

   The second is called a *type II error.* This occurs when the researcher claims that a relationship does not exist when in fact the relationship does exist in the populations. In a Type II error, the researcher should reject the null hypothesis and accept the research hypothesis, but the opposite occurs. The probability of committing a Type II error is noted as β.

   Generally, reducing the possibility of committing a Type I error increases the possibility of committing a Type II error and vice versa, reducing the possibility of committing a Type II error increases the possibility of committing a Type I error. Researchers generally try to minimize Type I errors, because when a researcher assumes a relationship exists when one really does not, things may be worse off than before. In Type II errors, the researcher misses an opportunity to confirm that a relationship exists, but is no worse off than before.

   In this example, which type of error would you prefer to commit?

   Research Hypothesis: The new drug is better at lowering blood pressure than the old drug

Null Hypothesis: The new drug is no better at lowering blood pressure than the old drug

If a *type I error* is committed, then drug is assumed to be effective when it really is not. People may be treated with the new drug, when they would have been better off with the old one. If a *type II error* is committed, then the new drug is assumed to be no better when it really is better (the null hypothesis should be rejected). People may not be treated with the new drug, although they would be better off than with the new one.

Probability of Error Level Researchers generally specify the probability of committing a Type I error that they are willing to accept, i.e., the value of alpha. In the social sciences, most researchers select an α =.05. This means that they are willing to make a type I error five percent of the time when the null hypothesis is true. In research involving public health, however, an alpha of .01 is not unusual. In general, researchers do not want to have a probability of a type I error bigger than .001 or one time in a thousand.

If the relationship between the two variables is strong (as assessed by a measure of association), and the level chosen for alpha is .05, then moderate or small sample sizes will detect it. For weaker relationships, however, and/or as the level of alpha gets smaller, larger sample sizes will be needed for the research to reach statistical significance.

2.3 T-test

Using T-Tests

Today's researchers have in their toolbox what is probably the most commonly performed statistical procedure, the *t-test.* T-tests are tests for statistical significance comparing two population means. T-tests can be used in several different settings. They can be used to test whether there are differences between two groups on the same variable, based on the mean (average) value of that variable for each group; for example, do students at private schools score higher on the SAT test than students at public schools? They can also be used to test whether a group's mean (average) value is greater or less than some standard; for example, is the average speed of cars on freeways in California higher than 65 mph? Lastly, they can be used to test whether the same group has different mean (average) scores on different variables; for example, are the same clerks more productive on IBM or Macintosh computers?

A standardized effect size, a test statistic (e.g., t and F scores) is computed by combining (unstandardized) effect and variation. In a t-test, for example, the standardized effect is effect (deviation of a sample mean from a hypothesized mean) divided by standard error. An effect size in actual units of responses is the degree to which the phenomenon exists (Cohen 1988).

Variation (variability) is the standard deviation or standard error of population. Cohen (1988) calls it the reliability of sample results. This variation often comes from previous research or pilot studies; otherwise, it needs to be estimated. Sample size (N) is the number of observations (cases) in a sample. The test size or significance level is the probability of rejecting the null hypothesis that is true. The power of the test is the probability of correctly rejecting a false null hypothesis.

To calcuate a value of t,

1. ***State the research hypothesis***; The new drug is better at blood pressure than the old drug

2. ***State the null hypothesis***; The new drug is no better at treating blood pressure than the old drug

3. ***Select the level of alpha***; such as p=.05, p=.01, p=.001

4. ***Stipulate whether the t-test will be a one-tailed test for significance***; Like other statistics, the t-test has a distribution that approaches the normal distribution as the sample size approaches infinity. Since we know the properties of the normal curve, we can use it to tell us how far away from the mean of the distribution our calculated t-score is. The normal curve is distributed about a mean of zero, with a standard deviation of one. A t-score can fall along the t-curve either above or below the mean; that is, either plus or minus some standard deviation units from the mean.

   A t-score must fall far from the mean in order to achieve statistical significance. That is, it must be quite different from the value of the mean of the distribution, something that has only a low probability of occurring by chance if there is no relationship between the two variables. If we have chosen a value of α =.05, we look for a value of t that falls into the extreme percent of the distribution.

   If we have a hypothesis that states the expected direction of the results, e.g., that the new drug is better at blood pressure than the old drug, then we expect the calculated t-score to fall into only one end of the t distribution. We expect the calculated t-score to fall into the extreme 5 percent of the distribution.

   If we have a hypothesis, however, that only states that there is some difference between two groups, but does not state which group is expected to have the higher score, then the calculated t-score can fall into either end of the t distribution. For example, our hypothesis could be that we expect to find a difference between the new and the old drug for blood pressure average (but we do not know which is going to be higher (better), or which is going to be lower, (less effective)).

For a hypothesis which states no direction, we need to use a “two-tailed” t-test. That is, we must look for a value of t that falls into either one of the extreme ends (“tails”) of the distribution. But since t can fall into either tail, if we select α =.05, we must divide the 5 percent into two parts of 2.5 percent each. So a two-tailed test requires t to take on a more extreme value to reach statistical significance than a one-tailed test of t.

5. ***Calculate t***

A t-score is calculated by comparing the average value on some variable obtained for two groups. Statistically, we represent these as $\overline{X}_1$ and $\overline{X}_2$. The calculation also involves the variance of each group, SD_1 and SD_2 and the number of observations in each group is represented as n_1 and n_2. The following is the formula used to calculate t:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

The null hypothesis here is that the means are equal, and the alternative hypothesis is that they are not. A big t, with a small p-value, means that the null hypothesis is discredited, and we would assert that the means are significantly different. However, a small t, with a big p-value indicates that they are not significantly different meaning they are significant with the null hypothesis.

Interpret the value of t: If the computed t-score equals or exceeds the value of t indicated in the table, then the researcher can conclude that there is a statistically significant probability that the relationship between the two variables exists and is not due to chance, and reject the null hypothesis. This lends support to the research hypothesis.

Remember, however, that this is only one statistic, based on just one sample, at one point in time, from one research project. It is not absolute, conclusive proof that a relationship exists, but rather support for the research hypothesis. It is only one piece of evidence that must be considered along with many other pieces of evidence on the same subject.

2.3.1 The Wilcoxon Rank-Sum Test

The Wilcoxon Rank-Sum (WRS) test is a nonparametric alternative to the two-sample t-test which is based solely on the order in which the observations from the two samples fall. Suppose that we have samples of observations from each of two populations A and B containing n_A and n_B observations respectively. We wish to test the hypothesis that the distributions of X-measurements in population Ais the same as that in B, which we will write symbolically as H_0: A=B. The departures from H_0 that the Wilcoxon test tries to detect are location shifts. If we expect to detect that the distribution of A is shifted to the right of distribution of B, we will write this as H_1: $A > B$. The other two possibilities are H_1: $A < B$ (A is shifted to the left of B), and the two sided-alternative, which we will write as H_1: $A \neq B$, for situations in which we have no strong prior reason for expecting a shift in a particular direction.

The Wilcoxon Rank-Sum test is based upon ranking the $n_A + n_B$ observations of the combined sample. Each observation has a **rank**: the smallest has rank 1, the 2nd smallest rank, and so on. The Wilcoxon rank-sum test statistic is the sum of the ranks for observations from one of the samples.

2.4 Summary

Tests for statistical significance are used because they constitute a common yardstick that can be understood by a great many people, and they communicate essential information about a research project that can be compared to the findings of other projects. They also serve to filter out unpromising hypotheses. However, they do not assure that the research has been carefully designed and executed. In fact, tests for statistical significance may be misleading, because they are precise numbers. But they have no relationship to the practical significance of the findings of the research.

Why does this matter? The objective of statistics is to extract information from data in order to be able to say something about a larger population. When we take samples or experiments of a number of people, we do not care about those people in the sample, we care about the population in general. That is what inference is all about. Traditionally, we use the methods presented above to be able to make inferences on sampling distributions of a statistics. However, supposing that there is not difference, calculate and interpret something like a p-value is not a mode of thinking that comes naturally to people. The following chapter will demonstrate why distributions of samples are more intuitive.

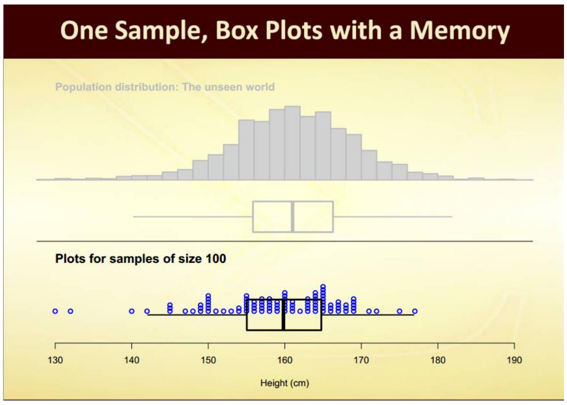
Chapter 3

Using intuition to make inferences

3.1 Distribution of Samples

The link between a sampling distribution and inference can be obscure and vague. Everyone is taught to narrow in on the "are they different?" question and to hypothesize "there is no difference". Statistical inference is backwards logic, you assume the thing you don't want and try to reject it. In 2011, Horton and colleagues introduced ways considering sampling distributions by investigating distributions of boxplots instead of distributions of sample statistics.

The goal of Horton et al., is to devise simpler versions of statistical inferences in order to equip students with a solid conceptual foundation on which to build formal inference [Horton et al., 2011]. They do so by providing students with simple inferential tools that they can use in order to quickly make an inference claim. Because Horton et al. believe that movement provides a powerful means of displaying the nature of variation they have introduced students to applets on line that illustrate the distributions of boxplots. The animations are impossible to convey in a static, print, thesis paper, the Web page can be found on http://www.censusatschool.org.nz/2009/informal-inference/WPRH/ which contains all the animations that will be described in this paper [Horton et al., 2011]. The following figures, Figure 3.1 - 3.4 are images of the animations on the website which can provide a better understanding of what I will be referring to.



One Sample, Box Plots with a Memory

Figure 3.1: One Sample of size 100

Figure 3.1 displays data on the heights of a sample of 100 girls aged 13 years taken from the 'Census At School' New Zealand database [Horton et al., 2011]. The motivation for this study is to compare the heights of girls aged 13 years with the height of girls aged 14 years where the sample of 13 year-old girls will look taller on average. Figure 3.1 combines a boxplot and dot plot of one sample of 14 year-old girls which allows the students to visualize an entire sample. The applet is set up so as the students takes more samples, the boxplots will vary which is a natural consequence of the variability in different samples.

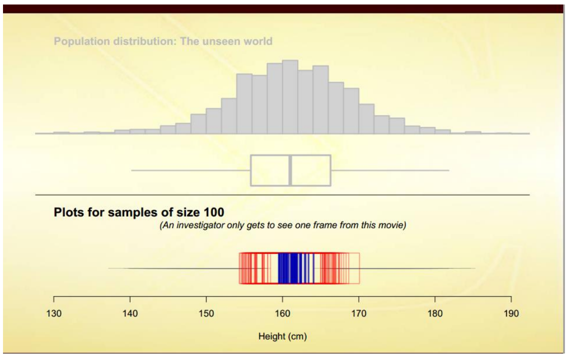


Figure 3.2: Boxplots with a memory over repeated sampling

In figure 3.2, Horton and colleagues display the population in the top half of the image in order to remind the students that the samples are being taken from a population. In the animations, all the boxplots that are seen over time leave behind ‘footprints’ with the most recent plot super imposed over the set of footprints. They use color to distinguish between the current and historical boxes and the median versus the rest of the box.The animations show clearly the variation in centers and spreads as they take new samples, and the effect of sample size on these features. Centers specify where the data lie by using either the mean or median. The spread is how far the data lie from the center in other works the range or difference from the max and the min. Figure 3.2 includes the footprint boxplots which is representative of a sampling distributions for the entire sample, as opposed to a sampling distribution for statistic. This is different from what students are used to seeing.

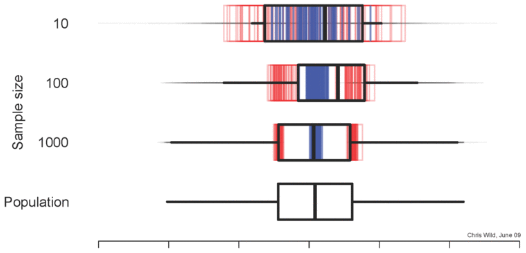


Figure 3.3: Sample Size Variation

Figure 3.3 also displays all the boxplots that are seen over time. The coloring of the red lines is to represent the footprints of the spread of the historical boxes. The blue lines represent the centers of the historical boxplots which are classified as the median. Additionally, the effect of sample size is visible in Figure 3.3 where, as the sample size increases, the variability of the box plots decreases. Visually, the students are able to see more spread out, less defined red lines and blue lines form smaller sample sizes. However, for larger sample sizes, the box plots contain compressed bold red and blue lines. Although this figure only illustrates sampling from one population, it allows the students to recall the varying box plots.

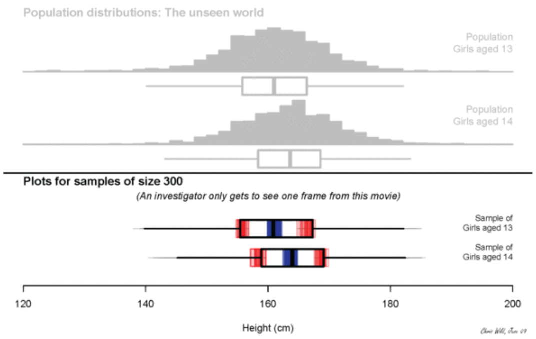


Figure 3.4: Comparing Two Populations

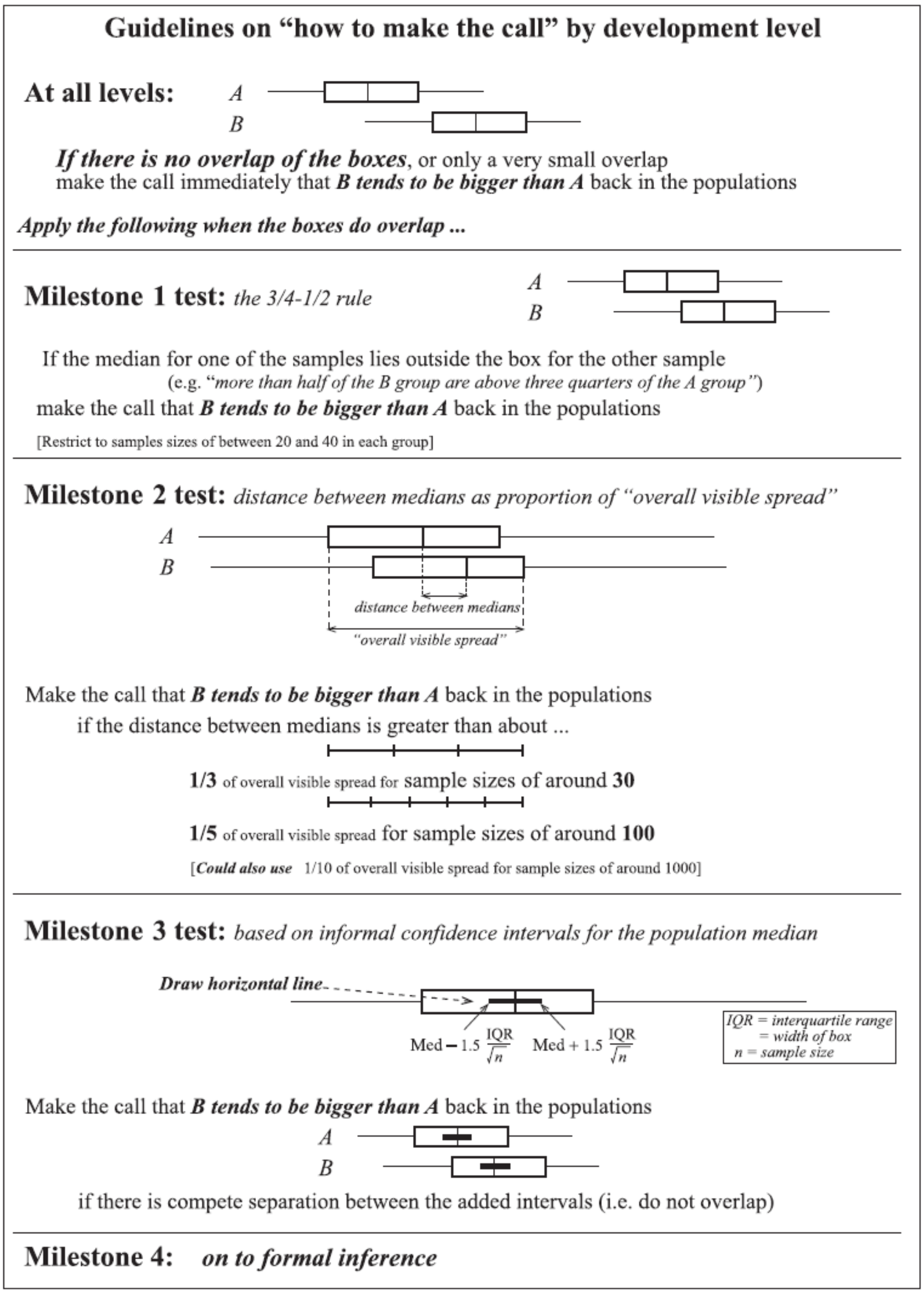
In Figure 3.4 they use the same vibrating box plots used in Figure 3.2 and Figure 3.3 to illustrate samples of size 300 of girls from two populations aged 13 and 14 years respectively. Again, Horton, et al. display the population in the top half of the image in order to remind the students that the samples are being taken from a population. Their motive for stacking the boxplots on top of each other is to compare the heights of girls aged 13 years with the height of girls aged 14 years where the sample of 14 year-old girls will look taller on average.

3.2 Making the call

The boxplots which represent sampling distribution of samples bring us back to an inferential question: Can students conclude from Figure 3.4 that 14 year old girls tend to be taller than 13 year old girls in the populations from which they sampled? The applets illustrate these image of sampling variation with vibrating boxplots but it might seem unclear as to how students can use these images to make inferential claims. The box plots do not have the central limit theorem to describe them like sampling distributions of average statistics, but Horton et. al did propose ad hoc rules for making inferences. The diagram displayed in Figure 3.5 communicates whether or not one can "Make the call that B's values tend to be larger than A's values back in the population(s)?"

The basic idea underlying Figure 3.5 is that one should only make the call if the location shift that they see between the two boxplots is sufficiently large to override the uncertainties that arise from natural variation. When one sees these rules they should think back to the applets and think about the vibrating boxplots and ask themselves how separate do the boxplots need to be in order to believe that B is bigger than A in general? One will have an easier time answering this question when they are focusing on large samples because the variability will be less. This means that large samples are less variable from sample to sample, but the width of the boxplot does not become more narrow. What changes in Horton et. al's proposal is that as students progress through the milestones there is a gradual refinement of how to determine whether an observed shift is sufficiently large to make the call. The overall objective of these rules is to allow students virtually "make the call" without taking their "eyes off the graph" [Horton et al., 2011].

The method of Horton et al. speaks to the intuition of inference based on samples versus based on a simple statistic. However, their method is ad hoc and because their method is not based on probabilistic derivations there is no obvious way to bound or even compute probabilities associated with type I and type II errors. Our particular interest, then, is whether their rules have comparable size and power as other standard tests of center.



Guidelines on "how to make the call" by development level

At all levels:

If there is no overlap of the boxes, or only a very small overlap
make the call immediately that ***B tends to be bigger than A*** back in the populations

Apply the following when the boxes do overlap ...

Milestone 1 test: *the 3/4-1/2 rule*

If the median for one of the samples lies outside the box for the other sample
(e.g. "*more than half of the B group are above three quarters of the A group*")
make the call that ***B tends to be bigger than A*** back in the populations

[Restrict to samples sizes of between 20 and 40 in each group]

Milestone 2 test: *distance between medians as proportion of "overall visible spread"*

Make the call that ***B tends to be bigger than A*** back in the populations
if the distance between medians is greater than about ...

1/3 of overall visible spread for sample sizes of around **30**

1/5 of overall visible spread for sample sizes of around **100**

[***Could also use*** 1/10 of overall visible spread for sample sizes of around 1000]

Milestone 3 test: *based on informal confidence intervals for the population median*

Draw horizontal line
$\text{Med} - 1.5\,\frac{\text{IQR}}{\sqrt{n}}$
$\text{Med} + 1.5\,\frac{\text{IQR}}{\sqrt{n}}$
IQR = interquartile range
= width of box
n = sample size

Make the call that ***B tends to be bigger than A*** back in the populations

if there is compete separation between the added intervals (i.e. do not overlap)

Milestone 4: ***on to formal inference***

Figure 3.5: How to make the call by level of development

Chapter 4

Methods

We investigate the ad hoc rules of Horton et al. associated with inferring that population B tends to have larger values than population A. In particular, we compare their rules to standard tests of center, t-tests and Wilcoxon Sum Tests. Our analysis take into account effect size, sample size, standard deviation, balance, population distribution, and amount of noise. Our motivation is to determine whether the proposed method actually controls size and has reasonable power.

4.1 Power

The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis. The statistical power is the ability of a test to detect an effect, if the effect actually exists [Cohen, 1988]. Cohen [1988] says, it is the probability that it will result in the conclusion that the phenomenon exists. A statistical power analysis is either retrospective (post hoc) or prospective (a priori). A prospective analysis is often used to determine a required sample size to achieve target statistical power, while a retrospective analysis computes the statistical power of a test given sample size and effect size. In this case we will be using a retrospective analysis.

Statistical power analysis explores relationships among following four components:

1. Standard effect size

2. Sample size (N)

3. Test size (significance level)

4. Power of the test

How are these components related each other? First, as a standardized effect size increases, statistical power increases (positive relationship), holding other components constant. A large standardized effect means that a statistic is substantially different from a hypothesized value; therefore, the test becomes more likely to detect the effect. Conversely, if the standardized effect size is small, it is difficult to detect effects even when they do exist; this test is less powerful. How does test size (significance level) affect statistical power? There is a trade-off between Type I error and Type II error . If a researcher has a significance level .10, the statistical power of a test becomes larger (positive relationship). Conversely, if we have a stringent significance level like .01, the Type II error will increase and the power of a test will decrease.

How does sample size N affect other components? When sample size is large, the samples are less variable from sample to sample. The variance of the statistic decreases when N increases but the variance of the sample does not.

In general, the most important component affecting statistical power is sample size in the sense that the most frequently asked question in practice is how many observations need to be collected. In our case we only differentiate between three sample sizes. There is a little room to change the size of the test (significance level) since conventional .05 or .01 levels are widely used. In this case we use an type I error rate of .05. It is difficult to control effect sizes in many cases. It is very time-consuming to get more observations, of course. However, if too many observations are used (or if a test is too powerful with a large sample size), even a trivial effect will be mistakenly detected as a significant one. Thus, virtually anything can be proved to come from different population or to have an effect regardless of actual effects [Cohen, 1988]. But if too few observations are used, a hypothesis test will be weak and less convincing.

4.2 Normal and T Distribution

In R, a free software environment for statistical computing and graphics, we simulated from normal and t-distributed data. We also added noise to normal data by adding large uniform outliers to see calculate the power of the milestones and t-test for that data. We ran samples with different standard deviation. We also used two different sample sizes, 10, 50, 100.

4.2.1 Normal Distribution

The normal (or Gaussian) distribution is a commonly used continuous probability distribution function that tells the probability that an observation in some context will fall between any two real numbers. For example, the distribution of grades on a test administered is often considered to be normally distributed.The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution: physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to the normal.

4.2.2 T-Distribution

In Section 2.3 we introduced a little about the calculating t and what it meant to have a large or small value of t. In this section I will simply give some historical background about the Student's T Distribution. The t distributions were discovered by William S. Gosset in 1908. Goesset was a statistician employed by the Guinness Brewing company which had stipulated that he not publish under his own name. Instead, he wrote under the pen name "Student" which is why it is often referred to Student's T distribution. These types of distribution arise in the following

Suppose we have a simple random sample of size n drawn from a Normal population with mean μ and standard deviation σ. Let $\overline{X}$ denote the sample mean sn s, the sample standard deviation. Then the following quantity has a t-distribution with n-1 degrees of freedom.

$$t = \frac{\overline{\chi} - \mu_0}{s/\sqrt{n}}$$

Note that there is a different t distribution for each sample size. When we speak of a specific t distribution, we have to specify the *degrees of freedom.* The degrees of degrees of freedom for this t statistic comes from the sample standard deviation s in the denominator of the equation.

The t density curves are symmetric and bell-shaped like the normal distribution and have their peak at 0. However, the spread is more than that of the standard normal distribution. This is due to the fact that in the above formula, the denominator is s rather than σ. Since$s$ is a random quantity varying with various samples, the variability in t is more, resulting in a larger spread.

The larger the degrees of freedom, the closer the t density is to the normal density. This reflects the fact that the standard deviation s approaches σ for large sample size n. In this particular case, we will be using a t-distribution with 2 degrees of freedom.

Chapter 5

Results

Our motive for creating these plots was to show which test has more power. As mentioned before, power is the probability of rejecting the null hypothesis when the alternative is true. When they mean difference is equal to 0, we have set the population so that the null hypothesis is true. We are forcing the two samples to come from the same populations but the goal here is whether or not these tests can tell that they came from different populations. The more the test is able to correctly tell that the samples come from different populations, the greater the power. In addition, as mean difference gets bigger, the probability of rejecting the null hypothesis goes up, meaning we reject the idea that the samples come from same populations.

5.1 Normal Data

The power plot shown in Figure 5.1 depicts our simulations for normal populations. The red color depicts the power from the t-test, the green color depicts the power of the Wilcoxon Rank-Sum test and the blue color depicts the power given the milestone set of rules. Power is measured on the y-axis and difference of means in the populations is measured on the x-axis. The solid lines with the circle across the lower half of the plot illustrate the simulations for sample size of 10. The dashed lines with a triangle which are concentrated in the top upper left-hand side of the plot are indicators of sample size 50. Lastly, the large dashed lines with squares represent a sample size of 100. The fact that the simulations for a sample size of 100 are along the top while the simulations for 10 are along the bottom portray how larger sample sizes increase the power of the test and makes it easier for them to reject the null hypothesis. The difference in big and small shapes is to depict the standard deviation; the smaller shape represent a sample standard deviation of five, while the larger shapes represents a sample standard deviation of two. samples.

Here the sample sizes did not make a difference.

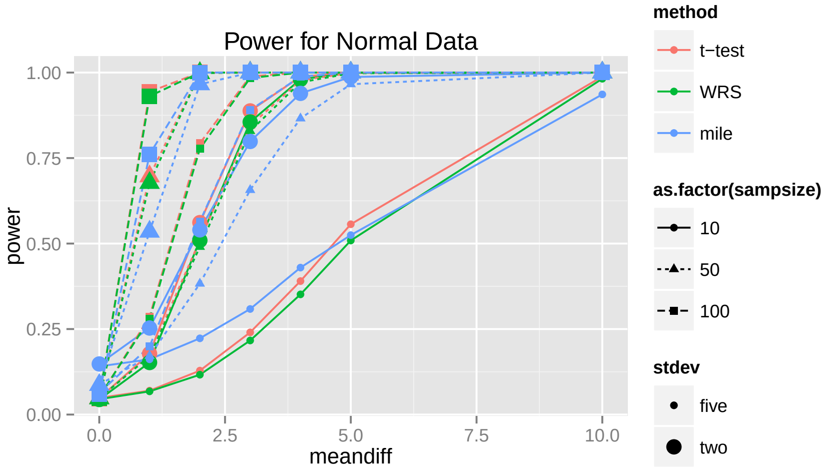


Figure 5.1: Normal Data

Figure 5.1 demonstrates that the t-test is more powerful when there is a small mean difference. Additionally, the Wilcoxon Rank-Sum test appears to be more powerful than the mile markers for small sample sizes. However, for large effect sizes, the milestones test is just as good as the standard test and the Wilcoxon test. It is important to note that the size of the mile markers is immensely bigger than the standard t-test and the Wilcoxon test. A table displaying the size of each test will be illustrated in section 5.4.

5.2 T-Distributed Data

Figure 5.2 displays the power plot for simulations using T-distributed data. The results are quite different from figure 5.1. For T-distributed data, the standard t-test had no power for a sample size of 10. That is, the t-test failed to reject the null hypothesis even when the alternative was true. The t-test gets more powerful for larger sample sizes, but clearly it is not the best test for this type of data. For a sample size of 10 and standard deviation of 5, the Wilcoxon Rank-Sum test and the mile markers appear to be equally powerful. Ultimately, as we increase the mean difference, the WRS test proves to be more powerful. The distinction between the power of the Wilcoxon Rank-Sum test and the mile markers remains the same for larger sample sizes.

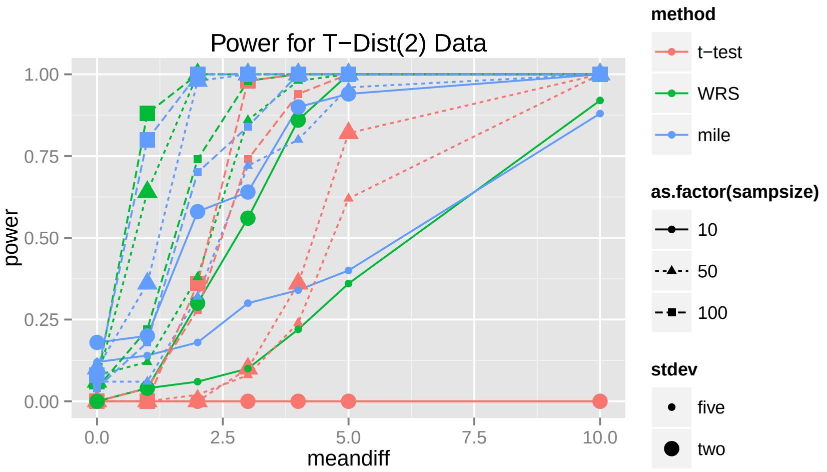


Figure 5.2: T-Distribution with Two Degrees of Freedom

5.3 Normal data with uniform outliers

Figure 5.3 displays the power plot for simulations using normal data with uniform outliers. Once again, the t-test failed to reject the null hypothesis even when the alternative was true. The t-test gets more powerful for larger sample sizes, but it is not the appropriate test for this type of data. When the sample size is 10, the mile markers appears to be more powerful than the WRS test. However, with sample size 10, the Wilcoxon Rank-Sum proves to be more powerful. It is also important to note the size of the test is different for the Wilcoxon Rank-Sum and mile markers. For large number of observations, it does not matter which test you use because they all appear to be equally powerful.

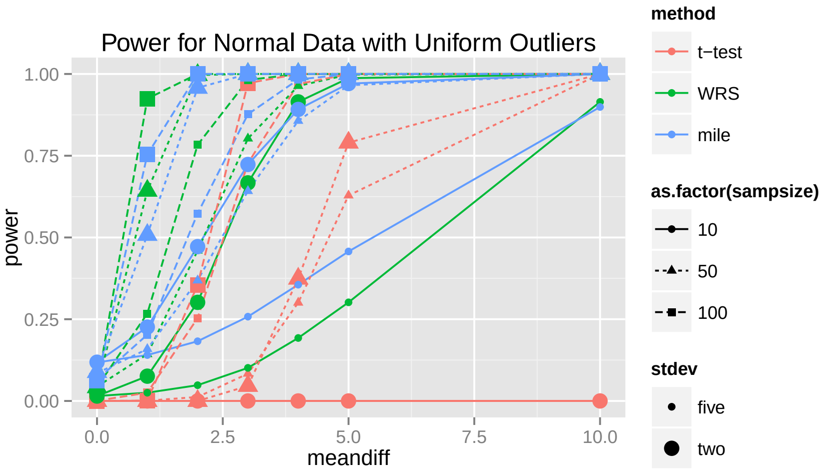


Figure 5.3: Normal Data with Uniform Outliers

5.4 Size

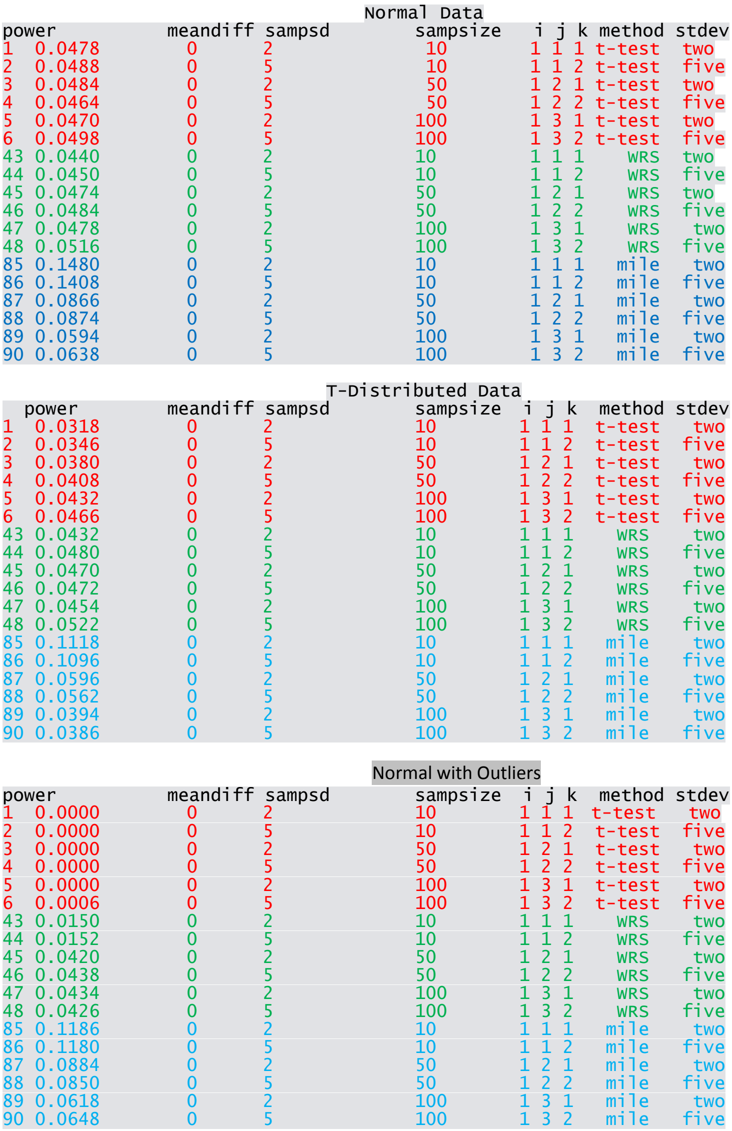
The *size of a test*, often called significance level, is the probability of committing a Type I error. A Type I error occurs when the null hypothesis is rejected when it is true. In a two-tailed test, the test size is the sum of two symmetric areas of both tails of a probability distributions. The areas surrounded by the probability distribution curve, x-axis, and a particular critical value, are called *rejection regions* in the sense that we reject the null hypothesis if a test statistic falls into these regions. As test size decrease, critical values shift to the extremes; the rejection areas become smaller; and thus, it is less likely to reject the null hypothesis.

The following page contains Figure 5.4 which illustrates the size of each test for different data. The size of each test can be determined by looking at each test when the mean difference is set to 0. This makes sense because at a mean difference of 0, each test must correctly fail to reject the null hypothesis since the alternative is not true. The red color represents the t-test, the green color is for the Wilcoxon Rank-Sum Test, and the blue represents the mile markers test. For normal data, the t-test has a size that ranges from 0.0478 - 0.0498. This is close to .05 which is the significance level/size that is typically used for t-tests. The Wilcoxon Rank-Sum test has a size that ranges from 0.0440 to 0.0516. This is also consistent with the traditional significance level of $\alpha = .05$. On the other hand, the mile markers test has a size that ranges from 0.0594 to 0.1480. This is a larger size than the size of the t-test and the Wilcoxon Rank Sum.

For T-distributed data, the size of the mile markers ranges from 0.0386 to 0.1118. Although the lower bound of this range (0.0386) is smaller than the one for normal data (0.0594), the size for the mile markers remains bigger than both tests. The size of the t-test for t-distributed data has range of 0.0318 to 0.0466. This is not consistent with the traditional size of $\alpha = 0.05$. We already knew that the t-test is strongest form normal distributed data from the Neyman-Pearson Lemma. Thus, it makes sense that it does not have the correct size. Similarly, the Wilcoxon Rank-Sum test has a size that ranges from 0.0432 to 0.0522. The Wilcoxon test does not have the correct size, but it is a bit more closer to the appropriate size.

For normal data with outliers, the size of the t-test is 0. This true even for larger sample sizes and different sample standard deviations. The size of the Wilcoxon Rank-Sum test ranges from 0.0150 to 0.0436. This is an extremely small size. Clearly, both the t-test and Wilcoxon test are not appropriate for this data. The size of the mile markers ranges from 0.0648 to 0.1186. The range of the mile markers for this data also appears to be smaller than it's size for normal and t-distributed data. Nonetheless, the size remains to be greater than the other two tests.

Evidently, this affects the results on the plots and may be a reason why the mile markers appear to be compatible in power with the t-test and the Wilcoxon Rank-Sum test. This result provides direction for future research.



```
                              Normal Data
power              meandiff sampsd        sampsize  i j k method stdev
1  0.0478           0        2              10      1 1 1 t-test  two
2  0.0488           0        5              10      1 1 2 t-test five
3  0.0484           0        2              50      1 2 1 t-test  two
4  0.0464           0        5              50      1 2 2 t-test five
5  0.0470           0        2             100      1 3 1 t-test  two
6  0.0498           0        5             100      1 3 2 t-test five
43 0.0440           0        2             10       1 1 1    WRS  two
44 0.0450           0        5             10       1 1 2    WRS five
45 0.0474           0        2             50       1 2 1    WRS  two
46 0.0484           0        5             50       1 2 2    WRS five
47 0.0478           0        2             100      1 3 1    WRS   two
48 0.0516           0        5             100      1 3 2    WRS  five
85 0.1480           0        2             10       1 1 1   mile   two
86 0.1408           0        5             10       1 1 2   mile  five
87 0.0866           0        2             50       1 2 1   mile   two
88 0.0874           0        5             50       1 2 2   mile  five
89 0.0594           0        2             100      1 3 1   mile   two
90 0.0638           0        5             100      1 3 2   mile  five
```

```
                          T-Distributed Data
  power            meandiff sampsd        sampsize i j k  method stdev
1  0.0318           0        2             10      1 1 1  t-test   two
2  0.0346           0        5             10      1 1 2  t-test  five
3  0.0380           0        2             50      1 2 1  t-test   two
4  0.0408           0        5             50      1 2 2  t-test  five
5  0.0432           0        2             100     1 3 1  t-test   two
6  0.0466           0        5             100     1 3 2  t-test  five
43 0.0432           0        2             10      1 1 1     WRS   two
44 0.0480           0        5             10      1 1 2     WRS  five
45 0.0470           0        2             50      1 2 1     WRS   two
46 0.0472           0        5             50      1 2 2     WRS  five
47 0.0454           0        2             100     1 3 1     WRS   two
48 0.0522           0        5             100     1 3 2     WRS  five
85 0.1118           0        2             10      1 1 1    mile   two
86 0.1096           0        5             10      1 1 2    mile  five
87 0.0596           0        2             50      1 2 1    mile   two
88 0.0562           0        5             50      1 2 2    mile  five
89 0.0394           0        2             100     1 3 1    mile   two
90 0.0386           0        5             100     1 3 2    mile  five
```

```
                               Normal with Outliers
power              meandiff sampsd        sampsize i j k  method stdev
1  0.0000           0        2             10      1 1 1 t-test    two
2  0.0000           0        5             10      1 1 2  t-test  five
3  0.0000           0        2             50      1 2 1  t-test   two
4  0.0000           0        5             50      1 2 2  t-test  five
5  0.0000           0        2             100     1 3 1  t-test   two
6  0.0006           0        5             100     1 3 2  t-test  five
43 0.0150           0        2             10      1 1 1     WRS   two
44 0.0152           0        5             10      1 1 2     WRS  five
45 0.0420           0        2             50      1 2 1     WRS   two
46 0.0438           0        5             50      1 2 2     WRS  five
47 0.0434           0        2             100     1 3 1     WRS   two
48 0.0426           0        5             100     1 3 2     WRS  five
85 0.1186           0        2             10      1 1 1    mile   two
86 0.1180           0        5             10      1 1 2    mile  five
87 0.0884           0        2             50      1 2 1    mile   two
88 0.0850           0        5             50      1 2 2    mile  five
89 0.0618           0        2             100     1 3 1    mile   two
90 0.0648           0        5             100     1 3 2    mile  five
```

Figure 5.4: Size for All Data

Chapter 6

Discussion

Using computer simulation methods in statistics can be beneficial to students by helping them better understand the underlying logic of the techniques. Traditional methods used in statistics involve more calculations and disengage students from the true objective of statistics. Since the logic of statistical inference does not come naturally to many, Horton et. al. came up with these novel ways of making statistical inference more meaningful to students. However, their rules are ad hoc and have not been assessed or compared to traditional methods. In and attempt to investigate Horton's rule we developed some interesting results.

The milestones test proved to be reasonably powerful when compared to the t-test and the Wilcoxon-Rank Sum test. In particular, for t-distributed data and normal data with added noise, the mile markers was slightly less powerful than the Wilcoxon Rank-Sum test. This is due primarily to the fact that the milestones are more robust to outliers than the t-test.

In general, the milemarkers are not as powerful as the t-test or the Wilcoxon Rank-Sum test. It is clear that for larger sample sizes, the mile markers proves to be less powerful than the other two tests. Although it is less powerful, it does not fall behind by orders of magnitude. The results showed that this test is slightly less powerful and has a reasonable size of about $\alpha = 0.10$. This means that the mile markers reject the null hypothesis 10 percent of the time. Additionally, the mile marker method proves to be more intuitive than more traditional methods. Our goal was not to prove that this method would be more powerful than any of two tests, we simply wanted to compare its power to standards center of tests.

Future directions for research on this topic would incorporate size in order to be able to compare all the three test appropriately. The mile markers had a bigger size which was greater than .05. Since the size of the t-test and the Wilcoxon Rank-Sum test is 0.05, we need to find an α such as 0.10 that can be used to compare the three tests at a level that is the same for all tests.

We hesitate to say that the mile markers are comparably powerful because it does not have the right size. Here, our effect size was fairly small. However, we know that a large standardized effect size means that a sample statistic is substantially different from a hypothesized value; therefore, it becomes more likely to detect the effect. Conversely, if the standardized effect size is small, it is difficult to detect effects even when they exist; this test is less powerful.

Bibliography

Jim Albert and Mark Beliner. Review of the resampling method of teaching statistics. *The American Statistician*, 48:129–131, 1994.

Anne Boomsma. Resampling with more care. *Chance*, 4:25–29, 1990.

B. Chance, R. delMas, and Garfield J. *Reasoning about sampling distributions.* The challenge of developing statistical literacy, reasoning, and thinking, 2004.

B. Chance, D. Ben-Zvi, Garfield J., and E. Medina. The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, pages 1–26, 2007.

J. Cohen. Statistical power analysis for the behavioral sciences. pages 1–17, 1988.

Gerard E. Dallal. Statistical computing packages: Dare we abandon their teaching to others? *The American Statistician*, 44:265–266, 1990.

Wiliam M. Duckworth and W. Robert Stephenson. Resampling methods: Not just for statisticians anymore. *Iowa State University Department of Statistics*, pages 1–6, 2003.

A. Hawkins. Training teachers to teach statistics. *Voorbutg: International Statistical Institute*, 1990.

N. J. Horton, M. Pfannkich, M. Regan, and N. J. HortonC.J. Wild. Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society*, 174:247–295, 2011.

Robin Lock, Patti Frazer Lock, Kari Lock Morgan, Eric Lock, and Dennis Lock. *Statistics: Unlocking the Power of Data.* John Wiley & Sons, Inc., 2013.

M. G. Marasinghe, W. Q. Meeker, D. Cook, and T Shin. Using graphics and simulation to teach statistical concepts. *The American Statistician*, 1996.

V. M. Ng and K.Y. Wong. Using simulation on the internet to teach statistics. *The Mathematics Teacher*, 1999.

Ivars Peterson. Pick a sample. *Science News*, 140:56–58, 1991.

R. Scheaffer. Statistics in the schools: The past, present and future of the quantitive literacy project. *American Statistical Association Proceedings of the Section on Statistical Education*, 1988.

Julian L. Simon and Peter C. Bruce. Resampling: A tool for everyday statistical work. *Chance*, 4:22–32, 1991.