



SENIOR THESIS IN MATHEMATICS

**A Critical Comparison of Methods in
Statistical Inference Education**

Author:

Rebecca Baiman

Advisor:

Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 10, 2015

Abstract

According to the College Board Program Summary Report 184, 173 students took the AP Statistics test in 2014, an increase of 9% from the previous year [4]. As the number of students studying statistics in the United States increases, innovations emerge in statistics education. One such innovation was a paper published in 2011 by Wild et al. out of Auckland University entitled “Towards More Accessible Conception of Statistical Inference”. The Wild et al. paper addressed the issue of student comprehension in statistical inference education by introducing new educational methods for teaching sampling and statistical inference. In contrast to the tradition t-test method of statistical inference education, the box plot statistical inference method developed by Wild et al. employs visual representation of the distribution of samples and the predicted variability of a sample to make a generalization about a population. The Wild et al. method of teaching statistical inference is innovative and visual. The question remains as to whether or not the new statistical inference curriculum is superior to the tradition curriculum. The following paper investigates the type I error rates and the intuitiveness of Wild et al.’s box plot based statistical inference pedagogy and compares the innovation of Wild et al. with the tradition t-test pedagogy.

Contents

1	Background	1
1.1	Statistical Inference	1
1.2	The T-test	2
1.3	The T-test in History	4
1.4	The Milestone Method	4
2	Type I Errors	9
2.1	T-test	9
2.2	Milestone Method	9
3	Intuition	14
3.1	T-test	14
3.2	Milestone Method	16
4	Future Research and Conclusion	20

Chapter 1

Background

1.1 Statistical Inference

Statistical Inference is broadly defined as making a generalization about a population based on data from a representative sample. Traditional statistical inference methods use the distribution of one statistic over many samples to make inferential claims. In particular, the following paper will focus on statistical inference used to analyze two populations in order to make the claim that those populations are different. One illustrative example is the testing of a new prescription drug by the FDA. Such a study samples one population taking the drug and another population taking a placebo pill. The FDA uses the data to make a claim about the effectiveness of the drug: whether the population using the drug has different average effectiveness than the population not using the drug.

Two hypotheses are defined when testing whether two populations are the same or different. The **null hypothesis** states that the populations are not different and the **alternative hypothesis** states that the populations are different. Statistical Inference is the tool used to reject the null hypothesis.

Based on the two hypotheses, statisticians define two types of errors and the rates at which they occur. A **type I error** is made when a true null hypothesis is rejected; α is the probability of a type I error. In other words, α is the probability that a statistical inference method will claim that two populations are different when they are the same. A **type II error** is made when a true alternative hypothesis is not accepted; β is the probability of a type II error. β is the probability that the statistical inference method will be unable to differentiate two populations when they are in fact

different.

In the example of testing the effectiveness of a new drug, an incorrect rejection of the null hypothesis (claiming a drug is effective when it is not) should be ranked as less desirable than incorrectly failing to reject the alternative hypothesis (having no evidence to claim a drug is effective when it really is). In other words, it is worse to put a useless, and possibly harmful, drug on the market than it is to deem the drug not significantly effective and continue to try to improve it. Statistic educators often explain the preference of an incorrect rejection of the null as the difference between incorrectly shifting the status quo and incorrectly maintaining the status quo.

Along with the value judgement described above, β is often not computable and is never controllable. Because β is difficult to predict, impossible to control, and less problematic based on the value judgement described above, statisticians typically choose to control and strictly limit α . When deciding on a test that will infer a generalization about a population, it is standard to use a test that has a set α level. It is widely accepted to use an α of **.05**. A **.05** α value implies that a statistical inference method incorrectly claims that an ineffective drug is effective with a probability of only **.05**. Tests that maintain defined α values are called **α level tests**.

In a seminal paper, R.A. Fisher describes the use of **.05** as α . He writes, “It is common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials” [7]. Fisher goes on to explain that an α of **.05** does not mean that a statistician should be wrong one out of twenty times. Instead, α indicates to a statistician when to ignore a test. When observed data could have been observed randomly more often than 1 in 20 trials, it should be set aside [7].

Once α levels are set, the quality of a statistical inference method testing whether two populations are different or the same can be assessed by looking at the rate of type II errors. Other criterion such as the power of the test and the intuitiveness of the test all influence the quality of the test as an educational tool. When focusing on statistical inference education, it is imperative to consider the intuitiveness of a hypothesis test as a measure of its quality as an educational tool.

1.2 The T-test

The **t-test** is one of the most common statistical inference methods taught at an introductory level of statistics. Given the assumption of random samples and either big sample sizes or normal observations, the t-test is an α level test. An α level test means one can choose and set the type I

error rate and the t-test will maintain that α . A standard t-test assesses the claim that the expected value in one population is different from the expected value in another. In a standard t-test the null hypothesis is that the expected values of the two populations are equal. To compare two populations, one calculates the t-score (t) which is the difference in sample means divided by standard error of the difference in sample means.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}} \quad (1.1)$$

Where \bar{x}_i is the mean of sample i , SD_i is the standard deviations of sample i , and n_i is the size of sample i .

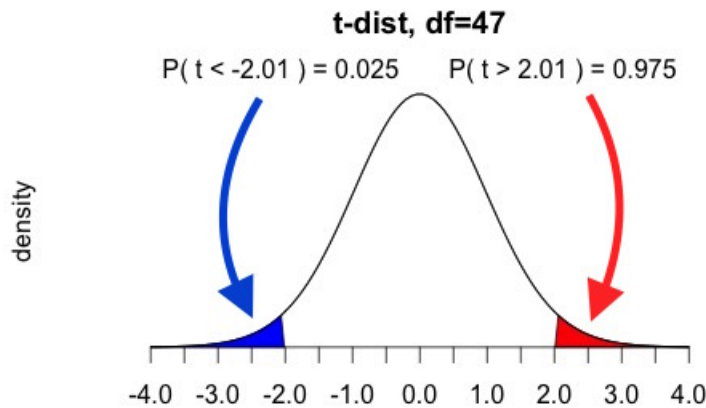


Figure 1.1: Density of t with 47 degrees of freedom

If the null hypothesis is true, as the sample size n goes to infinity the distribution of t approaches the standard normal distribution as a consequence of the central limit theorem. If the null hypothesis is true (the two populations are the same) the expectation of t is zero. If t falls close to the mean of the distribution, zero, then the two populations are not significantly different. The t-test maintains $\alpha = .05$ by defining the maximum distance from zero where t is significant under the null hypothesis. Figure 1.1 shows the density of t with 47 degrees of freedom. The blue and red sections delineate the **.025** and the **.975** quantiles. If the null is true and the two populations are identical, the probability that the t will lie outside the **.025** or **.975** quantiles of the t-distribution, represented by the blue and red areas in Figure 1.1, is **.05**. Thus the t-test defines the values of t that support

the null hypothesis as the middle **.95** of the t-distribution, the white area under the curve in figure 1.1. Alternatively, if the two populations are different and the alternative is true t should be further from zero, outside the range for t that would likely occur if the populations are identical. Therefore the t-test rejects the null hypothesis if t falls outside the **.05** or **.95** quantiles of the t-distribution.

A t-table is used to find the critical t value based on the degrees of freedom and the set α . If t is larger in magnitude than the critical t value then the null hypothesis is rejected. If the t-score calculated is less in magnitude than the critical t value, the test supports the null.

1.3 The T-test in History

The t-test was developed in 1908 by William Sealy Gosset and published under his pen name Student, sparking the full name for the test: Student's t-test [6]. For the statistical inference addressed here, the t-test is the hegemonic method taught in the classroom. In addition to its prevalence in education, the t-test is utilized in much statistical analysis. For example, **26%** of original articles in The New England Journal of Medicine from 2004 to 2005 used the t-test [8]. The prevalence of the t-test as an educational tool has not been widely questioned.

1.4 The Milestone Method

In 2011, C.J. Wild et al. published a paper entitled "Towards More Accessible Conceptions of Statistical Inference." Wild et al. argue that statistical inference should be taught earlier in students' careers. They assert that the current introduction to statistical theory does not give students an intuitive understanding of statistical theory. Wild et al. propose a new method of teaching ideas of statistical inference using box plots and a guide of milestones. I will refer to the new method proposed by Wild et al. as the milestone method. In the new proposed curriculum, students would learn to compare two populations using one milestone at a time. Wild et al. insinuate that the milestones will build upon each other, starting with simple intuitive understandings and slowly adding in statistical complexities. After students have worked with each milestone, they can then learn traditional statistical inference, like the t-test, with a greater appreciation for what the computations mean [11].

Today, the milestone method is presented in the resource section of the New Zealand Ministry of Education's website for teachers. However, the last national education standards for New Zealand

were published in 2010. Given that the Wild et al. paper was not published till 2011, it is not a surprise that the state standards for New Zealand statistical education does not explicitly include the milestone method [2]. However, because the government suggests the method as a resource for teachers there is a strong possibility that the milestone method will be part of the next set of New Zealand national education standards. If the milestone method is adopted as the hegemonic statistical inference curriculum in New Zealand, it could play a huge part in innovating math education. Contextualizing the Wild et al. paper, one sees that it is imperative to assess what we value in education and how innovative curriculums hold up under those values.



Figure 1.2: Various images of Wild et al.'s box plots with a memory gif [11]

In order to learn statistical inference through the milestone method, students are introduced to box plots: a visual representation of a set of data. A box plot shows a rectangle with one edge drawn at the 2nd quartile of a data set and the other edge of the box drawn at the 4th quartile of a data set. A line at the median of the data is drawn through the middle of the box. The introduction to statistical inference curriculum designed by Wild et al. begins with a visual tool the authors refer to as box plots with a memory. To show students how sampling variability of an entire sample occurs, they display a moving illustration that depicts many box plots of repeated sampling taken from one population. Figure 1.2 shows various still images of the box plot with a memory gif. They continually show sample box plots from the same population. When a new sample box plot is outlined in black, the old one is changed to a red coloring with the median line in blue. As seen in Figure 1.2, the representation of sampling as colored box plots gives students an idea of how samples from one population can differ. Wild et al.'s primary goal for student understanding is depicted below in Figure 1.3.

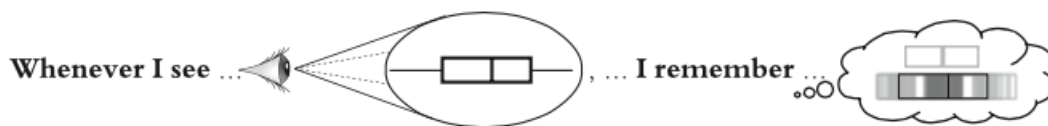


Figure 1.3: Primary goal in students' understanding of sampling [11]

Box plots with a memory can also give students a good understanding of how sampling changes

with differing sample sizes. As sample size increases, students can see the layered box plots becoming more and more similar. They can imagine as the sample size increases to the population size, every single layered box plots would represent the real population.

Once students have an intuitive understanding of sampling variability, they can compare sample box plots from two populations. To make inferences about populations using box plots Wild et al. suggests a set of milestones pictured in figure 1.4.


The milestone method is Wild et al.'s alternative method for hypothesis testing of the equality of the center of two populations. Each milestone adds more complexity to making the call as to whether two population centers are the same or different. Milestone 0, or the "At all levels" test, has the goal of prompting students to compare the overlap between two populations. The intuition behind milestone 0 is the sense that the variability of two samples from the same population (the null hypothesis is true) would rarely produce box plots that did not overlap at all. Thus milestone 0 urges students to make the call that two populations are different (reject the null hypothesis) if the box plots representing a sample from each population do not overlap at all.

Milestone 1 has a bit more depth in that it compares the median of one box plot to the 2nd and 4th quartile of the other box plot. Numerically, if **50%** of one sample is greater than **75%** or less than **25%** of the other sample, milestone 1 prompts a student to make the call that the populations are different. Wild et al. explains that using a sample size of 20-40 observations will maintain the utility of the test and allow students to focus on only a few new concepts. When teaching milestone 1, teachers would limit examples to sample sizes of 20-40. The paper claims that simulations with normal data have α values of about **15%** for $n = 20$, **7%** for $n = 30$, **3%** for $n = 40$, and **0.4%** for $n = 100$. The authors conclude that within the suggested sample sizes, students will generally make the same call as traditional statistical inference would suggest.

Wild et al. claim to move intuitively towards the traditional t-test methods in milestone 2. It urges students to compare the difference in box plot medians to the overall spread of the 2 middle quartiles. Visually, milestone 2 is intuitive and simple to calculate. According to the paper, the α values for milestone 2 are about **8%** for sample sizes 30 and 100. Wild et al. admits in the second milestone that the authors sacrifice a lower α rate for an extremely simple rule, but argues that the higher α is worth providing students with a useful rule of thumb and teaching students intuition used in the t-test.

Finally, milestone 3 compares the spread of a multiple of the interquartile range divided by the square root of the sample size. Milestone 3 is most complicated because it involves calculation

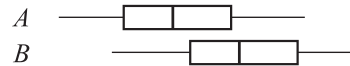
Guidelines on “how to make the call” by development level

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 ...

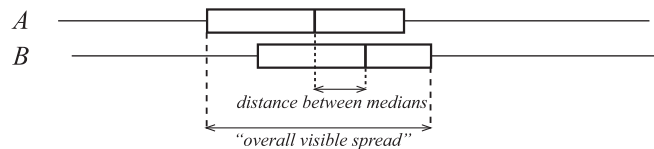
Milestone 1 test: *the 3/4-1/2 rule*




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”*



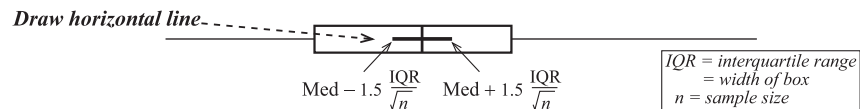
Make the call that ***B tends to be bigger than A*** back in the populations
if the distance between medians is greater than about ...


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

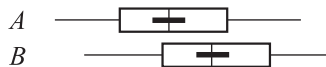

1/5 of overall visible spread for sample sizes of around **100**

[*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*



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



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

Milestone 4: *on to formal inference*

Figure 1.4: Visual presentation of Wild et al.’s milestone method [11]

rather than just visual measurements. Milestone 3 urges students to think about the variability of the sample median. After milestone 3, milestone 4 is formal statistical inference, like the traditional t-test method. Wild et al. claim that when students learn the milestones in order, they are prepared to understand formal statistical inference [11].

In 2014 Thalia Rodriguez, an undergraduate at Pomona College, wrote her thesis on the Wild et al. milestone method. She created R code which tested random samples run through every milestone consecutively. If any one milestone rejected the null, she considered the milestone test claiming a rejection of the null. She used code to compare the power (the probability that a test will correctly support the null) of the milestone method to the traditional t-test method [10]. Unfortunately, generating that data showed huge type I errors in the milestone method. Controllable type I errors are an essential part of traditional statistical inference and traditionally are presented early-on in statistical inference education. Thus if the milestone method is adopted as an educational tool, the unrestricted type I errors in the milestone method challenges one of the core tenants of statistical inference education. The following paper will further investigate the type I error rates as well as the intuitiveness of the milestone method, and compare the traditional t-test to the milestone method as tools for teaching statistical inference.

Chapter 2

Type I Errors

2.1 T-test

Under modest assumptions such as normal or close to normal data, the traditional t-test comparing two populations is an α level test. In other words, one can choose an α for a particular situation and the t-test will maintain that type I error level over time. The critical t value is determined based on the sample size and the confidence level, or desired α . As a student in introductory statistics, one is taught that the fundamental criterion of an effective and useful hypothesis test is the ability to set the type I error rate. Once a student recognizes that a test is an α level test, he or she can assess the usefulness of that test by calculating power. **Power** is defined as $1 - \alpha$ — the type I error rate. Many statistics courses will have students prove that the t-test is uniformly most powerful for a test of equality of means on normally distributed data.

Implicit in traditional method of statistical inference education is the definitive importance of a controllable and constant type I error rate. By teaching the t-test as the first statistical inference method, students learn that a maintainable type I error rate level is imperative and fundamental to hypothesis testing.

2.2 Milestone Method

Unlike the t-test, the milestone method does not claim controllable or consistent α levels. Additionally, the milestone method includes no intuition or explanation of type I errors as part of the curriculum. In their paper, Wild et al. claim that each milestone has approximately an $\alpha = .05$.

According to the paper [11], the approximate α values are as follows:

- **.15** to **.004** for milestone 1 and sample size between **20** and **40**
- **.05** for milestone 2
- **.02** for milestone 3

One can explore the theory behind the α levels for normal data in the simpler milestones, such as milestone 0. Recall from the background chapter that milestone 0 dictates that if box plots representing two samples do not overlap at all, students should reject the null hypothesis. If there are two populations **A** and **B** both distributed normally with $\mu = 0$ and $\sigma = 1$, the α level for milestone 0 is defined as the probability that the third quartile of the sample from population **A** is less than the first quartile of the sample from population **B** plus the probability that the third quartile of the sample from population **B** is less than the first quartile of the sample from population **A**. Both **A** and **B** are normal and symmetric over 0. Thus for the following calculation assume that α is two times the probability that the first quartile of the sample from population **A** is less than the third quartile of the sample from population **B**.

For some integer i let Q_{i_m} be the i^{th} quartile of sample m .

Let $\mathbf{A} \sim N(0, 1)$ and $\mathbf{B} \sim N(0, 1)$ independent,

and name samples A_1, A_2, \dots, A_n and B_1, B_2, \dots, B_n .

We define $\hat{Q}3_A$ as the center of the distribution of the third quartile of A_1, A_2, \dots, A_n and we define $\hat{Q}1_B$ as the center of the distribution of the first quartile of B_1, B_2, \dots, B_n

For milestone 0, $\alpha = 2 * P(\hat{Q}3_A < \hat{Q}1_B)$

For $X = \hat{Q}3_A$ and $Y = \hat{Q}1_B$ we know

$$\begin{aligned}
P(X < Y) &= \iint_{X < Y} f_{xy}(x, y) dx dy \\
\text{by independence,} &= \iint_{X < Y} f_x(x) f_y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^y f_x(x) f_y(y) dx dy \\
&= \int_{-\infty}^{\infty} f_y(y) \left[\int_{-\infty}^y f_x(x) dx \right] dy \\
&= \int_{-\infty}^{\infty} f_y(y) [F_x(y)] dy \\
&= \int_{-\infty}^{\infty} (\text{the pdf of } \hat{Q}1_B \text{ evaluated at } y) (\text{the cdf of } \hat{Q}3_A \text{ evaluated at } y) dy
\end{aligned} \tag{2.1}$$

the cdf of $\hat{Q}3_A = F_x(y)$

$$\begin{aligned}
&= P(X \leq y) \\
&= P(\hat{Q}3_A \leq y) \\
&= P\left(\frac{3}{4} \text{ of } \mathbf{A} \leq y\right) \\
&= \text{Binomial with probability } \Phi(y) \\
&= \binom{n}{\frac{3}{4}n} (\Phi(y))^{\frac{3}{4}n} (1 - \Phi(y))^{\frac{1}{4}n}
\end{aligned}$$

the pdf of $\hat{Q}1_B = f_y(y)$

$$\begin{aligned}
&= \frac{d}{dy} F_Y(y) \\
&\text{plugging in from left side,} \\
&= \frac{d}{dy} \binom{n}{\frac{1}{4}n} (\Phi(y))^{\frac{1}{4}n} (1 - \Phi(y))^{\frac{3}{4}n} \\
&= \binom{n}{\frac{1}{4}n} \left[\frac{1}{4}n (\Phi(y))^{\frac{1}{4}n-1} \varphi(y) (1 - \Phi(y))^{\frac{3}{4}n} \right. \\
&\quad \left. + (\Phi(y))^{\frac{1}{4}n} \left(\frac{3}{4}n\right) (1 - \Phi(y))^{\frac{3}{4}n-1} (-\varphi(y)) \right]
\end{aligned}$$

Plugging these results into 2.1 we get

$$\begin{aligned}
P(X < Y) &= \int_{-\infty}^{\infty} \left[\binom{n}{\frac{1}{4}n} \left[\frac{1}{4}n (\Phi(y))^{\frac{1}{4}n-1} \varphi(y) (1 - \Phi(y))^{\frac{3}{4}n} \right. \right. \\
&\quad \left. \left. + (\Phi(y))^{\frac{1}{4}n} \left(\frac{3}{4}n\right) (1 - \Phi(y))^{\frac{3}{4}n-1} (-\varphi(y)) \right] \right] \left[\binom{n}{\frac{3}{4}n} (\Phi(y))^{\frac{3}{4}n} (1 - \Phi(y))^{\frac{1}{4}n} \right] dy \tag{2.2}
\end{aligned}$$

Using the calculation above and plugging in for different sample sizes, we find the α presented in the second column of table 2.1.

Sample Size	Theorized α	Simulated α
$n = 10$.0485	.0196
$n = 50$.0000	.0000
$n = 100$.0000	.0000
$n = 500$.0000	.0000

Table 2.1: Theorized and simulated α for milestone 0

An alternative method for measuring α levels for milestones is simulation. Using R, one can code the tests of comparison for each milestone and run random samples of normal data to find the α values through simulation. For the results displayed in the table, **10,000** samples of differing sizes are drawn from two standard normal distributed populations. The quartiles of the samples are compared as described in each milestone and the number of times the milestone method would reject the null (despite the null being true) is counted, giving us a simulation of the α for each milestone. Using R to calculate α values for milestone 0 using normally distributed sampling, one gets the results displayed in the third column of Table 2.1.

For milestones beyond milestone 0, the theoretical type I error rates are calculated using identical but substantially less trackable integration. However, as seen in milestone 0, using simulation in R to find α values allows us to find approximate results and compare these with the results Wild et al. reports in their paper. Tables 2.2, 2.3, and 2.4 show simulation-determined α values for each milestone and for a variety of sample sizes.

Sample Size	Wild et al. α	Simulated α
$n = 10$.004 to .15	.4149
$n = 50$.004 to .15	.0189
$n = 100$.004 to .15	.0007
$n = 500$.004 to .15	.0000

Table 2.2: Predicted and simulated α for milestone 1

Sample Size	Wild et al. α	Simulated α
$n = 10$	$\sim .05$.1438
$n = 50$	$\sim .05$.0783
$n = 100$	$\sim .05$.0612
$n = 500$	$\sim .05$.0203

Table 2.3: Predicted and simulated α for milestone 2

Sample Size	Wild et al. α	Simulated α
$n = 10$.02	.0568
$n = 50$.02	.0274
$n = 100$.02	.0277
$n = 500$.02	.0231

Table 2.4: Predicted and simulated α for milestone 3

It is important to note that Wild et al. make no claim that the milestone method is an *alpha* level test. The authors explain “There is a trade off between more conventional type I error rates... and having an extremely simple rule. We gave more weight to the latter [11].” However, we believe that an assessment of the type I error rate is a vital part of understanding the effectiveness of a statistical test.

Chapter 3

Intuition

3.1 T-test

The t-test is the accepted method of teaching introductory statistical inference to compare the centers of two populations. However, there is not a defined curriculum for how teachers use the t-test in a classroom. The following is a generalized description of a method of statistical inference education using the t-test. While the particular curriculum differs classroom to classroom, the pedagogy of statistical inference is consistent with the following example curriculum.

The first step in students understanding the t-test is developing knowledge about sampling. Many statistics classes use applets to demonstrate sampling. One example of such an applet is depicted in figure 3.1. In the applet depicted below, students are given a histogram of the true population, and they choose a sample size. As they click the “draw samples” button the applet generates a sample as well as the sample statistic. Below is an example of an educational applet measuring the mean of samples of a population. Students are able to see where a particular mean of a sample draw lies in the distribution of a sample statistic. They also can observe that as the applet draws more and more samples or as the sample size increases, the distribution of the sample statistic centers around the true mean of the population.



Figure 3.1: Applet demonstrating variation of a sample statistic [5].

The applet depicted above and others like it give teachers a hands-on tool that demonstrates sampling and clarifies the idea of the sampling distribution of a statistic. After students have an intuitive understanding of sampling and sampling distribution, the teacher can introduce the t statistic.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

To understand the t-statistic intuitively, one can break down the numerator and denominator. The numerator measures the difference between the means of the two samples $(\bar{x}_1 - \bar{x}_2)$ as compared to the predicted difference $(\mu_1 - \mu_2)$. In the two-hypothesis test, the predicted difference, or the null hypothesis, is that $(\mu_1 - \mu_2) = 0$, in which the numerator is simply the difference between means of the samples. The denominator is more difficult to understand technically but it can be easily thought of as the variability associated with the samples moderated by the sample size.

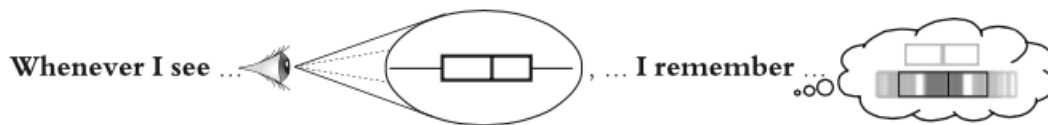
The intuition established by the sampling applet theoretically gives students a sense of how the t-statistic can vary. Then using a t-distribution the students can determine at what value of the t-statistic they will reject the null hypothesis. The hope is that students will picture the sampling

distribution of the statistic shown in the above applet while working with the t-statistic equation. Many people, including Wild et al., feel that statistical inference intuition often does not occur and as a result students do not gain a good understanding of the t-test.

3.2 Milestone Method

While the milestone method has not been formally tested in a classroom, one can turn to experts and logic to assess the validity of Wild et al.'s claim that the milestone method is more intuitive and can lead students to a deeper understanding of the t-test.

In the responses to the Wild et al. paper, educators and statisticians responded positively to the use of box plots to further students' understanding of sampling. The gifs developed by Wild et al. parallel the applets traditionally used to demonstrate sampling. Unlike the applet described in the previous section, the Wild et al. gifs demonstrating sampling have the same visual tools (box plots) as the inference test (milestones method). While most of Wild et al.'s sampling demonstrations are animated gifs and cannot be displayed statistically, the figure below shows the final goal of these gifs.



Due to the overwhelming positive responses to these visual representations of sampling from the educators and statisticians reviewing the Wild et al. paper, one can conclude that the intuition of sampling is better taught through box plots than through traditional methods.

As compared to the applets used in teaching the t-test, the Wild et al. gifs on sampling directly correspond to the hypothesis testing. The intuition students develop about variation of a sample is represented by box plots. Thus when they see the testing based on box plots, they are potentially more likely to connect their intuition about sampling to the milestone method of testing.

However, Wild et al.'s claim that the milestone method guides students towards a deeper understanding of the t-test (the build up to milestone 4) is subject to more debate. In order to determine the validity of Wild et al.'s claim, one must first understand the logic behind it. The following pages will describe each milestone and its connection to the t-test. The following breakdown of the theory behind each milestone was not included in the original Wild. et al. paper. The authors wrote very little about the logic behind the milestones and the intuition bells is my own analysis.

Milestone 0 and milestone 1 draw on similar intuitions. Upon seeing two box plots, a student's immediate reaction is to compare the two box plots with one another. In particular, they will look to the median lines of each box plot in order to compare the two populations [11]. Wild et al. attempts to encourage students to think instead about the variability of each box plot. The authors explain that students are more likely to identify the differences between the quartiles outlined by the two box plots rather than the horizontal spread of each box plot. milestone 0 and milestone 1 both should prompt students to recall the red and blue outlines of the quartiles of each sample shown in the sampling gifs. milestone 1 does not depend on sample size, but it has a restriction of sample size for teachers using the milestone method.

As noted by Wild et al., milestone 2 does indeed draw students closer to understanding the t-test. milestone 2 compares the difference in the medians to the spread of the samples. According to the paper, milestone 2 draws students closer to the t-test because the t value also represents the difference between the centers of each sample scaled by the spread (standard deviation) of the sample. The comparison in milestone 2 between distance of medians and overall spread of each sample depends on sample size, with the ratio of difference in centers to difference in range of spread increasing as sample size decreases. Milestone 2 refers to the median rather than the mean as the center of the sample and measures the spread of data using the center half of the sample rather than standard deviations scaled by sample size. In other words, milestone 2 introduces the concepts used in the t-test but uses more more visual versions of the center of the data and the spread of the data that students can easily identify on box plots. Wild et al. use the middle half of the data to represent the spread of the sample because it is a simple visual range. The similarities between milestone 2 and the t-test are shown below. In order to claim that population **1** and population **2** are different, the data must give either

$$\frac{|\mathbf{m}_1 - \mathbf{m}_2|}{\text{visible spread}} > \frac{1}{3} \qquad t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}} > \sim 2$$

for the milestone method and the t-test, respectively.

As students advance in school they learn the milestones sequentially. milestone 3 should, according to Wild et al., further expose students to theory behind the t-test. milestone 3 is the first milestone to incorporate sample size into the test. It uses a comparison of interquartile range (the distance between the first and third quartile) for each sample. The comparison in milestone 3 is based on Tukey's notched box plots idea. Tukey et al. published an article entitled "Variations of Box Plots" introducing more descriptive box plots in 1978. An example of Tukey et al.'s notched

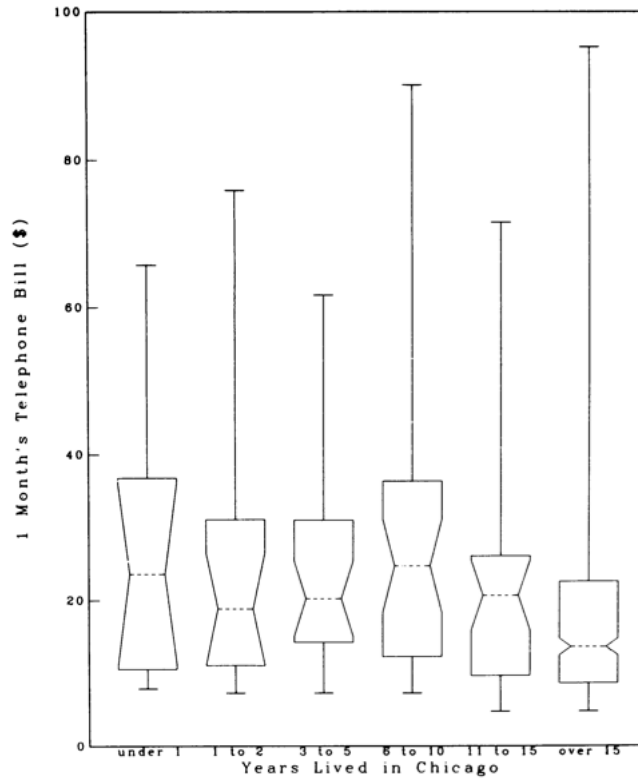


Figure 3.2: Example of Tukey's notched box plots [9]

box plots is depicted below [9].

By adding the indentations, or notches, around the median line for each box plot, Tukey et al. incorporated variability of the median adjusted by sample size into the box plot comparison. The paper claims that if the notches of two box plots do not overlap than the medians are significantly different with a **95%** confidence level [9]. Tukey et al. explains the calculation of the width of the notches using the Gaussian-based asymptotic approximation of the the standard deviation, s , and a **95%** confidence level. The notch width should be

$$M \pm 1.7 * \frac{1.25 * R}{1.35 * \sqrt{N}} \quad (3.1)$$

Where M is the median, R is the interquartile range, and N is the number of observations in the sample. Tukey et al. used **1.7** as the constant based on the asymptotic theory used to obtain the **95%** confidence interval for the true median. A **1.96** constant would work nicely if the standard deviations of the two samples were very different and did not overlap. If the standard deviation

of the two samples were very similar, a constant even smaller than **1.7** should be used in order to maintain the 95% confidence interval [9]. Tukey et al. empirically chose **1.7**.

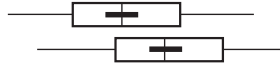


Figure 3.3: Wild et al.'s notched box plots [11]

Wild et al. adapted Tukey et al.'s notched box plots using a slightly different visual tool, the thick barred line intersecting the median of each box plot as seen above in figure 3.3. In their calculation of the barred horizontal line, Wild et al. used a smaller constant than Tukey et al., **1.5**. Wild et al. argues that the **1.5** multiplier is more memorable to students and will produce approximate **90%** confidence intervals. They hope that milestone 3 will give students intuition about why comparing two populations depends on uncertainty of each median value [11].

Chapter 4

Future Research and Conclusion

An exploration of the ways in which the Wild et al. milestone method is different and/or superior to the traditional t-test as a method of teaching statistical inference resulted in the following conclusions.

- The Wild et al. paper does provide useful and intuitive tools for understanding sampling. The paper received feedback from a multitude of educators and statisticians. Across the board these experts agree that the visual tools, in particular the moving box plots gif representing sampling, are useful tools that help students understand sampling[11].
- These intuitive and visual tools represent a movement in math education to break down the “black-box” method of learning. As reflected in the recent Common Core State Standards, many believe that curriculum should focus on students’ deep understanding of core ideas in math [3]. The Common Core website claims “rigor refers to deep authentic command of mathematical concepts” and that students should “pursue conceptual understanding” [3]. Thus in the context of a general movement towards students gaining a deeper understanding of math through standards-based education, Wild et al. provides valuable visual tools for students to understand sampling.
- While the tools conveying sampling are almost universally considered useful, as a method for statistical inference education the milestones are more controversial. As demonstrated in chapter 2, the milestone method does not preserve one fundamental piece of the structure of the t-test: the α levels. In fact, the proposed milestone curriculum does not include a lesson on α , a major deviation from traditional statistical inference education where students are

exposed to α before learning two sample hypothesis testing. Wild et al. propose a curriculum that values intuitive understanding over introduction of type I error rates. Thus the extent to which the milestone method is superior to the traditional t-test depends on what education should prioritize.

- Finally, the superiority of the milestone method depends on the intuitiveness of the method. Wild et al. claim that box plot based statistical inference curriculum will enhance students' understanding of statistical inference and enable them to better use the t-test in the future. In the chapter on intuition we investigated how the milestone method depends on similar theory as the t-test. Our investigation revealed that each milestone did share intuition with the t-test, but discovering the connections took a significant amount of work to uncover. The relationship between the milestone method and the t-test is not obvious and could be incredibly hard for teachers to communicate to students in a classroom.

Wild et al. simply argue that because the milestone method shares theory with the t-test and provides more visual and simple rules to test a hypothesis, the milestone method will provide a rigorous foundation for students to learn formal statistical inference. I question that argument. The intuition used in creating the milestones was incredibly complicated and difficult to unpack. Do Wild et al. really expect teachers to have the time and content knowledge to understand those connections? In addition, do Wild et al. expect students to absorb that knowledge even if teachers have a deep understanding of it? I see a large gap between the intuition and theory used to create the milestone method and the intuitive understanding that students absorb from the milestone method.

The evaluation of the milestone method would greatly benefit from contributions from education specialists and testing of students' understanding. Such contributions could argue more convincingly that students are or are not able to gain intuition about the t-test through the milestone method.

In the New Zealand state school system, statistics education begins at level 1 with students generally 5 years of age. For each year of school, New Zealand has state standards for what students should master with regard to statistics. Wild et al. suggest the milestone method for use in level 5 through level 7, or ages 9 through 12. Under resources for each year, the New Zealand Ministry of Education's website links to a guideline for teachers to use the milestone method effectively [2]. Not only does the Wild et al. resource suggest what milestone to teach when, it also addresses the issue of sample size when using the milestone method [1]. Using examples of box plots representing

different sample sizes from the same population, Wild et al. show teachers how the milestone method becomes a stronger tool for statistical inference as sample size increases and the variability of the sample decreases. Thus an official resource for teaching the milestone method does address some issues that I have identified with the milestones. It is important to note that the adoption of New Zealand statistical inference education to teaching the milestone method signifies that this paper could have deep and lasting consequences on the way students are taught and therefore presumably understand statistics. In other words, the Wild et al. milestone method is already affecting real students in classrooms.

As reflected in the comments on the Wild et al. paper, I believe the strongest contribution the Wild et al. paper made was questioning the traditional curriculum for statistical inference [11]. They challenged the historically predominant education system and proposed a possible new curriculum that is arguably more intuitive. However, after investigating the Type I errors and the intuition in each milestone, I conclude that while the visual tools explaining sampling are helpful to students, the milestone method is not necessarily superior to the t-test due to lack of explanation, control of Type I errors, and unclear conveyance of statistical inference theory.

In order to truly understand the effect of the milestone method on statistical inference education, educational evaluation is necessary. Educational assessment would include studies of classrooms learning statistical inference with a traditional curriculum and classrooms learning through the milestone method. Before and after learning statistical inference, students would be interviewed and tested on their understanding and ability to apply statistical inference. Because New Zealand is already using the milestone method in public education, this could be the best place to start assessing the value of the milestone method as an educational tool. A study comparing New Zealand students' understanding of statistical inference and ability to effectively use statistical inference after a milestone method curriculum versus a t-test curriculum could help assess the quality of the milestone method as an educational tool. A study measuring the effectiveness of each type of curriculum would give us a more tangible result as to whether or not the milestone method presented by Wild et al. is a superior curriculum for statistical inference education.

Bibliography

- [1] census at school. <http://www.censusatschool.org.nz/>.
- [2] Ministry of education. <http://www.minedu.govt.nz/>.
- [3] Governor’s Board. common core. <http://www.corestandards.org/>.
- [4] The College Board. Program summary report. <http://media.collegeboard.com/digitalServices/pdf/research/2014/Prog-Summary-Report-2014.pdf>.
- [5] Beth L Chance and Allan J Rossman. *Investigating Statistical Concepts, Applications, and Methods*. Duxbury Press.
- [6] Churchill Eisenhart. On the transition for “student’s” z to “student’s” t. *The American Statistician*, 33(1):6–10, February 1979.
- [7] R. A. Fisher. The statistical method in psychical research. *Proceedings of Society for Psychical Research*, pages 189–192, 1929.
- [8] Nicholas J. Horton and Suzanne S. Switzer. Statistical methods in the journal. *New England Journal of Medicine*, 353(18):1977–1979, 2005. PMID: 16267336.
- [9] Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of box plots. *The American Statistician*, 1978.
- [10] Thalia Rodriguez. Towards a more conceptual way of understanding and implementing inferential rules. Master’s thesis, Pomona College, 2014.
- [11] C.J. Wild, M. Pfannkuch, and M. Regan. Towards more accessible conceptions of statistical inference. *Journal of Royal Statistical Society*, 2011.