CHAPTER 5

Comparing the Performance of Two Groups and what it "Means"

Comparing a Smaller Group to the Overall Group (A Sample to a Population)

It's possible that we might want to compare some average measurement for a group of our students to a larger population. For example, we might want to compare how the 4th graders in our school performed in comparison to all of the other 4th graders in the state and if they performed, on average, the same or significantly different. In order to do this, we would have to perform a statistical test that compares the *mean* score for our group to that of the overall population *mean*. There will more than likely be a difference in the *mean* scores but to determine if that difference is statistically significant, in other words, that it did not happen due to chance or error, we would have to test for a statistically significant difference. We can do this by utilizing either a *z-Test* or a *Single Sample t-Test*, depending upon the information available to us.

We use a *z-Test* when we know the population *mean* AND population *standard deviation.* Usually obtaining the population *mean* is pretty easy since most test vendors will report this by school, district, state, etc. However, sometimes the population's *standard deviation* is not that easy to obtain. If we do have the population's *standard deviation,* we can then calculate the *z-Test*; if not, we will have to calculate a *Single Sample t-Test*. The formulas for each test are as follows:

***z-Test*:**

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

(Hinkle, Wiersma, & Jurs, 2003, p. 183)

Where:

Z is the calculated value of the test statistic
$\overline{X}$ s the sample mean
μ is the population mean
$\sigma_{\overline{X}}$ Standard error, which is based on the known population standard deviation (Sigma – σ)

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Where: σ = population standard deviation and n = size of the sample

t-Test:

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

(Hinkle, Wiersma, & Jurs, 2003, p. 194)

Where:

t is the calculated value of the test statistic
$\overline{X}$ is the sample mean
μ is the population mean

$S_{\overline{X}}$ is the estimated standard error, which uses the sample standard deviation (*S)* to estimate the standard error

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

Where:
S = sample standard deviation and n = size of the sample

(Hinkle, Wiersma, & Jurs, 2003, p. 194)

Although these formulas may look complicated, relax, Excel is going to do the heavy lifting. By learning how to calculate a *z-Test* in Excel, we will for all intents and purposes also be learning how to compute a *Single Sample t-Test*.

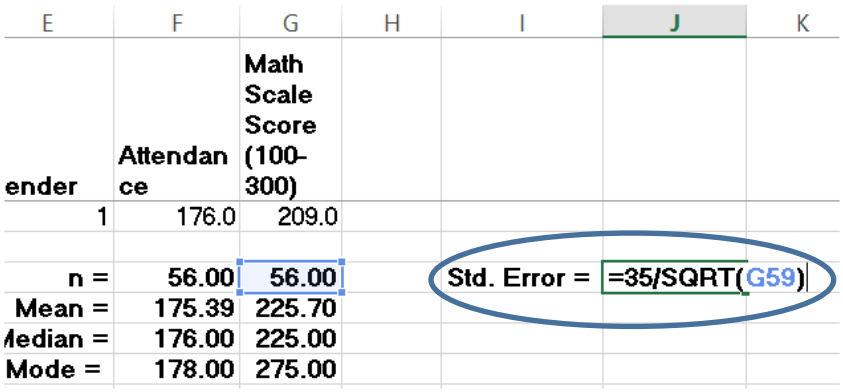
Calculating the *z-Test* in Excel
(Use Excel Chapter 5 Workbook, **Gd. 4 Math Scores at:**
http://www.ncpeapublications.org/index.php/ncpea-press-author-showcase)

Let's say we want to compare our school's 4th grade student performance on the state Math assessment to that of all the 4th graders in the state. In order to do this, we would have to know the *mean* score for the entire population of all 4th graders in the state and that population's *standard deviation*. Let's say the *mean* score for all 4th graders in the state is 210 and the *standard deviation* is 35. We select a random sample of 56 (n=56) of our 4th grade scores and calculate a *mean* of 225.7 and a *standard deviation* of 38.09. **(Note: see Chapters 2 & 3 and be sure to use Sample Standard Deviation, STDEV.S, for this set of data).**

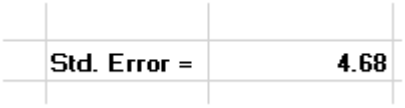
Gender	Attendance	Math Scale Score (100-300)
1	176.0	209.0
n =	56.00	56.00
Mean =	175.39	225.70
Median =	176.00	225.00
Mode =	178.00	275.00
SD =	3.71	38.09

In order to see if this score is significantly different from all 4th graders in the state, we can perform a simple *z-Test.* To do this, we first have to calculate something called *standard error.* If you refer back to our *z-Test* formula on page 107, you will see that we calculate the *z-Test* statistic by computing the difference between our sample mean and the population mean by subtracting one from the other and then dividing that by the *standard error.* To calculate the *standard error* we simply divide the population *standard deviation* by the square root of the sample size. Piece of cake!

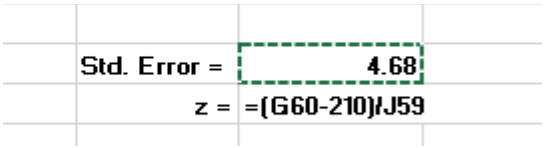
Open the Chapter 5 Workbook and go to the tab labeled, **Gd. 4 Math Scores.** Select cell **I59** and type in **Std. Error =**. In cell **J59** type in the following formula **=35/SQRT(G59)**. In this case **35** is the population *standard deviation* and **G59** is the sample size.



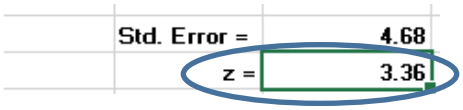
Hit the **Enter** key and you will get 4.677071733. Decrease the decimals to 2 places to get the following:



Now we can calculate our *z-Test* statistic. We simply subtract our sample mean (225.7) in cell **G60** from our known population mean of 210 and divide by the *standard error*, which we just calculated in cell **J59**. In cell **I60** type in **z** =. In cell **J60** type in the following formula **=(G60-210)/J59.**

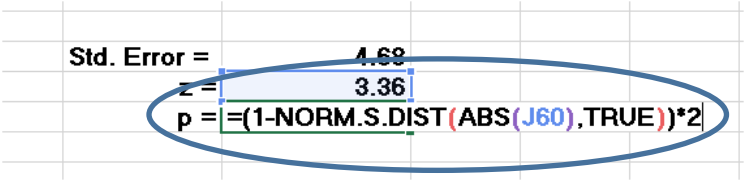


Hit the **Enter** key and you get 3.356037595. Decrease the decimals to two places to get the following:

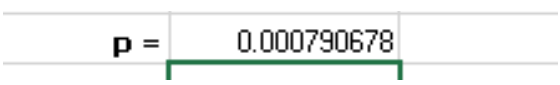


We know what you are thinking, "SO WHAT!" Well, all you need to know is that if you calculate a *z-Test* statistic that is larger than the absolute value of 1.96 than you have found a significant difference. In this case it means that your local 4th graders scored significantly different than the entire population of 4th graders. As a matter of fact, they scored significantly higher.

If you don't take our word for it that by calculating a *z-Test* statistic that is greater than 1.96 is statistically significant, you can actually calculate the precise *p value* in Excel by using a simple formula. In cell **I61** type in ***p*** =. Select cell **J61** and type in the following formula **=(1-NORM.S.DIST(ABS(J60),TRUE))*2**



Hit the **Enter** key and you get the following:



What will appears in cell **J61** is the *p value* or *level of significance* for this comparative statistical test; in this case it is .000790678, which we could simply round to .0008 by decreasing the decimal places.

What does this mean? It means that the probability that the difference between our 4th graders' *mean* Math score (225.7) as compared to the *mean* math score for all of the state's 4th graders (210) happened by chance or error is actually 8 in 10,000. WOW, what are the odds? Oh wait, we just figured that out. What this is actually telling us is that, on average, our 4th graders *mean* Math score is significantly higher than the *mean* score for all 4th graders in the state (210). Why this is the case could be based on a myriad of reasons, one of which could be the curriculum, the teachers, where the school is located, etc. In other words, we need to take this new information with a grain of salt and be humble in the fact that this result could actually be attributed to roughly 2.5% of all the possible sample groups of 4th graders from the overall state population.

Whenever we run a statistical test that provides us with a *p value,* we are always looking to see if the *p value* calculated is equal to or less than .05 (≤ .05), which is the default level for statistical significance for the social sciences. In other words, whenever we get a *p value* of .05, it tells us that the probability that the difference occurred by chance or error is at least less than 5 in 100 or 5%, which is rare.

[Refer to the link *Video 5.A – z-Test* listed at: http://www.ncpeapublications.org/index.php/ncpea-press-author-showcase]

What to do When We Do Not Have the Population Standard Deviation? Calculate a *t-Test.*

Unfortunately, 9 times out of 10 we will not know or be able to obtain the *standard deviation* for the entire population to which we want to compare our local group/sample. Therefore, we cannot calculate the *standard error*. When that occurs we then need to calculate a *Single Sample t-Test,* which is not a big deal. However, we will have to calculate a new statistic called the *estimated standard error.* We basically do this the same way we calculated the *standard error* for a *z-Test* but this time we use the information provided to us by our sample, our 4th grade scores.

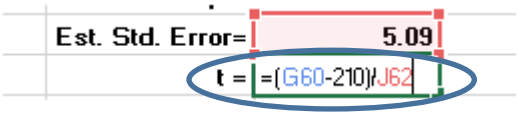
Using the same worksheet directly below where we calculated our *z-Test* and select cell **I62** and type in **Est. Std. Error** =. In cell **J62** type in the following formula **=G63/SQRT(G59)**. In this case **G63** is the *sample standard deviation* for our 4th graders and **G59** is again the sample size.

n =	56.00	56.00	Std. Error =	4.68
Mean =	175.39	225.70	z =	3.36
Median =	176.00	225.00	p =	0.0008
Mode =	178.00	275.00	Est. Std. Error =	=G63/SQRT(G59)
SD =	3.71	38.09	t =	

Hit the **Enter** key and we will get 5.08945431. Decrease the decimals to 2 places to get the following:

Est. Std. Error=	5.09
t =	

Now we can calculate our *t-Test* statistic. We simply subtract our sample mean in cell **G60** from our known population mean of 210 and divide by the *estimated standard error*, which we just calculated in cell **J62**. In cell **I63** type in **t =**. In cell **J63** type in the following formula **=(G60-210)/J62**

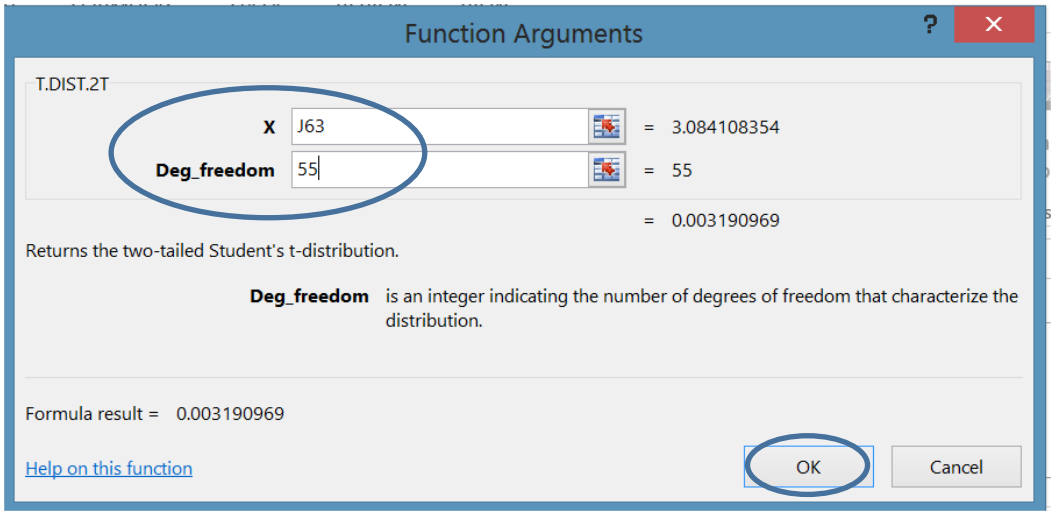


Hit the **Enter** key and you get 3.084108354. Decrease the decimals to two places to get the following:

Est. Std. Error=	**5.09**
t =	**3.08**

In order to determine if the value for the *t-Test* statistic is significant we have to calculate a *p value* for the test value. Unlike the *z-Test*, there is no absolute value (i.e., 1.96) that would tell us if the value we just calculated is statistically significant. So, we are going to have to input a function to tell us if it is significant.

Select cell **I64** and type in **p =**. Select cell **J64** and select the **Insert Function** icon to the left of the formula bar. When the **Insert Function** box appears select **T.DIST.2T** from the **Select a function:** menu and click on **OK.** The **Function Arguments** dialog box will appear. Make sure the cursor is blinking in the space next to **X** and type or click in **J63**, which is the value for *t* that we just calculated. In the box next to **Deg_freedom** type in the number **55.** The number 55 represents the degrees of freedom for a t-*Test*, which we need to determine significance. It is simply calculated by subtracting 1 from the total sample size, in this case it would be 56-1 = 55.



Click **OK** and we get a *p value* of .003190969, which is simply rounded to .003**:**

Est. Std. Error =	**5.09**
t =	**3.08**
p =	**0.003190969**

This new result tells us that the probability that the difference between our 4th graders' *mean* Math score (225.7) as compared to the *mean* math score for all of the state's 4th graders (210) happened by chance or error is approximately 3 in 1,000. Not quite the odds we previously calculated (8 in 10,000) but nonetheless, still a statistically significant result since we more than met the threshold where our *p value* is less than .05 (≤.05).

[Refer to the link *Video 5.B – Single Sample t-Test* listed at:
http://www.ncpeapublications.org/index.php/ncpea-press-author-showcase]

Statistical Significance and/or Practical Significance – Effect Size (*Cohen's d)*

Just because we might obtain statistically significant results, it does not mean that these results are practically significant. In other words, we might find that, on average, statistically significant results occur based on some treatment or condition. However, the difference may only be marginal and even though statistically significant, the practical significance might be small and not warrant our attention. To determine if this might be the case, we need to calculate what is referred to as an *effect size.* Almost every statistical analysis that we can do has a corresponding effect size. The effect size tells us just how much of an effect the treatment or condition impacts our outcome or dependent variable. However, be advised that it is only necessary for us to calculate an effect size when we find statistical significance. If the results of our statistical test are not significant, then effect size is moot.

The most commonly used effect size estimate in statistics is *Cohen's d. Cohen's d* provides us with a measure that helps us to interpret our results for practical purposes. In other words, does the significant difference we find actually have an effect that we should take seriously or do something about? The accepted interpretation for the *Cohen's d* range of *effect size* values is:

$\leq .20$ = small
.25 – .70 = medium
$\geq .80$ = large

(Hinkle, Wiersma, & Jurs, 2003; Warner, 2008)

The value we calculate for *Cohen's d* is literally interpreted as a unit of *standard deviation*. Specifically, a unit of *standard deviation* with regard to where the *mean* of the treatment/condition/experimental group falls in relation to the comparison group or population. So, if we calculate an effect size of .8 for a *t-Test,* it tells us that the *mean* for the experimental group is almost one full *standard deviation* different than the comparison or control group; pretty big difference, don't you think?

For example, let's say that we found that students who attended an SAT Math prep class score significantly higher on the Math portion of the SAT then all students in general. When we calculate the *Cohen's d,* we get a value of .96, which is considered a very strong effect. In fact, what this is telling us is that, on average, the *mean* SAT Math score for those students who attended the SAT prep class is almost one full *standard deviation* above the entire population who took the SAT at the same time. If the average SAT Math score for all students who took the SAT that year was 550 and the *standard deviation* was 110, this would mean that the average SAT math score for all those who attended the SAT prep class would be approximately 660. In this case, our statistical significance is now supported by practical significance.

To calculate a *Cohen's d* effect size for a *Single Sample t-Test,* we simply divide the difference between the sample *mean* and the population *mean* by the sample *standard deviation* or:

$$d = \frac{\overline{X} - \mu_{hyp}}{S}$$

Where:
d = Cohen's effect size
$\overline{X}$ = mean for the sample
μ_{hyp} = mean for the population
S = sample standard deviation

http://www.unt.edu/rss/class/Jon/ISSS_SC/Module008/isss_m8_introttests/node2.html

OR
for our previous example using 4th grade Math scores

$$d = \frac{225.7 - 210}{38.09}$$

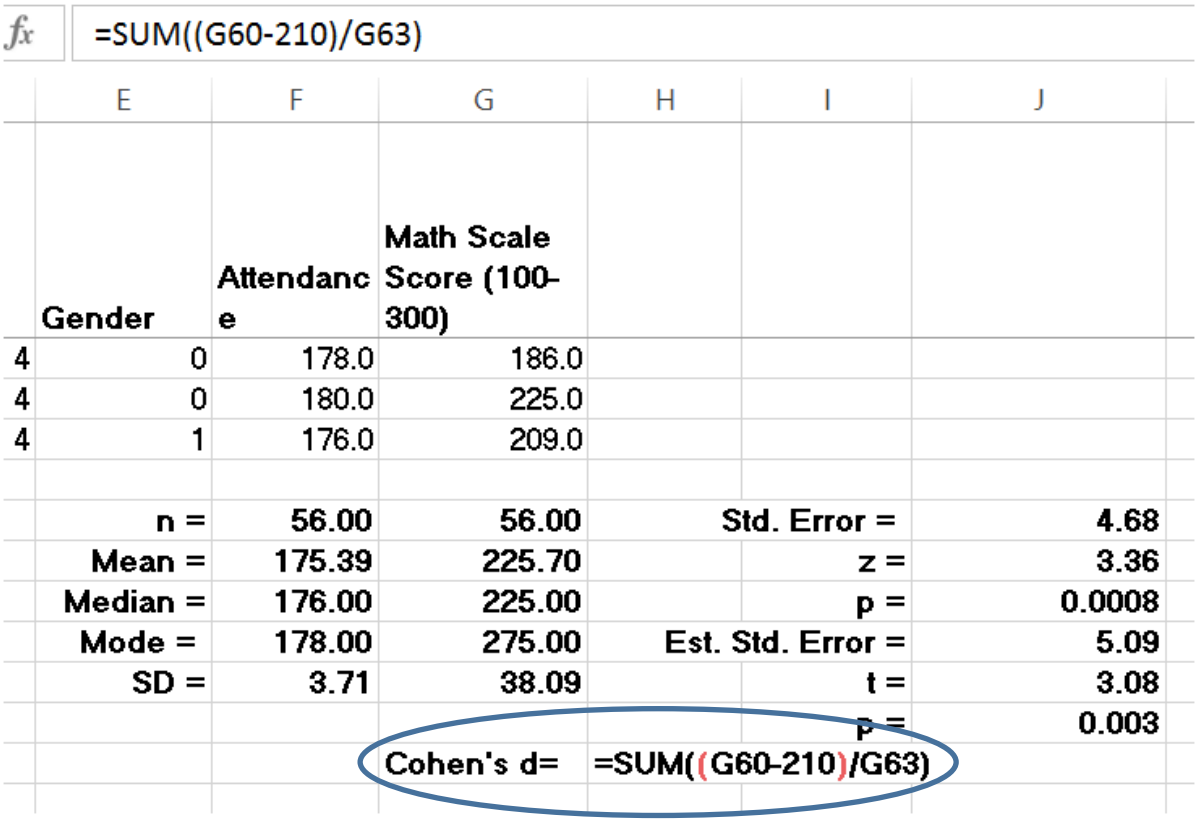
$$d = \frac{15.7}{38.09}$$

$$d = .41$$

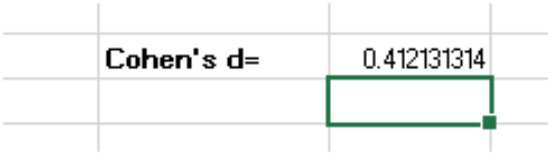
So, in the case of the *statistically significant* 4th grade Math scores, we would get an effect size of .41, which would be considered basically a medium or moderate effect size. What this tells us is that, on average, our 4th graders' *mean* Math score on the state's standardized assessment is approximately a little less than a half a *standard deviation* higher than the average Math score on the state's standardized assessment for all 4th graders.

Calculating a *Cohen's d* effect size for a *Single Sample t-Test*
(Use Chapter 5 Workbook, **Gd. 4 Math Scores at:**
http://www.ncpeapublications.org/index.php/ncpea-press-author-showcase)

In cell **G65** we'll type in **Cohen's d=** then place the cursor in cell **H65** and type in **=SUM((G60-210)/G63)**. Where cell **G60** represents the value for the sample *mean* (225.7), 210 equals the population *mean,* and cell **G63** is the sample *standard deviation* (38.09).



And press **Enter**



As you can see, we are left with the same value that we calculated by hand, .412131314, or simply .41. As we will see, the *Cohen's d* effect size can be calculated for each of our *t-Tests* discussed in this chapter with some slight modifications needed for each.

[Refer to the link *Video 5.C – Single Sample t-Test Effect Size* listed at: http://www.ncpeapublications.org/index.php/ncpea-press-author-showcase]

Comparing the Differences Between Two Means

School administrators often want to know if an action that they have taken resulted in positive results. For example:

- Did the change in start time result in a lower frequency of tardies?
- Did the new reading series result in higher language arts literacy scores on the state standardized test?
- Did the students in Mr. A's class do better than the students in Ms. B's class on the departmental midterm?
- Did the class do better on the posttest than they did on the pretest after a period of instruction?

To answer questions like these, we usually calculate and compare the *means* of the two groups that are either independent or dependent. Independent groups are comprised of different subjects (students) like two separate classes. Dependent or paired groups are comprised of the same subjects measured twice as in a pretest/posttest situation or pairs of subjects selected to be very similar to each other measured once at the conclusion of the treatment. This is referred to in Excel as a **t-Test: Paired.** Now, be advised that depending on the statistics book you pick up, this analysis is also sometimes referred to as a *matched/paired sample t-Test* or a *repeated measures t-Test*. It's just another case of statisticians trying to confuse the matter. The important thing to understand with the **t-Test: Paired** is that in most cases this test is used when we are testing the differences between *means* within the same group/sample based on two measurements that occur at two, separate distinct times to determine if the difference is significant. In the field of education this test is primarily used in a pre-posttest design, the same subjects measured twice, once before the treatment is given and once after the treatment is given. Consequently, this can be a valuable quantitative calculation for schools that incorporate Student Growth Objectives (SGOs) into their summative evaluation protocol for teachers.

The formula for a **t-Test:Paired** is:

$$t = \frac{\overline{D} - \mu_{hypD}}{S_{\overline{D}}}$$

(Witte & Witte, 2015, p. 333)

Where:

t = the t statistic
$\overline{D}$ = mean of the difference scores between groups
μ_{hypD} = hypothesized mean of the difference scores between groups
$S_{\overline{D}}$ = estimated standard error of the difference scores between groups