Equivalence Testing

Authored by: Aaron Ramert and Emily Westphal

28 August 2020

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

Hypothesis testing is a common practice for comparing two data samples with the intent of determining if the two source populations are different. Equivalence testing is an adjustment to this process to determine if the source populations are equivalent. Much of the process is the same, but there are some changes to the method and the theory. With equivalence testing there is also the added process of determining how close is close enough.

Keywords: JMP, hypothesis testing, consumer risk, producer risk, equivalence acceptance criterion

Introduction

There are situations in test and evaluation (T&E) where the goal of the test is to show that nothing changed. We can imagine that a new type of battery is used as a power source or standard parts are made from a new, cheaper material or some other change has been introduced into a system. The goal of the testing is then to show that the new and slightly altered system performs just as well as the legacy system. A common method of comparing two data populations is to conduct a statistical hypothesis test on two representative samples. If the reader is unfamiliar with hypothesis testing we recommend the best practice *Statistical Hypothesis Testing* by Jennifer Kensler. Hypothesis testing can be used to show a difference between two data samples and, if the proper assumptions are met, apply the results to the parent populations with statistical inference. A tester will only conclude there is a difference when there is an abundance of evidence to show they are different. When faced with a small amount of evidence that they are different the tester will be unable to conclude that there is a difference and continue as if they are the same. Continuing as if the samples are the equivalent is not the same as proving that they are. Equivalence tests are, "based on the desire to show that something is close enough to ideal to be acceptable" (Pardo, 2014). Equivalence testing uses the same approach as standard hypothesis testing but changes the null and alternate hypothesis so that the null is that the systems are different and the alternate is that they are the same. In short, for equivalence testing, the going-in presumption is that the populations are different; compelling evidence is necessary for the test team to conclude otherwise.

Equivalence tests are commonly used in the pharmaceutical industry to show, for example, that a generic drug has equivalent efficacy as a name-brand drug. There are several potential applications in the DOD as well: showing that simulation results were equivalent to live test results or that an upgraded systems performs at least as well the legacy system. This best practice explores equivalence testing in depth, including how the change in premise affects the test process and the associated test risks and metrics.

Review of Hypothesis Testing

Hypothesis testing supports decision making through a process which allows us to compare a sample statistic against a specified value or compare two different sample statistics to each other in order to

draw conclusions between two samples. We can use statistical inference to extend that conclusion to the populations which produced the samples, provided that the assumptions for statistical inference have been met. Hypothesis testing is relied upon because we have the ability to set the required confidence prior to testing and to determine the number of samples we require to achieve our desired power.

Hypothesis Testing Starts with a Theory

Hypothesis tests are conducted for a purpose. The chain of events starts with a theory. For example, we may believe that a system under test (SUT) produces an output greater than some reference value. Or perhaps we believe that after making some alteration to the SUT that it now has an output that is different in some way – in this case, the theory in question is that when we change a factor level in the SUT, the response value will also change. We actually investigate a theory every time we conduct a designed experiment.

Convert the Theory into Hypothesis Statements

The next step in the process is to convert the theory into two competing claims, expressed as hypothesis statements, which are written in a way so that they cannot both be true. The null hypothesis (H_0) represents, "the status quo, conventional thinking, or historical performance" (Kensler, 2018). The claim to be tested is the alternate hypothesis (designated by H_1 in this best practice but sometimes written as H_A in other documents). There are several possible hypotheses that can be made with regard to any measurable statistic. Three example hypothesis statements with a null and alternate hypothesis measured about the mean are shown in Figure 1.

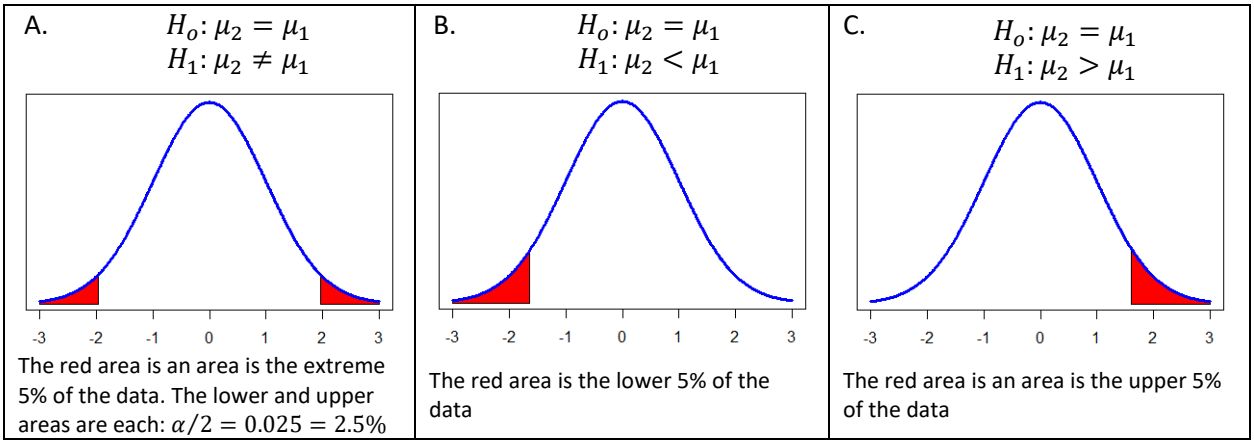


Figure 1: Three sample distributions with α shaded in red

In Figure 1 the reader can see that example A is a two-sided hypothesis test, meaning that we do not care if μ_2 is greater than or less than μ_1, we only want to show the mean is different between these two populations. In the other two examples we seek to show that μ_2 is specifically less than or greater than μ_1, respectively. Both examples B and C are one-sided tests. It most cases H_0 has an equal sign so there is only value that makes it true. There are an infinite number of values that can make H_1 true. In none of these examples (or any properly constructed pair of hypothesis statements) can H_0 and H_1 both be true.

There are two possible decisions. The tester can *reject* the null hypotheses and conclude that the alternative hypothesis is true or *fail to reject* the null hypothesis. The decision will be made based on the data collected and the accepted level of Type I risk. A more thorough investigation of these concepts is in the STAT COE best practice, *Statistical Hypothesis Testing* (Kensler, 2018).

Type I Risk

A Type I error is rejecting the null hypothesis when the null hypothesis is true. The probability that a Type I error is made is denoted as α (alpha) and is also called the significance level. This risk probability is used for test planning and is always set a priori, that is prior to testing. It is also common to refer to the confidence of a test. Confidence is expressed as a percentage and is the compliment of significance: $\text{Confidence} = (1-\alpha) \times 100\%$.

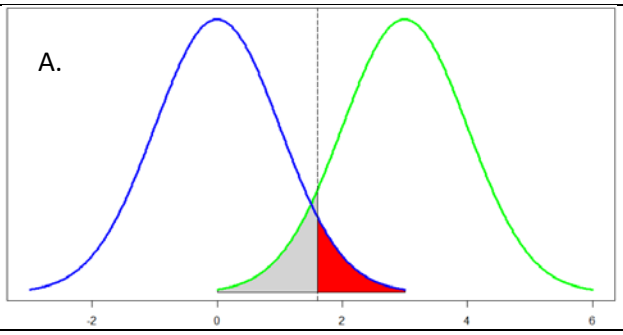
		Decision	
		Fail to Reject H_0	Reject H_0
Truth	H_0 is True	Confidence $(1-\alpha)$ Correct Decision	Type I Error (α) Incorrect Decision
	H_0 is False	Type II Error (β) Incorrect Decision	Power $(1-\beta)$ Correct Decision

Figure 2: Hypothesis test decision matrix

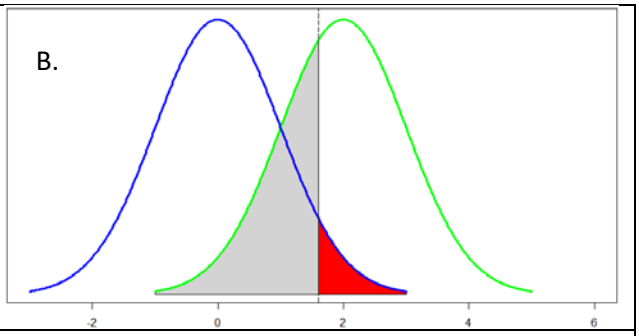
Type II Risk

A Type II error is failing to reject the null hypothesis (thus proceeding as if it is true) when the null hypothesis is false. The probability of a Type II error is denoted by β. It is typical to discuss the power of a test during test planning. The power is the complement of β: $\text{Power} = (1-\beta) \times 100\%$. Like α, power is used for test planning. However, unlike a Type I error, we do not directly set β. Instead it is calculated based on the selected α, the assumed difference between the data collected in the null configuration of the SUT and the altered configuration of the SUT, and the sample size. The calculations are usually performed by software and are not covered in this best practice.

The relationship between confidence and power can be further explained in one-sided upper-tail test shown in Figure 3. In this test $H_0: \mu = 0,\ H_1: \mu > 0$. The status quo data is represented by the blue curve on the left of both illustrations and is a standard normal distribution with $\alpha = 0.05$. To further highlight the α the upper 5% of the area under the blue curve has been colored red. The green curve on the right of both illustrations is also normal with $\text{SD} = 1$, but $\mu = 3$ in example A and $\mu = 2$ in example B. The gray area is β, which makes the remaining area under the green curve (colored white and red) representative of the power. The reader can clearly see that Figure 3A has greater power. For further information on increasing power review the best practice *Understanding the Signal to Noise Ratio in Design of Experiments* (Ramert, 2019).



The data from the status quo SUT is under the blue curve and $\alpha = 0.05$. The green curve represents the altered SUT. In this example $\beta = 0.09$ so the power is 91%.



The data from the status quo SUT is under the blue curve and $\alpha = 0.05$. In this example $\beta = 0.36$, which means that the power is only 64%.

Figure 3: An illustration of confidence and power when comparing two distributions

Examine the Data and Calculate the *p*-value

The next step in the process is to look at the data collected from the SUT under the alternative hypothesis and determine if the hypothesized change occurred. A simple interpretation of this process is to look at the data and ask, "What is the probability that I get a result this extreme or more extreme when the SUT is in the null state?" This probability is called the *p*-value. The actual value is calculated with statistical tables or software and then compared to α. If the *p*-value is lower than α, we assume that the results are too extreme to happen under the null state with a high probability and we choose to reject H_0 and conclude the alternative hypothesis must be true. Graphically we can draw a line at our test statistic and calculate the area to the edge of the distribution. That area is the *p*-value. If it less than α, we reject H_0. In Figure 4 there are two example *p*-values shown. Under the distribution shown with $\mu = 0$ and $\alpha = 0.05$.

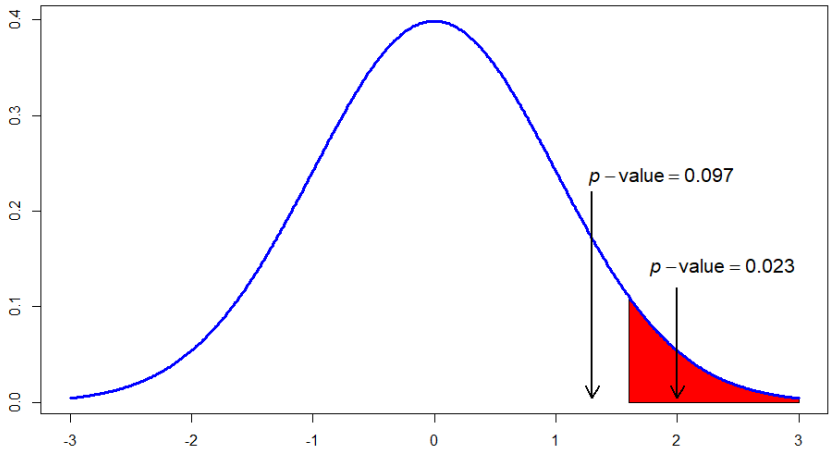


Figure 4: An illustration of *p*-values and their relationship to $\alpha = 0.05$

When the *p*-value is equal to or greater than α we fail to reject H_0. This does not mean that the test proved that the null hypothesis is true or that the alternate hypothesis is false. It only signifies that the data gathered was not compelling enough for us to conclude that it came from the alternate state.

In hypothesis testing, an ambiguous data set or even a slightly convincing data set will not allow us to reject the null with any meaningful amount of confidence and we continue to define our SUT under the premise of the null hypothesis. With lack of clear evidence to the contrary, the presumption of "no difference" stands. However, this is not an ideal method for proving that system performance has **not** changed. In this context, we would want to fail to reject the null hypothesis; however, this is a weak conclusion. Failing to reject the null hypothesis just means that we did not have sufficient data to conclude the alternative hypothesis was true. To do this we need to revise our theory, reverse our hypotheses, and adjust our definitions. We need to presume the data sets come from different populations, unless the data convincingly show they are the same.

Hypothesis Testing for Equivalence

We now consider the situation where we want to show that two population parameters are equivalent. The theory which prompted the hypothesis testing has the reverse premise, and consequently many of the metrics now have a reverse meaning, but the underlying hypothesis testing procedure remains the same.

An Equivalence Theory

The test process is still initiated with a theory, but in equivalence testing we have only one basic premise: "Despite the changes made to the SUT, the output remains unchanged." This is useful in the T&E for situations where there is no need for the SUT to improve performance, only sustain previously demonstrated performance but in some slightly altered configuration (e.g., an updated system that is cheaper or simpler to build, but performs the same as the original system).

The Null and Alternate Hypothesis

In equivalence testing, we maintain the same philosophy used in traditional hypothesis testing; the null hypothesis represents the status quo. However, that status quo belief is now that the system performs different when the inputs are different. We assume that if we *change* the system then we *change* its performance. In a physical system this could be because a part has been manufactured from a new material and is now lighter or heavier or more or less flexible. In a digital system this could be because a new algorithm in the code or perhaps a new coding language was used to compute some part of the output. In any case, something contributing to the SUT has changed, but testers desire to show that the performance has not changed. This desire means that the alternate hypothesis, still the theory that we want to show is correct, is that the SUT's performance is the same. The pair of hypotheses are written as:

$$H_o: \mu_1 \neq \mu_2$$
$$H_1: \mu_1 = \mu_2$$

Type I Risk

In the case of equivalence testing we continue to use α to indicate the probability of a Type I error, but the implications of that error have changed. Type I error is defined generically as the probability we

reject the null hypothesis when the null hypothesis is actually true. In the equivalence testing context, type I error rate is the probability we conclude the means are the same when they are actually different.

In equivalence testing the meaning and mathematical principles behind α and β have not changed, but the affected party does change. In the scenario we have changed something in the manufacturing process with the desire that the product meets the same standards. Now our hypothesis statements are:

H_o: $\mu_1 \neq \mu_2$: The product's performance is different than it was before
H_1: $\mu_1 = \mu_2$: The product's performance is the same as it was before

If a type I error is made the test team will reject the null even though it is true. That scenario is now a risk to the consumer, who expects to receive a product that performs just as well, but does not.

Because of this change in the party affected some literature uses the term β is used instead of α to show denote Type I error in equivalence testing. In this document we continue to use α.

Type II Risk

Continuing in the quality control testing scenario, the probability of a Type II error is also called the "consumer's risk" in traditional hypothesis testing. Recall that a type II error occurs when we fail to reject the null hypothesis when the alternative hypothesis is actually true. In the traditional hypothesis testing scenario, this would imply for example, that we miss a significant change in the system. If a Type II error is made then a substandard batch of the product will be erroneously determined to be within standard and delivered to the consumer. In equivalence testing, a Type II error is made the test team concludes that the system is different (null hypothesis) when it is truly equivalent (alternative hypothesis). That is a risk to the producer.

The Equivalence Acceptance Criterion

Whenever comparing two continuous results one can always conclude they are different if they are measured to a precise enough degree. At some point the difference becomes an academic difference but not a practical difference. This sensitivity level is controlled in equivalence testing by deriving the equivalence acceptance criterion, Δ (delta).

In some fields there may be industry standards to use as a guide for Δ. If not, the equivalence acceptance criterion is best developed from a consensus among stakeholders such as decision makers, system engineers, operators, and subject matter experts (SMEs). Deciding what difference in SUT performance is "good enough" and determining Δ for the equivalence test is not a trivial task and adequate time should be allotted for this in the planning phase. The value of Δ should be based on how big a difference is practically important to consider the means not equivalent. Δ is similar to the difference to detect when we size a traditional designed experiment. Once Δ is derived, the hypothesis can be stated as:

H_o: $|\mu_1 - \mu_2| > \Delta$: The difference in means is greater than the equivalence acceptance criterion
H_1: $|\mu_1 - \mu_2| \leq \Delta$: The difference in means is less than or equal to Δ (the means are equivalent)

One-Sided Equivalence Testing

A one-sided test is appropriate if the test team only cares if the response differs in one direction. For example, if a new rubber compound is used to make a tire with a mileage guarantee the producers may only care if the tire gets fewer miles of use. This is often called a test for noninferiority (Pardo, 2014). In these situations, the null hypothesis is often termed the "inferiority" hypothesis and the alternate is the "noninferiority" hypothesis. The following equation is used to compare the *t*-ratio to the critical value. Note that we no longer use μ, because it represents the true population parameter, which we will likely never know. Instead we use $\bar{y}$, which denotes the sample mean. The critical value $(t_{\alpha,\mathrm{df}})$ can be calculated with software or looked up in statistics tables.

$$\text{We reject } H_0 \text{ if: } \frac{|\bar{y}_1 - \bar{y}_2| + \Delta}{SE} > t_{\alpha,n_1+n_2-2}$$

The standard error (SE) of two different means can be calculated with the following equation where s^2 is the sample variance and n is the sample size.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two One-Sided Tests (TOST)

In equivalence testing, there is no two-sided test. Instead, we conduct two noninferiority tests from both sides. We expand the hypotheses statements as shown in Figure 5:

$H_o: \lvert\mu_1 - \mu_2\rvert > \Delta$	$H_o: \mu_1 - \mu_2 > \Delta$	Null hypothesis expanded for the upper side test
	$H_o: \mu_1 - \mu_2 < -\Delta$	Null hypothesis expanded for the lower side test
$H_1: \lvert\mu_1 - \mu_2\rvert \leq \Delta \longrightarrow H_1: -\Delta \leq \mu_1 - \mu_2 \leq \Delta$		The alternate hypothesis is expanded

Figure 5: Expansion of hypothesis for two one-sided tests

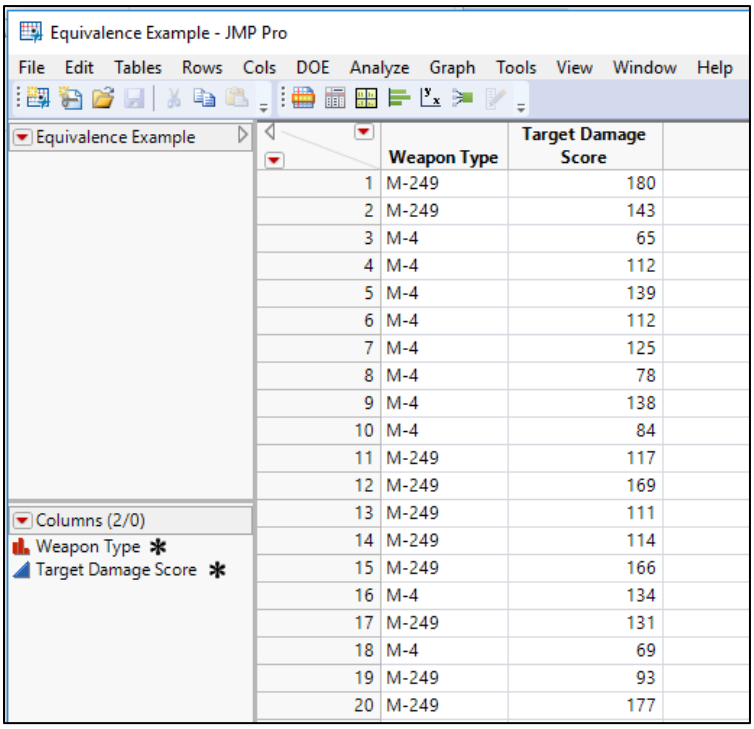
When the hypotheses are expanded to create the test statistics we get the following criterion.

$$\text{We reject } H_0 \text{ if: } \frac{|\bar{y}_1 - \bar{y}_2| + \Delta}{SE} > t_{\alpha,n_1+n_2-2} \quad \text{AND} \quad \frac{|\bar{y}_1 - \bar{y}_2| - \Delta}{SE} < -t_{\alpha,n_1+n_2-2}$$

Note that the two means are considered equivalent if, and only if, both null hypotheses are rejected. We must show that difference in the means is less than Δ and greater than −Δ to show that the means are equivalent; it cannot be true in just one direction. It is also prudent to point out that this is different than a traditional two-sided hypothesis test. When conducting a two-sided test with a Type I probability of α, we use $\alpha/2$ to compare to the *p*-value on each side. In this situation we use the full value of α on both tests because they are both one-sided.

Example Problem 1

Suppose we are analyzing the effectiveness of the door gunner in a helicopter. It is normal for the door gunner to use the M-249 but a unit wants to instead arm them with the M-4. The M-249 is a machine gun capable of a higher volume of fire, but it is heavier, harder to move, and more prone to jam. Can the M-4 provide equivalent combat power? Further suppose a test was conducted and the results were loaded into JMP statistical software as displayed in Figure 6. The test team determined that the M-4 was equivalent to the M-249 if the target damage score was within 20 points.



	Weapon Type	Target Damage Score
1	M-249	180
2	M-249	143
3	M-4	65
4	M-4	112
5	M-4	139
6	M-4	112
7	M-4	125
8	M-4	78
9	M-4	138
10	M-4	84
11	M-249	117
12	M-249	169
13	M-249	111
14	M-249	114
15	M-249	166
16	M-4	134
17	M-249	131
18	M-4	69
19	M-249	93
20	M-249	177

Figure 6. Data for 20 trials of weapon effectiveness

The reader can see the weapon used and the damage score for each of the 20 trials. This example is small enough that we can go through the calculations before we use JMP. Three basic statistics are calculated and shown in Table 1.

Table 1. Statistics for TOST example

Weapon	Sample size	Sample mean	Sample variance
M-249	$n_1 = 10$	$\bar{y}_1 = 140.1$	$s_1^2 = 981.21$
M-4	$n_2 = 10$	$\bar{y}_2 = 105.6$	$s_2^2 = 849.6$

Using the values calculated in Table 1 we can calculate the test statistic and critical value. First for the upper end of the curve.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{981.21}{10} + \frac{849.6}{10}} = \sqrt{183.0811} \approx 13.53$$

$$\text{We reject } H_0 \text{ if: } \quad \frac{|\bar{y}_1 - \bar{y}_2| + \Delta}{SE} > t_{\alpha, n_1+n_2-2}$$

$$\frac{|140.1 - 105.6| + 20}{13.53} > t_{0.05,18}$$

$$\frac{|34.5| + 20}{13.53} > t_{0.05,18}$$

$$4.03 > t_{0.05,18}$$

$$4.03 > 1.73 \xrightarrow{yields} TRUE$$

The statement is true so we reject H_0 in the first one-sided test. Next, we conduct the second test.

$$\frac{|\bar{y}_1 - \bar{y}_2| - \Delta}{SE} < -t_{\alpha, n_1+n_2-2}$$

$$\frac{|140.1 - 105.6| - 20}{13.53} < -t_{0.05,18}$$

$$1.07 < -1.734 \xrightarrow{yields} FALSE$$

The statement is false and we fail to reject H_0 in the second one-side test. Therefore, we also fail to reject the overall null hypothesis that the M-4 is not equivalent to the M-249. This is not a surprising outcome because our initial difference in means was 34.5.

To conduct the same test in JMP we start with the data entered in Figure 6. To conduct the equivalence test, click on the **Analyze** tab at the top of the window and choose **Fit Y by X** from the drop down as shown in Figure 7.

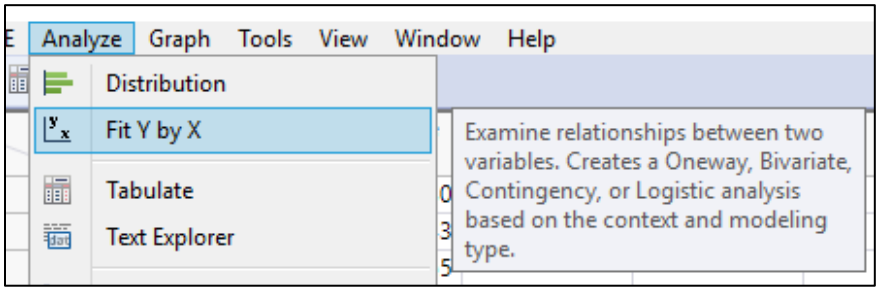


Figure 7. Screenshot from JMP while conducting TOST

Figure 8 shows the window that will then open. When it does, the factor and response must be selected and then select the **OK** button.

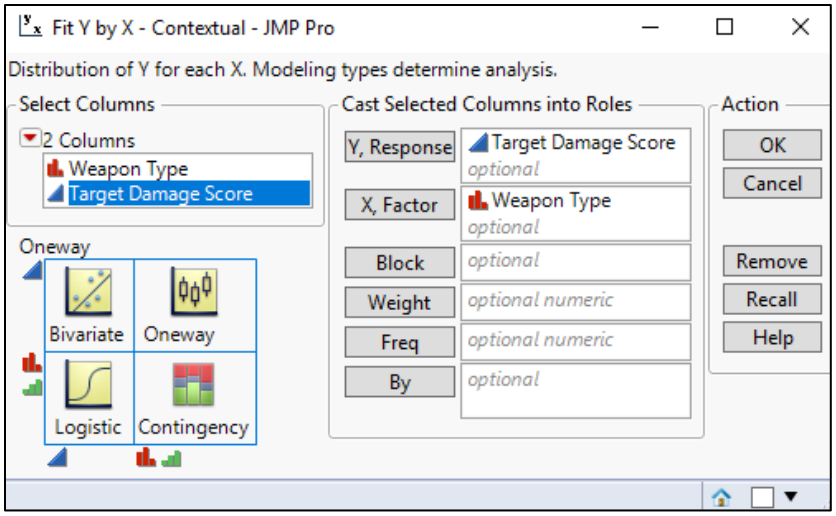


Figure 8. Screenshot from JMP while conducting TOST

Once the **OK** button is selected JMP will produce a dot plot similar to Figure 9. A visual inspection of the dots shows us that the M-4 scores tend to be lower than the M-249 scores.

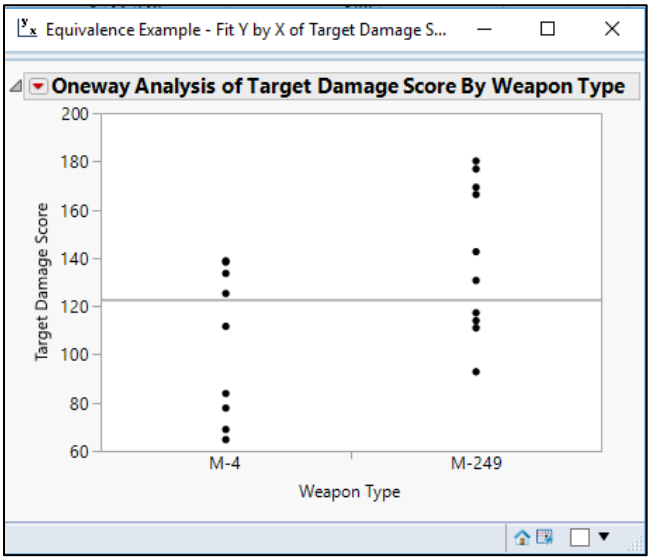


Figure 9. Screenshot of dot plot from JMP while conducting TOST

The final step to initiate the equivalence test is to select the red triangle in the upper left corner and select **Equivalence Test** from the drop-down menu as seen in Figure 10.

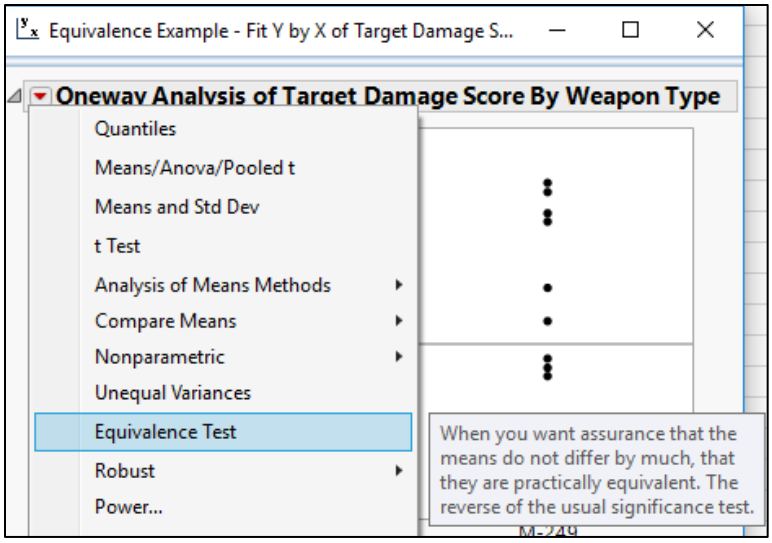


Figure 10. Screenshot of drop down menu from JMP

After **Equivalence Test** is chosen the Equivalence Acceptance Criteria needs to be set in the window that opens. It will look like Figure 11.

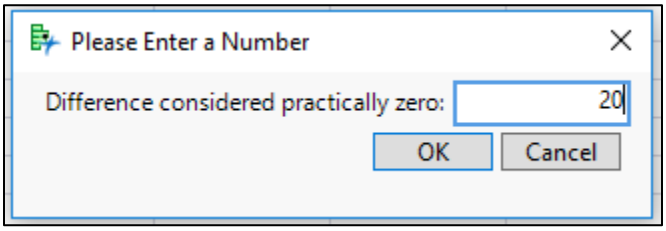


Figure 11. Screenshot of Δ selection

The test is run and a report will be displayed as Figure 12 shows. There are small differences in some of the values from the earlier calculations due to rounding errors, but the result is the same. In this case the upper threshold caused the TOST to fail. The diagram on the right displays that. The red vertical line is the actual difference in means. The two distributions from each one-sided test are centered on Δ and −Δ. The blue shaded area on the right hand *t*-distribution illustrates the upper threshold *p*-value. The blue area represents an answer to the question, "*If the true difference is 20, or higher, what is the probability we calculate 34.65 or less?*" The answer is 85.29%.

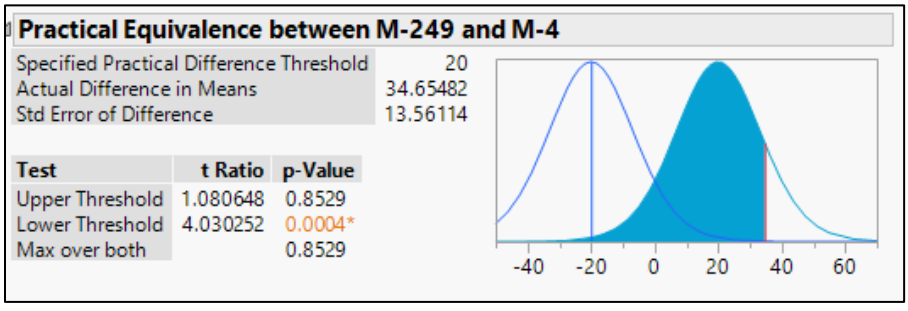


Figure 12. Screenshot of the TOST report from JMP

To adjust the value of α select the red triangle in the upper left again. The same drop-down menu will materialize but this time select **Set α Level** as shown in Figure 13. Then choose a value provided or select **Other** to write in a different specific value.

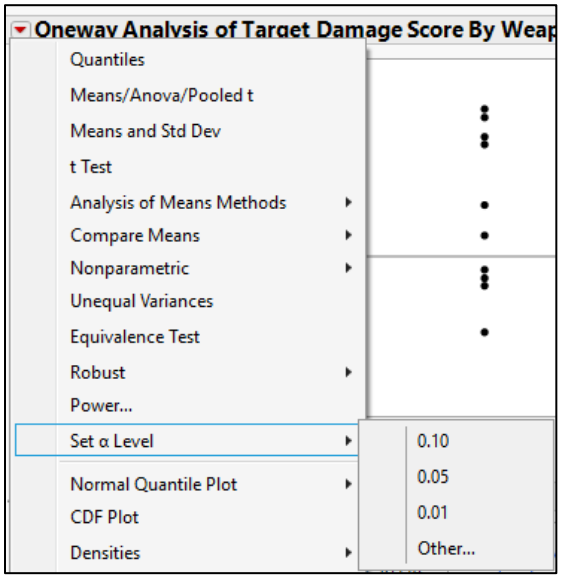


Figure 13. Screenshot of drop down menu

Example Problem 2

Another example is shown in Figures 14 and 15. In this example the comparison is between some real and simulated shots. This is a common scenario when attempting to validate modeling and simulation (M&S) data. Model validation is an ideal application for equivalence testing. The equivalence acceptance criterion for this example is 3.

	Source	Miss Distance			
1	Sim	26.28	16	Real	23.74
2	Sim	19.2	17	Real	20.26
3	Sim	26.39	18	Real	26.45
4	Sim	27.52	19	Real	22.14
5	Sim	20.86	20	Real	25.27
6	Sim	30.18	21	Real	27.95
7	Sim	27.2	22	Real	26.27
8	Sim	21.75	23	Real	25.79
9	Sim	25.5	24	Real	24.13
10	Sim	24.48	25	Real	22.56
11	Sim	24.51	26	Real	23.88
12	Sim	22.7	27	Real	23.79
13	Sim	25.58	28	Real	21.47
14	Sim	25.21	29	Real	17.98
15	Sim	20.37	30	Real	22.99

Figure 14. Screenshot of example 2 data in JMP

The results are shown in Figure 15. In this example we reject the null hypothesis that the mean miss distance is different between the model and live test data since both *p*-values for the one-sided tests are small (*p*-values < 0.05). We conclude that the means are equivalent, providing us with strong evidence that the model matches the live data overall. The calculated difference between means was 0.87, so it is not surprising that we reject the null hypothesis.

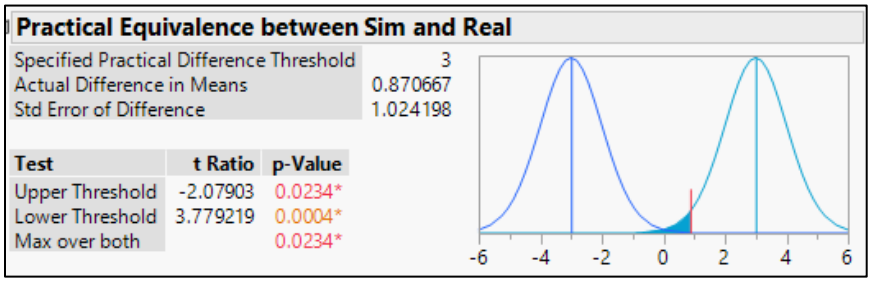


Figure 15. Screenshot of results of second example

If α is changed to $\alpha = 0.01$ then the sensitivity is changed and the result is different. At $\alpha = 0.01$ we fail to reject the null hypothesis because the upper threshold *p*-value, which is 0.0234, does not change, but it is now greater than α. Note that it is not good test practice to change α after the test-it is done here only to show the affect α has on the result.

Conclusion

Equivalence testing is a powerful, but under-utilized STAT tool which allows the tester to determine the equivalence of two groups beyond a simple difference of means. The logic and fundamental process of hypothesis testing does not change when adjusting to equivalence testing, but the theory behind the hypothesis does. Little time in this best practice was spent on the selection of the equivalence acceptance criteria, but it is not a trivial task and must be selected through an informed and deliberate method. When conducting an equivalence test the tester must also determine if it is an inferiority test or TOST. It has proven to be a useful tool in the test process because it allows the testers to examine the difference between two SUTs and account for the variation within each one.

References

Kensler, Jennifer. "Statistical Hypothesis Testing." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 30 Aug 2013.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017.

Pardo, Scott. *Equivalence and Noninferiority Tests for Quality, Manufacturing and Test Engineers*. Boca Raton: CRC Press, 2014. Print.

Ramert, Aaron. "Understanding the Signal to Noise Ratio in Design of Experiments." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 27 Aug 2018.

Vickers, Andrew. *What Is a P-value Anyway?: 34 Stories to Help You Actually Understand Statistics*. Boston: Pearson Education, 2010. Print.