Equivalence Testing

You are looking to replace an older measurement system with a new one. You want to know if the new measurement system is the "same" or "equivalent" to the old one. Note that you don't want to know if they are different. You want to know if they are the same. In this case, the "same" means that the difference between the two measurement systems are within a predefined margin of error. If the difference is within that margin of error, then the observed differences are not meaningful in the practical sense.

There are several equivalence testing methods including one sample, two sample and paired samples equivalence testing. This publication introduces equivalence testing, in particular the two sample equivalence test

In this publication:

- Introduction
- Example Data
- Calculations for Two Sample Equivalence Test
- Conclusions from the Test
- Other Alternate Hypotheses
- Comparison to the Two Sample t Test
- Other Equivalence Tests
- Summary
- Quick Links

Introduction

Suppose you work in a heat treating facility. You routinely measure the hardness of steel samples using a Rockwell hardness tester. The tester you are using has been around for a number of years and you are having more and more problems with it. You ordered a new hardness tester that has arrived and been set up. You want to know if the new hardness tester gives the same results as the old hardness tester.

To use equivalence testing, you must decide on the range within which the differences in the two testers are trivial or not of practical significance. This is your margin of error or equivalence interval. Suppose you decide that it is practically insignificant if the difference in the means of the two testers is within ± 0.5 Rockwell hardness.

You are performing a hypothesis test with equivalence testing. The null hypothesis (H_0) and the alternate hypothesis (H_1) for this example are given below.

H_0: The difference in the means of the two testers is outside the equivalence interval

H_1: The difference between the means is inside the equivalence interval and the means are equivalent

© 2024 BPI Consulting, LLC
www.spcforexcel.com

To perform equivalence testing, you collect your samples and run the samples in the old tester and the new tester. You then perform the calculations as shown below. Two key calculations are key to interpreting the results.

One is a confidence interval that is calculated. This interval is a range that contains the difference between the two means. If the confidence interval fits entirely within the equivalence interval (± 0.5 in this example), you conclude that the two testers are the same. If the confidence interval does not fit entirely within the equivalence limits, you conclude that the two tests are considered different.

The other key calculation involves the p-value. To perform this calculation, you have to decide on the value of alpha. Alpha (α) is called the significance level. Typical values of alpha are 0.05 and 0.10. The significance level gives the risk of accepting the null hypothesis when you should not. A value of 0.05 means that risk is 5%. This is also related to the confidence interval. If alpha = 0.05, the confidence interval is a 95% confidence interval.

A t-value is calculated. The p-value represents the probability of getting that t statistic if the null hypothesis is true. If the value of p is small, then we conclude that the probability of getting that t-value, if the null hypothesis is true, is small and reject the null hypothesis. If the p-value is large, we conclude that the probability of getting that t-value if the null hypothesis is true is large and we accept the null hypothesis.

We examine how this works with the example below.

Example Data

To test this, you take 50 samples at random from your process and test 25 of them in the old hardness tester and test 25 of them in the new hardness tester. The results are given in Table 1.

Table 1: Old and New Hardness Testers Results

Old	New	Old	New
30.2	30.3	31.1	28.7
30.0	29.4	28.8	31.1
29.4	29.1	29.7	29.6
30.4	29.4	28.2	29.7
29.9	28.2	29.7	27.7
29.5	29.0	28.7	29.6
30.2	29.0	29.1	29.8
30.7	29.7	30.4	31.0
29.8	30.0	29.9	28.3
29.8	30.1	29.7	29.5
29.1	29.7	29.4	30.2
31.4	28.1	30.3	29.5
29.9	29.1		

The data will be used to show how the calculations for the two sample equivalence test are done.

© 2024 BPI Consulting, LLC
www.spcforexcel.com

Calculations for Two Sample Equivalence Test

The SPC for Excel software was used to analyze the data. How those results were calculated is given below. The first step is to calculate the following sample statistics.

Sample Statistics

Variable	Old (Test)	New (Reference)
Sample Size	25	25
Average	29.81	29.43
Standard Deviation	0.725	0.829
SE Mean	0.145	0.166

The statistics calculated include the sample size, the average (the mean), the standard deviation, and the standard error the mean (SE Mean). The SE Mean estimates the differences between means you would obtain if you repeated taking samples from the same population. It is given by the square root of the standard deviation squared divided by the sample size:

$$SE\ Mean = \sqrt{\frac{s^2}{n}}$$

where s = the standard deviation and n = sample size.

The next step is to calculate the difference statistics:

Difference Statistics

Difference = Old Average - New Average	0.380
SE	0.220

The difference is simply the difference in the old tester average and the new tester average. SE is the standard error of the difference and is given by:

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

Subscript 1 refers to the test sample, while subscript 2 refers to the reference sample (old and new, respectively). The standard error of the difference estimates the variation in the differences between the two means if repeated samples are taken from the same populations and the difference in the averages calculated.

The next step is to compare the equivalence limits (or interval) to the calculated confidence limit for the difference in the averages.

The equivalence limits are what you selected, in this example +/- 0.5.

© 2024 BPI Consulting, LLC
www.spcforexcel.com

	Lower	Upper
Equivalence Limit	-0.5	0.5
95% Confidence Interval	0.000	0.749

To calculate the confidence interval, we need to select a value of alpha. We will use 0.05. The confidence interval gives a range of values for the difference between the two averages. The lower limit is the lowest difference you would expect while the upper limit is the largest difference you would expect.

You can perform two-sided or one-sided confidence intervals. We will do a two-sided confidence interval here. The equations for calculating the two-sided confidence intervals are given below.

$$\text{Lower Confidence Limit} = \text{minimum of } (C, D_L)$$

$$\text{Upper Confidence Limit} = \text{maximum of } (C, D_u)$$

where

$$C = (LEL+UEL)/2$$

$$D_L = \text{Difference in averages} - t(SE)$$

$$D_u = \text{Difference in averages} + t(SE)$$

where LEL is the lower equivalence limit (-0.5 in this example), UEL is the upper equivalence limit (0.5 in this example) and t is the t value from the t distribution. The t value depends on 1 – alpha and the degrees of freedom.

The degrees of freedom (df) is given by the following:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_1)^2}{n_2 - 1}}$$

If you enter the values in the above equation and round down, you will get df = 47. You can use the Excel function T.INV to find the value of t:

$$t = \text{T.INV}(1 - \text{alpha}, df) = \text{T.INV}(0.95, 47) = 1.678$$

Now we can calculate the confidence interval.

$$C = (LEL+UEL)/2 = (-0.5 + 0.5)/2 = 0$$

$$D_L = \text{Difference in averages} - t(SE) = 0.38 - 1.678(0.220) = 0.011$$

$$D_u = \text{Difference in averages} + t(SE) = 0.38 + 1.678(0.220) = 0.749$$

and

$$\text{Lower Confidence Limit} = \text{minimum of } (C, D_L) = \text{minimum of } (0, 0.011) = 0$$

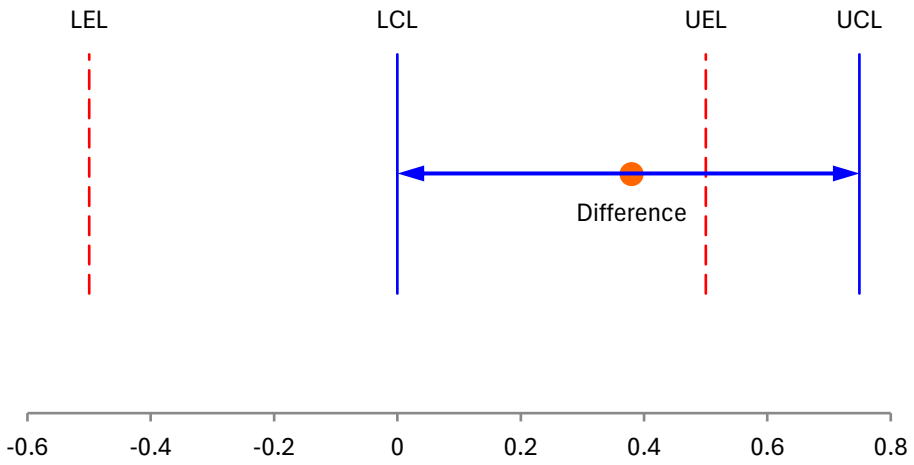
© 2024 BPI Consulting, LLC
www.spcforexcel.com

Upper Confidence Limit = maximum of (C, D_u) = maximum of (0, 0.749) = 0.749

So, the confidence interval is 0 to 0.749.

This gives you the first indication if the two testers are the "same" when you compare the confidence interval to the equivalence interval. If the confidence interval fits within the equivalence interval, you conclude that the two tests are the same. Figure 1 compares the two.

Figure 1: Confidence Interval Compared to Equivalence Interval



The figure plots the difference in the averages. The blue line shows the confidence interval. You can see it does not fit within the equivalence interval. So, it does not look like the two testers are the same.

The next step is to test the null hypothesis. The output from the SPC for Excel software is shown below.

Hypothesis

H_0: Difference <= LEL or Difference >= UEL

H_1: LEL < Difference < UEL

H_0 Tests	df	t-Value	p-Value
Difference <= LEL	47	3.9972	0.0001
Difference >= UEL	47	-0.5451	0.2941

As shown in the results, there are two null hypothesis tests. One hypothesis is that the difference in the two averages is less than LEL; the other is that the difference is greater than the UEL.

© 2024 BPI Consulting, LLC
www.spcforexcel.com

The equations for the t-values are given below, with subscript 1 for the hypothesis difference ≤ LEL and subscript 2 for the hypothesis difference ≥ UEL.

$$t_1 = (\text{Difference} - \text{LEL})/SE = (0.38 - -0.5)/0.22 = 3.997$$

$$t_2 = (\text{Difference} - \text{UEL})/SE = (0.38 - 0.5)/0.22 = -0.5451$$

Now, we calculate the p-value for each t-value. Remember the p-value is the probability of getting the t-value if the null hypothesis is true. We can use two of Excel's t distribution functions to determine the p-value as shown below:

$$p_1 = \text{T.DIST.RT}(t_1, df) = \text{T.DIST.RT}(3.997, 47) = 0.0001$$

$$p_2 = \text{T.DIST}(t_1, df, \text{TRUE}) = \text{T.DIST}(3.997, 47, \text{TRUE}) = 0.2941$$

Both these p-values must be less than the value of alpha we selected (0.05). One of them is not, so we cannot assume that the two testers are the same.

Conclusions from the Two Sample Equivalence Test

Based on the analysis above, we conclude that the old tester and the new tester are not equivalent. The two points that allow us to reach this conclusion are the following:

1. The confidence interval does not fit entirely within the equivalence interval (see Figure 1).
2. The maximum p-value for the null hypothesis is greater than the value of alpha selected (0.05 in this example).

Other Alternate Hypothesis

The example above used the following alternate hypothesis:

$$H_1: \text{LEL} < \text{Difference in the averages} < \text{UEL}$$

You can also use any of the following alternative hypotheses depending on what you are trying to find out:

$$H_1: \text{Test Average} > \text{Reference Average}$$

$$H_1: \text{Test Average} < \text{Reference Average}$$

$$H_1: \text{Difference} > \text{LEL}$$

$$H_1: \text{Difference} < \text{UEL}$$

Comparison to the Two Sample t Test

The two sample equivalence test appears to be very similar to the two sample t test. The two sample t test is used to determine if there is a statistically significant difference between the averages of two populations. The null and alternate hypotheses for the two sample t test are:

$$H_0: \mu_1 = \mu_2$$

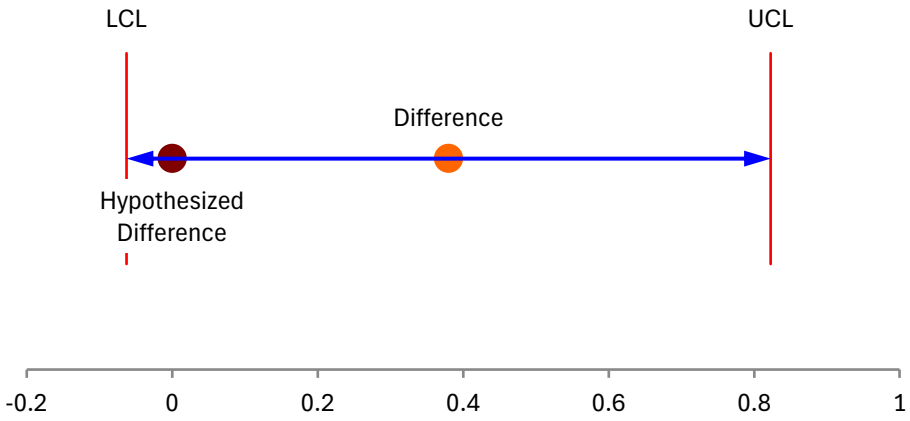
$$H1: \mu_1 \neq \mu_2$$

© 2024 BPI Consulting, LLC
www.spcforexcel.com

where μ represents the average. If the p-value for the two sample t test is less than alpha, you conclude that the two averages are different.

The data in Table 1 was analyzed using the t test for the difference in two means with the SPC for Excel software. Figure 2 plots difference in the two averages with the confidence interval.

Figure 2: t Test for the Difference in Means



You can see from Figure 2, that the hypothesized difference of zero lies with the confidence interval, so we conclude that the difference in the old tester and new tester averages can be zero – that the testers are the same. The p-value in this example is 0.091, which is greater than 0.05. So, more evidence that the two tests have the same average.

With the equivalence testing, we defined a range of values for the difference in the averages that was not practically important. It didn't matter to us if the difference ranged from -0.5 to 0.5. The difference of 0.38 falls in that range, so from an equivalence point of view, it is not significant.

If you have a range that is not significantly important to you, you should consider using the equivalence test instead of the standard t test.

Other Equivalence Tests

There are other equivalence tests than the two sample test. Two of these are:

1. One-Sample Equivalence Test: this test is used to determine if the average of a population is close enough to a target value.
2. Paired Samples Equivalence Test: this test is used to determine if the mean of a test population is close enough to the mean of a reference population using paired samples (each sample tested in each population).

In these tests, the "close enough" refers to the equivalence interval you select.

© 2024 BPI Consulting, LLC
www.spcforexcel.com

Summary

This publication introduced equivalence tests with a focus on the two sample equivalence test. In equivalence testing, the user selects a range of values that the user considers are not important in the practical terms. This is called the equivalence interval. In the two sample case, if the confidence interval falls within the equivalence interval, the two populations are considered the same. The analysis in this publication was done using the SPC for Excel software.

Quick Links

Visit our home page

SPC for Excel Software

Download SPC for Excel Demo

SPC Training

SPC Consulting

SPC Knowledge Base

Ordering Information

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.

Sincerely,

Dr. Bill McNeese

© 2024 BPI Consulting, LLC
www.spcforexcel.com