A Method for Determining Equivalence in Industrial Applications

By: Scott J. Richter and Carrie Richter.

Richter, S. J. & Richter, C. (2002). A method for determining equivalence in industrial applications. *Quality Engineering*, 14(3), 375-380.

Abstract:

This article considers the problem of determining whether the results of measurements at two different measurement times are equivalent. Quite often in an industrial quality assurance experiment, the goal is to provide evidence of equivalence, rather than difference. Although determining equivalence has long been a staple in biological and chemical applications, it is often useful in industrial situations, such as comparing the characteristics of products measured at different points in time, or produced using different formulations. Methods for assessing equivalence, however, are rarely taught to engineers and other scientists. The purpose of this article is to demonstrate a procedure for assessing equivalence, as well as to demonstrate the inappropriateness of a method commonly used in these situations. The authors suggest that equivalence testing methodology should be among the statistical tools at the disposal of quality and process control engineers.

Keywords: Equivalence test | Test reduction | Quality assurance

Article:

INTRODUCTION

Determining equivalence between two parameters is frequently of interest in industrial applications, especially when evaluating product and process quality. One example concerns a product that is measured regarding a certain characteristic at different points in time, and it is of interest to determine if the measurements at these points in time yield equivalent information about the product. This article addresses the following problem: A particular type of adhesive tape is measured immediately after production and again 24 h later to observe adhesive properties. The question here is whether the measurements of the adhesive properties are equivalent at the two points in time. If there is evidence that these measurements are equivalent, resources can be saved by eliminating one of the measurement times.

Unfortunately, appropriate methods for determining equivalence are usually absent from the curriculum of statistics courses taught to engineers (1), and the literature makes virtually no mention of these procedures. The result is that practitioners usually resort to ad hoc methods instead. The most common approach to assess evidence of equivalent means, for example, is to perform a test of the hypotheses $\mathrm{H}_0 : \mu_1 = \mu_2$ versus $\mathrm{H}_1 : \mu_1 \neq \mu_2$; and conclude "equivalence" if H_0 is not rejected. In this article, we will first examine this standard testing approach and point out the problems with its use for assessing equivalence. Next, we will demonstrate a more appropriate method which uses confidence intervals to assess equivalence and demonstrate its use on the equivalence problem described.

COMPARISON OF METHODS FOR ASSESSING EQUIVALENCE

Problem

In the manufacturing process of a certain type of adhesive tape, a sample of thickness measurements is taken immediately after production ("unconditioned") and another sample 24 h later ("conditioned"). If the results of these two samples tend to be "equivalent," then an argument could be made to eliminate one of the sampling episodes, saving time and resources. Let μ_{uc} be the mean thickness for tape samples measured immediately after production and μ_c be the mean thickness for tape samples measured 24 h after production. One hundred eight measurements taken immediately after production had a mean of 11.153 mils and standard deviation of 0.512 mils. One hundred seventy-two measurements taken 24 h after production yielded a mean thickness of 11.256 mils and a standard deviation of 0.651 mils. This sample information was to be used to determine if the mean thickness for unconditioned tape samples was equivalent to that for conditioned tape samples. Two possible methods of addressing this problem are considered: the standard hypothesis testing approach and the equivalence testing approach.

Standard Hypothesis Testing Approach

Test the hypotheses $\mathrm{H}_0 : \mu_1 = \mu_2$ versus $\mathrm{H}_1 : \mu_1 \neq \mu_2$. If insufficient evidence exists to reject the null hypothesis, declare the two means to be "equivalent." Note that to declare the means equivalent involves the decision to "accept" the null hypothesis, and the probability of incorrectly accepting H_0 (i.e., making a type II error) is unknown. Unfortunately, insufficient evidence to reject the null hypothesis is not the same as evidence that the null hypothesis is true. It may be argued, however, that $\mu_{\mathrm{uc}} = \mu_c$ is not strictly required— only that $|\mu_{\mathrm{uc}} - \mu_c| < \theta$ (i.e., that μ_{uc} and μ_c are "close enough" to be considered equivalent for practical purposes). Does that mean that this is a reasonable method for determining equivalence? The answer is No. It has been shown (2) that this method has the undesirable property of penalizing higher precision. In other words, a sample mean difference which is declared "equivalent" for a given sample size may be declared "not equivalent" for a larger sample size! This should not be surprising, because the standard testing procedure is set up to determine whether the sample evidence supports the alternative hypothesis. Thus, the method is merely behaving as it should if the observed difference were evidence of nonequivalence, rather than of equivalence: More precision should eventually lead to a conclusion that the means are different. However, this method is less than desirable if the goal is to establish equivalence.

Equivalence Testing Approach

The most common method for determining equivalence is an approach known as the two one-sided tests (TOST) approach, generally attributed to Westlake (3) and Schuirmann (4). First, determine an interval (θ_1, θ_2) such that if $\theta_1 < \mu_{uc} - \mu_c < \theta_2$, the means can be considered equivalent. Then, for a given significance level, $\alpha,$ declare the means equivalent if a 100 $(1-2\alpha)\%$ confidence interval for $\mu_{uc} - \mu_c$ falls completely between θ_1 and θ_2. Otherwise, conclude that the evidence is insufficient to declare the means equivalent. The method is computationally identical to testing the following sets of hypotheses:

$$\begin{aligned} &\mathrm{H}_{01} : \mu_1 - \mu_2 \leq \theta_1 \text{ versus } \mathrm{H}_{11} : \mu_1 - \mu_2 > \theta_1 \\ &\mathrm{H}_{02} : \mu_1 - \mu_2 \geq \theta_2 \text{ versus } \mathrm{H}_{12} : \mu_1 - \mu_2 < \theta_2 \end{aligned}$$

To see this, consider the case in which the samples can be treated as independent random samples from Normally distributed populations with common variance.

Then, H_{01} will be rejected if

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \theta_1}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} > t_{1-\alpha}(n_1 + n_2 - 2\mathrm{df}),$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and $t_{1-\alpha}(n_1 + n_2 - 2\mathrm{df})$ is the $1 - \alpha$ quantile of the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom (df), or, equivalently, if

$$(\bar{Y}_1 - \bar{Y}_2) - t_{1-\alpha}(n_1 + n_2 - 2\mathrm{df})\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} > \theta_1.$$

Similarly, H_{02} will be rejected if

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \theta_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} < -t_{1-\alpha}(n_1 + n_2 - 2\text{df})$$

or, equivalently, if

$$(\bar{Y}_1 - \bar{Y}_2) + t_{1-\alpha}(n_1 + n_2 - 2\text{df})\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} < \theta_2.$$

Thus, both H_{01} and H_{02} will be rejected at level of significance α if the $100(1-2\alpha)\%$ confidence interval

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{1-\alpha}(n_1 + n_2 - 2\text{df})\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is contained entirely between θ_1 and θ_2. Because both null hypotheses can be rejected at the level of significance and thus both $\theta_1 < \mu_1 - \mu_2$ and $\mu_1 - \mu_2 < \theta_2$ it can be concluded that $\theta_1 < \mu_1 - \mu_2 < \theta_2$. So, alternatively, the above hypotheses can be stated more succinctly for the equivalence problem:

$\mathrm{H}_0 : \mu_1 - \mu_2 \leq \theta_1 \;\text{ or }\; \mu_1 - \mu_2 \geq \theta_2$ versus $\mathrm{H}_1 : \theta_1 < \mu_1 - \mu_2 < \theta_2$.

Because of this property, this method is often referred to as the “two one-sided tests” (TOST) procedure. Note that this has intuitive appeal immediately because, here, the evidence of the test will either support or fail to support that the means are equivalent. Thus, if the means are declared equivalent, the probability that this conclusion is false (now a type I error, or α) is known.

To illustrate this approach, we return to the problem of determining if the mean thickness of the adhesive tape at two different times can be considered equivalent. The specifications state that the means of these measurements can be considered equivalent if they differ by no more than 0.3 mil (i.e., $-0.3 \leq \mu_{\text{uc}} - \mu_c \leq 0.3$). So, we wish to test

$\mathrm{H}_0 : \mu_{\text{uc}} - \mu_c < -0.3 \;\text{ or }\; \mu_{\text{uc}} - \mu_c > 0.3$ versus

$\mathrm{H}_1 : -0.3 \leq \mu_{\text{uc}} - \mu_c \leq 0.3.$

If the samples can be treated as independent random samples from Normally distributed populations with equal variance, a 90% confidence interval for the difference in population means is given by

$$(\bar{Y}_{\text{uc}} - \bar{Y}_c) \pm t_{0.05}(n_{\text{uc}} + n_c - 2\text{df})\sqrt{S_p^2\left(\frac{1}{n_{\text{uc}}} + \frac{1}{n_c}\right)},$$

where

$$S_p^2 = \frac{(n_{\text{uc}} - 1)S_{\text{uc}}^2 + (n_c - 1)S_c^2}{n_{\text{uc}} + n_c - 2}.$$

This yields the interval $-0.1031 \pm 1.6579(0.0738)$ or -0.103 ± 0.122, so $-0.225 \leq \mu_{\text{uc}} - \mu_c \leq 0.019$

with 90% confidence. Because this interval satisfies the alternative hypothesis (i.e., the interval is contained entirely between 20.3 and 0.3), the null hypotheses can be rejected at the 5% significance level and the means can be declared equivalent. Thus, based on the sample evidence, one of the sampling times can be eliminated.

This same result would result from the standard testing approach as well, however. Because the interval includes 0, the null hypothesis cannot be rejected. To see the advantage of the equivalence testing approach, suppose we had taken two equal sized samples of 200, instead of 108 and 172, respectively, with the same resulting sample statistics $(\bar{Y}_1 = 11.153,\ \ S_1 = 0.512, \bar{Y}_2 = 11.256,\ \ S_1 = 0.651)$. With more precision, we should have even stronger evidence of equivalence, and, indeed, the 90% confidence interval now becomes

$$-0.205 \leq \mu_{\text{uc}} - \mu_c \leq -0.001,$$

which gives even more evidence to support equivalence because the interval is not only once again entirely contained between 20.3 and 0.3, but also is narrower than the previous interval. The standard testing approach, however, now concludes that the means are different, because the width of the interval has decreased and no longer includes 0. This contradictory behavior makes the standard testing approach undesirable for establishing equivalence.

IMPLEMENTATION OF THE TOST EQUIVALENCE TESTING METHOD

Suppose that it is desired to try to demonstrate that two population means $(\mu_1 \text{ and } \mu_2)$ are equivalent, and suppose that it can be assumed that $\mu_1 \text{ and } \mu_2$ can be considered equivalent as long as the difference between μ_1 and μ_2 is no more than θ (i.e., $|\mu_1 - \mu_2| \leq \theta$). Implementation of the above-described method is as follows:

1. Determine the value θ such that if $|\mu_1 - \mu_2| \leq \theta,$ the means will be considered equivalent.
2. Determine the significance level (α) of the test.
3. Gather data to calculate sample means and standard deviations.
4. Construct a $100(1-2\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.
5. If the interval is contained entirely between $-\theta$ and $+\theta,$ then μ_1 and μ_2 can be declared equivalent at the α level of significance.

Note: For the more general case that μ_1 and μ_2 are considered equivalent if $\theta_1 < \mu_1 - \mu_2 < \theta_2,$ equivalence may be declared if the interval constructed is contained entirely between θ_1 and θ_2.

Implementing the Equivalence Test Using Computer Software

Many statistical software packages, such as Minitab, SAS, SPSS, and so forth will compute a confidence interval to compare the means of two independent samples. In these cases, the method illustrated earlier can be implemented directly.

Other commonly used software, such as the Data Analysis ToolPak in Microsoft Excel, will not construct confidence intervals, but will only produce a *p*-value for a significance test. However, these *p*-values may be used to perform the equivalence test, using the fact that the equivalence test consists of two one-sided tests, both of which must be rejected at level α .

Example

We will use a small hypothetical dataset to illustrate the equivalence test using Data Analysis ToolPak in Microsoft Excel. Suppose that data from the two samples are as follows:

Sample 1	9	8	9	8	9	10	10	9	8	9		
Sample 2	9	10	8	8	9	9	10	9	8	8	8	9

and suppose that we determine that the population means are equivalent if the mean difference is no more than $\theta = 0.8.$

Step 1. If the data are not already entered into columns of the spreadsheet, enter the data from the two samples into two spreadsheet columns. In this example, we might enter

$$A1{:}A10 \leftarrow 9 \quad 8 \quad 9 \quad 8 \quad 9 \quad 10 \quad 10 \quad 9 \quad 8 \quad 9$$

$$B1{:}B12 \leftarrow 9 \quad 10 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 9 \quad 8 \quad 8 \quad 8 \quad 9$$

Then choose:
Tools
Data Analysis
t-test: Two Sample Assuming Equal Variances.
Variable 1 Range: A1: A10
Variable 2 Range: B1: B12
Hypothesized Mean Difference: 0.8.
Alpha: 0.05 (for the equivalent test to constructing a 90% confidence interval)

Step 2. Repeat Step 1, reversing the variable order (enter Variable 1 Range: B1: B12, and Variable 2 Range: A1: A10)

Step 3. In the resulting calculations, if the calculation corresponding to [$P(T \leftarrow$ t) one-tail] is less than 0.05 for both tests, then equivalence can be declared at significance level 0.05. Otherwise, equivalence cannot be concluded.

For the above data, the resulting calculations are as follows

Step 1: See Table 1.
Step 2: See Table 2.
Step 3: Because the p-values [$P(T \leftarrow$ t) one-tail], 0.027762 and 0.003772, are both less than 0.05, the population means can be declared equivalent.

FURTHER COMMENTS ON THE EQUIVALENCE TESTING PROCEDURE

An interesting property of the TOST method is that a $100(1-2\alpha)\%$ confidence interval yields an α size, not 2α size, test for equivalence (similarly, two α -level one-tailed tests combine to produce an overall α -level test). However, this property is only guaranteed for “equal-tailed” confidence intervals (5), so care must be taken to ensure that an “equal-tailed” confidence interval be used when testing for equivalence (i.e., as in the example, where a 90% confidence interval was used to produce α =0.05 level test). Berger and Hsu (5) recommend an alternative procedure which guarantees that the size of the test is α, regardless of whether the confidence interval is equal tailed or not. However, most authors continue to recommend the use of the above-described method. Hauck and Anderson (6) argue that the TOST method is easily understood by nonstatisticians who do much of the analysis and interpretation of the analysis. Meredith and Heise (7) conclude that the method of Berger and Hsu “is of questionable practical value.” The more important practical consideration is addressing the common error of equating lack of statistical significance with “no difference” (6). The TOST procedure addresses this concern and provides an easy-to-implement and easy-to-interpret procedure for assessing equivalence.

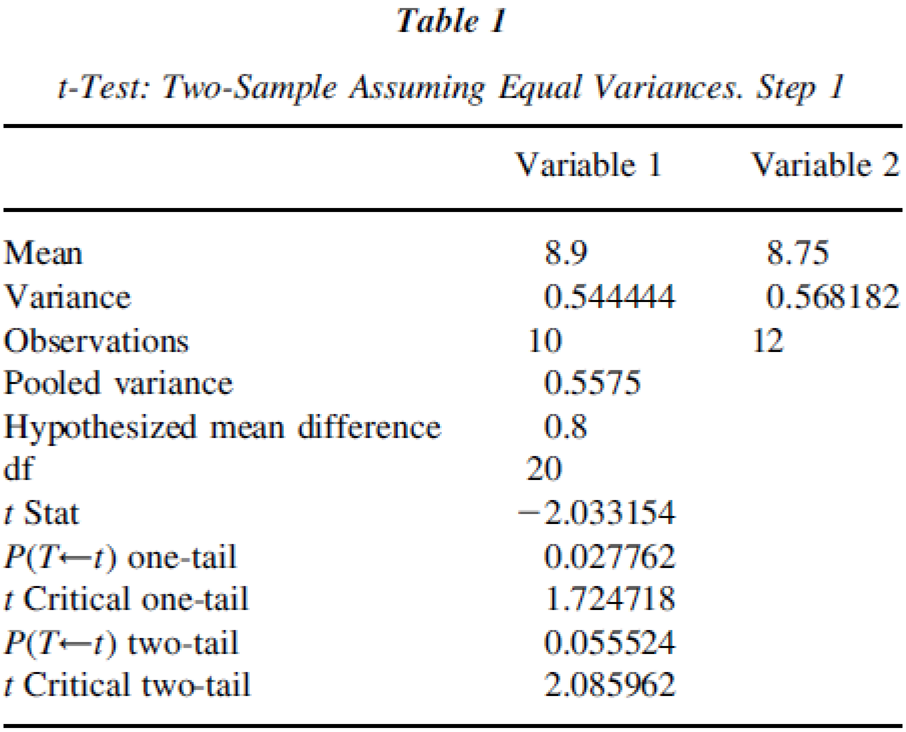

Table 1

t-Test: Two-Sample Assuming Equal Variances. Step 1

	Variable 1	Variable 2
Mean	8.9	8.75
Variance	0.544444	0.568182
Observations	10	12
Pooled variance	0.5575	
Hypothesized mean difference	0.8	
df	20	
t Stat	−2.033154	
P(*T*←*t*) one-tail	0.027762	
t Critical one-tail	1.724718	
P(*T*←*t*) two-tail	0.055524	
t Critical two-tail	2.085962	

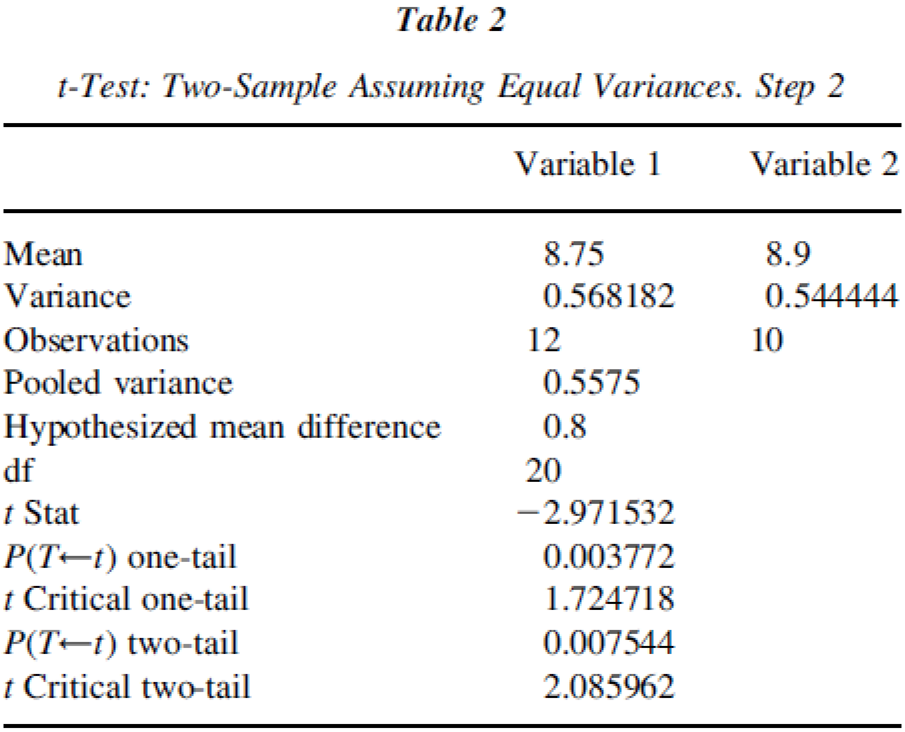

Table 2

t-Test: Two-Sample Assuming Equal Variances. Step 2

	Variable 1	Variable 2
Mean	8.75	8.9
Variance	0.568182	0.544444
Observations	12	10
Pooled variance	0.5575	
Hypothesized mean difference	0.8	
df	20	
t Stat	−2.971532	
P(*T*←*t*) one-tail	0.003772	
t Critical one-tail	1.724718	
P(*T*←*t*) two-tail	0.007544	
t Critical two-tail	2.085962	

DETERMINING EQUIVALENCE CONSTRAINTS

In the above illustrations, constraints $\theta_1 \text{ and } \theta_2$ were required for the equivalence testing approach, but not for the standard testing method. It might be argued, then, that this need for additional information would be a drawback to using the equivalence testing approach. However, we argue that determining the values $\theta_1 \text{ and } \theta_2$, which define the equivalence constraints, is a practical problem that should be addressed by the experimenter regardless of the method used to help assess equivalence. Indeed, the introduction of a practical range $\theta_1 \leq \mu_1 - \mu_2 \leq \theta_2$ which defines equivalence, could also be used to improve the standard testing approach. An approach that is sometimes used is as follows: If the null hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2$ is not rejected then estimate the power of the procedure to detect a difference in the means of at least θ_2 . Then, declare the means equivalent only if this estimated power is sufficiently large (typically, power at least 0.80). Although this approach is better (i.e., more cautious), it does not relieve the inherent problem that the procedure is not designed to detect equivalence. Thus, the method still penalizes increased precision. In addition, estimating the power is difficult in many applied situations and involves uncertainty as well, because the true variance of the populations must be estimated from the data and, thus, the exact power of the test cannot be known. The equivalence testing approach is easier to implement and gives more reliable results. Regardless of the method employed, however, failure to address this issue adequately can result in meaningless or misleading results, as the presence or absence of statistical significance may or may not be sufficient evidence of practical significance.

CONCLUSION

Using the equivalence testing method described, it was determined that there was sufficient evidence at the 5% significance level that thickness measurements of the conditioned and unconditioned tape were equivalent. Thus, the decision was made to eliminate one of the testing episodes. The advantage of this method over the standard testing method is that the probability of incorrectly concluding “equivalence” (a type I error) is known (in this case, the probability is 0.05). Thus, more complete information is available, and engineers can have more confidence when recommending that a testing episode be eliminated.

Many situations arise in industrial quality assurance applications in which it is desired to demonstrate equivalence rather than difference. The equivalence testing approach described in this article is easy to implement in industrial applications and addresses correctly whether results are equivalent. Equivalence testing procedures should be used instead of the standard hypothesis testing approach whenever it is desired to determine equivalence.

MORE ABOUT THE AUTHORS

Scott Richter is an assistant professor in the Department of Mathematics at Western Kentucky University in Bowling Green, KY.

Carri Richter was formerly a Quality Engineer at Tyco Adhesives in Franklin, KY, and is a member of the American Society for Quality.

REFERENCES

1. Reeve, R.; Giesbrecht, F. Dissolution Method Equivalence. In Statistical Case Studies: A Collaboration Between Academe and Industry; Peck, R., Haugh, L., Goodman, A., Eds.; American Statistical Association/Society for Industrial and Applied Mathematics: Philadelphia, 1998; 37–44.

2. Schuirmann, D.J. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. J. Pharmacokinet. Biopharmaceut. 1987, 15 (6), 657–680.

3. Westlake, W.J. Symmetric Confidence Intervals for Bioequivalence Trials. Biometrics 1976, 32, 741–744.

4. Schuirmann, D.J. On Hypothesis Testing to Determine If the Mean of a Normal Distribution Is Contained in a Known Interval. Biometrics 1981, 37, 617.

5. Berger, R.L.; Hsu, J.C. Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. Statist. Sci. 1996, 11 (4), 283–319.

6. Hauck, W.W.; Anderson, S. Comment on Berger and Hsu: Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. Statist. Sci. 1996, 11 (4), 303.

7. Meredith, M.P.; Heise, M.A. Comment on Berger and Hsu: Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. Statist. Sci. 1996, 11 (4), 304–306.