

The Two-Sample t-Test and Randomization Test

by Keith M. Bower, M.S.

Reprinted with permission from the American Society for Quality.

Many Six Sigma practitioners use “Student’s” independent two-sample t-test when investigating differences in means. This type of test is based upon drawing random samples from two independent normal (Gaussian) distributions. Often, however, the assumption of normality and/or random sampling is violated. This article discusses when and why the two-sample t-test may be considered robust to these assumptions by comparing t-test results to randomization test results.

Assumptions of Normality and Random Sampling

The two-sample t-test is widely used when addressing quality issues pertaining to business, engineering, and elsewhere. The theoretical foundation for this test includes the following assumptions:

- The two samples are each drawn from normal distributions.
- The two samples are drawn randomly from their respective populations.

However, for many investigations, one or both of these assumptions may be unrealistic.

For example, consider the investigation by Alfred Hallstrom et al.¹ The experimenters wanted to assess whether artificial respiration (in addition to chest compressions) is necessary before an Emergency Medical Services unit provides expert care. To test for a difference in survival rates using the two methods, telephone dispatchers randomly provided the caller at the scene instructions in either:

- Chest compressions *and* mouth-to-mouth ventilation.
- Chest compressions only.

In this situation, it may not be reasonable to assume that the experimenters have obtained a “representative” sample from a specific population because the subjects in this study are not strategically drawn from a larger population of interest; they are self-selecting.

Importantly, however, the two methods of treatment were applied *randomly* to the subjects. The caller had an equal probability of being instructed to perform either method on the patient. This concept of applying treatments at random is the assumption behind *randomization tests*, which were first described by Ronald A. Fisher.

An Example Comparing t-Test and Randomization Test Results

Consider two fertilizers (A and B) that are randomly applied to a type of sunflower seed. The maximum heights reached (in feet) are recorded after some time period. Assume that all other factors are held constant in this study. The data are shown in the table in Figure 1.

Fig. 1 Sunflower Heights Using Fertilizers A and B

Sample	Fertilizer	Height (ft)
1	A	9.9
2	B	9.6
3	B	9.7
4	B	9.4
5	A	10.1
6	B	9.5
7	A	9.9
8	B	9.6
9	A	9.5
10	A	10.2
11	B	9.4

Null hypothesis (H_0): *no difference* between fertilizers A and B with respect to sunflower height.

If H_0 is really true, then the first sunflower will reach 9.9 feet, for example, regardless of which of the two fertilizers was provided. Therefore, we can consider the observed results as merely one of the 462 ways that the 5 A fertilizers and 6 B fertilizers could have been assigned (N.B. $11!/5!6! = 462$).

Alternative hypothesis (H_1): fertilizer A is *superior* to fertilizer B on average with respect to sunflower height.

When all permutations are calculated, only 5 of the 462 permutations result in a difference *greater than or equal to* the observed mean difference of $9.920 - 9.533 = 0.387$. Divide 5 by 462 to find the probability value (p-value) for the test. The p-value for this randomization test is $5/462 = 0.0108$. We would, therefore, reject H_0 in favor of H_1 at the $\alpha = 0.05$ significance level, and conclude that fertilizer A outperforms fertilizer B with respect to maximum sunflower height.

Randomization Tests

The test procedure in the previous example did not consider normality, random sampling, equal variances, or other assumptions. The conclusion was based solely on the observed results, and the fact that the fertilizers were randomly assigned.

One may wonder why randomization tests are not widely used, nor addressed in many statistical texts. A key reason is the number of computations required for this procedure. In the sunflower example, the number of possible permutations was only 462. With larger sample sizes this value quickly becomes astronomical. For example, with two samples, each of size 30, there are over 1.18×10^{17} possible permutations! As noted by George E.P. Box, William G. Hunter, and J. Stuart Hunter:

It would be tedious [sic] to compute the randomization distribution every time a test of significance was made. However, ... *provided that we randomize*, we can employ t tests... as *approximations to exact randomization tests*, and we will be free of the random sampling assumption as well as the assumption of exact normality.²

With regard to the sunflower example, the results using a two-sample t-test (assuming unequal variances) are shown in Figure 2.

Fig. 2 Two-Sample t-Test

```
Two-Sample T-Test and CI: A, B  
  
Two-sample T for A vs B  


|   | N | Mean  | StDev | SE Mean |
|---|---|-------|-------|---------|
| A | 5 | 9.920 | 0.268 | 0.12    |
| B | 6 | 9.533 | 0.121 | 0.049   |

  
Difference = mu A - mu B  
Estimate for difference: 0.387  
95% lower bound for difference: 0.125  
T-Test of difference = 0 (vs >): T-Value = 2.98  
P-Value = 0.015 DF = 5
```

The output displays a p-value of 0.015. Therefore, using either testing procedure, the null hypothesis of no difference is rejected at the $\alpha = 0.05$ significance level.

Summary

Generally speaking, many statisticians would agree that the two-sample t-test provides a reasonably good approximation to the corresponding randomization test, the use of which does *not* depend on random sampling and normal distribution assumptions.

Though the above example uses the two-sample t-test, the argument can be extended to other statistical procedures (for example, the one-way ANOVA). For further information, and to learn when t-procedures may not provide useful approximations, see George E.P. Box, William G. Hunter, and J. Stuart Hunter's *Statistics for Experimenters*² and Joshua M. Tebbs and my "Some Comments on the Robustness of Student t Procedures."³

References

1. Alfred Hallstrom, Leonard Cobb, Elise Johnson, and Michael Copass, "Cardiopulmonary Resuscitation By Chest Compression Alone Or With Mouth-To-Mouth Ventilation," *The New England Journal of Medicine* 342, no. 21 (2000): 1546-1553.
2. George E. P. Box, William G. Hunter, and J. Stuart Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building* (New York: John Wiley & Sons, Inc., 1978), 96.
3. Joshua M. Tebbs and Keith M. Bower, "Some Comments on the Robustness of Student t Procedures," *Journal of Engineering Education* 92, no. 1 (2003): 91-94.

Bibliography

1. Fisher, Ronald A. *The Design of Experiments*. 8th ed. New York: Hafner Publishing Company Inc., 1966.
2. Kempthorne, Oscar. "Some Aspects of Statistical Inference." *Journal of the American Statistical Association* 61, no. 313 (1966): 11-25.
3. Salsburg, David. *The Lady Tasting Tea*. New York: W.H. Freeman and Co., 2001.
4. "Student" [William S. Gosset]. "The Probable Error of a Mean." *Biometrika* 6 (1908): 1-25.