The Paired *t*-Test

What is the paired *t*-test?

The paired *t*-test is a method used to test whether the mean difference between pairs of measurements is zero or not.

When can I use the test?

You can use the test when your data values are paired measurements. For example, you might have before-and-after measurements for a group of people. Also, the distribution of differences between the paired measurements should be normally distributed.

What are some other names for the paired *t*-test?

The paired *t*-test is also known as the dependent samples *t*-test, the paired-difference *t*-test, the matched pairs *t*-test and the repeated-samples *t*-test.

What if my data isn't nearly normally distributed?

If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. Or, you can perform a *nonparametric* test that doesn't assume normality.

Using the paired *t*-test

The sections below discuss what is needed to perform the test, checking our data, how to perform the test and statistical details.

What do we need?

For the paired *t*-test, we need two variables. One variable defines the pairs for the observations. The second variable is a measurement. Sometimes, we already have the paired differences for the measurement variable. Other times, we have separate variables for "before" and "after" measurements for each pair and need to calculate the differences.

We also have an idea, or hypothesis, that the differences between pairs is zero. Here are three examples:

- A group of people with dry skin use a medicated lotion on one arm and a non-medicated lotion on their other arm. After a week, a doctor measures the redness on each arm. We want to know if the medicated lotion is better than the non-medicated lotion. We do this by finding out if the arm with medicated lotion has less redness than the other arm. Since we have pairs of measurements for each person, we find the differences. Then we test if the mean difference is zero or not.
- We measure weights of people in a program to quit smoking. For each person, we have the weight at the start and end of the program. We want to know if the mean weight change for people in the program is zero or not.
- An instructor gives students an exam and the next day gives students a different exam on the same material. The instructor wants to know if the two exams are equally difficult. We calculate the difference in exam scores for each student. We test if the mean difference is zero or not.

Paired *t*-test assumptions

To apply the paired *t*-test to test for differences between paired measurements, the following assumptions need to hold:

- Subjects must be independent. Measurements for one subject do not affect measurements for any other subject.
- Each of the paired measurements must be obtained from the same subject. For example, the before-and-after weight for a smoker in the example above must be from the same person.
- The measured differences are normally distributed.

Paired *t*-test example

An instructor wants to use two exams in her classes next year. This year, she gives both exams to the students. She wants to know if the exams are equally difficult and wants to check this by looking at the differences between scores. If the mean difference between scores for students is "close enough" to zero, she will make a practical conclusion that the exams are equally difficult. Here is the data:

Table 1: Exam scores for each student

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	6
Nina	65	65	0
Tim	56	62	6
Kate	100	91	-9
Alonzo	88	78	-10
Jose	83	87	4
Nikhil	77	79	2
Julia	92	88	-4
Tohru	90	85	-5
Michael	84	92	8
Jean	68	69	1
Indra	74	81	7
Susan	87	84	-3
Allen	64	75	11
Paul	71	84	13
Edwina	88	82	-6

If you look at the table above, you see that some of the score differences are positive and some are negative. You might think that the two exams are equally difficult. Other people might disagree. The statistical test gives a common way to make the decision, so that everyone makes the same decision on the same data.

Checking the data

Let's start by answering: Is the paired *t*-test an appropriate method to evaluate the difference in difficulty between the two exams?

- Subjects are independent. Each student does their own work on the two exams.
- Each of the paired measurements are obtained from the same subject. Each student takes both tests.
- The distribution of differences is normally distributed. For now, we will assume this is true. We will test this later.

We decide that we have selected a valid analysis method.

Before jumping into the analysis, we should plot the data. The figure below shows a histogram and summary statistics for the score differences.

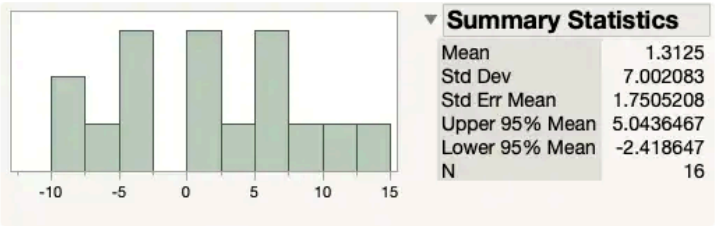

Figure 1: Histogram and summary statistics for the difference in test scores

From the histogram, we see that there are no very unusual points, or *outliers*. The data are roughly bell-shaped, so our idea of a normal distribution for the differences seems reasonable.

From the statistics, we see that the average, or mean, difference is 1.3. Is this "close enough" to zero for the instructor to decide that the two exams are equally difficult? Or not?

How to perform the paired *t*-test

We'll further explain the principles underlying the paired *t*-test in the Statistical Details section below, but let's first proceed through the steps from beginning to end. We start by calculating our test statistic. To accomplish this, we need the average difference, the standard deviation of the difference and the sample size. These are shown in Figure 1 above. (Note that the statistics are rounded to two decimal places below. Software will usually display more decimal places and use them in calculations.)

The average score difference is:

$$\overline{x_d} = 1.31$$

Next, we calculate the standard error for the score difference. The calculation is:

$$\text{Standard Error} = \frac{s_d}{\sqrt{n}} = \frac{7.00}{\sqrt{16}} = \frac{7.00}{4} = 1.75$$

In the formula above, n is the number of students – which is the number of differences. The standard deviation of the differences is s_d.

We now have the pieces for our test statistic. We calculate our test statistic as:

$$t = \frac{\text{Average difference}}{\text{Standard Error}} = \frac{1.31}{1.75} = 0.750$$

To make our decision, we compare the test statistic to a value from the *t*-distribution. This activity involves four steps:

1. We decide on the risk we are willing to take for declaring a difference when there is not a difference. For the exam score data, we decide that we are willing to take a 5% risk of saying that the unknown mean exam score difference is zero when in reality it is not. In statistics-speak, we set the significance level, denoted by α, to 0.05. It's a good practice to make this decision before collecting the data and before calculating test statistics.
2. We calculate a test statistic. Our test statistic is 0.750.
3. We find the value from the *t*-distribution. Most statistics books have look-up tables for the distribution. You can also find tables online. The most likely situation is that you will use software for your analysis and will not use printed tables.

   To find this value, we need the significance level (α = 0.05) and the *degrees of freedom*. The degrees of freedom (*df*) are based on the sample size. For the exam score data, this is:

   $$df = n - 1 = 16 - 1 = 15$$

   The *t* value with α = 0.05 and 15 degrees of freedom is 2.131.
4. We compare the value of our statistic (0.750) to the *t* value. Because 0.750 < 2.131, we cannot reject our idea that the mean score difference is zero. We make a practical conclusion to consider exams as equally difficult.

Statistical details

Let's look at the exam score data and the paired *t*-test using statistical terms.

Our null hypothesis is that the population mean of the differences is zero. The null hypothesis is written as:

$$H_o : \mu_{\mathrm{d}} = 0$$

The alternative hypothesis is that the population mean of the differences is not zero. This is written as:

$$H_o : \mu_{\mathrm{d}} \neq 0$$

We calculate the standard error as:

$$StandardError = \frac{s_d}{\sqrt{n}}$$

The formula shows the sample standard deviation of the differences as s_d and the sample size as n.

The test statistic is calculated as:

$$t = \frac{\mu_{\mathrm{d}}}{\frac{s}{\sqrt{n}}}$$

We compare the test statistic to a t value with our chosen alpha value and the degrees of freedom for our data. In our exam score data example, we set $\alpha = 0.05$. The degrees of freedom (df) are based on the sample size and are calculated as:

$$df = n - 1 = 16 - 1 = 15$$

Statisticians write the t value with $\alpha = 0.05$ and 15 degrees of freedom as:

$$t_{0.05,15}$$

The t value with $\alpha = 0.05$ and 15 degrees of freedom is 2.131. There are two possible results from our comparison:

- The test statistic is lower than the t value. You fail to reject the hypothesis that the mean difference is zero. The practical conclusion made by the instructor is that the two tests are equally difficult. Next year, she can use both exams and give half the students one exam and half the other exam.
- The test statistic is higher than the t value. You reject the hypothesis that the mean difference is zero. The practical conclusion made by the instructor is that the tests are not of equal difficulty. She must use the same exam for all students.

Testing for normality

The normality assumption is more important for small sample sizes than for larger sample sizes.

Normal distributions are symmetric, which means they are equal on both sides of the center. Normal distributions do not have extreme values, or outliers. You can check these two features of a normal distribution with graphs. Earlier, we decided that the distribution of exam score differences were “close enough” to normal to go ahead with the assumption of normality. The figure below shows a normal quantile plot for the data and supports our decision.

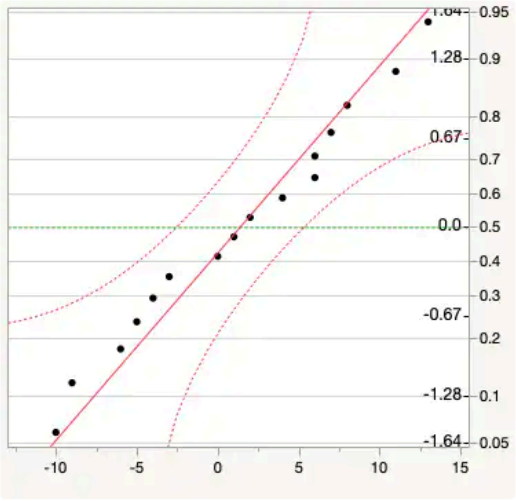

Figure 2: Normal quantile plot for exam data

You can also perform a formal test for normality using software. Figure 3 below shows results of testing for normality with JMP. We test the distribution of the score differences. We cannot reject the hypothesis of a normal distribution. We can go ahead with the paired *t*-test.

Goodness-of-Fit Test for Normality

	A2	Prob > A2
Anderson-Darling	0.197118	0.8730

Figure 3: Testing for normality in JMP software

What if my data are not from a normal distribution?

If your sample size is very small, it is hard to test for normality. In this situation, you need to use your understanding of the measurements. For example, for the test scores data, the instructor knows that the underlying distribution of score differences is normally distributed. Even for a very small sample, the instructor would likely go ahead with the *t*-test and assume normality.

What if you know the underlying measurements are not normally distributed? Or what if your sample size is large and the test for normality is rejected? In this situation, you can use nonparametric analyses. These types of analyses do not depend on an assumption that the data values are from a specific distribution. For the paired *t*-test, a nonparametric test is the Wilcoxon signed-rank test.

Understanding p-values

Using a visual, you can check to see if your test statistic is a more extreme value in the distribution. The *t*-distribution is similar to a normal distribution. The figure below shows a *t*-distribution with 15 degrees of freedom.

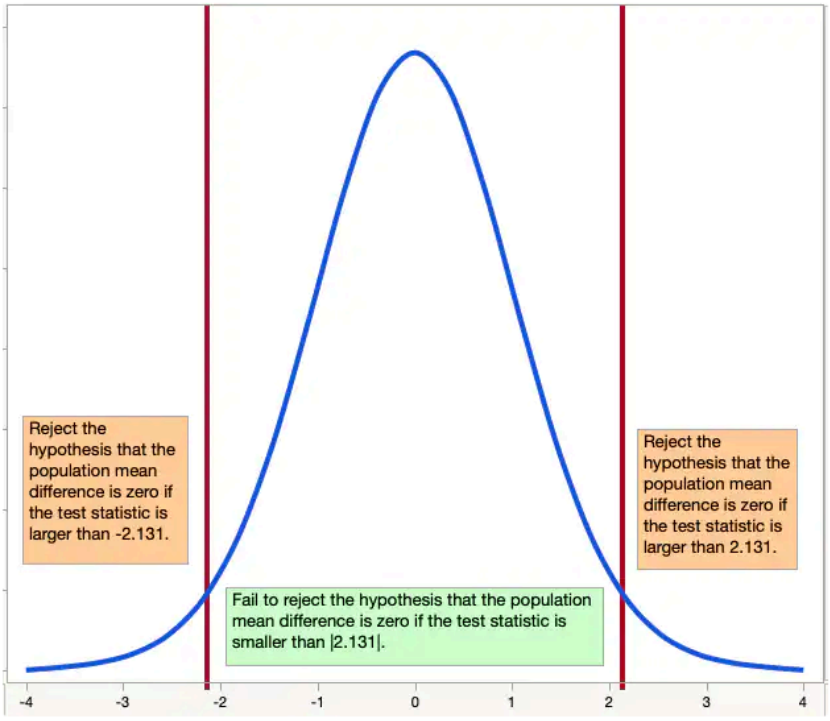

Figure 4: t-distribution with 15 degrees of freedom and $\alpha = 0.05$

Since our test is two-sided and we set $\alpha = 0.05$, the figure shows that the value of 2.131 "cuts off" 2.5% of the data in each of the two tails. Only 5% of the data overall is further out in the tails than 2.131.

Figure 5 shows where our result falls on the graph. You can see that the test statistic (0.75) is not far enough "out in the tail" to reject the hypothesis of a mean difference of zero.

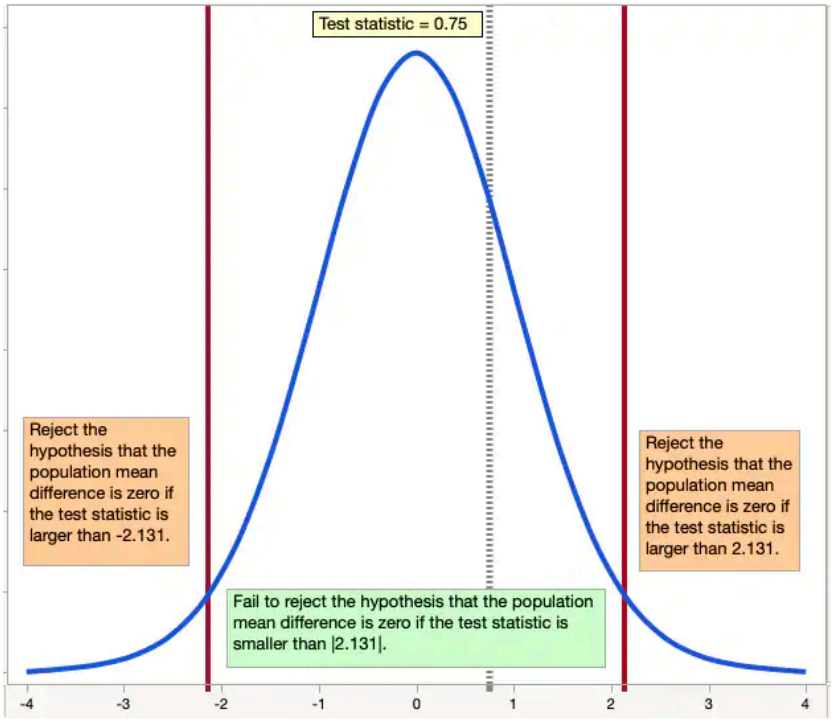

Figure 5: Results of t-test – test statistic is smaller than |2.131|

Putting it all together with software

To perform the paired *t*-test in the real world, you are likely to use software most of the time. The figure below shows results for the paired *t*-test for the exam score data using JMP.

Test Mean

Hypothesized Value	0
Actual Estimate	1.3125
DF	15
Std Dev	7.00208

	t Test
Test Statistic	0.7498
Prob > |t|	0.4650
Prob > t	0.2325
Prob < t	0.7675

Figure 6: Paired t-test results for exam score data using JMP software

The software shows results for a two-sided test (Prob > |t|) and for one-sided tests. The two-sided test is what we want. Our null hypothesis is that the mean difference between the paired exam scores is zero. Our alternative hypothesis is that the mean difference is not equal to zero.

The software shows a *p*-value of 0.4650 for the two-sided test. This means that the likelihood of seeing a sample average difference of 1.31 or greater, when the underlying population mean difference is zero, is about 47 chances out of 100. We feel confident in our decision not to reject the null hypothesis. The instructor can go ahead with her plan to use both exams next year, and give half the students one exam and half the other exam.