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.

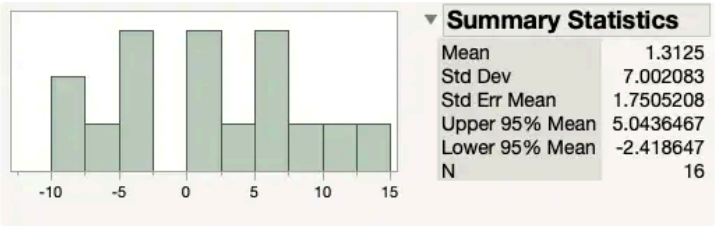

Figure 1: Histogram and summary statistics for the difference in test scores

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.

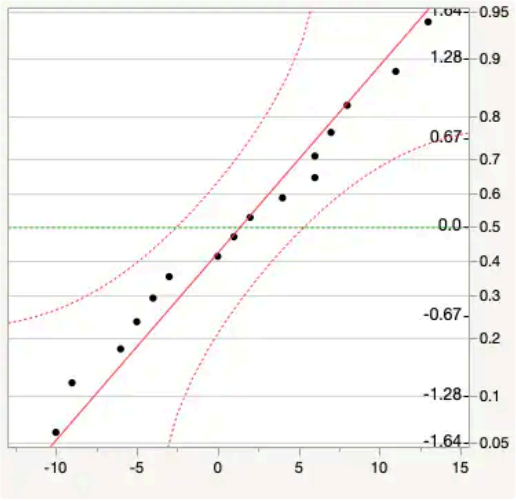

Figure 2: Normal quantile plot for exam data

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.

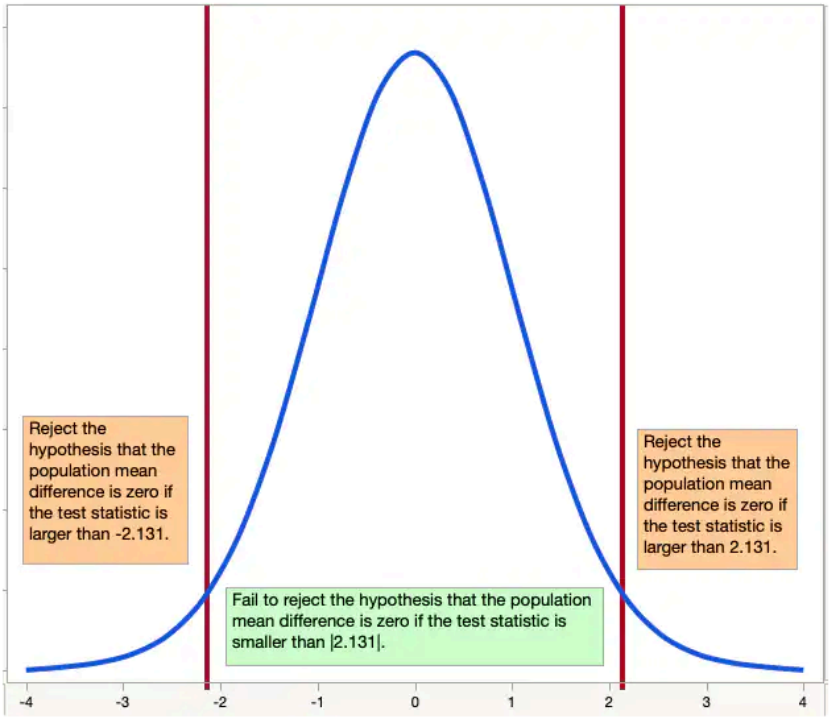

Figure 4: t-distribution with 15 degrees of freedom and $\alpha = 0.05$

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.

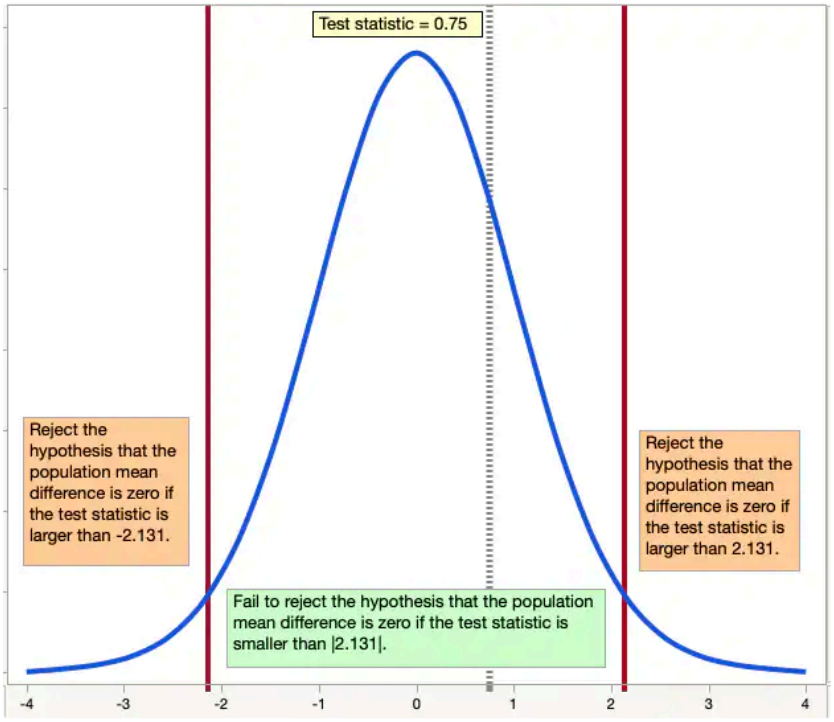

Figure 5: Results of t-test – test statistic is smaller than |2.131|

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.

The Chi-Square Test

What is a Chi-square test?

A Chi-square test is a hypothesis testing method. Two common Chi-square tests involve checking if observed frequencies in one or more categories match expected frequencies.

Is a Chi-square test the same as a χ^2 test?

Yes, χ is the Greek symbol Chi.

What are my choices?

If you have a single measurement variable, you use a Chi-square goodness of fit test. If you have two measurement variables, you use a Chi-square test of independence. There are other Chi-square tests, but these two are the most common.

Types of Chi-square tests

You use a Chi-square test for hypothesis tests about whether your data is as expected. The basic idea behind the test is to compare the observed values in your data to the expected values that you would see if the null hypothesis is true.

There are two commonly used Chi-square tests: the Chi-square goodness of fit test and the Chi-square test of independence. Both tests involve variables that divide your data into categories. As a result, people can be confused about which test to use. The table below compares the two tests.

Visit the individual pages for each type of Chi-square test to see examples along with details on assumptions and calculations.

Table 1: Choosing a Chi-square test

	Chi-Square Goodness of Fit Test	Chi-Square Test of Independence
Number of variables	One	Two
Purpose of test	Decide if one variable is likely to come from a given distribution or not	Decide if two variables might be related or not
Example	Decide if bags of candy have the same number of pieces of each flavor or not	Decide if movie goers' decision to buy snacks is related to the type of movie they plan to watch
Hypotheses in example	H_o: proportion of flavors of candy are the same H_a: proportions of flavors are not the same	H_o: proportion of people who buy snacks is independent of the movie type H_a: proportion of people who buy snacks is different for different types of movies

Theoretical distribution used in test	Chi-Square	Chi-Square
Degrees of freedom	Number of categories minus 1 • In our example, number of flavors of candy minus 1	Number of categories for first variable minus 1, multiplied by number of categories for second variable minus 1 • In our example, number of movie categories minus 1, multiplied by 1 (because snack purchase is a Yes/No variable and 2-1 = 1)

How to perform a Chi-square test

For both the Chi-square goodness of fit test and the Chi-square test of independence, you perform the same analysis steps, listed below. Visit the pages for each type of test to see these steps in action.

1. Define your null and alternative hypotheses before collecting your data.
2. Decide on the alpha value. This involves deciding the risk you are willing to take of drawing the wrong conclusion. For example, suppose you set α=0.05 when testing for independence. Here, you have decided on a 5% risk of concluding the two variables are independent when in reality they are not.
3. Check the data for errors.
4. Check the assumptions for the test. (Visit the pages for each test type for more detail on assumptions.)
5. Perform the test and draw your conclusion.

Both Chi-square tests in the table above involve calculating a test statistic. The basic idea behind the tests is that you compare the actual data values with what would be expected if the null hypothesis is true. The test statistic involves finding the squared difference between actual and expected data values, and dividing that difference by the expected data values. You do this for each data point and add up the values.

Then, you compare the test statistic to a theoretical value from the Chi-square distribution. The theoretical value depends on both the alpha value and the degrees of freedom for your data. Visit the pages for each test type for detailed examples.

Chi-Square Goodness of Fit Test

What is the Chi-square goodness of fit test?

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

When can I use the test?

You can use the test when you have counts of values for a categorical variable.

Is this test the same as Pearson's Chi-square test?

Yes.

Using the Chi-square goodness of fit test

The Chi-square goodness of fit test checks whether your sample data is likely to be from a specific theoretical distribution. We have a set of data values, and an idea about how the data values are distributed. The test gives us a way to decide if the data values have a "good enough" fit to our idea, or if our idea is questionable.

What do we need?

For the goodness of fit test, we need one variable. We also need an idea, or hypothesis, about how that variable is distributed. Here are a couple of examples:

- We have bags of candy with five flavors in each bag. The bags should contain an equal number of pieces of each flavor. The idea we'd like to test is that the proportions of the five flavors in each bag are the same.
- For a group of children's sports teams, we want children with a lot of experience, some experience and no experience shared evenly across the teams. Suppose we know that 20 percent of the players in the league have a lot of experience, 65 percent have some experience and 15 percent are new players with no experience. The idea we'd like to test is that each team has the same proportion of children with a lot, some or no experience as the league as a whole.

To apply the goodness of fit test to a data set we need:

- Data values that are a simple random sample from the full population.
- Categorical or nominal data. The Chi-square goodness of fit test is not appropriate for continuous data.
- A data set that is large enough so that at least five values are expected in each of the observed data categories.

Chi-square goodness of fit test example

Let's use the bags of candy as an example. We collect a random sample of ten bags. Each bag has 100 pieces of candy and five flavors. Our hypothesis is that the proportions of the five flavors in each bag are the same.

Let's start by answering: Is the Chi-square goodness of fit test an appropriate method to evaluate the distribution of flavors in bags of candy?

- We have a simple random sample of 10 bags of candy. We meet this requirement.
- Our categorical variable is the flavors of candy. We have the count of each flavor in 10 bags of candy. We meet this requirement.
- Each bag has 100 pieces of candy. Each bag has five flavors of candy. We expect to have equal numbers for each flavor. This means we expect 100 / 5 = 20 pieces of candy in each flavor from each bag. For 10 bags in our sample, we expect 10 x 20 = 200 pieces of candy in each flavor. This is more than the requirement of five expected values in each category.

Based on the answers above, yes, the Chi-square goodness of fit test is an appropriate method to evaluate the distribution of the flavors in bags of candy.

Figure 1 below shows the combined flavor counts from all 10 bags of candy.

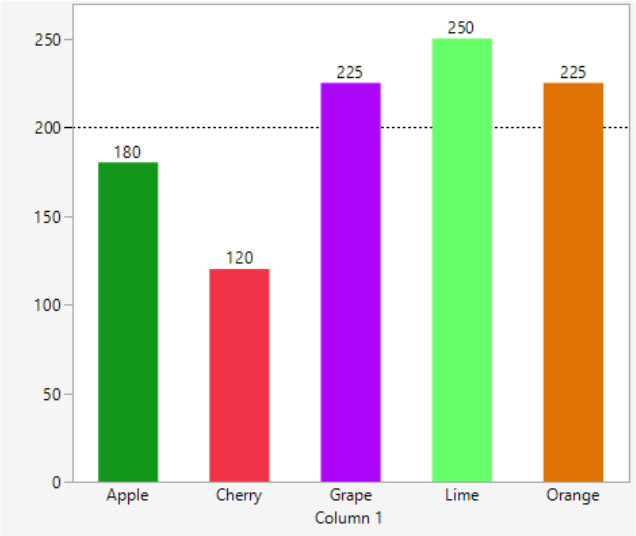

Figure 1: Bar chart of counts of candy flavors from all 10 bags

Without doing any statistics, we can see that the number of pieces for each flavor are not the same. Some flavors have fewer than the expected 200 pieces and some have more. But how different are the proportions of flavors? Are the number of pieces "close enough" for us to conclude that across many bags there are the same number of pieces for each flavor? Or are the number of pieces too different for us to draw this conclusion? Another way to phrase this is, do our data values give a "good enough" fit to the idea of equal numbers of pieces of candy for each flavor or not?

To decide, we find the difference between what we have and what we expect. Then, to give flavors with fewer pieces than expected the same importance as flavors with more pieces than expected, we square the difference. Next, we divide the square by the expected count, and sum those values. This gives us our test statistic.

These steps are much easier to understand using numbers from our example.

Let's start by listing what we expect if each bag has the same number of pieces for each flavor. Above, we calculated this as 200 for 10 bags of candy.

Table 1: Comparison of actual vs expected number of pieces of each flavor of candy

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy
Apple	180	200
Lime	250	200
Cherry	120	200
Cherry	225	200
Grape	225	200

Now, we find the difference between what we have observed in our data and what we expect. The last column in Table 2 below shows this difference:

Table 2: Difference between observed and expected pieces of candy by flavor

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected
Apple	180	200	180-200 = -20
Lime	250	200	250-200 = 50
Cherry	120	200	120-200 = -80
Orange	225	200	225-200 = 25
Grape	225	200	225-200 = 25

Some of the differences are positive and some are negative. If we simply added them up, we would get zero. Instead, we square the differences. This gives equal importance to the flavors of candy that have fewer pieces than expected, and the flavors that have more pieces than expected.

Table 3: Calculation of the squared difference between Observed and Expected for each flavor of candy

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference
Apple	180	200	180-200 = -20	400
Lime	250	200	250-200 = 50	2500
Cherry	120	200	120-200 = -80	6400
Orange	225	200	225-200 = 25	625
Grape	225	200	225-200 = 25	625

Next, we divide the squared difference by the expected number:

Table 4: Calculation of the squared difference/expected number of pieces of candy per flavor

Flavor	Number of Pieces of Candy (10 bags)	Expected Number of Pieces of Candy	Observed-Expected	Squared Difference	Squared Difference / Expected Number
Apple	180	200	180-200 = -20	400	400 / 200 = 2
Lime	250	200	250-200 = 50	2500	2500 / 200 = 12.5
Cherry	120	200	120-200 = -80	6400	6400 / 200 = 32
Orange	225	200	225-200 = 25	625	625 / 200 = 3.125
Grape	225	200	225-200 = 25	625	625 / 200 = 3.125

Finally, we add the numbers in the final column to calculate our test statistic:

$$2 + 12.5 + 32 + 3.125 + 3.125 = 52.75$$

To draw a conclusion, we compare the test statistic to a critical value from the Chi-Square distribution. This activity involves four steps:

1. We first decide on the risk we are willing to take of drawing an incorrect conclusion based on our sample observations. For the candy data, we decide prior to collecting data that we are willing to take a 5% risk of concluding that the flavor counts in each bag across the full population are not equal when they really are. In statistics-speak, we set the significance level, α , to 0.05.
2. We calculate a test statistic. Our test statistic is 52.75.
3. We find the theoretical value from the Chi-square distribution based on our significance level. The theoretical value is the value we would expect if the bags contain the same number of pieces of candy for each flavor.

   In addition to the significance level, we also need the *degrees of freedom* to find this value. For the goodness of fit test, this is one fewer than the number of categories. We have five flavors of candy, so we have 5 – 1 = 4 degrees of freedom.

   The Chi-square value with α = 0.05 and 4 degrees of freedom is 9.488.
4. We compare the value of our test statistic (52.75) to the Chi-square value. Since 52.75 > 9.488, we reject the null hypothesis that the proportions of flavors of candy are equal.

We make a practical conclusion that bags of candy across the full population do not have an equal number of pieces for the five flavors. This makes sense if you look at the original data. If your favorite flavor is Lime, you are likely to have more of your favorite flavor than the other flavors. If your favorite flavor is Cherry, you are likely to be unhappy because there will be fewer pieces of Cherry candy than you expect.

Understanding results

Let's use a few graphs to understand the test and the results.

A simple bar chart of the data shows the observed counts for the flavors of candy:

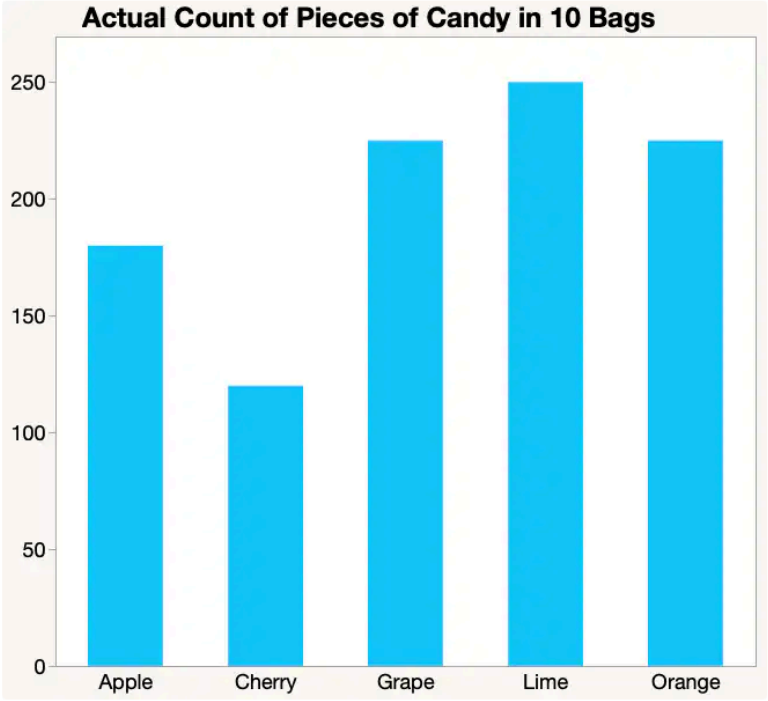

Figure 2: Bar chart of observed counts for flavors of candy

Another simple bar chart shows the expected counts of 200 per flavor. This is what our chart would look like if the bags of candy had an equal number of pieces of each flavor.

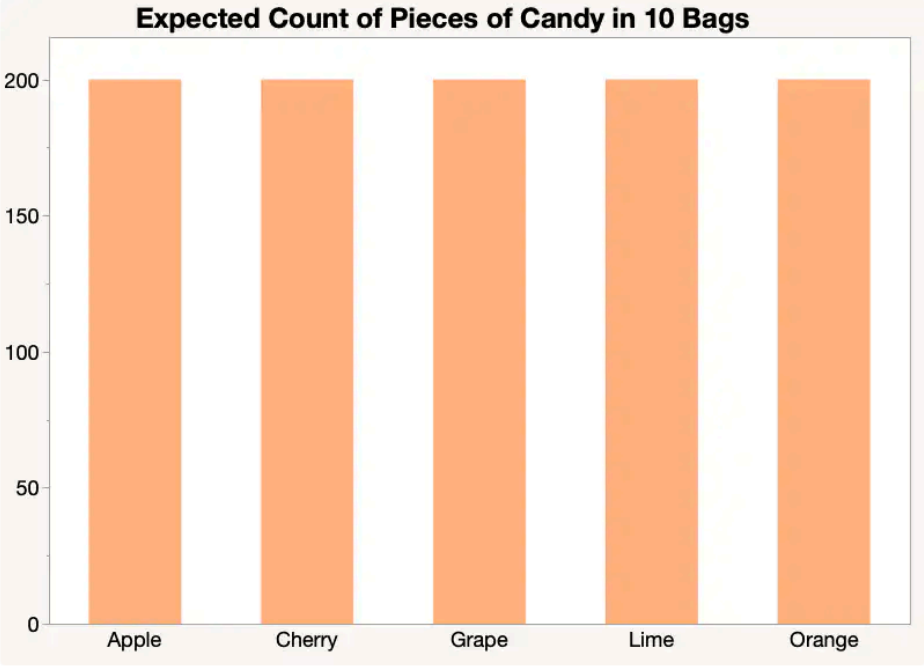

Figure 3: Bar chart of expected counts of each flavor

The side-by-side chart below shows the actual observed number of pieces of candy in blue. The orange bars show the expected number of pieces. You can see that some flavors have more pieces than we expect, and other flavors have fewer pieces.

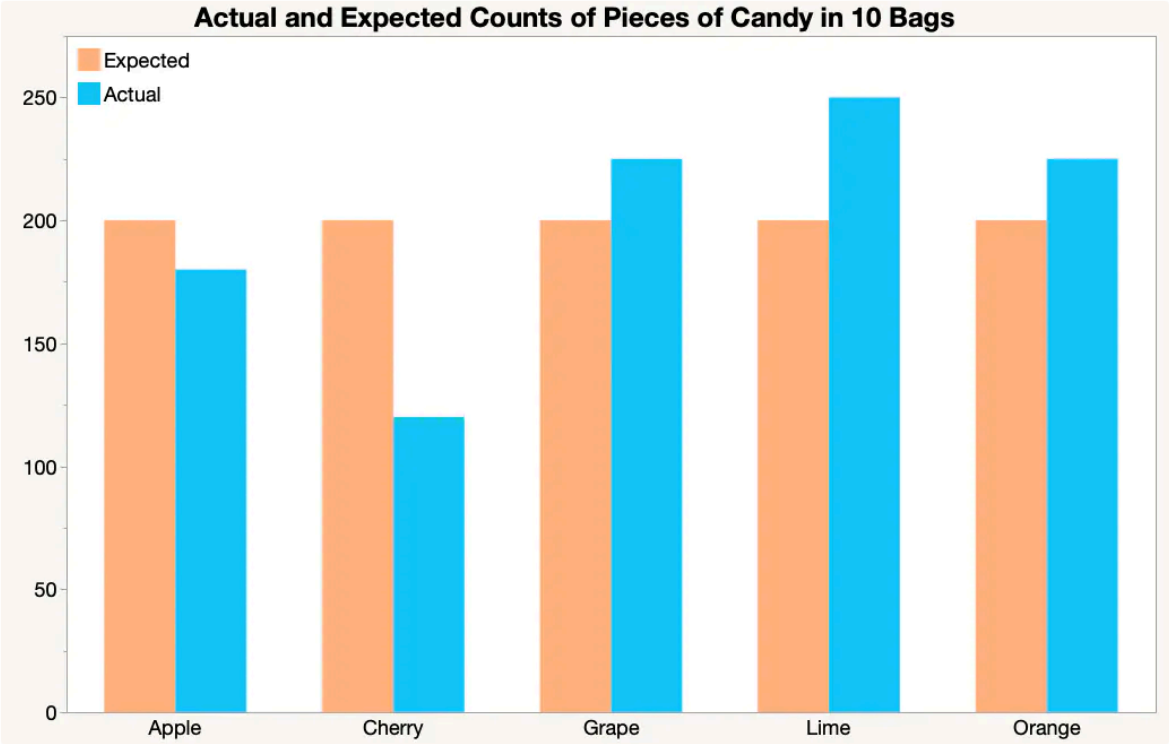

Figure 4: Bar chart comparing actual vs. expected counts of candy

The statistical test is a way to quantify the difference. Is the actual data from our sample "close enough" to what is expected to conclude that the flavor proportions in the full population of bags are equal? Or not? From the candy data above, most people would say the data is not "close enough" even without a statistical test.

What if your data looked like the example in Figure 5 below instead? The purple bars show the observed counts and the orange bars show the expected counts. Some people would say the data is "close enough" but others would say it is not. The statistical test gives a common way to make the decision, so that everyone makes the same decision on a set of data values.

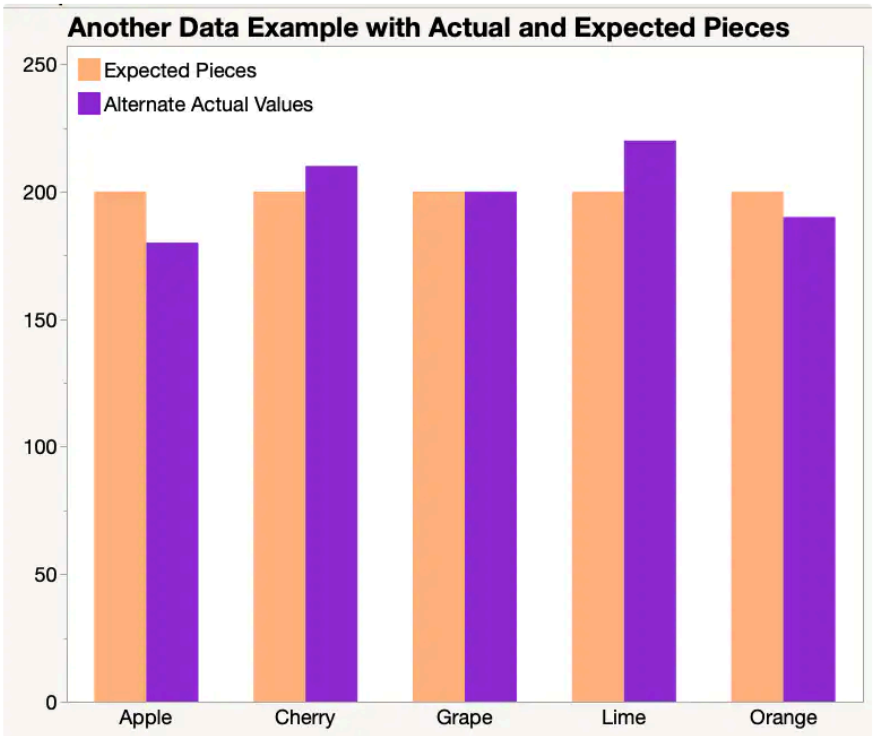

Figure 5: Bar chart comparing expected and actual values using another example data set

Statistical details

Let's look at the candy data and the Chi-square test for goodness of fit using statistical terms. This test is also known as Pearson's Chi-square test.

Our null hypothesis is that the proportion of flavors in each bag is the same. We have five flavors. The null hypothesis is written as:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5$$

The formula above uses *p* for the proportion of each flavor. If each 100-piece bag contains equal numbers of pieces of candy for each of the five flavors, then the bag contains 20 pieces of each flavor. The proportion of each flavor is 20 / 100 = 0.2.

The alternative hypothesis is that at least one of the proportions is different from the others. This is written as:

$$H_a : at\ least\ one\ p_i\ not\ equal$$

In some cases, we are not testing for equal proportions. Look again at the example of children's sports teams near the top of this page. Using that as an example, our null and alternative hypotheses are:

$$H_0 : p_1 = 0.2, p_2 = 0.65, p_3 = 0.15$$

$$H_a : at\ least\ one\ p_i\ not\ equal\ to\ expected\ value$$

Unlike other hypotheses that involve a single population parameter, we cannot use just a formula. We need to use words as well as symbols to describe our hypotheses.

We calculate the test statistic using the formula below:

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

In the formula above, we have *n* groups. The $\sum$ symbol means to add up the calculations for each group. For each group, we do the same steps as in the candy example. The formula shows O_i as the Observed value and E_i as the Expected value for a group.

We then compare the test statistic to a Chi-square value with our chosen significance level (also called the alpha level) and the degrees of freedom for our data. Using the candy data as an example, we set $\alpha = 0.05$ and have four degrees of freedom. For the candy data, the Chi-square value is written as:

$$\chi^2_{0.05,4}$$

There are two possible results from our comparison:

- The test statistic is lower than the Chi-square value. You fail to reject the hypothesis of equal proportions. You conclude that the bags of candy across the entire population have the same number of pieces of each flavor in them. The fit of equal proportions is "good enough."
- The test statistic is higher than the Chi-Square value. You reject the hypothesis of equal proportions. You cannot conclude that the bags of candy have the same number of pieces of each flavor. The fit of equal proportions is "not good enough."

Let's use a graph of the Chi-square distribution to better understand the test results. You are checking to see if your test statistic is a more extreme value in the distribution than the critical value. The distribution below shows a Chi-square distribution with four degrees of freedom. It shows how the critical value of 9.488 "cuts off" 95% of the data. Only 5% of the data is greater than 9.488.

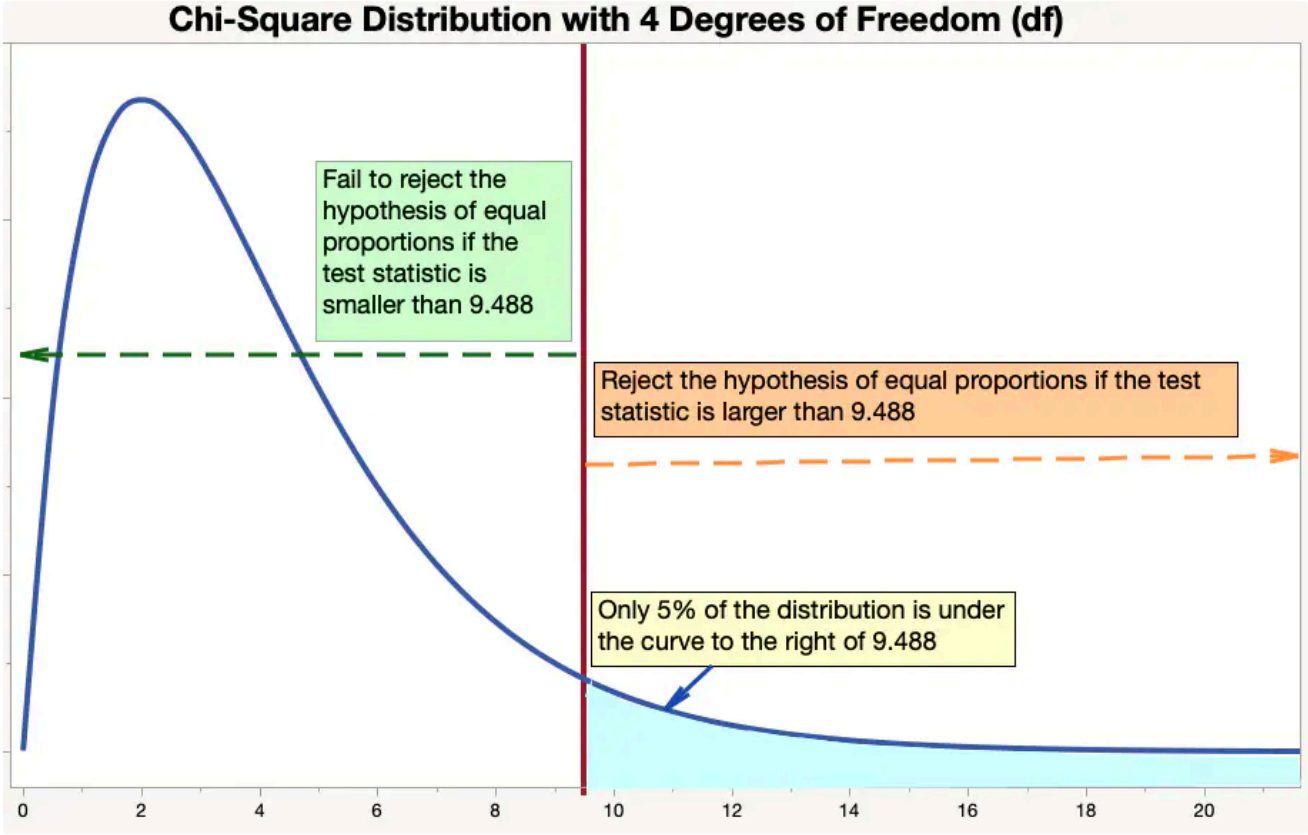

Figure 6: Chi-square distribution for four degrees of freedom

The next distribution plot includes our results. You can see how far out "in the tail" our test statistic is, represented by the dotted line at 52.75. In fact, with this scale, it looks like the curve is at zero where it intersects with the dotted line. It isn't, but it is very, very close to zero. We conclude that it is very unlikely for this situation to happen by chance. If the true population of bags of candy had equal flavor counts, we would be extremely unlikely to see the results that we collected from our random sample of 10 bags.

The next distribution plot includes our results. You can see how far out "in the tail" our test statistic is, represented by the dotted line at 52.75. In fact, with this scale, it looks like the curve is at zero where it intersects with the dotted line. It isn't, but it is very, very close to zero. We conclude that it is very unlikely for this situation to happen by chance. If the true population of bags of candy had equal flavor counts, we would be extremely unlikely to see the results that we collected from our random sample of 10 bags.

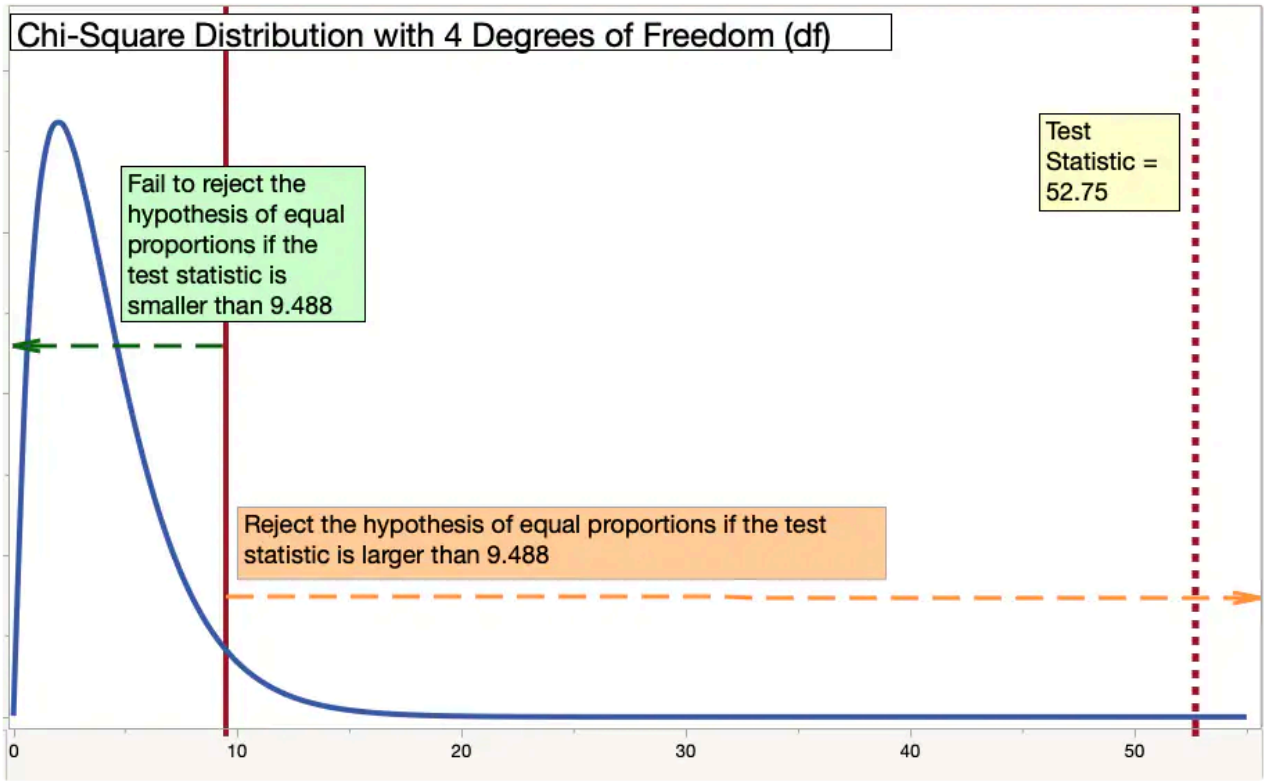

Figure 7: Chi-square distribution for four degrees of freedom with test statistic plotted

Most statistical software shows the p-value for a test. This is the likelihood of finding a more extreme value for the test statistic in a similar sample, assuming that the null hypothesis is correct. It's difficult to calculate the p-value by hand. For the figure above, if the test statistic is exactly 9.488, then the p-value will be p=0.05. With the test statistic of 52.75, the p-value is very, very small. In this example, most statistical software will report the p-value as "p < 0.0001." This means that the likelihood of another sample of 10 bags of candy resulting in a more extreme value for the test statistic is less than one chance in 10,000, assuming our null hypothesis of equal counts of flavors is true.