Google

Reading: Chi-squared tests: Goodness of fit versus independence

In the previous course, you learned how hypothesis tests are used to see significant differences among groups. Chi-squared tests are used to determine whether one or more observed categorical variables follow expected distribution(s). For example, you may expect that 50% more movie goers attend movies on weekends in comparison to weekdays. After observing movie goers attendance for a month, you then can perform a chi-squared test to see if your initial hypothesis was correct.

This reading will cover the two main chi-squared tests—goodness of fit and test for independence—which can be used to test your expected hypothesis against what actually occurred. Data professionals perform these hypothesis tests to offer organizations actionable insights that drive decision making.

The Chi-squared goodness of fit test

Chi-squared (χ^2) goodness of fit test is a hypothesis test that determines whether an observed categorical variable with more than two possible levels follows an expected distribution. The null hypothesis (H0) of the test is that the categorical variable follows the expected distribution. The alternative hypothesis (Ha) is that the categorical variable does not follow the expected distribution. Consider the scenario in this reading that will define the null and alternative hypotheses based on the scenario, set up a Goodness of Fit test, evaluate the test results, and draw a conclusion.

Chi-squared goodness of fit scenario

Imagine that you work as a data professional for an online clothing company. Your boss tells you that they expect the number of website visitors to be the same for each day of the week. You decide to test your boss's hypothesis and pull data every day for the next week and record the number of website visitors in the table below:

Day of the Week	Observed Values
Sunday	650
Monday	570
Tuesday	420
Wednesday	480
Thursday	510
Friday	380
Saturday	490
Total	3,500

Here are the main steps you will take:

1. Identify the Null and Alternative Hypotheses
2. Calculate the chi-square test statistic ($\chi 2$)
3. Calculate the p-value
4. Make a conclusion

Step 1: Identify the null and alternative hypotheses

The first step in performing a chi-squared goodness of fit test is to determine your null and alternative hypothesis. Since you are testing if the number of website visitors follows your boss's expectations, the below are your null and alternative hypotheses :

H_0: The week you observed follows your boss's expectations that the number of website visitors is equal on any given day

H_a: The week you observed does not follow your boss's expectations; therefore, the number of website visitors is not equal across the days of the week

Step 2: Calculate the chi-squared test statistic ($\chi 2$)

Next, calculate a test statistic to determine if you should reject or fail to reject your null hypothesis. This test statistic is known as the chi-squared statistic and is calculated based on the following formula:

$$\sum \frac{(observed - expected)^2}{expected}$$

The intuition behind this formula is that it should quantify the extent of any discrepancies between observed frequencies and expected frequencies for each categorical level. Squaring these differences does two things. First, it ensures that all discrepancies between observed and expected contribute positively to the chi-squared statistic. Second, it penalizes larger discrepancies. Dividing the sum of the squared differences by the expected frequency of each category level standardizes the differences. In other words, it accounts for the fact that larger discrepancies are more significant when the expected frequencies are small, and less so when the expected frequencies are large.

Returning to the example, since there were a total of 3,500 website visitors you observed; your boss's expectation is that 500 visitors would visit each day (3,500/7). In the formula above, 500 would serve as the "expected" value. A column has been added to your original table to include the test statistic calculation for each weekday:

Day of the Week	Observed Values	Chi-Squared Test Statistic
Sunday	650	$\frac{(650-500)^2}{500} = 45$
Monday	570	$\frac{(570-500)^2}{500} = 9.8$
Tuesday	420	$\frac{(420-500)^2}{500} = 12.8$
Wednesday	480	$\frac{(480-500)^2}{500} = 0.8$
Thursday	510	$\frac{(510-500)^2}{500} = 0.2$
Friday	380	$\frac{(380-500)^2}{500} = 28.8$
Saturday	490	$\frac{(490-500)^2}{500} = 0.2$

The X^2 statistic would be the sum of the third column above:

X^2 = 45 + 9.8 + 12.8 + 0.8 + 0.2 + 28.8 + 0.2

X^2 = 97.6

Note that the X^2 goodness of fit test does not produce reliable results when there are any expected values of less than five.

Step 3: Find the p-value

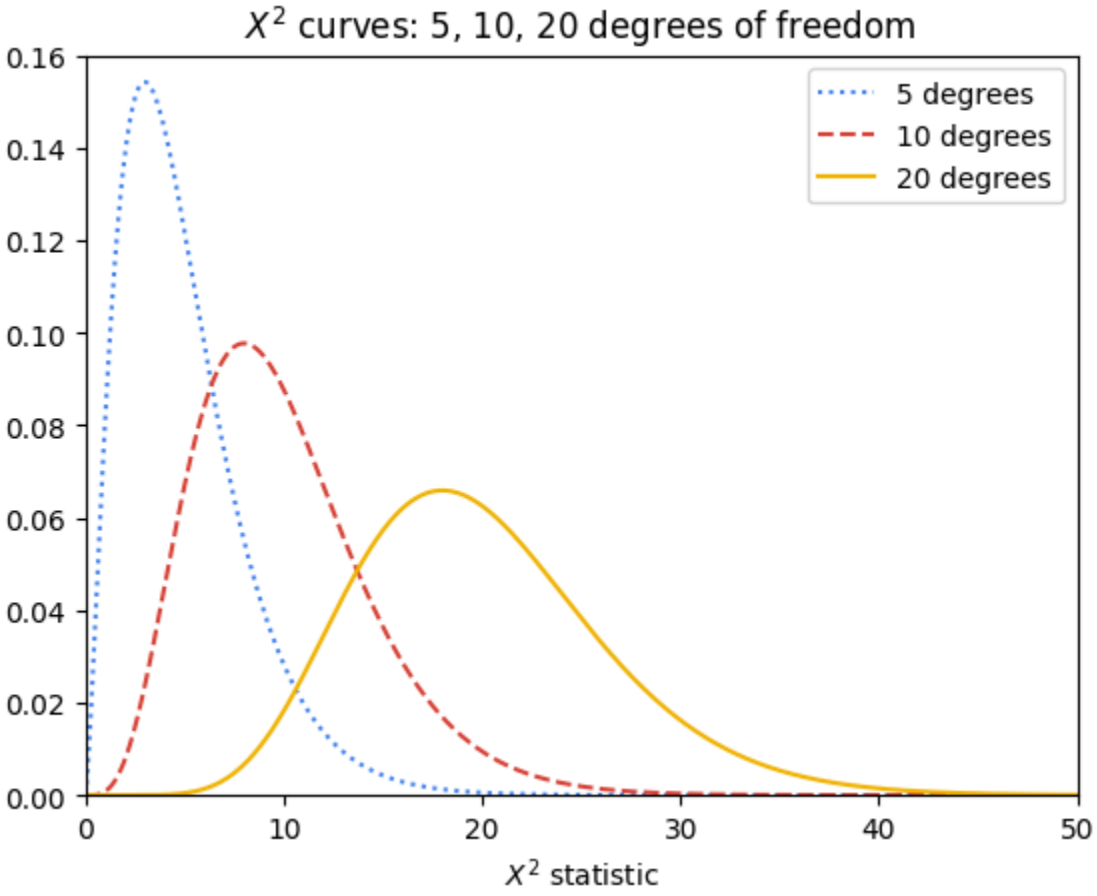
Now that you have calculated the X^2 statistic, consider the following question: What is the probability of obtaining a X^2 statistic of 97.6 or greater when examining 3,500 website visits if they null hypothesis is true? This is the question that the p-value—or "observed significance level"—will answer.

A long time ago, Pearson realized that p-values for X^2 statistics very closely corresponded with areas under certain curves, known as X^2 curves. X^2 curves represent probability density functions, and their shapes vary based on how many degrees of freedom are present in the experiment. Degrees of freedom are determined by the model, not by the data. This means that, in the website traffic example, the degrees of freedom are determined by how many different days a given visit can occur on—not by how many visits are sampled nor by the daily frequencies of the samples themselves. When the model is fully specified (i.e., you know all the possible categorical levels), then:

degrees of freedom = number of categorical levels – 1.

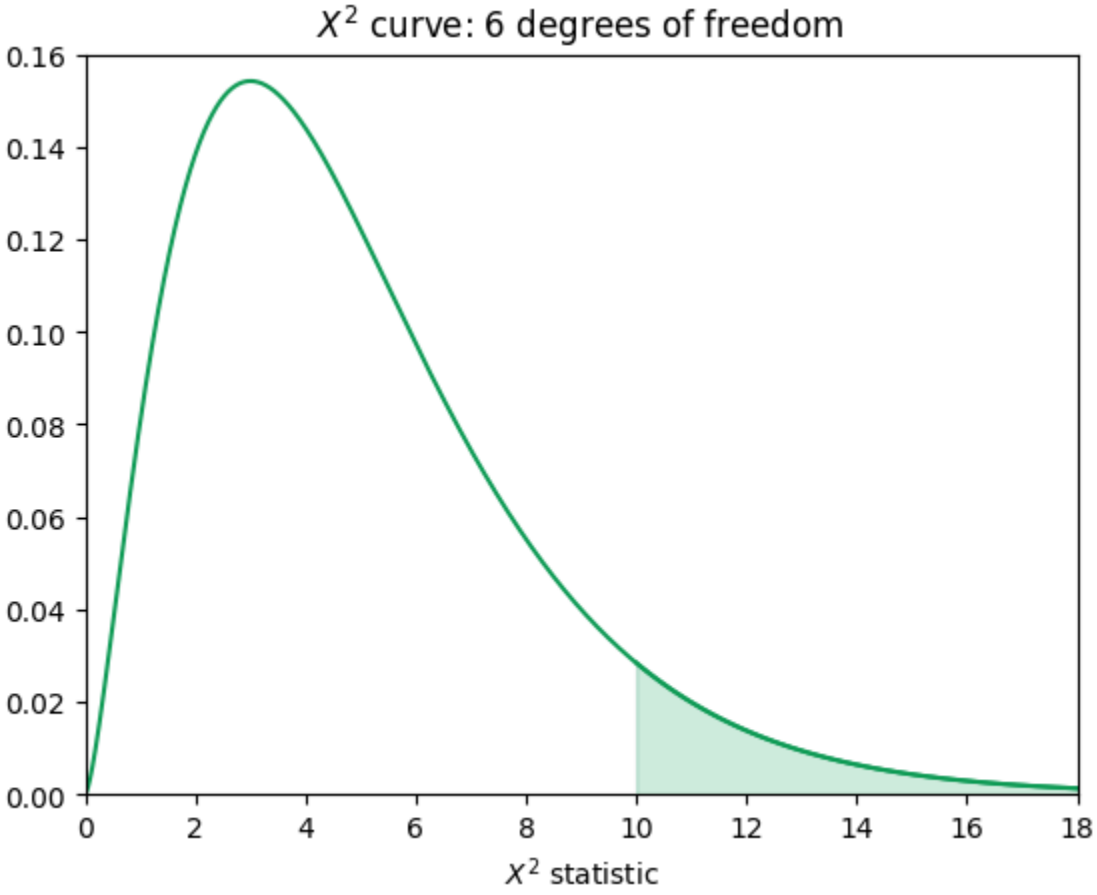
In this example, there are seven categorical levels (one for each day of the week). Therefore, there are six degrees of freedom. This is because the counts of each level (day) are free to fluctuate, but once you know the counts for six days, the seventh day cannot vary. It must result in a total of 3,500 when summed with the other six days.

The following figure depicts the X^2 curves for three different degrees of freedom: five, 10, and 20.



The p-value for a given X^2 test statistic is very closely approximated by the area to its right beneath the X^2 curve of the appropriate degrees of freedom. Notice that the more degrees of freedom there are in the experiment, the greater the area under the right tail of the curve for any given X^2 test statistic, and therefore the greater the probability of getting a given X^2 test statistic if the null hypothesis is true.

The following figure contains the X^2 curve for six degrees of freedom. For a X^2 test statistic of, for instance, 10, the value of P is approximated by the shaded area under the curve where x ≥ 10.



In the case of the website visits, there are six degrees of freedom, but the X^2 test statistic is 97.6—far along the x-axis in the right-skewed tail of the curve. The area under this interval is miniscule: 7.94e-19.

In other words, the chances of getting a X^2 test statistic ≥ 97.6 from 3,500 website visits if the null hypothesis were true are 7.94e-17%—practically zero.

Step 4: Make a conclusion

Since the p-value is far less than 0.05, there is sufficient evidence to suggest that the number of visitors is not equal per day.

Coding

Thankfully, you don't need to manually calculate your X^2 test statistic or determine P by hand. You can use the chisquare() function from Python's **scipy.stats** package to do this. The following code uses your observed and expected values to calculate the chi-squared test statistic and the p-value. Note that the degrees of freedom are set to the number of observed frequencies minus one. This can be adjusted using the **ddof** parameter, but note that this parameter represents **k - 1 - ddof** degrees of freedom, where **k** is the number of observed frequencies. So, by default, **ddof=0** when you call the function, and setting **ddof=1** means that your degrees of freedom are reduced by two.

```
import scipy.stats as stats
observations = [650, 570, 420, 480, 510, 380, 490]
expectations = [500, 500, 500, 500, 500, 500, 500]
result = stats.chisquare(f_obs=observations, f_exp=expectations)
result
```

Output:

```
Power_divergenceResult(statistic=97.599999999999994, pvalue=7.9438869233438347e-19)
```

The output confirms your calculation of the chi-square test statistic in Step 2 and also gives you the associated p-value. Because the p-value is less than the significance level of 5%, you can reject the null hypothesis.

The Chi-squared test for independence

Chi-squared (χ^2) Test for Independence is a hypothesis test that determines whether or not two categorical variables are associated with each other. It is valid when your data comes from a random sample and you want to make an inference about the general population. The null hypothesis (H_0) of the test is that two categorical variables are independent. The alternative hypothesis (H_a) is that two categorical variables are not independent.

Chi-squared test for independence scenario

Now suppose that you have been asked to expand your analysis to look at the relationship between the device that a website user used and their membership status. To do this, you must use the X^2 test for independence. In this example, the X^2 test of independence determines whether the type of device a visitor uses to visit the website (Mac or PC) is dependent on whether he or she has a membership account or browses as a guest (member or guest).

Step 1: Identify the null and alternative hypotheses

Just like the goodness of fit scenario, the first step is to determine your null and alternative hypotheses. You are comparing if the device used to visit your clothing store (Mac or PC) is independent from the visitor's membership status (member or guest). From that information you can determine that your null and alternative hypotheses are as follows:

H_0: The type of device a website visitor uses to visit the website is independent of the visitor's membership status.

H_a: The type of device a website visitor uses to visit the website is not independent of the visitor's membership status.

Step 2: Calculate the chi-squared test statistic (χ2)

To calculate the X^2 test statistic, arrange the data as a table that contains *m* x *n* values, where *m* and *n* are the number of possible levels contained within each respective categorical variable. The following table below breaks down the website visitors based on the device they used and their membership status.

Observed Values	Member	Guest	Total
Mac	850	450	1,300
PC	1,300	900	2,200
Total	2,150	1,350	3,500

Notice that the table starts with 2 x 2 known values (two levels for each category), from which totals are derived. These totals can be used to calculate the expected values, which are necessary to get the X^2 test statistic.

To calculate the expected values, use the following formula:

$$expected\ value = \frac{column\ total * row\ total}{overall\ total}$$

For example, the expected value for a Mac member would be:

$$expected\ value = \frac{2{,}150 * 1{,}300}{3{,}500} = 799$$

The logic of this calculation is as follows: if device and membership status are truly independent, then the rate of Mac users who are members should be the same as the rate of Mac users who are guests. The percentage of users who use Macs out of *all* the users is 1,300 / 3,500 = 0.371 * 100 = 37.1%. Accordingly, 37.1% of members and 37.1% of guests would be expected to use Macs. So, 0.371 * 2,150 members ≈ 799.

The following table contains all the expected values:

Expected Values	Member	Guest
Mac	799	501
PC	1,351	849

Step 3: Find the p-value

Finding the p-value associated with a particular X^2 test statistic is similar to the process outlined already for the goodness of fit test. The only minor difference is how to determine the number of degrees of freedom. For an independence test with two categorical variables with *m* x *n* possible levels, there are (*m* – 1) (*n* – 1) degrees of freedom, assuming there are no other constraints on the probabilities. So, in the working example, this means there is (2 – 1)(2 – 1) = 1 degree of freedom. The p-value in this example is 0.00022. It was determined using Python.

Step 4: Make a conclusion

Because the p-value is 0.00022, you can reject the null hypothesis in favor of the alternative. You conclude that the type of device a website user uses is not independent of his or her membership status. You may recommend to your boss to dive into the reasons behind why visitors sign up for paid memberships more on a particular device. Perhaps the sign-up button appears differently on a particular device. Or maybe there are device-specific bugs that need to be fixed. These are just a couple examples of things you might consider for further exploration.

Coding

You can use the **scipy.stats** package's chi2_contingency() function to obtain the X^2 test statistic and p-value of a X^2 independence test. The **chi2_contingency()** function only needs the observed values; it will calculate the expected values for you. Here is the Python code:

```
import numpy as np
import scipy.stats as stats
observations = np.array([[850, 450],[1300, 900]])
result = stats.contingency.chi2_contingency(observations, correction=False)
result
```

Output:

```
(13.660757846804358, 0.00021898310129108426, 1, array([[  798.57142857,   501.42857143],
       [ 1351.42857143,   848.57142857]]))
```

The output above is in the following order: the X^2 statistic, p-value, degrees of freedom, and expected values in array format. One thing to note is that when degrees of freedom = 1 (i.e., you have a 2 x 2 table), the default behavior of the **stats.chi2_contingency()** function is to apply Yates' correction for continuity. This is to make it less likely that small discrepancies will result in significant X^2 values. It is designed to be used when it's possible for an expected frequency in the table to be small (generally < 5). In the given example, it is known that the expected values are all well over five. Therefore, the **correction** parameter was set to False.

Key takeaways

- The X^2 goodness of fit test is used to test if an observed categorical variable follows a particular expected distribution.
- The X^2 test for independence is used to test if two categorical variables are independent of each other or not (when samples are drawn at random and you want to make an inference about the whole population).
- Both X^2 tests follow the same hypothesis testing steps to determine whether you should reject or fail to reject the null hypothesis to drive decision making, as you have explored elsewhere in this program.
