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 follows an expected distribution. The null hypothesis (H_0) of the test is that the categorical variable follows the expected distribution. The alternative hypothesis (H_a) 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 (χ^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 (χ^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:

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

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 chi-squared 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

Step 3: Find the p-value

Previously, you learned about setting a significance level (alpha) for your hypothesis test and how to use Python's Scipy stats package to determine p-values. You can use the same module's chisquare function to pass in your data to obtain the test statistic and p-value. The following code uses your observed and expected values to calculate the chi-squared test statistic and the p-value:

```
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)
print(Result)

# Output: Power_divergenceResult(statistic = 97.6, pvalue = 7.9438869e-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.

Step 4: Make a conclusion

Since the p-value is less than 0.05, there is sufficient evidence to suggest that the number of visitors is not equal per day. Now that you have made this conclusion, you now have been asked to expand your analysis to look at the relationship between the device that a website user used and their membership status. In order to expand your analysis you must use the Chi-Squared Test for Independence.

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. 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. You will utilize the chi-squared test of independence to compare if 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 (χ^2)

The table below now breaks down our 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

In order to get the expected value under the independence assumption, you will 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 table below shows all the expected values:

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

Step 3: Find the p-value

You can use the Python Scipy package' chi2_contingency function to obtain the chi-square test statistic and p-value. The chi2_contingency function only needs the observed values - it will calculate the expected values for you. Here is the python code:

```
Observations = np.array([[850, 450],
                         [1300, 900]]
Result = stats.contingency.chi2_contingency(Observations)
print(Result)

...
Output: (13.3964,
      0.000252,
      1,
      array([[798.57, 501.42],
         [1351.42, 848.57]]))
...
```

The output above is in the following order: the chi-square statistic, p-value, degrees of freedom, and expected values in array format. Looking at the p-value compared to a significance level of 5%, you can REJECT the null hypothesis in favor of the alternative.

Step 4: Make a conclusion

Based on the above p-value, 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. Is the sign-up button showing up differently on a particular device? Are there device specific bugs that need to be fixed? These are a couple of many questions you should seek into next to help

Key takeaways

- The Chi-squared Goodness of Fit test is used to test if an observed categorical variable follows an expected distribution.
- The Chi-squared Test for Independence is used to test if two categorical variables are independent of each other or not.
- Both Chi-squared 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.

Mark as completed

Like Dislike Report an issue