Discovering Insights with Chi Square Tests: A Hands-on Approach in Python

BEGINNER PANDAS PYTHON STATISTICS

Introduction

Let me take you into the universe of chi-square tests and how we can involve them in Python with the scipy library. We'll be going over the chi-square integrity of the fit test. Whether the reader is just starting or an accomplished information examiner, this guide will outfit you with pragmatic models and experiences so you can unhesitatingly apply chi-square tests in your own work.

Learning objectives

By the end of this article, readers will have:

1. Understood what a Chi-Square Test is and its purpose.

2. Recognized the different types of Chi-Square Tests.

3. Calculated Chi-Square in order to test any relation between two categorical variables.

4. Understand a project implemented in Chi-Square in Python using step-by-step instructions.

This article was published as a part of the Data Science Blogathon.

What is Chi-Square Test?

The Chi-Square test is one of the fact-based interactions used to assess the connection between two all-out factors to figure out the connection between them. This test is extremely straightforward including looking at the noticed frequencies of the factors with their normal frequencies under the supposition that there is no relationship between them. The Chi-Square trial of freedom is usually utilized kind of Chi-Square test. It is applied in circumstances where we have two straight-out factors – like obesity and heart failure event, and we need to research on the off chance that there is an association between them. By doing this we can decide if the example falls into classes in light of our assumptions for the variable dissemination.

Types of Chi-Square Tests

There are several types of Chi-Square Tests, including the chi-square goodness of fit test, the chi-square test of independence, and the chi-square test for homogeneity. The type of test used will depend on the specific research question being addressed and the type of data being analyzed.

1. **Chi-square Goodness of Fit Test:** This type of test is used to find out how the observed value of a given condition is significantly (or not) different from the expected value

2. **Chi-square Test of Independence:** This type of test is a statistical hypothesis test that can be utilized to determine if 2 categorical and nominal variables are (likely) related or not.

3. **Chi-square Test for Homogeneity:** This type of test is used by statisticians to check whether different columns and/or rows of data in a table belong to the same population (or not).

> A Comprehensive Guide to Using Chi Square Tests for Data Analysis (Updated 2023)

Calculating Chi-Square

To calculate the Chi-Square statistic, the observed frequencies are compared to the expected frequencies. The formula for the Chi-Square statistic is:

Chi-Square = Σ((Observed – Expected)^2 / Expected)

Where Observed is the observed frequency for each category and Expected is the expected frequency for each category.

Real World Example

Let me talk through a real-world example of the Chi-Square test to understand how it can help one determine if there's a relationship between obesity and heart failure rates. As a result, I used a sample of patients diagnosed with heart failure who had their body mass index (BMI) data to categorize them as obese or non-obese.

Now, to calculate the Chi-Square statistic, I created a contingency table showing the number of patients in each category for both obesity (based on BMI) and heart failure variables. After that I'm estimating the expected frequencies for each cell in this table assuming no association between these variables. In the end using the Chi-Square formula I compared the observed and expected frequencies to find if there was any significant association between the two variables.

If my calculated Chi-Square statistic value is greater than the critical value, I reject our null hypothesis that there's no link between obesity and heart failure. This indicates that obesity is indeed a risk factor for heart failure. Conducting such tests helps us gain valuable insights into relationships within a sample population and develop preventative measures to improve patient outcomes.

Now that we have seen how the process works in theory, let me show you practically, how the calculations and the process works:

Frequency of Heart Failure by Obesity

	Experienced Heart Failure	Did Not Experience Heart Failure
Obese	2245	1089
Not Obese	1431	428

Our Hypothesis

H0: Obesity and heart failure are independent
HA: Obesity and heart failure are not independent

Frequencies

	Experienced Heart Failure	Did Not Experience Heart Failure	Total
Obese	2245	1089	3334
Not Obese	1431	428	1859
Total	3676	1517	5193

Here, we calculated the total frequencies by summing up the observed frequencies.

To understand the number of obese patients who would not have undergone heart failure in our sample by chance, we will use the expected values. This is calculated by multiplying each row total by each column total, then dividing the result by the overall sample total. This will give us the expected values of obese patients who did not experience heart failure in our sample population.

Now, let's calculate the Chi-Square value using the below formula:

Chi-Square = Σ((Observed – Expected)^2 / Expected)

	Experienced Heart Failure	Did Not Experience Heart Failure
Obese	(2245 - 2360.06)2/2360.06	(1089-973.94)2/973.94
Not Obese	(1431-1315.94)2/1315.94	(428-543.05)2/543.05

And here are the results:

	Experienced Heart Failure	Did Not Experience Heart Failure
Obese	5.6	13.59
Not Obese	10.06	24.36

Finally, let's add all the values to find out Chi-Square

Chi-Square = 53.63

Now, we need to determine an alpha level for our test. Let's set 0.05 as an alpha level and find the critical value of Chi-Square (p). I used the Chi-Square calculator to calculate the p-value.

The p-value is less than .00001 which is, obviously, less than .05 (our alpha value)

Hence, the result is significant. In other words, we reject the null hypothesis which tells us that there is a relationship between Obesity and Heart Failure.

Read more: How to select best split in decision trees using Chi-Square?

Calculating Chi-Square in Python

Now that we did all the fun calculations manually, let's see if we can do the same using Python. As I previously mentioned, we will be using the scipy package's chi2_contingency function in Python to do this.

Step 1: Create a Contingency Table

I'm using the crosstab() function from the pandas library. I'm using Heart Failure to group by in the rows and the Obesity variable to group by in the columns. We also need to set margins to true to add row and column subtotals.

```
heartfailure_crosstab = pd.crosstab(df['obesity'], df['heart_failure'], margins=True, margins_name="subtotal")
```

It returns a contingency table that has this data:

	Experienced Heart Failure	Did Not Experience Heart Failure	Total
Obese	2245	1089	3334
Not Obese	1431	428	1859
Total	3676	1517	5193

Step 2: Compute Chi-Square and p-values

I used the scipy.stats.chi2_contingency function to calculate both my chi-square and p values.

To use this function, I use the following line:

```
chi, p, dof, expected = chi2_contingency(heartfailure_crosstab)
```

On a successful run of this above, the function returns the chi-square value to chi, p-value to p, degrees of freedom to dof, and expected values to expected variables respectively.

Step 3: Print P Value

In the above step we stored the output values to variables chi, p, dof, and expected respectively. To find out if p value in this calculation is less than our alpha value (0.05) we use the following command:

print(p)

The output of the above command will be:

0.0000000000004257

All I need to do now is to look at the p value and compare it with my alpha to make my conclusion. The above value is less than 0.0001 which is clearly less than 0.05 (our alpha value) and that allows us to conclude that the result is significant. Hence, we reject the null hypothesis which tells us that there is a relation between Obesity and Heart Failures.

Conclusion

The manual calculation of chi-square tests, as you have seen, requires quite a bit of time and manual effort, though Python's auto calculation using a command is much simpler and more efficient. In this article, I discussed what a chi-square test is, different types of chi-square tests, and how to perform a sample chi/square test. Additionally, we learned how to handle similar computations in Python with a single capability by doing so, saving time and effort.

5 Upcoming Python Libraries You Don't Want to Miss in 2023

Article Url - https://www.analyticsvidhya.com/blog/2023/02/discovering-insights-with-chi-square-tests-a-hands-on-approach-in-python/

Aashish