Member-only story

Chi-Square Test, with Python

The Complete Beginner's Guide to perform Chi-Square Test (with code!)

Chao De-Yu · Follow
Published in Towards Data Science
4 min read · May 22, 2021

Listen Share More

Photo by Kalen Emsley on Unsplash

In this article, I will introduce the fundamental of the chi-square test (χ2), a statistical method to make the inference about the distribution of a variable or to decide whether there is a relationship exists between two variables of a population.

The inference relies on the $\chi 2$ distribution curve, dependent upon the number of degrees of freedom d.f.

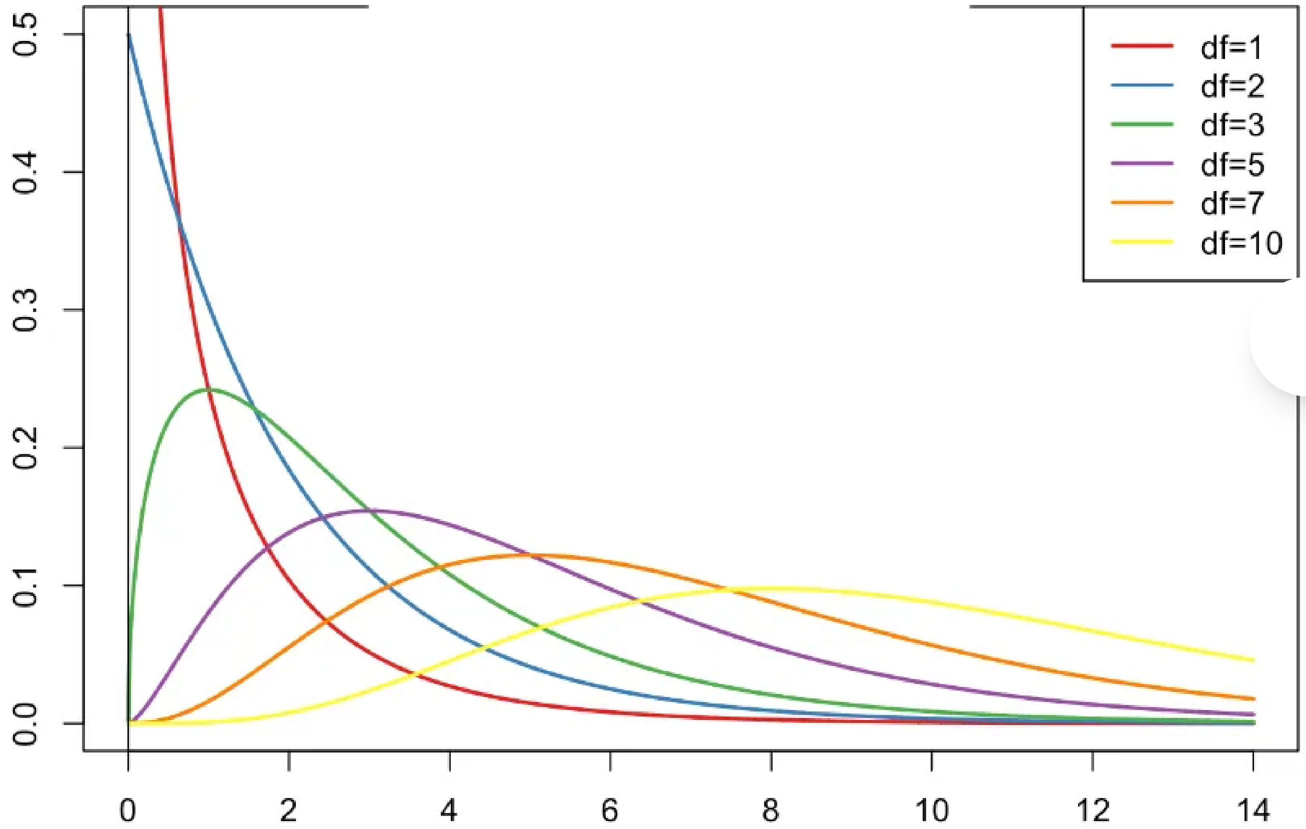

Figure 1: Chi-square distribution with different degree of freedom [1]

The $\chi 2$ distribution curve is right-skewed and as the number of degrees of freedom becomes larger, the $\chi 2$ curve will more similar to the normal distribution.

A: $\chi 2$ test of Independence

It is used to decide whether there is a relationship exists between two variables of a population. Useful when analyzing survey results of 2 categorical variables.

- H_0: The two categorical variables have **no relationship**
  H_1: There is **a relationship** between two categorical variables

- The number of degrees of freedom of the $\chi 2$ independence test statistics:
  d.f. = (# rows -1) *(#columns-1)

X_1 = row categorical variable with r levels X_2 = column categorical variable with c levels

Row Variable X_1	C				Row Totals
R_1	C				R_1 Total
R_2	O_{21}	O_{22}	.	O_{2c}	R_2 Total
. .	.	.	.	.	. .
R_r	O_{r1}	O_{r2}	.	O_{rc}	R_r Total
Column Totals	C_1 Total	C_2 Total	…	C_c Total	Grand Total

Table 1: **rxc** Contingency Table for 2 Categorical Variable

- If H_0 is true, each cell of the value in the contingency table above will contain a theoretical or **expected frequency E_{ij}**, as opposed to the **observed value O_{ij}** for each cell.

Assumed independent:

$$\mathbf{E_{ij}} = (R_i \cap C_j) = P(R_i) \times P(C_j) \times \text{Grand Total}$$

$$= \frac{R_i \text{ Total}}{\text{Grand Total}} \times \frac{C_j \text{ Total}}{\text{Grand Total}} \times \text{Grand Total}$$

$$= \frac{\mathbf{R_i\ Total \times C_j\ Total}}{\mathbf{Grand\ Total}}$$

Figure 2: Derivation of the expected frequency

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{(r-1)(c-1)}$$

over **ALL the cells** in the rxc contingency table

> *The table below is an exi… …o 2 categorical variables: people in categories from … >65 years, and their political affiliation, which is "Conservative", "Socialist" and "Other". Is there any evidence of a relationship between the age group and their political affiliation, at 5% significant level?*

	Conservative	Socialist	Other	Total
18-29	141	68	4	213
30-44	179	159	7	345
45-64	220	216	4	440
65 & older	86	101	4	191
Total	626	544	19	1189

Table 2: Exit poll survey [2]

According to five steps process of hypothesis testing:
H_0: whether age group and their political affiliation are independent, i.e. no relationship
H_1: whether age group and their political affiliation are dependent, i.e. ∃ a relationship
$\alpha = 0.05$
Following χ2 independence test statistics:

```
import pandas as pd
import scipy.stats as

# create sample data
data = [['18-29', 'Conservative'] for i in range(141)] + \
        [['18-29', 'Socialist'] for i in range(68)] + \
        [['18-29', 'Other'] for i in range(4)] + \
        [['30-44', 'Conservative'] for i in range(179)] + \
        [['30-44', 'Socialist'] for i in range(159)] + \
        [['30-44', 'Other'] for i in range(7)] + \
        [['45-65', 'Conservative'] for i in range(220)] + \
        [['45-65', 'Socialist'] for i in range(216)] + \
        [['45-65', 'Other'] for i in range(4)] + \
        [['65 & older', 'Conservative'] for i in range(86)] + \
        [['65 & older', 'Socialist'] for i in range(101)] + \
        [['65 & older', 'Other'] for i in range(4)]
df = pd.DataFrame(data, columns = ['Age Group', 'Political Affiliation'])

# create contingency table
data_crosstab = pd.crosstab(df['Age Group'],
                            df['Political Affiliation'],
                           margins=True, margins_name="Total")

# significance level
alpha = 0.05

# Calcualtion of Chisquare
chi_square = 0
rows = df['Age Group'].unique()
columns = df['Political Affiliation'].unique()
for i in columns:
    for j in rows:
        O = data_crosstab[i][j]
        E = data_crosstab[i]['Total'] * data_crosstab['Total'][j] / data_crosstab['Total']['Tot
        chi_square += (O-E)**2/E

# The p-value approach
print("Approach 1: The p-value approach to hypothesis testing in the decision rule")
p_value = 1 - stats.chi2.cdf(chi_square, (len(rows)-1)*(len(columns)-1))
conclusion = "Failed to reject the null hypothesis."
if p_value <= alpha:
    conclusion = "Null Hypothesis is rejected."

print("chisquare-score is:", chi_square, " and p value is:", p_value)
print(conclusion)

# The critical value approach
print("\n
```

```
print("\n------------------------------------------------------------------------------------
print("Approach 2: The                                                    g in the decision rule")
critical_value = stat                                                     )-1))
conclusion = "Failed t
if chi_square > criti
    conclusion = "Null Hypothesis is rejected."

print("chisquare-score is:", chi_square, " and critical value is:", critical_value)
print(conclusion)
```



```
Approach 1: The p-value approach to hypothesis testing in the decision rule
chisquare-score is: 24.367421717305202  and p value is: 0.0004469083391495099
Null Hypothesis is rejected.

--------------------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision rule
chisquare-score is: 24.367421717305202  and critical value is: 12.591587243743977
Null Hypothesis is rejected.
```

Conclusion: We have enough evidence that there is an association between age group and their political affiliation, at 5% significance level.

B: χ2 Goodness-Of-Fit Test

It is used to make the inference about the distribution of a variable.

- H_0: The variable has the specified distribution, normal
  H_1: The variable does not have the specified distribution, not normal
- The number of degrees of freedom of the χ2 Goodness-Of-Fit test statistics: d.f. = (# categories -1)
- It compares the **observed frequencies O** of a sample with the **expected frequencies E.**
  E = probability of the event * total sample size

> *The table below displays the more than 44 million people voting result for 2013 German Federal Election. 41.5% of German vote for the* Christian Democratic Union (CDU), *25.7% for the* Social Democratic Party (SPD) and *the remaining 32.8% as* Others.

> *Assume the researcher take a random sample and pick 123 students of FU Berlin about their party affiliation. O* [...] *PD and 40 for Others. These number correspond*

Party	Percentage	Relative frequency
CDU	41.5	0.415
SPD	25.7	0.257
Others	32.8	0.328
	100	1

Table 3: 2013 German Federal Election [3]

According to five steps process of hypothesis testing:
H_0: The variable has the specified distribution, i.e. the observed and expected frequencies are roughly equal
H_1: The variable does not have the specified distribution, not normal
$\alpha = 0.05$
Following χ2 Goodness-Of-Fit test statistics:

```
# Creation of data
data = [['CDU', 0.415 0]]
df = pd.DataFrame(dat d_freq'])
df['expected_freq'] =

# significance level
alpha = 0.05

# Calcualtion of Chisquare
chi_square = 0
for i in range(len(df)):
    O = df.loc[i, 'observed_freq']
    E = df.loc[i, 'expected_freq']
    chi_square += (O-E)**2/E

# The p-value approach
print("Approach 1: The p-value approach to hypothesis testing in the decision rule")
p_value = 1 - stats.chi2.cdf(chi_square, df['Varname'].nunique() - 1)
conclusion = "Failed to reject the null hypothesis."
if p_value <= alpha:
    conclusion = "Null Hypothesis is rejected."

print("chisquare-score is:", chi_square, " and p value is:", p_value)
print(conclusion)

# The critical value approach
print("\n--------------------------------------------------------------------------------
print("Approach 2: The critical value approach to hypothesis testing in the decision rule")
critical_value = stats.chi2.ppf(1-alpha, df['Varname'].nunique() - 1)
conclusion = "Failed to reject the null hypothesis."
if chi_square > critical_value:
    conclusion = "Null Hypothesis is rejected."

print("chisquare-score is:", chi_square, " and critical value is:", critical_value)
print(conclusion)
```



```
Approach 1: The p-value approach to hypothesis testing in the decision rule
chisquare-score is: 1.693614940576721  and p value is: 0.42878164729702506
Failed to reject the null hypothesis.

--------------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision rule
chisquare-score is: 1.693614940576721  and critical value is: 5.991464547107979
Failed to reject the null hypothesis.
```

Conclusion: We do not have enough evidence that the observed and expected frequencies are not equ

Recommended Reading

ANOVA Test, with Python

The Complete Beginner's Guide to perform ANOVA Test (with code!)

towardsdatascience.com

Two-Way ANOVA Test, with Python

The Complete Beginner's Guide to perform Two-Way ANOVA Test (with code!)

towardsdatascience.com

McNemar's Test, with Python

The Complete Beginner's Guide to perform McNemar's Test (with code!)

towardsdatascience.com

One-Sample Hypothesis Tests, with Python

The Complete Beginner Guide to perform One-Sample Hypothesis Tests (with code!)

levelup.gitconnected.com

Two-Sample Hypothesis Tests, with Python

The Complete Beginner Guide to perform Two-Sample Hypothesis Tests (with code!)

levelup.gitconnected.com

References

[1] "Chi-Square Tests • SOGA • Department of Earth Sciences." [Online]. Available: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Hypothesis-Tests/Ch Square-Tests/index.html

[2] "The Chi-Square Independence Test • SOGA • Department of Earth Sciences." [Online]. Available: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Hypothesis-Tests/Chi-Square-Tests/Chi-Square-Independence-Test/index.html

[3] "Chi-Square Goodness-of-Fit Test • SOGA • Department of Earth Sciences." [Online]. Available: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Hypothesis-Tests/Chi-Square-Tests/Chi-Square-Goodness-of-Fit-Test/index.html

Chi Square Test | Statistics | Data Science | Data Analysis

Follow

Written by Chao De-Yu

318 Followers · Writer for Towards Data Science

Data Analyst | MSc. Artificial Intelligence | LinkedIn — https://www.linkedin.com/in/thet-thet-yee-deyu/