

Independence Testing using Chi-Square Test

Abdul Khader, Syed

31 October 2022

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
import random

from scipy.stats import chi2_contingency, chi2

df = pd.read_excel("./data/grocery_database.xlsx", sheet_name = "campaign_data", engine = "openpyxl")
df = df[df['mailer_type'] != 'Control']
df.head()

```

	customer_id	campaign_name	campaign_date	mailer_type	signup_flag
0	74	delivery_club	2020-07-01	Mailer1	1
1	524	delivery_club	2020-07-01	Mailer1	1
2	607	delivery_club	2020-07-01	Mailer2	1
3	343	delivery_club	2020-07-01	Mailer1	0
4	322	delivery_club	2020-07-01	Mailer2	1

```
df.shape
```

```
(711, 5)
```

```
df.groupby("mailer_type")["signup_flag"].value_counts(normalize=True)
```

		signup_flag
mailer_type	signup_flag	
Mailer1	0	0.672000
	1	0.328000
Mailer2	0	0.622024
	1	0.377976

Test of Independence

using Chi-Square Test χ^2

Null Hypothesis

H_0 : There is **NO** significance relation between the *Mailer Type* and *User Signing Up*

Alternate Hypothesis

H_1 : There is a significance relation between the *Mailer Type* and *User Signing Up*

Creating the Contingency Table

```
contingency_table = pd.crosstab(df["mailer_type"], df["signup_flag"])
contingency_table
```

signup_flag	0	1
mailer_type		
Mailer1	252	123
Mailer2	209	127

```
contingency_mailer = contingency_table[1]
```

```
d1 = df[df.mailer_type == "Mailer1"]
mailer1_count = d1.shape[0]
```

```
d2 = df[df.mailer_type == "Mailer2"]
mailer2_count = d2.shape[0]
```

```
print("The signup rate for mailer1 is: " + str(contingency_mailer[0]/mailer1_count))
print("The signup rate for mailer2 is: " + str(contingency_mailer[1]/mailer2_count))
```

The signup rate for mailer1 is: 0.328

The signup rate for mailer2 is: 0.37797619047619047

```
stat,p,dof = chi2_contingency(contingency_table)[0:3]
confidence_level = 0.95
```

```
print("The chisquare stat is: ", round(stat, 3))
print("chi-square critical value is: ", round(chi2.ppf(confidence_level, df=dof), 3))
print("The chisquare p-value is: ", round(p,3))
```

```
The chisquare stat is:  1.728
chi-square critical value is:  3.841
The chisquare p-value is:  0.189
```

Conclusion

The p-value is *greater than 0.05* or *chisquare value is less than critical value*, we **CAN NOT REJECT** the null hypothesis.

Therefore, there is no significant relationship between the mailer type used(Fancy or Classical) and the Signing up of User.

Why are we using chi-square distribution in this case?

Here we have,

- Atleast one Categorical Variable(The type of Mail)
- Mutually Exculsive values of the Categorical Variable
- The observations are independent

How it looks with Guassian Distribution

We use gaussian distribution when we need to compate an observation variable with a number value like that of mean.