Academic Year 2024-25



# DEPARTMENT OF INFORMATION TECHNOLOGY

**COURSE CODE: DJS22ITL5013 DATE: 17-10-24**

**COURSE NAME: Statistical Analysis Lab CLASS: T.Y. BTech**

**NAME: Anish Sharma ROLL NO: I011**

**SAP ID: 60003220045**

# EXPERIMENT NO.08

**CO 2:** Perform Test of Hypothesis for independence and appropriateness of distribution using various statistical techniques.

**AIM / OBJECTIVE:** To implement chi-square test for independence and goodness of fit.

**DESCRIPTION OF EXPERIMENT:**

In statistics, there are two different usages of Chi-Square test:

1. The Chi-Square Goodness of Fit Test – Used to determine whether or not a categorical variable follows a hypothesized distribution.
2. The Chi-Square Test of Independence – Used to determine whether or not there is a significant association between two categorical variables.

Note that both of these tests are only appropriate to use when you’re working with categorical variables. These are variables that take on names or labels and can fit into categories. Examples include:

Eye color (e.g. “blue”, “green”, “brown”)

Gender (e.g. “male”, “female”)

Marital status (e.g. “married”, “single”, “divorced”)

• The Chi-Square Goodness of Fit Test

You should use the Chi-Square Goodness of Fit Test whenever you would like to know if some categorical variable follows some hypothesized distribution. • The Chi-Square Test of Independence

You should use the Chi-Square Test of Independence when you want to determine whether or not there is a significant association between two categorical variables.

Chi-Square Test of Independence: Formula

A Chi-Square test of independence uses the following null and alternative hypotheses:

H0: (null hypothesis) The two variables are independent.

H1: (alternative hypothesis) The two variables are not independent. (i.e. they are associated) We use the following formula to calculate the Chi-Square test statistic X2: X2 = Σ(O-E)2 / E where:

Σ: is a fancy symbol that means “sum” O: observed value

E: expected value

If the p-value that corresponds to the test statistic X2 with (#rows-1)\*(#columns-1) degrees of freedom is less than your chosen significance level then you can reject the null hypothesis. The formula for the chi-square goodness of fit test is:

χ2= ∑(Oi – Ei)2/Ei Steps:

1. Determine the expected numbers. For example, in genetics, you can use a Punnett square to calculate the theoretical expected values.

Academic Year 2024-25



1. Use the formula for each observed and expected category: ((Oi-Ei)^2/Ei)
2. Add the results together to get the final χ2 value.



1. Compare the calculated value to the critical value.

**INPUT DATA / DATASET:**

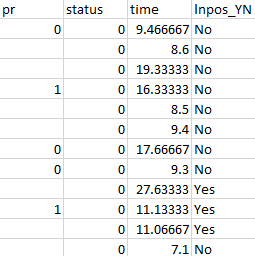
1. Select a dataset.
2. Apply chi-square test for independence and goodness of fit.

**SOURCE CODE:**

**DATASET 1:**

**Considered A Breast Cancer survival dataset.**

**The two categories considered are ‘Status’ and ‘lnpos\_YN (lymphatic node presence)’.**



import pandas as pd

from scipy.stats import chi2\_contingency

# Load the dataset

file\_path = "/content/BCSprep.csv"

data = pd.read\_csv(file\_path)

# Create a contingency table for 'lnpos\_YN' and 'status'

contingency\_table = pd.crosstab(data['lnpos\_YN'], data['status'])

# Perform the Chi-square test for independence

chi2\_stat, p\_value, dof, expected = chi2\_contingency(contingency\_table)

from scipy.stats import chi2

# Define degrees of freedom and significance level

dof = 1

alpha = 0.05

# Find the critical value using the inverse cumulative distribution function

critical\_value = chi2.ppf(1 - alpha, dof)

print("Critical value at alpha = 0.05 and df = 1:", critical\_value)

# Display the results

print("Contingency Table:")

print(contingency\_table)

print("\nChi-square Statistic:", chi2\_stat)

print("Degrees of Freedom:", dof)

print("P-value:", p\_value)

print("\nExpected Frequencies:")

print(expected)

# Decision based on the p-value

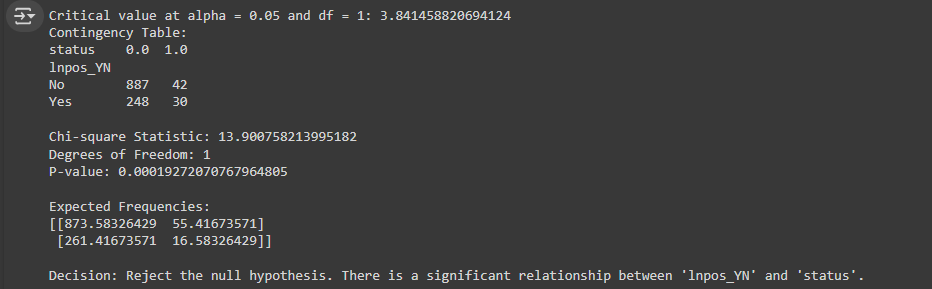
alpha = 0.05

if p\_value < alpha:

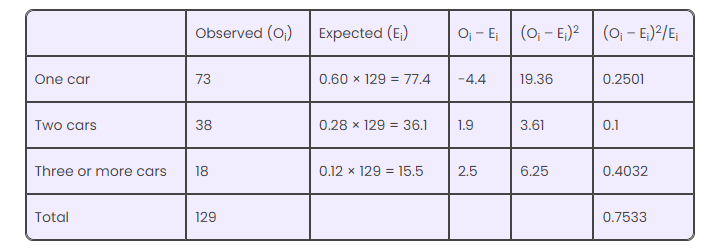
    print("\nDecision: Reject the null hypothesis. There is a significant relationship between 'lnpos\_YN' and 'status'.")

else:

    print("\nDecision: Fail to reject the null hypothesis. There is no significant relationship between 'lnpos\_YN' and 'status'.")



**DATASET 2:**



from scipy.stats import chi2

# Observed and expected frequencies from the provided data

observed = [73, 38, 18]

expected = [77.4, 36.1, 15.5]

# Calculate the Chi-square statistic manually

chi\_square\_stat = sum([(o - e) \*\* 2 / e for o, e in zip(observed, expected)])

# Calculate degrees of freedom (df)

# df = number of categories - 1

df = len(observed) - 1

# Calculate the p-value using the chi-square distribution's survival function (sf)

p\_value = chi2.sf(chi\_square\_stat, df)

# Define the significance level

alpha = 0.05

# Calculate the critical value for Chi-square at df and alpha

critical\_value = chi2.ppf(1 - alpha, df)

# Print the results

print("Chi-square statistic:", round(chi\_square\_stat, 4))

print("Degrees of Freedom:", df)

print("P-value:", round(p\_value, 4))

print("Critical Value at alpha = 0.05:", round(critical\_value, 4))

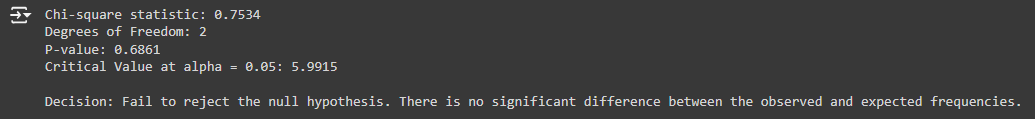
# Decision based on the p-value

if p\_value < alpha:

    print("\nDecision: Reject the null hypothesis. There is a significant difference between the observed and expected frequencies.")

else:

    print("\nDecision: Fail to reject the null hypothesis. There is no significant difference between the observed and expected frequencies.")



**CONCLUSION**:

In this experiment, we learnt to implement chi-square test for independence and goodness of fit.

**Website References:**

1. When to Use a Chi[-S](https://www.statology.org/when-to-use-chi-square-test/)quare Test (With Examples) - Statology
2. Chi-Square Test of Independence: Definition, Formula, and Example (statology.org)