SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

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: $X2 = \Sigma(O-E)2 / E$ 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:

$\chi 2 = \sum(O_i - E_i)^2/E_i$ Steps:

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

2. Use the formula for each observed and expected category: $((O_i-E_i)^2/E_i)$
3. Add the results together to get the final $\chi 2$ value.
4. 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)'.

pr	status	time	lnpos_YN
0	0	9.466667	No
	0	8.6	No
	0	19.33333	No
1	0	16.33333	No
	0	8.5	No
	0	9.4	No
0	0	17.66667	No
0	0	9.3	No
	0	27.63333	Yes
1	0	11.13333	Yes
	0	11.06667	Yes
	0	7.1	No

```
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'.")
```

```
Critical value at alpha = 0.05 and df = 1: 3.841458820694124
Contingency Table:
status    0.0  1.0
lnpos_YN
No        887   42
Yes       248   30

Chi-square Statistic: 13.900758213995182
Degrees of Freedom: 1
P-value: 0.00019272070767964805

Expected Frequencies:
[[873.58326429  55.41673571]
 [261.41673571  16.58326429]]

Decision: Reject the null hypothesis. There is a significant relationship between 'lnpos_YN' and 'status'.
```

DATASET 2:

	Observed (O_i)	Expected (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
One car	73	0.60 × 129 = 77.4	-4.4	19.36	0.2501
Two cars	38	0.28 × 129 = 36.1	1.9	3.61	0.1
Three or more cars	18	0.12 × 129 = 15.5	2.5	6.25	0.4032
Total	129				0.7533

```
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.")
```

```
Chi-square statistic: 0.7534
Degrees of Freedom: 2
P-value: 0.6861
Critical Value at alpha = 0.05: 5.9915

Decision: 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-Square Test (With Examples) - Statology
2. Chi-Square Test of Independence: Definition, Formula, and Example (statology.org)