Hypothesis Testing - V

Chi Square Test For Goodness of Fit

Chi-Square Distribution

- A Chi-square distribution is a continuous distribution with k degrees of freedom. They're widely used in hypothesis tests including the chi-square goodness of fit test and the chi-square test of independence.

- It is also used to test the goodness of fit of a distribution of data, whether data series are independent, and for estimating confidences surrounding variance and standard deviation for a random variable from a normal distribution.

- Chi-square distribution is a special case of the gamma distribution.

Chi-Square distribution Statistics

Notation	$\chi(k)$
Parameter	$k = 1,2,\ \dots$
Distribution	$x \geq 0$
Pdf	$\left(x^{\frac{k}{2}-1}e^{-\frac{x}{2}}\right)\ /\ \left(2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)\right)$
Cdf	$\gamma\left(\frac{k}{2},\frac{x}{2}\right)\ /\ \Gamma\left(\frac{k}{2}\right)$
Mean	k
Variance	$2k$
Skewness	$\sqrt{8\ /\ k}$
Kurtosis	$12\ /\ k$

Chi-Square table

Degrees of freedom *(df)*	χ^2 value [9]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Nonsignificant								**Significant**		

Chi-Square goodness of fit test

1. Define the hypothesis.
2. Calculate the Chi-square test statistic –

$$X^2 = \sum \frac{(O - E)^2}{E}$$

3. Find the degrees of freedom(*df*) - For chi-square goodness of fit tests, the *df* is the number of groups minus one.
4. Find Critical Chi-square value from the Chi-Square table using df and significance level.
5. Compare the chi-square value to the critical value i.e., if the Chi-square statistic value is greater, reject the null hypothesis, otherwise accept the null hypothesis.

When to use chi-square goodness of fit test

The following conditions are necessary if you want to perform a chi-square goodness of fit test:

- The **sample was randomly selected** from the population.
- There are a **minimum of five observations expected** in each group.

Example 1:

Problem Statement

A die was thrown 600 times and the following frequencies were observed

Face	1	2	3	4	5	6
Frequency	97	99	97	105	101	101

Test the hypothesis that the die is unbiased.

Example 1:

Solution

- Step1:Define the hypothesis

Null hypothesis H0 : die is unbiased

Alternative hypothesis H1 : die is biased

- Step 2: Set the Significance Level (alpha) to 0.05

Example 1:

- Step 3:Calculate Chi-Square Value

Observed(O)	Expected(E)	$(O-E)^2 / E$
97	100	0.09
99	100	0.01
97	100	0.09
105	100	0.25
101	100	0.01
101	100	0.01
600	600	0.46

Since we have assumed that die is unbiased, for each face we will take expected frequency as total no. of thrown divided by no. of face.

Example 1:

- Step 4:

  Calculate the Critical Value from the Chi square table with significance level 0.05 and degrees of freedom (6-1)=5 which is 11.07.

- Step 5:

  Now critical value is 11.07 and chi-value we got is 0.46

  As $0.46 < 11.07$

  Therefore we accept the Null hypothesis H0

Example 2:

Problem Statement

You have a bag of candies with different colors, and you want to test whether the observed distribution of candy colors in the bag matches the expected distribution. The expected distribution is provided by the candy manufacturer.

	Red	Green	Blue	Yellow	Orange
Expected Distribution	20%	30%	25%	15%	10%
Observed	38	72	60	25	5

Example 2:

Solution

- Step1:Define the hypothesis

Null Hypothesis (H0): The observed distribution of candy colors in the bag matches the expected distribution.

Alternative Hypothesis (H1): The observed distribution of candy colors in the bag does not match the expected distribution.

- Step 2: Set the Significance Level (alpha) to 0.05

Example 2:

- Step 3:Calculate Chi-Square Value

Observed(O)	Expected(E)	$(O-E)^2 / E$
38	0.20 * 200 = 40	00.10
72	0.30 * 200 = 60	02.40
60	0.25 * 200 = 50	02.00
25	0.15 * 200 = 30	00.83
5	0.10 * 200 = 20	11.25
200	200	16.58

Example 2:

- Step 4:

  Calculate the Critical Value from the Chi square table with significance level 0.05 and degrees of freedom (5-1)=4 which is 9.49.

- Step 5:

  Now critical value is 9.49 and chi-value we got is 16.58

  As 16.58 > 9.49

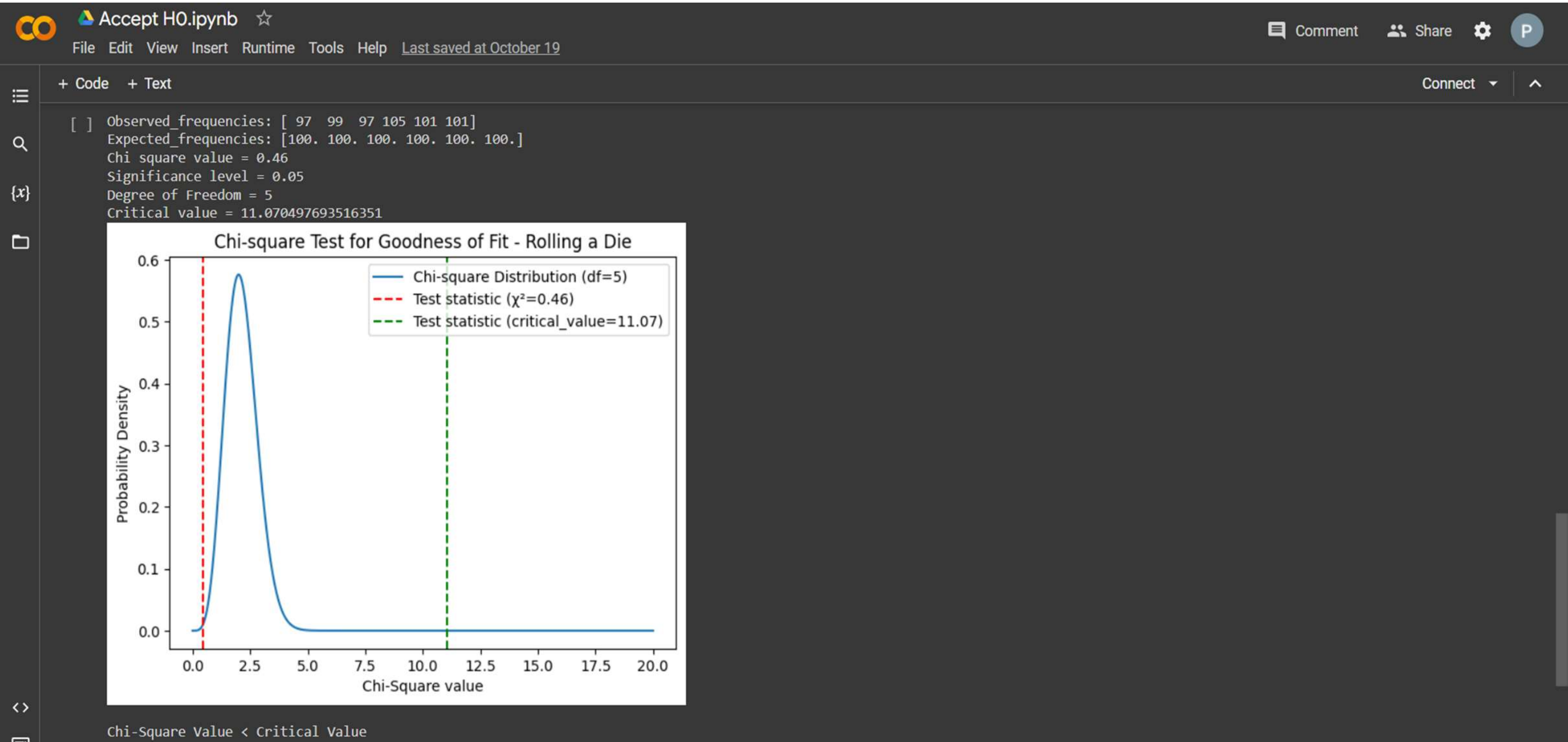
  Therefore we reject the Null hypothesis H0

```python
A die was thrown 600 times and the following frequencies were observed

Face:       1   2   3       4       5       6
Frequency: 97   99  97      105     101     101

Test the hypothesis that the die is unbiased.

'''
import numpy as np

# Define onserved frequencies ( from rolling the die 600 times )
observed_frequencies = np.array ( [ 97, 99, 97, 105, 101, 101 ] )
print(f'Observed_frequencies: {observed_frequencies}')

# Define expected frequencies ( uniform distribution for a fair die )
# Consider Null hypothesis  H0 : die is unbiased
# Alternative hypothesis   H1 : die is biased

# Since we have assumed that die is unbiased, for each face we will take frequency as total no. of thrown divided by no. of face.
avg = 600/6

expected_frequencies = np.array( [ 600/6, 600/6, 600/6, 600/6, 600/6, 600/6 ])
print(f'Expected_frequencies: {expected_frequencies}')

import scipy.stats as stats

# calculate the chi-square statistic
chi2, p_value = stats.chisquare( observed_frequencies, expected_frequencies )
print(f'Chi square value = {chi2}')
```

Code 1: Python code for Example 1

```python
critical_value = stats.chi2.ppf ( 1 - alpha, df)
print(f'Significance level = {alpha}')
print(f'Degree of Freedom = {df}')
print(f'Critical value = {critical_value}')


# Determin whether to accept or reject the null hypothesis
if chi2 < critical_value:
    result = "Accept H0: observed frequencies match the expected uniform distrubution"
else:
    result = "Reject H0: observed frequencies do not match the expected uniform distribution"

#Plot the chi- square distribution
import matplotlib.pyplot as plt

x = np.linspace(0,20,1000)
y = stats.chi.pdf(x, df)
plt.plot(x, y, label = f'Chi-square Distribution (df={df})')

# Mark the test statistics on the graph
plt.axvline(x=chi2, color = 'red', linestyle = '--', label = f'Test statistic (χ²={chi2:.2f})')
plt.axvline(x=critical_value, color = 'green', linestyle = '--', label = f'Test statistic (critical_value={critical_value:.2f})')
plt.legend()
plt.title('Chi-square Test for Goodness of Fit - Rolling a Die')
plt.xlabel("Chi-Square value")
plt.ylabel('Probability Density')
plt.show()

print(end='\n')
```

Code 1: Python code for Example 1



Output 1: Output of Code 1

+ Code + Text Connect

```python
'''
You have a bag of candies with different colors, and you want to test whether the
observed distribution of candy colors in the bag matches the expected distribution.
The expected distribution is provided by the candy manufacturer.
Red: 20%
Green: 30%
Blue: 25%
Yellow: 15%
Orange: 10%

'''

import scipy.stats as stats
import matplotlib.pyplot as plt
import numpy as np

# Step 1: Define the Hypotheses
# Null Hypothesis (H0): The observed distribution of candy colors in the bag matches the expected distribution.
# Alternative Hypothesis (H1): The observed distribution of candy colors in the bag does not match the expected distribution.

# Step 2: Set the Significance Level (alpha)
alpha = 0.05

# Step 3: Collect Data and Define Expected Frequencies
observed_frequencies = [38, 72, 60, 25, 5]  # Observed frequencies of candy colors
expected_frequencies = [0.20 * sum(observed_frequencies), 0.30 * sum(observed_frequencies), 0.25 * sum(observed_frequencies), 0.15 * sum(observed_frequencies), 0.10 * s

# Step 4: Perform the Chi-Square Test
chi2, p_value = stats.chisquare(observed_frequencies, expected_frequencies)

# Degrees of freedom
df = len(observed_frequencies) - 1
```

Code 2: Python code for Example 2

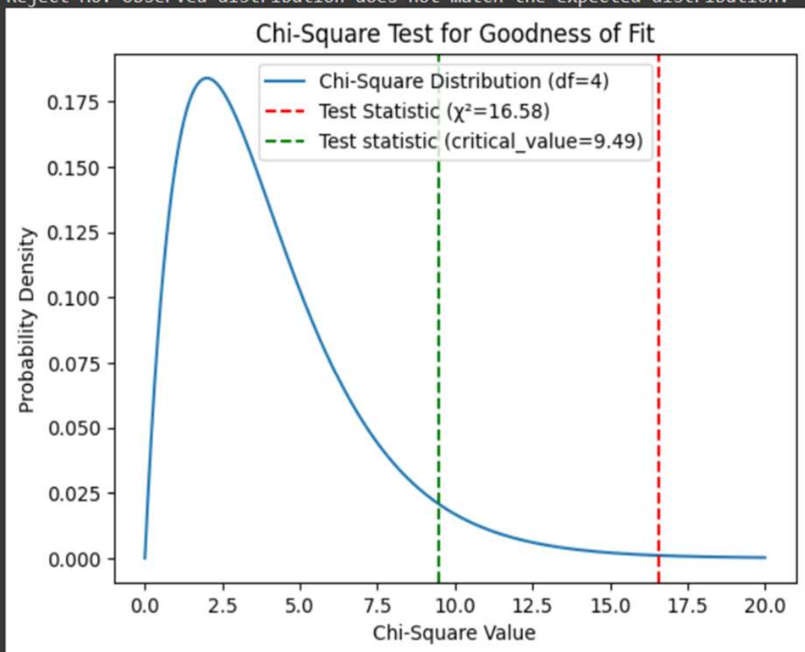
+ Code + Text Connect

```python
# Step 4: Perform the Chi-Square Test
chi2, p_value = stats.chisquare(observed_frequencies, expected_frequencies)

# Degrees of freedom
df = len(observed_frequencies) - 1

# Calculate the critical value
critical_value = stats.chi2.ppf(1 - alpha, df)

# Step 5: Make a Decision
if chi2 > critical_value:
    result = "Reject H0: Observed distribution does not match the expected distribution."
else:
    result = "Accept H0: Observed distribution matches the expected distribution."

# Step 6: Display the Results
print(f"Chi-Square Statistic: {chi2:.2f}")
print(f"P-Value: {p_value:.4f}")
print(f"Degrees of Freedom: {df}")
print(f"Critical Value: {critical_value:.2f}")
print(result)

# Step 7: Create a Probability Distribution Graph
x = np.linspace(0, 20, 1000)
y = stats.chi2.pdf(x, df)
plt.plot(x, y, label=f'Chi-Square Distribution (df={df})')

# Mark the test statistic on the graph
plt.axvline(x=chi2, color='red', linestyle='--', label=f'Test Statistic (χ²={chi2:.2f})')
plt.axvline(x=critical_value, color = 'green', linestyle = '--', label = f'Test statistic (critical_value={critical_value:.2f})')

plt.legend()
plt.title("Chi-Square Test for Goodness of Fit")
```

Code 2: Python code for Example 2

```
plt.xlabel("Chi-Square Value")
plt.ylabel("Probability Density")
plt.show()
```

```
Chi-Square Statistic: 16.58
P-Value: 0.0023
Degrees of Freedom: 4
Critical Value: 9.49
Reject H0: Observed distribution does not match the expected distribution.
```



Output 2: Output of Code 2

References

- YouTube: https://www.youtube.com/
- NITC Statistics Study Material

Team

- Prakash Singh(M210677CA)
- Debayan Ghosh(M210656CA)
- Prashant(M210704CA)

The End