# Analysis of Categorical Data

Bijay Lal Pradhan, PhD

# Analysis of Categorical Data

Categorical data refers to variables that represent distinct groups or categories, typically described using labels or names rather than numerical values.
Categorical data involves variables that indicate different categories, where each category represents a qualitative attribute.

# Example of Categorical Data

**Gender:** Male, Female
**Marital Status:** Single, Married, Divorced, Widowed
**Blood Type:** A, B, AB, O
**Country of Residence:** Nepal, India, USA, Canada
**Eye Color:** Blue, Brown, Green, Hazel
**Type of Pet:** Dog, Cat, Bird, Fish
**Education Level:** High School, Bachelor’s, Master’s, PhD
**Car Type:** Fuel, Electric
**Employment Status:** Employed, Unemployed, Student, Retired
**Favorite Fruit:** Apple, Banana, Orange, Mango

# Types of Categorical Variables

**Nominal Variables**: These variables represent categories with no inherent order or ranking. The categories are mutually exclusive and can't be ordered in any meaningful way. Examples include gender (male, female), blood type (A, B, AB, O), and eye color (blue, brown, green).

**Ordinal Variables**: These variables represent categories with a meaningful order or ranking, but the intervals between the categories are not consistent or known. Examples include education level (high school, bachelor's, master's, PhD) and customer satisfaction (satisfied, neutral, dissatisfied).

# Hypothesis test for single variables (Univariate)

Gender: Male and Female

Hypothesis: Whether proportion of male and female are equal?

Faculty: Science, Management, Humanities, Engineering

Hypothesis: Whether the proportion is 2:5:3:1 in respective faculty?

Birth of Baby in Month: Baisakh, Jestha, ……., Chaitra

Hypothesis: birth months matches a uniform distribution (each month equally likely).

# Example

You have surveyed 200 people in a population and want to test whether the proportion of males and females is equal.

**Observed Data**:

- Males: 120
- Females: 80

**Hypotheses**:

- **Null Hypothesis ($H_0$)**: The proportion of males and females is equal (50% male and 50% female).
- **Alternative Hypothesis ($H_1$)**: The proportion of males and females is not equal.

# Example

### Calculate the Expected Frequencies

Under the null hypothesis, you expect 50% males and 50% females in the sample of 200 people.

- Expected number of males = 50% of 200 = 100
- Expected number of females = 50% of 200 = 100

### Perform the Chi-Square Goodness-of-Fit Test

The Chi-Square statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(120 - 100)^2}{100} + \frac{(80 - 100)^2}{100} = \frac{800}{100} = 8$$

# Example

Critical values of chi-square (right tail)

| Degrees of freedom (*df*) | Significance level (α) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | .99 | .975 | .95 | .9 | .1 | .05 | .025 | .01 |
| 1 | ------- | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 |

## Determine Critical Value

- Degrees of freedom (df) = number of categories - 1 = 2 - 1 = 1
- Look up the critical value for Chi-Square with 1 degree of freedom at the 0.05 significance level: 3.841.

## Make a Decision

- Since the calculated Chi-Square statistic (8) is greater than the critical value (3.841), you reject the null hypothesis.

## Conclusion:

There is significant evidence to conclude that the proportion of males and females in the population is not equal.

# Question

You have surveyed 400 students and want to test if the distribution across four faculties matches the ratio 2:5:3:1.

**Observed Data:**

- Science: 80
- Management: 200
- Humanities: 90
- Engineering: 30

Expected Ratio: 2:5:3:1 total 11 (out of 80+200+90+30 = 400)

Calculate the expected frequencies:

$$\text{Science} = \frac{2}{11} \times 400 = 72.73$$

$$\text{Management} = \frac{5}{11} \times 400 = 181.82$$

$$\text{Humanities} = \frac{3}{11} \times 400 = 109.09$$

$$\text{Engineering} = \frac{1}{11} \times 400 = 36.36$$

**Observed Data:**

Science: 80 Management: 200 Humanities: 90 Engineering: 30 Total = 400

$$\text{Science} = \frac{2}{11} \times 400 = 72.73 \qquad \text{Management} = \frac{5}{11} \times 400 = 181.82$$

$$\text{Humanities} = \frac{3}{11} \times 400 = 109.09 \qquad \text{Engineering} = \frac{1}{11} \times 400 = 36.36$$

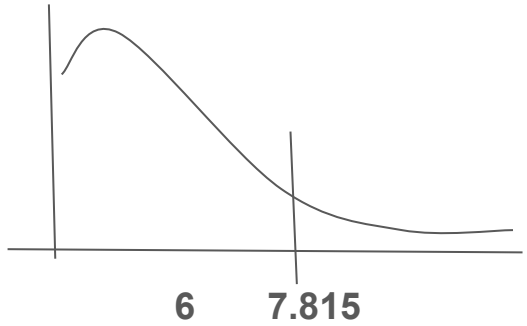
Calculated Chi Square

$$\chi^2 = \frac{(80-72.73)^2}{72.73} + \frac{(200-181.82)^2}{181.82} + \frac{(90-109.09)^2}{109.09} + \frac{(30-36.36)^2}{36.36} = \frac{52.95}{72.73} + \frac{331.72}{181.82} + \frac{364.43}{109.09} + \frac{40.45}{36.36} = 6$$

Tabulated Chi Square: $\chi^2_{0.05,3}$ = 7.815 (Critical Value)

Decision: Since $\chi^2$calculated value is < $\chi^2$table.

H1 is rejected. I.e. The the distribution across four faculties matches the ratio 2:5:3:1.



# Example 3

### Define the Hypotheses

- **Null Hypothesis (Ho)**: The births are uniformly distributed across the 12 months. This means each month has an equal number of births.
- **Alternative Hypothesis (H1)**: The births are not uniformly distributed across the 12 months.

Assume you have the observed frequencies of births for each month as follows:

| Month | Baisakh | Jeshta | Asar | Shrawan | Bhadra | Aswin | Kartik | Mangsir | Poush | Magh | Falgun | Chaitra |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Births | 45 | 50 | 48 | 47 | 42 | 44 | 53 | 46 | 41 | 49 | 47 | 48 |

**Expected Frequencies** $[E_i = \frac{\sum O_i}{12} = \frac{600}{12} = 50]$

**Chi-square Statistic:**

$$[\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(45-50)^2}{50} + \frac{(50-50)^2}{50} + \cdots + \frac{(48-50)^2}{50}]$$

$$[\chi^2 = \frac{25}{50} + \frac{0}{50} + \frac{4}{50} + \frac{9}{50} + \frac{64}{50} + \frac{36}{50} + \frac{9}{50} + \frac{16}{50} + \frac{81}{50} + \frac{1}{50} + \frac{9}{50} + \frac{4}{50}]$$

$$\chi^2 \approx 5.16$$

Tabulated Chi Square: $\chi^2_{0.05,11}$ = 19.675 (Critical Value)

Decision: Since $\chi^2$calculated value is < $\chi^2$ table.

H1 is rejected. i.e. The births are uniformly distributed across the 12 months.

| Month | Baisakh | Jeshta | Asar | Shrawan | Bhadra | Aswin | Kartik | Mangsir | Poush | Magh | Falgun | Chaitra |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Births | 45 | 50 | 48 | 47 | 42 | 44 | 53 | 46 | 41 | 49 | 47 | 48 |

# Contingency Table

A contingency table (also known as a cross-tabulation or cross-tab) is a type of table in a matrix format that displays the frequency distribution of variables. It is commonly used in statistics to explore the relationship between two or more categorical variables by showing how the variables interact with each other.

Components of Contingency

**Rows and Columns:** For each variables

**Cells:**Each cell in the table shows the frequency or count of occurrences

**Marginal Totals:**The sum of the frequencies in each row and each column.

**Grand Total:** The sum of all the frequencies in the table, representing the total number of observations.

# Example of contingency table

| Gender \ Preference | Prefer Product | Not Prefere Product | Row Total |
|---|---|---|---|
| Male | 40 | 10 | 50 |
| Female | 30 | 30 | 60 |
| Column Total | 70 | 40 | 110 |

The above table is called as contingency table
And it is 2x2 contingency table ( has two option in row and two option in column)

# Chi-square test of independence

The "test of independence" in the context of a contingency table is a statistical method used to determine whether two categorical variables are independent of each other. This is commonly tested using the **Chi-square test of independence**.

1. **Null Hypothesis (Ho)**: The two variables are independent. In other words, there is no association between the variables.
2. **Alternative Hypothesis (H1)**: The two variables are not independent. There is an association between the variables.

P(A&B)=P(A)*P(B) = $\frac{N(A \cap B)}{N} = \frac{N(A)}{N} \times \frac{N(B)}{N}$ A & B independent

$$E_{ij} = \frac{(\text{Row Total for row } i) \times (\text{Column Total for column } j)}{\text{Grand Total}}$$

# Example

| Gender \ Preference | Prefer Product | Not Prefere Product | Row Total |
|---|---|---|---|
| Male | 40 | 10 | 50 |
| Female | 30 | 30 | 60 |
| Column Total | 70 | 40 | 110 |

**Null Hypothesis (Ho)**: Gender and product preference are independent. There is no association between gender and product preference.

**Alternative Hypothesis (H1)**: Gender and product preference are not independent. There is an association between gender and product preference.

Calculation of Chi Square Value

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(40 - 31.82)^2}{31.82} + \frac{(10 - 18.18)^2}{18.18} + \frac{(30 - 38.18)^2}{38.18} + \frac{(30 - 21.82)^2}{21.82}$$

$$\chi^2 \approx 2.10 + 3.68 + 1.75 + 3.07 \approx 10.60$$

$$E_{11} = \frac{50 \times 70}{110} \approx 31.82$$

$$E_{12} = \frac{50 \times 40}{110} \approx 18.18$$

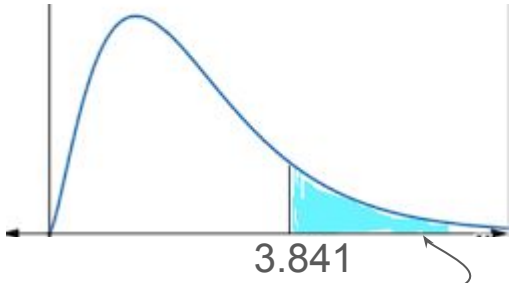
$$E_{21} = \frac{60 \times 70}{110} \approx 38.18$$

$$E_{22} = \frac{60 \times 40}{110} \approx 21.82$$

Degree of freedom: **(R-1)*(C-1) = (2-1)*(2-1) = 1**

Level of significance: **α = 0.05**

Chi Square table value $\chi^2_{0.05,1}$ **= 3.841**




$\chi^2$ Calculated value = **10.6**

Decision: Since $\chi^2$ calculated value is > $\chi^2$ table value. Ho is rejected. I.e. Gender and product preference are not independent. There is an association between gender and product preference.

In case of 2x2 contingency table (df=1) we can apply the formula below to find $\chi^2$

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

where df = 1

| | | | Row Total |
|---|---|---|---|
| | A | B | A+B |
| | C | D | C+D |
| Column T | A+C | B+D | N=A+B+C+D |

Suppose we are investigating the relationship between customer satisfaction (measured on a 3-point scale: Satisfied, Neutral, Dissatisfied) and frequency of visits to a restaurant (measured on a 4-point scale: Daily, Weekly, Monthly, Rarely).

| Frequency of visit | Daily | Weekly | Monthly | Rarely | Total |
|---|---|---|---|---|---|
| Satisfied | 30 | 40 | 50 | 20 | 140 |
| Neutral | 20 | 25 | 30 | 25 | 100 |
| Dissatisfied | 10 | 15 | 10 | 25 | 60 |
| Total | 60 | 80 | 90 | 70 | 300 |

## Expected Frequencies ??

| Frequency of visit | Daily | Weekly | Monthly | Rarely | Total |
|---|---|---|---|---|---|
| Satisfied | | | | | 140 |
| Neutral | | | | | 100 |
| Dissatisfied | | | | | 60 |
| Total | 60 | 80 | 90 | 70 | 300 |

| Frequency of visit | Daily | Weekly | Monthly | Rarely | Total |
|---|---|---|---|---|---|
| Satisfied | 30 | 40 | 50 | 20 | 140 |
| Neutral | 20 | 25 | 30 | 25 | 100 |
| Dissatisfied | 10 | 15 | 10 | 25 | 60 |
| Total | 60 | 80 | 90 | 70 | 300 |

Observed Frequencies

| Frequency of visit | Daily | Weekly | Monthly | Rarely | Total |
|---|---|---|---|---|---|
| Satisfied | | | | | 140 |
| Neutral | | | | | 100 |
| Dissatisfied | | | | | 60 |
| Total | 60 | 80 | 90 | 70 | 300 |

Expected Frequencies

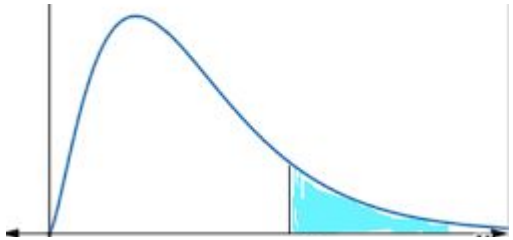
| Frequency of visit | Daily | Weekly | Monthly | Rarely | Total |
|---|---|---|---|---|---|
| Satisfied | | | | | |
| Neutral | | | | | |
| Dissatisfied | | | | | |

$$[\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi square calculated value

Chi square tabulated value

Decision



# Yate's Correction

| Gender \ Preference | Prefer Product | Not Prefere Product | Row Total |
|---|---|---|---|
| Male | 9.5 | 4.5 | 14 |
| Female | 30.5 | 29.5 | 60 |
| Column Total | 40 | 34 | 74 |

| Gender \ Preference | Prefer Product | Not Prefere Product | Row Total |
|---|---|---|---|
| Male | 10 | 4 | 14 |
| Female | 30 | 30 | 60 |
| Column Total | 40 | 34 | 74 |

# Contingency table with less than 5 expected frequency

| Gender \ Preference | Prefer Product | Not Prefere Product | Row Total |
|---|---|---|---|
| Male | 9.5 | 4.5 | 14 |
| Female | 30.5 | 29.5 | 60 |
| Column Total | 40 | 34 | 74 |

We have to use Yate's Correction