Date / /
Page No.

Chi-Square Test for Contingency Tables

The Chi-Square test determines whether there is a significant relationship between nominal (categorical) variables.

Null Hypothesis (H_0):
There is no relationship between the two categorical variables (they are independent).

Steps to compute Chi-square (χ^2):

1. Compute Expected frequencies

Each expected frequency is calculated using

$$E_{i,j} = \frac{T_i \cdot T_j}{T}$$

where

$E_{i,j}$ = expected frequency for a specific cell

T_i = row total

T_j = column total

T = grand total

Example: Diet and Health study

Data from the Mediterranean Diet and Health study

Diet	Cancers	Fatal Heart Disease	Nonfatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

Example calculation (AHA Diet & Cancers):

Proportion of total cases with cancer $\frac{22}{605} = 0.0364$

Date / /
Page No.

Expected cases for AHA diet $(0.864) \times (303) = 11.02$

Diet	Cancers (C)	Fatal Heart Disease (F)	Non-Heart Disease (N)	Healthy (H)	Total
AHA	15 (11.02)	24 (9.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

2 Compute Chi-Square Statistic

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

for this study: $\chi^2 = 16.55$

3 Compute degrees of freedom (df)

$$df = (r-1)(c-1)$$

for 2 diets and 4 health outcomes:

$$(2-1)(4-1) = 3$$

4 Find the p-value

Using a Chi-square distribution table, for $\chi^2 = 16.55$ and $df = 3$

P-value $= 0.0009$

Conclusion:

since $p < 0.05$, we reject the null hypothesis. This indicates a significant relationship between diet and health outcomes

Chi-Square Assumptions:

1 Each subject contributes to only one cell (eg Anagram pattern example is invalid)

2 Sample size should be ≥ 20 for valid results.

3 Yates's correction is ~~not~~ recommended for 2 x 2 tables

Date / /
Page No.

Chi Square Distribution and Goodness of fit

Chi-Square Distribution: Key Concepts

The Chi-square distribution (χ^2) arises from the sum of squared standard normal deviates and is widely used in statistical tests.

1 Definition: If Z is a standard normal variable (mean = 0, variance = 1), then the Chi-Square variable with K degrees of freedom is

$$\chi^2(K) = Z_1^2 + Z_2^2 + \dots + Z_K^2$$

Degrees of freedom (df) = the number of squared normal variables summed.

Example: $\chi^2(1)$ is simply a squared normal deviate

2 Properties of the Chi-Square distribution

Mean = df

Variance = 2 * df

Shape: Positively skewed, skewness decreases as df increases

Approximation to Normal: When df is large, the χ^2 distribution approaches a normal distribution

3 Example calculation

suppose we sample two standard normal scores, square them and sum them:

$$\chi^2(2) = Z_1^2 + Z_2^2$$

we want $P(\chi^2(2) \geq 6)$

Using a Chi-Square calculator: $p = 0.050$

Date / /
Page No.

4) Applications of chi-square
Goodness of fit test (one way tables)
Contingency tables (relationship between categorical variables)
Many other statistical tests are based on the chi-square distribution

Chi-Square Goodness-of-Fit Test
The Chi-Square Goodness of Fit Test determines whether observed data significantly deviate from expected frequencies under a specific theoretical distribution

1) Hypothesis Setup
Null Hypothesis (H_0): The observed data follow the expected distribution
Alternative Hypothesis (H_1): The observed data significantly differ from the expected distribution.

2) Steps to compute Chi-Square
- Compute Expected Frequencies
If all categories are equally likely, the expected frequency for each is E = Total count × Expected Proportion
- Compute Chi-Square Statistic:

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

Where O is the observed frequency, and E is the expected frequency

Date / /
Page No.

Determine Degrees of freedom (df) : $df = K - 1$
where K is the number of categories.

Compare to critical value/ compute p-value:
If $p < 0.05$, reject H_0 (significant difference)

Introduction to Linear Regression
Linear Regression is a statistical method used to pred the value of one variable (Y, the criterion variable) ba on another variable (X, the predictor variable). When only one predictor variable is used, it is called simple linear regression

Understanding Regression Lines
The best fitting line in regression is the one that minim the sum of squared prediction errors
The formula for regression line is:

$$Y' = bX + A$$

where; Y' = Predicted value of Y, b = Slope of the line
A = Y-intercept

Example Dataset

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Plotting these points show a positive relationship between X and Y.

Date / /
Page No.

2 Computing the Regression line
Using statistics:
Mean of X (M_x) = 3
Means of Y (M_y) = 2.06
Standard deviation of X (S_x) = 1.581
Standard deviation of Y (s_y) = 1.072
Correlation (r) = 0.627
The slope (b) is calculated as

$$b = \frac{r s_y}{s_x} = \frac{(0.627)(1.072)}{1.581} = 0.425$$

The intercept (A) is:

$$A = M_y - bM_x = 2.06 - (0.425)(3) = 0.785.$$

Thus the regression equation is:

$$Y' = 0.425X + 0.785.$$

3 Errors of Prediction
The error of prediction for each point is:

$$\text{Error} = Y - Y'$$

The sum of squared errors determines the best fitting line.

X	Y	Y'	Error (Y-Y')	(Error)^2
1	1.00	1.21	-0.21	0.044
2	2.00	1.64	0.36	0.133
3	1.30	2.06	-0.76	0.578
4	3.75	2.49	1.26	1.600
5	2.25	2.91	-0.66	0.436

4 Standardized Regression Equation
If X and Y are standardized (mean=0, standard deviation=1)
the regression equation simplifies to:

$$z'_y = r z_x$$

Date / /
Page No.

where $z_{y'}$ is the predicted standard score and r is the correlation.

5. Real Example : Predicting University GPA from High school GPA

Regression Equation:

University of GPA' = (0.675)(High school GPA) + 1.0

For a high school GPA of 3, predicted University GPA is

(0.675)(3) + 1.097 = 3.12

6. Assumptions of Regression :

The method works without assumptions, but inferential regression statistics require.

- Linearity : Relationship between X and Y is linear.
- Homoscedasticity : Variance of residuals is constant.
- Independence : Observations are independent of each other
- Normality : Errors follow a normal distribution