Statistical Tests in Election Analysis with Examples

Dr. Ratnesh Prasad Srivastava, CSIT, GGV, Bilaspur

September 17, 2025

1 Introduction

This document explains the statistical tests used in election analysis, including why and how each test is applied to understand voting patterns and predict election outcomes. Each test is accompanied by a practical example and solution.

2 Chi-square Test

2.1 Why it's used

The Chi-square test is used to determine if there is a significant association between two categorical variables. In election analysis, it helps identify relationships between voter demographics (age, gender, income) and voting preferences.

2.2 How it's used

The test compares observed frequencies in contingency tables with expected frequencies under the null hypothesis of independence. A significant result indicates that the variables are associated.

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

Where O_i are observed frequencies and E_i are expected frequencies.

2.3 Example: Age Group vs. Party Preference

Suppose we want to test if there's a relationship between age groups and preference for a particular political party. We survey 500 voters and get the following results:

	BJP	Congress	AAP	Others	Total
18-30	40	30	50	20	140
31-45	60	40	30	20	150
46-60	70	35	20	15	140
60+	50	40	10	10	110
Total	220	145	110	65	540

Test the hypothesis at = 0.05 that age group and party preference are independent.

2.4 Solution

2.4.1 Step 1: Set up hypotheses

H: Age group and party preference are independent

H: Age group and party preference are not independent

2.4.2 Step 2: Calculate expected frequencies

For each cell: $E_{ij} = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$

2.4.3 Step 3: Compute Chi-square statistic

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

2.4.4 Step 4: After calculations, we find ² = 32.45

2.4.5 Step 5: Degrees of freedom

Degrees of freedom = (rows - 1) × (columns - 1) = (4-1) × (4-1) = 9

2.4.6 Step 6: Critical value

Critical value for = 0.05 with 9 df is 16.92

2.4.7 Step 7: Decision

Since 32.45 ¿ 16.92, we reject the null hypothesis.

2.4.8 Conclusion

There is a significant association between age group and party preference.

3 T-test

3.1 Why it's used

The T-test compares the means of two groups to determine if they are statistically different. In election analysis, it might be used to compare support levels for a candidate between two regions or demographic groups.

3.2 How it's used

The test calculates a t-value based on the difference between means, accounting for variability and sample size. A significant t-value suggests a real difference between groups.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where $\bar{X}_1$ and $\bar{X}_2$ are sample means, s_p is the pooled standard deviation, and n_1, n_2 are sample sizes.

3.3 Example: Urban vs Rural Support for a Candidate

Suppose we want to compare support for a candidate between urban and rural areas. We survey 30 urban voters and 35 rural voters, with the following results:

Urban: Mean support = 65%, Standard deviation = 8%, Sample size = 30

Rural: Mean support = 58%, Standard deviation = 10%, Sample size = 35

Test at = 0.05 if there's a significant difference in support between urban and rural areas.

3.4 Solution

3.4.1 Step 1: Set up hypotheses

H: _urban = _rural (no difference in support)

H: _urban _rural (difference in support)

3.4.2 Step 2: Calculate pooled standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(30-1)8^2 + (35-1)10^2}{30+35-2}} = \sqrt{\frac{29 \times 64 + 34 \times 100}{63}} = \sqrt{\frac{1856 + 3400}{63}} = \sqrt{\frac{5256}{63}} = \sqrt{83.43} = 9.13$$

3.4.3 Step 3: Compute t-statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{65 - 58}{9.13\sqrt{\frac{1}{30} + \frac{1}{35}}} = \frac{7}{9.13\sqrt{0.0333 + 0.0286}} = \frac{7}{9.13\sqrt{0.0619}} = \frac{7}{9.13 \times 0.2488} = \frac{7}{2.27} = 3.08$$

3.4.4 Step 4: Degrees of freedom

Degrees of freedom = n + n - 2 = 30 + 35 - 2 = 63

3.4.5 Step 5: Critical t-value

Critical t-value for = 0.05 with 63 df is approximately 2.00

3.4.6 Step 6: Decision

Since —3.08— ¿ 2.00, we reject the null hypothesis.

3.4.7 Conclusion

There is a significant difference in support for the candidate between urban and rural areas.

4 ANOVA

4.1 Why it's used

Analysis of Variance (ANOVA) is used to compare means across three or more groups. In election analysis, it could determine if voting patterns differ significantly across multiple age groups, income brackets, or regions.

4.2 How it's used

ANOVA partitions the total variance into between-group and within-group components. The F-ratio tests whether between-group variance is significantly larger than within-group variance.

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{MS_{between}}{MS_{within}}$$

4.3 Example: Support for a Policy Across Income Groups

Suppose we want to compare support for a new policy across three income groups (low, middle, high). We survey voters and get the following support percentages:

Low income: 45, 48, 42, 50, 47 (n=5, Mean=46.4)
Middle income: 55, 58, 60, 57, 55 (n=5, Mean=57.0)
High income: 70, 72, 68, 75, 70 (n=5, Mean=71.0)
Test at = 0.05 if there's a significant difference in support across income groups.

4.4 Solution

4.4.1 Step 1: Set up hypotheses

H: _low = _middle = _high (no difference in support)
H: At least one mean is different

4.4.2 Step 2: Calculate overall mean

Grand mean = (46.4 + 57.0 + 71.0)/3 = 58.13

4.4.3 Step 3: Calculate Sum of Squares Between (SSB)

SSB = n_i(mean_i - grand_mean)^2 = 5×(46.4-58.13)^2 + 5×(57.0-58.13)^2 + 5×(71.0-58.13)^2
SSB = 5×(-11.73)^2 + 5×(-1.13)^2 + 5×(12.87)^2 = 5×137.59 + 5×1.28 + 5×165.64 = 687.95 + 6.40 + 828.20 = 1522.55

4.4.4 Step 4: Calculate Sum of Squares Within (SSW)

SSW = (x_ij - mean_i)^2
Low: (45-46.4)^2 + (48-46.4)^2 + (42-46.4)^2 + (50-46.4)^2 + (47-46.4)^2 = 1.96 + 2.56 + 19.36 + 12.96 + 0.36 = 37.2
Middle: (55-57)^2 + (58-57)^2 + (60-57)^2 + (57-57)^2 + (55-57)^2 = 4 + 1 + 9 + 0 + 4 = 18
High: (70-71)^2 + (72-71)^2 + (68-71)^2 + (75-71)^2 + (70-71)^2 = 1 + 1 + 9 + 16 + 1 = 28
SSW = 37.2 + 18 + 28 = 83.2

4.4.5 Step 5: Calculate Mean Squares

MSB = SSB / (k-1) = 1522.55 / 2 = 761.28
MSW = SSW / (N-k) = 83.2 / (15-3) = 83.2 / 12 = 6.93

4.4.6 Step 6: Calculate F-statistic

F = MSB / MSW = 761.28 / 6.93 = 109.85

4.4.7 Step 7: Compare with critical F-value

Critical F-value (=0.05, df=2, df=12) = 3.89

4.4.8 Step 8: Decision

Since 109.85 ¿ 3.89, we reject the null hypothesis.

4.4.9 Conclusion

There is a significant difference in policy support across income groups.

5 Regression Analysis

5.1 Why it's used

Regression analysis models the relationship between a dependent variable and one or more independent variables. In election analysis, it predicts voting behavior based on demographic factors, past voting patterns, or economic indicators.

5.2 How it's used

The method estimates coefficients that represent the relationship between predictors and the outcome variable. It helps identify which factors most influence voting decisions and to what extent.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where y is the dependent variable, x_i are independent variables, β_i are coefficients, and ϵ is the error term.

5.3 Example: Predicting Voting Probability

Suppose we want to predict the probability of voting for a particular party based on income, education, and age. We collect data from 100 voters and run a multiple regression analysis.

Dependent variable: Probability of voting for Party X (0-1 scale)

Independent variables:

- Income (in thousands)
- Education (years of schooling)
- Age (in years)

After running the regression, we get the following output:

Variable	Coefficient	Std. Error	t-value	p-value
Intercept	0.15	0.05	3.00	0.003
Income	0.002	0.001	2.00	0.048
Education	0.025	0.008	3.13	0.002
Age	-0.003	0.001	-3.00	0.003

$R^2 = 0.45$, Adjusted $R^2 = 0.43$, F-statistic = 15.8 (p ¡ 0.001)

Interpret the results and predict the voting probability for a voter with income = 60,000, education = 16 years, and age = 45.

5.4 Solution

5.4.1 Step 1: Interpret the coefficients

- Intercept (0.15): The baseline probability when all independent variables are zero
- Income (0.002): For each additional $1,000 income, voting probability increases by 0.002
- Education (0.025): For each additional year of education, voting probability increases by 0.025
- Age (-0.003): For each additional year of age, voting probability decreases by 0.003

5.4.2 Step 2: Check statistical significance

All variables have p-values ¡ 0.05, indicating they are statistically significant predictors.

5.4.3 Step 3: Assess model fit

$R^2 = 0.45$ means 45% of the variance in voting probability is explained by the model.

5.4.4 Step 4: Make a prediction

For a voter with income = 60, education = 16, age = 45:

$$y = 0.15 + 0.002 \times 60 + 0.025 \times 16 - 0.003 \times 45$$

$$y = 0.15 + 0.12 + 0.40 - 0.135 = 0.535$$

5.4.5 Conclusion

The predicted probability of this voter supporting Party X is 53.5%.

The model suggests that education has the strongest positive effect on voting probability, while age has a negative effect.