3.3 Inferring association between categorical variables - Chi-squared test for Independence

Monday, 10 October 2022 18:00

Summary

- Independence of variables
- Chi-Square distribution
  - Formulae for:
    - Expected frequency
    - Chi-Square statistic
    - df

Inferencing about association

Example: Brand preferences

- Suppose a survey is conducted in Mumbai and Chennai asking respondents their preferences about three brands. The result is summarized below.

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	279	73	225	577
Chennai	165	47	191	403
Total	444	120	416	980

- Independent (explanatory) variable is the city.
- Dependent (response) variable is the brand preference.

- There are two categorical variables here:

1. Brand (A, B, C), and
2. City (Mumbai, Chennai)

- City => Independent (explanatory) variable
- Brand => Dependent (response) variable

Example: Brand preferences

- We know how to summarize the data by calculating the marginal and joint probabilities.
- What are the marginal probabilities? Joint probabilities?
- Now we want to answer the question: "Whether brand preference associated with city?" We use the basis of statistical independence/dependence for this.
- Two categorical variables are statistically independent if the population conditional distributions on one of them is identical to each category of the other.
- In the example, the two conditional distributions are not identical. e.g. Brand A is preferred more in Mumbai than in Chennai.

- Joint and Marginal Probabilities

	Brand A	Brand B	Brand C	
Mumbai	0.28	0.07	0.24	0.59
Chennai	0.17	0.05	0.19	0.41
	0.45	0.12	0.42	1

- Conditional Distribution: P(City | Brand)

	Brand A	Brand B	Brand C	
Mumbai	279 (48%)	73 (13%)	225 (39%)	577 (100%)
Chennai	165 (41%)	47 (12%)	191 (47%)	403 (100%)

- From conditional distribution, we can observe:
  - Brand A is preferred more in Mumbai than in Chennai
  - Brand B preference is identical in both cities
  - Brand C is preferred more in Chennai than in Mumbai

- Since the conditional distributions are not identical, so we conclude that brand preference is associated with city.
- Hence, both categorical variables are dependent on each other.

<u>How to find conditional distribution</u>

Given a joint distribution or contingency table:

	Y1	**Y2**	**Y3**	
X1	*	*	*	**A**
X2	*	*	*	**B**
	K	**L**	**M**	**(total)**

- Find Conditional Distribution

$$\mathbf{P}\left(\mathbf{Y} \mid \mathbf{X}\right) = \frac{\mathbf{P}(\mathbf{X},\mathbf{Y})}{\mathbf{P}(\mathbf{X})}$$

	Y1	Y2	Y3	
X1	***/A**	***/A**	***/A**	**A**
X2	***/B**	***/B**	***/B**	**B**

and

$$\mathbf{P}\left(\mathbf{X} \mid \mathbf{Y}\right) = \frac{\mathbf{P}(\mathbf{X},\mathbf{Y})}{\mathbf{P}(\mathbf{Y})}$$

	Y1	Y2	Y3

X1	*/K	*/L	*/M
X2	*/K	*/L	*/M
	K	L	M

- Two variables are independent if conditional distributions on one of them is identical to each category of the other.

P (Y | X)

	Y1	Y2	Y3	
X1	*/A	*/A	*/A	A
X2	*/B	*/B	*/B	B

P (X | Y)

	Y1	Y2	Y3
X1	*/K	*/L	*/M
X2	*/K	*/L	*/M
	K	L	M

- That is, you find any one of the conditional distributions, and the values in same shaded cells should be equal.
- Then we say both variables are independent of each other.

Example: Brand preferences

- Refer to the same example extended to a third city:

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Chennai	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Delhi	110 (44%)	35 (14%)	105 (42%)	250 (100%)

- Conditional distributions is same across the cities. Hence we can conclude that brand preference is independent of the cities.
- However, statistical independence is a symmetric property between two categorical variables.

- Here, brand preference does not depend on city.
- This is a sample data.
- Statistical independence is a symmetric property, so:
  - If brand preference is independent of city, P(City | Brand) = P(City), then
  - City is also independent of the brand, P(Brand | City) = P(Brand)
    - Proof:

	Brand A	Brand B	Brand C
Mumbai	440 (74%)	140 (74%)	420 (74%)
Chennai	44 (7%)	14 (7%)	42 (7%)
Delhi	110 (19%)	35 (19%)	105 (19%)
Total	594 (100%)	189 (100%)	567 (100%)

- Conclusion:
  - If X is independent of Y, then
  - Y is also independent of X

Example: Brand preferences

- If the conditional distributions within the rows are identical, then so are the distributions within the columns.
- One can verify that the conditional distribution amongst columns equals (74%, 7%, 19%).

- However, the example was a sample data. What about the population?
- Based on this single sample information, can we draw inferences about the population, as we have been doing?
- Answer is in testing our hypothesis, of course!

- Expected frequency = ?
- Can you tell why we assume the variables to be independent in the null hypothesis?

Chi-square distribution

- Null hypothesis –

H_0: The categorical variables are independent.

- Alternate hypothesis –

H_1: The categorical variables are not independent.

Let f_o be the observed frequencies (from the sample)

Let f_e be the expected frequencies, if the variables were independent.

The expected frequency for a cell equals the product of row and column totals for that cell, divided by the total sample size.

- Null hypothesis is always the no effect null hypothesis. Alternate hypothesis says the opposite thing.
- f_e, the expected frequencies, are calculated assuming that the null hypothesis is true.
- $Expected\ frequency,\ f_e = \frac{\textbf{Row total} \times \textbf{Column total}}{\textbf{Total Sample size}}$

- Chi-square formula

Example: Brand preference

- Brand preference example, with expected frequencies in brackets for each cell.

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	279 (261.4)	73 (70.7)	225 (244.9)	577
Chennai	165 (182.6)	47 (49.3)	191 (171.1)	403
Total	444	120	416	980

- Chi-squared test statistic:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}.$$

- One example:
  - $261.4 = \frac{444 \times 577}{980}$

- $\frac{(f_o - f_e)^2}{f_e}$

	Brand A	Brand B	Brand C
Mumbai	1.185	0.075	1.617
Chennai	1.696	0.107	2.314

- Chi-square formula:

- $\boldsymbol{\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}}$

$\chi^2 = 6.994 \approx 7$

- How do we calculate df in a contingency table?

Chi-square distribution

- When the H_0 is true, expected and observed frequencies tend to be close for each cell, and the test statistic value is relatively small.
- If H_0 is false, at least some cells have a big gap between expected and observed frequencies, leading to a large test statistic value.
- The larger the χ^2 value, greater is the evidence against the null hypothesis of independence.
- Degrees of freedom for the chi-squared distribution is given by the expression: *df = (r-1)*(c-1)*. *r* and *c* are the # of rows and columns respectively.

- **Degrees of freedom**,

$$Degrees\ of\ freedom = (Number\ of\ rows\ - 1) \times (Number\ of\ columns\ - 1)$$

$$\boldsymbol{df = (r - 1) \times (c - 1)}$$

- Given tabular and calculated chi-squared statistic, when do we accept H_0?

Chi-square distribution

- For the brand preference example, calculated test statistic value is the $\chi^2 = 7.0$.
- Degrees of freedom *df = 2*. So at *α = 0.05* (95% confidence), the tabular value of test statistic, $\chi^2 = 5.99$.
- So we reject the null hypothesis of independence.
- However, at *α = 0.01* (99% confidence), the tabular value of test statistic, $\chi^2 = 9.21$, and we can not reject the null hypothesis.

- $df = (2-1) \times (3-1) = 1 \times 2 = 2$
- Calculated Chi-square statistic: $\chi^2 = 7$
- At: $df = 2$ and $\alpha = 0.05\ (95\%\ confidence)$
  - Tabular Chi-squared statistic: $\chi^2 = 5.99$

    Tabular value < Calculated value

    » We reject the null hypothesis

  - Conclusion: cities and brand preferences are dependent.
- At: $df = 2$ and $\alpha = 0.01\ (99\%\ confidence)$
  - Tabular Chi-squared statistic: $\chi^2 = 9.21$

    Tabular value ≥ Calculated value

    » We accept the null hypothesis

  - Conclusion: cities and brand preferences are independent.

- When we are concluding about hypotheses, we're essentially inferencing about the entire population.