Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

ANALYSIS OF CATEGORICAL VARIABLES

ANALYSIS OF CATEGORICAL VARIABLES

- Pearson's χ^2 (chi-squared) test is used to test hypotheses regarding the distribution of a categorical variable or the relationship between two categorical variables.
- Pearson's χ^2 test uses a test statistic, which we denote as Q, to measure the discrepancy between the observed data and what we expect to observe under the null hypothesis.
- Higher levels of discrepancy between data and H_0 results in higher values of Q.
- We use q to denote the observed value of Q based on a specific sample of observed data.

Pearson's χ^2 Test for One Categorical Variable

- Let us denote the binary variable of interest as X, based on which we can divide the population into two groups depending on whether $X = 1$ or $X = 0$.
- Further, suppose that the null hypothesis H_0 states that the probability of group 1 is $\boldsymbol{\mu_{01}}$ and the probability of group 2 is $\boldsymbol{\mu_{02}}$.
  - Here $\mu_{02} = 1 - \mu_{01}$.
- If the null hypothesis is true, we expect that, out of n randomly selected individuals, $E_1 = n\mu_{01}$ belong to the first group, and $E_2 = n(1 - \mu_{01})$ belong to the second group.
- We refer to E_1 and E_2 as the expected frequencies under the null.
- We refer to the observed number of people in each group as the observed frequencies and denote them $\boldsymbol{O_1}$ and $\boldsymbol{O_2}$ for group 1 and group 2, respectively.

Pearson's χ^2 Test for One Categorical Variable

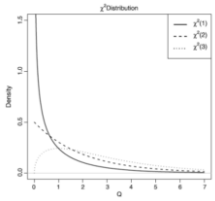
- Pearson's $\chi 2$ test measures the discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies as follows:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

- The value of Q will be zero only when the observed data matches our expectation under the null exactly.
- When there is some discrepancy between the data and the null hypothesis, Q becomes greater than zero.
- The higher discrepancy between our data and what is expected under H_0, the larger Q and therefore the stronger the evidence against H_0.

Pearson's χ^2 Test for One Categorical Variable

- If the null hypothesis is true, then the approximate distribution of Q is χ^2.
- Like the t-distribution, the χ^2-distribution is commonly used for hypothesis testing and denoted $\chi^2(df)$.
- The plot of the pdf for a χ^2-distribution with various degrees of freedom



- The observed significance level p_{obs} is calculated using the χ^2-distribution with 1 degree of freedom.
- This corresponds to the upper tail probability of q from the $\chi^2(1)$ distribution.

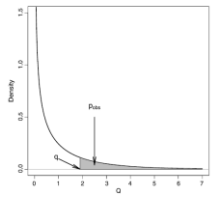
Example

- We use the heart attack survival rate (i.e., the probability of survival after heart attack) within one year after hospitalization.
- Suppose that $\boldsymbol{H_0}$ specifies that the probability of surviving is $\mu_{01} = 0.70$ and the probability of not surviving is $\mu_{02} = 0.30$.
- If we take a random sample of size $n = 40$ from the population, we expect that $E_1 = 0.70 \times 40 = 28$ and $E_2 = 0.30 \times 40 = 12$.
- Now suppose that the observed number of people in each group as the observed frequencies: $O_1 = 24$ and $O_2 = 16$.
- For the heart attack survival example, the observed value of the test statistic is

$$q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(24 - 28)^2}{28} + \frac{(16 - 12)^2}{12} = 1.90$$

- The $p_{obs} = P(Q \geq 1.90) = 0.17$ is obtained from a χ^2-distribution with 1 degree of freedom

Example



- The sampling distribution for Q under the null hypothesis: $Q \sim \chi^2(1)$.
- The p-value is the upper tail probability of observing values as extreme or more extreme than $q = 1.90$

- Therefore, the results are not statistically significant, and we cannot reject the null hypothesis at commonly used significance levels (e.g., 0.01, 0.05, and 0.1).
- In this case, we believe that the difference between observed and expected frequencies could be due to chance alone.

Categorical Variables with Multiple Categories

- Pearson's χ^2 test can be generalized to situations where the categorical random variable can take more than two values.
- In general, for a categorical random variable with k possible categories, we calculate the test statistic Q as

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

- The approximate distribution of Q is $\chi 2$ with the degrees of freedom equal to $df = k - 1$.
- Therefore, to find p_{obs}, we calculate the upper tail probability of q (the observed value of Q) from the $\chi^2(k-1)$ distribution.

Example

- Suppose that we monitor heart attack patients for one year and divide them into three groups:
  - patients who did not have another heart attack and survived,
  - patients who had another heart attack and survived,
  - patients who did not survive.
- Suppose that $\mu_{01} = 0.5$, $\mu_{02} = 0.2$, and $\mu_{03} = 0.3$.
- The expected frequencies of each category for a sample of $n = 40$:

  $E_1 = 0.5 \times 40 = 20$, $E_2 = 0.2 \times 40 = 8$, $E_3 = 0.3 \times 40 = 12$.
- This time, suppose that the actual observed frequencies based on a sample of size $n = 40$ for the three groups are

  $O_1 = 13$, $O_2 = 11$, $O_3 = 16$.

Example

- The amount of discrepancy:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}$$

- The observed value of this test statistic:

$$q = \frac{(13 - 20)^2}{20} + \frac{(11 - 8)^2}{8} + \frac{(16 - 12)^2}{12} = 4.91$$

- Using R-Commander, we find $p_{obs} = P(Q \geq 4.91) = 0.086$ using the χ^2-distribution with 2 degrees of freedom.
- Therefore, we can reject the null hypothesis at 0.1 level but not at 0.05 level.
- At the 0.1 significance level, we can conclude that the difference between observed and expected frequencies is statistically significant, and it is probably not due to chance alone.

Pearson's χ^2 Test of Independence

- We now discuss the application of Pearson's χ^2 test for evaluating a hypothesis regarding possible relationship between two categorical variables.
- More specifically, we measure the difference between the observed frequencies and expected frequencies under the null.
- The null hypothesis in this case states that the two categorical random variables are independent.
  - For two independent random variables, the joint probability is equal to the product of their individual probabilities.
    - In what follows, we use this rule to find the expected frequencies.

Pearson's χ^2 Test of Independence

- We use the following general form of Pearson's χ^2 test, which summarizes the differences between the expected frequencies (under the null hypothesis) and the observed frequencies over all cells of the contingency table:

$$Q = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} and E_{ij} are the observed and expected values in the ith row and jth column of the contingency table.
- The double sum simply means that we add the individual measures of discrepancies for cells by going through all cells in the contingency table.

13

Pearson's χ^2 Test of Independence

- As before, higher values of Q provide stronger evidence against H_0.
- For $I \times J$ contingency tables (i.e., I rows and J columns), the Q statistic has approximately the χ^2-distribution with $(I-1)\times(J-1)$ degrees of freedom under the null.
- Therefore, we can calculate the observed significance level by finding the upper tail probability of the observed value for Q, which we denote as q, based on the χ^2-distribution with $(I-1)\times(J-1)$ degrees of freedom.

14

Example 1

- The probability that the mother is smoker (i.e., smoke =1) and the baby has low birthweight (i.e., low =1) is the product of smoker and low-birthweight probabilities.
- For the baby weight example, we can summarize the observed and expected frequencies in the contingency tables.

	Observed frequency			Expected frequency	
	Normal	Low		Normal	Low
Nonsmoking	86	29	Nonsmoking	79.1	35.9
Smoking	44	30	Smoking	50.9	23.1

15

Example 1

- Then Pearson's test statistic is

$$Q = \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}}$$

$$q = \frac{(86-79.1)^2}{79.1} + \frac{(29-35.9)^2}{35.9} + \frac{(44-50.9)^2}{50.9} + \frac{(30-23.1)^2}{23.1} = 4.9$$

- Since the table has $I = 2$ rows and $J = 2$ columns, the approximate null distribution of Q is χ^2 with $(2-1)\times(2-1) = 1$ degrees of freedom.
- Consequently, the observed p-value is the upper tail probability of 4.9 using the $\chi^2(1)$ distribution.
- We find $p_{obs} = P(Q \geq 4.9) = 0.026$.
- Therefore, at the 0.05 significance level (but not at 0.01 level), we can reject the null hypothesis that the mother's smoking status and the baby's birthweight status are independent.

16

Example 2

- Suppose that we would like to investigate whether the race of mothers is related to the risk of having babies with low birthweight.
- The race variable can take three values: 1 for white, 2 for African-American, and 3 for others.
- As before, the low variable can take 2 possible values: 1 for babies with birthweight less than 2.5 kg and 0 for other babies.
- Therefore, all possible combinations of race and low can be presented by a 3×2 contingency table.
- The following Table provides the observed frequency of each cell and the expected frequency of each cell if the null hypothesis is true.

17

Example 2

	Observed frequency			Expected frequency	
Groups	Normal (low=0)	Low (low=1)	Groups	Normal (low=0)	Low (low=1)
1	73	23	1	66	30
2	15	11	2	18	8
3	42	23	3	46	21

- For example, there are 73 babies in the first row and first column.
- This is the number of babies in the intersection of race = 1 (mother is white) and low = 0 (having a baby with normal birthweight).
- If the null hypothesis is true, the expected number of babies in this cell would have been 66.

18

Example 2

- The observed value of the test statistic Q is obtaned as $q = 5.0$ using the following equations.

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} + \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}}$$

- The distribution of Q under the null hypothesis is χ^2 with $(3 - 1) \times (2 - 1) = 2$ degrees of freedom.
- To find the corresponding p-value, we need to find the probability of observing values as or more extreme than 5.0.
- This is the upper-tail probability of 5 from the $\chi^2(2)$ distribution: $p_{obs} = P(Q \geq 5)$.
- The value of p_{obs} is 0.08.
- We can reject the null hypothesis at 0.1 level but not at 0.05 level.
- At 0.05 level, the relationship between the two variables (i.e., race of mothers and birthweight status) is not statistically significant.