STA258
University of Toronto Mississauga
Analysis of Categorical Data

Al Nosedal, Asal Aslemand, Omid Jazi
University of Toronto.
Winter 2023

March 14, 2023

We now consider applications of the very important chi-square statistic, first proposed by Karl Pearson in 1900.
So we can get some idea as to why Pearson first proposed his chi-square statistic, we begin with the Binomial case. That is, let Y_1 be $Bin(n, p_1)$, where $0 < p < 1$. According to the Central Limit Theorem,

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

has a distribution that is approximately $N(0, 1)$ for large n, particularly when $np_1 \geq 5$ and $n(1 - p_1) \geq 5$.

Thus, it is not surprising that $Q_1 = Z^2$ is approximately $\chi^2(1)$. If we let $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$, we see that Q_1 may be written as

$$Q_1 = \frac{[Y_1 - np_1]^2}{np_1(1-p_1)} = \frac{[Y_1 - np_1]^2}{np_1} + \frac{[Y_2 - np_2]^2}{np_2}$$

or

$$Q_1 = \sum_{i=1}^{2} \frac{[Y_i - np_i]^2}{np_i}$$

To generalize, we let an experiment have k (instead of only two) mutually exclusive and exhaustive outcomes, say, $A_1, A_2, ..., A_k$. Let $p_i = P(A_i)$, and thus $\sum_{i=1}^{k} p_i = 1$. The experiment is repeated n independent times, and we let Y_i represent the number of times the experiment results in $A_i, i = 1, 2, ..., k$. This joint distribution of $Y_1, Y_2, ..., Y_{k-1}$ is a generalization of the Binomial distribution.

The joint pmf of Y_1, $Y_2, ..., Y_{k-1}$ is

$$f(y_1, y_2, ..., y_{k-1}) = \frac{n!}{y_1!y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

$y_1 \geq 0, y_2 \geq 0, \cdots, y_{k-1} \geq 0,\ y_1 + y_2 + ... + y_{k-1} \leq n$.
We say that Y_1, $Y_2, ..., Y_{k-1}$ have a **multinomial distribution** with parameters n and $p_1, p_2, ..., p_{k-1}$.

Pearson then extended Q_1 to an expression that we denote by Q_{k-1},

$$Q_{k-1} = \sum_{i=1}^{k} \frac{[Y_i - np_i]^2}{np_i}$$

He argued that Q_{k-1} has an approximate chi-square distribution with $k-1$ degrees of freedom in a similar way we argued that Q_1 is approximately $\chi^2(1)$.

Multinomial Response Model

- When a categorical (qualitative) variable of interest results in one of two responses ($k = 2$) (e.g., “Yes” or “No” to had epidural to reduce pain during child birth, “Yes” or “No” still breastfeeding at six months) the data ? called counts ? can be analyzed with a Binomial probability distribution.
- Categorical data with more than two levels (classes, categories) often result from a multinomial model.
- Binomial model is a multinomial model with $k = 2$.

Properties of the Multinomial Model

1. The are n identical trials.
2. There are k (or: $r \times c$ "r for number of rows" and "c for number of columns") possible outcomes to each trial. These outcomes are sometimes called classes, categories, or cells.
3. The probabilities of the k outcomes, denoted by $p_1, p_2, p_3, ..., p_k$, where $p_1 + p_2 + p_3 + ... + p_k = 1$, remain the same from trial to trial.
4. The trials are independent.
5. The random variables of interest are the cell counts $n_1, n_2, ..., n_k$ of the number of observations that fall into each of the k categories.

Example

Suppose that customers can purchase one of the three brands of milk at a supermarket. In a study to determine whether one brand is preferred over another, a record is made of a sample of $n = 300$ milk purchases. The data are shown below. Do the data provide sufficient evidence to indicate a preference for one or more brands?

Brand 1	Brand 2	Brand 3	Total
78	117	105	300

Step 1. State Hypotheses.

If all the brands are **equally** preferred, then the probability that a purchaser will choose any one brand is the same as the probability of choosing any other - that is, $p_1 = p_2 = p_3 = 1/3$. Therefore, the null hypothesis of "no preference" is

$$H_0 : p_1 = p_2 = p_3 = 1/3$$

If p_1, p_2, and p_3 are not all equal, the brands are not equally preferred; in other words, the purchasers must have a preference for one (or possibly) two brands. The alternative hypothesis is

$$H_a : p_1, p_2, \text{ and } p_3 \text{ are not all equal}$$

Therefore, we seek a test statistic that will detect a **lack of fit** of the observed **cell counts** to our hypothesized (null) expected cell counts based on the hypothesized cell probabilities.

These expected values are:

$E(n_1) = np_1 = (300)\left(\frac{1}{3}\right) = 100$

$E(n_2) = np_2 = (300)\left(\frac{1}{3}\right) = 100$

$E(n_3) = np_3 = (300)\left(\frac{1}{3}\right) = 100$

Table of Expected Counts

Brand 1	Brand 2	Brand 3	Total
100	100	100	300

Step 2. Computing test statistic

The test statistic for comparing the observed and expected cell counts (and, consequently, testing $H_0 : p_1 = p_2 = p_3 = 1/3$ is the **$\mathbf{X^2}$ statistic**:

$$X^2 = \sum_{all\ \ cells} \frac{(\text{observed - expected})^2}{\text{expected}} = \sum_{all\ \ cells} \frac{(n_i - E(n_i))^2}{E(n_i)}$$

$$X^2 = \frac{(78-100)^2}{100} + \frac{(117-100)^2}{100} + \frac{(105-100)^2}{100}$$

$$X^2 = 4.84 + 2.89 + 0.25 = 7.98$$

Step 3. Finding P-value

To find the P-value, compare X^2 with critical values from the chi-square distribution with degrees of freedom one fewer than the number of values the brand can take. That's $3 - 1 = 2$ degrees of freedom. From Table D, we see that $X^2 = 7.98$ falls between 0.02 and 0.01 critical values of the chi-square distribution with 2 degrees of freedom. So the P-value of $X^2 = 7.98$ is between 0.01 and 0.02 $(0.01 < P - value < 0.02)$.

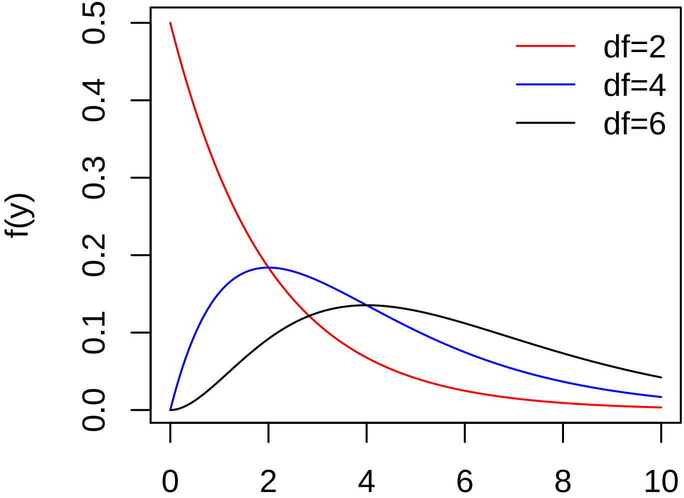
Step 4. Conclusion

If we used $\alpha = 0.05$, since our $P - value < \alpha = 0.05$, we could reject H_0 at the 5% significance level. We would conclude that the three brands of milk are **not** equally preferred.

Chi-Square Distributions

The **chi-square distributions** are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by giving its **degrees of freedom**.

Helpful fact: the mean of any chi-square distribution is equal to its degrees of freedom.


f(y)
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
df=2
df=4
df=6

Example

Raymond Weil is about to come out with a new watch and wants to find out whether people have special preferences of the color of the watchband, or whether all four colors under consideration are equally preferred. A random sample of 80 prospective watch buyers is selected. Each person is shown the watch with four different band colors and asked to state his or her preference. The results (observed counts) are given below.

Tan	Brown	Maroon	Black	Total
12	40	8	20	80

Use $\alpha = 0.01$.

Step 1. State Hypotheses.

If all the brands are **equally** preferred, then the probability that a purchaser will choose any one color is the same as the probability of choosing any other - that is, $p_1 = p_2 = p_3 = p_4 = 1/4$. Therefore, the null hypothesis of "no preference" is

$$H_0 : p_1 = p_2 = p_3 = p_4 = 1/4$$

If p_1, p_2, p_3 and p_4 are not all equal, the colors are not equally preferred. The alternative hypothesis is

$$H_a : p_1, p_2, p_3 \text{ and } p_4 \text{ are not all equal}$$

Therefore, we seek a test statistic that will detect a **lack of fit** of the observed **cell counts** to our hypothesized (null) expected cell counts based on the hypothesized cell probabilities.

These expected values are:

$E(n_1) = np_1 = (80)\left(\frac{1}{4}\right) = 20$
$E(n_2) = np_2 = (80)\left(\frac{1}{4}\right) = 20$
$E(n_3) = np_3 = (80)\left(\frac{1}{4}\right) = 20$
$E(n_4) = np_4 = (80)\left(\frac{1}{4}\right) = 20$

Table of Expected Counts

Tan	Brown	Maroon	Black	Total
20	20	20	20	80

Step 2. Computing test statistic

The test statistic for comparing the observed and expected cell counts (and, consequently, testing $H_0 : p_1 = p_2 = p_3 = p_4 = 1/4$ is the $\mathbf{X^2}$ **statistic**:

$$X^2 = \sum_{all\ \ cells} \frac{\text{(observed - expected)}^2}{\text{expected}} = \sum_{all\ \ cells} \frac{(n_i - E(n_i))^2}{E(n_i)}$$

$$X^2 = \frac{(12-20)^2}{20} + \frac{(40-20)^2}{20} + \frac{(8-20)^2}{20} + \frac{(2-20)^2}{20}$$

$$X^2 = 64/20 + 400/20 + 144/20 + 0 = 30.4$$

Step 3. Finding P-value

To find the P-value, compare X^2 with critical values from the chi-square distribution with degrees of freedom one fewer than the number of values the color can take. That's $4 - 1 = 3$ degrees of freedom. From Table D, we see that $X^2 = 30.4$ is greater than the greatest entry in the df $= 3$ row, which is the critical value for tail area 0.0005. The P-value is therefore smaller than 0.0005.

Step 4. Conclusion

Since our $P - value < \alpha = 0.01$, we conclude that there is evidence to reject the null hypothesis that all four colors are equally likely to be chosen. Some colors are probably preferable to others. Our P-value is very small.

```
x = c(12, 40, 8, 20);

chisq.test(x);

# chisq.test(x) gives you test statistic;
# degrees of freedom and P-value;

chisq.test(x)$expected;

# gives you expected counts;
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 30.4, df = 3, p-value = 1.137e-06
## [1] 20 20 20 20
```

Exercise

Consider a multinomial experiment involving $n = 150$ trials and $k = 5$ cells. The observed frequencies resulting from the experiment are shown in the accompanying table, and the null hypothesis to be tested is as follows:

$$H_0 : p_1 = 0.1,\ \ p_2 = 0.2,\ \ p_3 = 0.3,\ \ p_4 = 0.2,\ \ p_5 = 0.2$$

Test the hypothesis at the 1% significance level.

Cell	1	2	3	4	5
Frequency	12	32	42	36	28

Therefore, we seek a test statistic that will detect a **lack of fit** of the observed **cell counts** to our hypothesized (null) expected cell counts based on the hypothesized cell probabilities.

These expected values are:

$$E(n_1) = np_1 = (150)\left(\frac{1}{10}\right) = 15$$
$$E(n_2) = np_2 = (150)\left(\frac{2}{10}\right) = 30$$
$$E(n_3) = np_3 = (150)\left(\frac{3}{10}\right) = 45$$
$$E(n_4) = np_4 = (150)\left(\frac{2}{10}\right) = 30$$
$$E(n_5) = np_5 = (150)\left(\frac{2}{10}\right) = 30$$

Table of Expected Counts

Cell	1	2	3	4	5
Frequency	15	30	45	30	30

Step 2. Computing test statistic

$$X^2 = \sum_{all\ \ cells} \frac{\text{(observed - expected)}^2}{\text{expected}} = \sum_{all\ \ cells} \frac{(n_i - E(n_i))^2}{E(n_i)}$$

$$X^2 = \frac{(12-15)^2}{15} + \frac{(32-30)^2}{30} + \frac{(42-45)^2}{45} + \frac{(36-30)^2}{30} + \frac{(28-30)^2}{30}$$

$$X^2 = 9/15 + 4/30 + 9/45 + 36/30 + 4/30 = 2.2667$$

Step 3. Finding P-value

To find the P-value, compare X^2 with critical values from the chi-square distribution with degrees of freedom one fewer than the number of "columns". That's $5 - 1 = 4$ degrees of freedom. From Table D, we see that $X^2 = 2.2667$ is smaller than the smallest entry (5.39) in the df $= 4$ row, which is the critical value for tail area 0.25. The P-value is therefore greater than 0.25.

Step 4. Conclusion

Since our $P - value > 0.25 > \alpha = 0.01$, we conclude that there is NOT enough evidence to reject the null hypothesis H_0. There is not enough evidence to infer that at least one p_i is not equal to its specified value.

The Chi-square Test for Goodness of fit

A categorical variable has k possible outcomes, with probabilities $p_1, p_2, ..., p_k$. That is , p_i is the probability of the ith outcome. We have n independent observations from this categorical variable.
To test the null hypothesis that the probabilities have specified values

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, ..., p_k = p_{k0}$$

use the **chi-square statistic**

$$X^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

(expected counts are equal to np_i).
The P-value is the area to the right of X^2 under the density curve of the chi-square distribution with $k - 1$ degrees of freedom.

Sample Size Assumption

We must have enough data for the methods to work. We usually check the following:
Expected Cell Frequency Condition: We should expect to see at least five individuals in each cell.
The expected cell frequency condition sounds like the condition that np and $n(1-p)$ be at least 10 when we tested proportions, it is similar to that.