

Chapter 15

Chi-Squared Tests

What do we do here?

Until now, our data has been ratio type variables. We could find mean, variance, etc.. We also assumed normal distribution. We also assumed equal variances in our populations. Basically, we were performing parametric tests.

- Now, we test the Hypothesis for Nominal / Categorical variables.
- The data can be classified into categories & there is no particular category order.
- The goal is to test if the observed frequencies are significantly different from the expected frequencies.
- We don't need to worry about the population distribution or the shape of the population. Even outliers are OK.
- Basically, we are performing non-parametric tests here.

Chi-Square vs Normal Distribution

If x is normally distributed with mean μ and variance $\sigma^2 > 0$, then:

$v = ((x-\mu)/\sigma)^2 = z^2$ is distributed as a chi-square random variable with 1 degree of freedom.

Example: What is the area between $z = -1.96$ and $z = 1.96$?

From Normal distribution table, we have $P(-1.96 < Z < 1.96) = 0.95$

And from the chi-square table (1 degree of freedom):

$$P(-1.96 < Z < 1.96) = P(|Z| < 1.96) = P(\chi^2(1) < 1.96^2) = 0.95$$

Excel function: $\text{CHISQ.DIST}(1.96^2, 1, 1) = 0.95$

Two Techniques ... both Chi-Squared ...

The first is a *goodness-of-fit test* applied to data produced by a *multinomial experiment*, a generalization of a binomial experiment and is used to describe one population of data. Example: has the increased gas price impacted the car sales?

Effect of increased gas price		
Market Share	Last year	This year (n = 400)
Small car	35%	168
Medium Car	40%	192
Large Car	15%	40

The second uses data arranged in a *contingency table* to determine whether two classifications of a population of nominal data are *statistically independent*; this test can also be interpreted as a comparison of two or more populations. Example: Does the choice of beer depend on the gender? Does the voting pattern depend on the party affiliation?

	Male	Female	Total
Light	51	39	90
Regular	56	21	77
Dark	25	8	33
Total	132	68	200

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

The Multinomial Experiment...

Unlike a binomial experiment which only has two possible outcomes (e.g. heads or tails), a ***multinomial experiment***:

- Consists of a fixed number, **n**, of trials.
- Each trial can have one of **k** outcomes, called cells.
- Each probability **p_i** remains constant.
- Our usual notion of probabilities holds, namely:

$$p_1 + p_2 + \dots + p_k = 1, \text{ and}$$

- Each trial is ***independent*** of the other trials.

Goodness of Fit Test: Multinomial Probability Distribution

1. State the null and alternative hypotheses.

H0: The population follows a multinomial distribution with specified probabilities for each of the k categories where a_1, a_2, \dots, a_k are the values we want to test.

H1: The population does not follow a multinomial distribution with specified probabilities for each of the k categories

$$H_0: p_1 = a_1, p_2 = a_2, \dots, p_k = a_k$$

H₁: At least one p_i is not equal to its specified value

2. Select a random sample and record the observed frequency, f_i , for each of the k categories.
3. Assuming H_0 is true, compute the expected frequency, e_i , in each category by multiplying the category probability by the sample size.
4. Compute the value of the test statistic.

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

where f_i = observed frequency; e_i = expected frequency. The test statistic has a chi-square distribution with $k - 1$ degrees of freedom provided that the expected frequencies are 5 or more for all categories.

5. Rejection Rule:

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$

Where α is the significance level and there are $k - 1$ degrees of freedom.

Example 15.1

Two companies, A and B, have recently conducted aggressive advertising campaigns to maintain and possibly increase their respective shares of the market for fabric softener. These two companies enjoy a dominant position in the market. Before the advertising campaigns began, the market share of company A was 45%, whereas company B had 40% of the market. Other competitors accounted for the remaining 15%.

Example 15.1

To determine whether these market shares changed after the advertising campaigns, a marketing analyst solicited the preferences of a random sample of 200 customers of fabric softener. Of the 200 customers, 102 indicated a preference for company A's product, 82 preferred company B's fabric softener, and the remaining 16 preferred the products of one of the competitors. Can the analyst infer at the 5% significance level that customer preferences have changed from their levels before the advertising campaigns were launched?

Example 15.1...

We compare market share *before* and *after* an advertising campaign to see if there is a *difference* (i.e. if the advertising was effective in improving market share). We hypothesize values for the parameters equal to the before-market share. That is,

$$H_0: p_1 = .45, p_2 = .40, p_3 = .15$$

The alternative hypothesis is a denial of the null. That is,

H_1 : At least one p_i is not equal to its specified value

Example 15.1...

Test Statistic

If the null hypothesis is true, we would expect the number of customers selecting brand A, brand B, and other to be 200 times the proportions specified under the null hypothesis. That is,

$$e_1 = 200(.45) = 90$$

$$e_2 = 200(.40) = 80$$

$$e_3 = 200(.15) = 30$$

In general, the expected frequency for each cell is given by

$$e_i = np_i$$

This expression is derived from the formula for the expected value of a binomial random variable, introduced in Section 7.4.

Example 15.1...

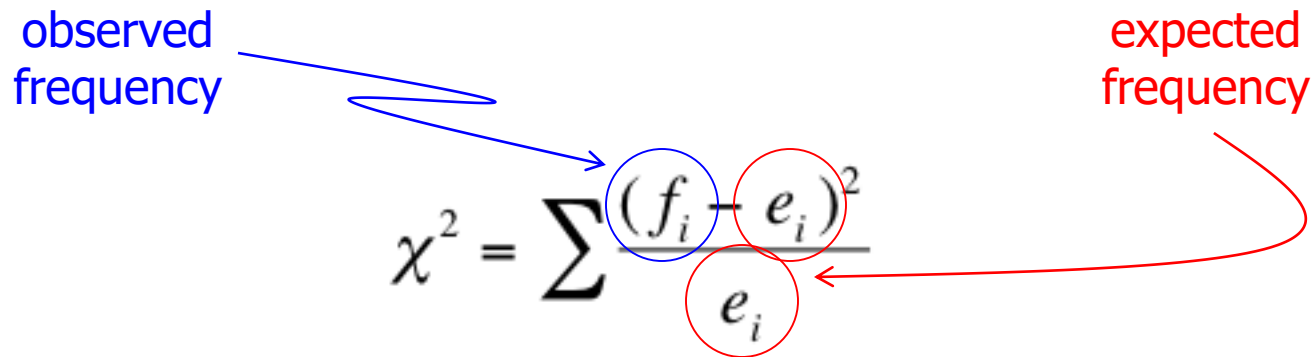
If the expected frequencies and the observed frequencies are quite different, we would conclude that the null hypothesis is false, and we would reject it.

However, if the expected and observed frequencies are similar, we would not reject the null hypothesis.

The test statistic measures the similarity of the expected and observed frequencies.

Chi-squared Goodness-of-Fit Test...

Our Chi-squared goodness of fit test statistic is given by:



The diagram shows the formula $\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$. A blue arrow points from the text "observed frequency" to the f_i term in the numerator, which is circled in blue. A red arrow points from the text "expected frequency" to the e_i term in the denominator, which is circled in red. The e_i term in the numerator is also circled in red.

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

Note: this statistic is *approximately* Chi-squared with $k-1$ degrees of freedom provided the sample size is large. The rejection region is: $\chi^2 > \chi_{\alpha, k-1}^2$

Example 15.1...

COMPUTE

In order to calculate our test statistic, we lay-out the data in a tabular fashion for easier calculation by hand:

Company	Observed Frequency	Expected Frequency	Delta	Summation Component
	f_i	e_i	$(f_i - e_i)$	$(f_i - e_i)^2 / e_i$
A	102	90	12	1.60
B	82	80	2	0.05
Others	16	30	-14	6.53
Total	200	200		8.18

Check that these are equal

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

Example 15.1...

INTERPRET

Our rejection region is:

$$\chi^2 > \chi_{\alpha, k-1}^2 = \chi_{.05, 3-1}^2 = 5.99147$$

Since our test statistic is 8.18 which is greater than our critical value for Chi-squared, we reject H_0 in favor of H_1 , that is, “*There is sufficient evidence to infer that the proportions have changed since the advertising campaigns were implemented*”

Note: Excel’s function CHISQ.TEST(actual frequency, observed frequency) can be used to get the p-value directly from the dataset.

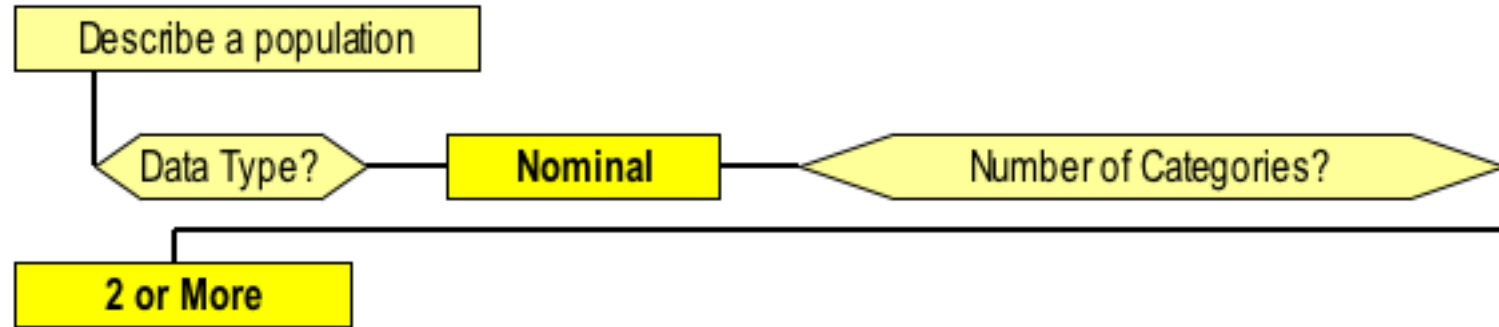
Required Conditions...

In order to use this technique, the sample size must be *large enough* so that the expected value for each cell is 5 or more.
(i.e. $n \times p_i \geq 5$)

If the *expected frequency* is less than five, combine it with other cells to satisfy the condition.

Identifying Factors...

Factors that Identify the Chi-Squared Goodness-of-Fit Test:



Test Statistic:

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

$$e_i = (n)(p_i)$$

Parameters of interest:

$$p_1, p_2, \dots, p_k$$

Required Condition: $e_i \geq 5$

Chi-squared Test of a Contingency Table

The *Chi-squared test of a contingency table* is used to:

- determine whether there is enough evidence to infer that *two nominal variables are related*, and
- to infer that *differences exist* among two or more populations of nominal variables.

In order to use these techniques, we need to classify the data according to two different criteria.

Example 15.2

The MBA program was experiencing problems scheduling their courses. The demand for the program's optional courses and majors was quite variable from one year to the next.

In desperation the dean of the business school turned to a statistics professor for assistance.

The statistics professor believed that the problem may be the variability in the academic background of the students and that the undergraduate degree affects the choice of major.

Example 15.2

As a start he took a random sample of last year's MBA students and recorded the undergraduate degree and the major selected in the graduate program.

The undergraduate degrees were BA, BEng, BBA, and several others.

There are three possible majors for the MBA students, accounting, finance, and marketing. Can the statistician conclude that the undergraduate degree affects the choice of major?

Example 15.2

Xm15-02

The data are stored in two columns. The first column consist of integers 1, 2, 3, and 4 representing the undergraduate degree where

1 = BA

2 = BEng

3 = BBA

4 = other

The second column lists the MBA major where

1= Accounting

2 = Finance

3 = Marketing

Example 15.2

IDENTIFY

The problem objective is to determine whether two variables (undergraduate degree and MBA major) are related. Both variables are nominal. Thus, the technique to use is the chi-squared test of a contingency table. The alternative hypotheses specifies what we test. That is,

H_1 : The two variables are **dependent**

The null hypothesis is a denial of the alternative hypothesis.

H_0 : The two variables are **independent**.

Test Statistic

The test statistic is the same as the one used to test proportions in the goodness-of-fit-test. That is, the test statistic is

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

Note however, that there is a major difference between the two applications. In this one the null does not specify the proportions p_i , from which we compute the expected values e_i , which we need to calculate the χ^2 test statistic. That is, we cannot use $e = np_i$ because we don't know the p_i (they are not specified by the null hypothesis). It is necessary to estimate the p_i from the data.

The degrees of freedom will be $(r - 1).(c - 1)$ where r and c are the respective numbers of rows and columns.

Example 15.2

The first step is to count the number of students in each of the 12 combinations. The result is called a cross-classification table.

	MBA Major			
Undergrad Degree	Accounting	Finance	Marketing	Total
BA	31	13	16	60
BEng	8	16	7	31
BBA	12	10	17	39
Other	10	5	7	22
Total	61	44	47	152

Example 15.2

If the null hypothesis is true (Remember we always start with this assumption.) and the two nominal variables are independent, then, for example

$$P(\text{BA and Accounting}) = [P(\text{BA})] [P(\text{Accounting})]$$

Since we don't know the values of $P(\text{BA})$ or $P(\text{Accounting})$

We need to use the data to estimate the probabilities.

Test Statistic

There are 152 students of which 61 who have chosen accounting as their MBA major. Thus, we estimate the probability of accounting as

$$P(\text{Accounting}) \approx \frac{61}{152} = .401$$

Similarly

$$P(\text{BA}) \approx \frac{60}{152} = .395$$

Example 15.2...

If the null hypothesis is true

$$P(\text{BA and Accounting}) = (60/152)(61/152)$$

Now that we have the probability we can calculate the expected value. That is,

$$\begin{aligned} E(\text{BA and Accounting}) &= 152(60/152)(61/152) \\ &= (60)(61)/152 = 24.08 \end{aligned}$$

We can do the same for the other 11 cells. Basically, we apply the formula below to each cell in the table.

$$\begin{aligned} e_{ij} &= \frac{(\text{Row } i \text{ Total})(\text{Column } j \text{ Total})}{\text{Sample Size}} \\ &= \frac{60 \times 61}{152} = 24.08 \end{aligned}$$

Example 15.2

COMPUTE

We can now compare *observed* with *expected* frequencies...

Undergrad Degree	MBA Major					
	Accounting		Finance		Marketing	
BA	31	24.08	13	17.37	16	18.55
BEng	8	12.44	16	8.97	7	9.59
BBA	12	15.65	10	11.29	17	12.06
Other	10	8.83	5	6.37	7	6.80

and calculate our test statistic:

$$\chi^2 = \frac{(31 - 24.08)^2}{24.08} + \frac{(13 - 17.37)^2}{17.37} + \dots + \frac{(7 - 6.80)^2}{6.80} = 14.70$$

Testing the Hypothesis

The degrees of freedom will be $(4 - 1) \times (3 - 1) = 3 \times 2 = 6$ since we have 4 rows and 3 columns. Since $\alpha = .05$, we can use Chi-Square table to find χ^2_{critical} .

In Excel, we can use *CHISQ.INV(0.95,6)* to get the required number which is 12.59159. Our Test statistics χ^2_{Stat} has been calculated to be 14.70. So, we reject the Null Hypothesis.

We can also find the p-value using Excel. *1-CHISQ.DIST(14.7,6,1) = 0.0227*.

Since p-value is less than α , we reject the Null. There is enough evidence to infer that the MBA major and the undergraduate degree are related.

Note: Excel's function *CHISQ.TEST(actual frequency, observed frequency)* can be used to get the p-value directly from the dataset.

Example 15.2...

INTERPRET

The p-value is .0227. There is enough evidence to infer that the MBA major and the undergraduate degree are related.

We can also interpret the results of this test in two other ways.

1. There is enough evidence to infer that there are differences in MBA major between the four undergraduate categories.
2. There is enough evidence to infer that there are differences in undergraduate degree between the majors.

Required Condition – Rule of Five...

In a contingency table where one or more cells have *expected values* of ***less than 5***, we need to combine rows or columns to satisfy the rule of five.

Note: by doing this, the degrees of freedom must be changed as well.