Outlier

Section Summary | Introduction to Statistics

Categorical Data Analysis: The Chi-Square Distribution

An Introduction to the Chi-Square Distribution

1. The Chi-Square distribution is a distribution with one parameter (degrees of freedom) that is used for testing categorical data. We are looking to see if there is a relationship between two categorical variables.

2. There are 3 types of chi-square tests that you will see in chapter 10.

    a. Tests of Independence

    b. Tests of Homogeneity

    c. Tests of Goodness of Fit

3. Tests of Independence: Test used to determine if there is a relationship between two categorical variables. Does being in one category affect the likelihood of being in another category? Below are some examples of this type of test.

    a. Are non-citizens more likely to be employed than citizens? Such an analysis seeks to determine if citizenship status affects employment status.

    b. Are women more likely to get diabetes than men? This study is trying to determine if biological sex increases the likelihood of developing diabetes.

    c. Distribution of income level by marital status, does your relationship status affect your income level?

4. Test of Goodness of Fit: test used to determine if data fits a hypothesis. Below is an example of such a test.

    a. Testing to see if $\frac{3}{4}$ of a student population is passing and $\frac{1}{4}$ is not. The hypothesis is that $\frac{3}{4}$ of a 5^{th} grade class can pass a test on 4^{th} grade material.

5. Tests of Homogeneity: A test used to determine if a sample is representative of the population. Below is an example where such a test would be used.

a. A California Health Interview survey is composed of 46% men and 54% women. Is this survey representative of the state of California?

6. Before a chi-square test can be used, the following assumptions must be met:

   a. Data are randomly independently sampled.

   b. The expected number of items for each category must be greater than or equal to 5.

   c. The same item can't be measured twice.

   d. Multiple raters rating the same thing is not allowed.

Test for Independence

1. The test for independence uses the chi-squared test statistic

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

   d. O_i is the observed number.

   e. E_i is the expected number.

   f. Σ means to sum up over all the cells in the table.

   g. If the χ^2 is small, then there is a low probability of the two categorical variables being related.

2. The set up for finding the test statistic: Smoking vs Weight

   a. We are testing to see if there is a relationship between smoking and being overweight.

   b. H_0: There is no relationship between being overweight and smoking.

   c. H_a: There is a relationship between being overweight and smoking.

   d. Test for independence will all have similar null and alternative hypotheses. The null will always be "no relationship" (the variables are independent) and the alternative will always be "there is a relationship" (the variables are not independent).

e. The biggest part of the set up is to find the expected number for each cell in the table. Below is the table of observed numbers.

Overweight	Smoker		
	Yes	No	Row Total
Yes	14	124	138
No	10	52	62
Column Total	24	176	200

f. Finding the expected uses the assumption from the null hypothesis. We assume the variables smoker and overweight are independent. This means that the probability of being a smoker and overweight is equal to the probability of being a smoker times the probability of being overweight. We take this probability and multiply it by the sample size to get the expected number. For example, the expected number of smoker and overweight people is found by 200 times the probability of being a smoker times the probability of being overweight:

$$200 * \frac{24}{200} * \frac{138}{200} = 16.56$$

g. Now do this for each of the cells in the first table, this will give a list of expected numbers that will be used to compute the test statistic. Below is the table of expected numbers computed.

Expected	Smoker	
Overweight	Yes	No
Yes	$200 * \frac{24}{200} * \frac{138}{200} = 16.56$	$200 * \frac{176}{200} * \frac{138}{200} = 121.44$
No	$200 * \frac{24}{200} * \frac{62}{200} = 7.44$	$200 * \frac{176}{200} * \frac{62}{200} = 54.56$

h. Now go through and calculate $\frac{(O_i - E_i)^2}{E_i}$ for each cell and add the results together.

$$\chi^2 = \frac{(14 - 16.56)^2}{16.56} + \frac{(124 - 121.44)^2}{121.44} + \frac{(10 - 7.44)^2}{7.44} + \frac{(62 - 54.56)^2}{54.56} \approx 1.4507$$

i. There is an alternate way of getting the expected number. Notice the calculations used to find the expected number for overweight and smoker was $200*\frac{24}{200}*\frac{138}{200}$. The first pair of 200's would divide out leaving $200*\frac{24}{200}*\frac{138}{200}=\frac{24*138}{200}$. This means that the expected number of people that are overweight smokers is $\frac{24*138}{200}=\frac{(total\ number\ of\ smokers)*(total\ number\ of\ overweight\ people)}{Sample\ Size}$. The alternate way of finding the expected number is to take (column total)(row total)/sample size.

3. Finding the Critical Value

   a. The chi-squared test statistic uses the chi-squared distribution. This distribution has one parameter; degrees of freedom.

   b. The degrees of freedom is calculated by $(number\ of\ rows-1)*(number\ of\ columns-1)$. In the previous example, the degrees of freedom are $(2-1)(2-1)=1$.

   c. All chi-squared tests are essentially right-tailed tests. The closer the test statistic is to zero, the lower the probability of dependence between the two variables. The larger the test statistic, the more likely there is a relationship between the two variables. So, the critical number is going to be a number where the level of significance is equal to the area to the right. So, if $\alpha=0.05$, then the critical number will have 5% of the area to the right and 95% of the area to the left. The 95th percentile of a chi-squared distribution with 1 degree of freedom is approximately 3.84.

4. Finding the p-value

   a. The p-value for a chi-squared test will be the probability of observing a value larger than the test statistic in a chi-squared distribution with a given degree of freedom. Having a test statistic of 1.4507 and 1 degree of freedom, the p-value is approximately .2284.

5. Deciding, if the test statistic is larger than the critical value or the p-value is less than the level of significance, then we reject the null hypothesis and conclude that there is evidence to suggest a relationship between the two categorical variables. The previous example has a test statistic of 1.4507

which is less than 3.84. This means we fail to reject the null hypothesis and that there is little evidence that there is a relationship between smoking and obesity.

6. Computer Programs
    a. Most people do not make these computations by hand. Instead, they use programs like SAS or an applet found on the internet.
    b. Using SAS, the program that would give the results for the previous example is
        - Libname sasfile "/home/example/chis18/";
        - Proc freq data = sasfile.chis;
        - Tables ovrwt*smkcur/chisq;
    c. If an applet is used, one would enter the data and click calculate.
7. The most common errors encountered in this section are:
    a. The test statistic is negative when everything is supposed to be positive.
    b. The variables are not categorical.

Goodness of Fit Test

1. The purpose of this test is to determine if the data fit the hypothesized distribution. The test statistic uses the same formula: $\Sigma\frac{(O_i - E_i)^2}{E_i}$ and is always a right-tailed test. The only difference is in how the expected numbers are calculated using the hypothesized distribution and the degrees of freedom is calculated in a different manner.
2. Example: Dogs are more popular than cats.
    a. The null hypothesis is
    $H_0: 70\%\ of\ pet\ owners\ own\ dogs\ and\ 30\%\ of\ pet\ owners\ own\ cats$
    b. The alternative hypothesis is
    $H_a: The\ distribution\ of\ pet\ owners\ does\ not\ fit\ the\ given\ distribution$

. The alternative hypothesis will always be “does not fit the given distribution”.

c. The expected numbers are calculated by using the distribution from the null hypothesis. So in this example, the sample size was 50. There were 20 cat owners and 30 dog owners observed. The expected number of cat owners is $30\%*50 = 15$ and the expected number of dog owners is $70\%*50 = 35$.

d. The test statistic for this is $\frac{(20-15)^2}{15} + \frac{(30-35)^2}{35} \approx 2.38$

e. Now we have to find the critical value which comes from a chi-squared distribution. The degrees of freedom for this type of test is calculated by the number of categories minus 1. This example has two categories which means there are 1 degree of freedom. If the level of significance is 5%, then the critical value is 3.84.

f. Since the test statistic is 2.38 is less than 3.84, we fail to reject the null hypothesis.

3. 2nd Example: The distribution of cat, dog, and fish owners in a particular town.

a. In this example, we have a sample of 100 people. There are 3 categories in which the people are divided: cat owner, dog owner, and fish owner.

b. The survey of 100 people shows 30 cat owners, 60 dog owners, and 10 fish owners.

c. The null hypothesis is
$H_0: The\ distribution\ of\ pet\ owners\ is\ 20\%\ cat\ owners,\ 75\%\ dog\ owners,\ and\ 5\%\ fish\ owners$
.

d. The alternative is
$H_a: The\ distribution\ of\ pet\ owners\ does\ not\ fit\ the\ given\ distribution$
.

e. The expected number of cat owners, dog owners, and fish owners are $100*20\% = 20$, $100*75\% = 75$, and $100*5\% = 5$.

f. The test statistic is $\chi^2 = \frac{(30-20)^2}{20} + \frac{(60-75)^2}{75} + \frac{(10-5)^2}{5} = 13.$

g. There are 3 categories which means the critical value is going to come from a chi-squared distribution with 2 degrees of freedom. If $\alpha = 0.05$, then the critical value is about 5.99.

h. With a test statistic of 13, we reject the null hypothesis.

4. 3rd Example: Distribution of 5th graders passing a 4th grade level test

a. This was an example mentioned in section 10.1. A 5th grade class at a poor public school was examined to see if they could pass a 4th grade level test. 286 students were sampled with 184 passing and 102 did not.

b. The null hypothesis is
H_0: *75% of the students pass and 25% will not.*

c. The alternative hypothesis is
H_a: *The population does not fit the given distribution.*

d. The expected number of passing students and failing students are 214.5 and 71.5.

e. The test statistic is $\frac{(184-214.5)^2}{214.5} + \frac{(102-71.5)^2}{71.5} \approx 17.35$

f. With 1 degree of freedom, this test statistic would indicate that we should reject the null hypothesis.

Test of Homogeneity

1. This test is similar to the test of independence.

g. Degrees of freedom = (number of rows -1)(number of columns -1)

h. The test statistic has the same formula ($\chi^2 = \Sigma\frac{(O_i - E_i)^2}{E_i}$)

i. The test of independence uses a random sample from one population. While, the test of homogeneity uses 2 samples. Each sample comes from a distinct population.

2. Example: Does Florida really have a lot of old people?

   a. H_0: Age distribution of Florida residents is no different than the age distribution of California residents.

   b. H_a: The age distribution of Florida is different from the age distribution of California.

   c. Below is a table displaying the results of the two samples.

	California	Florida
Under 5	2412977	1128864
6 to 17	6566462	3088402
18 to 64	24920910	12715697
65+	5656651	4366362
Total	39557000	21299325

   d. Using SAS, the test statistic was found to be $\chi^2 = 400.9373$, this is a very large test statistic and therefore we reject the null hypothesis.

When Do You Use Each Test?

1. Which chi-squared test do you use? That depends on the null hypothesis.
2. Test of Independence

   a. H_0: Variables are independent

   H_a: Variables are dependent

   b. 1st example: the null hypothesis is H_0: kindergarten success is independent of attending preschool. The alternative hypothesis is H_a: kindergarten success is related to attending preschool.

   c. 2nd example: the null hypothesis is H_0: employment is independent of citizenship and the alternative hypothesis is H_a: employment is related to citizenship.

3. Test of Goodness of Fit

a. H_0: The data fits the given distribution

   H_a: The data does not fit the given distribution

b. Example: the null hypothesis is H_0: the distribution of educational outcomes is the same as 2010 and the alternative hypothesis is H_a: the distribution of educational outcomes is different from what it was is 2010.

4. Test of Homogeneity

   a. H_0: The two populations have the same distribution.

      H_a: the two populations have different distributions.

5. Things to remember

   a. The null hypothesis is a statement of no (0) relationship or difference between the variables

   b. Just because something is statistically significant does not mean that it is important.