

Math107: Multivariate
Categorical Data – The χ^2
Goodness-of-fit Test for
Independence

Dr. Richard Mikula

Spring 2010

At this point we will test for independence in multivariate (many variables) categorical data. Thus, we will consider experiments where the outcomes are categorical (non-numerical data); in particular experiments where we collect more than one type of categorical data. For simplicity, we will restrict ourselves to experiment where two different categorical data are collected.

We will examine whether the types of data we collect are independent or dependent. To do this we will consider a quantity called **Pearson's Chi-Square Statistic**, which is denoted by χ^2 .

We will organize the data collected from our experiment in what is called a **contingency table**. A particular place in the table is called a **cell**. Along one row will represent all the outcomes in one type of the categorical data, and along one column will represent all the outcomes in the other type of categorical data.

At this point, we consider an example to motivate our discussion:

A random sample of 500 persons is questioned regarding political affiliation and attitude toward a tax reform program. From the data collected, we would like to answer:

Do the data indicate that the pattern of opinion is different between the two political groups?

The collected data is organized in the following two way contingency table:

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

We will consider the possibility that the opinion regarding the tax reform is independent of the political affiliation. This assumption we are making is called the **null hypothesis**, and is denoted by H_0 . The other possibility is that that the opinion regarding the tax reform and political affiliation are dependent. This is referred to as the **alternate hypothesis**, and is denoted by H_1 .

We will use the collected data to test whether we should reject the null hypothesis in favor of the alternate hypothesis, or choose not to reject the null hypothesis*

*Keep in mind, we are not claiming to prove that the null hypothesis is true. We are simply testing whether or not we should reject the null hypothesis.

In statistical experiments, we decide on a low probability which will be our threshold for determining whether our sample data provides sufficient evidence to reject or not reject the null hypothesis. We set the **level of significance** α , which is the value for this small probability. We commonly use $\alpha = 0.05$, but this is not set in stone.

If we observe in our sample data, data whose probability of occurring (assuming the null hypothesis is true) is less than or equal to α , we will reject the null hypothesis. Otherwise, we choose not to reject the null hypothesis.

The idea behind testing these types of claims is to compare actual counts to the counts we would expect if the null hypothesis were true (if the variables are independent). If a significant difference between the actual counts and experimental counts exists, we take this as evidence against the null hypothesis*

The method for obtaining the expected counts requires that we compute the number of observations expected within each cell under the assumption that the null hypothesis is true.

*Since a significant difference between actual counts and experimental counts should have a small probability of occurring.

Recall, if two events E and F are independent, then

$$P(E \cap F) = P(E)P(F).$$

We can use this multiplication rule for independent events to obtain the expected proportion of observations within each cell, under the assumption of independence. We then multiply this by N , the sample size, to obtain the expected count or frequency within each cell.

Recall our original contingency table:

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

To get the expected count in a cell, we compute

$$\frac{(\text{row total})(\text{column total})}{N}.$$

Thus we get:

	Favor	Indifferent	opposed	total
Democrat	$\frac{285 \cdot 202}{500}$	$\frac{285 \cdot 150}{500}$	$\frac{285 \cdot 148}{500}$	285
Republican	$\frac{215 \cdot 202}{500}$	$\frac{215 \cdot 150}{500}$	$\frac{215 \cdot 148}{500}$	215
Total	202	150	148	500

That is:

	Favor	Indifferent	opposed	total
Democrat	115.14	85.50	84.36	285
Republican	86.86	64.50	63.64	215
Total	202	150	148	500

A useful measure for the overall discrepancy between the observed and expected frequencies is given by the χ^2 statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where we sum over all the cells in our table; O is an observed value from a particular cell in our table (that is, the original table), and E is the expected value from the same cell in the second table. In our case:

	Favor	Indifferent	opposed	
Dem.	$\frac{(138-115.14)^2}{115.14}$	$\frac{(83-85.50)^2}{85.50}$	$\frac{(64-84.36)^2}{84.36}$	
Rep.	$\frac{(64-86.86)^2}{86.86}$	$\frac{(67-64.50)^2}{64.50}$	$\frac{(84-63.64)^2}{63.64}$	
				χ^2

	Favor	Indifferent	opposed	
Democrat	4.539	0.073	4.914	
Republican	6.016	0.097	6.514	
				22.153

The statistic χ^2 has what is called a χ^2 probability distribution * with **degrees of freedom** *d.f.* In this case

$$d.f. = (r - 1)(c - 1)$$

where there are r rows and c columns in our contingency table.

The number of degrees of freedom associated with the χ^2 test is equal to the number of cells in the contingency table that may be filled in freely when we are given the row totals, the column totals and the grand total.

*The χ^2 distribution is a special case of what is referred to as a Gamma distribution.

We note here that we need to make the assumption that N is large enough so that

1. All expected counts (frequencies) are greater than or equal to 1.
2. No more than 20 percent of the expected frequencies are less than 5.

We will use a table to find a value called χ^2_α . The probability

$$P(\chi^2 \geq \chi^2_\alpha) = \alpha$$

is the probability that χ^2 is greater than or equal to the value χ^2_α is α .

We will pick α to be some small positive number, say $\alpha = 0.05$. Then we will look up the value χ^2_α in the table. It is the number in the column with value α and in the row with the correct number of degrees of freedom. We will reject the null hypothesis if $\chi^2 \geq \chi^2_\alpha$ and choose not to reject the null hypothesis if $\chi^2 < \chi^2_\alpha$.

In our example

$$d.f. = (2 - 1)(3 - 1) = 2.$$

Using $\alpha = 0.05$ and referring to the table below:

The χ^2 Distribution:

Values of χ^2_α for given values of α and degrees of freedom $d.f.$

d.f.	$\alpha = 0.99$	0.95	0.1	0.05	0.025	0.01
1	0.000	0.004	2.706	3.841	5.024	6.635
2	0.020	0.103	4.605	5.991	7.378	9.210
3	0.115	0.352	6.251	7.815	9.348	11.345
4	0.297	0.711	7.779	9.488	11.143	13.277
5	0.554	1.145	9.236	11.070	12.833	15.086
6	0.872	1.237	10.645	12.592	14.449	16.812
7	1.239	1.690	12.017	14.067	16.013	18.475
8	1.646	2.180	13.362	15.507	17.535	20.090
9	2.088	2.700	14.684	16.919	19.023	21.666
10	2.558	3.247	15.987	18.307	20.483	23.209
12	3.571	5.226	18.549	21.026	23.337	26.217
14	4.660	6.571	21.064	23.685	26.119	29.141
15	5.229	7.261	22.307	24.996	27.488	30.578
16	5.812	7.962	23.542	26.296	28.845	32.000
18	7.015	9.390	25.989	28.869	31.526	34.805
20	8.260	10.851	28.412	31.410	34.170	37.566

We see that

$$\chi_{\alpha}^2 = 5.991.$$

Since

$$\chi^2 = 22.153 \geq 5.991$$

we choose to reject the null hypothesis in favor of the alternate hypothesis. That is, the assumption that the two variables are independent appears to be improbable from the collected data. Thus, the political affiliation and the opinion on tax reform appear to be dependent.

A Short Discussion on Hypothesis Testing and the Role of Probability in Hypothesis Testing:

Note that probability plays a key role in hypothesis testing in statistics. If a sample is taken from a population and it is observed that an event E of probability less than α^* has occurred, then we are lead to believe that something is wrong out our working hypothesis H_0 .

If this occurs, there is a small probability that E may have occurred under the working hypothesis H_0 , and thus there could be nothing wrong with our working hypothesis H_0 . However, it is very unlikely that we observe this event E under the working hypothesis H_0 .

*Some small positive, preassigned number

Rejection of the working hypothesis H_0 based on the observation of an unlikely event E tends to all but rule out the working hypothesis H_0 . On the other hand, failure to reject the working hypothesis H_0 does not rule out other possibilities. As a result, the firm conclusion is established when we reject the working hypothesis H_0 .

In hypothesis testing, the **alternate hypothesis** H_1 usually represents the question to be answered, the theory to be tested, and thus its specification is crucial.

The **null hypothesis** H_0 nullifies or opposes the alternate hypothesis and is usually the logical negation of the alternate hypothesis or at least part of its logical complement.

The conclusions that are drawn in hypothesis testing are

- Reject H_0 in favor of H_1 because of evidence in the collected data.
- Fail to reject H_0 because of insufficient evidence in the data.

Thus we fail to rule out the possibility of the null hypothesis if we fail to reject it in our hypothesis testing.

The best illustration of this is trials in the American legal system.

In the American legal system, the indictment comes because of suspicion of guilt. Here H_0 stands in opposition to H_1 , and is maintained unless H_1 is supported by evidence "beyond a reasonable doubt."

However, failure to reject H_0 does not imply innocence, but merely that the evidence was insufficient to convict.

So here the court system does not necessarily accept H_0 , but fails to reject H_0 , the assumption of innocence.

rejection of H_0 when it is true is referred to as a **type I error**, and will occur with a probability of α .

Non-rejection of H_0 when it is false is called a **type II error**. The table below summarizes this:

	H_0 is True	H_0 is False
Do Not Reject H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

At this point we summarize the **Chi-Square Goodness-of-fit Test for Independence of Categorical Data**:

If a claim is made regarding the independence of two variables in a contingency table, we can use the steps that follow to test the claim.

1. A claim is made regarding independence of the data
 - H_0 : The row variable and column variable are independent.
 - H_1 : The row variable and column variable are dependent.
2. Choose a level of significance α .

- 3.(a) Calculate the expected frequencies (counts) for each cell in the contingency table as row total times column total divided by sample size N .
- (b) Verify that the requirements for the goodness-of-fit test are satisfied:
- All expected frequencies are ≥ 1 .
 - No more than 20 percent of the expected frequencies are < 5 .
- (c) Compute the test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

with summation taken over all the cells in the table. Here the O 's are the observed frequencies and the E 's are the expected frequencies under the assumption of independence.

4. Determine the critical value χ_{α}^2 with $d.f =$ (number of rows - 1)(number of columns - 1) degrees of freedom (this you get from the table for the Chi-square probability distribution).

5. If the computed value of χ^2 is greater than or equal to χ_{α}^2 we reject the null hypothesis H_0 in favor of the alternate hypothesis H_1 . Otherwise, we choose not to reject the null hypothesis H_0 .

We consider another example:

A survey is undertaken to determine the incidence of alcoholism in different professional groups. Random samples of the clergy, educators, executives and merchants are interviewed and the observed frequency counts are given below:

	Alcoholic	Non-alcoholic	Total
Clergy	32	268	300
Educators	51	199	250
Executives	67	233	300
Merchants	83	267	350
Total	233	967	1200

As in the last example, the null hypothesis H_0 shall be that the two variables are independent. That is, the proportions of alcoholics in each population of professional groups are the same.

We compute the value of the expected frequencies E assuming independence. Recall that E is the product of the row and column totals divided by the sample size N .

The following table lists these values:

	Alcoholic	Non-alcoholic	Total
Clergy	58.25	241.75	300
Educators	48.54	201.46	250
Executives	58.25	241.75	300
Merchants	67.96	282.04	350
Total	233	967	1200

Next, we compute the values of

$$\frac{(O - E)^2}{E}$$

for each cell in the table:

	Alcoholic	Non-alcoholic	Total
Clergy	11.83	2.85	
Educators	0.12	0.03	
Executives	1.31	0.32	
Merchants	3.33	0.80	
			$\chi^2 = 20.59$

Here the degrees of freedom is given by

$$d.f. = (4 - 1)(2 - 1) = 3.$$

Thus, using $\alpha = 0.05$ we look up the value of χ^2_{α} in the table above to get

$$\chi^2_{0.05} = 7.81.$$

Since our computed value of $\chi^2 = 20.59$ is larger than this, we choose to reject the null hypothesis.

Thus, it appears that the variables are dependent.

Homework Exercise:

A study of a new flu vaccine is conducted. A random sample of 818 individuals is selected and each is classified according to inoculation status and the state of health. The results of the study are given below. State and test the hypothesis that one's state of health is independent of one's inoculation status at the $\alpha = 0.05$ level.

	Had Flu	Didn't Have Flu	Total
Inoculated	276	3	
Not Inoculated	473	66	
Total			818

A sociologist is interested in the relationship between religious affiliation and attitude toward government-funded social welfare programs. Religious affiliation is split into three levels: actively affiliated with a religious organization, inactive but claiming an affiliation, and claiming no religious affiliation. Four attitudes toward social welfare are identified: cut out welfare completely, maintain it at a reduced level, maintain it at the current level, increase it. One thousand persons are randomly selected and interviewed. The results of the survey are given in the table below. State and test the null hypothesis that an individual's attitude toward social welfare programs is independent of his or her religious affiliation at the $\alpha = 0.05$ level.

	Stop	Dec.	Same	Inc.	Total
Act.	10	150	180	60	400
Inact.	36	141	158	15	350
No Aff.	28	98	115	9	250
Total	74	389	453	84	1000

We will assume the null hypothesis H_0 : That religious affiliation and attitude toward welfare are independent.

First, we compute the expected frequencies E 's; this is given the the table below:

	Stop	Dec.	Same	Inc.	Total
Act.	29.6	155.6	181.2	33.6	400
Inact.	25.9	136.15	158.55	29.4	350
None	18.5	97.25	113.25	21	250
Total	74	389	453	84	1000

Next, we compute χ^2 . We do this by computing $\frac{(O-E)^2}{E}$ for each cell in our contingency table:

	Stop	Dec.	Same	Inc.	
Act.	12.98	0.20	0.01	20.74	
Inact.	3.94	0.17	0.00	7.05	
No Aff.	4.88	0.01	0.02	6.86	
					$\chi^2 = 56.86$

Lastly, we compute the number of degrees of freedom, and then look up the value of χ^2_α :

$$d.f = (3 - 1)(4 - 1) = 6$$

Thus

$$\chi^2_\alpha = 12.592.$$

Because

$$\chi^2 \geq \chi^2_\alpha$$

we reject the null hypothesis in favor of the alternate hypothesis. Hence, religious affiliation and attitude toward welfare appear to be dependent.

Solution to Home Work Exercise:

Recall the observed frequency table:

	Had Flu	Didn't Have Flu	Total
Inoculated	276	3	279
Not Inoculated	473	66	539
Total	749	69	818

Step 1: We compute the expected frequency table:

	Had Flu	Didn't Have Flu	Total
Inoculated	255.47	23.53	279
Not Inoculated	493.53	45.47	539
Total	749	69	818

Step 2: Compute the table with values $\frac{(O-E)^2}{E}$ to get χ^2 :

	Had Flu	Didn't Have Flu
Inoculated	1.65	17.91
Not Inoculated	0.85	9.27
	$\chi^2 = 29.68$	

Here

$$d.f. = (2 - 1)(2 - 1) = 1.$$

Thus,

$$\chi_{0.05}^2 = 3.841.$$

Since

$$29.68 \geq 3.841$$

we reject the null hypothesis. That is, we reject the assumption of independence.