
27. One Way Tables

So far when we have considered proportions we have considered only two possible outcomes for our experiment: “success” and “failure.” Every member of our population must fall into one of these two categories. Often the population will be far more complex. For example consider the following table which shows the ethnic breakdown of the OU student population (in Fall 1992).

White	Black	Hispanic	Asian	Indian	Other
.758	.061	.023	.035	.047	.076

This is a *one way* table since we have only taken a cross-section of the student population in one way (ethnicity). We could have taken a cross-section in two ways (ethnicity and gender, for example) and produced a

two-way table:

	White	Black	Hisp.	Asian	Indian	Other
m	.402	.030	.014	.019	.022	.054
f	.356	.031	.009	.016	.025	.022
Tot	.758	.061	.023	.035	.047	.076

Initially we will deal only with one-way tables; the analysis of two way tables is similar. Consider the following example.

27.1. Example.

A random sample of 483 students is taken from the OU student body. It is found that this sample has the following ethnic breakdown:

	White	Black	Hispanic	Asian	Indian	Other
#	386	24	27	12	20	14
%	79.9	5.0	5.6	2.5	4.1	2.9

Does this sample differ significantly ($\alpha = 0.05$) from the overall student population with respect to ethnicity?

Notice that the sample has more whites and Hispanics than we might have expected and fewer of the other classifications. The question is whether this is due to some systematic error in the way the sample was taken or can be attributed to the natural random errors implicit in sampling. The Chi-squared test (χ^2 test) is a way of answering this question. In this case we are actually testing a hypothesis (note the word *significant*):

H_0 : sample is not biased with respect to ethnicity

against

H_A : sample is biased

We will step through the solution to this problem by way of introduction to the chi-squared test.

Solution. From our data we have an *outcomes* table:

	White	Black	Hispanic	Asian	Indian	Other
#	386	24	27	12	20	14

Based on what we know about the population, we can construct an *expectations* table:

	White	Black	Hispanic	Asian	Indian	Other
#						
	.758	.061	.023	.035	.047	.076

The first row in the expectations table is initially blank; the second row is the proportion of the of population which falls in each category. We know what proportion of the sample we would *expect* to fall in each category from the census data on the population. To fill in the first row (# row) in the expectations table, multiply the expected proportion times the total number in the sample. Thus the expected number of whites in the sample should be:

$$\text{expected whites} = 0.758 \times 483 = 366.11$$

and the expected number of Blacks in the sample should be:

$$\text{expected whites} = 0.061 \times 483 = 29.46$$

Continuing in this fashion, we can fill in the # row in the expectations table:

	White	Black	Hisp.	Asian	Indian	Other
#	366.11	29.46	11.11	16.91	22.70	36.71
	.758	.061	.023	.035	.047	.076

The chi-squared test now compares the actual outcomes with these expected outcomes by computing the following test statistic:

$$\chi^2 = \sum \frac{(\text{Observation} - \text{Expectation})^2}{\text{Expectation}}$$

Since we have six cells (ethnicities), the sum will have six terms. For

this problem the sum is:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}} \\ &= \frac{(386 - 366.11)^2}{366.11} + \frac{(24 - 29.46)^2}{29.46} + \dots \\ &\dots + \frac{(27 - 11.11)^2}{11.11} + \frac{(12 - 16.91)^2}{16.19} + \dots \\ &\dots + \frac{(20 - 22.70)^2}{22.70} + \frac{(14 - 36.71)^2}{36.71} \\ &= 1.08 + 1.01 + 22.73 + 1.43 + 0.32 + 14.05 \\ &= 40.62\end{aligned}$$

As usual, though, we have a spreadsheet that does all of the above for us. In this case, the spreadsheet is FORMULAS.XLSX and the tab is Chi-square, 1-way. It's only necessary to enter the basic data and the

census data:

	White	Black	Hispanic	Asian	Indian	Other
sample	386	24	27	12	20	14
census	.758	.061	.023	.035	.047	.076


The first line represents the actual counts of each ethnicity in our sample, while the second line represents the proportion of each ethnicity in the actual population, based on a census. With this information, the spreadsheet computes everything for us, and provides a p -value to test the null hypothesis against the alternative hypothesis.

Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 1-way.


The screenshot shows the following data in the spreadsheet:

Observations	Class A	Class B	Class C	Class D	Class D	Class E	Totals
Sample	386	24	27	12	20	14	483
Census	0.758	0.061	0.023	0.035	0.047	0.076	1
Observed	0.799172	0.049689	0.055901	0.024845	0.041408	0.028986	1
Expected	366.114	29.463	11.109	16.905	22.701	36.708	483

Test Statistic	26.56911
Degrees of Freedom	5
p-value	0.0069%



From the above, we see that the p -value is 0.0069%, so we have highly significant data to support the alternative hypothesis, that the sample is biased.



28. Two Way Contingency Tables

Sometimes the population can be partitioned on two directions. Consider the following table:

	Cancer	Heart Disease	Other
Smoker	135	310	205
Nonsmoker	55	155	140

The table lists the causes of death from 1000 randomly selected death certificates. Note that there are actually six categories instead of the two we have considered up to now. In addition, the categories themselves are in two groups: smokers versus nonsmokers and three different causes of death (so this is a 2×3 table). Each group (smok-

ers/nonsmokers and causes of death) are collectively exhaustive: each member of the population was either a smoker or not and (since we are sampling death certificates) each had a cause of death. We might be interested in testing the null hypothesis:

H_0 : Cause of death is unrelated to whether or not someone smokes.
against

H_A : Cause of death is related to smoking.

The Chi squared test (χ^2 test) is a vehicle for doing this. The basic setup is a contingency table like the one above with one or more rows and columns. The rows partition the population with respect to one variable (for example smoking) and the columns partition the population with respect to another variable (for example cause of death). The Chi squared statistic is a way of determining if the row effects and column effects are independent of one another. The more general form of the null and alternative hypotheses are:

H_0 : Row effects and column effects are independent.
against

H_A : Row and column effects are dependent on each other. In order to understand the difference between “dependent” and “independent” events, consider some examples.

28.1. Example.

Suppose that a clinic treats a total of 150 patients this month. Some of the patients are adults (over 18) and some are children. Some of the conditions involve trauma (bruises, broken bones or other injuries) and some involve other conditions. A contingency table of outcomes might look like:

	<i>trauma</i>	<i>non-trauma</i>
<i>adult</i>	20	30
<i>non-adult</i>	40	60

Totaling the rows and columns gives a slightly better view of the data:

	trauma	non-trauma	total
adult	20	30	50
non-adult	40	60	100
	60	90	150

Forty percent of all patients ($\frac{60}{150} \times 100\%$) were seen for trauma; forty percent of all adults ($\frac{20}{50} \times 100\%$) were seen for trauma; forty percent of all children ($\frac{40}{100} \times 100\%$) were seen from trauma. If we randomly select a patient file and discover that the patient was seen for trauma, we can't deduce from this information that it was more or less likely that the patient was an adult. Similarly, if we randomly select an adult patient, we can't deduce whether or not the patient was seen for a trauma-related condition.

Now let's suppose that we select a different group of 150 patients with a slightly different distribution.

28.2. Example.

Suppose that the clinic treated a total of 150 patients two months ago. Some of the patients are adults (over 18) and some are children. Some of the conditions involve trauma (bruises, broken bones or other injuries) and some involve other conditions. Suppose that the contingency table of outcomes looks like:

	<i>trauma</i>	<i>non-trauma</i>	<i>total</i>
<i>adult</i>	10	40	50
<i>non-adult</i>	50	50	100
	60	90	150

Notice we have the same total number of patients, total number of adults, total number of non-adults, total number of trauma patients and total number of non-trauma patients. However, the distribution inside the cells of the contingency table is now different. Fifty percent of the

children are seen for trauma whereas only 20 percent of the adults are seen for trauma. In this second sample, *age* and *type of condition* are dependent.

Of course, data are rarely as decisive as in the above examples. The chi-squared test is a way to decide if an appearance of non-independence in the data is statistically significant. To see how the test works, let's reconsider the smoking data.

28.3. Example.

A researcher randomly selects 1000 death certificates and, after interviewing the attending physician, records the following information about the deceased:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>
<i>Smoker</i>	135	310	205
<i>Nonsmoker</i>	55	155	140

At a significance of level of 5%, do these data show that smoking and cause of death of dependent?

Note: the data can't show that smoking *causes* death since everyone in the sample is already dead. What the data can show is that dying of cancer or heart disease is related to whether or not the deceased smoked.

Following the same process that we used for the emergency room above, we could produce an *observations* table.

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	135	310	205	650
<i>Nonsmoker</i>	55	155	140	350
<i>totals</i>	190	465	345	1000

Next build an *expectations* table:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>				<i>650</i>
<i>Nonsmoker</i>				<i>350</i>
<i>totals</i>	<i>190</i>	<i>465</i>	<i>345</i>	<i>1000</i>
<i>proportions</i>	<i>.19</i>	<i>.465</i>	<i>.345</i>	

The *cells* of the expectations table are initially blank; you have to fill them in with computations. In addition, a new row (proportions) has been added. In the first column, this is the proportion of all deaths attributed to cancer ($\frac{190}{1000}$); in the second column, the proportion of all deaths attributed to heart disease ($\frac{465}{1000}$); in the third column, the proportion of all deaths attributed to other causes ($\frac{345}{1000}$). In each case, the proportion row is filled in with the formula

$$\frac{\text{column total}}{\text{overall total}}$$

You use the proportion row to fill in the cells. The number which goes in the cells is what *you would expect the result to be if the row and*

column effects were independent. Since 19% of all deaths were attributable to cancer, if “cancer” and “smoking” were unrelated, we would expect that 19% of all smokers’ deaths would be caused by cancer. Thus, the upper right cell in the expectations table is

$$19\% \text{ of } 650 = 123.5$$

Similarly, the upper middle cell in the expectations table is

$$46.5\% \text{ of } 650 = 302.25$$

and the upper left cell is

$$34.5\% \text{ of } 650 = 224.25$$

More generally, the cells in the expectations table are filled in as follows:

Expectations Table

	Cancer	♡ Disease	Other	totals
Smkr	$.19 \times 650$	$.465 \times 650$	$.345 \times 650$	650
NonSmkr	$.19 \times 350$	$.465 \times 350$	$.345 \times 350$	350
totals	190	465	345	1000
prop's	.19	.465	.345	

which results in an expectations table which looks like:

	<i>Cancer</i>	♡ <i>Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	<i>123.5</i>	<i>302.25</i>	<i>224.25</i>	<i>650</i>
<i>Nonsmoker</i>	<i>66.5</i>	<i>162.75</i>	<i>120.75</i>	<i>350</i>
<i>totals</i>	<i>190</i>	<i>465</i>	<i>345</i>	<i>1000</i>
<i>proportions</i>	<i>.19</i>	<i>.465</i>	<i>.345</i>	

Notice that the rows and columns still add up to the marginal totals. This table gives what we would *expect* to observe if the row and column effects were independent. Notice that this differs from our actual observations:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	135	310	205	650
<i>Nonsmoker</i>	55	155	140	350
<i>totals</i>	190	465	345	1000

Next we need a rule to decide if the differences between the observations and the expectations are statistically significant. The next step in the process is to compute the test statistic:

$$\chi^2 = \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}}$$

the sum being taken over each data cell in the contingency tables. Fortunately, there is a spreadsheet to do all of this for us. For completeness, though, let's step through the computations that are hidden inside the spreadsheet. In our example there are six terms to sum:

$$\begin{aligned}
\chi^2 &= \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}} \\
&= \frac{(135 - 123.5)^2}{123.5} + \frac{(310 - 302.25)^2}{302.25} + \dots \\
&\dots + \frac{(205 - 224.25)^2}{224.25} + \frac{(55 - 66.5)^2}{66.5} + \dots \\
&\dots + \frac{(155 - 162.75)^2}{162.75} + \frac{(140 - 120.75)^2}{120.75} \\
&= 1.07 + 0.199 + 1.652 + 1.989 + 0.36 + 3.069 \\
&= 8.349
\end{aligned}$$

As usual, we must now compare the value of the test statistic against a cutoff which we find in a table. The test statistic in this case is not normal however: it is a “chi-squared” statistic which is tabulated on page 666 (Table A-4) in your text. In order to use the table, you need to know the

degrees of freedom for the test statistic. This is computed by

$$\text{degrees of freedom} = (\# \text{ of rows} - 1) \times (\# \text{ of cols} - 1)$$

Thus in our problem the degrees of freedom are

$$(2 - 1) \times (3 - 1) = 2$$

The degrees of freedom tell you the *row* in the table in which you need to look. The entries across the top correspond (for this type of problem) to the significance level. Thus the cutoff for this problem is 5.991. This cut-off corresponds to the pre-set significance level of 5%, but our test statistic is larger, so the associated *p*-value would be less. As a consequence, we'd reject the null hypothesis that the variables are independent and conclude that they are dependent. This is, of course, much easier with the spreadsheet.

Solution.

Step 1. Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 2-way.

Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	135	310	205			650
Group 2	55	155	140			350
Group 3						0
Group 4						0
Group 5						0
Group 6						0
Totals	190	465	345	0	0	1000
Proportion	0.19	0.465	0.345	0	0	1
0.10%						
Expected	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	123.5	302.25	224.25	0	0	650
Group 2	66.5	162.75	120.75	0	0	350
Group 3	0	0	0	0	0	0
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	190	465	345	0	0	1000
Test Statistic	8.348631					
Degrees of Freedom	2					
p-value	1.5386%					

Step 2. The p -value is 1.5386%, so we have *significant* (but not

highly significant) evidence that heart disease and smoking are *dependent* variables.

Notes:

1. The chi-squared statistic has other uses than the one described in this section. *Not every application of the chi-squared involves two-way contingency tables.*
2. In this unit our tests have not involved parameters (means, standard deviations) but instead have involved categories. The hypotheses related to issues of dependence or independence rather than magnitudes of parameters. For this reason, these kinds of tests are called *non-parametric*.

28.4. Example.

In a study of heart disease among males, the 356 subjects were classified according to socioeconomic status and smoking habits. The study recognized three levels of socioeconomic status (high, middle and low) and three smoking categories (current smoker, never smoked, former smoker). The data are summarized in the following contingency table:

	<i>high</i>	<i>middle</i>	<i>low</i>
<i>current</i>	51	22	43
<i>former</i>	92	21	28
<i>never</i>	68	9	22

At the 5% significance level do the data show that smoking habits and socioeconomic status are dependent or independent?

Solution.

Step 1. Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 2-way.

Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	51	22	43			116
Group 2	92	21	28			141
Group 3	68	9	22			99
Group 4						0
Group 5						0
Group 6						0
Totals	211	52	93	0	0	356
Proportion	0.592697	0.146067	0.261236	0	0	1
Expected						
Group 1	68.75281	16.94382	30.30337	0	0	116
Group 2	83.57022	20.59551	36.83427	0	0	141
Group 3	58.67697	14.46067	25.86236	0	0	99
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	211	52	93	0	0	356
Test Statistic		18.50974				
Degrees of Freedom		4				
p-value		0.0981%				

Step 2. The p -value is 0.0981%, so we have *highly significant ev-*

idence that smoking habits and socioeconomic status are *dependent* variables.

Two-way tables can also be used to do hypothesis tests for proportions:

$$H_0 : p_E = p_C \quad \text{against} \quad H_A : \begin{cases} p_E > p_C & \text{or} \\ p_E < p_C & \text{or} \\ p_E \neq p_C \end{cases}$$

In this case, we'd have two rows and two columns:

	<i>Experimental Group</i>	<i>Control Group</i>
<i>Number of Successes</i>		
<i>Number of Failures</i>		

and thus the test has one degree of freedom. This approach is slightly different from the one we used earlier, where we tested to see if two **parameters** were different, while the Chi-squared tests for **independence**.

28.5. Example.

A large Midwestern hospital tracked the 12-month survival rates for persons who were treated for cardiac arrest in the hospital ER. The hospital gathered the following data.

	<i>Non-Smokers</i>	<i>Smokers</i>
<i>Survived at least 12 month</i>	84	45
<i>Deceased within 12 months</i>	3123	2886

Is there a statistically significant difference in the 12-month survival rates for smokers and non-smokers?

Solution.

Now enter the list into the spreadsheet FORMULAS.XLSX, using the tab at the bottom labeled Chi-square, 2-way. The reported p-value of 0.31% means we reject the null hypothesis that the proportions—i.e., survival rates—are the same for the two groups.

Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	84	45				129
Group 2	3123	2886				6009
Group 3						
Group 4						
Group 5						
Group 6						
Totals	3207	2931	0	0	0	6138
Proportion	0.522483	0.477517	0	0	0	
	0.10%					

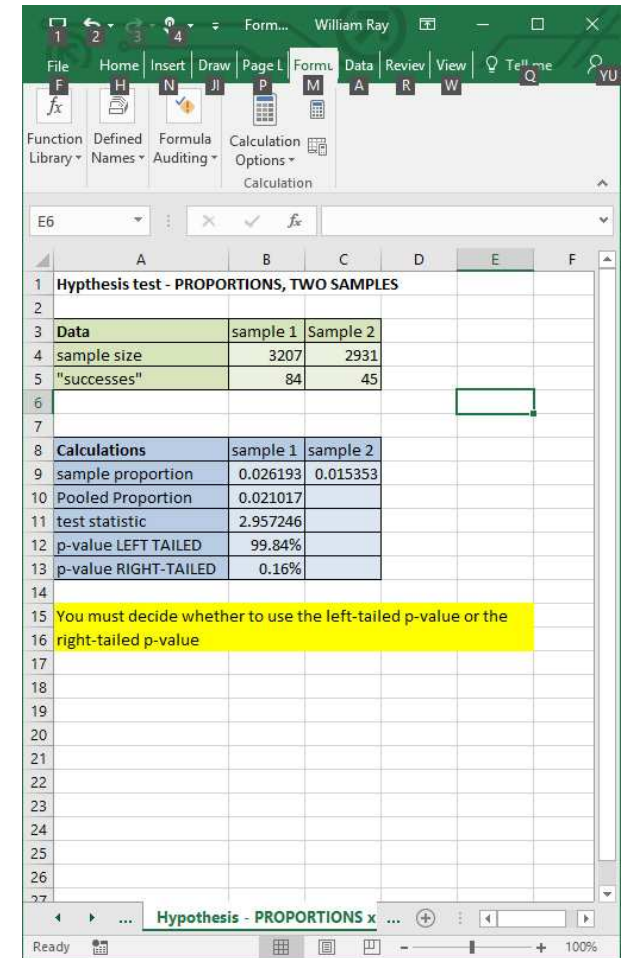
Expected	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	67.40029	61.59971	0	0	0	129
Group 2	3139.6	2869.4	0	0	0	6009
Group 3	0	0	0	0	0	0
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	3207	2931	0	0	0	6138


Test Statistic	8.745302
Degrees of Freedom	1
p-value	0.3104%

You can do the same test using the tab labeled Hypothesis - PROPORTIONS x2 and obtain a similar but not quite identical solution. Remember, the tests are not quite the same, one being parametric and the other non-parametric. To do the earlier parametric test, you need to know the sample sizes rather than the number in success/fail category:

	<i>Non-smokers</i>	<i>Smokers</i>
<i>Sample size</i>	<i>3207</i>	<i>2931</i>
<i>Survival Rate</i>	<i>84</i>	<i>45</i>

From the spreadsheet, this gives a p-value of 0.16%.





Both the Chi-squared and the normal test *approximate* nominal (attribute) data with a continuous (numerical) distribution (the normal distribution). The relative accuracy of this approximation depends on several factors, including the expected and observed cell frequencies and how "far" the true value of the population proportion (assuming the null hypothesis) is from 0.5. There is a more exact test due to Fisher that is not covered in this class, but for most applications either the normal or Chi-squared approach provides satisfactory results. par