Contingency Tables / Cross Tabulation Tables (*Draft 2010-04*)

This is part of a series of quantitative tools that will allow you to do some basic analyses in order to back up conclusions stated in your papers, proposals, or thesis [rob rucker. 2010-04, rob rucker].

Contingency Tables/Cross Tabulation Tables

Contingency Tables, also known as Cross Tabulation Tables, are often used to determine if two or more variables are independent, otherwise, you may be able to speak of 'dependence' or 'bias', a condition always interesting to discover. For example, are job promotions distributed proportionally among males and females, are budgets distributed proportionally among departments, does annual income influence drug convictions?

The formal way to say this is that the *null hypothesis* states: the variables are *independent*, there is no bias. This is the conservative stance, the status quo, the position that will take a lot to unseat. A lot of evidence will be required to reject the null hypothesis.

The *alternative hypothesis* is: The variables are *dependent*, there is bias. This is the role of the challenger who will need a great deal of evidence to change the status quo.

In our case we will calculate a number, called a test statistic, that will help to decide whether or not the variables are independent. The test statistic is called *Chi-square* and I will lead you through its calculation within the first 'voting' example.

Cut to the Chase: Understanding Contingency Tables

Let me start as simply as possible: A *two-way* contingency table is a two dimensional array of data where the row values are characterized by one classification and the columns are characterized by another. That is, there is one class of populations whose subclasses comprise the labels of each row and another population whose subclasses comprise the column labels. Below, the categories are Candidates and Gender. The two subclasses of Candidate are candidate A and candidate B and the two subclasses of Gender are Men and Women.

Example I: *Do the Women Like Candidate A better than B?*

Let's consider a voting situation where there are two candidates A and B and the question is, do the votes for candidate 'A' or 'B' *depend* on the gender of the voter or is it *independent* of the voters' gender? That is, do women favor one candidate *disproportionately*, over the other, or is the choice independent of gender?

O.k., here's how to tell: Suppose, after the election is over, we find that 75% of *all* voters favored A while 25% voted for B. Further we know there were 60 males and 40 females. Just knowing this much allows us to say that 75% of males should have cast ballots for A and 75% of females should have also, if there was no gender *bias*. If there is a large discrepancy from this expectation, then we have found evidence that there was gender bias, otherwise we have to stick with the null hypothesis of no bias.

TABLE 1. 2 X 2 Contingency Table (marginal values shown only)

Gender / Candidate	Male	Female	Overall results counted to yield Marginal Totals
A			75
B			25
Marginal Totals	60	40	100

Ok., for actual example numbers. If, for example, there were 100 voters total, then A got 75 votes and while B got 25. That is, the 75% and 25% assumed above.

Since the 100 voters were composed of 60 males and 40 females, we can calculate the *expected* number of males and females that would have voted for A, if there was no bias. That is, since 75% of all voters voted for 'A' then 60 * 0.75 = 45 males should have voted for 'A' and 40 * 0.75= 30 females should have voted for A. Similarly, B got 25% of the total vote and so she should have received 15 male votes and 10 female votes.

In the table below the row variable/category is the candidate, with values of A or B, while the column variable/category is gender, Male or Female. Look at the table below where I have shown the expected number of voters for A and B while also showing the actual number of voters according to gender. At this point I don't have how many Males or Females voted for candidate A, only that the total was 75. Similarly the number of voters voting for B would be 12 males and 10 females if there was no bias. These numbers in the margins are called ---- *marginal totals*.

So, just knowing this much, I can fill in what I would **expect** to be the numbers of voters in each cell, given that there is no bias.

TABLE 2. 2 X 2 Contingency Table (expected values)

Gender / Candidate	Male	Female	Total
A	(45)	(30)	75
B	(15)	(10)	25
Total	60	40	100

For example, the number in the A-Male cell comes from:

Since overall, 75% of voters voted for A then, of the 60 males, there should have been 0.75 * 60 = 45 males who voted for A. That is the **expected** value for that cell.

Similarly, there should have been 0.75 * 40 = 30 females voting for A and so on through the rest of the numbers.

Another way to look at this is to realize that the 75 voters must have come from 75% of 60 and 40. That is:

75 = 0.75 * (60 + 40) = 0.75 * 60 + 0.75 * 40 = 45 + 30

Similarly, the 25 votes must have come from:

25 = 0.25 * (60 + 40) = 0.25 * 60 + 0.25 * 40 = 15 + 10

But, What Were the Actual Values (Some texts call these Observed Values)?

From the overall outcome of the election we were able to calculate the expected voting pattern. But,

what were the actual numbers? Aha, suppose we also had exit polls on all the voters which told us which gender voted for which candidate (let's just suppose these were accurate)!

TABLE 3. 2 X 2 Contingency Table (actual values)

Gender / Candidate	Male	Female	Total
A	40	35	75
B	20	5	25
Total	60	40	100

Suppose the exit polls found that the actual number of males voting for A was 40, while the actual number of females was 35. For B, the actual numbers were 20 and 5. Compare these numbers with the expected values and you will see differences.

TABLE 4. 2 X 2 Contingency Table (observed - expected)/expected)

Gender / Candidate	Male	Female	Total
A	(40-45)/45	(35-30)/30	75
B	(20-15)/15	(5-10)/10	25
Total	60	40	100

Ok, the Observed and Expected Differ, So What?

Now that we have the last table which shows the discrepancies between expected and actual, scaled by the expected value. What are the implications of these differences? Start with an extreme case where the actual number *equals* the expected number in each cell. The difference in each cell would then be *zero*. This would mean that regardless of gender, the proportion of voters maintains the 75% /25% split. So we say: the variables are *independent*. *This is the criterion of independence - small differences of observed to expected.*

As the differences get larger and larger, we are less sure of independence and must rely on some kind of calculation to aid us in concluding when independence shades over into dependence. The usual calculation done in these cases is called the *Chi-square statistic*.

Chi-Square to the Rescue

The Chi-Square *statistic* is used to make decisions like this and makes use of the fact that if the observed and the expected numbers differ, then there is a chance of dependency, the more the difference, the higher the chance of dependency. To test whether or not I am dealing with independent variables, I calculate a special statistic called the Chi-Square statistic. If it's big, then I am less inclined to say the variable are independent. If small, then I am more inclined to say they are independent. (Recall the discussion above where the Chi-Square was zero and hence no chance of dependence. So to actually calculate this statistic what you do is:

How to Calculate the Chi-Square Statistic

For each cell, take the difference between expected and observed, square that difference and divide by the expected value. Do that for every cell and add them up, that's the *Chi-Square statistic.* Check the value you get against a standard table of Chi-square values to determine what are the chances of getting your number under the conditions of independence. If your number *exceeds* the

tabulated value, for the confidence level you have chosen, then you would reject the hypothesis of independence. Here is the calculation, first in words, and then in numbers:

Chi-square =

(observed male votes for candidate A - expected male votes for A)^2 / expected male votes for A +

(observed female votes for candidate A- expected females for A)^2 /expected female votes for A +

(observed male votes for candidate B - expected male votes for B)^2 /expected male votes for B +

(observed female votes for candidate B - expected female votes for B)^2/expected female votes for B

Chi-square = $(40-45)^2 / 45 + (35-30)^2 / 30 + (20-15)^2 / 15 + (5-10)^2 / 10$

Chi-Square = 50/9 = 5.56

Now we consult a standard table with degrees of freedom = 1, and confidence level at 95%, we see a value of 3.84. That is a *critical value.* Values to the right of this one occur only 5 times out of 100 repetitions of this experiment. This critical value represents our critical value for a 95% confidence level. Is our Chi-squared statistic larger than this number? Yes, and so we reject the hypothesis of independence and conclude that there is good evidence (not iron clad of course) for gender bias.

***End Cut to the Chase

Example II: A 2 X 2 Contingency Table of: Place of National Origin versus Gender

Consider a survey that asked 30 Males and 70 Females their place of national origin. Suppose there were three only categories of response: Asia, South America (SA), and Europe (Eu). Tabulating the overall totals of the responses gave the table below. Values in the margins of the table are called . . . *marginal values,* while values in the cells of the table are called . . . *cell values.* Suppose I was given this table of marginal values, and suppose that this was all I knew, since the individual cell values were not yet available to me.

:

TABLE 5. 2 X 2 Cross Tabulation (Contingency Table)

Gender / National Origin //	Male	Female	Total
Asia			20
Europe			30
South America			50
Total	30	70	100

Notice that this data doesn't give me cell details such as how many of the Asians were male or female, or how many Europeans were male or female, just marginal totals.

What I **Can** *Do While Waiting for the Individual Cell Values*

In spite of not knowing what the cell values are, I can still make some estimates of what these values *should* be, just working on the assumption that the place of origin and gender are *independent*. Independence means: knowing that someone in this survey is female, doesn't help me to bet on her nationality, similarly, knowing nationality doesn't help in guessing gender. If these variables were *dependent* though, knowing that a respondent is female *will* tell me something about her nationality. Ultimately we will be able to judge how independent/dependent these variables are by some

calculations using what is called the *Chi-Square test*.

Let's see how the *independence* assumption will give me table entries (note this is a pure calculation based on the independence assumption together with just the margin totals). O.K., here goes:

Using just the *nationality* percentages, calculated from the margin totals- notice that of the 30 males, 20% *should* be Asian (that is, 20/100), 30% *should* be European (that is, 30/100), and 50% *should* be South American. That means that out of the 30 total males, 20% of the 30 yields 6. So, there should be (I would *expect* to see) 6 males of Asian origin. The value '6' is called the *expected* cell value. I have placed these expected value within parentheses. It also follows that 30% of 30 yields an expected 9 Europeans, while 50% of 30 yields an expected 15 South Americans.

Similarly, using just the Gender marginal totals, of the 70 females, 20% should be Asian, 30% should be European, and 50% should be South American, again, assuming independence. That means that there should be 20/100 * 70 = 14, Asian females, 30/100 * 70 = 21 European females, and 50/100 * 70 = 35 South American females.

Below is the table of *expected values*, in parentheses.

:

TABLE 6. Assuming Independence - 2 X 2 Cross Tabulation (Contingency Table)

Gender/Origin	Male	Female	Nationality Marginal Totals
Asia	(6)	(14)	20
Europe	(9)	(21)	30
South America	(15)	(35)	50
Gender Marginal Totals	30	70	100

Now the observed Cell Totals Arrive

Suppose that I (finally) got the observed cell values from the surveyor as shown in the table below. How close are the observed values to the expected values? Just stop for a moment and suppose, just for argument sake that the values that the surveyor reported were exactly the expected values. I discuss this case next although, if that happened, I would hire another surveyor!

An Extreme Case: In an extreme case, suppose the survey values *exactly* matched the expected values! That is, my surveyor found 6 males from Asia, 9 from Europe, and 15 from South America. Similarly, 14 females from Asia, 21 from Europe, and 35 from South America. If that very rare event actually happened, I wouldn't need any calculations and could immediately conclude: The Gender and Place of Origin are Independent since what I got exactly matches what I would expect if the variables were independent.More formally: *There is no way I can reject the null hypothesis.*

Life gets a little more complicated when the survey data differs from the expected values as in the next table. Let me assume that the survey turned up the actual observed numbers shown in the table below. Parentheses are around the expected numbers and the actual numbers are without parentheses.

TABLE 7. - Expected and Observed Data: Contingency Table

Gender/Origin	Male	Female	Total
Asia	(6) 8	(14) 12	20
Europe	(9) 20	(21) 10	30
South America	(15) 2	(35) 48	50
Total	30	70	100

I notice there is some difference between what I expect and what I observe. The question is: is the difference large enough to cast real doubt on the hypothesis of independence?

Chi-Square to the Rescue

The Chi-Square *statistic* is used to make decisions like this and makes use of the fact that if the observed and the expected numbers differ, then there is a chance of dependency, the more the difference, the higher the chance of dependency.

To test whether or not I am dealing with independent variables, I calculate a special statistic called the Chi-Square statistic. If it's big, then I am less inclined to say the variable are independent. If small, then I am more inclined to say they are independent. (Recall the discussion above where the Chi-Square was zero and hence no chance of dependence. So to actually calculate this statistic what you do is:

For each cell, take the difference between expected and observed, square that difference and divide by the expected value. Do that for every cell and add them up, that's the Chi-Square.

For the numbers in the above table the calculations run as follows:

For example, for the first cell, observed = 8, expected = 6, so for this cell the calculation would be:

$(6-8)^2 /6$

Ch-Square= $(6-8)^2/6 + (14-12)^2/14 + (9-20)^2/9 + (21-10)^2/21 + (15-2)^2/15 + (35-49)^2/35$

$= 31.88$

O.k, so I got a Chi-square number, so what? To see if this number is large enough to cause me to reject the null hypothesis I need to consult a table, a math package or a captive statistician! Remember, if Chi-Square was zero, I could immediately conclude that the null hypothesis was as solid as it could possibly be and so the variables are independent. When that Chi-Square gets bigger though, I will need some help in deciding when I have to abandon the null hypothesis and accept the alternative hypothesis, that is, the variables are likely dependent.

Interpreting the Chi-Square Value

When I look at the Chi-Square tables, they are indexed by what is called *degrees of freedom*.

For right now, you can determine 'df' by the formula below:

df= (rows-1) x (columns -1)

For my 3x2 table, df = 2 * 1 = 2

So, looking at the tabulated values Chi Square for 2 degrees of freedom (df) and

95% confidence, I see the number 5.99.

Since my number was 31.88, which is way larger than 5.99, I have extremely strong evidence that the Gender and Place of Origin are NOT independent.That is, there is good reason to believe that knowing the gender of the respondent already suggests their place of origin.

The Chi-Square Distribution for degrees of freedom= 2 (If you wanted to know more)

Below I have written down the density function for Chi-Square with df=2, then the quantile number such that 95% of the curve lies to its left (that number is 5.99146). If I got a Chi-Square value greater than 5.99146 I would conclude that there was good reason to suspect dependence, that is Gender and Place of National Origin were not independent.

Finally as a check on that quantile number, I found the area under the curve from zero to 5.99146 and verified that it was indeed 95% (or 0.95 fraction of the total area).

```
chip = PDF[ChiSquareDistribution[2], t]
```

$$\frac{e^{-t/2}}{2}$$

This is the density function for Chi-Square df = 2

```
quantile = Quantile[ChiSquareDistribution[2],
  .95]
```

This is the number such that 95% of the curve lies to its' left

```
5.99146
```

```
Integrate[chip, {t, 0, 5.99146}]
```

Verify that the area from zero to 5.99146 is actually 0.95 (95%)

```
0.95
```

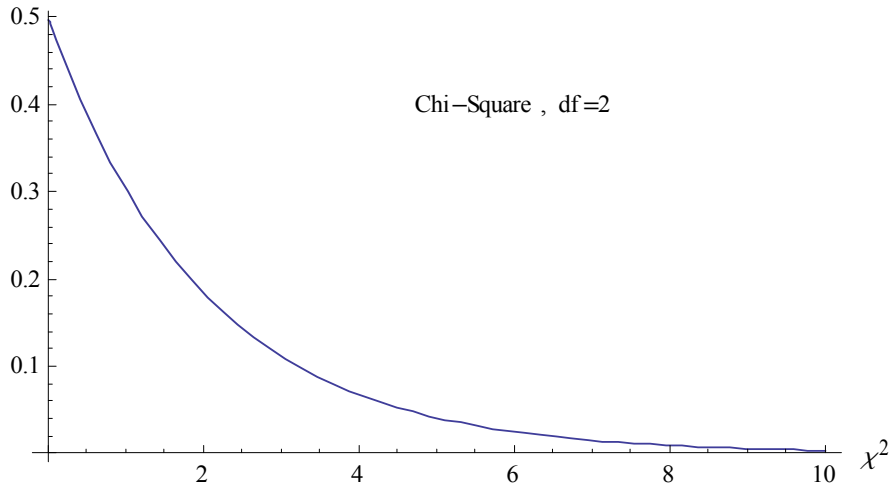

FIGURE 1. Chi-Square Detail for the Place of National Origin versus Gender

Example III: What Color is Your Helmet?

The entries in the table below are the frequency counts of a motorcycle accident study (from Triola 2006). Numbers in parentheses are *expected* values while bare numbers are actual observed counts. The observed values are counts of motorcycle riders in various states of 'dis-repair', who wore helmets of various colors. The class of population (the row variable) for the rows are Motorcycle Riders with two levels (subclasses): non-injured versus injured/killed. The column class of population is helmet color with three subclasses: Black, White, and Orange/Yellow.

The *control* row were cycle riders randomly selected at roadside stops and their helmet colors noted. The second row is cycle riders killed or injured together with their helmet colors. Is there a connection, are these variable independent, or does wearing a helmet make a difference?

Null Hypothesis: Helmet color is independent of injury/death incidence. That is, the null hypothesis says that your helmet color *doesn't* make a difference as far as injury or death goes.

Alternative Hypothesis: Helmet color is not independent of injury or death. That is, wearing different color of helmet *does* affect your chance of injury or death)

This amounts to calculating if the first variable is independent of the second variable. That is, are the proportions of black, white, and yellow/orange helmets of uninjured riders about the same as the injured/killed proportions? Lets see: In the table, the numbers in parentheses are the 'expected

values', calculated as if the variables were independent. For example, in column 1, the observed value is 491 while the expected value is 513.7 = 899/1232 * 704. The question here is whether or not this discrepancy, along with all the others in the table, can be accounted for by chance or is there some inter-dependence of the variables? Another way to look at these expected value calculations is to realize that the marginal total of 899 is 899/1232 of the grand total. So, the expected value of row 1 (Control) is:

899 = 899/1232 * (704 + 489 + 39) = 899/1232 * 704 + 899/1232 * 489 + 899/1232 * 39

899 = 513.7 + 356.8 + 28.5

TABLE 8. Case Control Study of Motorcycle Riders

	Black	White	Yellow/Orange	Totals
Control (not injured))	491 (513.7)	377 (356.8)	31 (28.5)	899
Case (injured/killed)	213 (190.3)	112 (132.2)	8 (10.5)	333
Totals	704	489	39	1232

For this table we calculate a sum of squares called that Chi-square statistic that determines to what extent the observed values match expected values. If they pretty much match, then the variables are probably independent. If they don't match, then the variables may not be independent, depending on the size of my calculated statistic. The result of my calculation gave me (which matched the article values):

Chi-squared = 8.76

df = (2-1)(3-1) = 2

From Chi-square math package calculations done after the first example, we see that the 95% quantile number is 5.99146. Since 8.76 is much larger, to the point of saying that it is extremely unlikely that the helmet color is independent of injuries. So, some conclusions can be inferred. According to the authors of the study, this finding led them to suggest that *visibility* of the riders decreased their chances of injury. (Hmmm, could you have concluded that without a study?) Notice though, that an expressed opinion, backed up by credible data, is much more powerful than opinion alone!

Test of Independence Conditions of Use

For the Chi-Square to be a reasonable test, the following conditions ought to be met.

Data in the tables was randomly selected. These data counts are denoted as 'O', observations.

For every cell in the table, the expected frequency 'E' is at least 5.

Note: the data don't have to be drawn from a normal population or any other specific distribution.

The test for independence is

Chi-square = Sum ($(O-E)^2$ /E)

The degrees of freedom: df = (rows-1)* (columns-1)

Example IV: An Exercise

Let me take a made up, but simplified example to see how these expected values can be calculated.

TABLE 9. A

	Black	White	Yellow/Orange	Totals
Control (no injuries)				22
Cases (injury/killed)				14
Totals	8	16	12	36

Just suppose that all I know was that there were a total of 36 cyclists (from direct questioning or from highway records) tallied where: 8 wore black, 16 white, and 12 yellow/orange helmets. Of those, 22 were noted as not injured, while records showed that 12 were injured or killed. At this point, helmet color and injury or not injury are not known in detail and only margin totals are known.

What are the Expected Values in This Table?

Suppose I don't yet know the actual observed cell values, and my question is: What should I *expect* to go into each cell, only knowing the row and column totals?

Well, first off, a total of 8 must be allocated within column 1 since it must total to 8. But what should I expect in the first cell of column 1? Of the total 8 in column 1, 22/36* 8 = 4.89 ought to go in that first cell while 14/3* 8 ought to go in the second cell of column 1, since the 8 units ought to be allocated in proportion to the row proportions.

TABLE 10. Expected Frequencies

	Black Helmet	White Helmet	Yellow/Orange Helmet	Totals
Control (no injuries)	8 *22/36 = 4.888	16 *22/36=9.7777	7.3333	22
Cases (injury/killed)	8 *14/36 =3.1111	6.2222	4.6666	14
Totals	8	16	12	36

TABLE 11. Actual Observed Values

	Black Helmet	White Helmet	Yellow/Orange Helmet	Totals
Control (no injuries)	3	8		22
Cases (injury/killed)	5			14
Totals	8	16	12	36

Notice that all of these entries are not arbitrary since the marginal total must be honored. For example, since the injury cases total 14 and I have filled in 5, that means 9 must be distributed across the rest of that row. Similarly, I must have 8 more in the white helmet column.

Sample Calculations

Chi-square = your turn!

df = 2

Example V: Pesticide Residues in Organic and Non-Organic Foods

The row totals and column totals are the frequencies for each category. Since not much organic food is grown, there is a disproportionate number of tests on conventional food products. The ques-

tion is, given the number of samples containing pesticide residue from organic versus non-organic foods, what health conclusions can be drawn?

. The table below is from Agresti and Franklin pg. 91. This is a USDA study in California. For example, out of 127 samples of organic product, there were 29 cases of pesticide residue. Out of 26571 conventional food product samples, 19485 pesticide residues were found.

In this case, pesticide residue, yes or no, is the response for the class of pesticide residues, having two levels (subclasses). The second class is type of food, with two levels (subclasses, organic and non-organic.

TABLE 12. Pesticide Observed Values

Food Type	Residue Present	Residue Not Present	Total
Organic Foods	29	98	127
Conventional	19485	7086	26571
Total	19514	7184	26698

If you would like to try your hand at this, calculate the expected values and then calculate the Chi Square value. (No surprise here, the organic food is overwhelmingly cleaner than the standard food).

Summary

These Chi-Square calculations are very helpful at the early stages of an analysis, often pointing to the need for further study. As I have noted above, all these calculations need to be verified in as many different ways as possible. There are numerous threats to validity that the researcher must be aware of and try to guard against, or at least lessen these threats via larger samples and randomization. I hope you like this tutorial and that it proves helpful in your future work. (Rob Rucker 2010-04)