Chi-Squared

This was found at http://pathwayscourses.samhsa.gov/eval201/eval201_3_pg7.htm

Contingency Tables

To determine chi-square, we need to set up a contingency table. A contingency table shows whether the likelihood of being in a category (such as marijuana user) is "contingent on" being in another category (such as the comparison group). This helps show whether a program makes a difference.

A contingency table presents the number and proportion of people falling into two or more categories on two or more nominal-level variables. In our case, we will be comparing two variables (drug use and program participation). Therefore, we will be looking at a 2x2 table.

A 2x2 table is the simplest kind of contingency table. Contingency tables can become much more complex, comparing several different levels of variables. Tables can be 2x4, 3x3, and so on.

Examples:

- 2x2 table
- 2x3 table
- 3x3 table
- 2x5 table

2x4? 3x3? Feel like you're in a lumber yard? It's not as confusing as it seems. Come, I'll walk you through the process.

Step 1. State your hypotheses.

Here are the hypotheses that relate to marijuana use:

$H_{0\ (null)}$: There is no difference between the participants' and comparison group's use of marijuana at posttest.

$H_{i\ (research)}$: There is a difference between the participants' and the comparison group's use of marijuana at posttest.

Step 2. Collapse your data.

In real life, when data are collected for the number of days someone used a drug in the past 30 days, many report no use and a few report many days. Because of the extreme scores reported, an evaluator collapses the data or recodes them.

What does that mean? Kids who report any drug use in the past 30 days are coded as "users." Kids who report no days are coded as "nonusers." Thus, if a participant has used a drug in the past 30 days, that participant will be coded as a user. It doesn't matter how many days the person used a drug. If the participant has not used any drugs within the past 30 days, he or she will be coded as a nonuser.

Because it uses data collapsed into nominal-level variables, chi-square is known as a "distribution free" statistic. It is a test that does not require normally distributed data. We can test for a relationship between program participation and drug use without looking at number of days used. We just need to know whether or not kids used drugs.

Step 3. Insert the collapsed data into your contingency table.

	Participants N=48	**Comparisons N=50**
Number and proportion of kids who smoked marijuana in past month	**Cell A** 20 (42%)	**Cell B** 20 (40%)
Number and proportion of kids who did not smoke marijuana in past month	**Cell C** 28 (58%)	**Cell D** 30 (60%)

Step 4. Add up totals for each row and column.

	Participants N=48	**Comparisons N=50**	**Total**
Number and proportion of kids who smoked marijuana in past month	**Cell A** 20 (42%)	**Cell B** 20 (40%)	
Number and proportion of kids who did not smoke marijuana in past month	**Cell C** 28 (58%)	**Cell D** 30 (60%)	
Total			

Step 5. Compare the frequencies.

As you can see from the table, 20 participants and 20 comparison group members smoked marijuana within the past month. Thirty comparison group members and 28 participants did NOT smoke marijuana. (Remember that two participants didn't answer the marijuana question at posttest.) The observed frequencies for the participant and comparison groups are almost identical.

However, we cannot just eyeball the numbers. We need to use statistics to make sure these differences are "real." For contingency tables, we always use a chi-square statistic to do this. Chi-square can help us determine if differences are **statistically significant**.

This is how we determine that the differences are not due to chance. Jack's evaluator used the chi-square test to determine statistical significance. He used this particular statistic because both the independent variable (program participation) and dependent variable (alcohol use) are nominal-level variables, in this case "yes" or "no."

If the percentage differences are small or the numbers of kids in the groups are small, you may worry that you won't see the same thing in another group getting the program. The way to resolve the issue is to find out whether the difference was statistically significant.

To determine if differences are statistically significant, we need to go through several more steps. Hang in there. It's not like climbing the Washington Monument--that has way more steps than this.

Step 6. Select a probability level.

In the social sciences, findings with more than 5 percent likelihood of happening by chance are generally considered to be "not significant." We express this likelihood as $p = 0.05$. This means that we are 95 percent sure that the differences are "real." (Think of it this way: 100 percent certainty - 5 percent chance = 95 percent certainty.)

Step 7. Calculate chi-square.

Yes, of course, there's a formula:

$$X^2 = \sum([f_o - f_e]^2 / f_e)$$

where

X^2 = chi-square

$\sum$

= summation f_o

= observed frequency
f_e = expected frequency

The formula looks complicated, but it isn't so hard. The steps are:

1. For each cell in your contingency table, subtract the expected frequency (what would be expected by chance) from the observed frequency (what actually happened).

2. Square the result.
3. Divide by the expected frequency.
4. Add up the resulting numbers.

O.K., so maybe it is tricky. Let's walk through the process step by step. We already know the observed frequencies of alcohol use (number of kids who drank and who did not drink). Next we need to calculate the expected frequencies.

To calculate the expected frequencies for all four cells-A, B, C, and D-we need to multiply the observed frequency totals. These are the column and row totals for the column and row in which the cell appears. Then we divide by the total number of kids. We'll go through these cell by cell.

Expected Frequency, Cell A: Participants Who Smoked Marijuana, Past 30 Days

Marijuana Use in Past Month	Participants	Comparison Group	Total
Yes	**Cell A** 20	20	
No	28	30	58
Total		50	98

Column Total x Row Total = 48 x 40
Expected Frequency (f_e) Calculation = (48x40)/100 = 19.2

Expected Frequency, Cell B: Comparison Group Who Smoked Marijuana, Past 30 Days

Marijuana Use in Past Month	Participants	Comparison Group	Total
Yes	20	**Cell B** 20	
No	28	30	58
Total	48		98

Column Total x Row Total = 50 x 40
Expected Frequency (f_e) Calculation = (50x40)/100 = 20

Expected Frequency, Cell C: Participants Who Did Not Smoke Marijuana, Past 30 Days

Marijuana Use in Past Month	Participants	Comparison Group	Total
Yes	20	20	40
No	**Cell C** 28	30	
Total		50	98

Column Total x Row Total = 48 x 58
Expected Frequency (f_e) Calculation = (48x58)/100 = 27.8

Expected Frequency, Cell D: Comparison Group Who Did Not Smoke Marijuana, Past 30 Days

Marijuana Use in Past Month	Participants	Comparison Group	Total
Yes	20	20	40
No	28	**Cell D** 30	
Total	48		98

Column Total x Row Total = 50 x 58
Expected Frequency (f_e) Calculation = (50x58)/100 = 29

Now we're ready to calculate the differences between observed frequencies and expected frequencies. Then we'll square them and divide by expected frequency. These calculations are shown in the table.

Cell	Observed Frequency (f_o)	Expected Frequency (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
A	20	19.2	.8	.64	.64/19.2=.333
B	20	20	0	0	0/20=0
C	28	27.8	.2	.04	.04/27.8=.001
D	30	29	1	1	1/29=.034
Total					X^2 = .368

So, our chi-square value is .368. Well, that and a dollar will buy us a cup of coffee. This does not really mean much to anyone without knowing whether the value is statistically significant. Let's see how we find that out.

Step 8: Calculate degrees of freedom.

Whether a particular chi-square value is statistically significant depends on something called degrees of freedom. Degrees of freedom depend on the number of categories for each variable. The formula for calculating it is: (Number of categories in the row variable - 1) x (Number of categories in column variable - 1). Using our marijuana use example above, we can plug in the numbers to find the degrees of freedom. Degrees of freedom = (2 categories in row variable - 1) x (2 categories in column variable - 1) = (2 - 1) x (2 - 1) = 1.

Step 9: Determine chi-square critical value.

To find out whether our chi-square value is significant, we need to check a table of critical values. This table shows the minimum chi-square value for various degrees of freedom. To be significant at a particular degree of freedom, the chi-square value must equal or exceed the critical value. The table shows that for 1 degree of freedom (which is what we always have in a simple 2x2 contingency table), we need a chi-square value of at least 3.84 to be significant at the .05 level.

Step 10. Determine if our chi-square value is greater than the critical value.

Our chi-square value for marijuana use is .368. Since .368 is less than the critical value (3.84), Jack's evaluator failed to reject the null hypothesis of no difference. He therefore concluded that there is not a difference between the participants' and the comparison group's use of marijuana. In other words, Jack's program did not appear to have an effect.

Program Evaluation 201

Alcohol Use

Chi-Square Calculations for Alcohol Use (steps 1-6)

Now let's take a look at the contingency table for alcohol use and work through the same steps.

Step 1. State your hypotheses.

Here are the hypotheses that relate to alcohol use:

H_0 (null): There is no difference between the participants' and comparison group's use of alcohol at posttest.

H_i (research): There is a difference between the participants' and the comparison group's use of alcohol at posttest.

Step 2. Collapse your data.

Kids who reported 0 times = nonusers
Kids who reported $\geq$1 time = users

Step 3. Insert the collapsed data into your contingency table.

	Participants N=50	**Comparisons N=50**
Number and proportion of kids who drank in past month	**Cell A** 20 (40%)	**Cell B** 38 (76%)
Number and proportion of kids who did not drink in past month	**Cell C** 30 (60%)	**Cell D** 12 (24%)

Step 4. Add up totals for each row and column.

	Participants N=50	**Comparisons N=50**	**Total**
Number and proportion of kids who drank in past month	**Cell A** 20 (40%)	**Cell B** 38 (76%)	
Number and proportion of kids who did not drink in past month	**Cell C** 30 (60%)	**Cell D** 12 (24%)	
Total			

Step 5. Compare the frequencies.

This table tells a different story for alcohol use than the one for marijuana. It looks like there is a real difference between the program participants and the comparison group.

Twenty kids from the participant group and 38 kids from the comparison group drank alcohol. From this, it would seem obvious that more kids from the comparison group were drinking alcohol.

However, we cannot just eyeball the numbers. We need to use statistics to make sure these differences are "real." For contingency tables, we always use a chi-square statistic to do this. chi-square can help us determine if differences are **statistically significant.**

This is how we determine that the differences are not due to chance. Jack's evaluator used the chi-square test to determine statistical significance. He used this particular statistic because both the independent variable (program participation) and dependent variable (alcohol use) are nominal-level variables, in this case "yes" or "no."

To determine if differences are statistically significant, we need to go through several more steps. Hang in there.

Step 6. Select a probability level.

In the social sciences, findings with more than 5 percent likelihood of happening by chance are generally considered to be "not significant." We express this likelihood as $p = 0.05$. This means that we are 95 percent sure that the differences are "real." (Think of it this way: 100 percent certainty - 5 percent chance = 95 percent certainty.)

Step 7. Calculate chi-square.

Here's the formula:

$$X^2 = \sum([f_o - f_e]^2 / f_e)$$

where

X^2 = chi-square

$\sum$ = summation

f_o

= observed frequency
f_e = expected frequency

The formula looks complicated, but it isn't so hard. The steps are:

1. For each cell in your contingency table, subtract the expected frequency (what would be expected by chance) from the observed frequency (what actually happened).
2. Square the result.
3. Divide by the expected frequency.
4. Add up the resulting numbers.

Let's walk through the process again step by step. We already know the observed frequencies of alcohol use (number of kids who drank and who did not drink). Next we need to calculate the expected frequencies.

To calculate the expected frequencies for all four cells-A, B, C, and D-we need to multiply the observed frequency totals. These are the column and row totals for the column and row in which the cell appears. Then we divide by the total number of kids. We'll go through these cell by cell.

Expected Frequency, Cell A: Participants Who Drank, Past 30 Days

Alcohol Use in Past Month	Participants	Comparison Group	Total
Yes	**Cell A** 20	38	
No	30	12	42
Total		50	100

Column Total x Row Total = 50 x 58
Expected Frequency (f_e) Calculation = (50x58)/100 = 29

Expected Frequency, Cell B: Comparison Group Who Drank, Past 30 Days

Alcohol Use in Past Month	Participants	Comparison Group	Total
Yes	20	**Cell B** 38	
No	30	12	42
Total	50		100

Column Total x Row Total = 50 x 58
Expected Frequency (f_e) Calculation = (50x58)/100 = 29

Expected Frequency, Cell C: Participants Who Did Not Drink, Past 30 Days

Alcohol Use in Past Month	Participants	Comparison Group	Total
Yes	20	38	58
No	**Cell C** 30	12	
Total		50	100

Column Total x Row Total = 50 x 42
Expected Frequency (f_e) Calculation = (50x42)/100 = 21

Expected Frequency, Cell D: Comparison Group Who Did Not Drink, Past 30 Days

Alcohol Use in Past Month	Participants	Comparison Group	Total
Yes	20	38	58
No	30	**Cell D** 12	
Total	50		100

Column Total x Row Total = 50 x 42
Expected Frequency (f_e) Calculation = (50x42)/100 = 21

Now we're ready to calculate the differences between observed frequencies and expected frequencies. Then we'll square them and divide by expected frequency. These calculations are shown in the table.

Cell	Observed Frequency (f_o)	Expected Frequency (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
A	20	29	-9	81	81/29=2.8
B	38	29	9	81	81/29=2.8
C	30	21	9	81	81/21=3.9
D	12	21	-9	81	81/21=3.9
Total					X^2 = 13.4

So, our chi-square value is 13.4. As you know, this does not really mean much to anyone without knowing whether the value is statistically significant. Let's see how we find that out.

Step 8: Calculate degrees of freedom.

Whether a particular chi-square value is statistically significant depends on something called degrees of freedom. Degrees of freedom depend on the number of categories for each variable. The formula for calculating it is:

(Number of categories in the row variable - 1) x (Number of categories in column variable -1)

Using our alcohol use example above, we can plug in the numbers to find the degrees of freedom.

Degrees of freedom = (2 categories in row variable -1) x (2 categories in column variable -1) = (2-1) x (2-1) = 1

Step 9: Determine chi-square critical value.

To find out whether our chi-square value is significant, we need to check a table of critical values. This table shows the minimum chi-square value for various degrees of freedom. To be significant at a particular degree of freedom, the chi-square value must equal or exceed the critical value.

The table shows that for 1 degree of freedom (which is what we always have in a simple 2x2 contingency table), we need a chi-square value of at least 3.84 to be significant at the .05 level.

Step 10. Determine if our chi-square value is greater than the critical value.

Our chi-square value for alcohol use is 13.4. Since 13.4 is larger than the critical value (3.84), Jack's evaluator rejected the null hypothesis of no difference. He therefore concluded that there is a difference between the participants' and the comparison group's use of alcohol.

From the information in the contingency table for alcohol use we could **see** that there were not as many kids drinking in the participant group as there were in the comparison group. After calculating the chi-square statistic, we now **know** that this difference is significant.