# Chapter 11:

# Chi Square

## What is the difference between:

1. Test of statistical significance
   (eg. t test, z test, chi square)

   and

2. Measures of association
   (e.g., lambda, gamma)

## Test of Statistical Significance:

Can the relationship between 2 variables (or 2 categories of the same variable), found in our sample, be generalized to the whole population or can the relationship found in our sample be attributed as a quirk of sampling?

In other words:

Does the relationship, found in our sample, actually exist in the population or is it due to sampling?

## Chi Square is a test of statistical significance.

Chi Square examines the probability that there is more than a random association between two variables.

## Measures of Association (such as Lambda and Gamma)

examine the size of the relationship or "association" between two variables regardless of the probability that the relationship found is due to a quirk in sampling (i.e., due to chance).

Does it make sense to report (or even examine) the measure of association if the test of statistical significance shows that the relationship found in the sample is due to sampling rather than being a real relationship likely to be found in the population?

Answer: it can depend. In the case of Chi Square, one limitation is Chi Square's sensitivity to sample size.

Consequently, if there is a small sample, chi square is more likely to suggest no relationship between two variables.

Therefore, if one finds a large measure of association with a small sample, it wouldn't hurt to note this.

How does Chi Square help us determine the probability that the relationship found in our sample also exists in the larger population?

Answer: Chi Square compares the observed relationship, found in the sample, to a "table of no relationship."

---

That is, it creates a table displaying the 2 variables as if there were no relationship and then compares this table to the table of actual data found from the sample.

---

If the values in the 2 tables are similar, then there is a high probability that, what relationship is seen in the sample, is due to sampling and not due to a real relationship in the population.

## More Specifically: How does Chi Square Work?

First, it examines the sample data for the two variables and their marginal totals.

Next, it creates a hypothetical table using the marginal totals and fills in all the columns and rows (the middle of the table) with what you would expect to find if there is no relationship between the two variables.

Finally, it compares this "table of no relationship" to the actually table

The more similar the actual table is to the table of no relationship the more likely that there is NO difference between the two variables in the population.

That is, the relationship found in the sample is likely due to chance.

Table 11.1 provides actual frequencies from a sample.

Table 11.1 Percentage of Men and Women Who Are First-Generation College Students

| *First-Generation* | *Men* | *Women* | *Total* |
|---|---|---|---|
| Firsts | 35.4%<br>(691) | 46.6%<br>(1,245) | 41.9%<br>(1,936) |
| Nonfirsts | 64.6%<br>(1,259) | 53.4%<br>(1,425) | 58.1%<br>(2,684) |
| Total (*N*) | 100.0%<br>(1,950) | 100.0%<br>(2,670) | 100.0%<br>(4,620) |

*Source:* Adapted from W. Elliot Inman and Larry Mayes, "The Importance of Being First: Unique Characteristics of First Generation Community College Students," *Community College Review* 26, no. 3 (1999): 8. Reprinted with permission.

Table 11.3 provides hypothetical frequencies assuming no relationship between variables.

Table 11.3 Expected Frequencies of Men and Women and First-Generation College Status

| *First-Generation* | *Men* | *Women* | *Total* |
|---|---|---|---|
| Firsts | 817.14 | 1,118.86 | 1,936 |
| Nonfirsts | 1,132.86 | 1,551.14 | 2,684 |
| Total (*N*) | 1,950 | 2,670 | 4,620 |

How are the hypothetical numbers created?

We need to calculate $f_e$ = Expected Frequency: the cell frequencies that would be expected in a bivariate table if the two variables were unrelated (statistically independent)

For each cell in the table:

$$f_e = \frac{\text{(column marginal) ( row marginal)}}{\text{Total N}}$$

Table 11.3 Expected Frequencies of Men and Women and First-Generation College Status

| *First-Generation* | *Men* | *Women* | *Total* |
|---|---|---|---|
| Firsts | 817.14 | 1,118.86 | 1,936 |
| Nonfirsts | 1,132.86 | 1,551.14 | 2,684 |
| Total (*N*) | 1,950 | 2,670 | 4,620 |

## How do we calculate Chi-Square?

$$X_2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where:

$f_o$ = observed frequencies
$f_e$ = expected frequencies

Chi Square is the test statistic that summarizes the differences between the observed and the expected frequencies in a bivariate table.

## How do we calculate Chi-Square?

Table 11.5 Calculating Chi-Square

| *First-Generation College Status and Gender* | $f_o$ | $f_e$ | $f_o - f_e$ | $(f_o - f_e)^2$ | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|---|---|
| Men firsts | 691 | 817.14 | –126.14 | 15911.2996 | 19.47 |
| Men nonfirsts | 1,259 | 1132.86 | 126.14 | 15911.2996 | 14.04 |
| Women firsts | 1,245 | 1118.86 | 126.14 | 15911.2996 | 14.22 |
| Women nonfirsts | 1,425 | 1551.14 | –126.14 | 15911.2996 | 10.26 |

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 57.99$$

Chi Square is the test statistic that summarizes the differences between the observed and the expected frequencies in a bivariate table.

How do we interpret the Chi Square Statistic?

That is, in our example what does the number, 57.99 mean?

Answer: we use a Chi Square distribution table to locate 57.99 and it's associated probability (Appendix D in text). Or, we observe computer results such as from SPSS.

To read the table we need to know the degrees of freedom.

With cross-tabulation data we find the degrees of freedom by using the following formula:

$$df = (r - 1)(c - 1)$$

Where:

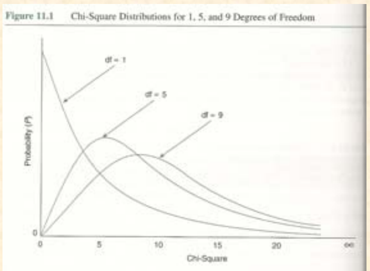
r = the number of rows
c = the number of columns

The df in a bivariate table can be interpreted as the number of cells in the table for which the expected frequencies are free to vary, given the marginal totals are already set.

In our example, there are four cells with only one cell free to vary.

Figure 11.1 shows the various chi-square distributions depending on the degrees of freedom.

As the df gets larger (above 30) the chi-square distribution begins to resemble the normal curve.




An examination of the Chi Square Distribution table, with a df of 1, shows us that:

the probability of obtaining a $X^2$ of 57.99 is less than .001 if the null hypothesis were true.

## Limitations of Chi Square:

1. Chi Square is sensitive to sample size. That is, the size of the calculated chi square is directly proportional to the size of the sample.
2. Chi Square is sensitive to small expected frequencies in one or more of the cells in the table.
3. While Chi Square shows us statistical significance it does not give us much information about the strength of the relationship or substantive significance. (This is left for measures of association)