Chapter 9: Chi-Square Analysis

- The hypothesis tests we have conducted so far have all made demanding assumptions of the data:
  - The data must be normally distributed
  - The tests required that the data be measured at the interval-ratio level
- In social science research, it is often difficult to satisfy both of these conditions.
  - In many instances the data researchers are asked to analyze are not measured at the interval-ratio level.
  - Moreover, even if the data are measured at this level, the distributions may be skewed.

Chapter 9: Chi-Square Analysis

- To allow researchers to make statements about the population in instances where the data do not meet these criteria, a series of statistics known as *non-parametric* measures, were developed.
- As mentioned, the non-parametric measures are useful because they allow us to test for significant differences, while relaxing some of the more rigorous data requirements.
- The non-parametric procedure that is covered in this chapter is called the Chi-Square test (X^2).
- Specifically, the manner of Chi-Square test we will be performing includes a comparison of 2 variables, thus, it is referred to as a 2-way Chi-Square.

Chapter 9: Chi-Square Analysis

- Essentially, when we perform a Chi-Square analysis, we are not going to compare differences between sample means, as was the case in the previous chapter. Rather, in this analysis, we compare frequency distributions of the two variables.
- Specifically, we use X^2 tests to determine whether observed frequencies in a distribution differ significantly from what is considered an even (or equitable) distribution.
- Examples of conditions in which a X^2 analysis is appropriate are as follows:
  - Does support for the death penalty vary across political parties, or is support for the death penalty consistent regardless of political party affiliation?
  - Do male and female students both have similar attitudes regarding campus security?

Chapter 9:
Chi-Square Analysis

For this example, imagine that we were interested in knowing whether support for mandatory minimum sentences varies by political party. In order to examine this question, we collect data for 300 individuals (100 Republicans, 100 Democrats, and 100 Independents). The data were distributed as follows:

	Republicans	Democrats	Independents	total
No	33	70	44	147
Yes	67	30	56	153
total	100	100	100	300

Chapter 9:
Chi-Square Analysis

- In order to perform this analysis, we begin by stating our Null and Research Hypotheses. However, we will have to write out each of the hypotheses.
  - H_0: Support for mandatory sentencing policies does not vary across political parties.
  - H_1: Support for mandatory sentencing policies differ by political party affiliation.
- Next, we will use the following formula to compute Chi-Square:

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

Chapter 9:
Chi-Square Analysis

- For ease of computation, we will next create a table that will allow us to derive Chi-Square. The first step is to create a table in which the observed frequencies (f_o) are listed as a single column:

fo	fe	(fo-fe)	(fo-fe)^2	(fo-fe)^2/fe
33				
70				
44				
67				
30				
56				
300				

- Note that in addition to the observed data, we also have to add four columns.
- The second column will contain the expected frequencies (f_e). The expected frequencies represent the number of observations we would expect to find if all cells were represented proportionally (i.e., if there is no association between the variables of interest).

Chapter 9:
Chi-Square Analysis

- The f_e values are computed for each of the observed frequencies. In this case, we will have to compute 6 expected frequencies. The formula required to calculate the expected values is as follows:

$$f_e = \frac{(R_t * C_t)}{N}$$

R_t=Row total for a given cell
C_t=Column total for a given cell
N=total sample size

- Again, we will need to use the above formula to compute the expected frequencies for each of the 6 observed values.

Chapter 9:
Chi-Square Analysis

	Republicans	Democrats	Independents	total
No	33	70	44	147
Yes	67	30	56	153
total	100	100	100	300

- The following describes how to calculate the necessary f_e values:

Republicans not in favor of mandatory sentences: (147*100)/300=49
Democrats not in favor of mandatory sentences: (147*100)/300=49
Independents not in favor of mandatory sentences: (147*100)/300=49

Republicans in favor of mandatory sentences: (153*100)/300=51
Democrats in favor of mandatory sentences: (153*100)/300=51
Independents in favor of mandatory sentences: (153*100)/300=51

- Note that the sum of the expected frequencies must equal the total sample size. The expected frequencies indicate the numbers we would expect if all of the categories are represented equitably, or if there is no association between these two variables.

Chapter 9:
Chi-Square Analysis

- Once we have computed the expected frequencies, we are then able to fill in the rest of the table.

fo	fe	(fo-fe)	(fo-fe)^2	(fo-fe)^2/fe
33	49	-16	256	5.22
70	49	21	441	9
44	49	-5	25	0.51
67	51	16	256	5.02
30	51	-21	441	8.65
56	51	5	25	0.49
300	300	0		28.89

- We obtain the computed Chi-Square value (X^2_{comp}) by summing all of the values in the final column. Thus, for this problem the X^2_{comp} is 28.89.
- In order to make a decision about our Null Hypothesis, we will have to compare X^2_{comp} *to a critical value* (X^2_{crit}).

Chapter 9: Chi-Square Analysis

- As with previous analyses, the critical X^2 value will be based on our alpha-level (either .05 or .01) and the degrees of freedom.
  - In this case we will use an alpha-level of .05.
  - The degrees of freedom in a Chi-Square analysis are the product of (r-1)*(c-1), where r is the number of rows in the table and c is the number of columns. In this case, we are analyzing a 2X3 table. Therefore, we have 2 degrees of freedom.
- Next, we refer to the Chi-Square table and use the above information to find the X^2_{crit} we will use for the purposes of our hypothesis test.
  In this case X^2_{crit}=5.991
- Finally, we compare X^2_{comp} and X^2_{crit} .
  If X^2_{comp} > X^2_{crit} we will reject the Null Hypothesis
  If X^2_{comp} < X^2_{crit} we will fail to reject (or retain) the Null Hypothesis

Chapter 9: Chi-Square Analysis

- In this case, X^2_{comp} (28.89) > X^2_{crit} (5.991), so we will reject the Null Hypothesis.
- Note that it is mathematically impossible to have a *negative* Chi-Square value. The Chi-Square can be zero, but never negative.
- In interpreting our results, it is not enough simply to state your decision regarding the Null and Research Hypotheses. Instead you will need to relate your decision back to the initial question. In this case, we will reject the Null Hypothesis, which assumes no difference in support for mandatory minimum sentences across political parties. Instead the data suggest that support for mandatory minimum sentencing policies differ significantly by political party affiliation.
- In this analysis, we have established that there is a significant association (or relationship) between support for sentencing laws and political party affiliation. However, what this analysis does not show is the strength of the relationship between these two factors.
- Simply because a significant relationship exists, this does not necessarily suggest that there is a strong association between the two variables. In order to determine the strength of the association, we have to perform a secondary analysis, called Cramer's V.

Chapter 9: Cramer's V

- The following calculation will allow us to determine the magnitude of the association between the variables included in the Chi-Square test.
- The procedure will generate a single value, which indicates the strength of the association. Before computing and interpreting this value, it is important to first discuss a few characteristics of this statistic.
  - Cramer's V is bound by 0 and 1.
  - Following convention, the strength of the relationship will be interpreted as follows:

    values between 0.0-.30 indicate a weak association.
    values between .31-.60 indicate a moderate association.
    values >.60 indicate a strong association.

Chapter 9: Cramer's V

- To compute Cramer's V, use the following equation:

$$V = \sqrt{\frac{X^2_{comp}}{(N)(min)}}$$

where "min" is a single value, the smaller of the two quantities (r-1) **or** (c-1).

- Note that we already have all of the information necessary to complete this calculation. Using the data from the previous problem, we can fill in all of the necessary values:

$$V = \sqrt{\frac{28.89}{(300)(1)}}$$

- Because the table we analyzed was a 3 (columns) by 2 (rows), the "min" value we need to use is 1 (2-1=1).

Chapter 9: Cramer's V

- Therefore, Cramer's V for this Chi-Square analysis is:

$$V = \sqrt{.0966} \qquad V = .31$$

- Based on this analysis, we will conclude that there is a statistically significant moderate relationship between political party affiliation and support for mandatory minimum sentencing policies.
- This example illustrates that there is a conceptual difference between a statistically significant relationship between two variables and the strength of the observed relationship.
- In other words, simply rejecting the Null Hypothesis we cannot necessarily assume that a strong relationship between the two variables exists.

Chapter 10: Correlation Analysis

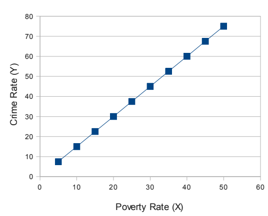
- This chapter introduces the concept of a correlation, also referred to as a zero-order correlation, or as a Pearson's correlation coefficient.
- A correlation is similar to the calculation of Cramer's V, which we discussed previously, in that both indicators provide information regarding the level of association between two factors.
- However, unlike Cramer's V, correlation coefficients indicate not only the strength of association between two factors, but also the direction of the underlying relationship.
- In addition, we will also be able to perform a hypothesis to determine whether the observed correlation is statistically significant. That is to say, whether the evidence suggests that the correlation is likely a product of sampling error, or if the coefficient reflects a true relationship in the population.

Chapter 10:
Correlation Analysis

- The prior inferential statistical procedures that we have studied this semester have provided information as to whether or not there were statistically significant differences between sample means, or if an observed distribution is significantly different from an expected distribution.
- In instances where we have found statistically significant differences (i.e., we have rejected our Null Hypothesis), the implicit conclusion is that there is an association between the variables being analyzed.
- Until we covered the calculation of the Cramer's V statistic, we have not been able to measure the strength of the association between two variables.
- Correlation coefficients, however, are designed to provide information about the strength and direction of the association between two variables.
- It is important to point out that despite their strengths, researchers are not able to draw conclusions regarding any *causal* association between variables. Recall, that "correlation does not imply causation."

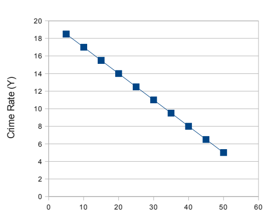
Chapter 10:
Correlation Analysis

- The direction of the association between two variables is either positive (i.e., direct) or negative (i.e., inverse).
  - A positive relationship is one in which similar values on the X and Y variables tend to cluster together. That is, high values on the X variable are associated with high values on the Y variable. Similarly, low values on the X variable tend to be associated with low values on the Y variable.
    - An example of a positive relationship is the correlation between neighborhood levels of poverty and criminal deviance.



Chapter 10:
Correlation Analysis

- A negative relationship is one in which opposing values on the X and Y variables tend to cluster together. That is, high values on the X variable are associated with low values on the Y variable. Similarly, low values on the X variable tend to be associated with high values on the Y variable.
  - An example of a negative relationship is the correlation between neighborhood levels of educational attainment (i.e., the percent of the population with at least a bachelor's degree) and criminal deviance.



Chapter 10:
Correlation Analysis

- It is important to point out that correlations that we will discuss in this course measure a linear association between two continuous variables. The authors describe curvilinear correlations, yet such associations will not be discussed in this class.
- In addition to the direction of association, correlation coefficients also provide information regarding the strength (or magnitude) of the observed relationship (similar to the Cramer's V statistic). The chart below indicates the conventional rules we will use to interpret the strength of a correlation coefficient:

1.0	Perfect positive relationship
.61 to .99	Strong positive relationship
.31 to .60	Moderate positive relationship
0 to .30	Weak positive relationship
-.30-0	Weak negative relationship
-.60 to -.31	Moderate negative relationship
-.99 to -.61	Strong negative relationship
-1.0	Perfect negative relationship

Chapter 10:
Correlation Analysis

- The following formula we use to calculate a zero-order correlation coefficient:

$$r_{xy}=\frac{\sum XY-N\bar{X}\bar{Y}}{\sqrt{(\sum X^2-N\bar{X}^2)(\sum Y^2-N\bar{Y}^2)}}$$

- The symbol used to describe the correlation coefficient "r_{xy}" simply indicates that the resulting value is the association between two variables (x and y).
- As you can see, the components of this equation are all based on calculations that we are familiar with, as similar procedures have been used in previous calculations.

Correlation Analysis
Example

The following example is taken from the textbook (Chapter 10; problem 6).

As a researcher, you are interested in studying whether there is an association between attitudes related to "victimless" crimes. Specifically, you want to know whether attitudes regarding the legalization of prostitution (X) are associated with attitudes regarding the legalization of marijuana (Y). You draw a sample of six individuals and obtained the following information. (Note: questions were based on a 10-point scale with lower values indicating more permissive attitudes).

X	Y
1	2
6	5
4	3
3	3
2	1
7	4

Correlation Analysis Example

- In order to calculate the correlation coefficient, we will first need to compute the mean for both X and Y.
  - Mean of X=(23/6) or 3.83
  - Mean of Y= (18/6) or 3.00
- The next step will be to add three additional columns to the table that contains the original data:

	X	Y	XY	X^2	Y^2
	1	2			
	6	5			
	4	3			
	3	3			
	2	1			
	7	4			
Sum:					

Where: XY is the product of X*Y
X^2 is the value of X-squared
Y^2 is the value of y-squared

Correlation Analysis Example

- Below is the completed table. Note that for the number we will include in the correlation equation are the *sums* from each of the three last columns:

	X	Y	XY	X^2	Y^2
	1	2	2	1	4
	6	5	30	36	25
	4	3	12	16	9
	3	3	9	9	9
	2	1	2	4	1
	7	4	28	49	16
Sum:			83	115	64

- With this information, we are now able to generate the correlation between attitudes related to victimless crimes.

Correlation Analysis Example

$$r_{xy}=\frac{\sum XY-N\bar{X}\bar{Y}}{\sqrt{(\sum X^2-N\bar{X}^2)(\sum Y^2-N\bar{Y}^2)}}$$

$$r_{xy}=\frac{83-(6*3.83*3)}{\sqrt{(115-6*14.67)(64-6*9)}}$$

$$r_{xy}=\frac{83-68.94}{\sqrt{(27)(10)}}$$

$$r_{xy}=\frac{14.06}{16.43}$$

$$r_{xy}=.86$$

- Based on this calculation, it appears that there is a strong, positive correlation (because r_{xy} > .60) between attitudes regarding legalization of prostitution and attitudes regarding legalization of marijuana.

Correlation Analysis Example

- The final step is to determine if the observed correlation is statistically significant. As with previous types of analyses, this requires that we conduct a hypothesis test. This test will allow us to make a determination whether the evidence suggests that the observed association reflects a true relationship in the population, or if it appears that the observed correlation is likely due to sampling error.
- Essentially, this hypothesis test involves a comparison of our computed correlation coefficient to a critical correlation coefficient.
- As was the case with previous calculations, the critical correlation coefficient is based on two factors, the alpha-level and the number of degrees of freedom.

Correlation Analysis Example

- For this problem, we will use an alpha-level of .05.
- The number of degrees of freedom (df) for a correlation analysis is determined by subtracting 2 from the sample size: df=n-2. In this example, df=4
- Next, we will obtain the critical correlation coefficient from Table F (Titled "Critical Values of r at the .05 and .01 levels of significance") in the text. The critical value (r_{crit}) for an alpha-level of .05 and 4 degrees of freedom is .8114.
- Finally, we can compare the computed and critical correlation values:
  - if $|r_{xy}| > r_{crit}$ then we will reject the Null Hypothesis, which assumes that there is no relationship between the variables of interest in the population. In other words, when we reject the Null Hypothesis, this indicates that the observed correlation **is** statistically significant.
  - if $|r_{xy}| < r_{crit}$ then we will fail to reject (or retain) the Null Hypothesis, which assumes that there is no relationship between the variables of interest in the population. In other words, when we fail to reject the Null Hypothesis, this indicates that the observed correlation **is not** statistically significant.

Correlation Analysis Example

- In this example, the absolute value of r_{xy} (.86) is greater than r_{crit} (.81). Therefore, we will conclude that the observed relationship is statistically significant.
- It is important to remember that when interpreting your correlation coefficient, I want you to refer to all three key pieces of information provided: strength, direction, and statistical significance.

Partial Correlation Coefficient

- The zero-order correlation coefficient provides information about the strength of the association between two variables. Correlation coefficients, while an important first step to understand the underlying relationship between two factors, must be interpreted with some caution.
- As the authors explain, it is possible that the observed correlation does not reflect a true relationship between two variables. Before researchers are comfortable concluding that there is an association between the two indicators, they will perform an additional test to determine if the observed correlation may be influenced by a third, unobserved factor.
- Based on the results from this secondary analysis, researchers will have a more confidence in the conclusion that is drawn about the relationship between two variables.

Partial Correlation Coefficient

- In order to gain additional information regarding the correlation between X and Y, we will calculate what is referred to as a partial correlation coefficient.
- A partial correlation coefficient is a useful next step because it allows a researcher to remove, or partial out, the common association that both X and Y have with a third variable Z. By removing the influence that Z has on both X and Y, we will be able to re-examine the strength of the correlation between X and Y.
- In other words, it is possible that our interpretation of the strength, direction, and significance of the correlation between X and Y will change depending on the relationship that both X and Y have with the third factor.
- Again, it is the responsibility of the researcher to try to rule out possible alternative or conflicting explanations before they conclude that a relationship exists between the two variables of interest.

Partial Correlation Coefficient

- A number of possible outcomes to the observed correlation between X and Y may occur once partial correlation coefficient is calculated:
  - r_{xy} remains largely unchanged. If the zero-order correlation and the partial correlation coefficients are nearly the same, this suggests that the third variable does not have much of an impact on the observed relationship
  - r_{xy} becomes dramatically weaker. If the magnitude of the partial correlation coefficient is much weaker, especially if it approaches zero, this is a condition called spuriousness. When a relationship is spurious, this suggests that the zero-order correlation is an artifact of the high correlation that both X and Y have with Z. In other words, there is no true relationship between X and Y, but rather, the initial correlation was masking the association that both factors have with Z.
  - r_{xy} changes direction. Although rare, it is possible for the zero-order and partial correlation coefficients to have different signs. When such a situation occurs it is important to point out that by partialling out the effect of a third factor, a researcher has an entirely different interpretation of the association between the two variables.

Partial Correlation Coefficient

- To calculate a partial correlation coefficient, we will need to include two additional correlation coefficients, specifically r_{xz} and r_{yz}. Note that as the symbols suggest, these additional coefficients represent the correlations between our two initial variables and the third factor, which we refer to as Z.
- Below is an example of a correlation matrix. As the book describes, because correlation coefficients are reflexive, only values above or below the main diagonal are reported.

	X	Y	Z
X	1.0	rxy	rxz
Y		1.0	ryz
Z			1.0

- As you can see in the above table, the correlation between any given pair of variables is determined by the column and row labels.

Partial Correlation Coefficient

- The formula we will use to calculate the partial correlation coefficient is as follows:

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^{2}}\sqrt{1 - r_{yz}^{2}}}$$

- To compute the partial correlation, we will use the information from the previous example. However, imagine that you have reason to suspect that there is another factor that may have influenced the correlation between attitudes regarding victimless crimes. Specifically, you have reason to suspect that educational attainment (i.e., years of schooling completed) is related to both attitudes related to prostitution and marijuana. To examine this question, you computed the correlation between X and Z and Y and Z. The resulting correlation matrix is below:

	X	Y	Z
X	1.0	0.86	0.87
Y		1.0	0.90
Z			1.0

- With this information, we can now recalculate the correlation between X and Y, removing the effect of Z.