Statistics Notes

Chapter 10 - Chi-Square Test & Measures of Association

- **Chi-square test** - inferential statistic technique designed to test for significant relationships between 2 variables organized in a bivariate table
- Chi-square requires **no assumptions** about the shape of the population distribution from which a sample is drawn (recall, the t-test required that the sample size be 50 or larger or the distribution was normal)
- chai-square test can be applied to **nominally** or **ordinally** measured vars

Statistical Independence:

- Independence (statistical) - the absence of association between 2 cross-tabulated vars. The percentage distributions of the dependent var within each category of the independent var are identical. In other words, 2 vars don't influence each other at all.

Hypothesis Testing w/Chi-Square:

1.) Making assumptions
2.) Stating the research & null hypothesis & selecting alpha level
3.) Selecting the sampling distribution & specifying the test statistic
4.) Computing the test statistic
5.) Making a decision & interpreting the results
NOTE: 3 usual alpha levels used in social sciences: .05, .01, .001

The Assumptions:

- The chi-square test requires no assumption about the shape of the populaton distribution from which the sample was drawn
- However, like all inferential techniques, it assumes **random sampling**
- chi-square test can be applied to vars measured at a **nominal** and/or **ordinal** level of measurement

Stating Research & Null Hypothesis:

- Research hypothesis (H1) states that the 2 vars are related in the population
- Null hypothesis (H0) states that no association exists between the 2 cross-tabulated vars in the population, & therefore the vars are statistically independent
- With T-test, you use numbers as well as symbols when stating hypotheses but with the Chi-test, you simply use words

The Concept of Expected Frequencies:

- **Expected frequencies (fe)** - the cell frequencies that would be **expected** in a bivariate table if the 2 vars were statistically independent, assuming vars are statistically independent.

- **Observed frequencies (fo)** - the cell frequencies actually observed in a bivariate table.
- Sample data is our observed frequencies
- Frequencies always add up to the column total

Calculating Expected Frequencies
column marginal - column total
row marginal - row total

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

Calculating the Obtained Chi-Square

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

f_o = observed frequencies

f_e = expected frequencies

The Sampling Distribution of Chi-square

- The sampling distribution of chi-square tells us the **probability** of getting values of chi-square, assuming no relationship exists in the population
- The chi-square sampling distributions depend on the degrees of freedom
- the X^2 (chi) sampling distribution is not one distribution, but a family of distributions
- The distributions are always **positively skewed**. The research hypothesis for the chi-square is always a 1-tailed test (right-tailed)
- Chi-square values are always positive, the min possible value is 0, with no upper limit to its max value.
- As the number of degrees of freedom increases, the X^2 (chi) distribution becomes more symmetrical
- if observed & expected frequencies are the same, you get a chai-square of zero
- smaller your chai-square, the more likely you aren't able to reject the null hypothesis. The larger, you're more likely to reject the null hypothesis

Determining the Degrees of Freedom:

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

where

r = the number of rows

c = the number of columns

Distribution of the Chi:

- pg 380
- At the top row, you see the alpha levels
- At the chi-critical value (values seen on pg 380), you can reject the null hypothesis (if it's at or above the critical value, you can reject but if it's below the critical value, we fail to reject)

Limitations of the Chi-Square Test:

- Tells us there is a relationship but doesn't tell us the strength of the relationship or its substantive significance in the population
- Sensitive to sample size. The size of the calculated chi-square is directly proportional to the size of the sample, independent of the strength of the relationship between the vars. To reject null hypothesis, increase sample size

Measure of Associations:

- The measure enables us to use a single summarizing measure or number for analyzing the relationship between 2 vars
- The measure tells us the strength of the relationship & at times its direction (positive or negative)
- The measure of association is a single summarizing number that reflects the strength of the relationship, indicates the usefulness of predicting the dependent var from the independent var, & often shows the direction of the relationship

Types of Measure of Association:

- In this chapter, we'll only talk about measures for nominal & ordinal vars only
- Lambda & Cramer's V (measure of association for nominal vars)
- Gamma (measure of association for ordinal vars)

Proportional Reduction of Error:

- **PRE** - concept that underlies the definition & interpretation of several measures of association. PRE measures are derived by comparing the errors made in predicting the depent var while ignoring the independent var with errors made when making predictions that use info about the independent var
- Measures the amount of errors when you predict dependent var.
- PRE tells you how well the independent var can be used to predict the value of the dependent var
- All measures of association are based on the concept of proportional reduction of

error (PRE)

- According to PRE, 2 vars are associated when info about one can help us improve our prediction of the other
- the formula for PRE is:

$$PRE = \frac{E1-E2}{E1}$$

E1 = errors of prediction made when the independent variable is ignored.

E2 = errors of prediction made when the prediction is based on the independent variable.

- When choosing the category for your prediction, choose the category with the larger amount

PRE Measures of Association:

- The PRE can range from 0.0 to +1.0
- A PRE of 0.0 indicates that the 2 vars aren't associated; info about the independent var will not improve prediction about the dependent var
- A PRE of + 1.0 indicates a perfect positive or negative association between the vars
- Measure the strength of the relationship with this chart:

-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00
Perfect negative Relationship	Very Strong negative relationship	Strong negative relationship	Moderate negative relationship	Weak negative relationship	No relationship at all	Weak positive relationship	Moderate positive relationship	Strong positive relationship	Very Strong positive relationship	Perfect positive Relationship

Two PRE Measures: Lambda, Cramer's V & Gamma:

	Appropriate for…
• *Lambda* λ	NOMINAL variables
• *Cramer's V*	Chi-Square related measure
• *Gamma* γ	ORDINAL & DICHOTOMOUS NOMINAL variables

Lambda: Measure of Association for Nominal:

- Asymmetrical measure of association, lambda is suitable for us w/nominal vars & may range from 0.0 to 1.0
- Provides us with an indication of the strength of an association between the independent & dependent vars

- Lambda is a PRE & follows the basic formula;

$$\text{Lambda} = \frac{E1 - E2}{E1}$$

where:

$$E1 = N_{total} - N_{\text{mode of dependent variable}}$$

$$E2 = \sum_{\substack{\text{for all} \\ \text{categories}}} (N_{category} - N_{\text{mode for category}})$$

- for the 1st E, you use the greatest row total
- for the 2nd E, you can use the greatest column total & for the category part, you simply add up all the mistakes you made, so if the column choices were 104, 85, & 50, you'd choose 104 then you'd say you made 135 mistakes by doing so (85+50), then you repeat this step for all the columns

Guidelines for Calculating Lambda:

- The numbers will change when you change an independent var to the dependent or vice versa
- Make sure you know which var is the independent & which var is the dependent so that you have the correct lambda calculation

Cramer's V:

- Based on the value of chi-square & ranges between 0 to 1. 0 indicating no association & 1 indicating perfect association
- Because it can't take negative values, it's considered a nondirectional measurement
- Formula:

$$\textit{Cramer's } V = \sqrt{\frac{\chi^2}{N \times m}}$$

where m = smaller of $(r \pm 1)$ or $(c \pm 1)$.

Symmetrical Measure of Association:

- A measure whose value will be the same when either var is considered the independent var or the dependent var
- **Gamma** is a symmetrical measure of association

Finding Gamma:

- Purpose of gamma is to determine whether there is an association between the independent & dependent vars
- We want to determine whether the association is positive or negative, gamma

can help us answer this question

Symmetrical Measures of Association:

- **Gamma** - suitable for use w/ordinal vars or with dichotomous nominal vars (which are vars that can only have 2 possible values, for example: sex). It can vary from 0 to pos. or neg. 1 & provides us w/an indication of the strength & direction of the association between the vars

Measures of Association:

- Measure of association - a single summarizing number that reflects the strength of the relationship. This statistic shows the **magnitude** and/or **direction** of a relationship between vars
- **Magnitude** - The closer to the absolute value of 1, the stronger the association. If the measure equals 0 then there is no relationship between the 2 vars.
- **Direction** - The sign on the measure indicates if the relationships positive or negative. In a positive relationship, when one var is high so is the other. In a negative relationship, when one var is high, the other is low (inverse relationship).