Geography 301a: Lab Number 1

MEASURES OF ASSOCIATION: Proportional Reduction in Error Methods (PRE)

Association means that knowing an individual's score on one variable (or the independent variable) helps reduce uncertainty or error in predicting the score on the other (dependent) variable.

More technically, two variables are associated if, and only if, the conditional distribution of Y varies with levels of X.

How two variables relate in a sample is one thing, but does that sample reflect the population at large. In geography, and the social sciences in general, one of our goals is to determine if our data sample reflects society with a certain confidence level.

The χ^2 test tells us whether our sample reflects the population as a whole. The null hypothesis would state that there is "no association" between the sample and the population at large. This means that the conditional distributions of Y are equal across X. We calculate the expected frequencies if the null hypothesis is true and compare those frequencies to the actual observations. Remember the rules for the Chi Square test: H_o states that $\chi^2=0$ for the whole population; the χ^2 table gives the critical values for various levels of alpha; and if $\chi^2 >$ the critical value from the table, then reject the H_o.

Association means a difference in the conditional distribution of Y; but how do we index the degree of difference? One idea is "how many errors do we make predicting the values of Y if we know X?" This can be expressed as a percentage of the errors we make predicting Y *not* taking X into account.

χ^2 measures in a vague manner the level of association between two variables. Goodman and Kruskal have developed a measure which allows one to improve upon the ability to predict the classification of one variable knowing the value of the other variable.

The measures of association which use χ^2 tell us little about the strength of the relationship - only whether there is a relationship or not. The values of measures such as Phi Φ, the coefficient of contingency, and Cramer's V provide us with a value of association, but it is more helpful to know how much we can improve upon our predictions of the data, and the accompanying population, if we know *by how much* we can improve our predictions through the Proportional Reduction in predictive Error.

There are many measures of PRE: Goodman-Kruskal tau, phi, lambda, Cramer's V, Pearson's contingency C, and the uncertainty or contingency coefficient. They all have different rules and procedures, but their purposes are similar - the proportional reduction of error.

The following are the formulae needed to complete this lab:

Chi Square:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Phi:

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

where N = the total number of observations, normally used for 2 x 2 tables

Coefficient of Contingency:

$$c = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Cramer's V:

$$V = \sqrt{\frac{\phi^2}{\min(r-1),(c-1)}}$$

where ϕ^2 = phi squared

Lambda:

$$\lambda = \frac{\text{misclassified in situation 1 - misclassified in situation 2}}{\text{misclassified in situation 1}}$$

To determine the dependent misclassified value, look at the rows. The row with the largest sum total value will dominate the rows, and therefore predicting the other values for the rows is harder - they would get misclassified. The dependent misclassified value = (N - the largest row total).

To determine the independent misclassified value, look to the columns. Follow the same procedure for each column as you did for the rows. The independent misclassified value = (the column sum - largest value within a column), you must do this for each column, then add up the misclassified cases for the columns.

QUESTIONS

Use the following data for all the questions. Please show your work.

	Work Experience (months)			
Attitude	6	7-9	10	**TOTAL**
Favorable	10	12	55	77
Indecisive	5	9	10	24
Hostile	8	12	26	46
TOTAL	23	33	91	147

source: fictional data

1: Calculate χ^2 at 95% confidence.

2: Calculate phi.

3: Calculate the Coefficient of Contingency.

4: Calculate Cramer's V.

5: What do these four above values tell us about the association between the variables?

6: Calculate lambda.

7: By knowing the values of Y, to what degree would we be able to better predict the values of X?

8: Comment on the magnitude of the value for lambda.

9. When is χ^2 a more appropriate method of analysis than a PRE approach?

Part 2

1. Now that you've done some of the calculations by hand its time to try the techniques on data that you have collected. The data should be a cross tabulation table with 2 variables. One of the variables should be designated as the independent variable and one as the dependent. While not all of the techniques distinguish between them, some do.

In you write up, you need to provide a short introduction as to where you variables come from and what is the relationship you hypothesize. Sources of error need to be identified.

Data can often be obtained from journal articles relevant of your subfield of geography. If you have difficulty in finding some data please come see me (the instructor), the TA or one of the instructors of your other geography courses (they are usually happy to help students). One of the things that this illustrates is that data is not always easy to come by.