CORRELATION COEFFICIENTS

Thus far we have discussed chi-square (X^2), a measure of whether or not two variables are independent.

Correlation coefficients or measures of association provide additional information about the strength of association between two variables. Like X^2 they are summary statistics, representing in a single number some characteristic of the distribution of numbers in a crosstabulation of two variables.

A Key Consideration:
Level of Measurement

Two questions must be answered in choosing the appropriate measure of association:

- Is the dependent variable (effect) nominal, ordinal or interval?

- Is the “independent variable” (cause) nominal, ordinal or interval?

Common Measures of Association

Nominal

Phi

Lambda

Ordinal

Yule's Q (Gamma)

Somer's d

Tau b

Interval

Pearson Product-Moment Correlation

Reducing Ignorance:
A "Gambling" Logic

Consider the following table. It records the percentages of 200 votes cast in a local school board election for candidate A and candidate B:

		Gender		
		M	F	
Vote	A	20%	60%	40%
	B	80%	40%	60%
		100%	100%	

Assume you are ignorant of every characteristic of the individual voters in an election and you read the newspapers the day after the election to find out who won.

Who won the above election?
What was the final vote?

Questions, Part I:

Assume you are to be introduced to all 200 voters who participated in this election. Before you meet them, your teacher asks you to guess how each one of them voted. What is your best guess about each one? How many times will you guess correctly?

Answers, Part I:

Your best guess will be to guess that the voter supported candidate B. You will be right 60% of the time and wrong 40% of the time.

Questions, Part II (New Information):

Once introduced to each voter, you are able to recognize their gender (or at least I hope so). Because you have the information from our survey summarized in the cross-tabulation between ***Vote*** and ***Gender*** above, you are able to make more precise predictions.

What is your new best guess about the vote of each individual you meet? How often will you be correct?

Answers, Part II (New Information):

If the voter is Male, your best guess is that he voted for B. If the voter is female, your best guess is that she voted for A.

If you guessed B for males you'd be right 80% of the time. If you'd guessed A for females you'd be right 60% of the time.

Overall, you’re be right more often knowing the gender of the voter than not knowing the gender of the voter. You’re now getting 80% of males right rather than 60% of them right. Your now getting 60% of the females correct after getting 60% of them wrong with your initial guess. Overall, your percent of correct guesses has gone up from 60% to 70%.

Let’s assume that men and women are equally represented in the school board election.

Males:	Guess B. 80 right, 20 wrong
Females:	Guess A. 60 right, 40 wrong
Totals	140 right (70%) 60 wrong (30%)

If new information provided by a second variable helps you to better understand the marginal frequencies of the first variable be say there is an "association" or a "correlation" between them.

Lambda is the correlation coefficient that captures this logic. It is also called a PRE statistic. PRE stands for "Proportional Reduction of Error". In ignorance you make 40% errors. With information about the gender of each voter, you make 30% errors. Thus you reduce your errors by 10% and lambda equals 0.25

Lambda

$$\lambda = \frac{\text{Errors (w/o info)} - \text{Errors (w/info)}}{\text{Errors (w/o info)}}$$

$$= \frac{80 - 60}{80} = .25$$

Problems with this approach:

It doesn’t work very well if both columns yield the same prediction.

Consider the following table:

		Gender		
		M	F	
Vote	A	20%	40%	30%
	B	80%	60%	70%
		100%	100%	

What is your best guess for each voter based only on the marginal frequencies of “Vote”?

What if you know whether each voter is male or female? Does this information help you improve your best guess or reduce your prediction error?

Talking about Correlations

Does Size Matter?

Yes, bigger is better - unless you hope to prove the null hypothesis that no relationship exists between two variables.

How do we describe size?

It really depends on your experience with certain types of data and specific correlation coefficients. While almost all correlation coefficients vary in absolute value from 0 to 1, the same data will produce very different numbers from one measure of association to the next. Gamma will often be the largest (because it doesn't take ties into account). Tau b and other correlation coefficients will inevitably be smaller.

And of course whether any coefficient is significant depends on the sample size. A weak coefficient may represent a relationship that is statistically significant (different from 0) if the sample size is large enough and a strong correlation may be insignificant if it is based on very few cases.

One possible typology:

0.0 to 0.1	"No" relationship
0.1 to 0.3	"Weak" relationship
0.3 to 0.5	"Moderate" relationship
0.5 to 0.8	"Strong" relationship
0.8 to 0.9	"Powerful" or "Trivial"
1.0	"Fake"