Proportional Reduction in Error & Measures of Association

Univariate Reminders

- **Level of Measurements**
  - Can the values be out of order?
  - If yes, can you meaningfully subtract the values from each other?
    - No, then it’s Ordinal.
    - If the values have labels, the answer is probably No.
    - If the values are labeled with ranges, it’s definitely No.
  - If Yes (values can be out of order AND subtracted), it’s Interval.

- **Univariate Analysis**
  - Central tendency (mean, median, mode)
  - Dispersion (standard deviation, range/IQR, variation ratio)

Four Basic Statistical Tests

- **Common pattern:**
  - All are based on some sort of standard error
  - All generate a probability of getting a difference larger than the one observed
    - Test for significance: difference between 2 numbers (usually an observed stat & some expected value), divided by some standard error
  - All follow the same 5 steps, and result in a p-value

- **Four main tests in Sociology:**
  - "T-test" for 1 interval and 1 categorical variable
  - ANOVA ("F test") for several categories & 1 interval
  - Regression ("r correlation") for several interval variables
  - "Chi-square" for 2 categorical (ordinal or nominal) variables
1. A relationship exists when...

- A. The DV changes across the IV
- B. The IV changes across the DV
- C. The CV changes across the IV
- D. The IV changes across the CV
- E. The DV changes across the CV

**Relationships**

Definition = distribution of DV varies by IV
- I.e. conditional distributions differ
- Could look at different DVs:
  - e.g. accounting for attitudes about abortion
- Could look at different IVs:
  - What accounts for partyid best (polyviews, educ, relig, or income)?

2. Which is NOT one of the three things you should do in any crosstab?

- A. Compare % across
- B. Compare % up/down
- C. Look for modal %s
- D. Look for main diagonal
- E. Look for chi-square’s p-value

**Chi-square tests...**

A. Association 21%
B. Dependence 47%
C. Prowess 3%
D. Robustness 3%
E. Strength 26%

**Problems w/ Chi-square**

- Varies w/ size of sample & of crosstab
  - Sometimes use alternative chi-square-based measurements
- Only significance, not strength
  - Can eyeball (compare across, modal %s)
  - Or can use other stats: PREs
Chi-square-based measures

- $\chi^2$, varies with df of $(r-1) \times (c-1) \sum \frac{(f_i - f_{oi})^2}{f_{oi}}$
- $\chi^2$ also varies with sample size, so Phi = $\sqrt{\frac{\chi^2}{n}}$
- Pearson’s contingency coefficient C = $\sqrt{\frac{\chi^2}{\chi^2 + n}}$
- And, it varies w/ table size, so Cramer’s V = $\frac{\chi^2}{\min(r, c) - 1}$

Chi-square = only part of picture

- All questions ask: Is there a relationship?
  - Compare Column%’s across; Modal %’s; Chi-square’s p
- But three key questions for crosstabs
  - Dependent relationship? (chi-square)
  - Size/Strength? (% comparisons, plus PREs)
  - Direction? (% comparisons, or gamma...)
- Three key questions for intervals, too (corr/reg)
  - Strength? – weak or strong?
  - Form? – linear? Curvilinear?
  - Direction? – positive? negative?
  - Remember valence!!

Distinguishing PREs

- Not same as strength, or significance
- Not necessarily a “measure of association”... but some overlap
  - Not all MOAs have a PRE interpretation
  - Some other statistics ($r^2$) do have a PRE interpretation

Proportional Reduction in Error (PRE)

- By what percentage do we reduce our error in predicting values of the dependent variable by knowing values of the independent variable?
- Imagine that I want to guess which students hadn’t finished an exam.
  - By how much is my guesswork improved if I know how early or late you submit homework?
  - If I know you’re usually early, I might guess that you’re already done with the exam.

PRE Formula

- All PREs follow same basic formula
  
  $\text{PRE} = \frac{(E1 - E2)}{E1}$

  where $E1$ is the number of errors when nothing is known about the independent, and $E2$ is the number of errors when the value of the dependent variable is known
All PREs have a p value

- Most have a distribution that matters; others are involved in a test statistic
- All p’s are interpreted basically the same:
  - The null = that there’s no relationship – e.g. the statistic, such as gamma, is zero in the population
  - P = probability we’d be wrong if we rejected the null
  - ~ risk in saying the statistic is what SPSS says it is
  - We want that risk to be lower than 5% (.05) so we can reject the null & find support for our claim
  - Remember: p never is zero, though might be close
  - If SPSS says .000, it’s <.001; if it says .00000, it’s <.00001

4. A p value is the probability of being wrong if we...

A. Reject the null 90%
B. Accept the null 8%
C. Reject the alternative 0%
D. Accept the alternative 3%
E. Give up 0%

Nominal Variables: Lambda

- Based on ability to guess values on one of the variables
  - PRE achieved through knowledge of values on the other variable
  - Represents reduction in errors as a proportion of the errors that would have been made on the basis of the overall distribution.
- Example: if we make 600 fewer errors predicting employment status, for a sample of 900, when we know the gender, lambda = 600/900 = .67
  - Measures the statistical association between gender and employment status. It indicates a strong relationship.
- Rough guideline: above 0.5 is strong? below is weak?
- Always positive; can’t be negative

Ordinal variables: Gamma

- PRE (guess values on one variable w/ values of another) BUT considers order of values
- So cannot use with nominals – i.e. if either is nominal
- Example: Religiosity and political conservatism.
  - If we know there is a strong relation between the two & they’re each ordinal(v. strong, strong, neutral, weak, v. weak), then
  - Knowing that someone is very strong in political conservatism would lead us to guess they’re very strong in religiosity.
- Guesses that a higher value on one variable connects w/ higher value on the other, for every possible pair of cases
- Allows us to guess direction as well as magnitude ("strength") of an association

The Idea Behind "Gamma" (γ)

- Compares most pairs of cases in the sample
- Focuses on those concordant w/ relationship
  - (those that show that higher values on one variable are associated with higher values on the other one)
- and those discordant w/ a relationship
  - (the reverse of or at discord with a positive relationship – a higher value on one variable but lower on the other)
- Gamma = (C – D) / (C + D)
  - the difference, as a fraction of the combination

C vs. D vs. “ties”

- Concordant pair:

<table>
<thead>
<tr>
<th></th>
<th>John</th>
<th>Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>Higher</td>
<td>Lower</td>
</tr>
<tr>
<td>Pay</td>
<td>Higher</td>
<td>Lower</td>
</tr>
</tbody>
</table>

- Discordant pair:

<table>
<thead>
<tr>
<th></th>
<th>John</th>
<th>Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>Higher</td>
<td>Lower</td>
</tr>
<tr>
<td>Pay</td>
<td>Lower</td>
<td>Higher</td>
</tr>
</tbody>
</table>

- Tied pairs:

<table>
<thead>
<tr>
<th></th>
<th>J</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educ</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Pay</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
Interpreting Gamma 1: Direction
- If gamma > 0, relationship is positive
  - predict that a higher value on one probably does mean a higher value on the other
- If gamma < 0, relationship is negative
  - predict that a higher value on one probably means a lower value on the other
- In above example:
  - 0.615 -> positive relationship

Disadvantage of Gamma
- Does not take into account tied pairs
- Therefore overstates strength
  - Possible for gamma to suggest a perfect positive association (+1) with only one concordant pair:
    \[
    (C-D)/(C+D) = (1-0)/(1+0) = 1
    \]
  - Somer’s D takes ties into account
    \[
    = (C-D) / (all tied pairs)
    \]

Interpreting Gamma 2: Strength
- Weak relationship if 0 to .3 (or 0 to -.3)
- Moderate if 0.3 to .6 (or -.3 to -.6)
- Strong if 0.6 to 1 (or -.6 to -1)

A Cautionary Note...
- Gamma assumes a higher value means a higher value label
  - E.g. 1 = low ses, 2 = middle, 3 = high ses
- If your variable is the reverse, you’ll need to either recode the variable to reverse the values or interpret gamma upside-down

Goodman & Kruskall’s tau b
- Like gamma, tau b is asymmetrical
  - Infers a ‘direction’ to the relationship
  - \( \chi^2 \) is symmetrical – does not distinguish between DV and IV (though column percents do)
  - (plain) tau also symmetrical – use for 2 nominals
  - Use tau b for 2 ordinals
  - Like Lambda, not absolute; relative only
    - below .5ish weaker; above .5ish is stronger

5 key PRE Measures, esp 2...
- **Gamma** (8.5, p.223) – \( \frac{(C-D)}{(C+D)} \)
- **Somer’s D** (8.4, p.224)
  - based on row marginals
  - based on column marginals
- **Goodman & Kruskall’s tau b** (8.4, p.224)
  - based on row marginals
  - based on column marginals
- **Lambda**
  - nominal only, based on modal category
  - Must use if either variable is nominal
Measures to use...

- Chi-square always, to establish dependence
- Lambda if either variable is nominal and has more than 2 categories
  - Also interpret the p-value – can you reject the null (that the lambda is simply 0)?
- Gamma if each variable is either ordinal or dichotomous
  - Also interpret the p-value – can you reject the null (that the gamma is simply 0)?

SPSS work...

- Demo: Crosstab w Abortion Index
  - Two possible explanations: sex & polviews
  - Don't forget:
    - get cell count, column %, chi-square, & gamma or lambda
    - use Pearson chi-square, and get Asymp Sig (= p value)
  - Summary table @ board
- Lab: is the US in Decline?
  - Now you try, 😊
  - Handout and data are on the web

5. Which is the correct choice?

A. Gamma for 2 nominals
   5%

B. Lambda for 2 intervals
   5%

C. Lambda for 1 nominal & 1 ordinal
   72%

D. Gamma for 1 ordinal & 1 nominal
   8%

E. Chi-square for one nominal & 1 interval
   10%