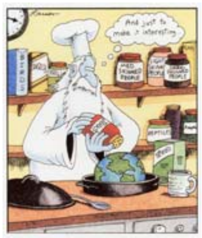
Proportional Reduction of Error



λ

γ

τ

- Why not χ^2 ?
- a) Doesn't tell much about strength or nature of relationship.
- b) Sample size influences value of χ^2 .
  - i.e. If you take a particular cross - section and multiply all cells by 10, you also increase the χ^2 value by 10 as the value of χ^2 depends on sample size as well as amount of departure from independence.

- c) Sum based an observed and expected frequencies there are many different tables may have same χ^2 value.
- Association for nominal variable

Phi

- Phi
- - measures based on χ^2

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

max value depends on size of table
if r > 2 and c > 2 . Φ can be > 1.0.

Coefficient Of Contingency

- Always < 1.0 but never = 1
- Always 0 - 1.0
- Largest value depends on number of rows and columns

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

Cramer's V

- It can attain a value of 1

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

or

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

Example: Canadian firms

		Firm type		Total
		domestic	foreign	
level of ownership	widely held	197221	44579	241800
	effective control	87984	15843	103827
	legal control	84414	60641	145055
Total		369619	121063	490682

χ^2=32920

$$\phi = \sqrt{\frac{32920}{490682}} = 0.259$$

$$c = \sqrt{\frac{32920}{32920 + 490682}} = 0.251$$

- Note that although values of Φ and C aren't equal, they are of the same magnitude but 'they are not particularly large' is not a very satisfactory way to have an interpretation.
- Alternatives to χ^2 are based on the idea of proportional reduction in error or PRE.
- They have a clean interpretation based on how well you can predict the value of dependent variables if you know the value of the independent.

lambda

- Its main advantage relates to its asymmetrical nature.
  - Contrary to other tests, the way variables are paired is of utmost importance; rows and columns are not interchangeable.
- Another advantage is the absence of constraints on the distribution of the variables

- Lambda (λ)
  - Very common measure of association.
  - Compares two types of errors
    - Error made while ignoring the independent variable (E1).
    - Errors made taking into account the independent variable (E2).

Calculating Lambda

- **Assuming the rows represent the dependent variable**
- Three Steps
  - Find (E1)
    - Subtract the largest row total from N.
  - Find (E2)
    - For each column, subtract the largest cell frequency from the column total and then add the subtotals together.
  - Calculate Lambda
    - Subtract (E1 - E2) and then divide by E1.

Example

		firm type		Total
		domestic	foreign	
level of ownership	widely held	197221	44579	241800
	effective control	87984	15843	103827
	legal control	84414	60641	145055
Total		369619	121063	490682

- E1 = 490682 – 241800 = 248882
- E2 = (369619 – 197221) = 172398

  \+

  (121063 – 60641) = 60422

  = 232820

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

$$\lambda = \frac{248882 - 232820}{248882}$$

$$\lambda = 0.065$$

- The λ tells you the proportion by which you reduce your error in predicting the dependent variable if you know the independent that's why its called a *Proportional Reduction In Error* measure.
- The largest the value can be is 1.
  - When variables are independent, λ = 0.
  - λ is not symmetric its value depends on which is the independent variable.

- Suppose we take the column as dependent

$$\lambda = \frac{121063 - 121063}{121063}$$

$$\lambda = 0$$

Symmetric λ

- If you have no reason to pick one as dependent or independent, use symmetric λ .
- Symmetric =Σ of 2 differences/Σ of denominator
- Example

$$\lambda = \frac{16062}{369945} = 0.043$$

Limitations of Lambda

- Lambda is asymmetric
  - Different values depending on which variable is the independent.
- Lambda can be misleading when one of the row totals is larger than the other.
  - It may be preferable to use a chi-square based measure when the rows are very unequal.

Example 2

Die/ Value	1	2	3	4	5	6	Total
Blue	10	15	30	30	15	20	120
Red	12	20	35	30	3	20	120
Green	17	17	21	20	26	19	120
Total	39	52	86	80	44	59	360

- E1= 240
  - All rows have an equal likelihood so you can take your choice

- E2 = 212
  - For value 1 39-17=22
  - For value 2 52-20=32
  - For value 3 86-35=51
  - For value 4 80-30=50
  - For value 5 44-26=18
  - For value 5 59-20=39

$$\lambda = \frac{240-212}{240} \approx .17$$

What do we conclude about the fairness of the dice?

Measures Of Association For Ordinal Variables

- Many measures are based on comparing pairs of case.
  - Using the classes of variable as 1 : high, 2 : medium, 3 = low

Example

City	Pop (000s)	Rank	Class	Retirees (000s)	Class
City A	672	7	3	3.3	3
City B	956	5	2	11.7	2
City C	5775	1	1	175.0	1
City D	3269	2	1	18.4	2
City E	795	6	3	11.0	2
City F	969	4	2	5.6	3
City G	1942	3	2	22.0	1

Cross tabulation form

	Retirees class		
Pop class	1	2	3
1	1	1	0
2	1	1	1
3	0	1	1

- A pair of cases is *concordant* if the value of each variable is larger (or smaller) for one case than for the other case.
- p is the number of concordant pairs
- They are *discordant* if the value of one variable for a case is larger than the value for the other case.
- q is the number of discordant pairs
- When 2 cases have identical values, they are *tied* on any one of the values

Goodman & Kruskal's Gamma

- A positive gamma say there are more
- like pairs than unlike pairs.
- The absolute value of gamma is the proportional reduction of error when using knowledge of concordance rather than a random choice.
- If variables are independent, gamma = 0; but if it equals 0, it does not necessarily mean independence.

$$\gamma = \frac{(P-Q)}{(P+Q)} = \frac{(9-2)}{11} = 0.636$$

- Let x be an independent variable with three values and let y be a dependent with two values, with a, b, ..., f being the cell counts in the resulting table

		X		
		1	2	3
Y	1	a	b	c
	2	d	e	f

Example for 2 by 3 table

Type of pair	Number of pairs	Symbol
Concordant	a(e+f)+b(f)	P
Disconcordant	c(d+e)+b(d)	Q
Tied on x	ad+be+cf	T_x
Tied on y	a(b+c)+bc+d(e+f)+ef	T_y

Kendall's tau - b

$$\mathrm{T}_b = \frac{P-Q}{\sqrt{(P+Q+T_x)(P+Q+T_y)}}$$

Where : T_x is the number of ties involving only the first variable
T_y is the number of ties involving only the second variable

$$T_b = \frac{9-2}{\sqrt{(9+2+5)(9+2+5)}} = 0.437$$

- No simple explanation in terms of proportional reduction of error.
- The statistics are more easily calculated if you lay them out in table like that below. Each pair of rows is only compared once. The comparison results in 1 of 3 outcomes; concordant (denoted P), disconcordant (denoted Q), or tied (where at least one set of ranks are tied). If the rows are tied on the X variable its entered as T and T_X, if its tied on variable Y its entered as T and T_Y.

	P	Q	T	T_x	T_y
1, 2	X				
1, 3	X				
1, 4	X				
1, 5			X	X	
1, 6			X		X
1, 7	X				
2, 3	X		X		X
2, 4			X		X
2, 5			X		X
2, 6			X	X	
2, 7			X	X	
3, 4			X	X	
3, 5	X				
3, 6	X				
3, 7			X		X
4, 5			X		X
4, 6	X				
4, 7		X			
5, 6		X			
5, 7	X		X	X	
6, 7	X		X	X	
	9	2	10	5	5

Tau - C

- Where : m is the smaller number of rows and columns
  - Note : There is no simple proportional reduction of error interpretation.

$$T_c = \frac{2m(P-Q)}{(P-Q)^2(m-1)}$$

$$T_c = \frac{2(3)(9-2)}{7^2(3-1)} = \frac{6(7)}{49(2)} = 0.438$$

m=3 because the cross tabulation table is a 3 by 3 table
(3 classes of population and 3 classes of retirees)

Somer's d

- gamma, T_b, T_c are all symmetric measures
- same as gamma except the denominator is sum of all pairs if cases that are not tied on independent variables.
- i.e.

$$d = \frac{(P-Q)}{P+Q+(pick\ T_x, T_y)} \quad d = \frac{(9-2)}{(9+2+5)} = 0.437$$