| 🏠 | **Syllabus** | **Statnotes** | **Websites** | **Lab** | **Instructor** | 🏠 |

Nominal-by-Interval Association
Eta, the Correlation Ratio

Eta is a coefficient of nonlinear association. For linear relationships, eta equals the correlation coefficient (Pearson's r). For nonlinear relationships it is greater -- hence the difference between eta and r is a measure of the extent of nonlinearity of relationship. Just as r^2 is interpreted as the percent of variance in one variable explained linearly by the other, eta^2 is the percent of variance in the dependent variable explained linearly or nonlinearly by the independent variable. This interpretation requires that the dependent variable be interval in level, and the independent variable be categorical (nominal, ordinal, or grouped interval).

In SPSS, select Analyze, compare Means, Means; click Options; select Anova table and eta. Alternatively, select Analyze, Crosstabs; click Statistics, select eta.

Key Concepts and Terms

- **Relationship to ANOVA**. Eta is a measure of strength of relationship based on sums of squares computed in analysis of variance. However, eta may be a useful coefficient outside the context of ANOVA. Eta equals the square root of the sum of squares for an interval variable y between classes (columns) divided by the total sum of squares. The numerator and denominator in this formula have meanings as in ANOVA, and to the extent that x and y are linearly or nonlinearly related, the numerator will be as large as the denominator and eta will approach 1.0.

- **Computation**. Eta^2 is computed as between-groups sum of squares divided by total sum of squares (see the section on ANOVA). In the example in the table below, data are given for two variables, x and y. Note that eta requires larger frequencies than are contained in this table, but small values are used for simplicity of computational example. The raw x-y data are shown in the upper left corner of this figure. The x data are grouped into ranges to increase the frequency count for each of the y classes. After grouping, x ranges from 0 to 9. Visual inspection of the table of joint frequencies of x and y shows a curvilinear relationship.

Eta Computing Chart

Joint frequencies table and computing chart (columns A–I):

x	y
11	14
24	48
31	35
42	39
81	46
74	41
66	37
98	09
50	51
25	32
06	04
51	61
86	15
90	14
92	25
38	34
61	55
28	27
15	04
50	44

Joint frequencies — x range:

y range	y'	0-9 (0)	10-9 (1)	20-9 (2)	30-9 (3)	40-9 (4)	50-9 (5)	60-9 (6)	70-9 (7)	80-9 (8)	90-9 (9)	f_y
60-69	6						1					1
50-59	5						1	1				2
40-49	4						1		1	1		4
30-39	3				2	1		1				5
20-29	2			1						1		1
10-19	1		1							1	2	4
0-9	0	1	1								1	3
f_x		1	2	3	2	1	3	2	1	2	3	
$\Sigma y'$		0	1	9	6	3	15	8	4	5	2	

Computing chart:

A f_y	B y'	C $f_y(y')$	D $f_y(y')^2$	E x'	F f_x	G $\Sigma y'$	H $(\Sigma y')^2$	I $(\Sigma y')^2/f_x$
1	6	6	36	0	1	0	0	0
2	5	10	50	1	2	1	1	.5
4	4	16	64	2	3	9	81	27
5	3	15	45	3	2	6	36	18
1	2	2	4	4	1	3	9	9
4	1	4	4	5	3	15	225	75
3	0	0	0	6	2	8	64	36
				7	1	4	16	18
				8	2	5	25	12.5
				9	3	2	4	1.3
$n = 20$		$\Sigma f_y(y') = 53$	$\Sigma f_y(y')^2 = 203$		$n = 20$	$\Sigma(\Sigma y') = 53$		$\Sigma[(\Sigma y')^2/f_x] = 195.3$

$$Eta^2_{yx} = \frac{\text{Between SS}}{\text{Total SS}}$$

$$Eta^2_{yx} = 44.8/200 = .224$$

$$Eta_{yx} = \sqrt{.224} = .47$$

where between groups $SS = \sum\left[\frac{(\Sigma y')^2}{f_x}\right] - \frac{[(\Sigma\Sigma y')]^2}{n}$
$= 195.3 - 53^2/20 \approx 44.8$

total sum of squares $= \Sigma f_y(y')^2 - [\Sigma f_y(y')]^2/n$
$= 203 - 53^2/20 \approx 200$

Below the table of joint frequencies is a manual computing chart with nine columns:

1. Column A if f_y. It lists f_y, which is the number of units in each range of y.
2. Column B is lists y', which is the grouped values of y, from 0 to the number of ranges minus 1 (that is, to 6).
3. Column C $f_y(y')$. It is obtained by multiplying column B by column A.
4. Column D is $f_y(y')^2$. It is obtained by multiplying column C by column B.
5. Column E lists x', which is grouped values of x, from 0 to the number of ranges minus 1 (that is, to 9).
6. Column F lists f_x, this is the number of units in each range of x.
7. Column G is SUMy'. It lists, for each value of x' (0 through 9), the sum of y'. The sum of y' equals the sum of y' values times the corresponding number of units in that class (column) of x'. For instance, this sum for column 9 equals $2(1) + 1(0) = 2$.
8. Column H is $(\text{SUM}y')^2$. It is obtained by squaring the values in column G.
9. Column I is $(\text{SUM}y')^2/f_x$. It is obtained by dividing the values in column H by the values in column F.

Eta-squared is computed by using the sums of columns A, C, D, G, and I in the appropriate places in the formulas shown in the figure above. In this example, eta$_{yx}$ is .47, which compares with a Pearson's r correlation of .11 for the grouped data in this example. Squaring, we can say that the linear relationship (reflected in r^2) accounts for about 1.2% of the variance, whereas the nonlinear relationship (reflected in eta^2) accounts for 22.4% of the variance.

Assumptions

- **Meaning of association:** Eta defines "perfect relationship" as curvilinear and defines "null relationship" as statistical independence, as described in the section on [association](#). Note that by defining perfect association as curvilinear, eta is not sensitive to the order of the classes of the categorical variable. Therefore an equally high (not perfect) eta would result from both patterns below.
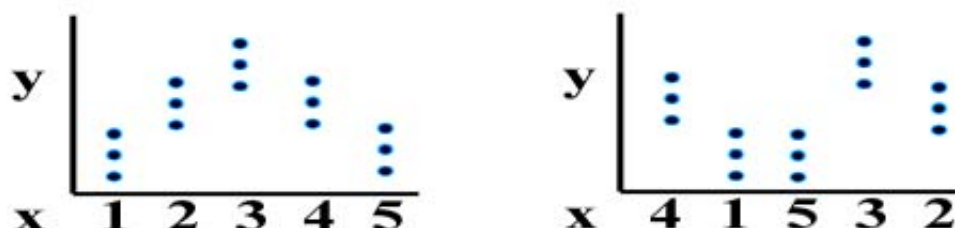
- **Asymmetry**. Eta is asymmetric: unlike Pearson's correlation (r), with the correlation ratio (eta) one will get a different value for the coefficient depending on which is the independent and which is the dependent variable. Eta^2 is interpreted as the percent of variance in the dependent variable explained by the independent variable. One can compute eta both ways, with each variable considered as dependent in turn, then average the two coefficients to obtain a "symmetric eta" coefficient, but this coefficient does not have a clear "variance explained" interpretation. Like other forms of correlation and association, eta cannot prove causal direction, only measure its level given the researcher's assumption of causal direction. For this reason, eta has no sign and varies from 0 to 1.0.

- **Interval dependent**. One variable must be interval or ratio in level. Typically, this is the dependent variable, particularly when one is giving a "variance explained" interpretation to eta. However, eta can be computed with either variable considered as dependent.

- **Categorical variable with several categories**. A second variable must be categorical, and its categories may be arranged in any order. That is, it may be of any data level, including nominal. There must be enough classes of the categorical variable to give the grand mean of column means stability. Typically, the categorical variable is the independent variable, particularly when one is giving a "variance explained" interpretation to eta.

- **Cell count not small**. The frequencies of each class of the categorical variable must be large enough to give statistical stability to the means of these classes. This often requires that the interval variable be grouped into ranges so as to assure that there are an adequate number of categorical y values corresponding to each interval x value. When the x variable is ranged, its crosstabulation with the categorical y variable is called the *table of joint frequencies*.

Bibliography

- Siegel, Sidney (1956). *Nonparametric Statistics For The Behavioral Sciences*. NY: McGraw-Hill.

Back