

Lecture 22: Correlation

Dewayne E Perry

ENS 623

Perry@ece.utexas.edu

Pearson r

- Most widely used index of relationship
- Short for: *Karl Pearson's product moment correlation coefficient*
- Values ranges between -1.00 and +1.00
 - ↳ .00 means there is no relationship
 - ↳ +1.00 - a perfect positive linear relationship
 - ↳ -1.00 - a perfect negative linear relationship
- May be correlated even though scores do not agree

Pearson r

→ Examples

$$\Rightarrow (8, 6, 4, 2) \text{ and } (16, 12, 8, 4) \quad r = 1.00$$

$$\Rightarrow (8, 6, 4, 2) \text{ and } (6, 4, 4, 6) \quad r = .00$$

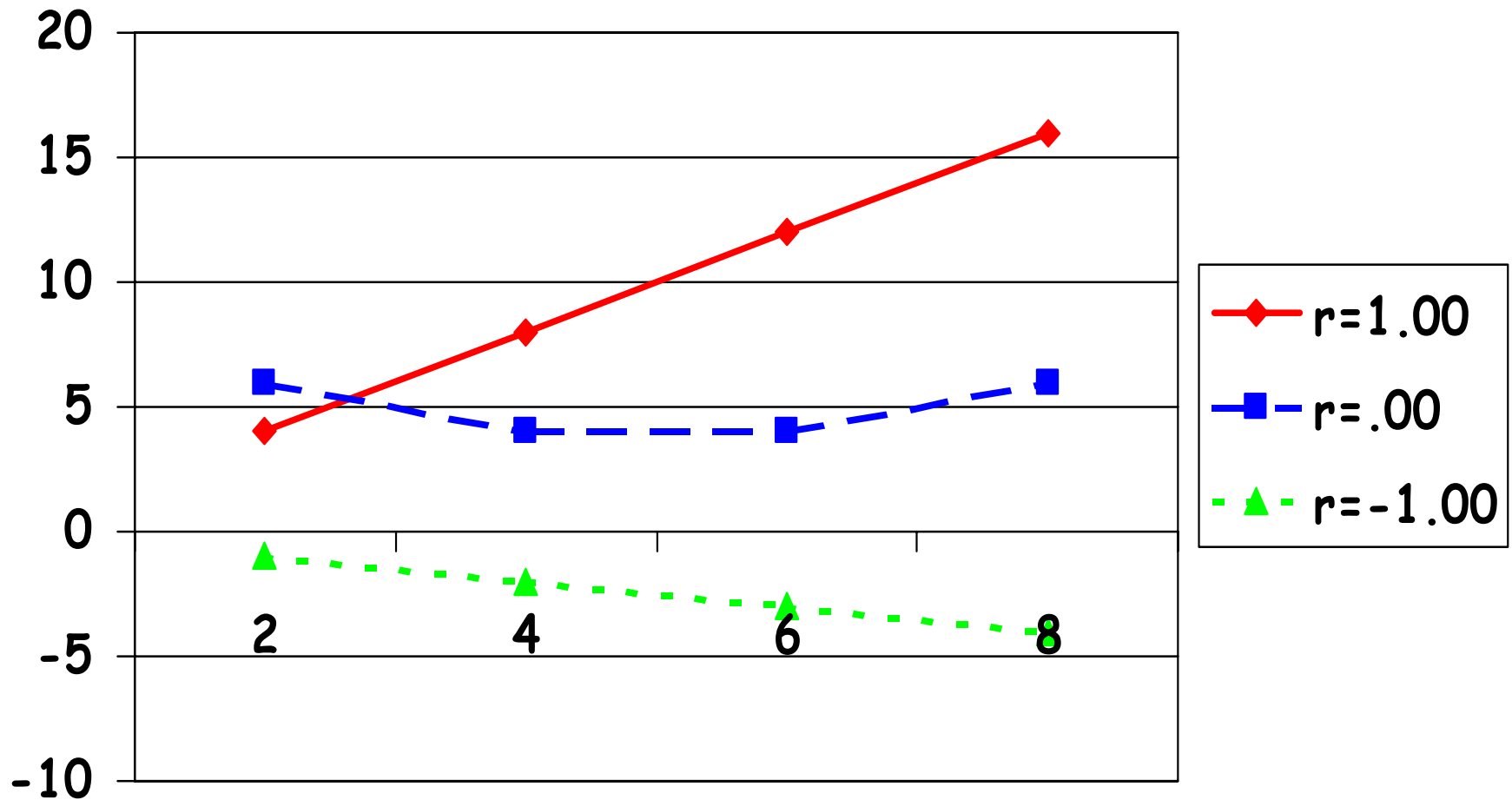
$$\Rightarrow (8, 6, 4, 2) \text{ and } (-4, -3, -2, -1) \quad r = -1.00$$

→ Results are what one would expect if standard scored (Z-scored): $Z = X - \bar{X} / \sigma$

$$\Rightarrow \text{product moment correlation: } r_{xy} = \sum Z_x Z_y / N$$

- Z's are distances from mean called *moments*
- multiplied by each other to form *products*

Pearson r Correlations



Pearson r

→ An easier formula that computes r from the raw data

$$r_{xy} = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{\left(\left[N \sum X^2 - (\sum X)^2 \right] \left[N \sum Y^2 - (\sum Y)^2 \right] \right)}}$$

$$r_{xy} = \frac{4(240) - (20)(40)}{\sqrt{\left(\left[4(120) - (20)^2 \right] \left[4(480) - (40)^2 \right] \right)}} = \frac{160}{\sqrt{(80)(320)}} = \frac{160}{160} = 1.00$$

Interpretations

- Prime interpretation: the larger r , the higher the degree of linear relationship
- The square of r : the proportion of the variance shared by X and Y
 - ↪ Proportion of variance of Y scores attributable to variation in the X scores
 - ↪ $r^2 + k^2 = 1.00$
 - ↪ r^2 is the *coefficient of determination*
 - ↪ k^2 is the *coefficient of non-determinism*
 - ↪ Though useful, it is a poor reflection of the practical value of any given correlation
 - ↪ More useful in regression (discussed later)

Interpretations

→ r as an indicator of practical importance

↪ *Binomial effect-size display (BESD) procedure*

↪ *Binomial* : research results cast as dichotomous

↪ Introduced because

- Interpretation is quite transparent
- Applicable whenever r is used
- Very conveniently computed

↪ BESD question: what is the effect on the *success rate* of the new treatment

- Displays the change attributable to treatment
- Converts effect size r into a success rate via table lookup (RR Table 14.6)

✓ $r=.30$, accounting for 9% of the variance

✓ shows an increase in the *success rate* from 35% to 65%

- Short form: $r \times 100 =$ percentage increase of success
- [Insight based on 50-50 probability of treatment effect]

↪ More clearly shows real-world importance of treatment than effect size estimates

Small but Important

→ While effect size may be small, the practical importance may be large

↳ May have important social, psychological or biological effects

→ Another way to compute r (or ϕ)

$$r = \frac{\text{difference between cross products}}{\sqrt{\text{product of all marginal totals}}}$$

→ Examples

↳ Vietnam versus non-Vietnam veterans, 50% more likely to have an alcohol problem, $r = .0698$

↳ Vietnam veterans about twice as likely to suffer depression as non-Vietnam, $r = .0597$

→ Small effects. But can reflect effects of enormous consequence

↳ Aspirin and heart attacks: $r = .0337$

↳ But this translates in a significant number of lives

Spearman Rank Correlation

→ ρ sometimes used as a quick index of correlation

↪ Easy and painless to compute

↪ Consider the following example (D is difference in rank)

X	Y	rank X	rank y	D	D-squared
6.8	79.713	2	1	1	1
12.2	47.691	1	2	-1	1
1.7	28.002	3	3	0	0
0.3	11.778	4	4	0	0

$$\rho = 1 - \frac{6(2)/4^3 - 4}{60} = 1 - \frac{12}{60} = .80$$

→ Nothing sacrosanct in scale used

↪ Reduces skewedness

↪ Choose for symmetry, lack of skewedness

↪ Tends to increase accuracy of analysis

➤ Sometimes leads to slightly higher r

➤ Sometimes to lower

✓ case of logarithmic transformations: .80 instead of .99 (RR 14.11)

Spearman Rank Correlation

- Most useful correlations are *product moment correlations*
- When data in rank form, apply *Spearman rho*
 - ↳ But nothing more than *Pearson r* computed on numbers that happen to be ranks
 - ↳ Ranks are more predictable
 - ↳ New ingredient: D - the difference between the ranks assigned to each pair of sampling units
- $$\rho = 1 - 6 \sum D^2 / N^3 - N$$

Point Biserial Correlation

→ Special case of product moment correlation r

↪ One variable continuous,

↪ One dichotomous,

➤ with arbitrarily applied numeric values

➤ Such as 0 and 1, or -1 and +1

→ Example: M vs F on verbal skills

↪ M=2,3,3,4 vs F=4,5,5,6

↪ X is implicit in M/F, Y is explicit

↪ Encode gender as 0,1

↪ Y mean = 4, X mean = 0.5

↪ X1 mean = 5, X2 mean = 3

Point Biserial Correlation

$$\rightarrow t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2_{\text{pooled}}\right)}} = \frac{5 - 3}{\sqrt{\left(\left(\frac{1}{4} + \frac{1}{4}\right) 0.6667\right)}} = 3.46$$

→ Which at 6 *df* is significant at $p < .01$, one-tailed test

$$\rightarrow r = .816$$

→ Significance test = size of effect X size of study

↳ Index for size of study varies with index of effect size:

➤ Eg, *N*, *df*, square root of *N* or *df*

↳ As either increases, significance test score increases



$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{df}$$

↳ First term is proportion of variance explained by *r* to the proportion not explained by 4 - ie signal to noise ratio

Phi Coefficient

→ Another special case of the product moment correlation r

↪ Both variables are dichotomous

↪ Arbitrarily applied numeric values 0,1 or +-1

→ Example - Dem/Rep answer Yes/No

↪ D: 1Y, 4N vs R: 4Y, 1N

↪ $r = .60$ for party membership and answer

↪ If sample size not too small ($N > 20$) and both variables are not too far from 50-50 split (no greater than 75/25), can use t test for significance

↪ $t = 2.12$ which is $p = .034$, one-tailed

↪ More common is chi-square test for significance of phi

$$\chi^2(1) = \phi^2 \times N$$

↪ since $phi = .60$ and $N = 10$, $chi-square = 3.60$

↪ which is significant at the .058 level

$$\chi^2$$

→ Three ways to compute χ^2

→ 1) $\chi^2(1) = \phi^2 \times N$

→ 2)
$$\chi^2(1) = \frac{N(BC - AD)^2}{(A + B)(C + D)(A + C)(B + D)}$$

	X1	X2
Y1	A	B
Y2	C	D

$$\chi^2$$

→ Expected frequency (E) vs observed (O):

↳ for each cell

↳ multiply total of the row by total of the column

↳ divide by the total number

$$\rightarrow 3) \chi^2(1) = \sum \frac{(O - E)^2}{E}$$

	X1	X2
Y1	E/O	E/O
Y2	E/O	E/O

$$\chi^2(1) = .202 + .735 + 2.028 + 7.2$$

$$\chi^2(1) = 10.165$$

	X1	X2
Y1	713/725	196/184
Y2	71/59	20/32

Equivalences

$$\rightarrow Z = \phi \times \sqrt{N}$$

$$\rightarrow \phi = Z / \sqrt{N}$$

$$\rightarrow \phi = \sqrt{\chi^2(1) / N}$$

Curvilinear Correlation

- Sometimes predictions are not linear but curvilinear (quadratic - U shaped)
 - ↳ higher (U shaped) or lower levels (upside down U) at ends
 - ↳ eg, extreme levels of arousal associated with great/poor performance

5 Product-Moment Correlations

→ Pearson r

- ↪ both variables continuous
- ↪ t test for significance

→ Spearman rho

- ↪ both variables ranked
- ↪ t test for significance
- ↪ or exact probability test if N is small ($N < 7$)

→ Point biserial ($r-pb$)

- ↪ one continuous, one dichotomous
- ↪ t test for significance

→ Phi

- ↪ both variables dichotomous
- ↪ chi -square, t and Z tests

→ Curvilinear r

- ↪ both continuous
- ↪ t test

Comparing Correlations

→ Primary question

- ↪ often not so much about relationship
- ↪ but about difference in such relationships
- ↪ comparison of independent correlation coefficients
 - based on different independent subjects
- ↪ comparison of non-independent correlation coefficients
 - based on the same subjects