Association and Correlation

Portland State University
USP 634 Data Analysis I
Spring 2017

Introduction

- Two variables are said to be associated when they vary together—that is, when one changes as the other changes.
- Association can be important evidence for causal relationships, particularly if the association is strong.

Introduction

- If variables are associated, the score (value) of one variable can be predicted from the score of the other variable.
- The stronger the association, the more accurate the predictions.
- The “predictor” is the *independent* variable
- The variable being “predicted” is the *dependent* variable

Association and bivariate tables

- Bivariate association can be investigated by finding answers to three questions:
  - Does an association exist?
  - How strong is the association?
  - What is the pattern and/or direction of the association?

Association and bivariate tables

- The table shows the relationship between authoritarianism of bosses (X) and the efficiency of workers (Y) for 44 workplaces.

	Low Authoritarian	High Authoritarian	TOTAL
Low Efficiency	10	12	22
High Efficiency	17	5	22
TOTAL	27	17	44

Is there an association?

- An association exists if the conditional distributions of one variable change across the values of the other variable.
- With bivariate tables, column percentages are the conditional distributions of Y for each value of X.
- If the column percentages change, the variables are associated.

Association and bivariate tables

- The column % is (cell frequency / column total) * 100.
  - (10/27)*100 = 37.04%
  - (12/17)* 100 = 70.59%
  - (17/27)*100 = 62.96%
  - (5/17)*100 = 29.41%

	Low authoritar.	High authoritar.	TOTAL
Low efficiency	10 (*37.04%*)	12 (*70.59%*)	22
High efficiency	17 (*62.96%*)	5 (*29.41%*)	22
TOTAL	27	17	44

Is there an association?

- The column %s show efficiency of workers (Y) by authoritarianism of supervisor (X).
- The column percentages change, so these variables are associated.

	Low	High
Low	37.04%	70.59%
High	62.96%	29.41%
	100%	100%

How strong is the association?

- The stronger the relationship, the greater the change in column %s (or conditional distributions).
  - In weak relationships, there is little or no change in column %s.
  - In strong relationships, there is marked change in column %s.

How strong is the association?

- One way to measure strength is to find the "maximum difference", the biggest difference in column percentages for **any row** of the table.

Difference	Strength
Between 0 and 10%	Weak
Between 10 and 30%	Moderate
Greater than 30%	Strong

How strong is the association?

- The maximum difference is 70.59 – 37.04 = 33.55.
- This is a strong relationship.

	Low	High
Low	37.04%	70.59%
High	62.96%	29.41%
	100%	100%

What is the pattern of the relationship?

- “Pattern” = which values of the variables go together?
- To detect, find the cell in each column which has the highest column percentage.

What is the pattern of the relationship?

- “Low” on authoritarianism goes with “High” on efficiency.
- “High” on authoritarianism goes with “Low” on efficiency.

	Low	High
Low	37.04 %	70.59 %
High	62.96 %	29.41 %
	100%	100%

What is the direction of the relationship?

- If *both* variables are ordinal, we can discuss *direction* as well as pattern.
- In *positive* relationships, the variables vary in the same direction.
  - As one increases, the other increases.
- In *negative* relationships, the variables vary in opposite directions.
  - As one increases, the other decreases.

What is the direction of the relationship?

- Relationship is *negative*.
- As authoritarianism increases, efficiency decreases.
- Workplaces high in authoritarianism are low on efficiency.

	Low	High
Low	37.04 %	70.59 %
High	62.96 %	29.41 %
	100%	100%

What is the direction of *this* relationship?

- This relationship is positive.
- Low on X is associated with low on Y.
- High on X is associated with high on Y.
- As X increases, Y increases.

	Low	High
Low	**60%**	**30%**
High	**40%**	**70%**
	100%	**100%**

Summary

- A strong, negative relationship between authoritarianism and efficiency.
- Consistent with the idea that authoritarian bosses cause inefficient workers (mean bosses make lazy workers).
- **But…**

	Low	High
Low	37.04%	70.59%
High	62.96%	29.41%
	100%	100%

Summary: Strength and direction

- …These results may also be consistent with the idea that inefficient workers *cause* authoritarian bosses (lazy workers make mean bosses).

	Low	High
Low	37.04%	70.59%
High	62.96%	29.41%
	100%	100%

Association vs. causation

- Association and causation are not the same things.
- Strong associations may be used as evidence of causal relationships **but** they do not prove variables are causally related.
- What else would we need to know to be sure there is a causal relationship between authoritarianism and efficiency?

NOMINAL MEASURES OF ASSOCIATION

Measures of association (MoAs)

- MoAs gauge strength of relationship (and do not address statistical significance).
- For nominal variables, MoAs are on 0 to 1 scale, where 0 is no relationship and 1 is strongest
- For ordinal and numeric variables, MoAs are on -1 to 1 scale,
  - where 0 is no relationship,
  - -1 is perfect negative relationship,
  - 1 perfect positive relationship

χ^2-based MoAs: Φ [phi]

- χ^2 is directly proportional to N, so can be normalized by dividing by N: $\phi = \sqrt{\chi^2/N}$
- Provides a measure of association ranging from 0 to 1 for 2x2 tables
- $\Phi = 1$ ➔ the case when the diagonally opposite cells are empty.
  - Problem with Φ is that when Table is bigger than 2x2, upper limit > 1. Difficult to interpret.

Φ vs. Cramer's V

- Cramer's V: Slightly modified Φ suitable for larger tables:
  - The upper limit of Φ is min (r-1,c-1), so divide by this term to get unity (to "normalize" to a maximum of 1).
  - Limitation: intermediate values somewhat hard to interpret because the formula is not linear. E.g., value of 0.5 not twice as strong as a value of 0.25.

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

Limitations of Φ and Cramer's V

- Φ is used for 2x2 tables only. For larger tables, use V.
- Φ and V are indexes of the *strength* of the relationship *only*. They do *not* identify the pattern.
- To analyze the pattern of the relationship, see the column percentages in the bivariate table.

Proportional reduction in error (PRE)

$$\frac{\text{(Error Rate Not Knowing)} - \text{(Error Rate Knowing)}}{\text{(Error Rate Not Knowing)}}$$

- Do your best to predict value of the dependent variable without knowing the independent variable; subtract correct predictions from total cases; this is E_1 (error rate not knowing)
- Do the same using information about the independent variable ("knowing")
- Apply the above formula

Lambda (λ)

- A way of implementing PRE using information from a table showing distribution of cases according to two variables
- $\lambda = (E_1 - E_2)/E_1$

$$\lambda = \frac{E_1 - E_2}{E_1}$$

  - E_1 = N – largest row total
  - E2 = For each column, subtract the largest cell frequency from the column total
- $0 \leq \lambda \leq 1$

Lambda (λ) – example

- How much does the road system "explain" urban form?
- Road system: independent variable; urban form: dependent variable
- Does knowledge of road system reduce prediction errors?
- $\lambda = (E_1 - E_2)/E_1$
- To compute λ, we must first find E_1 and E_2:
  - E_1 = N – largest row total = 100-60

		Road System		
		Radial	Grid	
Urban Form	Sector	15	45	60
	Concentric Circle	25	5	30
	Poly-centric	10	0	10
		50	50	100

Lambda (λ) – example

- E_2 = For each column, subtract the largest cell frequency from the col. total
- For radial road grid, assigning all 50 cases to concentric circle urban form yields error of 25;
- For grid, assigning all 50 cases to sector yields error of 5
- $E_2 = 25 + 5 = 30$

		Road System		
		Radial	Grid	
Urban Form	Sector	15	45	60
	Concentric Circle	25	5	30
	Poly-centric	10	0	10
		50	50	100

Lambda (λ) – example

- $\lambda = (E_1 - E_2)/E_1$
  $= (40\text{-}30)/40$
  $= 10 / 40$
  $= 25\%$
- …So knowing the road system lets us improve our prediction of urban form by 25 percent.

		Road System		
		Radial	Grid	
Urban Form	Sector	15	45	60
	Concentric Circle	25	5	30
	Poly-centric	10	0	10
		50	50	100

The limitations of lambda (λ)

- λ gives an indication of the *strength* of the relationship *only*.
  - It does *not* give all information about pattern.
- To analyze the *pattern* of the relationship, use the column percentages in the bivariate table.
- λ can be zero even when there is an association between the variables.

ORDINAL MEASURES OF ASSOCIATION

MoAs for Ordinal Variables

- Continuous ordinal variable (many possible values/scores):
  - Spearman's rho $$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

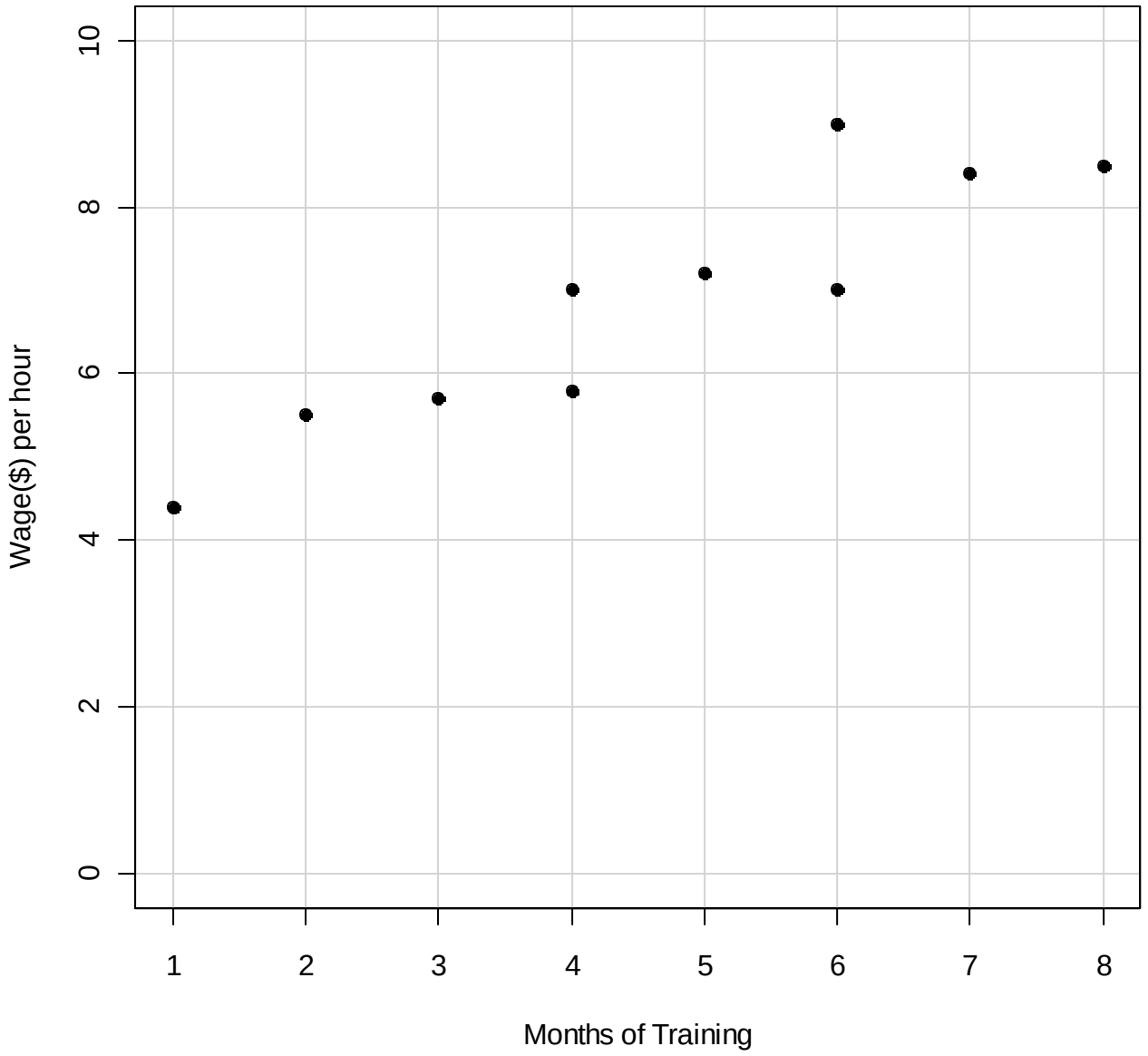
    where ΣD^2 = the sum of the differences in ranks, the quantity squared

- Collapsed ordinal variable (a few values or scores):
  - Gamma (PRE)

MEASURES OF ASSOCIATION
for Numeric Variables

Scatter plots

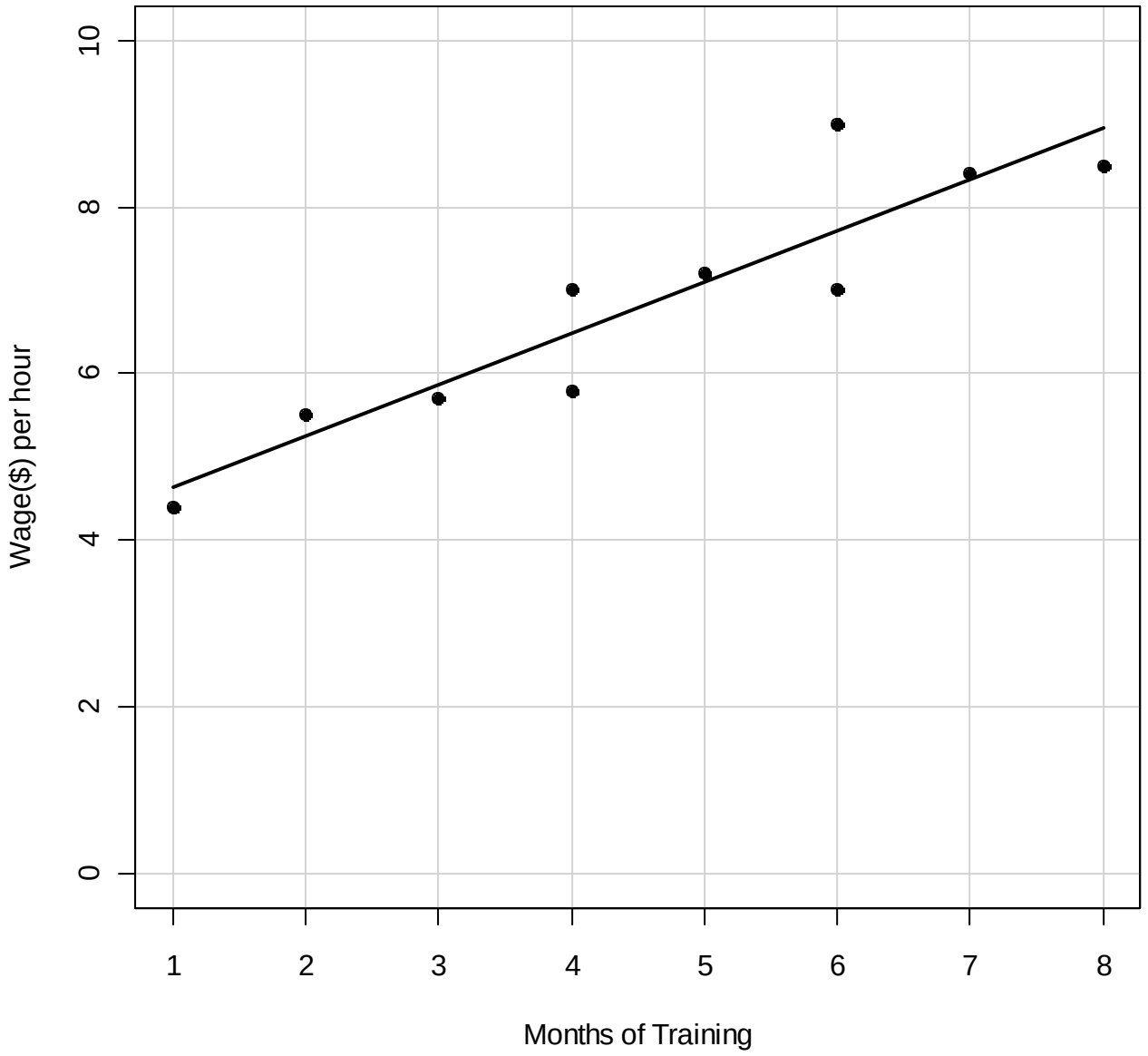
- Scatter plots have two dimensions:
  - The independent variable (X) is plotted along the horizontal axis (which is called “the X axis”).
  - The dependent variable (Y) is plotted along the vertical axis (which is called “the Y axis”).
- Each dot on a scatter plot is a case/an observation.
- The dot is placed at the intersection of the case’s scores on X and Y.



X	Y
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5

Scatter plot & regression line

- Inspection of the scatter plot should always be the first step in assessing the association between two numeric variables
- Regression line is a single straight line that comes “as close as possible” to all data points, which indicates **strength** and **direction** of the relationship



X	Y
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5

Regression line: Strength and direction

- Strength of association
  - The greater the extent to which dots are clustered around the regression line, the stronger the relationship
- Direction of association
  - Positive: regression line rises left to right.
  - Negative: regression line falls left to right.
- Slope of regression line
  - Steeper slope implies larger “effect”—but caution: this partly an artifact of variable *units* and outliers

How do we measure the association of X and Y?

- Use a calculated regression line, if linear relationship is appropriate
- Another way to measure the extent of clustering around the regression line is to using Pearson's r or R^2. These measures can be tested for statistical significance.

Pearson's r

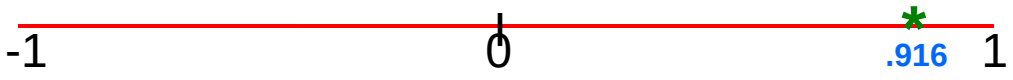
- AKA Pearson Product-Moment **Correlation**
- Pearson's r is a measure of association for numeric variables:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

- Ranges from -1 to 1:
  - 0 indicates no relationship,
  - -1 a perfect negative relationship
  - 1 a perfect positive relationship
- Limitation: No direct interpretation of intermediate values

Correlation: Pearson's r

- $r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{\sum X_i Y_i - N\bar{X}\bar{Y}}{\sqrt{\sum X_i^2 - N\bar{X}^2}\sqrt{\sum Y_i^2 - N\bar{Y}^2}}$
- $r = \frac{342.5 - 10 * 4.60 * 6.85}{\sqrt{256 - 10 * 4.60^2}\sqrt{489.4 - 10 * 6.85^2}} = 0.916$
- R code: `cor(X,Y)`

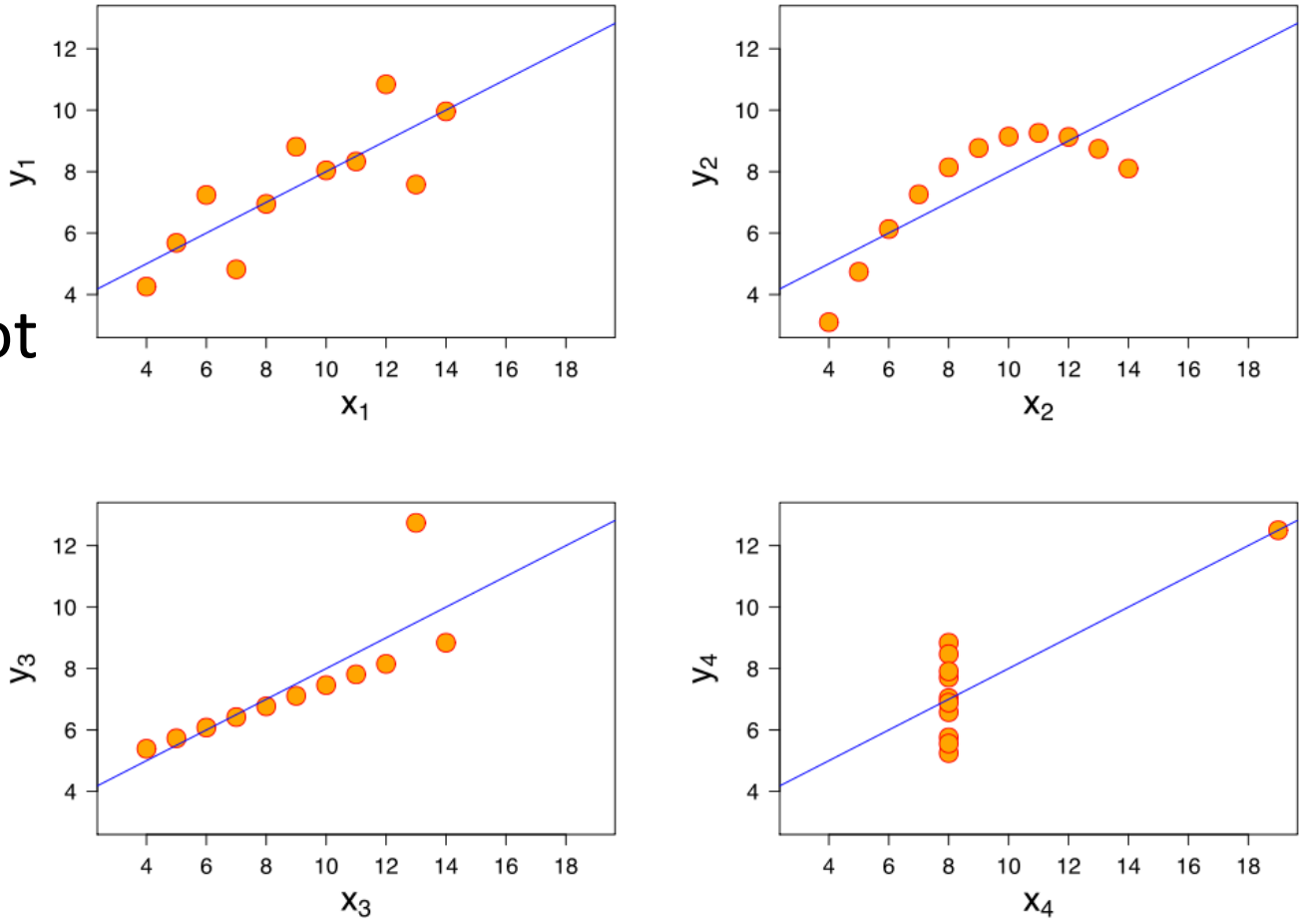


N = 10
ΣX_i = 46
ΣX_i^2 = 256
ΣY_i = 68.5
ΣY_i^2 = 489.4
$\Sigma X_i Y_i$ = 342.5
$\bar{X}$ = 4.60
$\bar{Y}$ = 6.85

X	Y
1	4.4
2	5.5
3	5.7
4	5.8
4	7
5	7.2
6	7
6	9
7	8.4
8	8.5

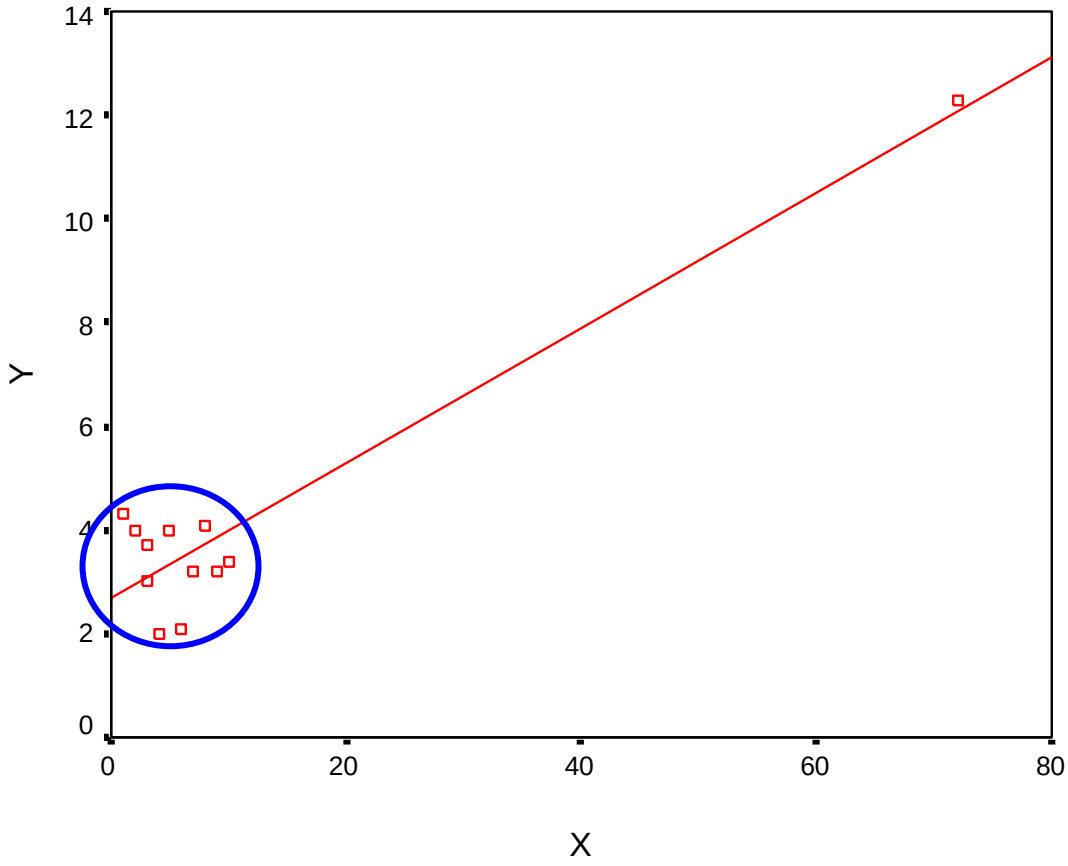
Correlation

- Assume a linear relationship
- Sensitive to outliers
- Always look at scatter plot, not just *r* statistic.



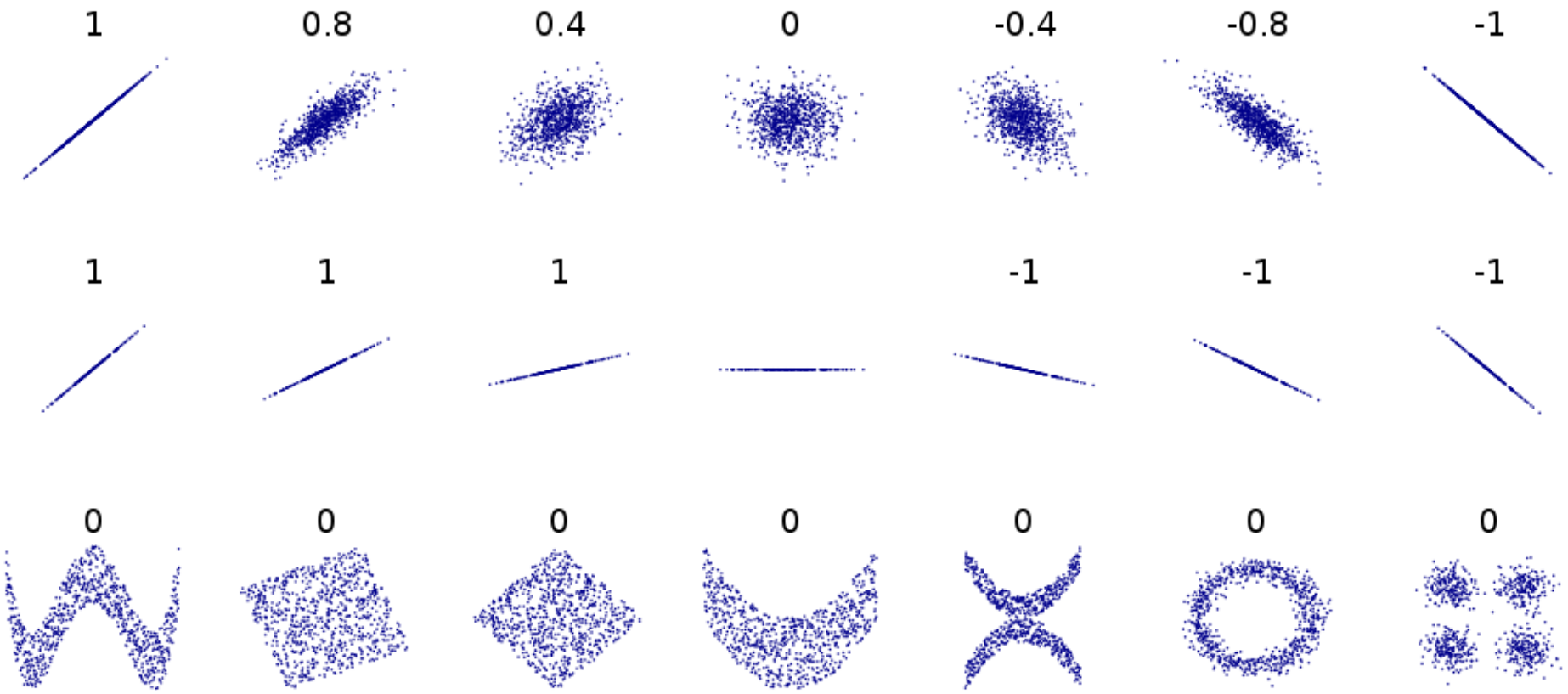
Four sets of data with the same correlation of 0.816

Correlation



Outlier: high r; but for most data points, no relationship.

Correlation



Subjective Strength

0 - .30	Weak
.30 - .60	Moderate
> .60	Strong

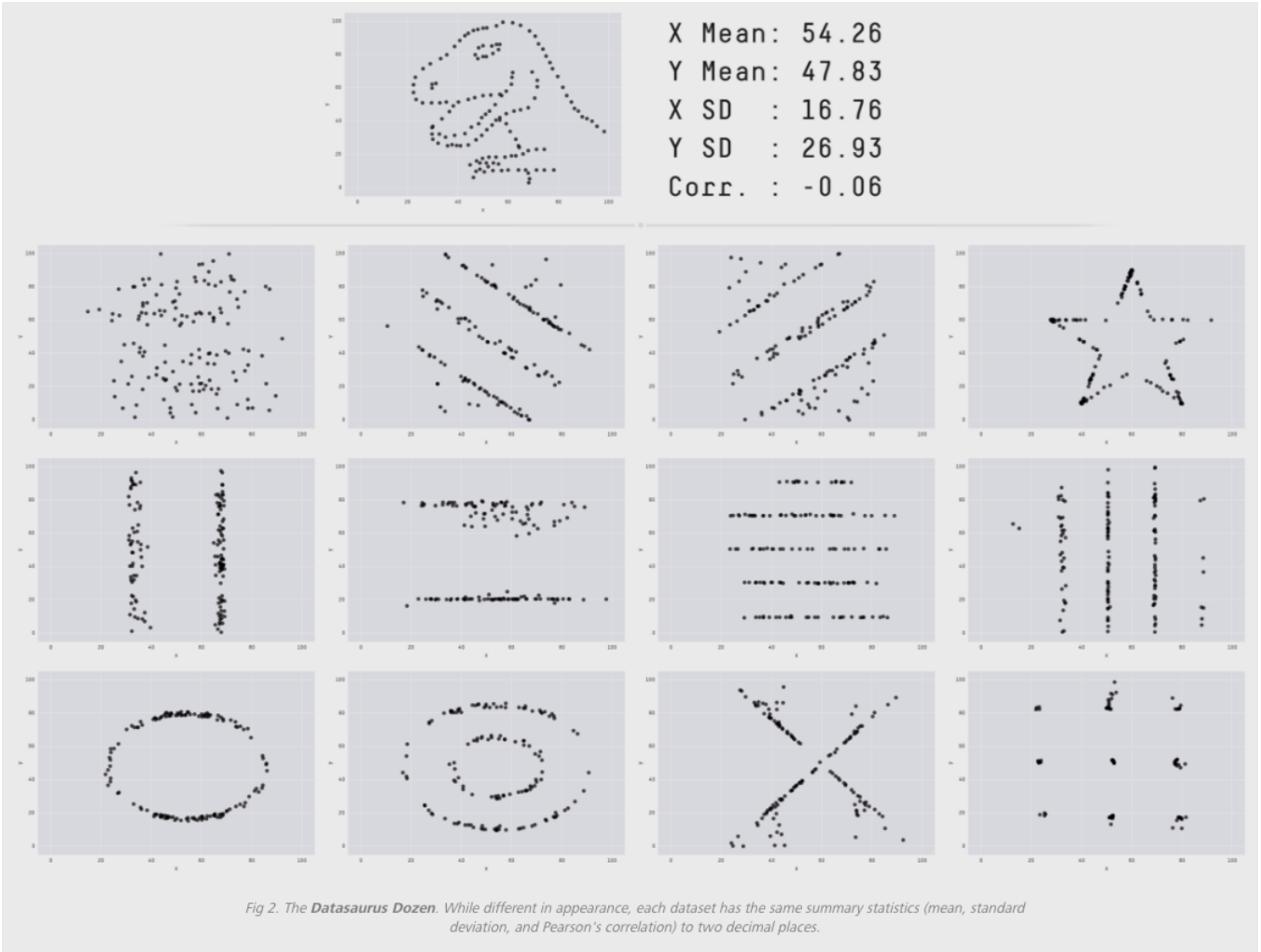
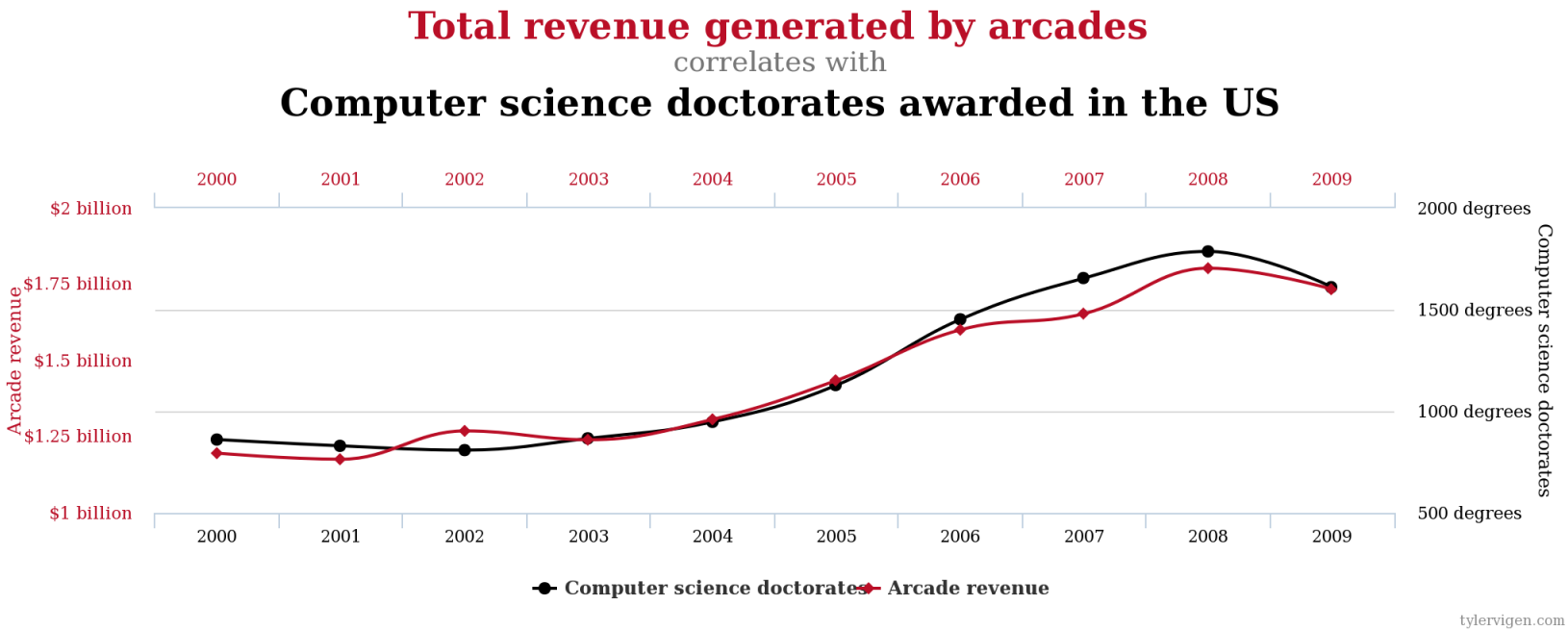


*Fig 2. The **Datasaurus Dozen**. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.*

Source: https://www.autodeskresearch.com/publications/samestats

Spurious Correlation



http://www.tylervigen.com/spurious-correlations

Coefficient of Determination (R^2)

- By squaring *r*, we obtain a PRE measure called the **coefficient of determination** (R^2)
- Can be interpreted as the proportion of variation in dependent variable (Y) explained by independent variable (X):

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

- *Our example*: $R^2 = (.916)^2 = .84$ (84% of variation in hourly wages is accounted for by months of training).

Testing Pearson's *r* for significance

1. Hypothesis:
   1. H_o: $r = 0$
   2. H_1: $r \neq 0$
2. Assumptions: 1) random sampling, 2) normal distributions, 3) a linear relationship, and 4) equal variance of y for all values of x ("homeskedasticity")
3. Sampling distribution is student's *t*, with d.f.=N-2=8: $t_{\alpha=0.05/2} = \pm 2.306$
4. Test Statistic is $t(\text{obtained}) = r\sqrt{\frac{N-2}{1-r^2}}$

   $t(\text{obtained}) = .916\sqrt{\frac{10-2}{1-.916^2}} = 6.45 > t_{\alpha=0.05/2}$:Reject H_o
5. Conclusion

Correlation Matrix

- A **correlation matrix** is a table that shows the relationships between all possible pairs of variables
- Using the matrix below:
  - What is the correlation between Birth Rate and Infant Mortality Rate?
  - Of all the variables correlated with Infant Mortality Rate, which has the strongest relationship? The weakest?

A Correlation Matrix Showing the Interrelationships of Four Variables for 74 Nations

	1	2	3	4
	Birth Rate	Infant Mortality Rate	Life Expectancy	Percent Urban
1 Birth Rate	1.00	0.88	−0.84	−0.66
2 Infant Mortality Rate	0.88	1.00	−0.89	−0.71
3 Life Expectancy	−0.84	−0.89	1.00	0.74
4 Percent Urban	−0.66	−0.71	0.74	1.00

A lot more measures out there

- Darlington, Richard. **An Outline for Choosing among 19 Measures of Association** (for categorical variables)**:**

http://www3.psych.cornell.edu/Darlington/crosstab/TABLE0.HTM

- Correlation (for numeric variables)

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression3.html

Measures of Bivariate Associations

		Independent Variable		
		Nominal 2 Groups	**Ordinal**	**Numeric**
Dependent Variable	**Nominal**	Phi, Cramer's V; Lambda;	Phi, Cramer's V; Lambda;	--
	Ordinal	Phi, Cramer's V; Lambda;	Gamma; Kendall's Tau; Somer's d	--
	Numeric	* Mann-Whitney U * Runs	--	Spearman's rho Pearson's r Kendall's Tau