Correlation

CSGE602013 –STATISTICS AND PROBABILITY

FACULTY OF COMPUTER SCIENCE UNIVERSITAS INDONESIA

Credits

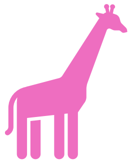
- The content was based on previous semester's course slides created by all previous lecturers.

References

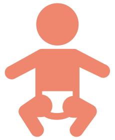
- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed., Sheldon M. Ross, Elsevier, 2009.
- Applied Statistics for the Behavioral Sciences, 5th Edition, Hinkle., Wiersma., Jurs., Houghton Mifflin Company, New York, 2003.

Introduction

- While studying babies’ development, we have the variables:



Height



Weight

Head Circumference

Type of milk consumed

Age of mother

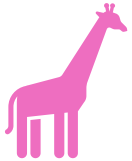
- We may obtain a statistic

The mean age of the mother is 25 years

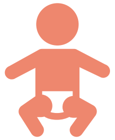
- This statistic is valid for **a single variable.**

Introduction

- While studying babies' development, we have the variables:



Height



Weight

Head Circumference

Type of milk consumed

Age of mother

- What if we want to describe **statistics / relationship of two variables?**
  - Age and height
  - Height and weight

Correlation

Scatterplot

- Pictures the relationship between variables

TABLE 5.1
Quantitative SAT Scores and Final Examination Scores for 15 Introductory Psychology Students*

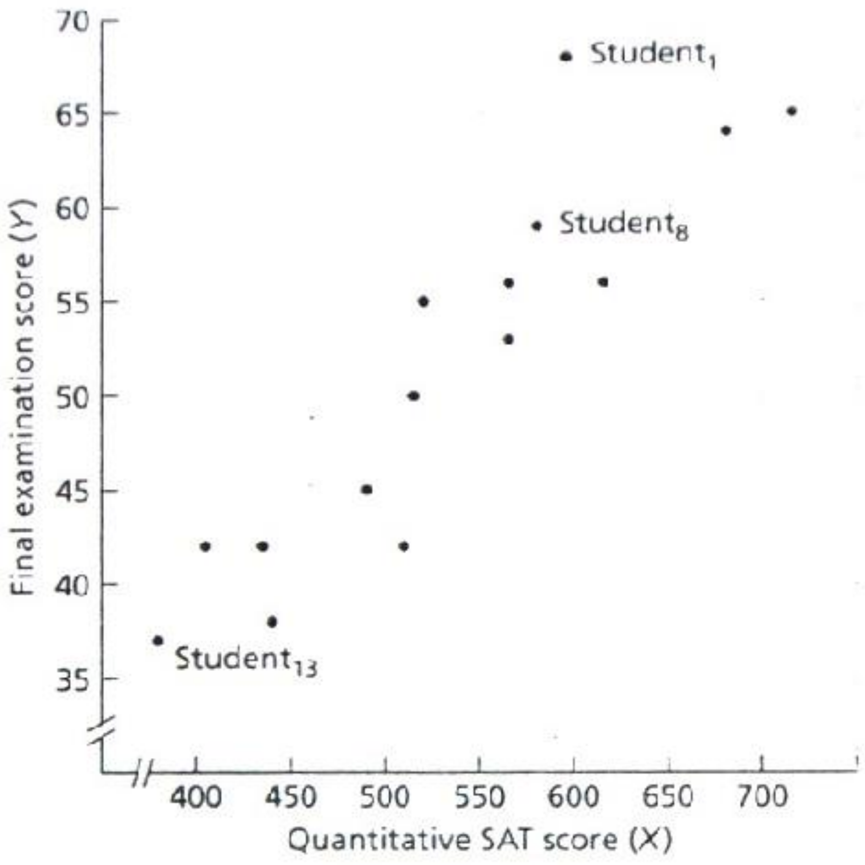
Student	*Quantitative SAT Score (X)*	*Final Examination Score (Y)*
1	595	68
2	520	55
3	715	65
4	405	42
5	680	64
6	490	45
7	565	56
8	580	59
9	615	56
10	435	42
11	440	38
12	515	50
13	380	37
14	510	42
15	565	53
Σ	8,010	772
	$\overline{X} = 534.00$	$\overline{Y} = 51.47$
	$s_X = 96.53$	$s_Y = 10.11$

**Note:* s_X and s_Y are sample standard deviations.

Scatterplot (2)

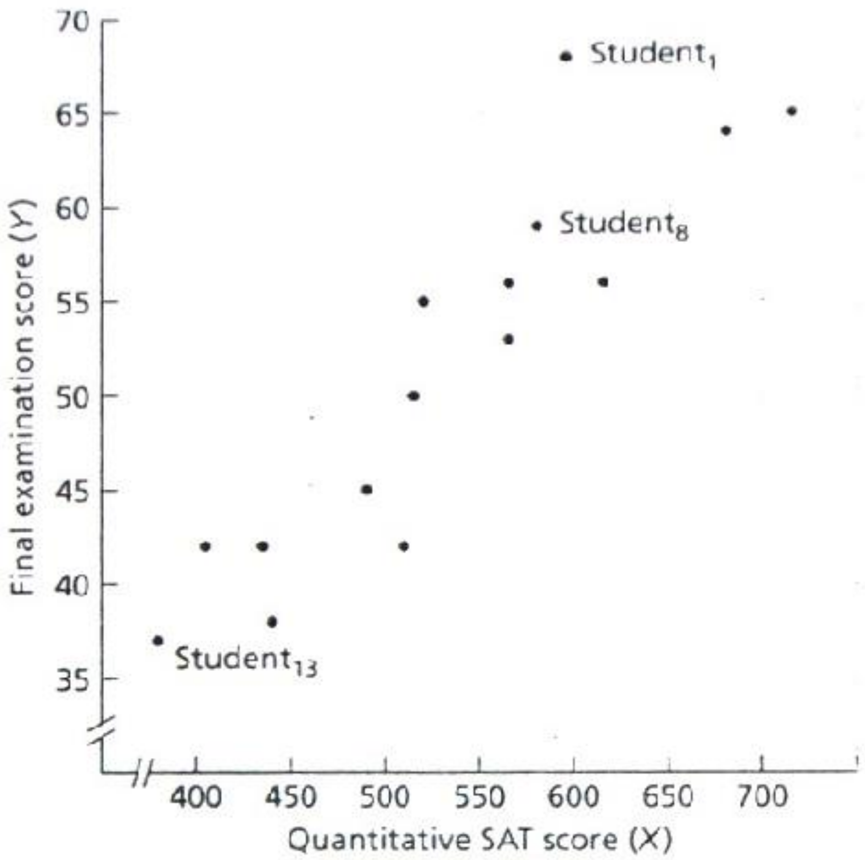
- What can we infer from this scatterplot?
- We can obtain **notion** of the relationship between two variables using scatterplot.
- But, it is **not precise**.
- How do we measure the relationship then?

We need a number!



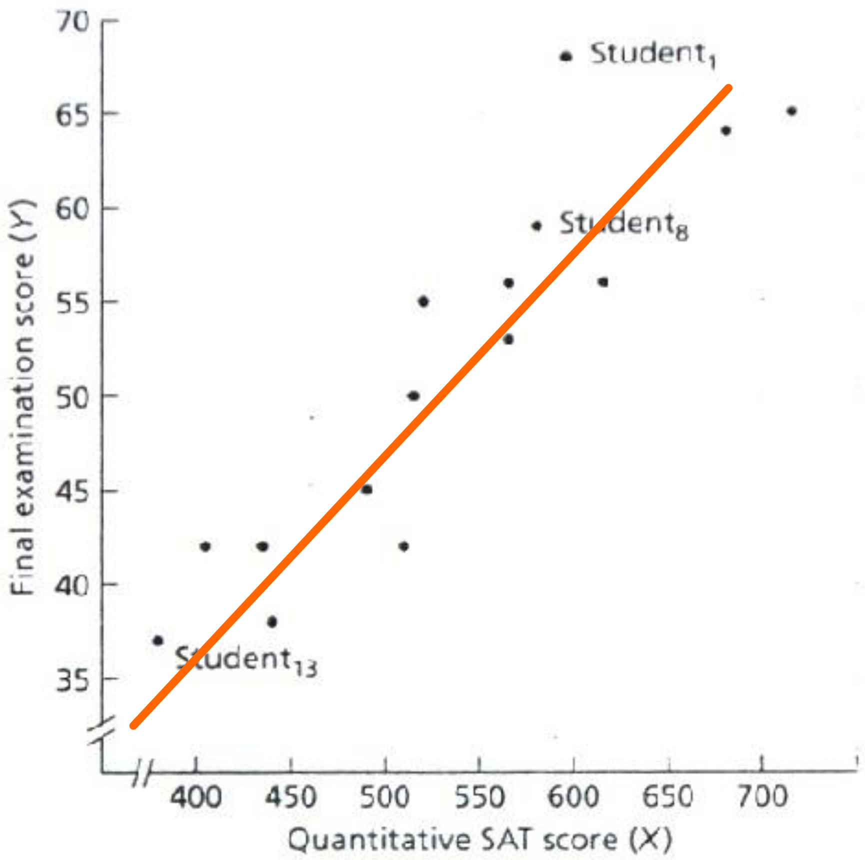
Correlation Coefficient

- Correlation coefficient is a measure of the relationship between two variables.



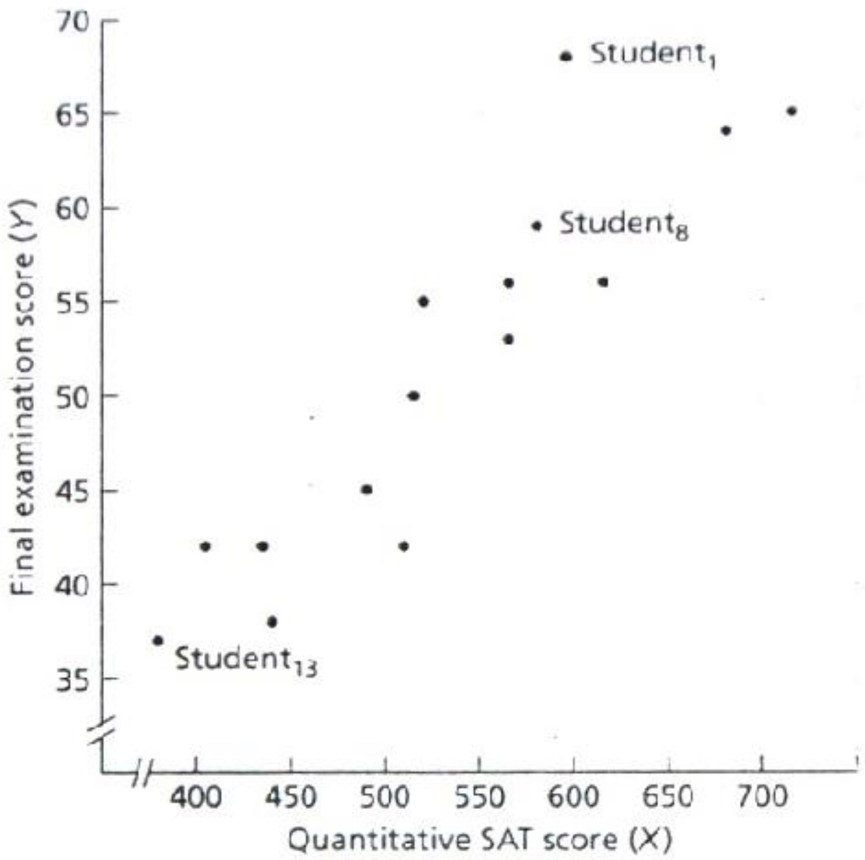
Correlation Coefficient (2)

- The **values** are between -1.0 and +1.0, inclusively.
- The **sign** shows the direction of relationship (slope).
  - + : positively correlated, lower-left-to-upper-right pattern
  - - : negatively correlated, upper-left-to-lower-right pattern



Correlation Coefficient (3)

- The absolute value of the coefficient indicates the magnitude of relationship.
  - 0 → there is no relationship
  - 1 → there is perfect relationship (linear relationship)



Correlation Coefficient (4)

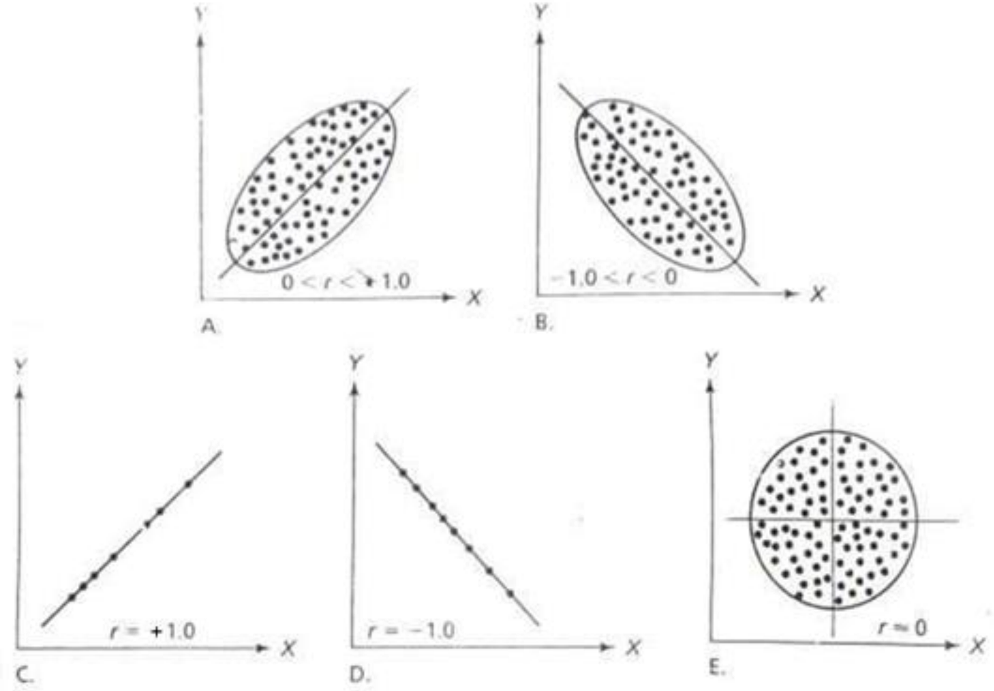
- The numeric value of a correlation coefficient is a function of
  - **Slope** (general direction of relationship): positive/negative
  - **Width of the ellipse** that encloses those points

Positively correlated

Negatively correlated

Positive perfect correlation

No relationship



Pearson r

- One of the well-known correlation coefficients is Pearson product-moment correlation coefficient, symbolized by r or **Pearson** r.
- Pearson r was developed by Karl Pearson (1857 - 1936).
- **Pearson** r is used most often in the behavioral sciences [Hinkle, 2003].

Pearson r

- Suppose there is positive correlation.

If an individual has a score on variable X that is above the mean of ($\bar{X}$), this individual is likely to have a score on the Y variable that is above the mean of Y ($\bar{Y}$).

- The same rationale can be applied for negative correlation (opposite direction).
- Karl Pearson defined a correlation coefficient between two variables (Pearson r) as:

$$r_{xy} = \frac{\sum (z_x z_y)}{n-1}$$

*) Standard scores are used rather than raw scores

Computing the Pearson r

TABLE 5.2
Data for Calculating the Pearson Product-Moment Correlation Coefficient Using Formula 5.1

	X	Y	z_X	z_Y	$z_X z_Y$
	595	68	0.63	1.64	1.03
	520	55	−0.15	0.35	−0.05
	715	65	1.88	1.34	2.52
	405	42	−1.34	−0.94	1.26
	680	64	1.51	1.24	1.87
	490	45	−0.46	−0.64	0.29
	565	56	0.32	0.45	0.14
	580	59	0.48	0.74	0.36
	615	56	0.84	0.45	0.38
	435	42	−1.03	−0.94	0.97
	440	38	−0.97	−1.33	1.29
	515	50	−0.20	−0.15	0.03
	380	37	−1.60	−1.43	2.29
	510	42	−0.25	−0.94	0.24
	565	53	0.32	0.15	0.05
Σ	8,010	772	0.00	0.00	12.67

$$r_{xy} = \frac{\sum(z_x z_y)}{n-1}$$

Computing the Pearson r (2)

TABLE 5.2
Data for Calculating the Pearson Product-Moment Correlation Coefficient Using Formula 5.1

	X	Y	z_X	z_Y	$z_X z_Y$
	595	68	0.63	1.64	1.03
	520	55	−0.15	0.35	−0.05
	715	65	1.88	1.34	2.52
	405	42	−1.34	−0.94	1.26
	680	64	1.51	1.24	1.87
	490	45	−0.46	−0.64	0.29
	565	56	0.32	0.45	0.14
	580	59	0.48	0.74	0.36
	615	56	0.84	0.45	0.38
	435	42	−1.03	−0.94	0.97
	440	38	−0.97	−1.33	1.29
	515	50	−0.20	−0.15	0.03
	380	37	−1.60	−1.43	2.29
	510	42	−0.25	−0.94	0.24
	565	53	0.32	0.15	0.05
Σ	8,010	772	0.00	0.00	12.67

$\bar{X} = 534$ $\bar{X} = 51.47$

$s_x = 96.53$ $s_y = 10.11$

Computing the Pearson r (3)

TABLE 5.2
Data for Calculating the Pearson Product-Moment Correlation Coefficient Using Formula 5.1

	X	Y	z_X	z_Y	$z_X z_Y$
	595	68	0.63	1.64	1.03
	520	55	−0.15	0.35	−0.05
	715	65	1.88	1.34	2.52
	405	42	−1.34	−0.94	1.26
	680	64	1.51	1.24	1.87
	490	45	−0.46	−0.64	0.29
	565	56	0.32	0.45	0.14
	580	59	0.48	0.74	0.36
	615	56	0.84	0.45	0.38
	435	42	−1.03	−0.94	0.97
	440	38	−0.97	−1.33	1.29
	515	50	−0.20	−0.15	0.03
	380	37	−1.60	−1.43	2.29
	510	42	−0.25	−0.94	0.24
	565	53	0.32	0.15	0.05
Σ	8,010	772	0.00	0.00	12.67

$$r_{xy} = \frac{12.67}{14} = 0.9$$

Computing the Pearson r (4)

- Using previous formula to compute Pearson r is really tedious
  - We need to convert each raw score to a z-score
- Instead, transform that formula into another formula that doesn't need to compute z-score directly
  - Deviation score formula
  - Raw score formula
  - Covariance to find Pearson r

Deviation Score Formula for Pearson r

- Using definition of z-score and standard deviation, we can transform original formula of Pearson r into

$$r_{xy} = \frac{\sum(xy)}{\sqrt{\left(\sum x^2 \sum y^2\right)}}$$

  - x_i and y_i (small x & y) are deviation scores
  - $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$

Deviation Score Formula for Pearson r (2)

	X	Y	x	y	xy	x^2	y^2
	595	68	61.0	16.53	1,008.33	3,721.0	273.24
	520	55	−14.0	3.53	−49.42	196.0	12.46
	715	65	181.0	13.53	2,448.93	32,761.0	183.06
	405	42	−129.0	−9.47	1,221.63	16,641.0	89.68
	680	64	146.0	12.53	1,829.38	21,316.0	157.00
	490	45	−44.0	−6.47	284.68	1,936.0	41.86
	565	56	31.0	4.53	140.43	961.0	20.52
	580	59	46.0	7.53	346.38	2,116.0	56.70
	615	56	81.0	4.53	366.93	6,561.0	20.52
	435	42	−99.0	−9.47	937.53	9,801.0	89.68
	440	38	−94.0	−13.47	1,266.18	8,836.0	181.44
	515	50	−19.0	−1.47	27.93	361.0	2.16
	380	37	−154.0	−14.47	2,228.38	23,716.0	209.38
	510	42	−24.0	−9.47	227.28	576.0	89.68
	565	53	31.0	1.53	47.43	961.0	2.34
Σ	8,010	772	0.0	0.0	12,332.00	130,460.0	1,429.72

$$r_{xy} = \frac{\sum(xy)}{\sqrt{\left(\sum x^2 \sum y^2\right)}}$$

$\bar{X} = 534 \quad \bar{Y} = 51.47$

Deviation Score Formula for Pearson r (3)

	X	Y	x	y	xy	x^2	y^2
	595	68	61.0	16.53	1,008.33	3,721.0	273.24
	520	55	−14.0	3.53	−49.42	196.0	12.46
	715	65	181.0	13.53	2,448.93	32,761.0	183.06
	405	42	−129.0	−9.47	1,221.63	16,641.0	89.68
	680	64	146.0	12.53	1,829.38	21,316.0	157.00
	490	45	−44.0	−6.47	284.68	1,936.0	41.86
	565	56	31.0	4.53	140.43	961.0	20.52
	580	59	46.0	7.53	346.38	2,116.0	56.70
	615	56	81.0	4.53	366.93	6,561.0	20.52
	435	42	−99.0	−9.47	937.53	9,801.0	89.68
	440	38	−94.0	−13.47	1,266.18	8,836.0	181.44
	515	50	−19.0	−1.47	27.93	361.0	2.16
	380	37	−154.0	−14.47	2,228.38	23,716.0	209.38
	510	42	−24.0	−9.47	227.28	576.0	89.68
	565	53	31.0	1.53	47.43	961.0	2.34
Σ	8,010	772	0.0	0.0	12,332.00	130,460.0	1,429.72

$$r_{xy} = \frac{12332}{\sqrt{(130460)(1429.72)}} = 0.90$$

$\bar{X} = 534$ $\bar{Y} = 51.47$

$s_x = 96.53$ $s_y = 10.11$

Raw Score Formula for Pearson r

- By algebraically manipulating deviation score formula, we can get the following formula:

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

- Advantages
  - We only need raw scores here
  - X, Y are raw scores for each variable

Raw Score Formula for Pearson r (2)

Student	X (SAT Score)	Y (Final Exam)	XY	X²	Y²
1	595,00	68,00	40.460	354.025	4.624
2	520,00	55,00	28.600	270.400	3.025
3	715,00	65,00	46.475	511.225	4.225
4	405,00	42,00	17.010	164.025	1.764
5	680,00	64,00	43.520	462.400	4.096
6	490,00	45,00	22.050	240.100	2.025
7	565,00	56,00	31.640	319.225	3.136
8	580,00	59,00	34.220	336.400	3.481
9	615,00	56,00	34.440	378.225	3.136
10	435,00	42,00	18.270	189.225	1.764
11	440,00	38,00	16.720	193.600	1.444
12	51500	50,00	25.750	265.225	2.500
13	380,00	37,00	14.060	144.400	1.369
14	510,00	42,00	21.420	260.100	1.764
15	565,00	53,00	29.945	319.225	2.809
Σ	8.010,00	772,00	424.580	4.407.800	41.162

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

$$r_{xy} = \frac{15(424.580) - (8.010)(772)}{\sqrt{\left(15(4.407.800) - 8010^2\right)\left(15(41.162) - 772^2\right)}} = 0.90$$

Covariance Formula for Pearson r (2)

- Definition of covariance x and y – (stay tuned for more details later on)

$$s_{xy} = \frac{\sum(\mathrm{X} - \overline{\mathrm{X}})(\mathrm{Y} - \overline{\mathrm{Y}})}{\mathrm{n} - 1} = \frac{\sum(\mathrm{xy})}{\mathrm{n} - 1}$$

- Using this definition, we transform previous formula into

$$r_{xy} = \frac{\mathrm{s_{xy}}}{\mathrm{s_x s_y}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)\mathrm{s_x s_y}}$$

Pearson r

- Two Conditions For Computing Pearson r
  - The two variables to be correlated must be **paired observations** for the same set of individuals or object
  - We use mean and variance in computing Pearson r → the variables being correlated must be measured on an interval or ratio scale.
- Factors affecting the size of Pearson r

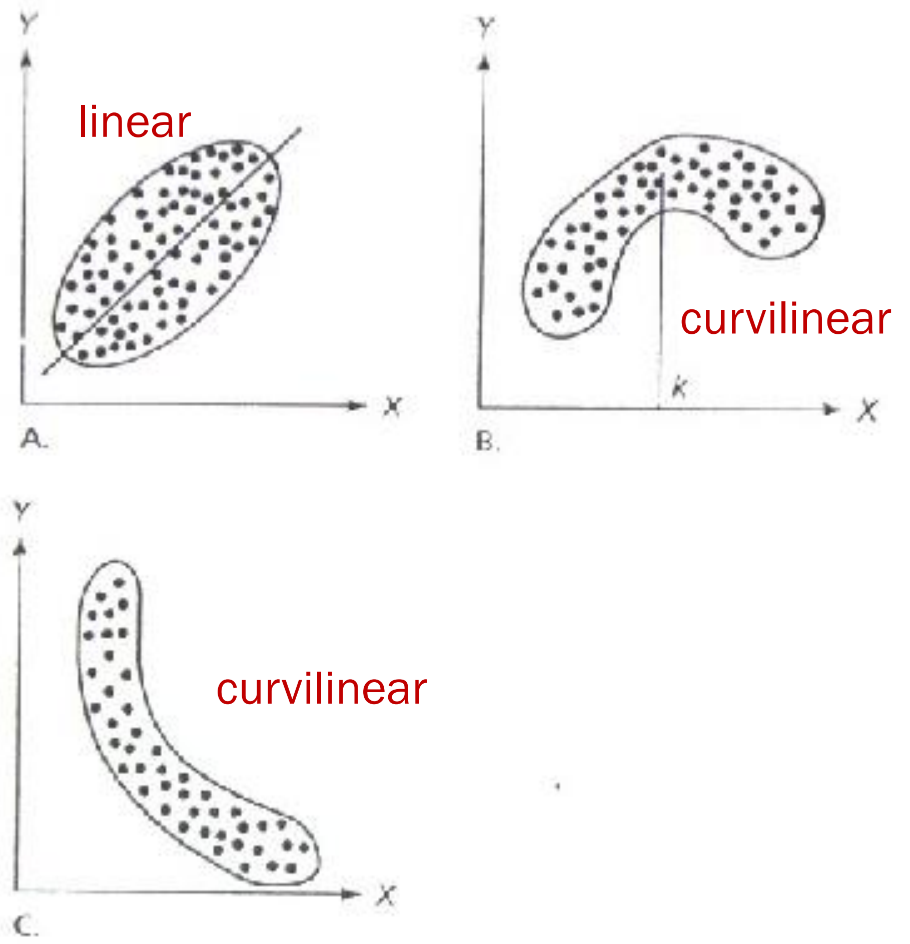
Linearity

Homogeneity of the group

Size of the group

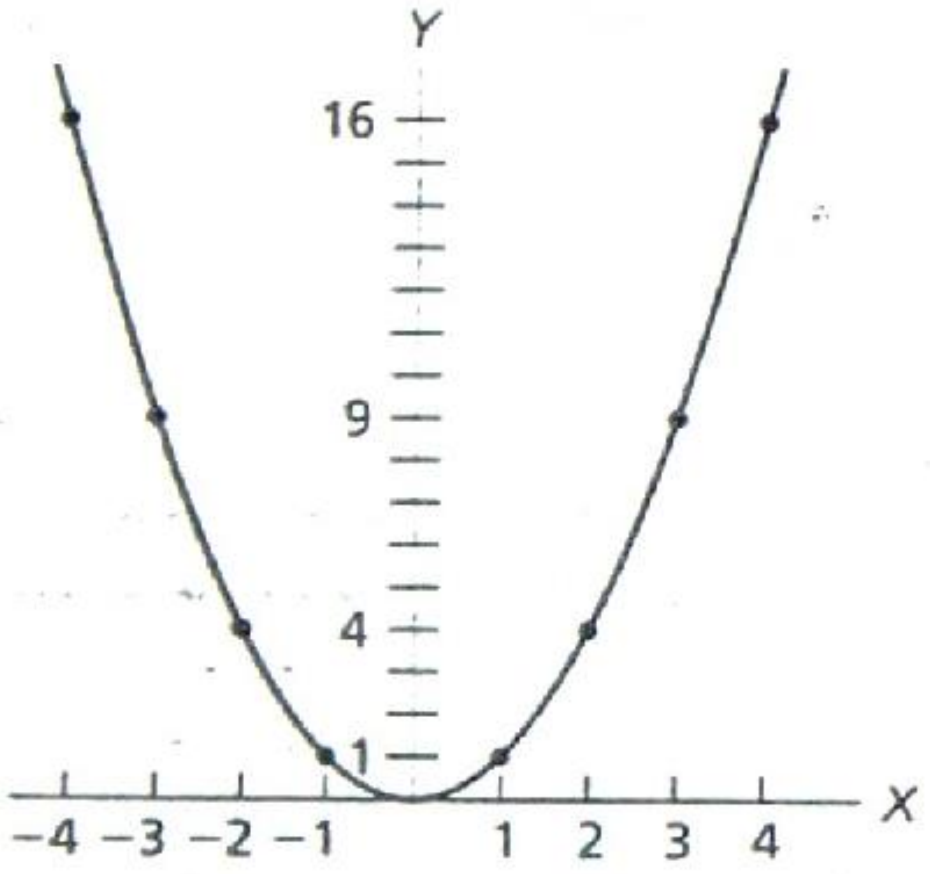
Linearity

- Relationship between two variables can be:
  - Linear
  - Curvilinear
- Pearson r is an index of the linear relationship between two variables.



Linearity (2)

- If the Pearson r is applied to variables that are curvilinear, it will underestimate the relationship between the variables
  - Pearson r of the data in the picture is 0.
  - But, it has a perfect relation $y = x^2$
  - Here, Pearson r is not suitable.



Homogeneity

Homogeneity vs Variance?

- As the homogeneity of a group increases, the variance decreases.
- As the homogeneity increases on one **or** both variables:
  - The absolute value of the correlation coefficient tends to become smaller.
- So?
  - If you are looking for relationships between variables, make sure that there is enough variation or heterogeneity in the scores.

Homogeneity of the Group

- We are investigating the relationship between IQ scores and performance on a cognitive task, we need to include a wide range of IQ scores.

- What if only individuals with IQ > 140 are included ?

We get low correlation !

- Does this mean that there is no relationship here?

Size of the Group

- In general, size of the group used in the calculation of the Pearson r does not influence the value of the coefficient.
- But, size of the group affects the **accuracy** of the relationship
- **Exception:**
  - When $n = 2$, what do you think will happen?

Properties of Pearson r [Ross, 2009]

- $-1 \leq r \leq 1$
- If for constants a and b, with $b > 0$, and $y_i = a + bx_i$, then $r = 1$
- If for constants a and b, with $b < 0$, and $y_i = a + bx_i$, then $r = -1$
- If r is the sample correlation coefficient for pairs $(x_i, y_i), i = 1, \dots, n$, Then, r is also correlation coefficient for the data pairs:

$$(a + bx_i, c + dy_i) \qquad\qquad i = 1,2, \dots \dots . n$$

Interpreting the Correlation Coefficient

- Rule of thumb
  - Pearson r is an ordinal scale [Hinkle, et al., 2003]

Size of Correlation	Interpretation
0.90 to 1.00(-0.90 to -1.00)	Very high positive (negative) correlation
0.70 to 0.90(-0.70 to -0.90)	High positive (negative) correlation
0.50 to 0.70(-0.50 to -0.70)	Moderate (negative) correlation
0.30 to 0.50(-0.30 to -0.50)	Low positive (negative) correlation
0.00 to 0.30(0.00 to -0.30)	Very low positive (negative) correlation

Pearson r in Terms of Variance

- Variance represents individual differences.
- Pearson r also indicates
  - The proportion of the variance in **one variable** that can be associated with variance in the **other variable.**
- Or,

$$s_Y^2 = s_A^2 + s_O^2$$

  - $s_Y{}^2$ = the total variance in Y
  - $s_A{}^2$ = the variance in Y associated with X
  - $s_O{}^2$ = the variance in Y associated with other factors

Example:
Pearson r = 0.69 between variable X and Y. This tells us that, there are factors other than X, could contribute to variance in Y.

Coefficient of Determination

- The square of correlation coefficient (r^2) equals the proportion of the total variance in Y that can be associated with variance in X, or the **coefficient of determination.**

$$r^2 = \frac{s_A^2}{s_Y^2}$$

  - Previously, $r = 0.69, r2 = 0.48$ so, 48% variance in Y can be associated with the variance in X.

Coefficient of Determination (2)

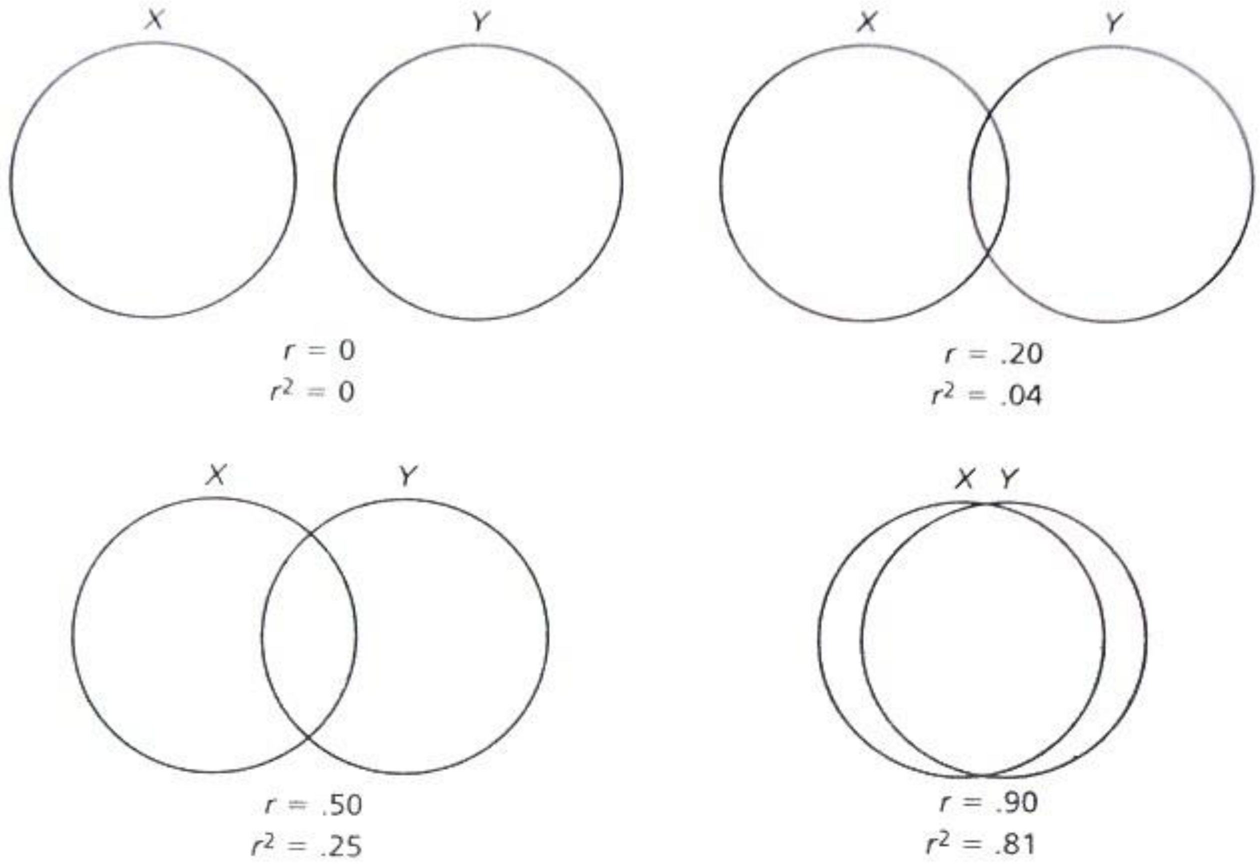


FIGURE 5.6
Illustration of the coefficient of determination (r^2) as overlapping areas representing variance

Spearman Rho (ρ)

- Spearman ρ is a special case of the Pearson r
- Spearman ρ is used when ***rank*** information is used:
  - Data itself consist of ranks
  - Where the raw scores are converted to rankings
- Why?
  - Rankings are ordinal data, the Pearson r is not applicable to them.

Spearman Rho (ρ) (2)

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

- Where
  - n : number of paired ranks
  - d : difference between the paired ranks
- If there exist same raw scores ? (or tied ranks) ?
  - Average of these rank positions

Spearman Rho (ρ) (3)

Student	X (SAT Score)	Y (Final Exam)	Xrank	Yrank	d	d^2
1	595,00	68,00	4	1	3	9
2	520,00	55,00	8	7	1	1
3	715,00	65,00	1	2	-1	1
4	405,00	42,00	14	12	2	4
5	680,00	64,00	2	3	-1	1
6	490,00	45,00	11	10	1	1
7	565,00	56,00	6,5	5,5	1	1
8	580,00	59,00	5	4	1	1
9	615,00	56,00	3	5,5	-2,5	6,25
10	435,00	42,00	13	12	1	1
11	440,00	38,00	12	14	-2	4
12	51500	50,00	9	9	0	0
13	380,00	37,00	15	15	0	0
14	510,00	42,00	10	12	-2	4
15	565,00	53,00	6,5	8	-1,5	2,25
Σ	8.010,00	772,00			0	36,50

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Spearman Rho (ρ) (4)

Student	X (SAT Score)	Y (Final Exam)	Xrank	Yrank	d	d^2
1	595,00	68,00	4	1	3	9
2	520,00	55,00	8	7	1	1
3	715,00	65,00	1	2	-1	1
4	405,00	42,00	14	12	2	4
5	680,00	64,00	2	3	-1	1
6	490,00	45,00	11	10	1	1
7	565,00	56,00	6,5	5,5	1	1
8	580,00	59,00	5	4	1	1
9	615,00	56,00	3	5,5	-2,5	6,25
10	435,00	42,00	13	12	1	1
11	440,00	38,00	12	14	-2	4
12	51500	50,00	9	9	0	0
13	380,00	37,00	15	15	0	0
14	510,00	42,00	10	12	-2	4
15	565,00	53,00	6,5	8	-1,5	2,25
Σ	8.010,00	772,00			0	36,50

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(36.5)}{15(225-1)}$$

$$\rho = 0.93$$

Recall that, previously, we got **Pearson** $r = 0.9$.

Pearson r and Spearman rho (ρ)

- Previously we obtained Pearson $r = 0.9$ and Spearman $\rho = 0.93$
- The difference between Pearson r and Spearman ρ because of some tied scores.
- Example:
  - Score 565 appears in 6th and 7th position, the rank will be average{6,7} = 6.5
  - Score 42 appears in 11th, 12th, 13th position, the rank will be average{11,12,13} = 12
- When there is no tied scores, Spearman ρ will equal Pearson r

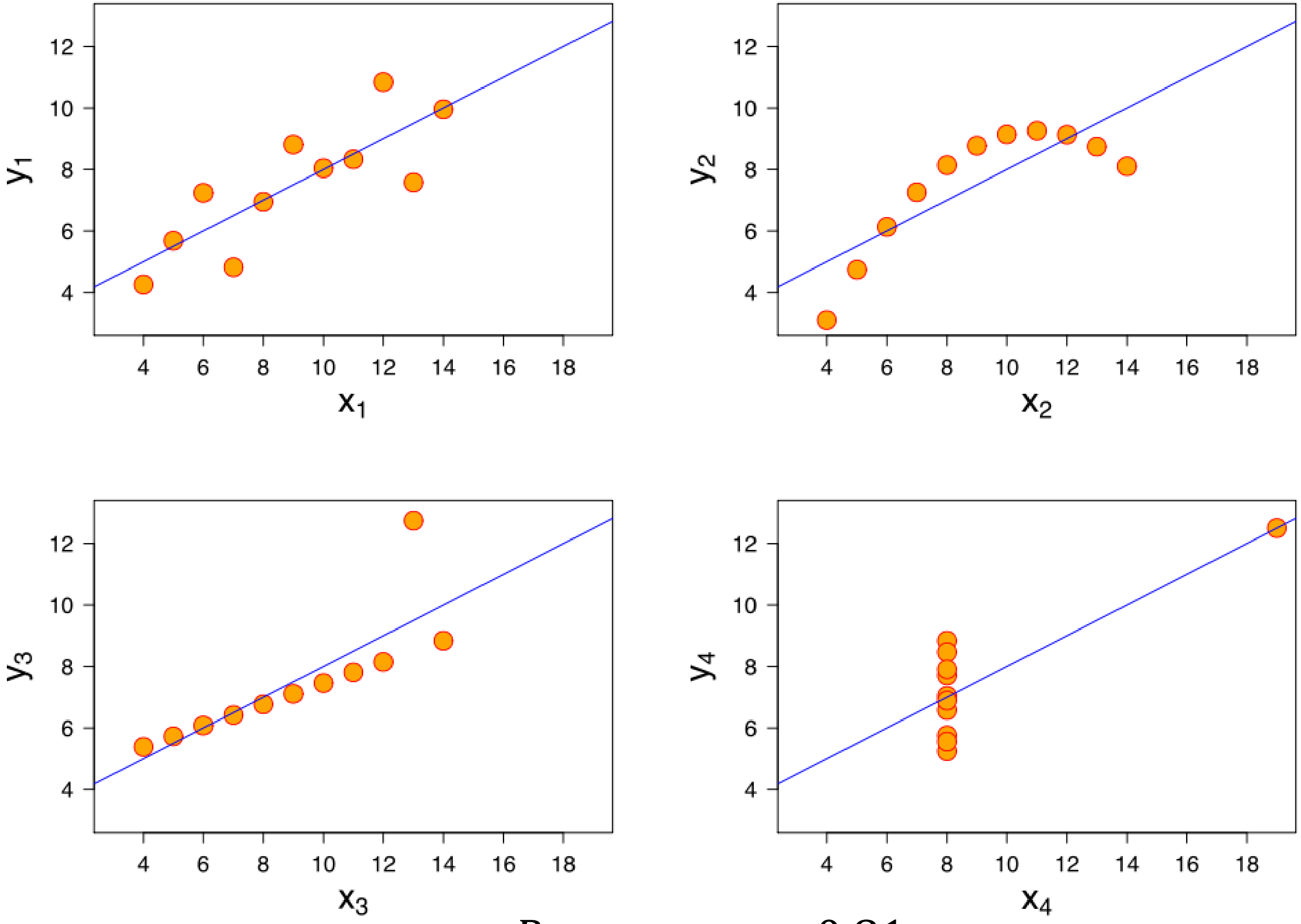
Correlation and Causality

Do not mistake correlation with causality!

- Two variables with high correlation show that they have strong association.
- But, it doesn't necessarily follow that scores on one variable are directly caused by scores on the other variable.
- A third, fourth, or a combination of other variables may be causing the two correlated variables.

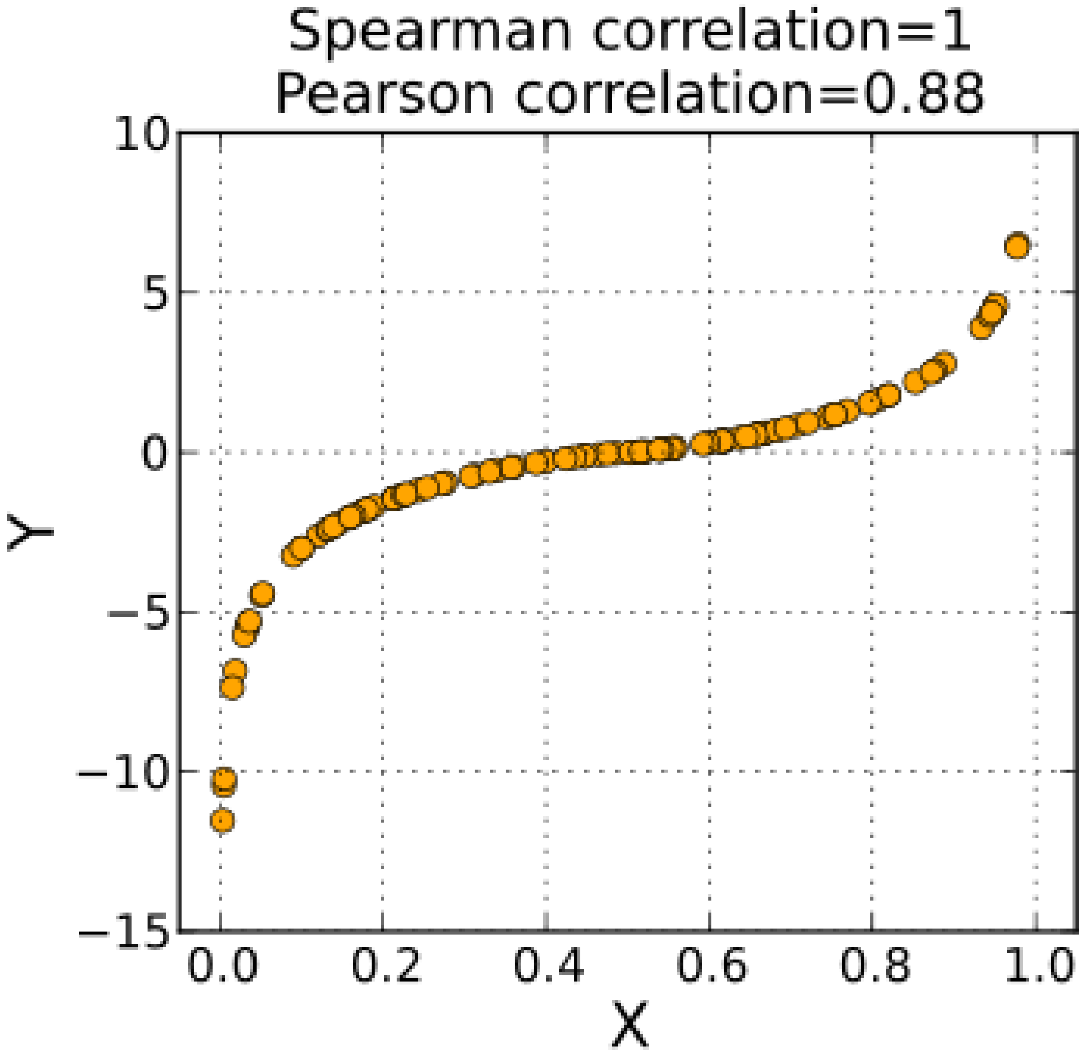
Examples?

Correlation Between 2 Variables



$$Pearson\ r\ =\ 0.81$$

Correlation Between 2 Variables (2)



Correlation Between 2 Variables (3)