Correlation:
A measure of relationship

CSGE602013 - Statistics and Probability

Credits

These course slides were prepared by Alfan F. Wicaksono. **Suggestions**, **comments**, and **criticism** regarding these slides are welcome. Please kindly send your inquiries to current lecturer.

The content was based on previous semester's (odd semester 2013/2014) 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.

Outline

- Meaning of Correlation
- Correlation Coefficient
- Computing the Pearson r
- Factor Affecting the Size of the Pearson r
- Interpreting the Correlation Coeficient
- Spearman rho (special case of Pearson r)
- Correlation and Causality

Meaning of Correlation

Previously, we describe some statistics **for a single variable**.

Now, we want to describe statistic for **two variables**, i.e., **relationship** between two variables !

For example, suppose we have table containing **paired values** (x_i, y_i) as follow:

- TOEFL Score & Math Score
- GPA & First Salary
- Age & Productive time a day
- ...

Is there any relationship between these two variables ?

Meaning of Correlation

SAT score & Final exam

We see that students with high SAT score tend to have high score on the final exam.

Those students with low SAT scores usually have low final exam score.

There is relationship !

There is **correlation** between two 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.

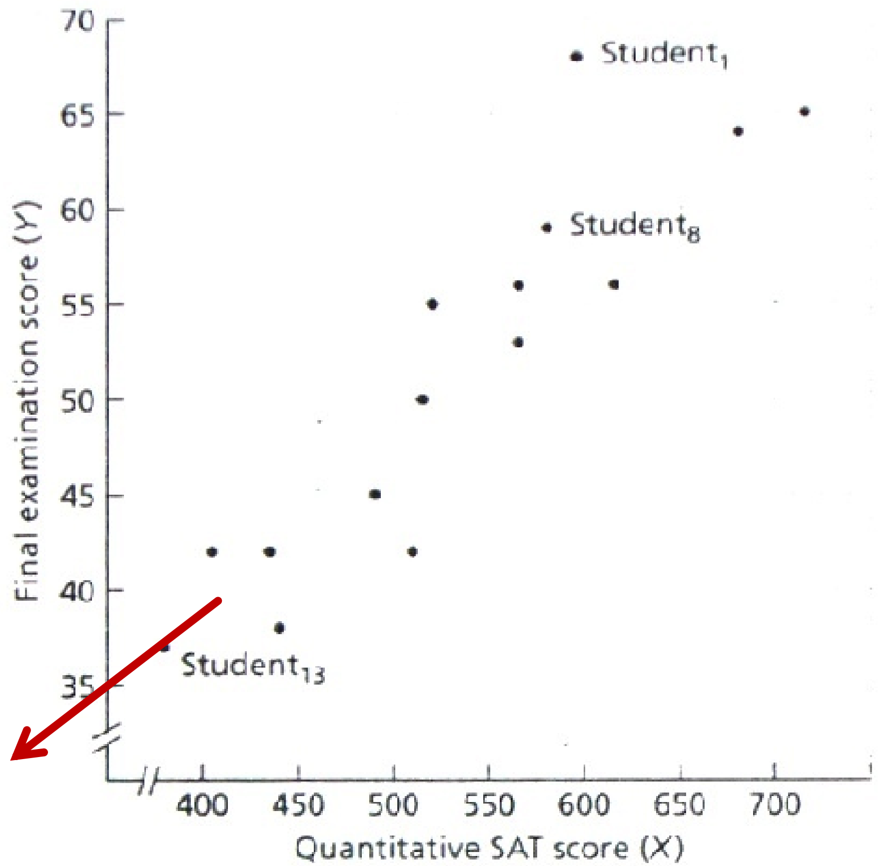
Meaning of Correlation

Scatterplot

Picture of relationship between variables

We can obtain notion of the relationship between two variables using scatterplot. But, it is **not precise** !

It seems that there is **positive correlation** between these two variables



Correlation Coefficient

Correlation Coefficient is a measure of the relationship between two variables.

This can take on values between **-1.0 and +1.0, inclusive**.

Sign indicates the direction of relationship (**slope**).

- + : positively correlated, lower-left-to-upper-right pattern
- - : negatively correlated, upper-left-to-lower-right pattern

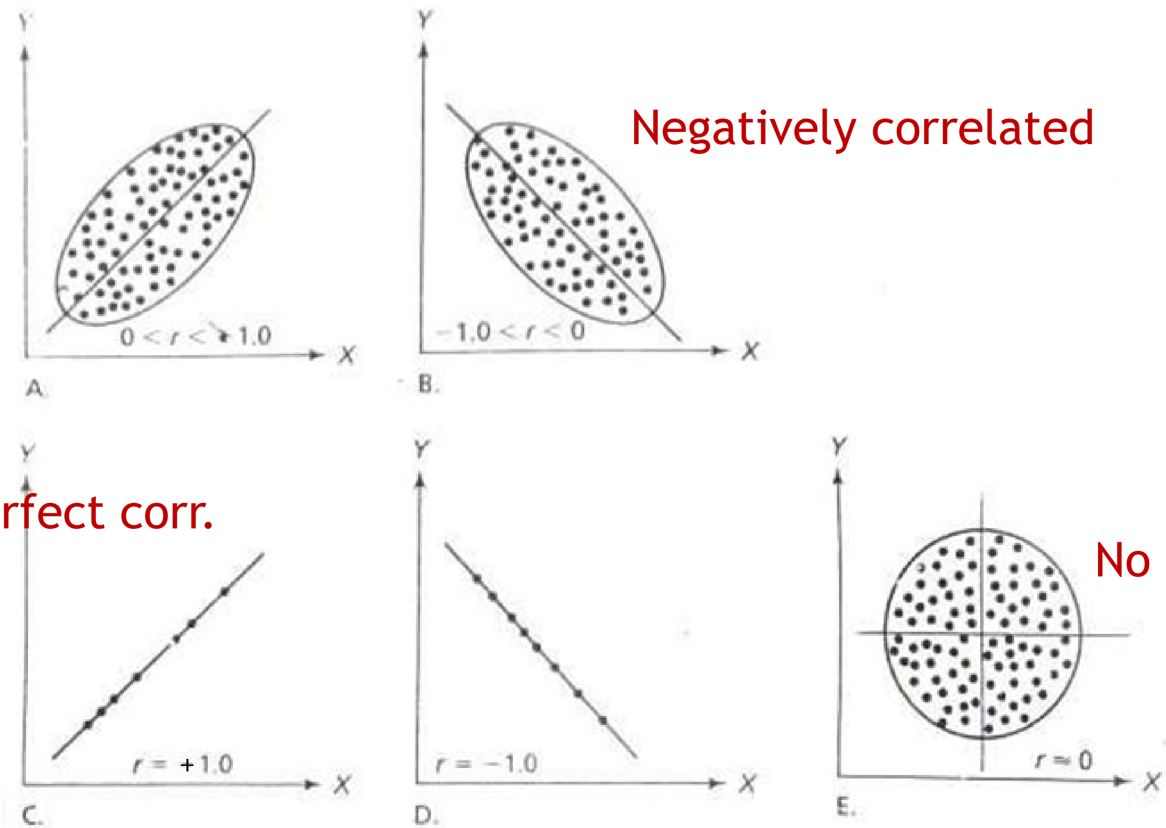
Absolute value of the coefficient indicates the magnitude of relationship.

- 0 -> there is no relationship
- 1 -> there is perfect relationship (**linear relationship**)

Correlation Coefficient

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



Negatively correlated

Positive perfect corr.

No relationship

Correlation Coefficient

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].

Computing the Pearson *r*

The rationale of Pearson *r*

Suppose there is positive correlation: if an individual has a score on variable X that is above the mean of X ($\bar{X}$), 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).

Computing the Pearson *r*

Based on the previous rationale, Karl Pearson defined a correlation coefficient between two variables (**Pearson *r***) as:

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

Standard scores are used rather than raw scores, because of the difference in the measurements for the two variables.

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

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

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

Strong enough ☺

X: SAT Score $\bar{X}$=534 $\bar{Y}$=51.47

Y: Final Exam s_x=96.53 s_y=10.11

Computing the Pearson *r*

Using previous formula to compute **Pearson *r*** is really tedious for some of you ☺

It's because we need to convert each raw score to a z-score ☺

Can we **transform that formula into another formula** that doesn't need to compute z-score directly ? **YES**

- Deviation score formula
- Raw score formula
- Covariance to find **pearson *r***

Computing the Pearson *r*

1. Deviation score formula

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

$$x_i = X_i - \overline{X} \qquad y_i = Y_i - \overline{Y}$$

Computing the Pearson *r*

1. Deviation score formula

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

$s_x=96.53$ $s_y=10.11$

	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

Computing the Pearson *r*

1. Deviation score formula

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

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

Same as before

Computing the Pearson *r*

2. Raw score formula

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

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

We only need raw scores here ! ☺
X, Y are raw scores for each variable

Computing the Pearson *r*

Raw score formula

Student	X (SAT Score)	Y (Final Exam)	XY	X^2	Y^2
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

Computing the Pearson *r*

2. Raw score formula

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

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

Same as before

Computing the Pearson *r*

3. Covariance to find Pearson *r*

Definition of **covariance x and y** (we'll discuss in next course):

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

Using this definition, we transform previous formula into:

Pearson r

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

Before computing Pearson *r* …

We must satisfy two conditions before computing **Pearson *r***

- The two variables to be correlated must be ***paired observations*** for the **same set** of **individuals or object.**
- Because we use mean and variance in computing **Pearson** ***r***, the variables being correlated must be measured on an **interval or ratio scale.**

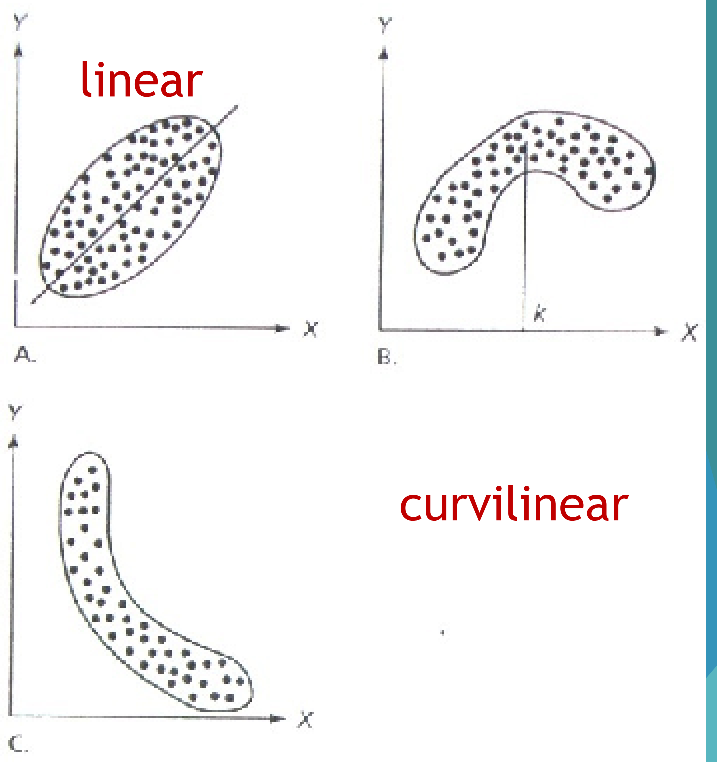
Factors affecting the size of **Pearson *r***

1. Linearity

Relationship between two variables can be:

- Linear
- Curvilinear

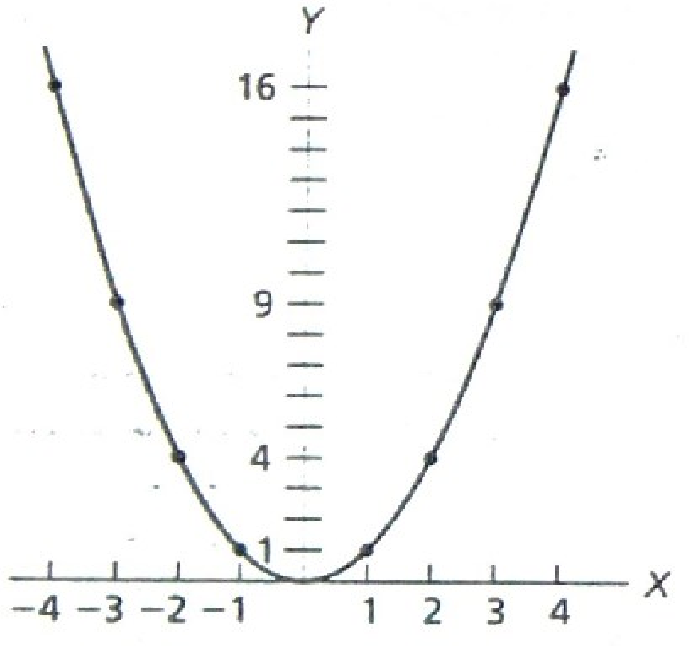
Pearson *r* is an index of the **linear relationship** between two variables.



Factors affecting the size of **Pearson *r***

1. Linearity

If the **Pearson *r*** is applied to variables that are **curvilinearly related**, 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.

Factors affecting the size of **Pearson *r***

2. Homogeneity of the group

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**.

Implication:

If you are looking for relationships between variables, make sure that there is **enough variation or heterogeneity** in the scores.

Factors affecting the size of **Pearson *r***

2. Homogeneity of the group

For example:

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

Now, what if **only** individuals **with IQ > 140** are included ?

We get low correlation !

Can we conclude that there is no relationship here ?? Why ?

Factors affecting the size of **Pearson *r***

3. 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, why ?

Interpreting the correlation coefficient

Properties of Pearson *r* [Ross, 2009]

1. $-1 \leq r \leq 1$
2. If for constants a and b, with **b > 0**,

$$y_i = a + b x_i$$

then **r = 1**

3. If for constants a and b, with **b < 0**,

$$y_i = a + b x_i$$

then **r = -1**

4. If **r** is sample correlation coeff. for pairs (x_i, y_i), i = 1, …, n.

Then, **r** is also correlation coeff. for the data pairs:

$$(a + b x_i, c + d y_i) \qquad i = 1, \ldots, n$$

Interpreting the correlation coefficient

Rule of thumb for interpreting the size of **Pearson *r***.

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

Interpreting the correlation coefficient

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**.

For 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.

Interpreting the correlation coefficient

Pearson *r* in terms of variance

Or, symbolically

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

- 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

Interpreting the correlation coefficient

Pearson *r* in terms of variance

The notion of **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 → coeff. determination.

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

Previously, r = 0.69, r^2 = 0.48. so, 48% variance in Y can be associated with the variance in X.

Interpreting the correlation coefficient

Various coefficient of determination

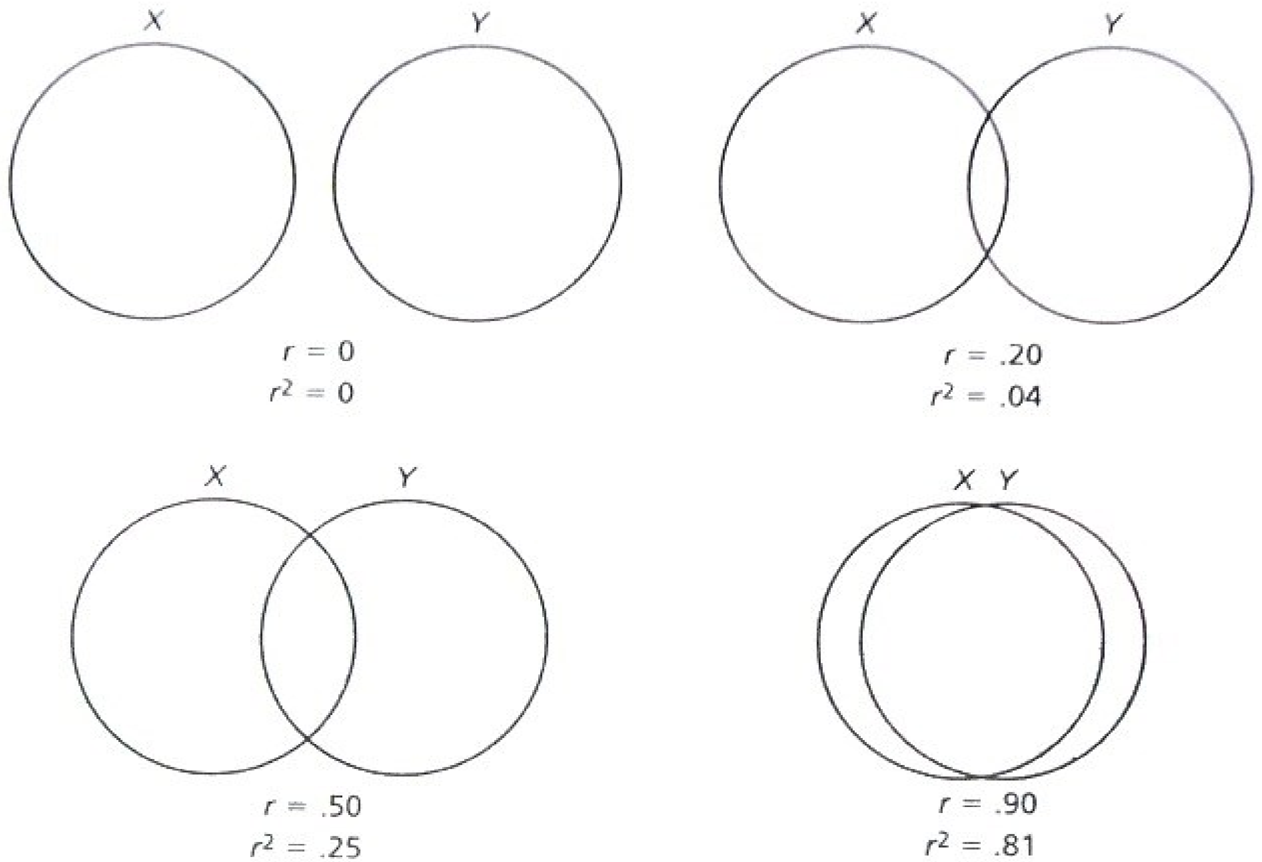


FIGURE 5.6
Illustration of the coefficient of determination (r^2) as overlapping areas representing variance

Spearman rho (ρ)

Spearman rho is a **special** case of the **Pearson** ***r***.

Spearman rho is used when “Ranking” information is used:

- Data consist of ranks
- Studies in which the raw scores are converted to rankings

Why ?

Rankings are ordinal data, the **Pearson** ***r*** is not applicable to them.

Spearman rho (ρ)

Formula for spearman rho

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

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
- Suppose 2 same scores would have occupied rank 6 & 7. so, both are assigned rank **6.5**.
- Suppose 3 same scores would have occupied rank 11, 12, & 13. so, they are assigned rank **12**.

Spearman rho (ρ)

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

Spearman rho (ρ)

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

$$\rho = 1 - \frac{6(36.50)}{15(225 - 1)} = 1 - 0.07 = 0.93$$

Previously, we got **Pearson *r*** = 0.9.

So, **Spearman rho** and **Pearson *r*** are different ???

Spearman rho (ρ)

- The difference between Pearson *r* and Spearman rho happened **because of some tied scores**.
  - Example:
  - *Score 565 appears in 6^{th} and 7^{th} position, the rank will be average{6,7} = 6.5*
  - *Score 42 appears in 11^{th}, 12^{th}, 13^{th} position, the rank will be average{11,12,13} = 12*

- When there is no tied scores, Spearman rho will equal Pearson *r*.

Correlation & Causality

Two variables with high correlation ?

- Indeed, 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.

Correlation & Causality

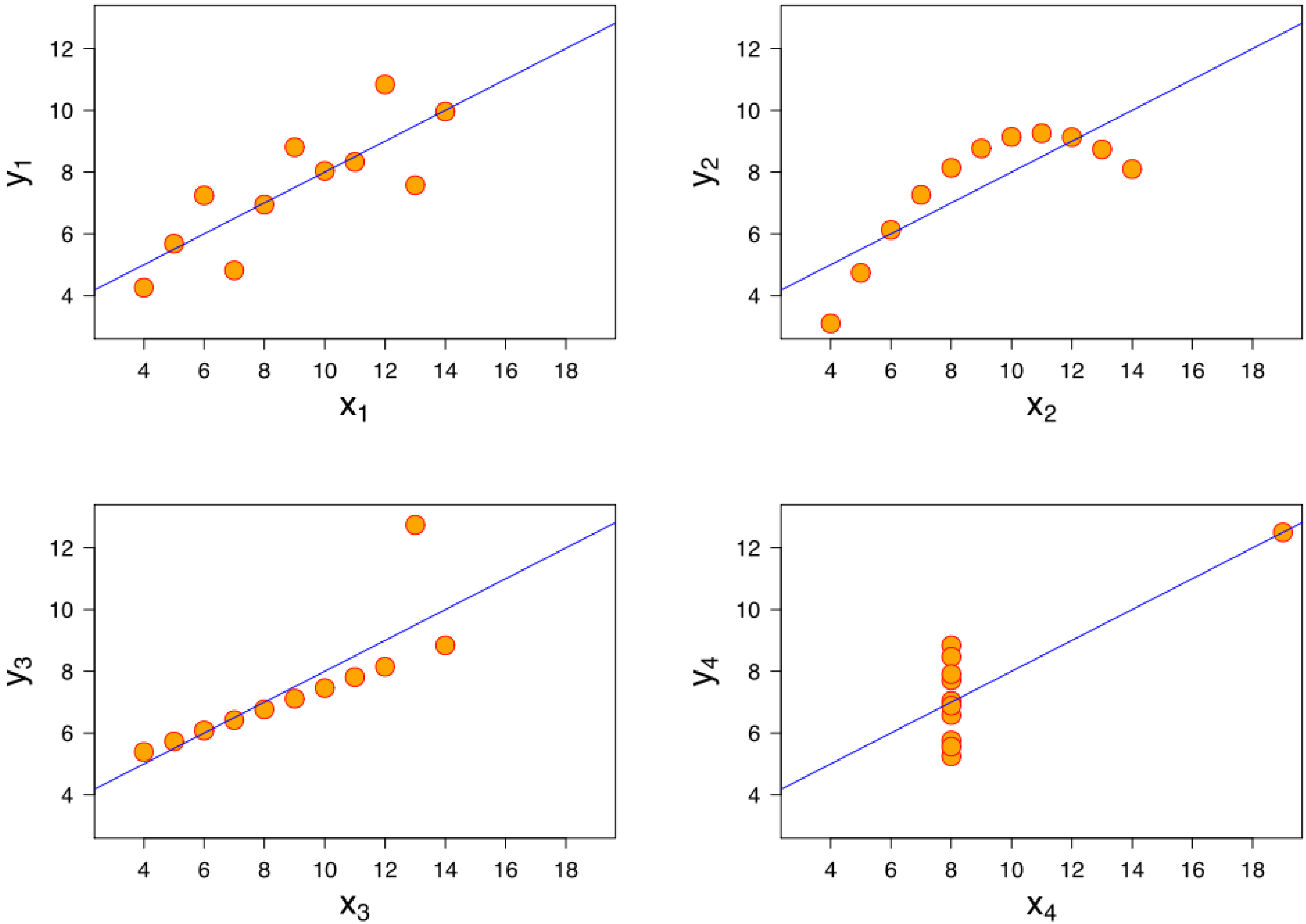
Example: survey on **K-6 elementary school**.

Reading comprehension and the **running speed** have a strong **positive correlation.**

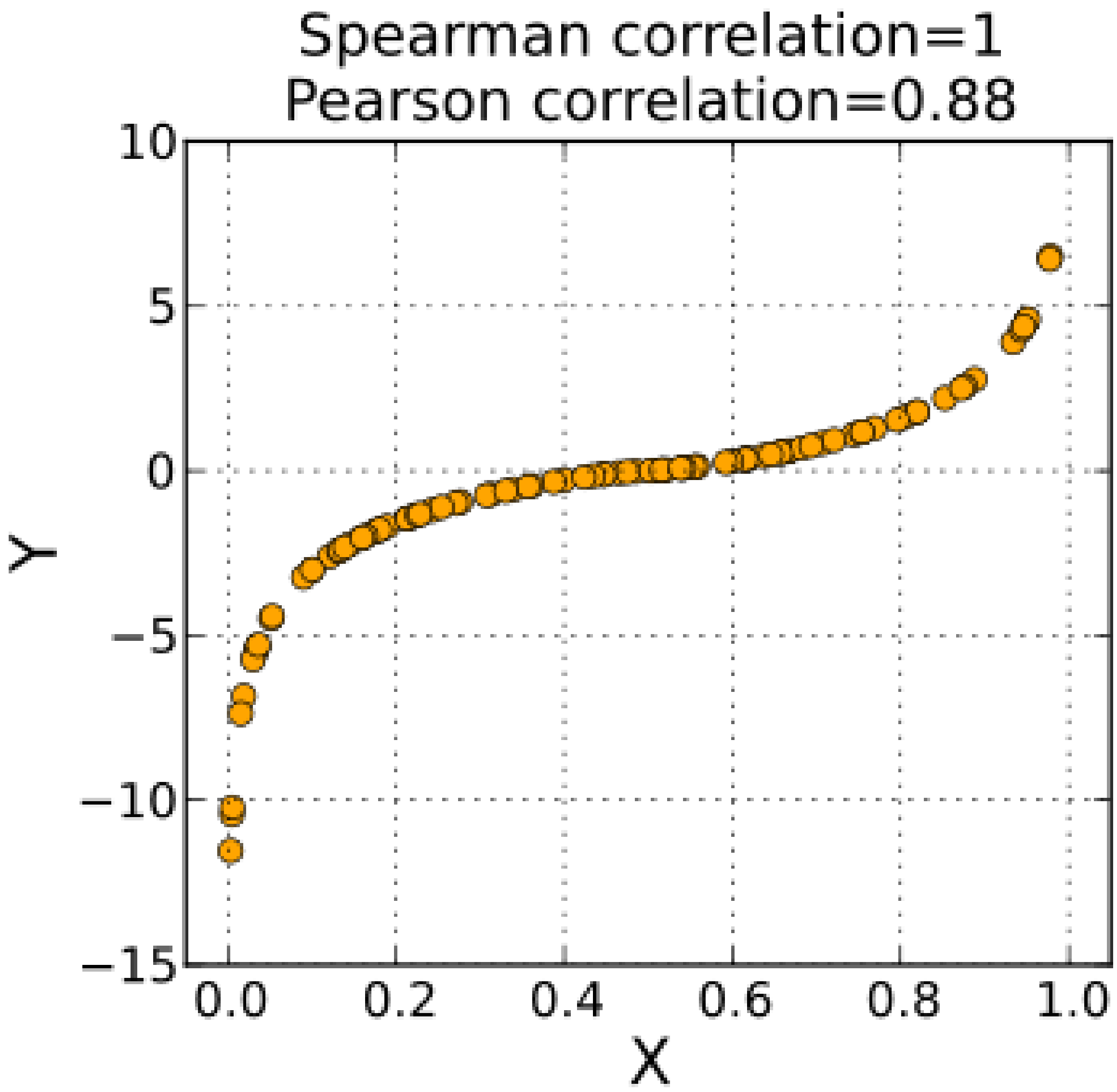
Student who reads better can run faster ??

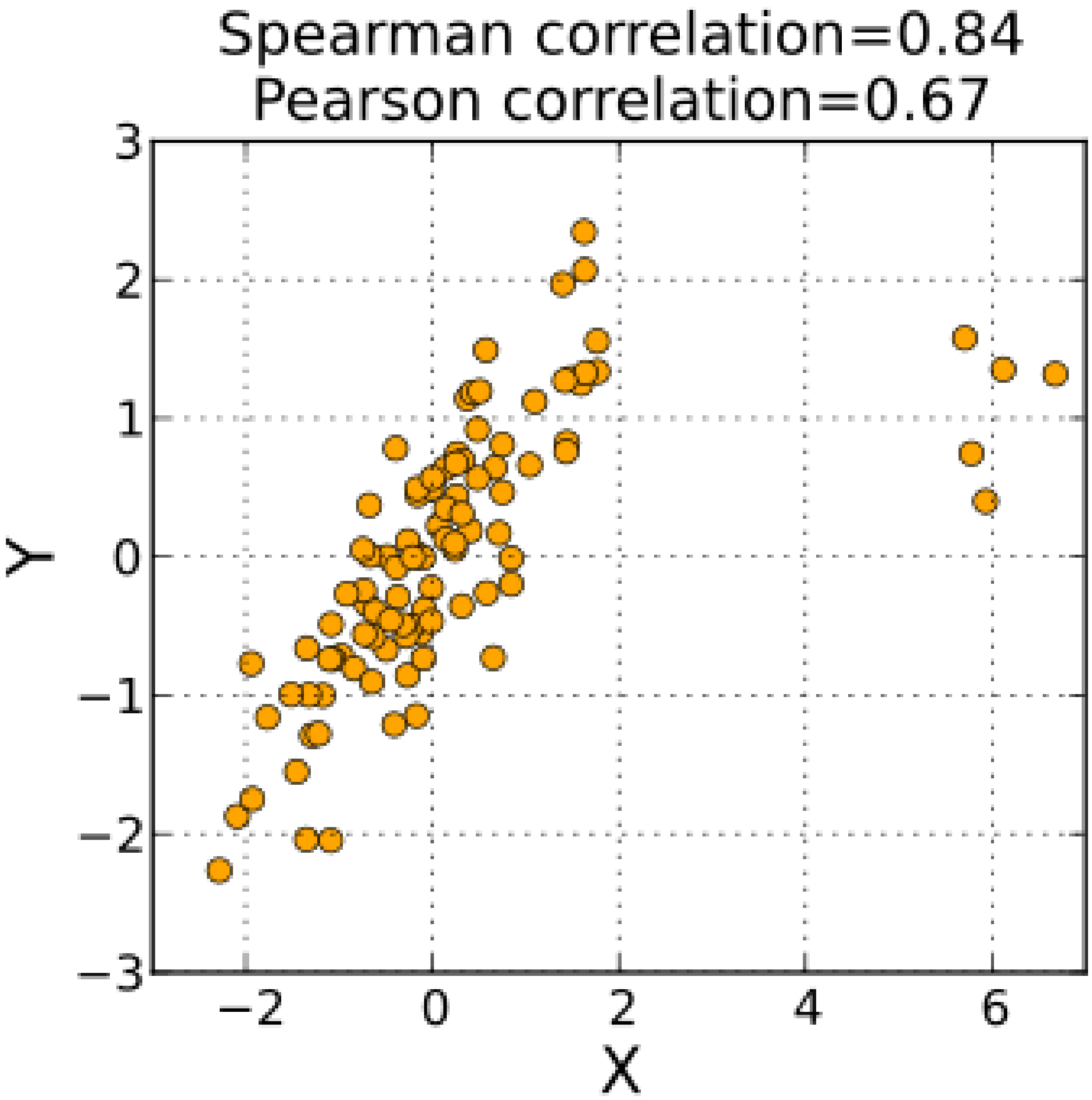
Age may be the key here...

6th graders who can read better than **1st graders**, of course can run faster than 1st graders ☺.



Pearson-r = 0.81


Spearman correlation=1
Pearson correlation=0.88
10
5
0
−5
−10
−15
Y
0.0
0.2
0.4
0.6
0.8
1.0
X


Spearman correlation=0.84
Pearson correlation=0.67
Y
X
3
2
1
0
−1
−2
−3
−2
0
2
4
6