# Describing Relationships: Scatterplots and Correlation

## Chapter 14

October 31, 2012

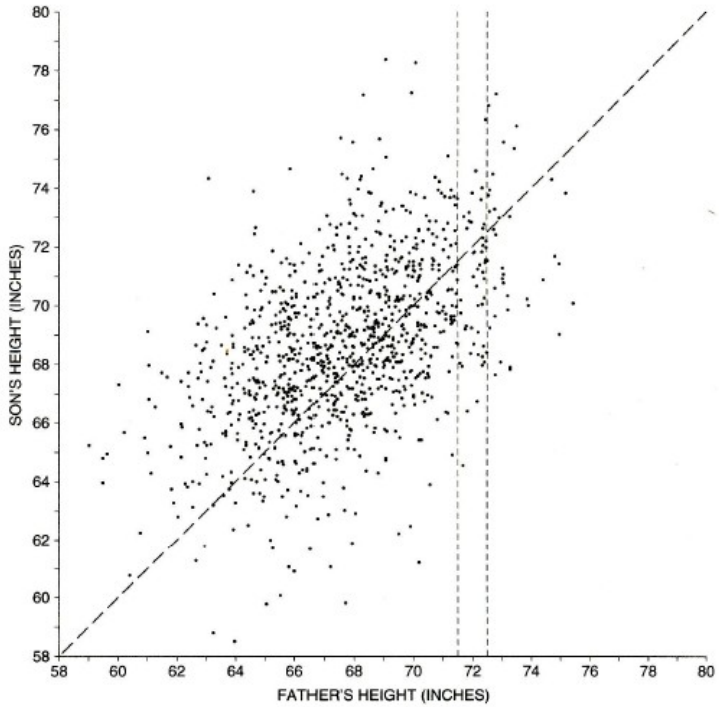
- Relationships Among Variables
- Scatter plots
- Pearson's Correlation *r*
- Examples
- Ecological Correlations

# 1.0 Relationships Among Variables

- Two quantitative variables.
- Data displays: scatterplots.
- Numerical summaries: correlation
  - concept, calculation, interpretation, cautions

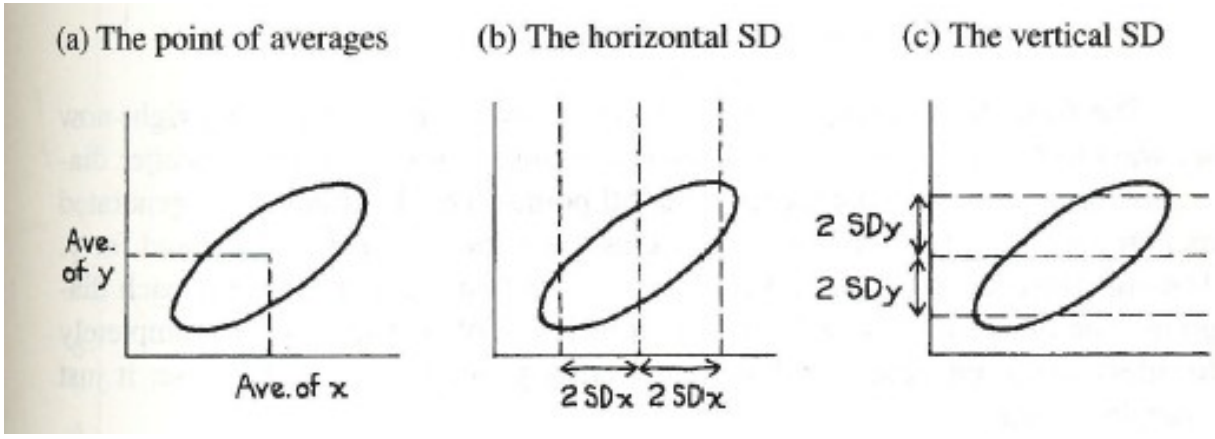
# 2.0 Scatterplots

- Heights of 1,078 fathers and sons. (Karl Pearson data)
- Shows positive association between son's and father's height. The association is not very strong.



# 2.1 Numerical Summaries

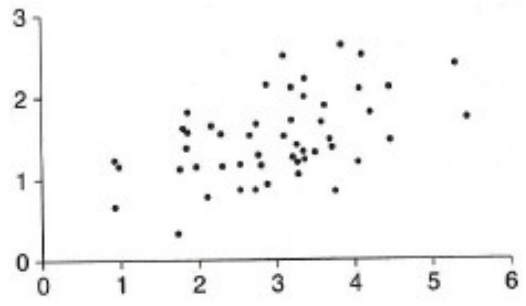
- The point of averages marks the center of the cloud.
- The horizontal and vertical standard deviations describe the spread from side to side or top to bottom.



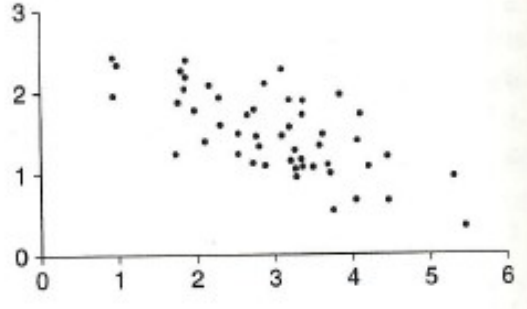
# 2.1 Example

For each scatter digram below:

1. The average of $x$ is around
   1.0 1.5 2.0 2.5 3.0 3.5 4.0
2. Same for $y$
3. The SD of $x$ is around: 0.25 0.5 1.0 1.5
4. Same for $y$
5. Is the association positive, negative or 0?




(i)

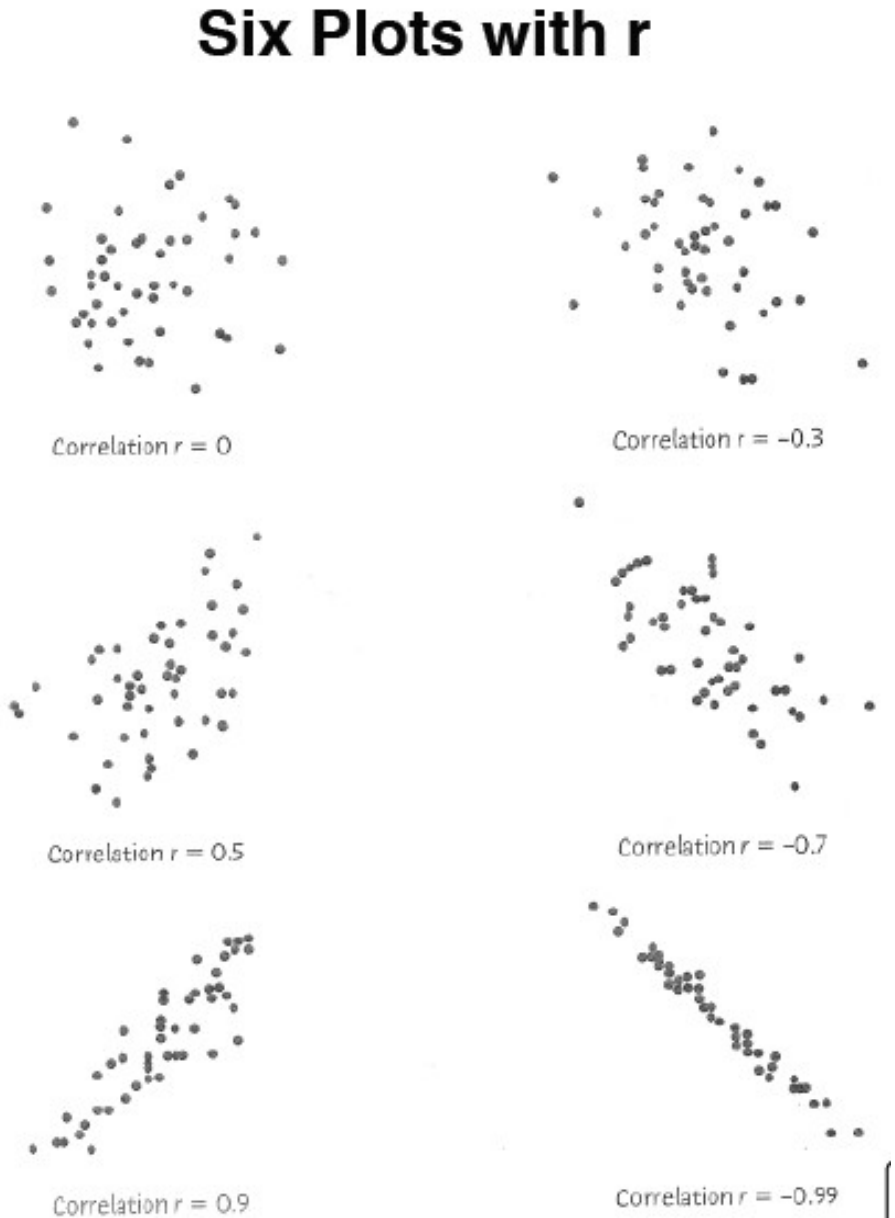



(ii)

# 2.2 Is this Enough?

- All scatter plots on next slide have the same point of averages and horizontal and vertical standard deviations.
- However, the strength of association clearly varies.

# 2.2 Is this Enough?



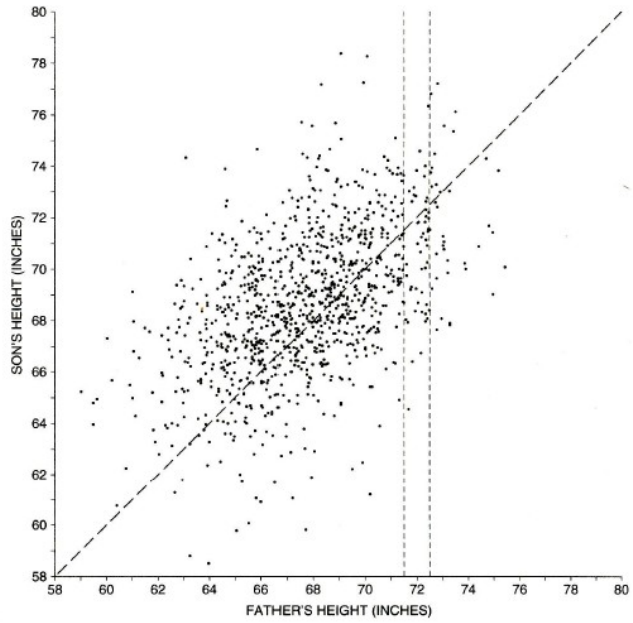

(From 221, by Prof. Morita)

# 3.0 Pearson's Correlation *r*

- Pearson's correlation coefficient takes a value between -1 and $+1$.
- A positive correlation means the cloud slopes up.
- A negative correlation means the cloud slopes down.
- A large correlation means the points cluster more tightly around a line.

# 3.1 Example

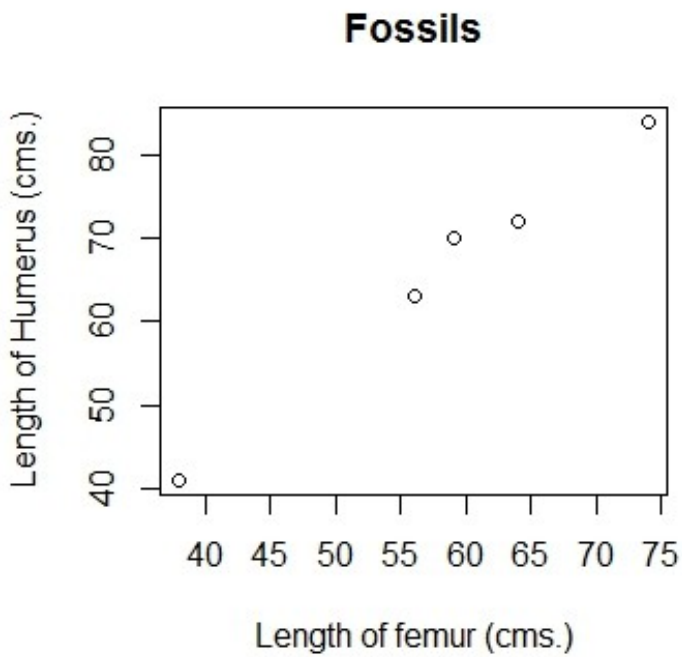
- For the father-son data, the correlation is around -0.3, 0, 0.5 or 0.8?
- If you took only the fathers' who were taller than 6 feet, and their sons, would the correlation between the heights be around -0.3, 0, 0.5, 0.8?



# 3.2 Calculating $r$

Table: Length of femur and humerus in 5 Archaeopteryx fossils

| | | | | | |
|---|---|---|---|---|---|
| Femur ($x$): | 38 | 56 | 59 | 64 | 74 |
| Humerus ($y$): | 41 | 63 | 70 | 72 | 84 |



# 3.2 Calculating *r*

- Step 1: Calculate the mean and S.D. for $x$ and $y$.

| | |
|---|---|
| Femur: | Avg. = 58.20 c.m., S.D. = 13.20 c.m. |
| Humerus: | Avg. = 66.0 c.m., S.D. = 15.89 c.m. |

- Step 2: Calculate standard scores for $x$ and $y$.

| x | Standard Score | y | Standard Score |
|---|---|---|---|
| 38 | (38-58.2)/13.20 = -1.530 | 41 | (41-66.0)/15.89 = -1.573 |
| 56 | (56-58.2)/13.20 = -0.167 | 63 | (63-66.0)/15.89 = -0.189 |
| 59 | (59-58.2)/13.20 = 0.061 | 70 | (70-66.0)/15.89 = 0.252 |
| 64 | (64-58.2)/13.20 = 0.439 | 72 | (72-66.0)/15.89 = 0.378 |
| 74 | (74-58.2)/13.20 = 1.197 | 84 | (84-66.0)/15.89 = 1.133 |

# 3.2 Calculating $r$

- Step 3: Average the products of these standard scores. Divide by $n - 1$ instead of $n$.

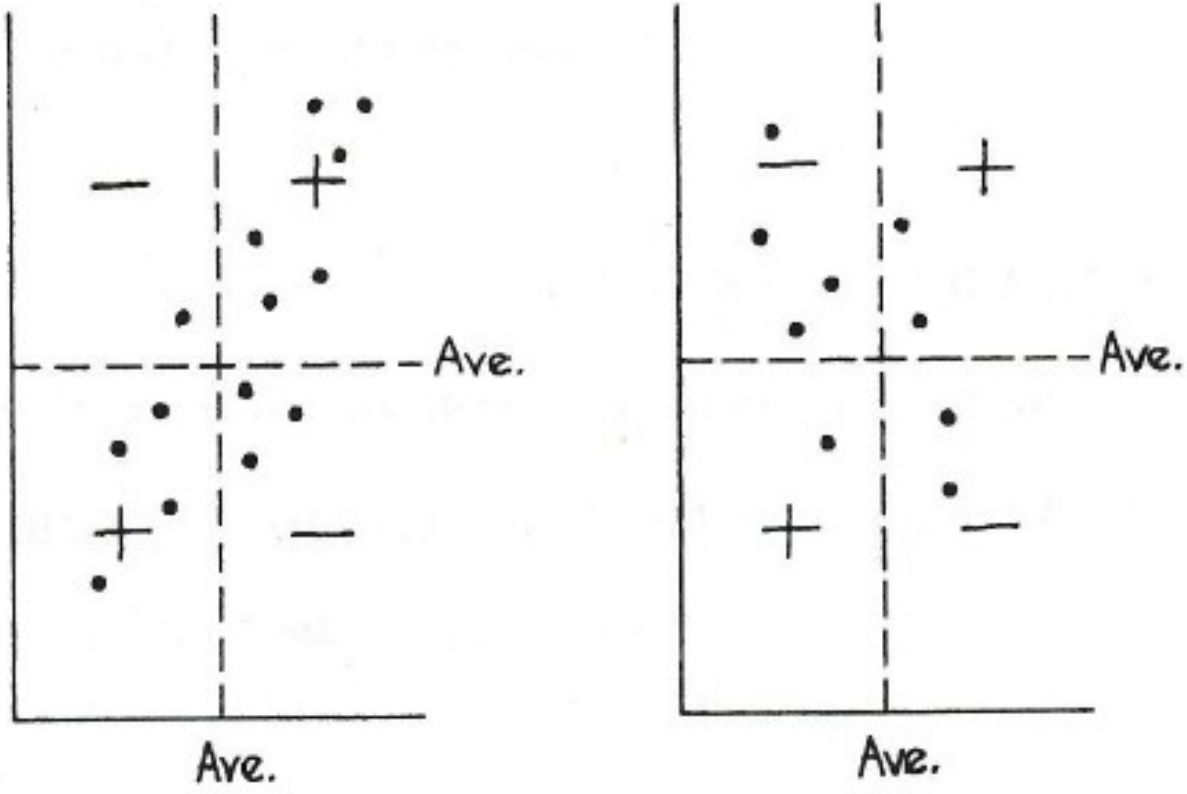
$$r = \tfrac{1}{4}\left[(-1.530)(-1.573) + (-0.167)(-0.189) + (0.061)(0.252) + (0.439)(0.378) + (1.197)(1.133)\right]$$

$$r = \tfrac{1}{4}\left[(2.4067 + 0.0316 + 0.0154 + 0.1659 + 1.3562)\right]$$

$$= 0.994$$

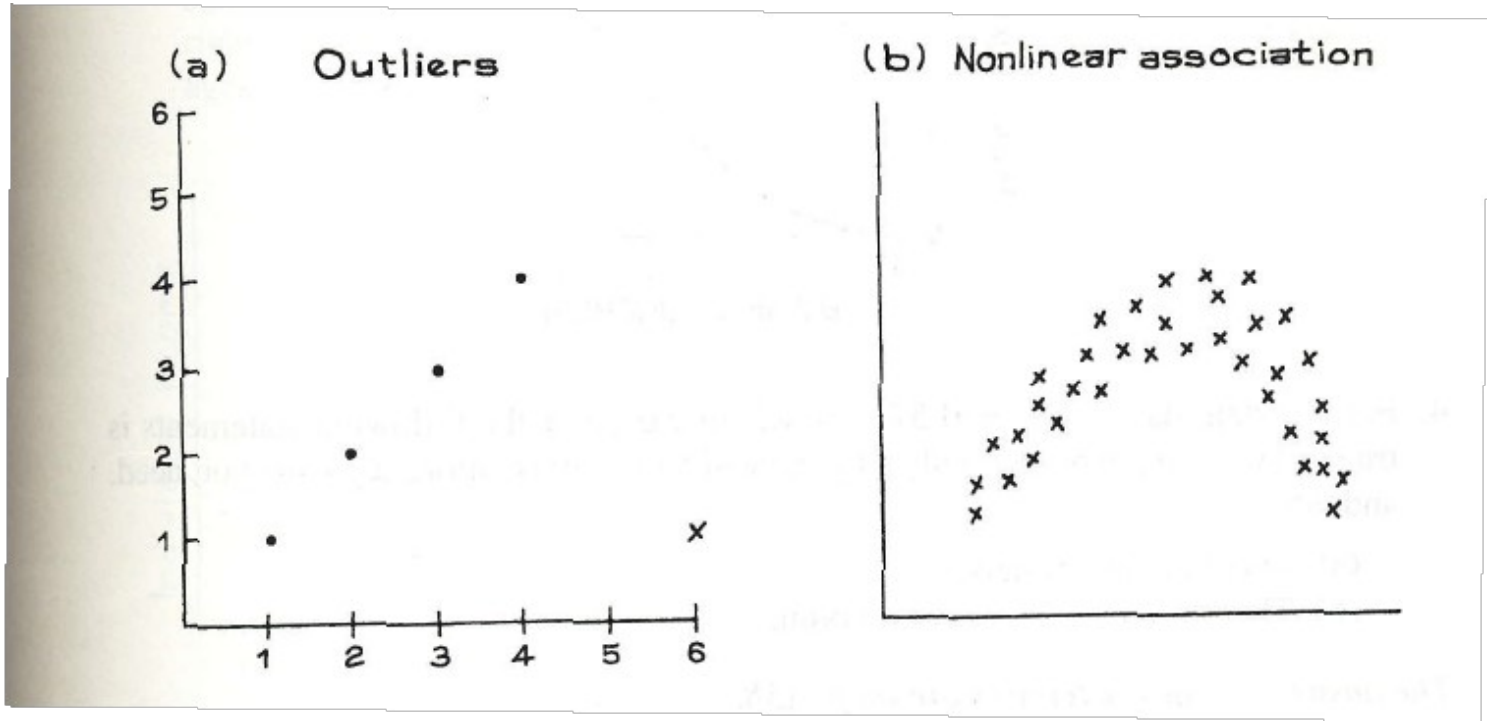
# 3.3 How *r* works

$r$= avg. [ (X in standard scores) $\times$ (Y in standard scores) ].



# 3.4 Some Cases when $r$ Fails

- $r$ is useful for FOOTBALL shaped diagrams.
- In other cases, it can be misleading.



# 4.0 Example

The correlation between height and weight among men age 18-74 in the U.S. is about 0.40. Say whether each of these conclusions follows.

- Taller men tend to be heavier.
- The correlation between weight and height for men age 18-74 is also about 0.4.
- Heavier men tend to be taller.
- If someone eats more and puts on 10 pounds, he is likely to get somewhat taller.

# 4.1 Example

For women age 25 and over in the U.S. in 2005, the relationship between age and education level (years of schooling completed) can be summarized as follows.

average age $\approx$ 50 years, S.D. $\approx$ 16 years,
average ed. level $\approx$ 13.2 years, S.D. $\approx$ 3.0 years, r $\approx$ -0.2.

True or false: as you get older, you become less educated. Explain.

# 4.2 Example

A sociologist is studying the relationship between suicide and literacy in 19th century Italy. He has data for each province, showing the % of literates and the suicide rate in that province. The correlation is 0.6.

1. Provinces with higher literacy rates tend to have higher suicide rates. True or false?
2. Does this give a fair estimate of the association between literacy and suicide? Explain.

# 5.0 Ecological Correlations

Ecological correlations are based on rates or averages. They often tend to over-state the strength of an association.

- The panel on the left shows income versus education for individuals from 3 states A, B, C. The correlation is moderate.
- The panel on the right shows the average for each state. The correlation between the averages is almost 1.