Chapter 14: Describing Relationships: Scatterplots and Correlation

How can we understand the relationship between two variables?

Thought Question

Do you think there is a relationship between the following quantities? If so, explain the relationship.

- *Height and weight of a person:*

- *Size and fuel efficiency of a car:*

- *Enrollment at Wayne State and proportion of women in Congress:*

Scatterplots

How can we visually represent the relationship between two variables?

Terminology

- **Response variable:** variable that measures the outcome of a study.
- **Explanatory variable:** variable that we think causes changes in the response variable.
- **Scatterplot:** a plot of the relationship between two quantitative variables measured on the same individuals.

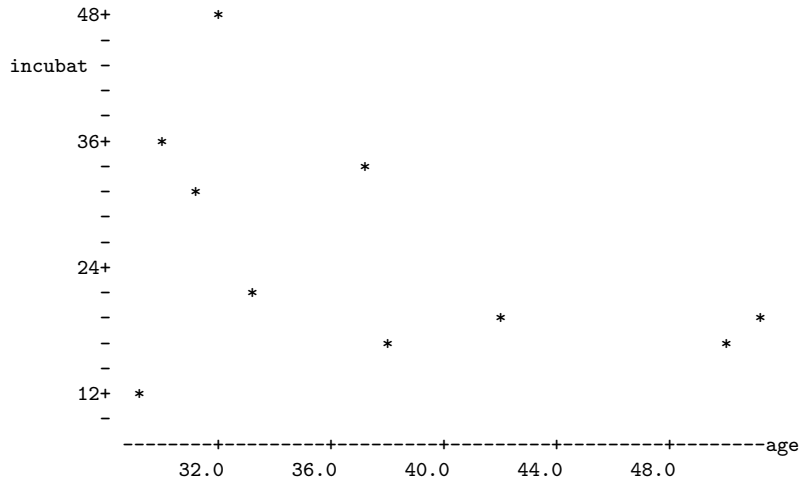
Remark: Not every situation has an explanatory and a response variable.

Example: Botulism

The following data refer to an outbreak of botulism. Each case is a person who died from botulism in the outbreak. The variables recorded are the subject's age (in years) and the incubation period (in hours).

Case	1	2	3	4	5	6	7	8	9	10	11
Age	29	37	42	38	51	30	32	33	31	32	50
Incubation	13	34	20	18	19	36	48	21	32	48	16

The following plot is a scatterplot of this data.



What is the explanatory variable?

age

What is the response variable?

incubation period

What pattern do you see in the scatterplot?

The incubation period roughly decreases as age increases up to about age 44, and then levels off.

Interpreting Scatterplots

How do we interpret the information presented in a scatterplot?

Terminology

- **Direction:** what happens to the relationship as we move from left to right.
- **Form:** shape of the relationship.
- **Strength:** how closely the points follow a clear form.

How to Interpret Scatterplots

- Look for an overall pattern and striking deviations from that pattern.
- Ignore outliers when looking for an overall pattern.
- Describe the overall pattern by the direction, form, and strength of the relationship.

Terminology

- **Positive Association:** as values of one variable tend to increase, the values of the other tend to increase.
- **Negative Association:** as values of one variable tend to increase, the values of the other tend to decrease.

Example: Index of Exposure

As part of the assessment of the consequences of radioactive contamination on human health, investigators in the 1960s calculated an "index of exposure" for each of the nine Oregon counties having frontage on either the Columbia River or the Pacific Ocean near the Hanford, WA facility of the Atomic Energy Commission. This index was based on several factors, including distance from Hanford and average distance of the population from water frontage. The cancer mortality rate (cancer mortality per 100,000 person-years from 1959 - 1964) was also determined for each of these counties.

County	Index of Exposure	Cancer Mortality Rate
Clatsop	8.34	210.3
Columbia	6.41	177.9
Gilliam	3.41	129.9
Hood River	3.83	162.3
Morrow	2.57	130.1
Portland	11.64	207.5
Sherman	1.25	113.5
Umatilla	2.49	147.1
Wasco	1.62	137.5

Make a scatterplot of this data to look for the relationship between the counties' exposure and cancer mortality rate.

What is the form, direction, and strength of the association? Are there any outliers?

The plot shows a moderately strong, positive, linear association. There are no outliers.

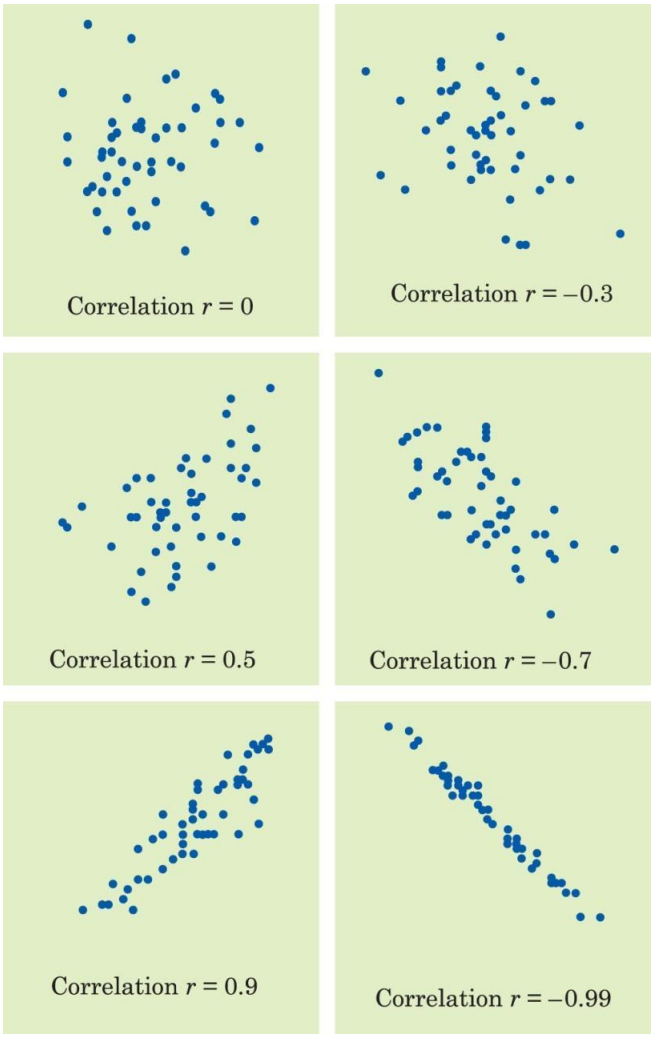
Correlation

How can we we numerically measure the strength and direction of a linear relationship?

Terminology

- **Correlation r:** describes the direction and strength of a linear relationship between two quantitative variables.

Example



How to Calculate the Correlation

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

Example: Botulism

Recall the following data on an outbreak of botulism. Each case is a person who died from botulism in the outbreak. The variables recorded are the subject's age (in years) and the incubation period (in hours).

Case	1	2	3	4	5	6	7	8	9	10	11
Age	29	37	42	38	51	30	32	33	31	32	50
Incubation	13	34	20	18	19	36	48	21	32	48	16

Determine the correlation between the age and incubation period of botulism in these subjects.

$r = -0.475$

Properties of Correlation

- Positive r indicates positive association and negative r indicates negative association.

- The value of r always falls between -1 and 1.

- The extreme values $r = -1$ and $r = 1$ occur only when the points lie exactly along a straight line.

- The correlation r has no unit of measurement; it is a dimensionless number between -1 and 1. It doesn't change if we change the units of measurement.

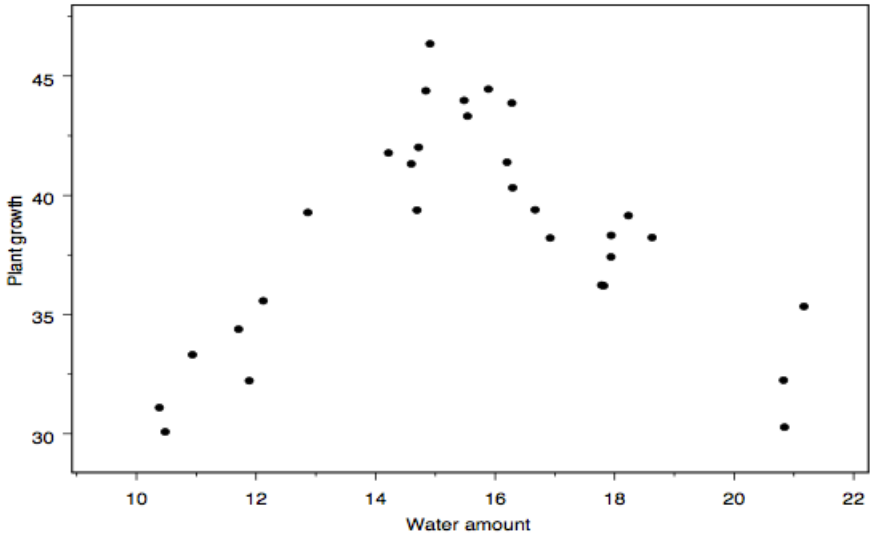
- Correlation ignores the distinction between explanatory and response variables.

- Correlation measures only the strength of **linear association** between the two variables.

- Correlation is strongly affected by outliers.

Example: Water and Plant Growth

Researchers are trying to understand the relationship between the amount of water applied to plots (measured in cm) and total plant growth (measured in cm). A sample of $n = 30$ plots is taken from different parts of a field. The data from the sample are plotted below.



Using statistical software, the correlation between plant growth and water amount is computed to be $r = 0.088$.

Do plant growth and water amount have a strong linear relationship?

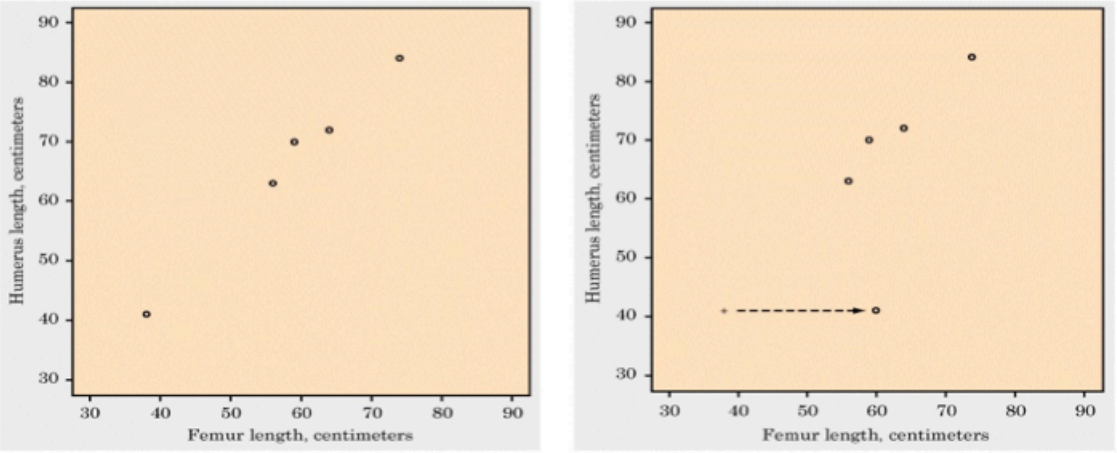
No

It is correct to conclude that plant growth and water amount are not related?

No - they have a strong quadratic relationship.

Example: Archaeopteryx Bones

Depicted here are the lengths of the femur (leg bone) and the humerus (upper arm bone) for five archaeopteryx. Plot (a) represents the original data and Plot (b) changes one data point.



(a) No outlier. (b) Outlier.

In (a), the correlation $r = 0.994$ and in (b), $r = 0.640$.

How does the outlier affect the correlation?

It significantly decreases the correlation.