Chapter 7: Scatterplots, Association, and Correlation

Explanatory/Response Variables

The explanatory variable attempts to "explain" the response variable. You would use the explanatory variable to predict the value of the response variable. In a scatterplot, the explanatory variable is always graphed on the horizontal axis.

1. Identify the explanatory and response variables, and state whether they are categorical or quantitative.
   a. Researchers measure the heights of children at age 6 and again at age 16.
   height (age 6)
   expl: ~~age~~, quantitative, ~~years~~ inches
   resp: height (age 16), quantitative, inches
   b. A political scientist selects a large sample of registered voters, both male and female, and asks each voter whether they voted for the Democratic or Republican candidate in the last congressional election.
   expl: gender, categorical
   resp: political affiliation, categorical
   c. Breast cancer patients received one of two treatments: (1) removal of the breast or (2) removal of the tumor and lymph nodes only followed by radiation. The patients were followed to see how long they lived following surgery.
   expl: type of treatment, categorical
   resp: survival rate, quantitative, years

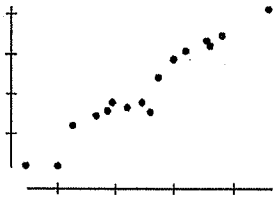
Association in Scatterplots

A scatterplot is used to graph the relationship between two quantitative variables for the same individuals. Each individual is represented on the graph by a point. Two variables have a positive association when the both increase or decrease together. Two variables have a negative association when an increase in one variable indicated a decrease in the other. Two variables have no association when the change in one variable cannot be determined from the change in the other.

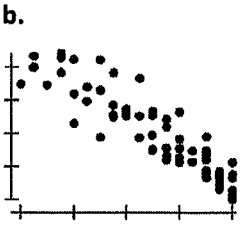
When describing scatterplots, we look for strength, direction, form, and unusual features.

Strength:

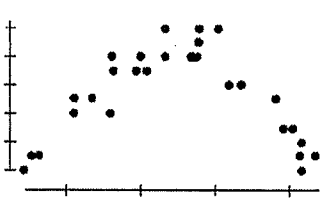
a.



b.

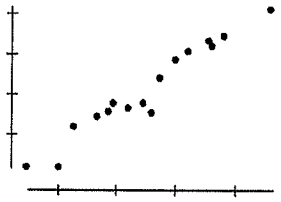


c.

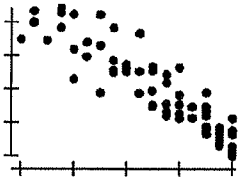


Direction:

a.



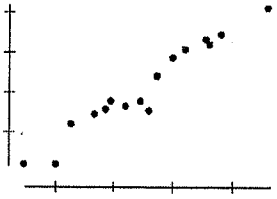
b.



c.

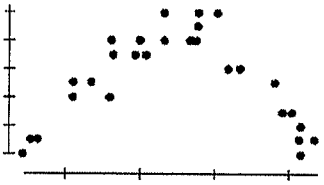
Form:

a.



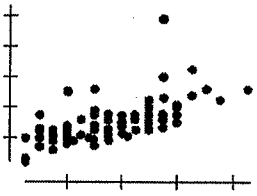
b.

c.

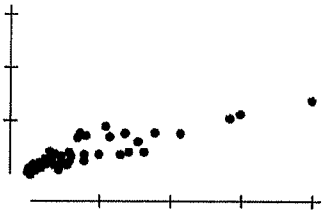


Unusual Features:

a.



b.



2. Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier would suggest different species. Here are data on the lengths in centimeters of the femur and the humerus for the specimens that preserve both bones:

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

Make a scatterplot. Do you think all five specimens come from the same species?

3. The presence of harmful insects in farm fields is detected by putting up boards covered with a sticky material and then examining the insects trapped on the board. Which colors attract insects best? Experimenters placed six boards of each of four colors in a field of oats and measured the number of cereal leaf beetles trapped.

Board Color	Insects Trapped					
Lemon Yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

a. Make a plot of the counts of insects trapped against board color (space the four colors equally on the horizontal axis).

b. Based on the data, what do you conclude about the attractiveness of these colors to the beetles?

c. What type of association exists between board color and insect count? Explain.
none, this is categorical data

Correlation

correlation measures the strength and direction of the **linear** relationship between two quantitative variables. It is possible to have a strong association but a weak correlation.

EX:

Another name for correlation is the r-value. When there is a positive association between variables, the r-value is positive. When there is a negative association between variables, the r-value is negative. When there is no association between variables, the r-value is ≈ 0.

The r-value is between -1 and 1. An r-value of 1 indicates a perfect positive linear relationship; an r-value of -1 indicates a perfect negative linear relationship.

An r-value is a standardized value and therefore has no units attached to it. Converting the units of measurement of the data values has no effect on the correlation.

The correlation of *x* with *y* is the same as the correlation of y with x.

The r-value is greatly affected by outliers. A single outlier can change the strength and direction of the correlation.

EX:

4. If women always married men who were two years older than themselves, what would be the correlation between the ages of husband and wife?

   1

5. Find the correlation between the length of the femur and humerus form specimens of Archaeopteryx (refer to #2). If the lengths of the bones were measured in inches instead of centimeters, how would the correlation change?

   $r = \frac{\sum z_x z_y}{n-1}$

6. Each of the following statements contains a blunder. In each case, explain what is wrong.
   a. "There is a high correlation between the sex of American workers and their income."

      categorical data

   b. "We found a high correlation (r = 1.09) between students' ratings of faculty teaching and ratings made by other faculty members."

      r cannot be > 1

   c. "The correlation between planting rate and yield of corn was r = 0.23 bushels."

      r has no units