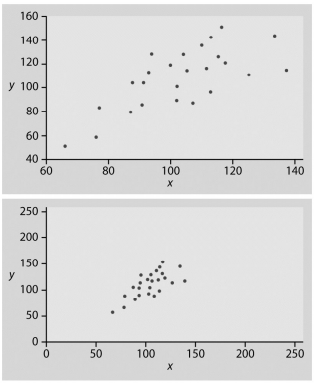
Section 2.2 - Correlation



Sometimes it is difficult to judge the strength of a relationship by eye

=> need a numerical measure of strength to supplement a scatterplot

CORRELATION (r) - measures the **strength** and **direction** of the *linear* relationship between two *quantitative* variables.

$$r = \left(\frac{1}{n-1} \right) \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $\bar{x}$ and $\bar{y}$ are the means of the *x*-values and *y*-values, respectively, and s_x and s_y are the corresponding standard deviations.

- Correlation is an "average" of the product of the standardized values of the two variables.
- The sign of *r* gives the direction of the relationship.

  r > 0 => positive association

  r < 0 => negative association
- Strength is on a scale 0 to 1.
- So *r* can range from **-1** to **+1**.

Properties of Correlation

1. Correlation makes no distinction between explanatory and response variables. The correlation between x and y is the same as the correlation between y and x.
2. Correlation requires that both variables be quantitative. We cannot compute a correlation between a categorical variable and a quantitative variable or between two categorical variables.
3. r does not change if the data are linearly transformed. For example, the correlation between height and weight is the same whether height is measured in feet or centimeters or weight is measured in kilograms or pounds. This happens because all the observations are **standardized** in the calculation of correlation.

Properties of Correlation cont.

4. The correlation r itself has no unit of measurement; it is just a number.

5. Positive r indicates positive association between the variables and negative r indicates negative association.

6. r is always between -1 and 1.
   Values of r near 0 indicate **weak linear association**.
   Values of r close to -1 or 1 indicate **strong linear association** (points fall close to a straight line).
   If r = 1 or r = -1 then **all the points fall exactly on a straight line.**

Properties of Correlation cont.

7. Correlation measures the strength of only a *linear* relationship between two variables. A relationship may be very strong, yet curved and have a correlation of zero.

8. Correlation is a **non-resistant** measure. r is **strongly** affected by outliers.

   Correlation and Regression Applet on text's website

What's wrong with the following statements?

1. At OSU there is no correlation between the ethnicity of students and their GPA.

2. The correlation between height and weight of a group of students is 2.61.

3. The correlation between height and weight of a group of students is 0.61 inches per pound.

4. The correlation between height and weight of a group of students is 0.61, so the correlation between weight and height is -0.61.

5. The correlation between height and weight of a group of students is 0.61 using inches and pounds, but converting inches to centimeters would make $r > 0.61$ (since an inch equals about 2.54 centimeters).