2 Summarizing Bivariate Data

- 2.1 The Correlation Coefficient
- 2.2 The Least-Squares Line
- 2.3 Features and Limitations of the Least-Squares Line

2.1 The Correlation Coefficient

Introduction

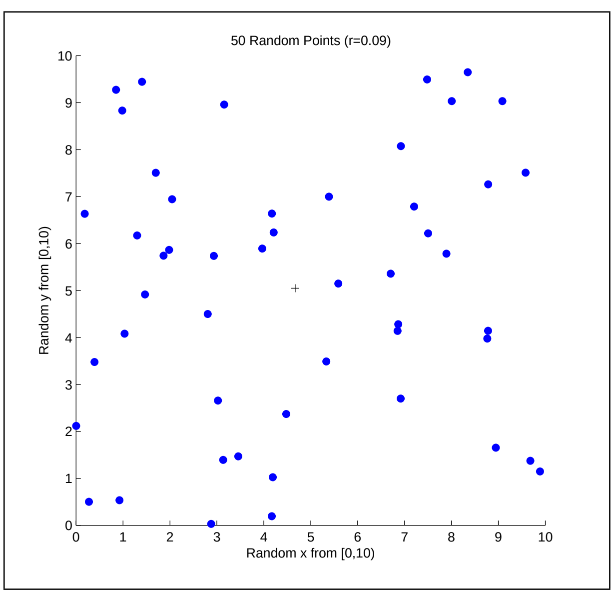
A *bivariate* data set consists of n ______________________, $(x_1, y_1), \cdots, (x_n, y_n)$.

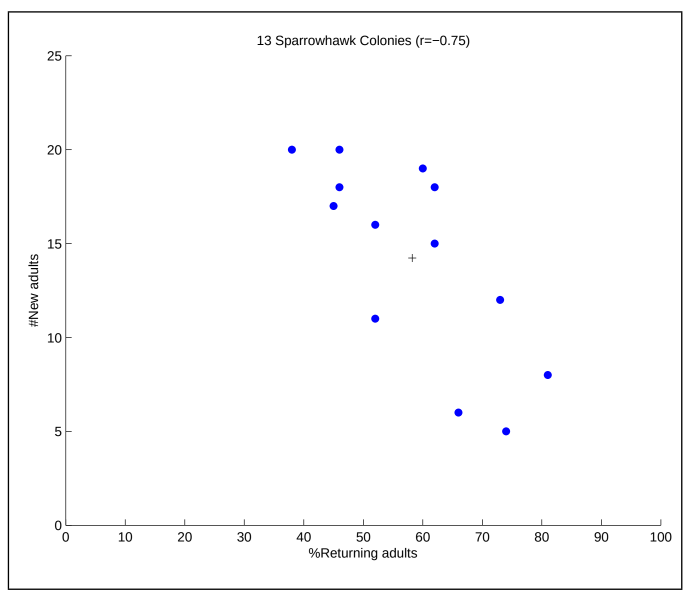
A *scatterplot* is a ________________ of a bivariate data set.

e.g. Here are data for 13 sparrowhawk colonies relating the % of adult sparrowhawks in a colony that return from the previous year and the number of new adults that join the colony:

%Returning adults	74	66	81	52	73	62	52	45	62	46	60	46	38
#New adults	5	6	8	11	12	15	16	17	18	18	19	20	20

The right-hand scatterplot, below, is from these data. It shows $\cdots$





The Correlation Coefficient

The *correlation coefficient*, r, measures the ____________ and ____________ of the linear relationship (if any) between x and y:

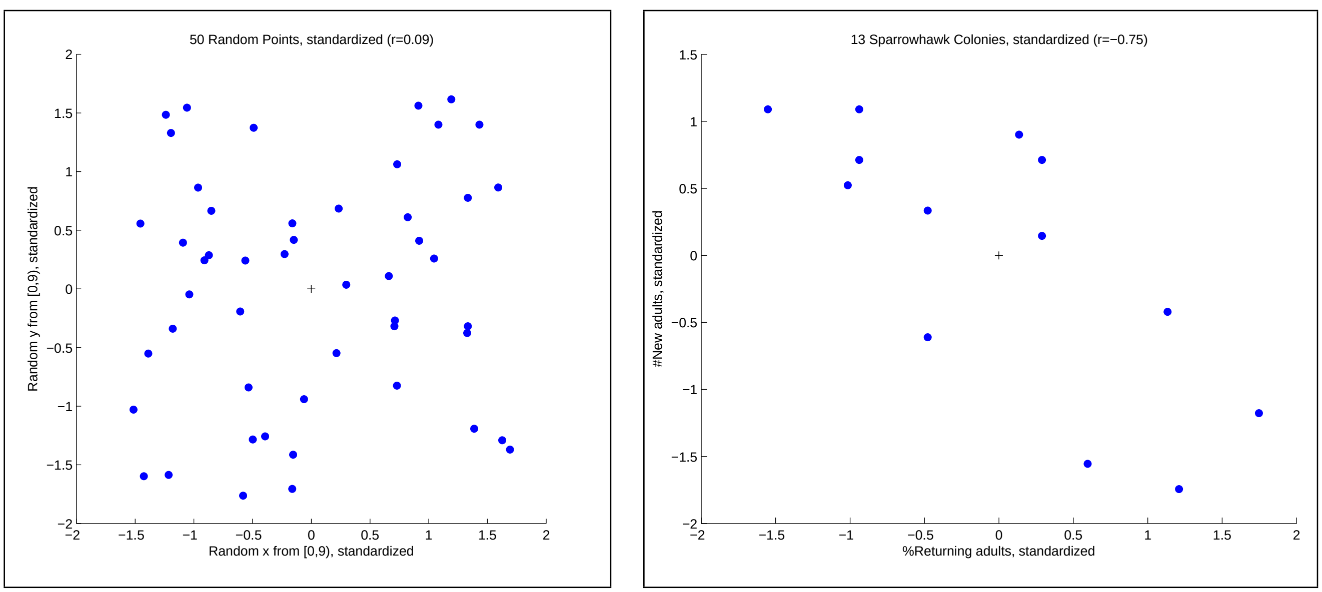
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \text{ (a form I prefer)}$$

An Informal Explanation of r

- Start with a scatterplot
- Shift origin to ____________ by subtracting $\bar{x}$ from each x_i and $\bar{y}$ from each y_i
- Rescale the x-axis by dividing each x coordinate by s_x, and rescale the y-axis by dividing each y coordinate by s_y

  Now x coordinates, $\frac{x_i - \bar{x}}{s_x}$, have mean ______ and standard deviation ______. y coordinates, $\frac{y_i - \bar{y}}{s_y}$, have the same mean and standard deviation.

- Analyze the sign of the i^{th} term in the last sum above, $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$, by quadrant:



e.g. For the sparrowhawk data, $r =$ ____________. For the random data, $r =$ ____________.

Properties of r

- $-1 \leq r \leq 1$, and

  $r = \pm 1 \implies$ data are _____________; $r \approx \pm 1 \implies$ data are __________________________

  $r \not\approx 0 \implies$ some linear relationship: x and y are *correlated*

  $r > 0 \implies$ slope of line is _____________

  $r < 0 \implies$ slope of line is _____________

  $r \approx 0 \implies$ no linear relationship: x and y are ____________________

- r doesn't distinguish between _______ and _______

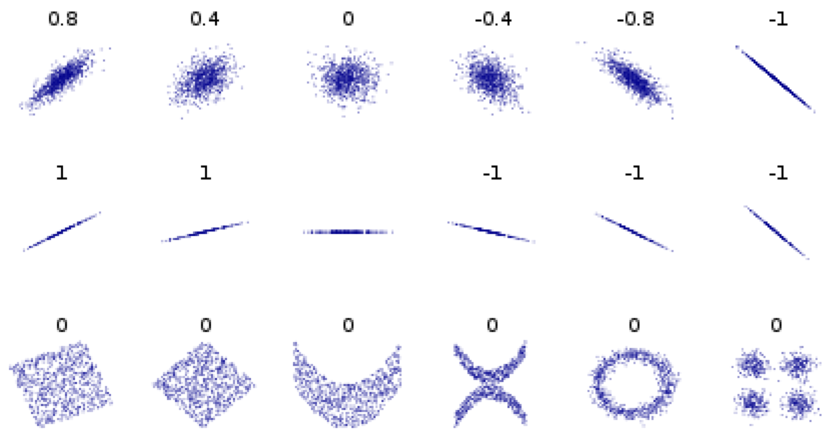
- r doesn't depend on _____________ or _____________

Cautions

- r measures strength of a *linear* relationship; check scatterplot to avoid using r for a ____________

  e.g. The data { (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4) } fit ____________, but $r = 0$ because the data have no ____________ relationship (draw).

  e.g. (from `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`)



- r is not resistant to the influence of ____________: don't use it for a data set with ____________

  e.g. Adding $(0, 0)$ to the sparrowhawk data changes r to ____________.

- Correlation does not imply causation:

  A __________________ (or *lurking*) *variable* is one ____________ under consideration that correlates with both the independent and dependent variables of interest.

  e.g.

  - Increasing ice cream sales are correlated with increasing ____________ rates. Does ice cream cause ____________? ______

    The confounding variable is ________________________.
  - Sleeping with shoes on is correlated with ____________________________________.
    Does sleeping with shoes on cause ____________? ______

    The confounding variable is ________________________.
  - A student wishing to understand the cause of ____________ drank, on successive nights, nothing but ...

  If either the independent variable under study, or a correlated confounding variable, affects the dependent variable, then both will seem to by the (___________) criterion of correlation.

The Least-Squares Regression Line

The *least-squares regression line* is the line that __________________ the data (according to a reasonable criterion). We'll study its basics in §2.2-2.3, and we'll use it for inference in Chapter 8.