Introduction to Probability and Statistics Fourteenth Edition

Chapter 3
Describing Bivariate Data

3.1 Describing bivariate categorical data

Bivariate Data

- When two variables are measured on a single experimental unit, the resulting data are called **bivariate data**.
- You can describe each variable individually, and you can also explore the **relationship** between the two variables.
- Bivariate data can be described with
  - **Graphs**
  - **Numerical Measures**

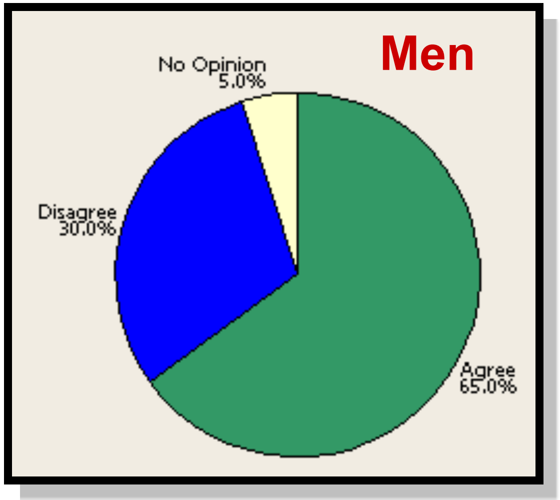
Graphs for Qualitative Variables

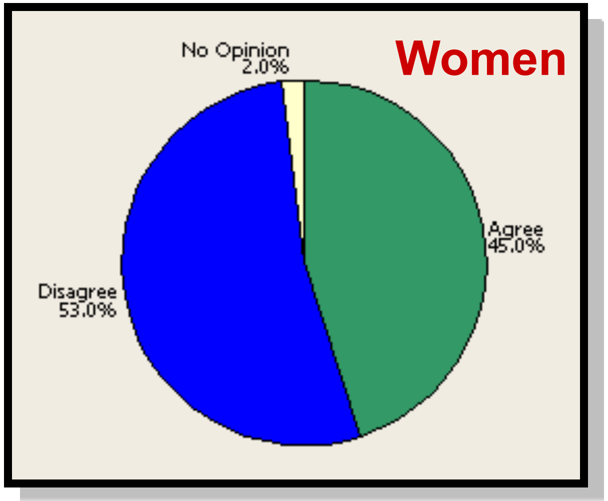
- When at least one of the variables is qualitative, you can use comparative pie charts or bar charts.

Do you think that men and women are treated equally in the workplace?

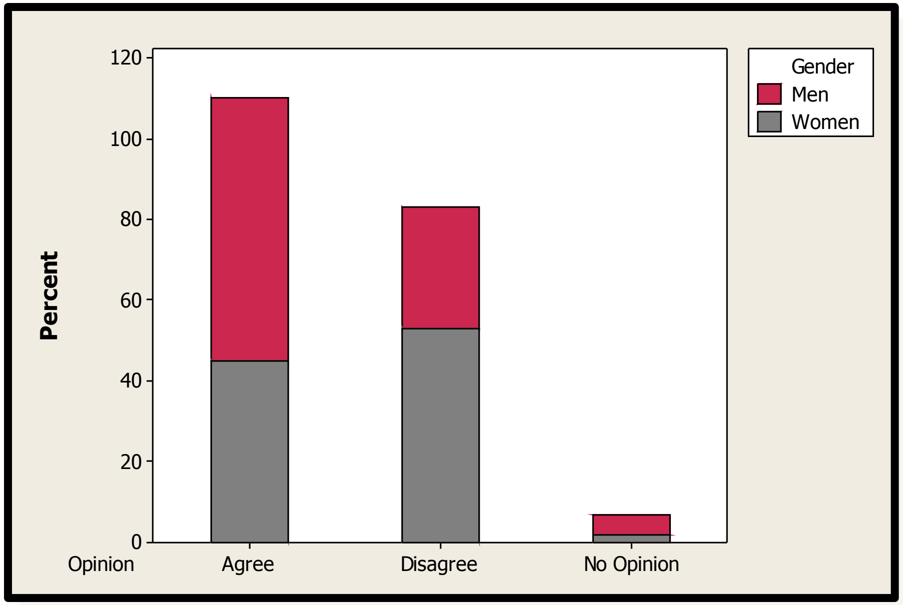
Variable #1 = Opinion

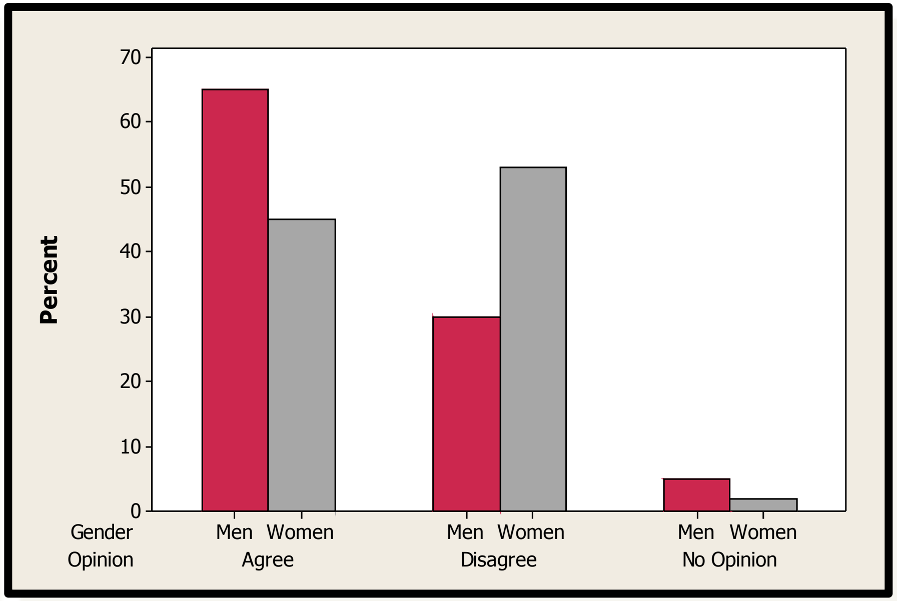
Variable #2 = Gender





Comparative Bar Charts





- Stacked Bar Chart
- Side-by-Side Bar Chart

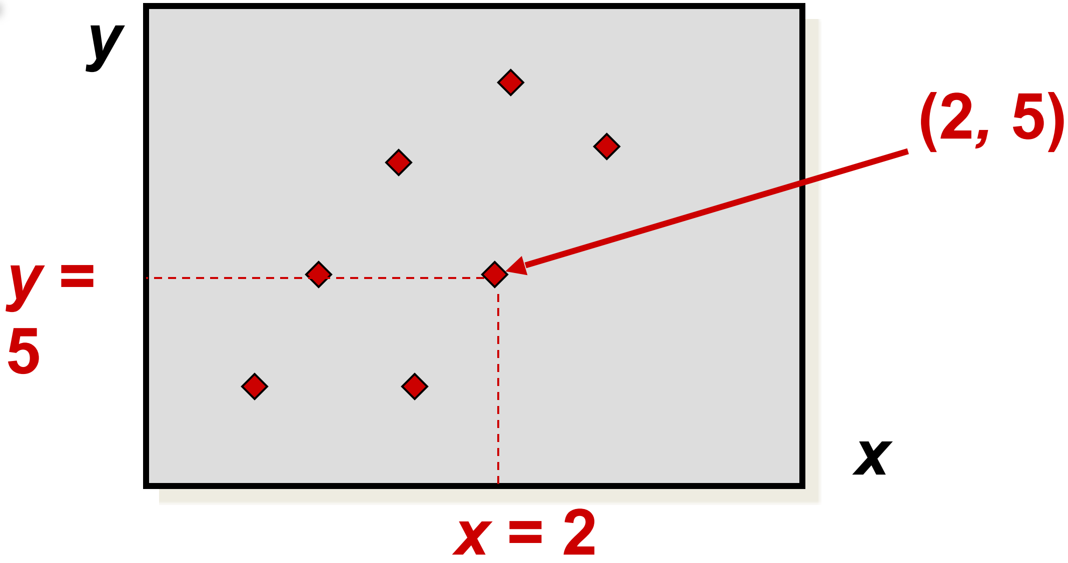
Describe the relationship between opinion and gender:

More women than men feel that they are not treated equally in the workplace..

3.2 Describing bivariate quantitative data

Two Quantitative Variables

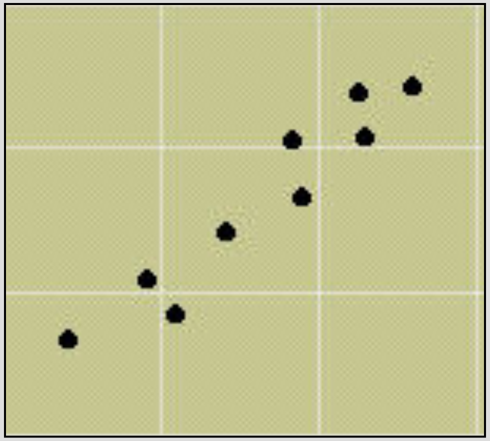
When both of the variables are quantitative, call one variable *x* and the other *y*. A single measurement is a pair of numbers (*x*, *y*) that can be plotted using a two-dimensional graph called a **scatterplot.**



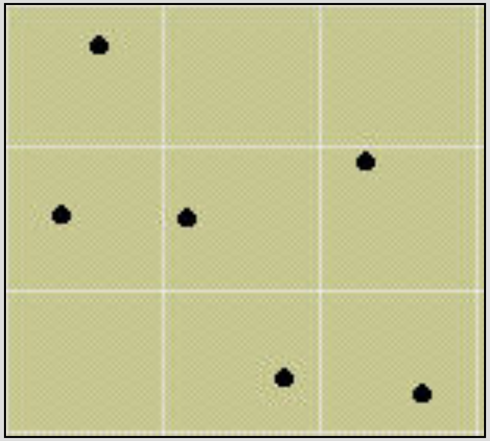
Describing the Scatterplot

- What **pattern** or **form** do you see?
  - Straight line upward or downward
  - Curve or no pattern at all
- How **strong** is the pattern?
  - Strong or weak
- Are there any **unusual observations**?
  - Clusters or outliers

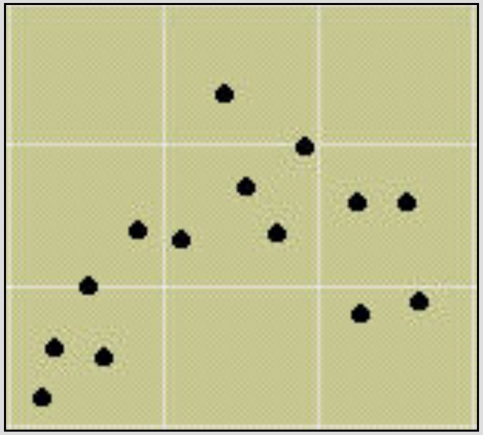
Examples



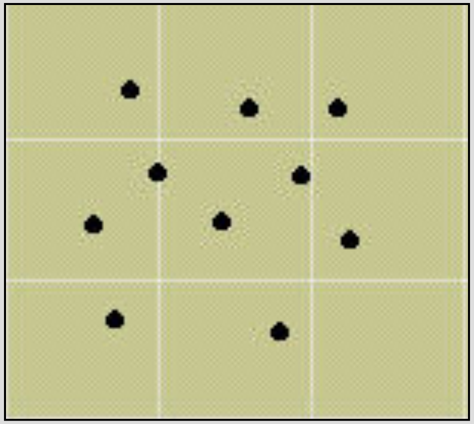
Positive linear - strong



Negative linear -weak



Curvilinear



No relationship

Numerical Measures for Two Quantitative Variables

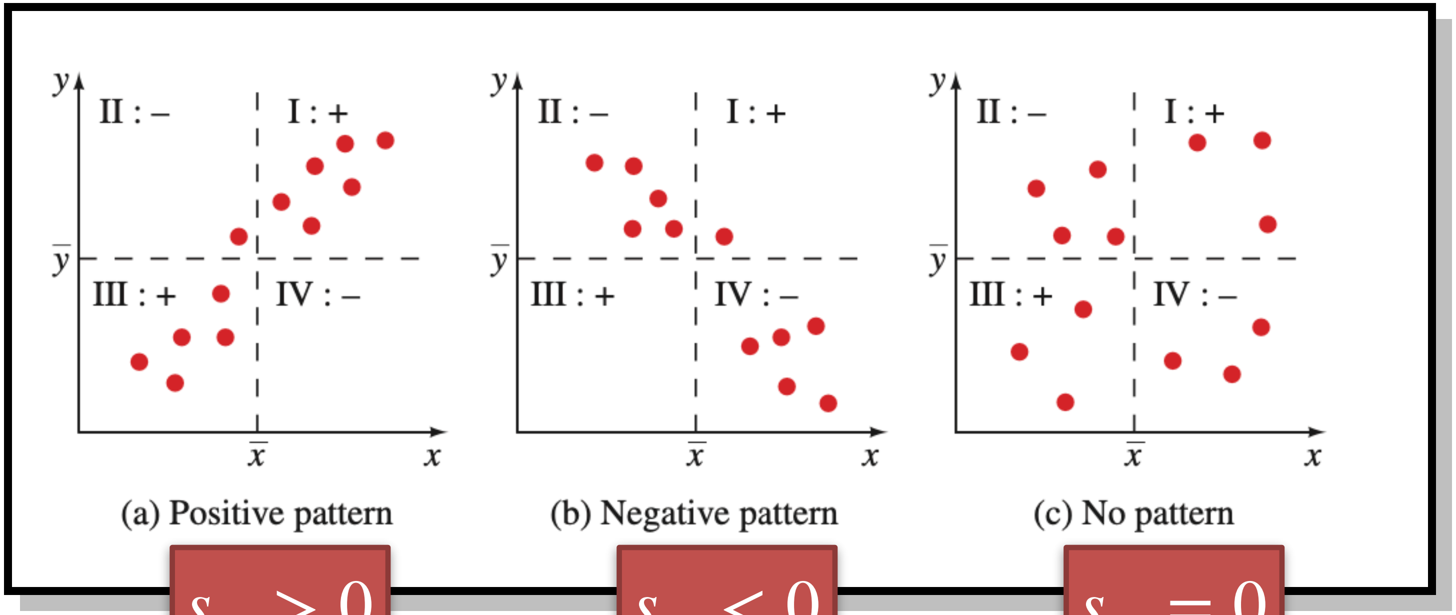
- Assume that the two variables *x* and *y* exhibit a **linear pattern** or **form**.
- The **covariance** between *x* and *y* is

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

- The covariance equals
-

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{(n-1)}$$

Fig 3.7 The signs of the cross-products $(x_i - \bar{x})(y_i - \bar{y})$ in the covariance formula



(a) Positive pattern

(b) Negative pattern

(c) No pattern

$s_{xy} > 0$

$s_{xy} < 0$

$s_{xy} = 0$

The Correlation Coefficient

- The strength and direction of the relationship between *x* and *y* are measured using the **correlation coefficient, *r***.

$$\boxed{r = \frac{s_{xy}}{s_x s_y}}$$

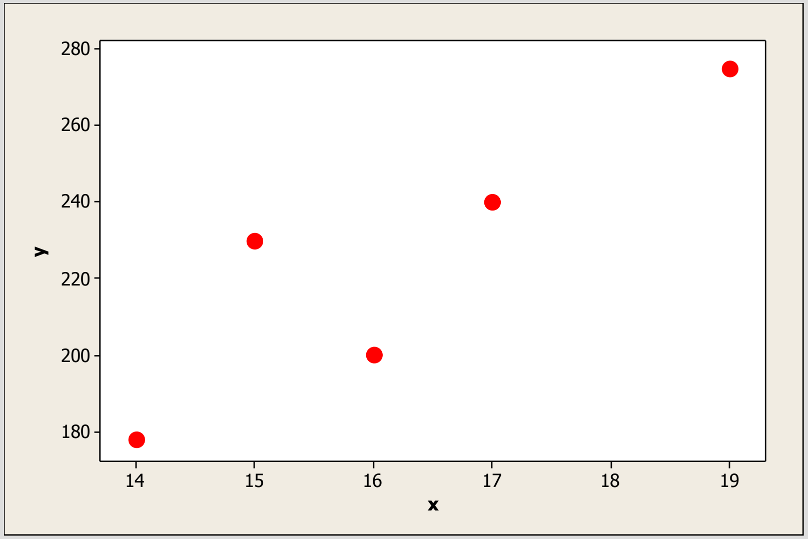
s_x = standard deviation of the *x*'s

s_y = standard deviation of the *y*'s

Example

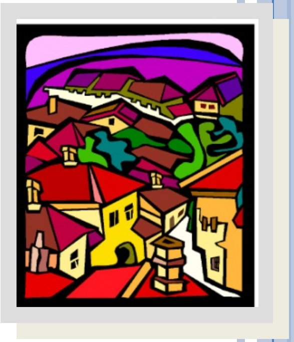
- Living area *x* and selling price *y* of 5 homes.

Residence	1	2	3	4	5
x (thousand sq ft)	14	15	17	19	16
y ($000)	178	230	240	275	200



- The scatterplot indicates a positive linear relationship.

Example



x	y	xy
14	178	2492
15	230	3450
17	240	4080
19	275	5225
16	200	3200
81	1123	18447

Calculate

$$\bar{x} = 16.2 \qquad s_x = 1.924$$

$$\bar{y} = 224.6 \qquad s_y = 37.360$$

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

$$= \frac{18447 - \frac{(81)(1123)}{5}}{4} = 63.6$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{63.6}{1.924(37.36)} = .885$$

INTERPRETING *r*

- $-1 \le r \le 1$ Sign of *r* indicates direction of the linear relationship.
- $r \approx 0$ Weak relationship; random scatter of points
- $r \approx 1$ or -1 Strong relationship; either positive or negative
- $r = 1$ or -1 All points fall exactly on a straight line.

The Regression Line

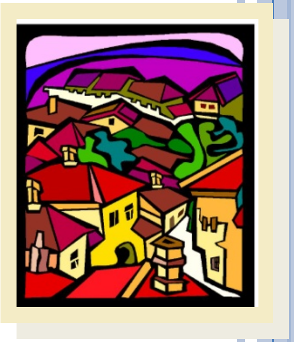
- Sometimes *x* and *y* are related in a particular way—the value of *y* **depends** on the value of *x*.
  - *y* = dependent variable
  - *x* = independent variable
- The form of the linear relationship between *x* and *y* can be described by fitting a line as best we can through the points. This is the **regression line,**

$$\boldsymbol{y = a + bx.}$$

  - *a* = *y*-intercept of the line
  - *b* = slope of the line

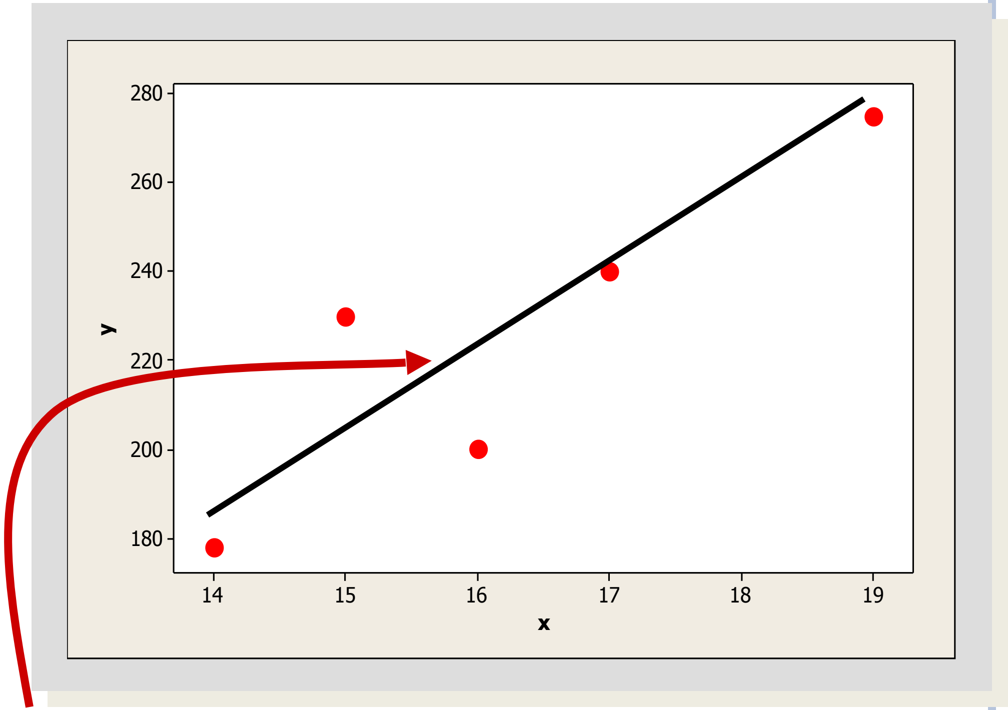
The Regression Line

- To find the slope and *y*-intercept of the best fitting line, use:



$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$



- The least squares
- regression line is $\boxed{y = a + bx}$

Example

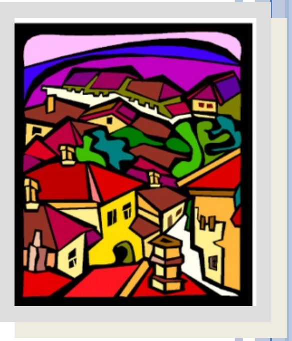
x	*y*	*xy*
14	178	2492
15	230	3450
17	240	4080
19	275	5225
16	200	3200
81	1123	18447

Recall

$$\bar{x} = 16.2 \qquad s_x = 1.9235$$

$$\bar{y} = 224.6 \quad s_y = 37.3604$$

$$r = .885$$

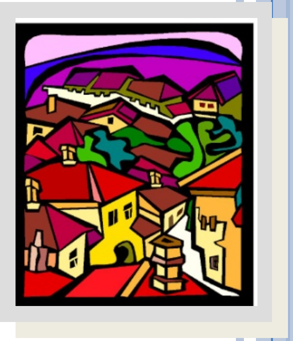


$$b = r\frac{s_y}{s_x} = (.885)\frac{37.3604}{1.9235} = 17.189$$

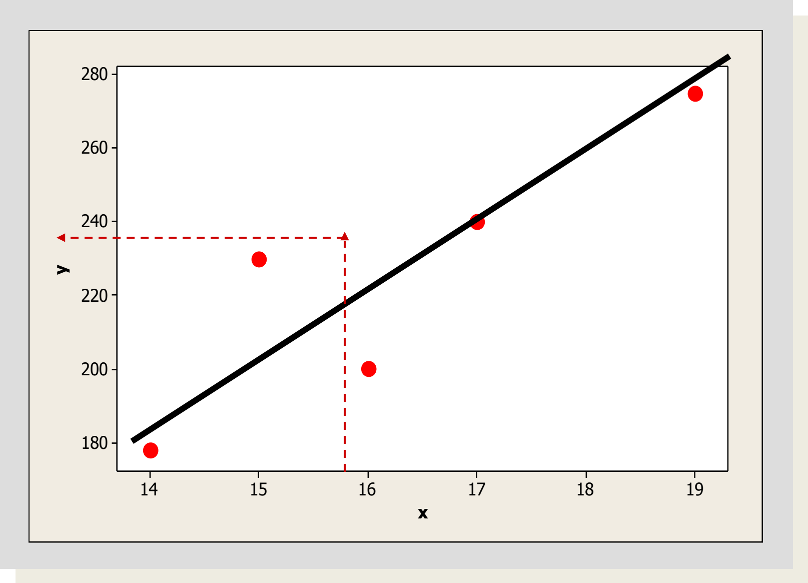
$$\mathrm{a} = \bar{y} - b\bar{x} = 224.6 - 17.189(16.2) = -53.86$$

$$\text{RegressionLine}: y = -53.86 + 17.189x$$

Example



- Predict the selling price for another residence with 1600 square feet of living area.



Predict: $y = -53.86 + 17.189x$

$= -53.86 + 17.189(16) = 221.16 \text{ or } \$221,160$

Key Concepts

I. Bivariate Data

1. Both qualitative and quantitative variables
2. Describing each variable separately
3. Describing the relationship between the variables

II. Describing Two Qualitative Variables

1. Side-by-Side pie charts
2. Comparative line charts
3. Comparative bar charts
   - ✓ Side-by-Side
   - ✓ Stacked
4. Relative frequencies to describe the relationship between the two variables.

Key Concepts

III. Describing Two Quantitative Variables

1. Scatterplots
   - ✓ Linear or nonlinear pattern
   - ✓ Strength of relationship
   - ✓ Unusual observations; clusters and outliers
2. Covariance and correlation coefficient
3. The best fitting line
   - ✓ Calculating the slope and *y*-intercept
   - ✓ Graphing the line
   - ✓ Using the line for prediction