Ch3. Describing Bivariate Data

Sometimes the data that are collected consist of observations for two variables on the same experimental unit. Special techniques that can be used in describing these variables will help you identify possible relationships between them.

윤 연 옥

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**

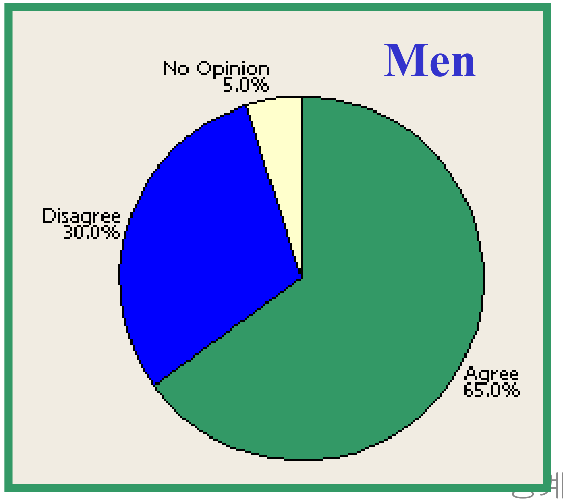
3.1 Describing **Bivariate Categorical Data**

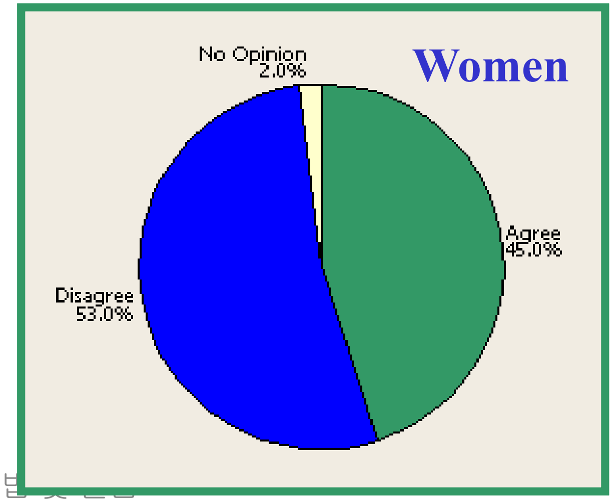
Graphs for Qualitative Variables

- When at least one of the variables is qualitative, you can use comparative pie charts or bar charts.
- Question: Do you think that men and women are treated equally in the workplace?

  - variable no. 1 : Opinion

  - variable no. 2 : Gender

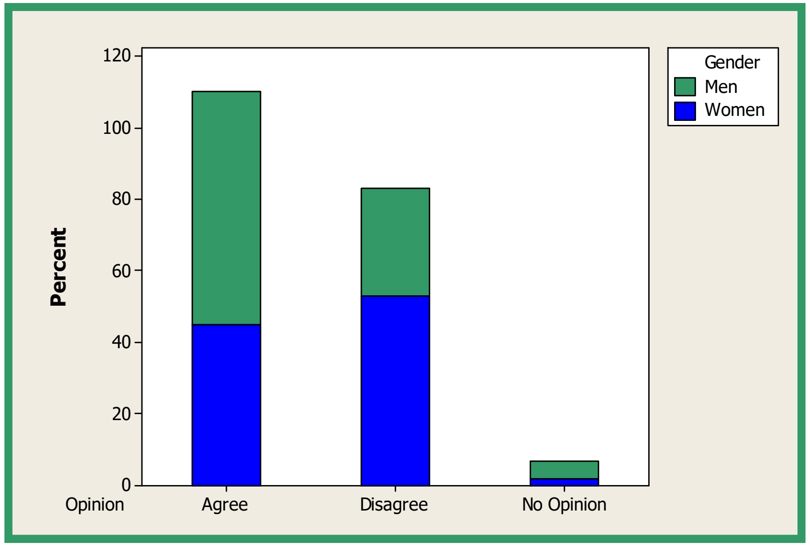




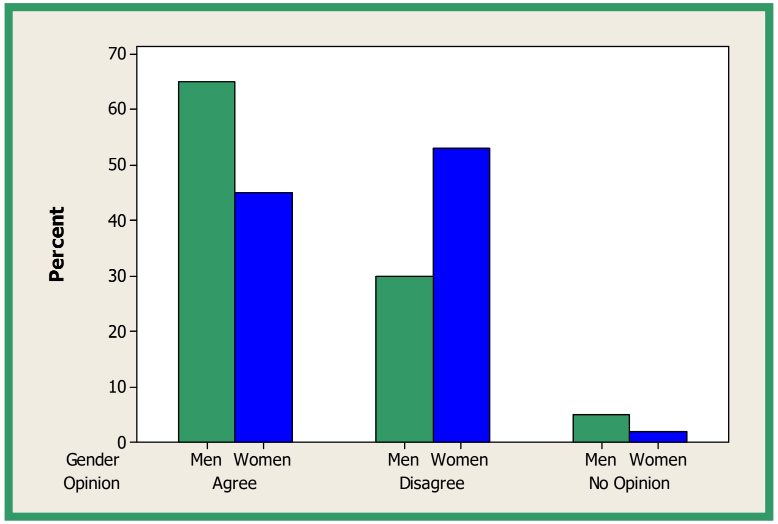
3.1 Describing **Bivariate Categorical Data**

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.

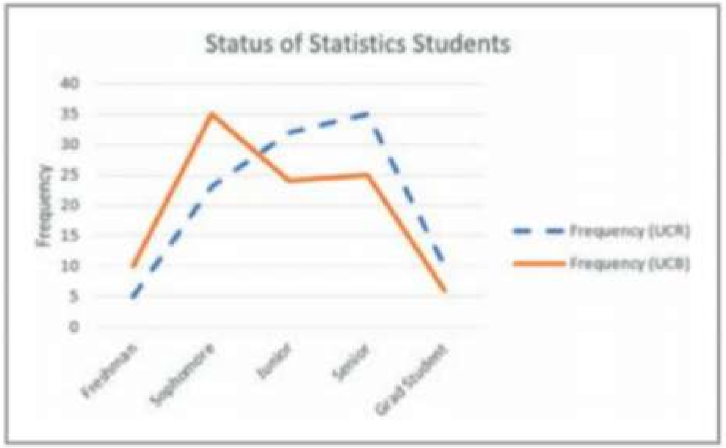
Comparative Line and bar Charts

Data: Status of students in Statistics Class at UCR(105) and UCB(100)

	Freshman	Sophomore	Junior	Senior	Grad Student
Frequency (UCR)	5	23	32	35	10
Frequency (UCB)	10	35	24	25	6

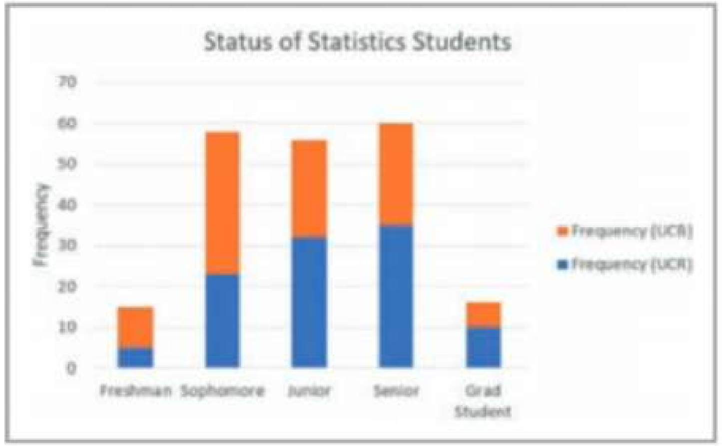
Line chart

(a)



Stacked bar chart

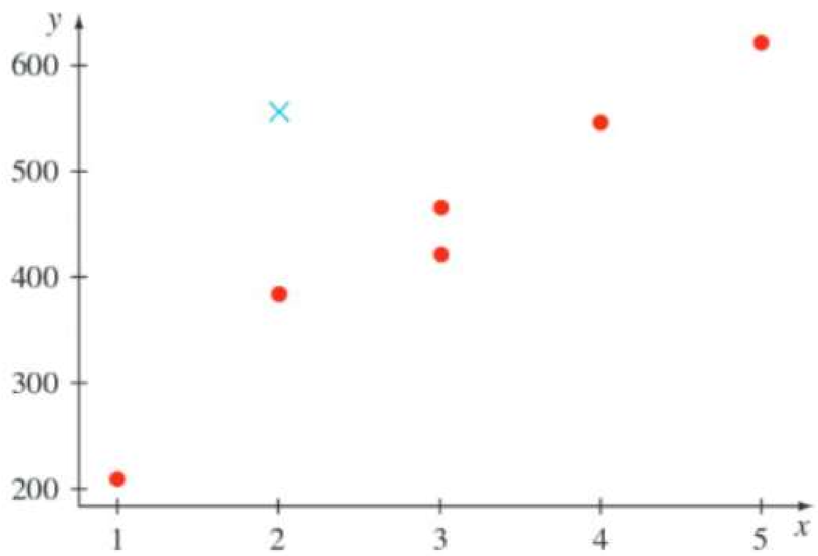
(b)



3.2 Describing **Bivariate Quantitative Data**

Scatterplot

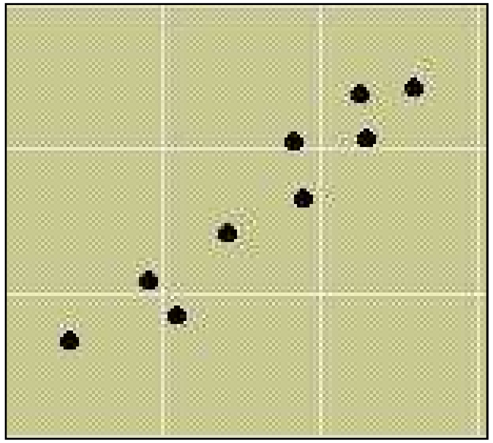
- When both of the variables are quantitative, call one variable *x* and the other *y*. A single measurement is a pair of numbers (*x*, *y*) that can be plotted using a two-dimensional graph called a **scatterplot.**



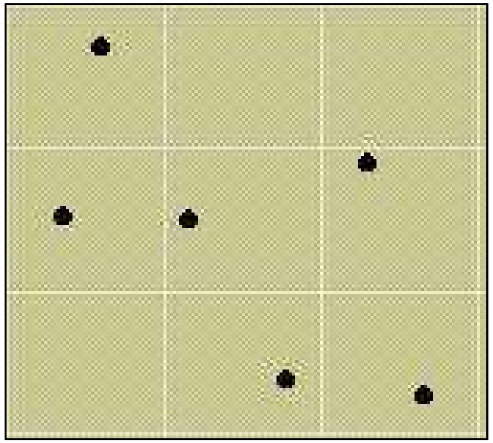
Describing the Scatterplot

- 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

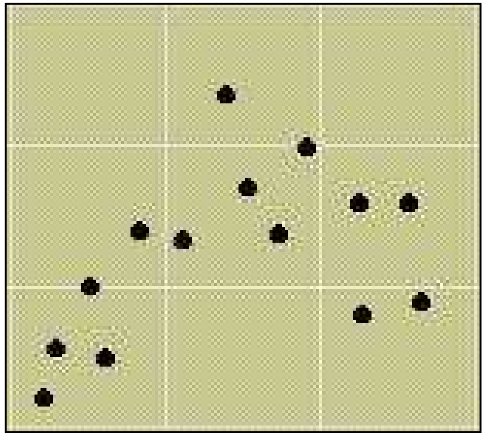
Example



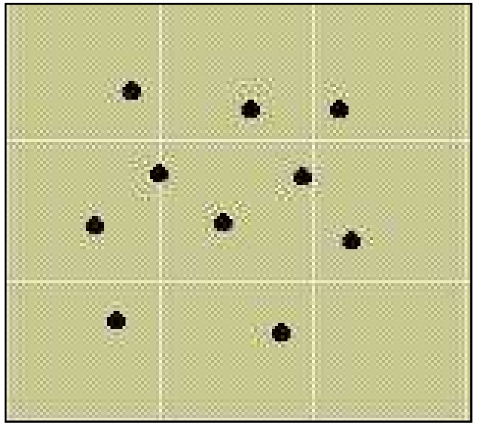
Positive linear - strong



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**.
- There are two numerical measures to describe
  - The **strength** and **direction** of the relationship between x and y.
  - The **form** of the relationship.

The Correlation Coefficient

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

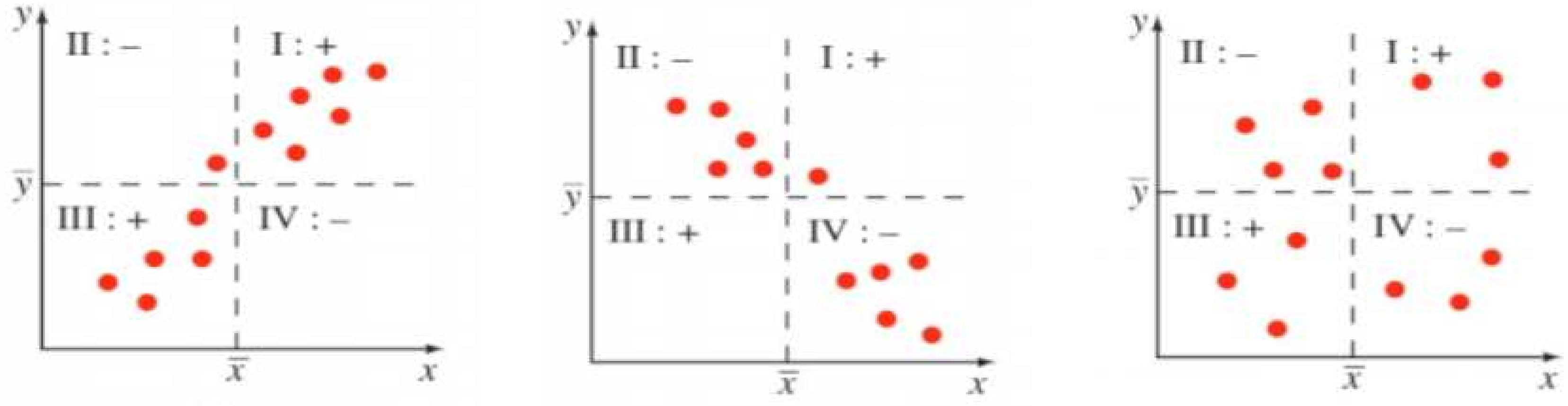
$$r = \frac{s_{xy}}{s_x s_y}$$

where $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{n-1}$ **:covariance**

s_x = standard deviation of the x's

s_y = standard deviation of the y's

- 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

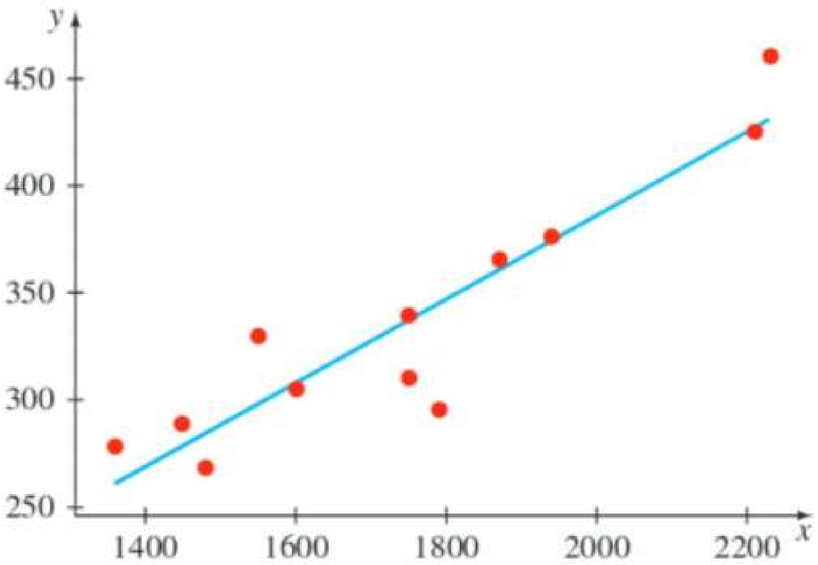
- If most of the points are in areas I and III (forming a positive pattern), s_{xy} and r will be positive.
- If most of the points are in areas II and IV (forming a negative pattern), s_{xy} and r will be negative.
- If the points are scattered across all four areas (forming no pattern), s_{xy} and r will be close to 0.

Example

Data : Living Area and Selling Price of 12 Residence

Residence	x (sq. ft.)	y (in thousands)
1	1360	$278.5
2	1940	375.7
3	1750	339.5
4	1550	329.8
5	1790	295.6
6	1750	310.3
7	2230	460.5
8	1600	305.2
9	1450	288.6
10	1870	365.7
11	2210	425.3
12	1480	268.8
sum	**20,980**	**4043.5**

Scatterplot



Indicates a positive linear relationship

$$\bar{x} = 1748.33 \quad s_x = 281.4842$$
$$\bar{y} = 336.9583 \quad s_y = 59.7592$$

$$s_{xy} = \frac{\sum x_i\, y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1} = \frac{7{,}240{,}383 - \frac{(20{,}980)(4043.5)}{12}}{11} = 15{,}545.19$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{15{,}545.19}{(281.48)(59.75)} = .9241$$

$$-1 \le r \le 1$$

Interpreting *r*

- $-1 \leq r \leq 1$: 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,**

$$y = a + bx.$$

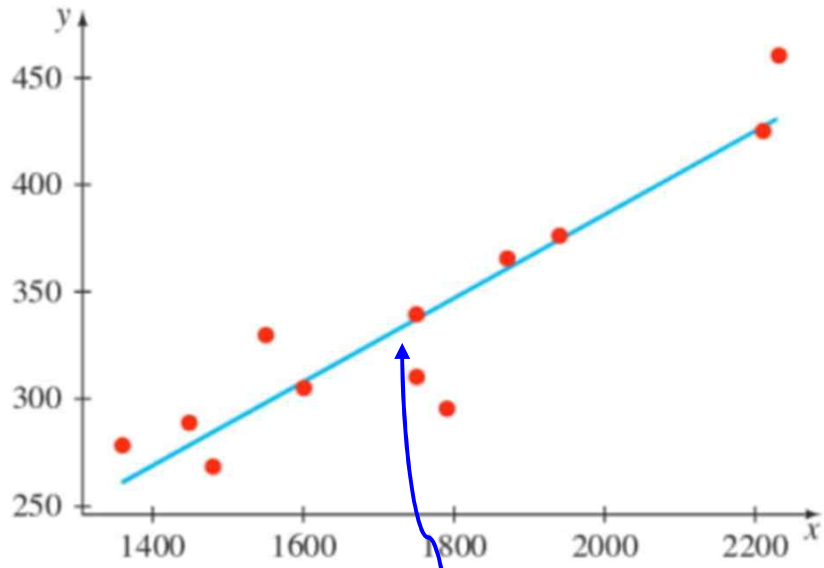
  - a = y-intercept of the line
  - b = slope of the line

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 $\hat{y} = a + bx$

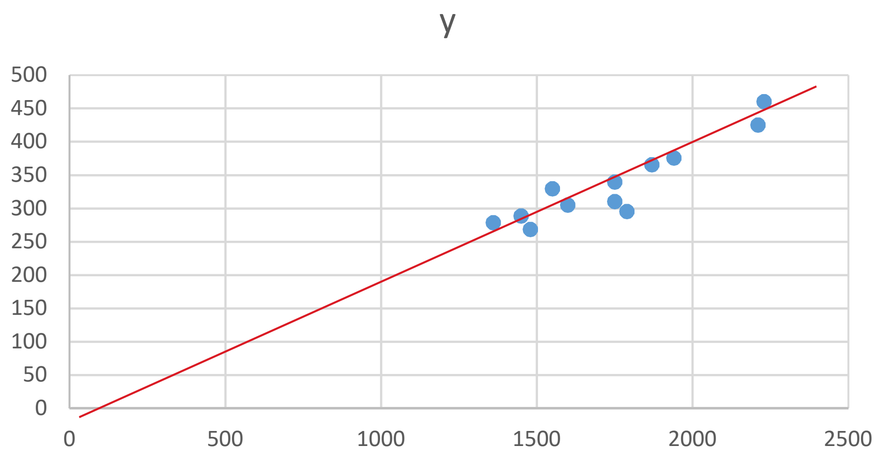
Example

Racall $\bar{x} = 1748.33$ $s_x = 281.4842$ $\bar{y} = 336.9583$ $s_y = 59.7592$
$r = 0.92414$

$$b = r\frac{s_y}{s_x} = (.92414)\frac{59.7592}{281.4842} = 0.196$$

$$a = \bar{y} - b\bar{x} = 336.95 - 0.196 * 1748.33 = -6.05$$

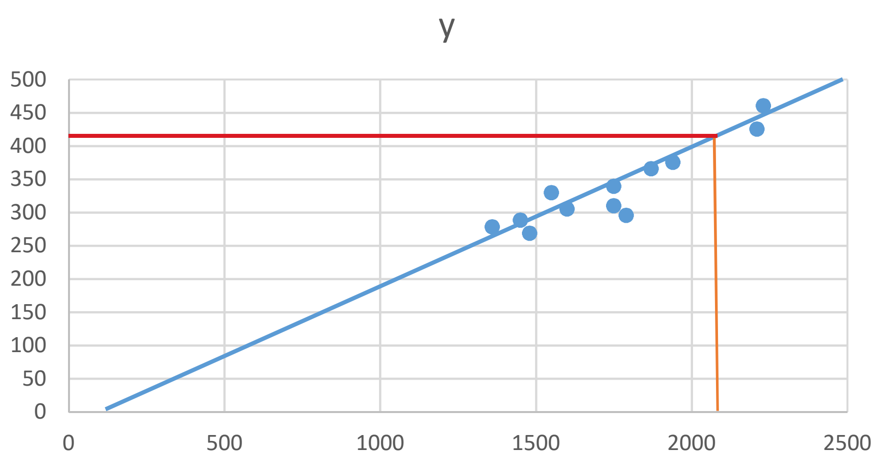
regression line : $\boldsymbol{\hat{y} = a + bx = -6.05 + 0.196x}$



Example (conti~)

- Predict the selling price for another residence with 2100 square feet of living area.
- $\hat{\boldsymbol{y}} = \boldsymbol{a} + \boldsymbol{bx} = -\mathbf{6.05} + \mathbf{0.196}\boldsymbol{x}$

  $= -6.05 + 0.196(2100) = \$405.95$(thousand)



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

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

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

Key Concepts

3. The best fitting line

- Calculating the slope and *y*-intercept

$$\boldsymbol{b = r\left(\frac{s_y}{s_x}\right)} \quad \text{and} \quad \boldsymbol{a = \bar{y} - b\bar{x}}$$

- Graphing the line
- Using the line for prediction