Chapter 2 : Data and Statistics

Data and Statistics

1. **Data :**

➢**Elements, Variables, and Observations**

a) **Elements** are the entities on which data are collected.

b) A **variable** is a characteristic of interest for the elements.

c) Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an **observation**.

Example

- The data set in Table 1.1 includes the following five variables: ***Fund Type , Net Asset Value ($) ,5-Year, Average Return (%) , Expense Ratio , Morningstar Rank***
- The set of measurements for the first observation (American Century Intl. Disc) is IE, 14.37, 30.53, 1.41, and 3-Star.
- The set of measurements for the second observation (American Century Tax-Free Bond) is FI, 10.73, 3.34, 0.49, and 4-Star, and so on.
- A data set with 25 elements contains 25 observations.

TABLE 1.1 DATA SET FOR 25 MUTUAL FUNDS

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds500 (2008).

1. Data :

➢Scales of Measurement

- Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio.
- The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

1. Data :

➢Scales of Measurement

- **Nominal scale :** When the data for a variable consist of labels or names used to identify an attribute of the element
- **Ordinal scale :** if the data has the properties of nominal data and the order or rank of the data is significant.
- **Interval scale :** if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric.
- **Ratio scale** if the data have all the properties of interval data and the ratio of two values is significant.

1. Data :

➢Categorical and Quantitative Data

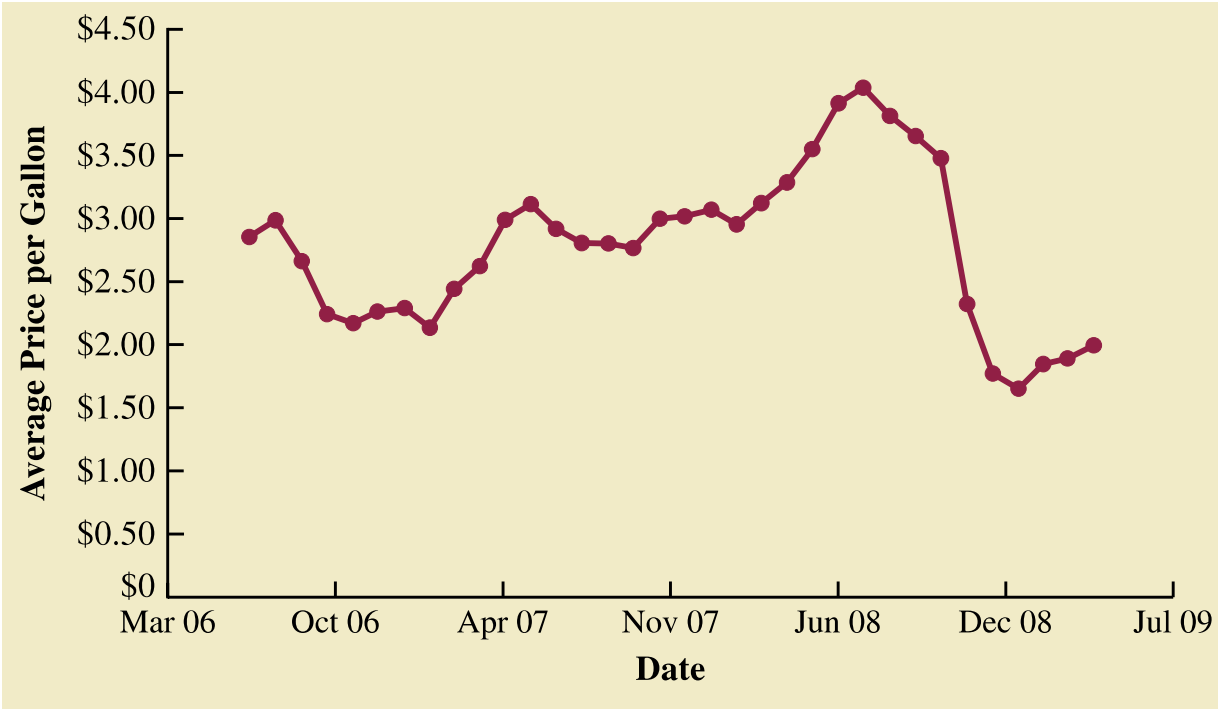
- Data that can be grouped by specific categories are referred to as **categorical data**. **Categorical data** use either the nominal or ordinal scale of measurement.
- Data that use numeric values to indicate how much or how many are referred to as **quantitative data**.
- **Quantitative data** are obtained using either the interval or ratio scale of measurement.
- A **categorical variable** is a variable with categorical data, and a **quantitative variable** is a variable with quantitative data.

1. Data :

➢Cross-Sectional and Time Series Data

- **Cross-sectional data** are data collected at the same or approximately the same point in time.
- Time series data are data collected over several time periods.

Example



1. Data :

➢Statistical Inference

- A population is the set of all elements of interest in particular study.
- A sample is a subset of the population.
- The process of conducting a survey to collect data for a sample is called a **sample survey**.
- Statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as **statistical inference**.
-

1. Descriptive Statistics :

➢ Summarizing Categorical Data

- **Frequency Distribution** is a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes.

Example : To develop a frequency distribution for these data, we count the number of times each soft drink appears in Table 2.1. Coke Classic appears 19 times, Diet Coke appears 8 times, Dr. Pepper appears 5 times, Pepsi appears 13 times, and Sprite appears 5 times.

TABLE 2.3 RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS OF SOFT DRINK PURCHASES

Soft Drink	Relative Frequency	Percent Frequency
Coke Classic	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

TABLE 2.1 DATA FROM A SAMPLE OF 50 SOFT DRINK PURCHASES

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

1. Descriptive Statistics :

➢ Summarizing Categorical Data

- **A frequency distribution** shows the number (frequency) of items in each of several nonoverlapping classes.

RELATIVE FREQUENCY

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

- A **percent frequency distribution** summarizes the percent frequency of the data for each class.
- A **bar chart** is a graphical device for depicting categorical data summarized in a frequency, relative frequency, or percent frequency distribution.

1. Descriptive Statistics :

➢ Summarizing Categorical Data

- **The pie chart** provides another graphical device for presenting relative frequency and percent frequency distributions for categorical data**.**

FIGURE 2.1 BAR CHART OF SOFT DRINK PURCHASES

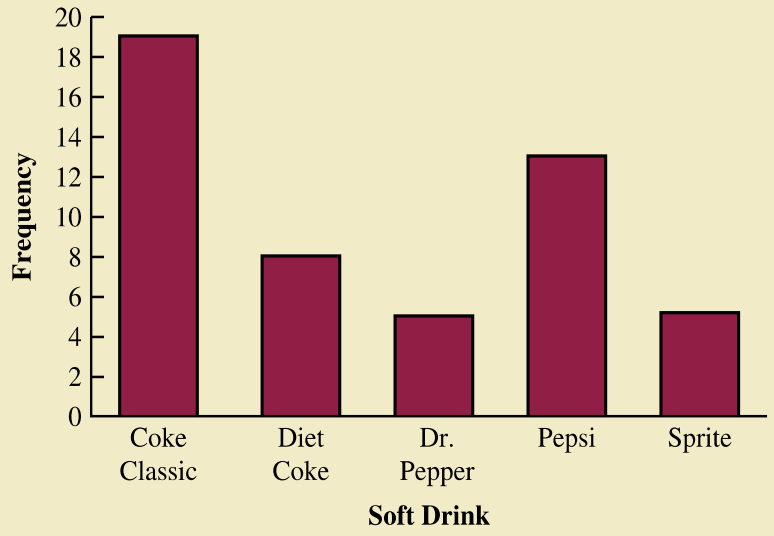
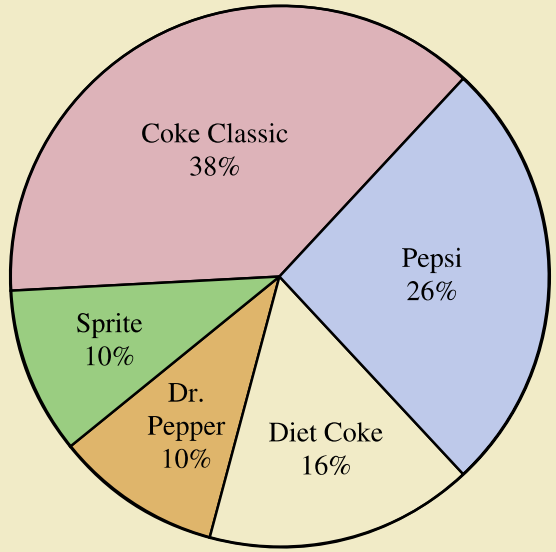


FIGURE 2.2 PIE CHART OF SOFT DRINK PURCHASES



1. Descriptive Statistics :

➢Summarizing Quantitative Data

- The three steps necessary to define the classes for a frequency distribution with quantitative data are:

1. **Determine the number of nonoverlapping classes.**
2. **Determine the width of each class.**
3. **Determine the class limits.**

TABLE 2.4

YEAR-END AUDIT TIMES (IN DAYS)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

TABLE 2.5

FREQUENCY DISTRIBUTION FOR THE AUDIT TIME DATA

Audit Time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

1. Descriptive Statistics :

➢Summarizing Quantitative Data

- We define the relative frequency and percent frequency distributions for quantitative data in the same manner as for qualitative data. First, recall that the **relative frequency** is the proportion of the observations belonging to a class. With n observations,

$$\text{Relative frequency of class} = \frac{\text{Frequency of the class}}{n}$$

- The percent frequency of a class is the relative frequency multiplied by 100.

TABLE 2.6 RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Relative Frequency	Percent Frequency
10–14	.20	20
15–19	.40	40
20–24	.25	25
25–29	.10	10
30–34	.05	5
Total	1.00	100

1. Descriptive Statistics :

➢Summarizing Quantitative Data

- One of the simplest graphical summaries of data is **a dot plot**. A horizontal axis shows the range for the data. Each data value is represented by a dot placed above the axis.
- A common graphical presentation of quantitative data is a **histogram**. A histogram is constructed by placing the variable of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis.

FIGURE 2.3 DOT PLOT FOR THE AUDIT TIME DATA

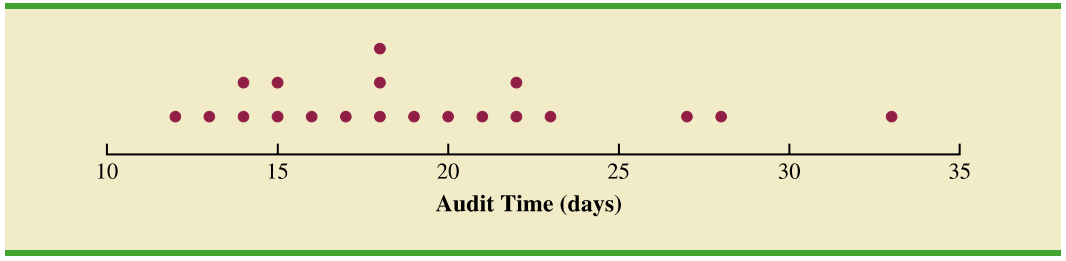
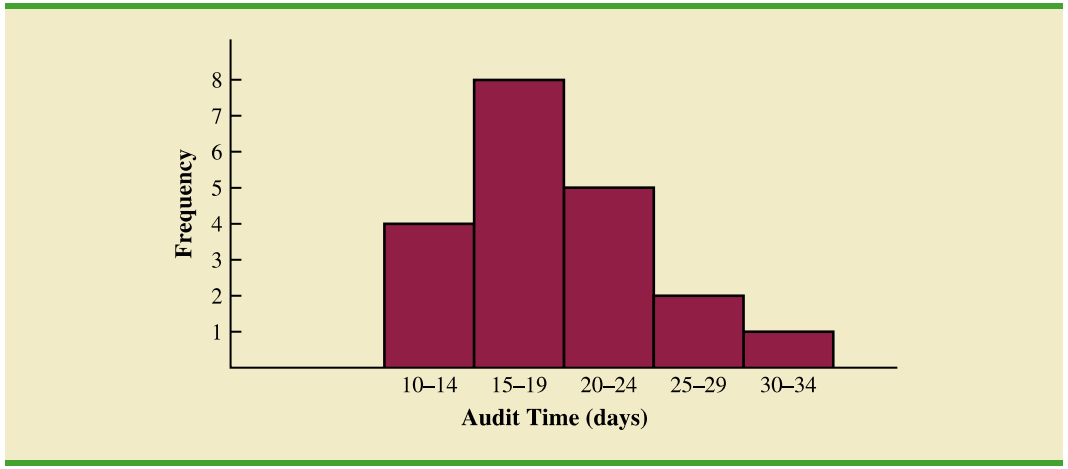


FIGURE 2.4 HISTOGRAM FOR THE AUDIT TIME DATA



1. Descriptive Statistics :

➢Summarizing Quantitative Data

- A variation of the frequency distribution that provides another tabular summary of quantitative data is the **cumulative frequency distribution**.
- The **cumulative frequency distribution** uses the number of classes, class widths, and class limits developed for the frequency distribution.
- A **cumulative percent frequency distribution** shows the percentage of data items with values less than or equal to the upper limit of each class.

TABLE 2.7 CUMULATIVE FREQUENCY, CUMULATIVE RELATIVE FREQUENCY, AND CUMULATIVE PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

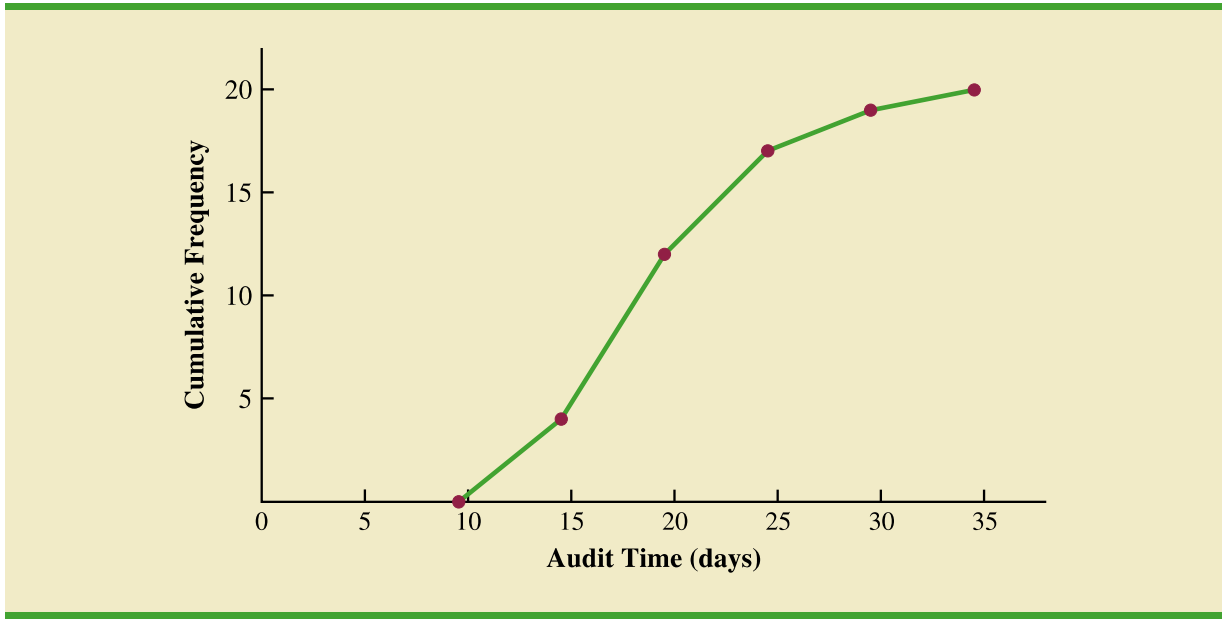
Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

1. Descriptive Statistics :

➢Summarizing Quantitative Data

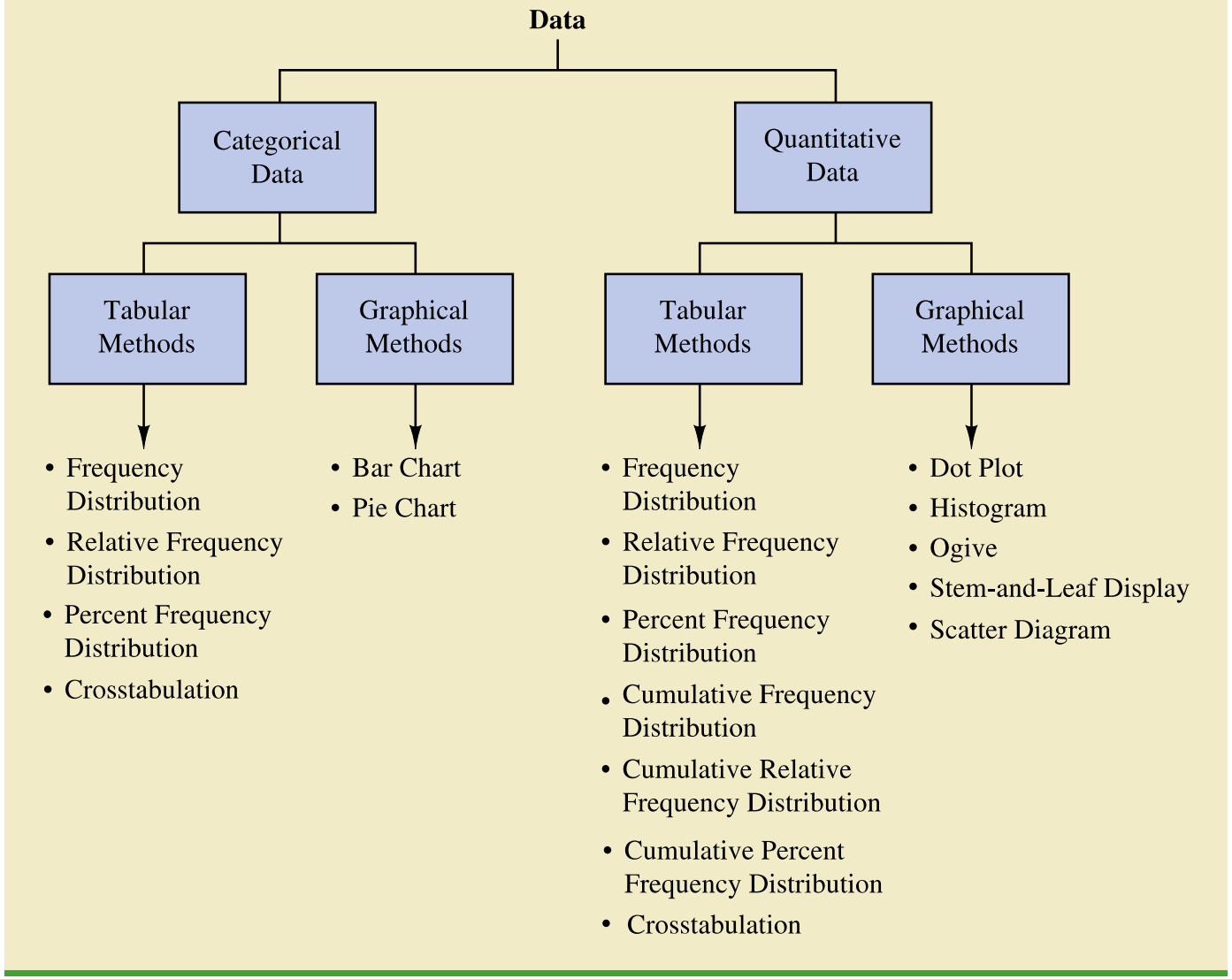
- A graph of a cumulative distribution, called an **ogive**, shows data values on the horizontal axis and either the cumulative frequencies, the cumulative relative frequencies, or the cumulative percent frequencies on the vertical axis.

FIGURE 2.6 OGIVE FOR THE AUDIT TIME DATA



Summary

FIGURE 2.9 TABULAR AND GRAPHICAL METHODS FOR SUMMARIZING DATA



1. Descriptive Statistics: Numerical Measures

➢ Measures of Location

- The mean provides a measure of central location for the data. If the data are for a sample, the mean is denoted by $\bar{x}$; if the data are for a population, the mean is denoted by the Greek letter μ.

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x_i}{n}$$

POPULATION MEAN

$$\mu = \frac{\Sigma x_i}{N}$$

TABLE 3.1 MONTHLY STARTING SALARIES FOR A SAMPLE OF 12 BUSINESS SCHOOL GRADUATES

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$
$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$
$$= \frac{42{,}480}{12} = 3540$$

1. Descriptive Statistics: Numerical Measures

➢Measures of Location

- The median is another measure of central location.
- MEDIAN
- Arrange the data in ascending order (smallest value to largest value).

a) For an odd number of observations, the median is the middle value.

b) For an even number of observations, the median is the average of the two middle values.

3310 3355 3450 3480 3480 $\underbrace{3490 \quad 3520}_{\text{Middle Two Values}}$ 3540 3550 3650 3730 3925

$$\text{Median} = \frac{3490 + 3520}{2} = 3505$$

1. Descriptive Statistics: Numerical Measures

➢Measures of Location

- A third measure of location is the **mode**.
- The **mode** is the value that occurs with greatest frequency.
- **Percentile** : the pth percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100 - p) percent of the observations are greater than or equal to this value.

1. Descriptive Statistics: Numerical Measures

➢Measures of Location

- **Calculating the *pth* percentile**
- **Step 1.** Arrange the data in ascending order (smallest value to largest value).
- **Step 2.** Compute an index i

$$i = \left(\frac{p}{100}\right)n$$

where p is the percentile of interest and n is the number of observations.

- **Step 3.**

a) If *i is not an integer, round up*. The next integer *greater* than i denotes the position of the *p*th percentile.

b) If *i is an integer*, the *p*th percentile is the average of the values in positions i and $i + 1$.

Example. As an illustration of this procedure, let us determine the 85th percentile for the starting salary data.

Step 1. Arrange the data in ascending order.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Step 2.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

Step 3. Because *i* is not an integer, *round up*. The position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730.

As another illustration of this procedure, let us consider the calculation of the 50th percentile for the starting salary data. Applying step 2, we obtain

$$i = \left(\frac{50}{100}\right)12 = 6$$

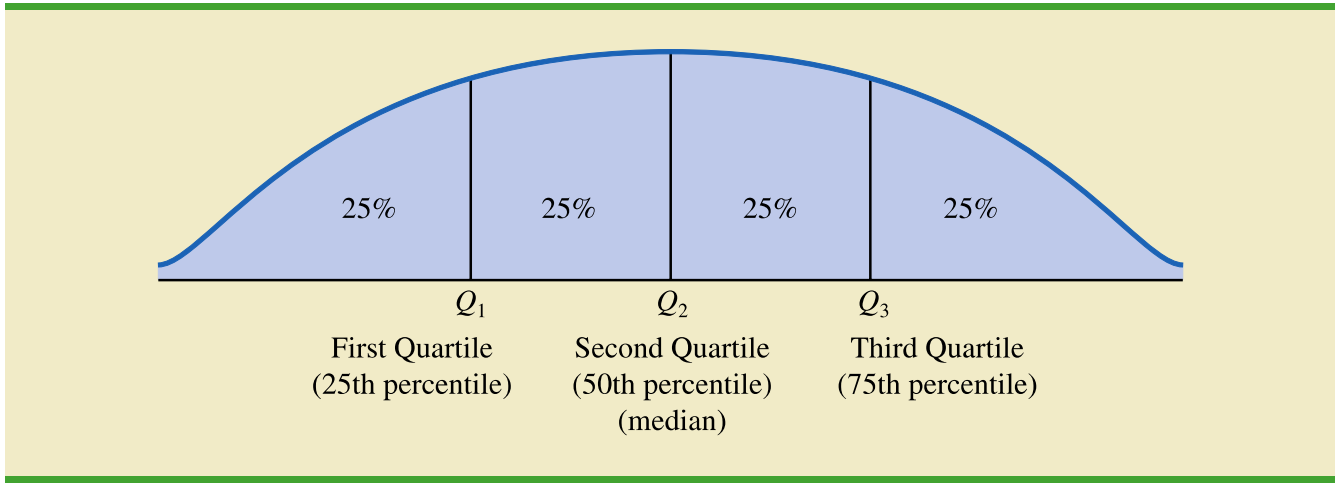
Because *i* is an integer, step 3(b) states that the 50th percentile is the average of the sixth and seventh data values; thus the 50th percentile is (3490 + 3520)/2 = 3505. Note that the *50th percentile is also the median.*

1. Descriptive Statistics: Numerical Measures

➢Measures of Location

- Quartiles: The division points are referred to as the quartiles and are defined as
  - Q1 = first quartile, or 25th percentile
  - Q2 = second quartile, or 50th percentile (also the median)
  - Q3 = third quartile, or 75th percentile.
- Figure 3.1 shows a data distribution divided into four parts.

FIGURE 3.1 LOCATION OF THE QUARTILES



Example. The starting salary data are again arranged in ascending order. We already identified Q2, the second quartile (median), as 3505. 3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

The computations of quartiles Q1 and Q3 require the use of the rule for finding the 25th and 75th percentiles.

For Q1,

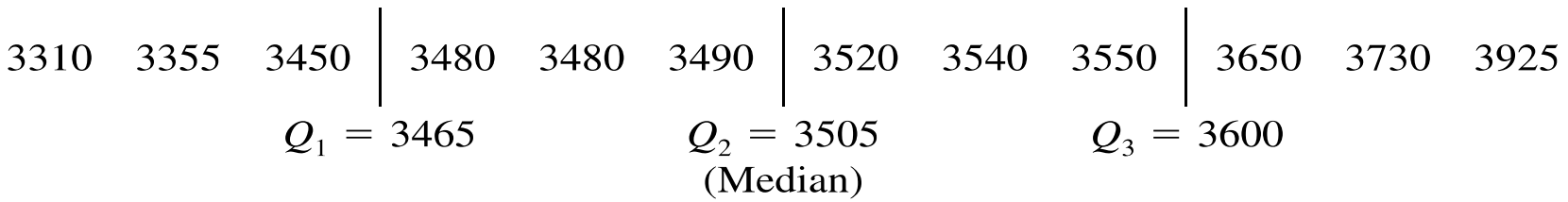
$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

Because i is an integer, step 3(b) indicates that the first quartile, or 25th percentile, is the average of the third and fourth data values; thus, Q1 = (3450 + 3480)/2 = 3465.

For Q3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

Again, because i is an integer, step 3(b) indicates that the third quartile, or 75th percentile, is the average of the ninth and tenth data values; thus, Q3 = (3550 + 3650)/2 = 3600. The quartiles divide the starting salary data into four parts, with each part containing 25% of the observations.



1. Descriptive Statistics: Numerical Measures

➢Measures of Variability

- **Range:** the simplest measure of variability is the **range.**

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

- **Interquartile Range**: This measure of variability is the difference between the third quartile, Q3, and the first quartile, Q1. In other words, the interquartile range is the range for the middle 50% of the data.

$$\text{IQR} = Q3 - Q1$$

For the data on monthly starting salaries, the quartiles are Q3 = 3600 and Q1 = 3465. Thus, the interquartile range is 3600 - 3465 = 135.

1. Descriptive Statistics: Numerical Measures

➢Measures of Variability

- **The variance** is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation (xi) and the mean**.**

POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

SAMPLE VARIANCE

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

TABLE 3.3 COMPUTATION OF THE SAMPLE VARIANCE FOR THE STARTING SALARY DATA

Monthly Salary (x_i)	Sample Mean ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean ($x_i - \bar{x}$)^2
3450	3540	−90	8,100
3550	3540	10	100
3650	3540	110	12,100
3480	3540	−60	3,600
3355	3540	−185	34,225
3310	3540	−230	52,900
3490	3540	−50	2,500
3730	3540	190	36,100
3540	3540	0	0
3925	3540	385	148,225
3520	3540	−20	400
3480	3540	−60	3,600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Using equation (3.5),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301{,}850}{11} = 27{,}440.91$$

1. Descriptive Statistics: Numerical Measures

➢Measures of Variability

- **The standard deviation** is defined to be the positive square root of the variance**.**
- **Coefficient of variation:** indicates how large the standard deviation is relative to the mean.

STANDARD DEVIATION

$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

1. Descriptive Statistics: Numerical Measures

➢Exploratory Data Analysis

- In a five-number summary, the following five numbers are used to summarize the data:

1. **Smallest value**
2. **First quartile (Q1)**
3. **Median (Q2)**
4. **Third quartile (Q3)**
5. **Largest value**

Example : The monthly starting salaries shown in Table 3.1 for a sample of 12 business school graduates are repeated here in ascending order.

3310 3355 3450 | 3480 3480 3490 | 3520 3540 3550 | 3650 3730 3925

$Q_1 = 3465$ $Q_2 = 3505$ (Median) $Q_3 = 3600$

1. Descriptive Statistics: Numerical Measures

➢Exploratory Data Analysis

- A **box plo**t is a graphical summary of data that is based on a five-number summary. A key to the development of a box plot is the computation of the median and the quartiles, Q1 and Q3. The interquartile range, IQR = Q3 - Q1, is also used.
- The steps used to construct the box plot follow.

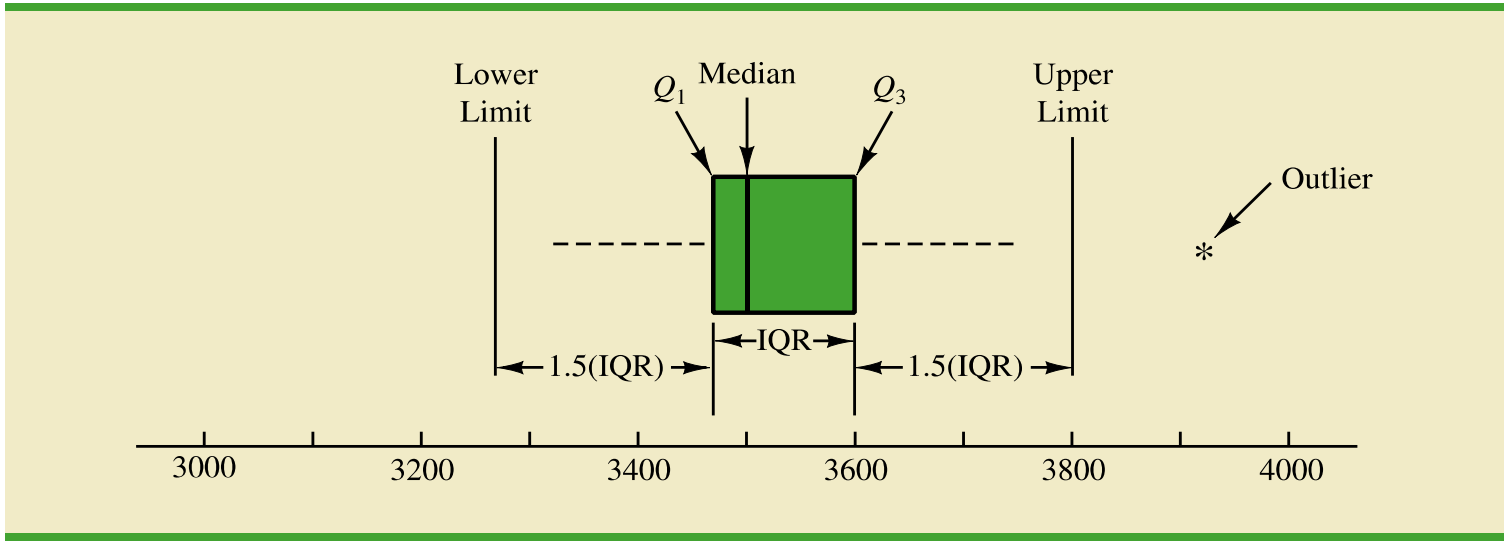
○ A box is drawn with the ends of the box located at the first and third quartiles. For the salary data,Q1 =3465 and Q3 =3600.This box contains the middle 50% of the data.

○ A vertical line is drawn in the box at the location of the median (3505 for the salary data).

○ By using the interquartile range, IQR = Q3 - Q1, limits are located. The limits for the box plot are 1.5(IQR) below Q1 and 1.5(IQR) above Q3.

- For the salary data, IQR = Q3 - Q1 = 3600 - 3465 = 135. Thus, the limits are 3465 - 1.5(135) = 3262.5 and 3600 + 1.5(135) = 3802.5. Data outside these limits are considered outliers.
- The dashed lines in Figure 3.5 are called whiskers. The whiskers are drawn from the ends of the box to the smallest and largest values inside the limits computed in step 3. Thus, the whiskers end at salary values of 3310 and 3730.
- Finally, the location of each outlier is shown with the symbol *. In Figure 3.5 we see one outlier, 3925.

FIGURE 3.5 BOX PLOT OF THE STARTING SALARY DATA WITH LINES SHOWING THE LOWER AND UPPER LIMITS



1. Descriptive Statistics: Numerical Measures

➢Measures of Association Between Two Variables

- covariance as a descriptive measure of the linear association between two variables.
- For a sample of size n with the observations (x1, y1), (x2, y2), and so on, the sample covariance is defined as follows:

SAMPLE COVARIANCE

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

POPULATION COVARIANCE

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N}$$

TABLE 3.7 CALCULATIONS FOR THE SAMPLE COVARIANCE

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

Correlation Coefficient

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT: SAMPLE DATA

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad (3.12)$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT: POPULATION DATA

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (3.13)$$

where

ρ_{xy} = population correlation coefficient
σ_{xy} = population covariance
σ_x = population standard deviation for x
σ_y = population standard deviation for y

TABLE 3.8 COMPUTATIONS USED IN CALCULATING THE SAMPLE CORRELATION COEFFICIENT

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totals	30	90	0	50	0	800	200

$\bar{x} = 10$ $\bar{y} = 30$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$