Data and Descriptive Statistics

Statistics

The term **statistics** refers to numerical facts such as averages, medians, percentages, and maximums that help us understand a variety of business and economic situations.

Statistics can also refer to the art and science of collecting, analyzing, presenting, and interpreting data.

We emphasize the use of statistics for business and economic decision-making.

Applications in Business and Economics

Accounting: public accounting firms use statistical sampling procedures when conducting audits for their clients.

Economics: economists use statistical information in making forecasts about the future of the economy or some aspect of it.

Finance: financial advisors use price-earnings ratios and dividend yields to guide their investment advice.

Marketing: electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.

Production: a variety of statistical quality control charts are used to monitor the output of a production process.

Information Systems: a variety of statistical information helps administrators assess the performance of computer networks.

Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

Elements are the entities on which data are collected.

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

A **data set** consists of all the data collected for a particular study.

- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.

An **observation** is the set of **measurements** obtained for a particular element.

- A data set with *n* elements contains *n* observations.

Data Set for 60 Nations in the World Trade Organization

DATAfile: *Nations*

Shown here are the first five nations out of 60 (rows) of a table including several variables (columns.)

- Observation #1 (Armenia) contains the measurements: Member, 4,267, B+, and Stable.
- Observation #2 (Australia) contains the measurements: Member, 51,812, AAA, and Negative.
- And so on.

Nation	**WTO Status**	**Per Capita GDP (S)**	**Fitch Rating**	**Fitch Outlook**
Armenia	Member	4,267	B+	Stable
Australia	Member	51,812	AAA	Negative
Austria	Member	48,328	AA+	Stable
Azerbaijan	Observer	4,214	BB+	Stable
Bahrain	Member	28,608	B+	Stable

Nominal and Ordinal Scales of Measurement

Nominal scale: the data are labels or names used to identify an attribute of the element.

Example: the WTO Status variable in the Nations table has a nominal scale of measurement because it uses the data “Member” and “Observer” are labels used to identify the status category for a nation.

Note: a numeric code may also be used. For example, we could use the label “1” to identify a “Member” status category and “2” to identify an “Observer” status category.

Ordinal scale: the data have the properties of nominal data, and the order or rank of the data is meaningful.

Example: the Fitch Rating variable in the Nations table has ordinal data because the labels, which range from AAA to F, are rank-ordered by credit rating.

Note: ordinal data can also be recorded by a numerical code, such as a student’s class rank in school.

Interval and Ratio Scales of Measurement

Interval scale: the data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure. Data with an interval scale of measurement are always numerical.

Example: College admission SAT scores have an interval scale of measurement.

Note: the difference between the values of a variable with an interval scale of measurement are always meaningful.

Ratio scale: the data have the properties of interval data, and the ratio of two values is meaningful.

Example: variables such as distance, height, weight, and time use the ratio scale of measurement.

Note: the ratio scale of measurement requires that a zero value be included to indicate that nothing exists for the variable at the zero point.

Categorical and Quantitative Data

Data can be further classified as **categorical** or **quantitative**.

The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.

Categorical Data

A **categorical variable** is a variable with categorical data.

Statistical analyses are rather limited.

We can summarize categorical data by:

- counting the number of observations in each category
- computing the proportion of the observations in each category

Quantitative Data

A **quantitative variable** is a variable with quantitative data.

Quantitative data indicate *how many* or *how much* and are always numerical.

Ordinary arithmetic operations are meaningful for quantitative data.

More alternatives for statistical analysis are available when the data are quantitative.

Cross-Sectional and Time Series Data

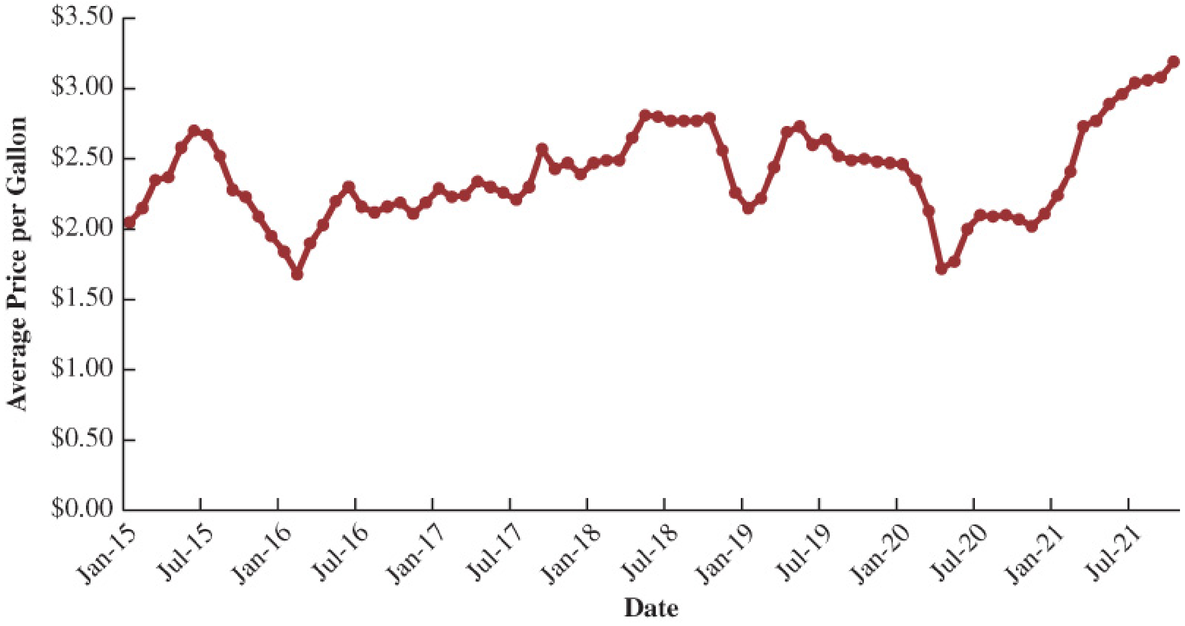
Cross-sectional data are collected at the same or approximately the same point in time.

> Example: The data in the previously shown Nations data file are cross-sectional because they describe the five variables for the 60 WTO nations at the same point in time.

Time series data are instead collected over several time periods.

Graphs of time series data, such as the U.S. average price per gallon of regular gasoline between 2015 and 2021 shown to the right, help understand:

- what happened in the past
- identify any trends over time
- project future levels for the time series



Considerations on Data Acquisition

Time Requirement

- Searching for information can be time-consuming.
- Information may no longer be useful by the time it is available.

Cost of Acquisition

- Organizations often charge for information even when it is not their primary business activity.

Data Errors

- Using any data that happen to be available.
- Data acquired with little care can lead to misleading information.

Descriptive Statistics

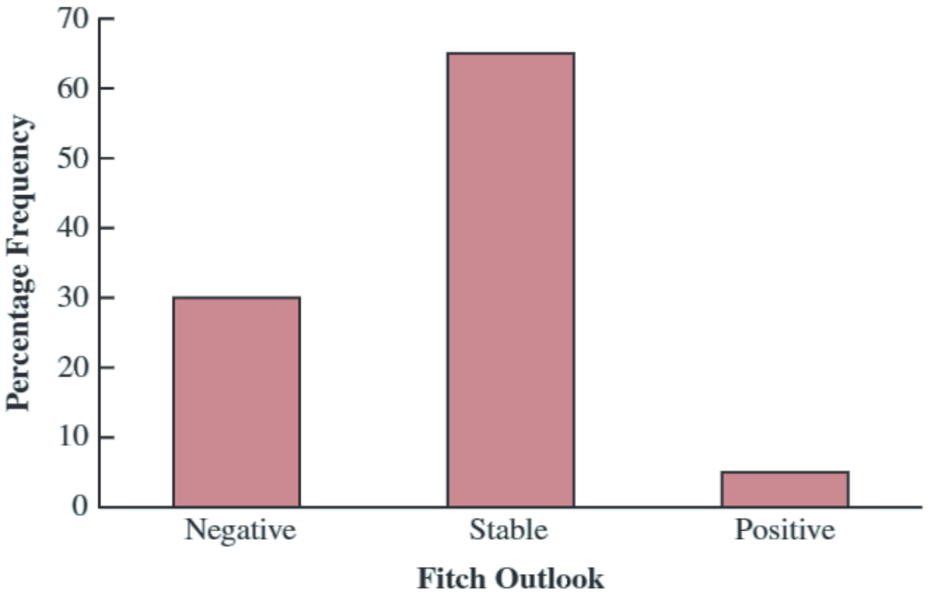
Most of the statistical information in publications consists of data that are summarized and presented in a form that is easy to understand.

Such summaries of data, which may be tabular, graphical, or numerical, are referred to as **descriptive statistics**.

Examples:

- The variable Fitch Outlook in the WTO Nations table can be summarized as a table or bar chart.
- Numerical descriptive statistics such as the mean (or average.)

Fitch Outlook	Frequency	Percent Frequency (%)
Positive	3	5.0
Stable	39	65.0
Negative	18	30.0



Statistical Inference

When the collection of information about a large group of elements (individuals, companies, voters, households, products, customers, and so on) is not feasible because of time, cost, and other considerations, data can be collected from only a small portion of the group.

Formally, we use the following definitions of the larger and smaller groups of elements:

Population: the set of all elements of interest in a particular study

Sample: a subset of the population

We also formally define the following statistical processes:

Statistical inference: the process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population

Census: the process of collecting data for the entire population

Sample survey: a process of collecting data for a sample

Analytics

Analytics is the scientific process of transforming data into insight for making better decisions.

Analytics is now generally thought to comprise three broad categories of techniques:

Descriptive analytics describes what has happened in the past.

- Examples – data queries, reports, descriptive statistics, data visualization, data dashboards, and basic what-if spreadsheet models

Predictive analytics uses models constructed from past data to predict the future or to assess the impact of one variable on another

- Examples – linear regression, time series analysis, forecasting models, and simulation.

Prescriptive analytics yield the best course of action.

- Examples – optimization models and decision analysis

Big Data and Data Mining

Big data are larger and more complex data sets that cannot be managed, processed, or analyzed with commonly available software in a reasonable amount of time.

Data warehousing is the process of capturing, storing, and maintaining data.

- Computing power and data collection tools have reached the point where it is now feasible to store and retrieve extremely large quantities of data in seconds.

Data mining methods help develop useful decision-making information from large databases.

- Analysts use a combination of automated procedures to “mine the data” and convert it into useful information.
- The most effective data mining procedures are multiple regression, logistic regression, and machine learning, which discover relationships in the data and predict future outcomes.

Computers and Statistical Analysis

Statisticians use computer software to perform tedious and time-consuming statistical computations and analyses.

Simple statistical analysis using Microsoft Excel

Big data requires special data manipulation and analysis tools such as:

- R and Python – open-source programming languages
- SAS and SPSS – commercially available packages

Graphical displays

Frequency Distribution

A **frequency distribution** is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.

Example DATAfile: *SoftDrink*

To develop a frequency distribution for the sample of 50 soft drink purchases, we count the number of times each soft drink appears.

The frequency distribution highlights Coca-Cola as the leader, followed by Pepsi, Diet Coke, Dr. Pepper, and Sprite.

Soft Drink	**Frequency**
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Relative Frequency and Percent Frequency Distributions

The **relative frequency** of a class is the fraction or proportion of the total number of data items belonging to the class.

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

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

The relative frequency distribution for the soft drink data shows a relative frequency of $19/50 = 0.38$ for Coca-Cola, $13/50 = 0.26$ for Pepsi, and so on.

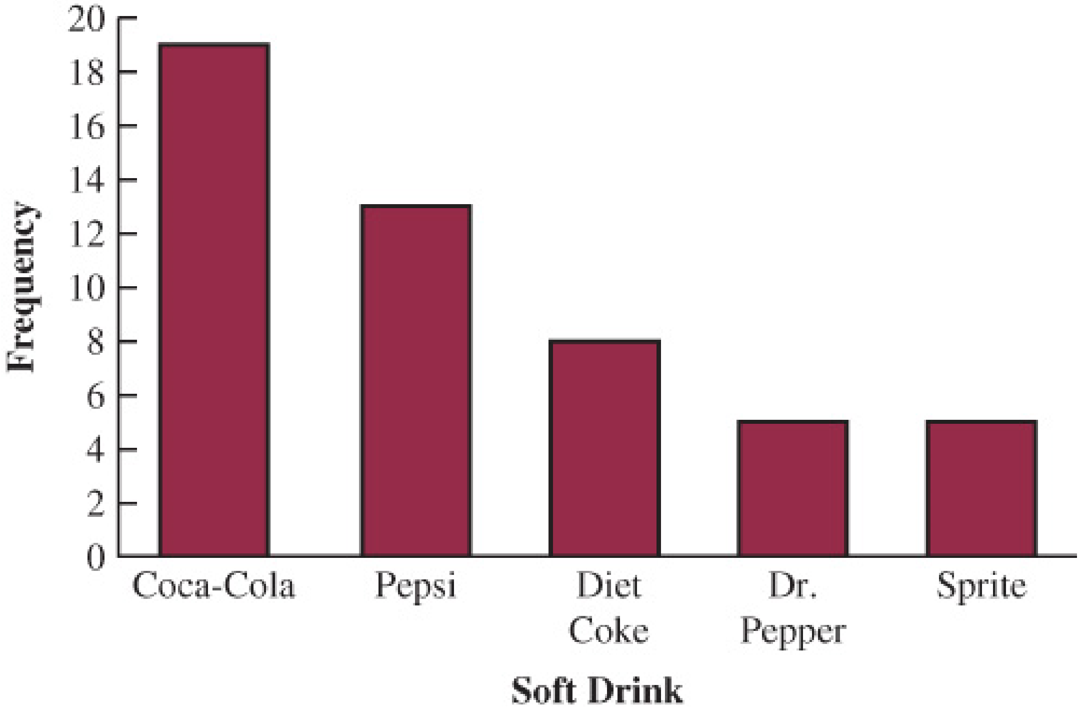
The percent frequency distribution, shows 38% Coca-Cola purchases, 16% Diet Coke purchases, and so on.

Soft Drink	Relative Frequency	Percent Frequency
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

Bar Chart

A **bar chart** is a graphical display for depicting categorical data summarized in a frequency, relative frequency, or percent frequency distribution.

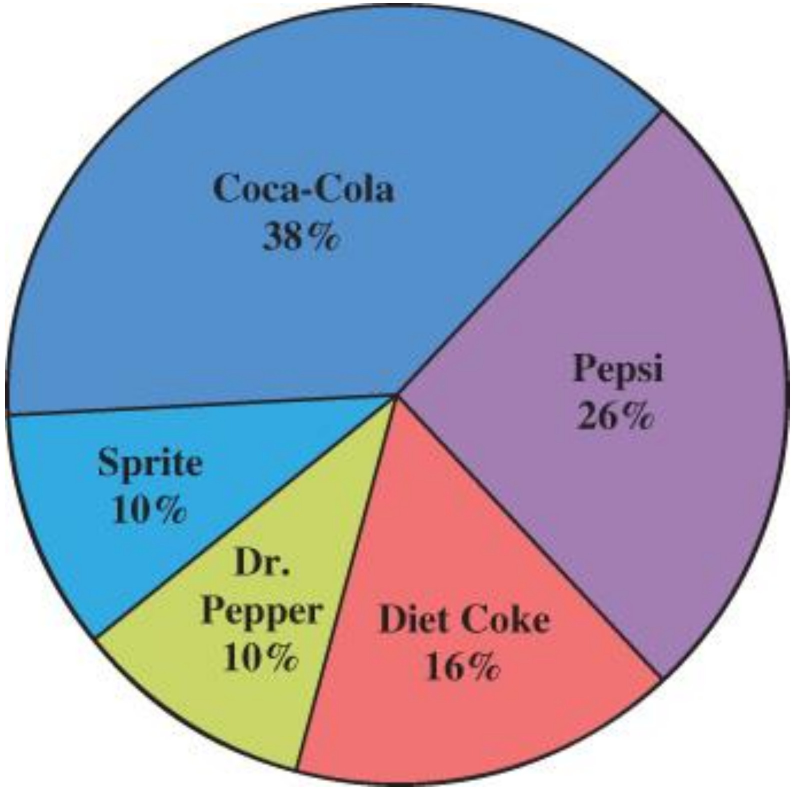
- On one axis (usually the horizontal axis), we specify the labels used for the classes (categories).
- A frequency, relative frequency, or percent frequency scale is used for the other axis (usually the vertical axis).
- Using a bar of fixed width, drawn above each class label, we extend the height appropriately.
- The bars are separated to emphasize the fact that each class is a separate category.
- In this example, the bars are also sorted.



Pie Chart

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

- First draw a circle.
- Then, use the percent frequencies to subdivide the circle into sectors that are proportional to the percent frequency for each class.
- Because there are 360 degrees in a circle, the sector of the pie chart labeled Coca-Cola consists of $0.38(360) = 136.8°$.
- Similar calculations for the other classes yield the pie chart shown.
- In general, pie charts are not the best way to present percentages for comparison.



Frequency Distribution for a Quantitative Variable

To build a frequency distribution for quantitative data, we must be more careful in defining the nonoverlapping classes to be used in the frequency distribution.

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

1. **Number of classes**: from 5, for small data sets, up to 20 for large data sets.
2. **Width of the classes**: $\text{Approx. class width} = \dfrac{\text{Largest Data Value – Smallest Data Value}}{\text{Number of classes}}$
3. **Class limits** are chosen so that classes do not overlap, and each data item belongs to only one class

Example DATAfile: *Audit*

The data shows the time in days required to complete year-end audits for a sample of 20 clients of a small public accounting firm.

Five classes were chosen, with class width $= (33 - 12)/5 = 4.2 \approx 5$. The resulting frequency distribution is shown to the right.

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

Relative Frequency and Percent Frequency Distributions for a Quantitative Variable

We define the relative frequency and percent frequency distributions for quantitative data in the same manner as for categorical data

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

Remember that the percent frequency of a class is the relative frequency multiplied by 100.

Based on the class frequencies for the 20 audit times, the table shows the relative frequency distribution and percent frequency distribution.

Note that 40% of the audits required between 15 and 19 days, and only 5% of the audit required 30 or more days.

Audit Time (days)	**Relative Frequency**	**Percent Frequency**
10-14	0.20	20
15-19	0.40	40
20-24	0.25	25
25-29	0.10	10
30-34	0.05	5
Total	1.00	100

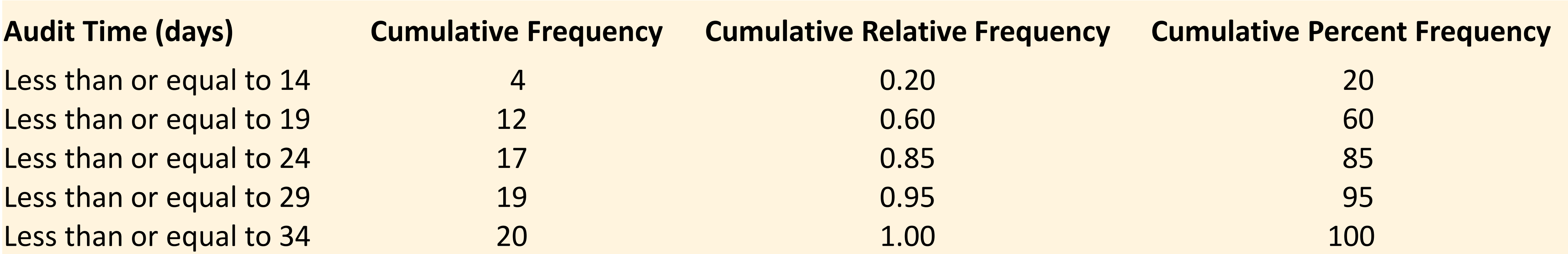
Cumulative Distributions

A **cumulative frequency distribution** shows the number of items with values less than or equal to the upper limit of each class

A **cumulative relative frequency distribution** shows the proportion of items with values less than or equal to the upper limit of each class.

A **cumulative percent frequency distribution** shows the percentage of items with values less than or equal to the upper limit of each class.

The distributions in the table show that 85% of the audits were completed in 24 days or less, 95% of the audits were completed in 29 days or less, and so on.



Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

Dot Plot

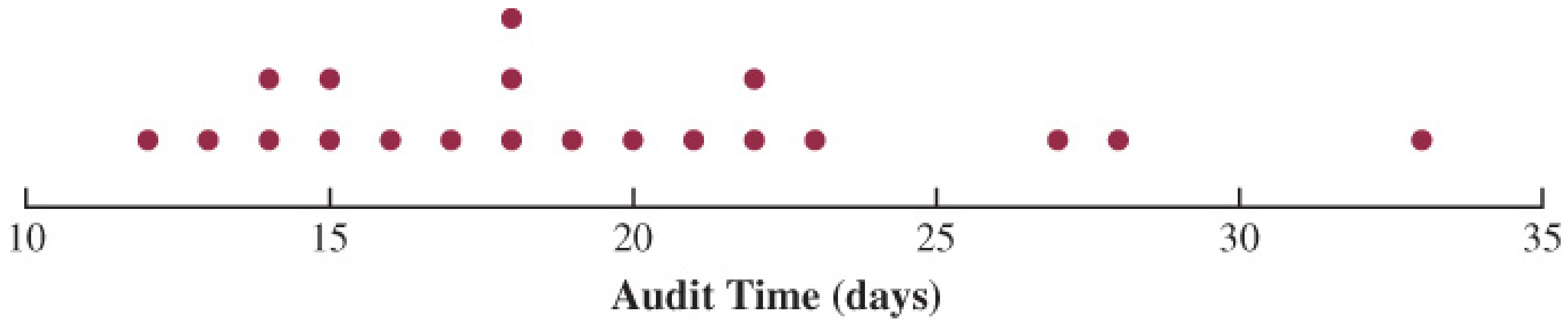
The **Dot Plot** is one of the simplest graphical summaries of quantitative data.

A horizontal axis shows the range for the data, and each data value is represented by a dot placed above the axis.

Dot plots show the details of the data and are useful for comparing the distribution of the data for two or more variables.

The dot plot for the audit time data is shown below.

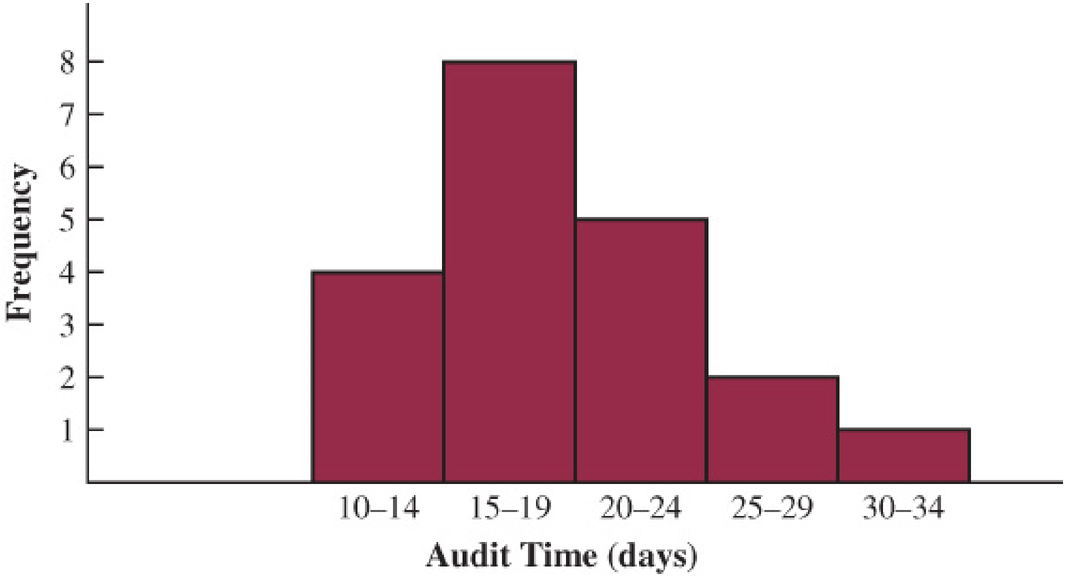
- The three dots located above 18 on the horizontal axis indicate that an audit time of 18 days occurred three times.



Histogram

The **histogram** is a common graphical display that can be prepared for quantitative data previously summarized in a frequency, relative frequency, or percent frequency distribution.

- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes



The histogram for the audit time data is shown.

- Note that the class of 15–19 days shows the greatest frequency with 8 audits.
- The class of 30-34 days shows the lowest frequency with only 1 audit.

Stem-and-Leaf Display

A **stem-and-leaf display** shows both the rank order and shape of a distribution of data.

It is similar to a histogram on its side, but it has the advantage of showing the actual values.

- The leading digits of each data item are arranged to the left of a vertical line.
- To the right of the vertical line, we record the last digit for each item in rank order.
- Each line (row) in the display is referred to as a *stem*.
- Each digit on a stem is a *leaf*.

Example DATAfile: *AptitudeTest*

The data indicate the number of questions answered correctly in a 150-question aptitude test given to 50 individuals recently interviewed for a position.

Stem-and-leaf displays for data with more than three digits are possible.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Crosstabulation

Consider Zagat's review of 300 restaurants in the Los Angeles area. The data set includes measurements on quality rating and typical meal price (DATAfile: *Restaurant*.)

A **crosstabulation** of the data is shown below, where:

- the Quality Rating, a categorical variable with categories *good*, *very good*, and *excellent*, is shown in the left margin, and its frequency distribution in the right margin.
- the Meal Price, a quantitative variable ranging from $10 to $49 and grouped into four classes, is shown in the top margin, and its frequency distribution in the bottom margin.
- the cells in the body of the table provide the counts for all the combinations of Quality Rating and Meal Price.

	Meal Price				
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**	**Total**
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

Crosstabulation: Row Percentages

Converting the entries in a crosstabulation into row percentages or column percentages can provide more insight into the relationship between the two variables.

For row percentages, the results of dividing each frequency in the crosstabulation by its corresponding row total are shown in the table below.

Each row in the table is a percent frequency distribution of meal price for one of the quality rating categories.

- Of the restaurants with the lowest quality rating (good), 50% have low meal prices.
- Of the restaurants with an excellent quality rating, most are in the $30-$39 and $40-$49 higher price ranges.

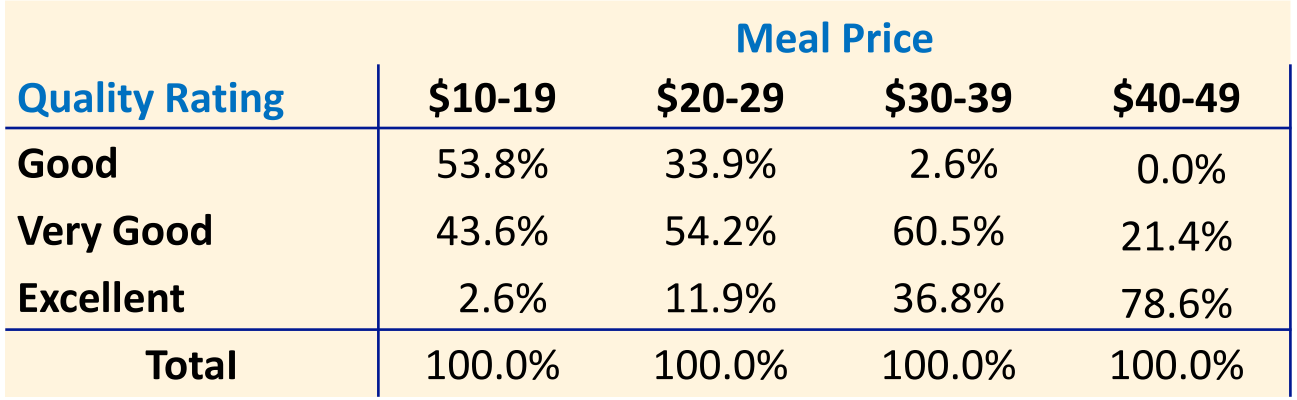
	Meal Price				
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**	**Total**
Good	50.0%	47.6%	2.4%	0.0%	100.0%
Very Good	22.7%	42.7%	30.6%	4.0%	100.0%
Excellent	3.0%	21.2%	42.4%	33.4%	100.0%

Crosstabulation: Column Percentages

For column percentages, the results of dividing each frequency in the crosstabulation by its corresponding column total are shown in the table below.

Each column in the table is a percent frequency distribution of quality rating for one of the classes of meal prices.

- Only 2.6% of restaurants with low ($10-19) meal prices have an excellent quality rating.
- As meal price increases, percent frequency distribution shifts toward higher quality ratings.



	Meal Price			
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**
Good	53.8%	33.9%	2.6%	0.0%
Very Good	43.6%	54.2%	60.5%	21.4%
Excellent	2.6%	11.9%	36.8%	78.6%
Total	100.0%	100.0%	100.0%	100.0%

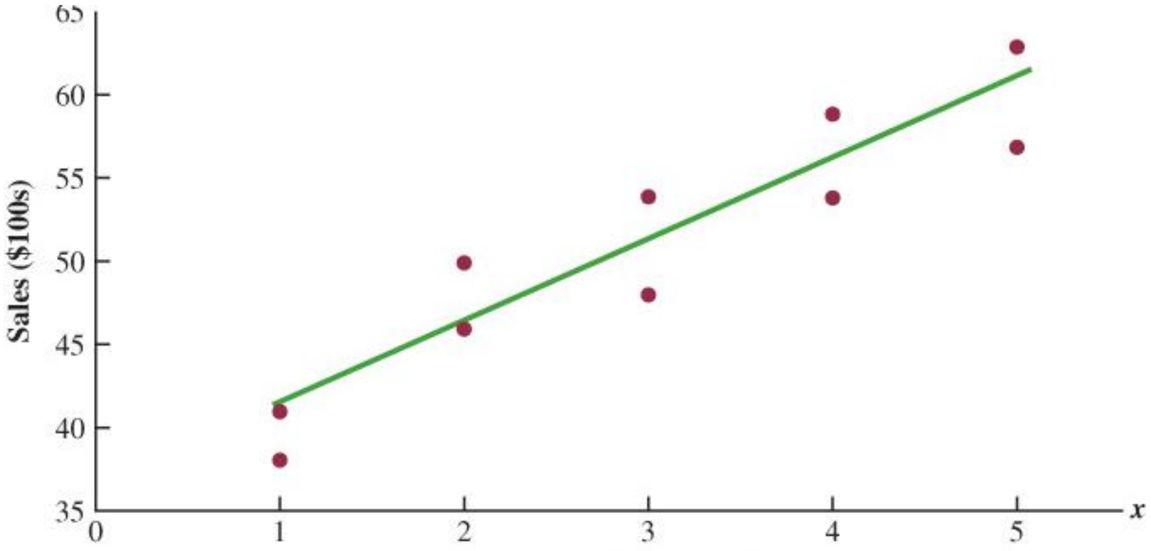
Scatter Diagram and Trendline

A **scatter diagram** is a graphical presentation of the relationship between two quantitative variables.

A **trendline** is a line that provides an approximation of the relationship.

Example DATAfile: *Electronics*

The horizontal axis (x) shows the number of commercials; the vertical axis (y), sales for an electronic store in San Francisco.



The relationship between the number of commercials and sales does not form a perfectly straight line. However, the general pattern of the points and the trendline suggest that the overall relationship is positive.

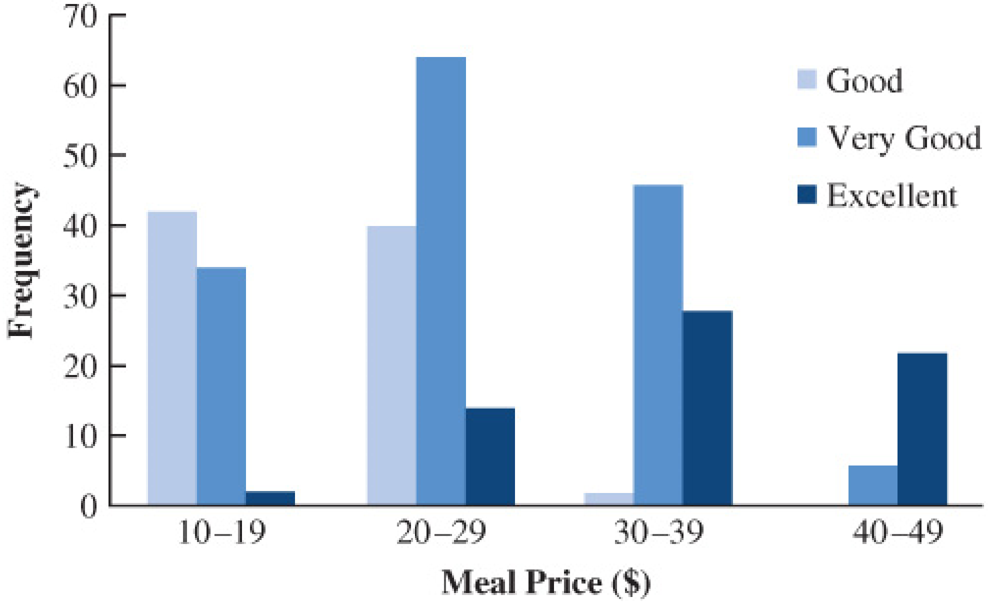
Side-by-Side Bar Chart

A **side-by-side bar chart** is a graphical display for depicting multiple bar charts on the same display.

We can use a side-by-side bar chart to represent the column percentages of the Zagat's restaurant reviews.

- Each cluster of bars represents one of the four classes of meal prices.
- Each bar within a cluster represents one of the three quality rating categories.

The chart makes even more apparent the shift toward higher quality ratings as the meal prices increase.



Stacked Bar Chart

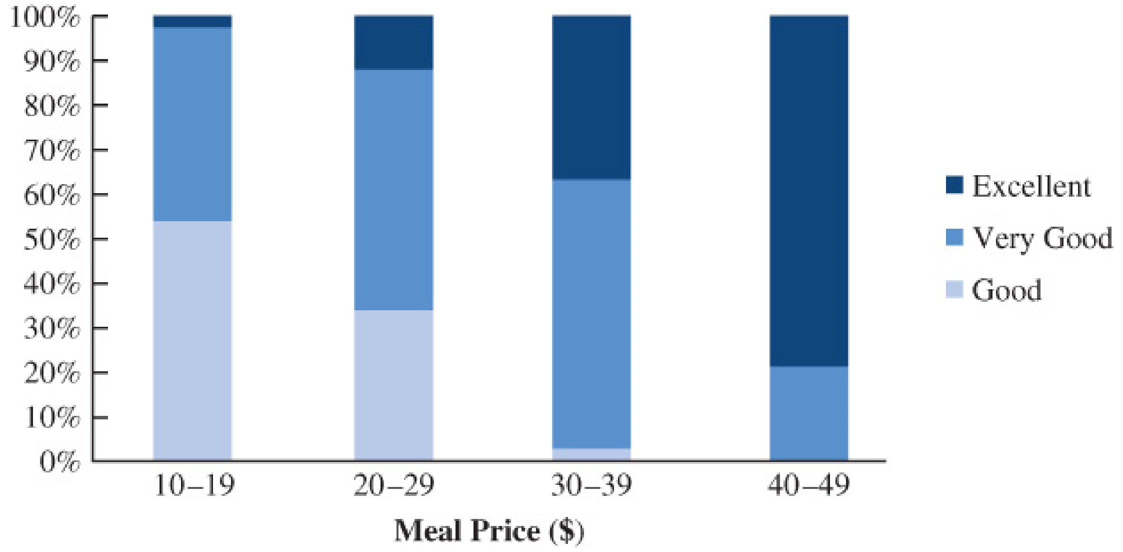
We can also use a **stacked bar chart** to compare two variables on the same display.

In a stacked bar chart, each bar is broken into rectangular segments of a different color showing the relative frequency of each class in a manner similar to a pie chart.

We can also use a stacked bar chart to represent the column percentages of the Zagat’s restaurant reviews.

Note that, because percentage frequencies are displayed, all bars are of the same height, extending to the 100% mark

The stacked bar chart shows even more clearly the shift toward higher quality ratings as meal prices increase.



Numerical Measures

Numerical Measures

If the measures are computed for data from a sample, they are called **sample statistics**.

If the measures are computed for data from a population, they are called **population parameters**.

A sample statistic is referred to as the **point estimator** of the corresponding population parameter.

Statistical software packages and spreadsheets can be used to develop the descriptive statistics presented in this chapter.

Mean

The most important measure of central location is the **mean**, the average of all the data values. For a sample with *n* observations, the formula for the sample mean is as follows.

$$\boldsymbol{\bar{x}} = \frac{\sum \boldsymbol{x_i}}{\boldsymbol{n}}$$ where x_i is the ith observation in a data set of size n

Example: consider the class size data for a sample of five college classes: 46, 54, 42, 46, 32.

Using the introduced notation, we have: x_1=46, x_2=54, x_3=42, x_4=46, and x_5=32.

To compute the sample mean, we write:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum x_1 + x_2 + x_3 + x_4 + x_5}{n} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

The sample mean $\bar{x}$ is the point estimator of the **population mean**, μ.

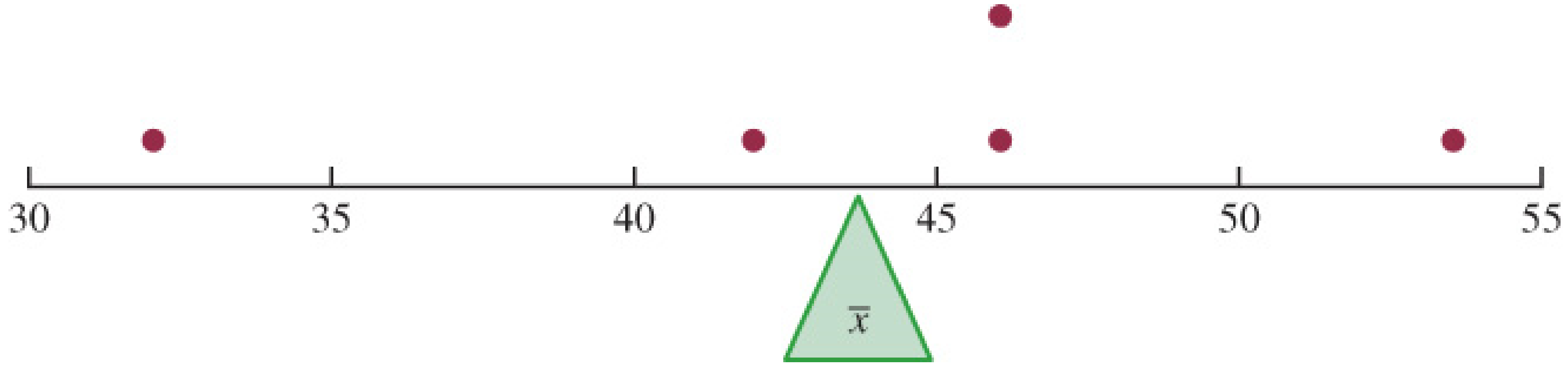
$$\boldsymbol{\mu} = \frac{\sum \boldsymbol{x_i}}{\boldsymbol{N}}$$ where N is the size of the population

The Mean as the Center of Balance for the Dot Plot

If you view the horizontal axis of the dot plot as a long narrow board in which each of the dots has the same fixed weight, the mean is the fulcrum (or pivot point) that balances the dot plot.

This is how a see-saw on a playground works; the only difference is that the see-saw is pivoted in the middle so that as one end goes up, the other end goes down.

In the dot plot, we are locating the pivot point based on the location of the dots.



Weighted Mean

In the formula for the mean, each data is given equal importance or weight.

In those instances where the mean is computed by giving each observation a weight that reflects its relative importance, we can calculate the **weighted mean** instead.

For a sample with *n* observations and weights w_i, the formula for the weighted mean is

$$\boldsymbol{\bar{x} = \frac{\sum w_i x_i}{\sum w_i}}$$

Example: calculate the weighted mean for a sample of five purchases of raw materials listed to the right.

Purchase	Cost per Pound ($)	Weight (lbs)
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.95	1000
5	3.25	800

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} =$$

$$\frac{1{,}200(3.00) + 500(3.40) + 2{,}750(2.80) + 1{,}000(2.95) + 800(3.25)}{1{,}200 + 500 + 2{,}750 + 1{,}000 + 800} = \frac{18{,}500}{6{,}250} = \$2.96$$

Median

The **median** is the value in the middle of a data set when data are arranged in ascending order.

To compute the median, first arrange the data in ascending order (smallest to largest value.)

a. If *n* is *odd*, the median is the middle value.

b. If *n* is *even*, the median is the average of the two middle values.

Example: consider the starting monthly salary for 12 business graduates.

First, we arrange the data in ascending order:

5710 5755 5850 5880 5880 5890 5920 5940 5950 6050 6130 6325

Because $n = 12$ is even, the median is the average of the 6th and 7th (middle) values.

$$Median = (5890 + 5920)/2 = 5905$$

Because the mean is influenced by extremely small and large data values, the median is the preferred measure of central location to report annual income and property value data.

Geometric Mean

The **geometric mean** is a measure of central location calculated by finding the nth root of the product of n values.

The general formula for the geometric mean, denoted $\bar{x}_g$, follows.

$$\bar{\boldsymbol{x}}_{\boldsymbol{g}} = \sqrt[\boldsymbol{n}]{(\boldsymbol{x}_{\boldsymbol{1}})(\boldsymbol{x}_{\boldsymbol{2}}) \ldots (\boldsymbol{x}_{\boldsymbol{n}})} = [(\boldsymbol{x}_{\boldsymbol{1}})(\boldsymbol{x}_{\boldsymbol{2}}) \ldots (\boldsymbol{x}_{\boldsymbol{n}})]^{\boldsymbol{1/n}}$$

The geometric mean is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results.)

It should be applied any time you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, etc.)

Other common applications include changes in populations of species, crop yields, pollution levels, and birth and death rates.

Mode

The **mode** of a data set is the value that occurs with the greatest frequency.

Example: consider again the starting monthly salary for 12 business graduates.

5710 5755 5850 5880 5880 5890 5920 5940 5950 6050 6130 6325

$5,880 is the mode because it is the only monthly starting salary that occurs more than once.

It may happen that the greatest frequency occurs at two or more different values.
In these instances, more than one mode exists.
In the presence of multiple modes, we define the following two cases:

- If the data have exactly two modes, the data are said to be *bimodal*.
- If the data have more than two modes, the data are said to be *multimodal*.

Percentile

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.

Admission test scores for colleges and universities are frequently reported in terms of percentiles.

The **$\boldsymbol{p}$th percentile** of a data set is a value such that at least *p* percent of the items take on this value or less and at least $(100–p)$ percent of the items take on this value or more.

To calculate the pth percentile of a data set, we must first arrange the data in ascending order so that the smallest value in the data set is in position 1, the next one in position 2, and so on.

The **location** of the pth percentile, denoted L_p, is computed using the following equation:

$$L_p = \frac{p}{100}(n+1)$$

Now, we are ready to calculate the pth percentile. Let us do that with an example.

3.1 An Application of the Percentile

Compute the 80th percentile for the sample of 12 business graduates' starting salaries.

With the data arranged in ascending order, we indicate the position of each observation directly below its value.

	5710	5755	5850	5880	5880	5890	5920	5940	5950	6050	6130	6325
Position	1	2	3	4	5	6	7	8	9	10	11	12

The location of the 80th percentile is

$$L_{80} = \frac{p}{100}(n+1) = \frac{80}{100}(12+1) = 10.4$$

The interpretation of $L_{80} = 10.4$ is that the 80th percentile is 40% of the way between the values in position 10 and 11.

$$\text{80th percentile} = 6050 + 0.4(6130 - 6050) = 6050 + 0.4(80) = 6082$$

Quartile

Quartiles are specific percentiles that divide the data set into four parts, with each part containing approximately 25% of the observations. Quartiles are defined as follows:

$Q_1 =$ first quartile, or 25th percentile

$Q_2 =$ second quartile, or 50th percentile (also the median)

$Q_3 =$ third quartile, or 75th percentile

The procedure for computing percentiles can be also used to compute quartiles. Let us calculate Q_1 and Q_3 for the sample of 12 business graduates' starting salaries.

First, we calculate the locations

$$L_{25} = \frac{25}{100}(12+1) = 3.25 \quad \text{and} \quad L_{75} = \frac{75}{100}(12+1) = 9.75$$

The calculations of Q_1 and Q_3 follow:

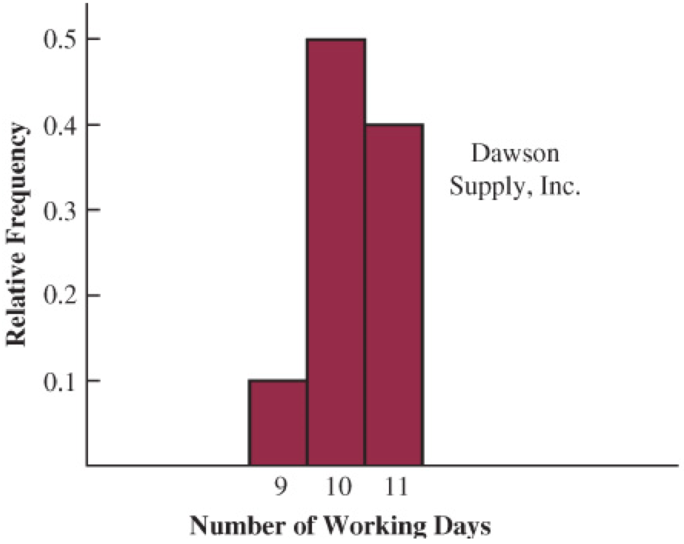
$$Q_1 = 5850 + 0.25(5880 - 5850) = 5857.5 \qquad Q_3 = 5950 + 0.75(6050 - 5950) = 6025$$

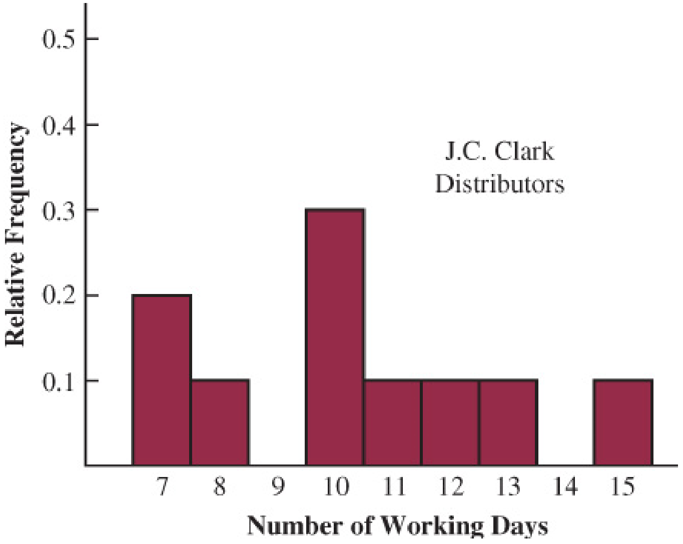
Measures of Variability

It is desirable to consider measures of variability (dispersion), as well as measures of location.

Consider the histograms for the number of days required to fill orders for two suppliers.

Although the mean number of days is 10 for both suppliers and the supplier to the right is able to fill some orders in as little as 7 to 8 days, the supplier to the left has a lower dispersion and demonstrates a higher degree of reliability in terms of making deliveries on schedule.





Range and Interquartile Range

Range

The **range** is the simplest measure of variability, and it is defined as

Range = Largest Value – Smallest Value

For the example of the business graduates' starting salaries, the range is

$$6325 - 5710 = 615$$

However, the **range sensitivity to extreme data values** makes it a poor choice to measure the dispersion in a data set.

Interquartile Range

The **interquartile range (IQR)** overcomes the dependency on extreme values considering the range for the middle 50% of the data.

The interquartile range is calculated as the difference between the third quartile, Q_3, and the first quartile, Q_1.

$$\boldsymbol{IQR = Q_3 - Q_1}$$

For the example of the business graduates' starting salaries, the *IQR* is

$$6025 - 5857.5 = 167.5$$

Variance

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 (x_i) and the mean ($\bar{x}$ for a sample, μ for a population.)

The difference between each x_i and the mean is called a *deviation about the mean*.

In the computation of the variance, the deviations about the mean are squared.

Population Variance

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

Where *N* is the population size.

Sample Variance

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

Where *n* is the sample size.

A Calculation of the Sample Variance

Let us calculate the sample variance of the class size for the sample of five college classes as presented in the previous section.

A summary of the data, including the computation of the deviations about the mean and the squared deviations about the mean, is shown in the table below.

The sum of squared deviations about the mean is $\sum(x_i - \bar{x})^2 = 256$.

With $n - 1 = 4$, the sample variance is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64$$

Note that, because the variance is a squared operator, its units are also squared.

In this case, the units are (students)^2, and not students.

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	4
54	44	100
42	44	4
46	44	4
32	44	144
		$\sum(x_i - \bar{x})^2 = 256$

Standard Deviation

The **standard deviation** is defined to be the positive square root of the variance.

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

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

In the previous example, we calculated the sample variance of the class size for the sample of five college classes as $s^2 = 64$.

Thus, the sample standard deviation of the class size is

$s = \sqrt{64} = 8$ students

Because the standard deviation is the square root of the variance, the units of the variance, $(\text{students})^2$, are converted to students in the standard deviation.

Thus, the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

Coefficient of Variation

The **coefficient of variation**, usually expressed as a percentage, measures how large the standard deviation is relative to the mean.

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

For the class size of the sample of five college classes, we found a sample mean of 44 and a sample standard deviation of 8.

The coefficient of variation is

$$\left(\frac{8}{44} \times 100\right)\% = 18.2\%$$

In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean.

z-Scores

In addition to measures of location, variability, and shape, *measures of relative location* help us determine how far a particular value is from the mean.

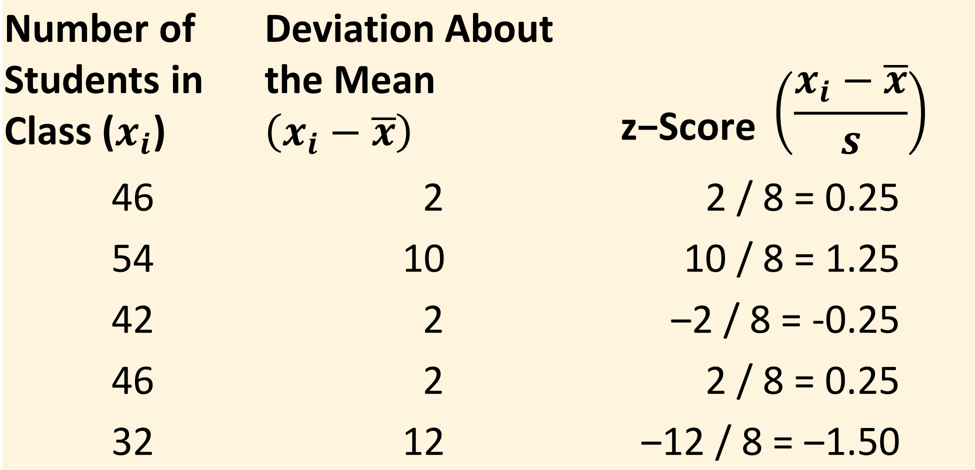
The ***z*-score**, often called the *standardized value*, denotes the number of standard deviations, *s* a data value x_i is from the mean, $\bar{x}$.

$$\boldsymbol{z_i = \frac{x_i - \bar{x}}{s}}$$

The *z*-scores for the class size data from the previous section are computed to the right.

For example, for $x_1 = 46$, the *z*-score is

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{46 - 44}{8} = \frac{2}{8} = 0.25$$



Number of Students in Class ($\boldsymbol{x_i}$)	**Deviation About the Mean ($\boldsymbol{x_i - \bar{x}}$)**	**z–Score $\boldsymbol{\left(\frac{x_i - \bar{x}}{s}\right)}$**
46	2	2 / 8 = 0.25
54	10	10 / 8 = 1.25
42	2	−2 / 8 = -0.25
46	2	2 / 8 = 0.25
32	12	−12 / 8 = −1.50

Distribution Shape

Skewness is an important numerical measure of the shape of a distribution.

Because of its complexity, skewness is usually calculated with the help of statistical software.

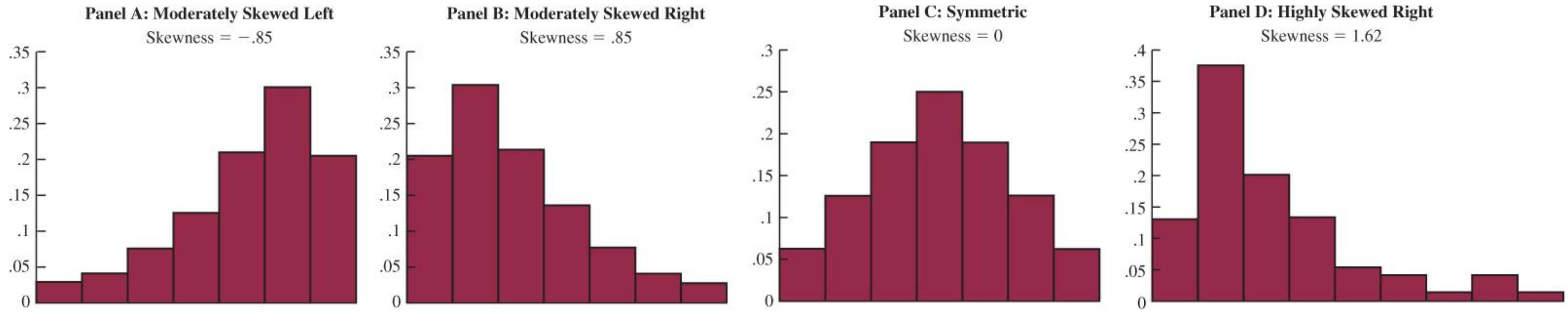
$$\text{skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Panel A: a distribution moderately skewed to the left has negative skewness.

Panel B: a distribution moderately skewed to the right has positive skewness.

Panel C: a symmetric distribution has the mean equal to the median, and skewness $= 0$.

Panel D: a distribution highly skewed to the right has a larger positive skewness.



Panel A: Moderately Skewed Left

Skewness = −.85

Panel B: Moderately Skewed Right

Skewness = .85

Panel C: Symmetric

Skewness = 0

Panel D: Highly Skewed Right

Skewness = 1.62

Chebyshev’s Theorem

Chebyshev’s theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean:

> **At least $(1 - 1/z^2)$ of the data values must be within *z* standard deviations of the mean, where *z* is any value greater than 1.**

Some of the implications of the theorem for $z = 2$ and 3 are:

- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89% of the data values must be within $z = 3$ standard deviations of the mean.

Example: the midterm test scores for 100 students in a college business statistics course had $\bar{x} = 70$ and $s = 5$. How many students had test scores between 58 and 82?

We have: $z = (58 - 70)/5 = -2.4$, and $z = (82 - 70)/5 = +2.4$

Applying Chebyshev’s Theorem with $z = 2.4$, we have: $1 - 1/z^2 = 1 - 1/2.4^2 = 0.826$

Thus, at least 82.6% of the scores are between 58 and 82.

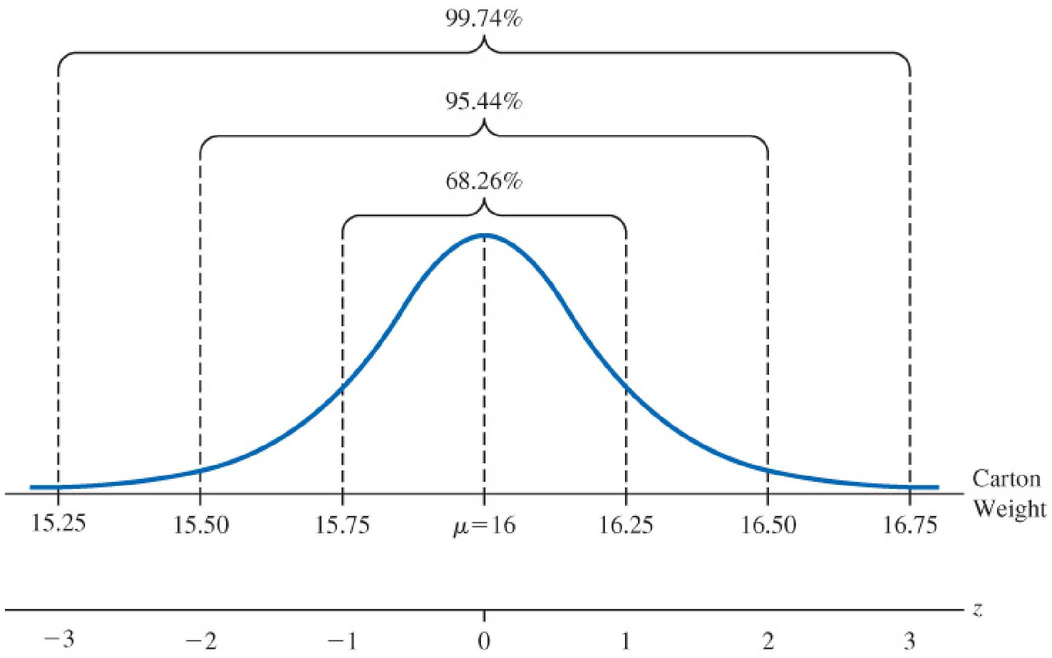
Empirical Rule

When the data are believed to approximate a symmetric bell-shaped distribution, the **empirical rule** can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

For data having a bell-shaped distribution:

- ~68% of the data values will be within one standard deviation of the mean.
- ~95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.

The empirical rule is based on the normal Distribution.



An Application of the Empirical Rule

Liquid detergent cartons are filled automatically on a production line. Filling weights have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is 0.25 ounces, we can use the empirical rule to conclude:

1. ~68% of the filled cartons will have weights between 15.75 and 16.25 oz.
2. ~95% of the filled cartons will have weights between 15.50 and 16.50 oz.

We can use this information to conclude approximately how many filled cartons will

- weigh between 16 and 16.25 oz:
  Because the distribution is symmetric, from point 1 we have $68\%/2 = 34\%$.
- weigh between 15.50 and 16 oz:
  Because the distribution is symmetric, from point 2 we have $95\%/2 = 47.5\%$.
- weigh less than 15.50 oz:
  Because 50% of the filled cartons have weights less than 16 oz. and 47.5% have weights between 15.50 and 16 ounces, it follows that 2.5% weigh less than 15.50 oz.

Detecting Outliers with *z*-Scores

An **outlier** is an unusually small or unusually large value in a data set.

Care should be taken when handling outliers, as they might be:

- an incorrectly recorded data value
- a data value that was incorrectly included in the data set
- a correctly recorded data value that belongs in the data set

A data value with a *z*-score less than –3 or greater than +3 might be considered an outlier.

For this example, refer to the starting monthly salary for the 12 business graduates, with mean, $\bar{x} = 5940$, and standard deviation, $s = 165.65$.

The smallest value in the data set, 5710, has $z = (5710 - 5940)/165.65 = -1.39$, and the largest value in the data set, 6325, has $z = (6325 - 5940)/165.65 = 2.32$.

Because all the values are within three standard deviations of the mean (*z*-scores within ±3), we conclude that there is no evidence of outliers.

Alternative Method for Detecting Outliers

Another approach to identifying outliers is based upon the values of the first and third quartiles (Q_1 and Q_3) and the interquartile range (*IQR*).

For this example, refer to the starting monthly salary for the 12 business graduates.

To use this method, we first compute the following lower and upper limits:

Lower Limit $= Q_1 - 1.5(IQR) = 5857.5 - 1.5(167.5) = 5606.25$

Upper Limit $= Q_3 + 1.5(IQR) = 6025 + 1.5(167.5) = 6276.25$

Looking at the starting monthly salary for the 12 business graduates, we see that:

There are no starting salaries lower than the Lower Limit = 5606.25.

There is one starting salary, 6325, that is greater than the Upper Limit = 6276.25.

Thus, using this alternate approach, 6325 is considered to be an outlier.

Five-Number Summary

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

1. Smallest value
2. First quartile (Q_1)
3. Median (Q_2)
4. Third quartile (Q_3)
5. Largest value

As an example, consider the monthly starting salary for the 12 business graduates.

The smallest value is 5710 and the largest value is 6325. We previously computed the quartiles ($Q_1 = 5857.5$; $Median = Q_2 = 5905$; and $Q_3 = 6025$).

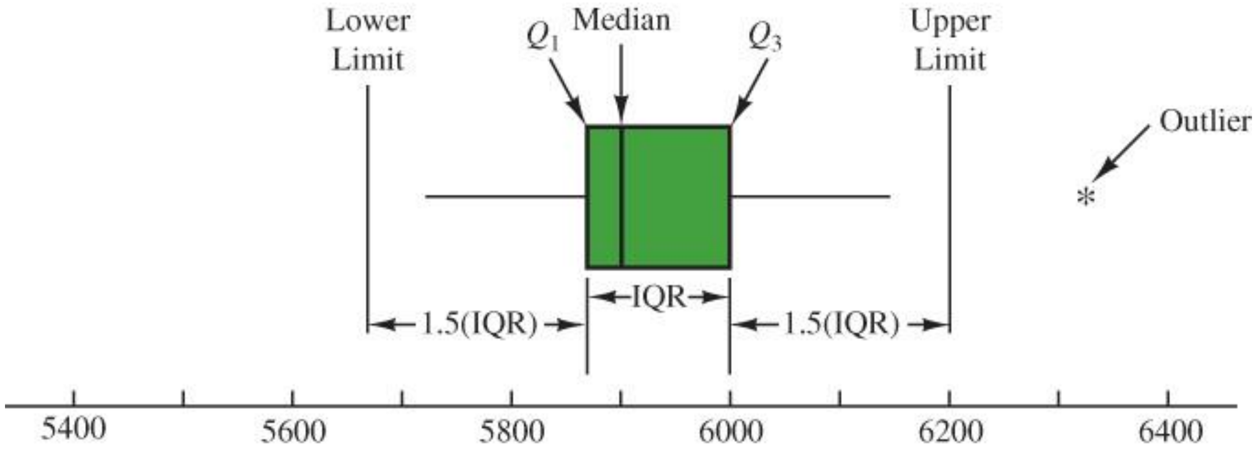
The five-number summary indicates that the starting salaries in the sample are between 5710 and 6325, approximately 50% of the starting salaries are between 5857.5 and 6025, and the median or middle value is 5905.

Boxplot

A **boxplot** is a graphical display of data based on a five-number summary.

The steps used to construct the boxplot for the monthly starting salary data are:

1. Draw a box between $Q_1 = 5857.5$ and $Q_3 = 6025$.
2. Draw a vertical line in the box at the location of the median, $Q_2 = 5905$.
3. Using $IQR = 167.5$, compute the lower and upper limits for detecting *outliers* (note that the limits are usually not drawn.)
4. Extend horizontal lines (*whiskers*) away from the box and up to the smallest and largest values that lie within the limits.
5. Use a small asterisk (*) to represent outliers.



A Comparative Analysis Using Boxplots

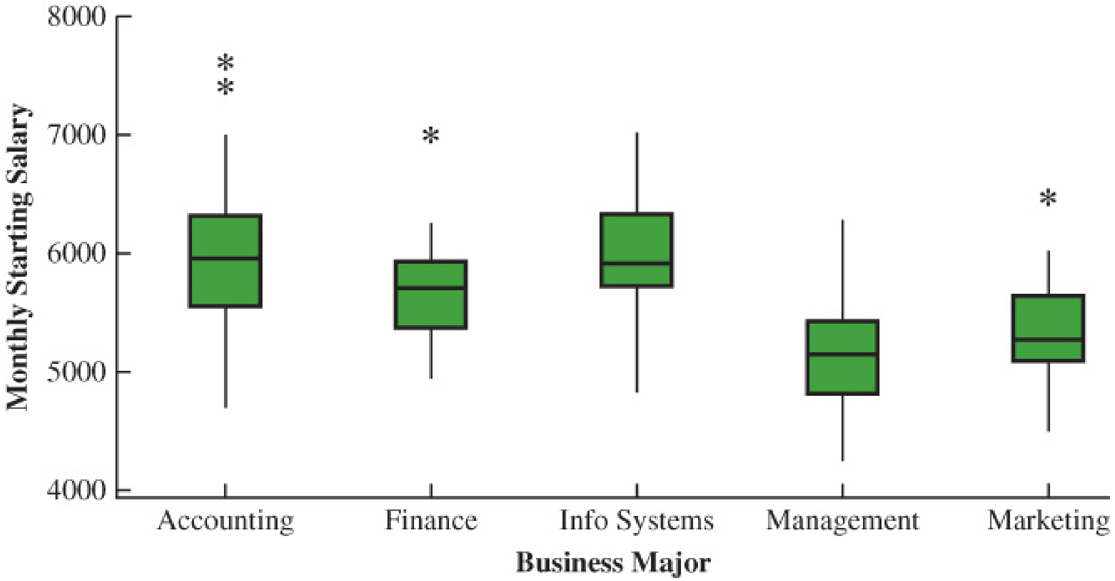
Boxplots can be used to provide a graphical summary of two or more groups and facilitate visual comparisons among the groups.

As an application, consider the major and starting salary data for a new sample of 111 recent business school graduates (DATAfile: *MajorSalaries*.)

Note that the boxplots for each major have been drawn vertically, as is often the case.

A comparative analysis of the boxplots reveals:

- Accounting has the higher salaries, management and marketing the lower.
- Based on the median, Accounting and Information Systems have the higher median salaries, followed by Finance.
- High salary outliers exist for accounting, finance, and marketing majors.



Measures of Association Between Two Variables

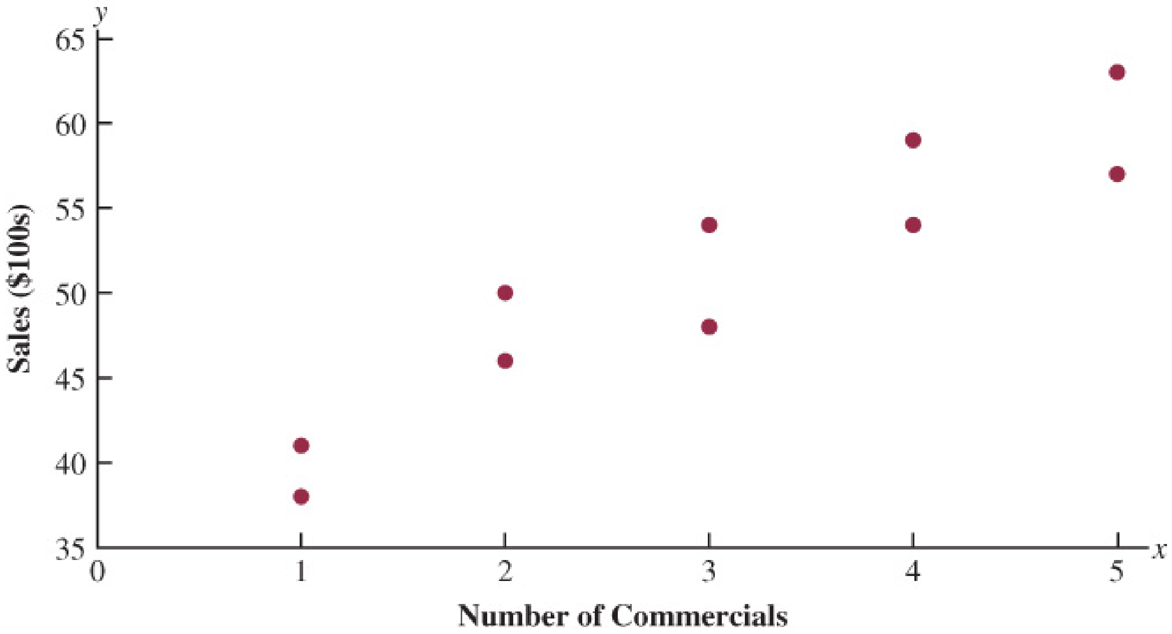
Often a manager or decision-maker is interested in the relationship between two variables.

Covariance and **correlation coefficient** are two descriptive measures of the *linear association* between two variables.

DATAfile: *Electronics*

As an application, we consider a sample of $n = 10$ weekly data for an electronics store in San Francisco.

The scatter diagram to the right suggests a *positive and linear* relationship between the number of weekend television commercials and the sales (in $100s) at the store during the following week.



Covariance

For a sample of size *n* with observations (x_i, y_i), and $i = 1 \dots n$, the **sample covariance** is defined as

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

The table to the right shows the detailed calculations for the sample covariance of the electronics store data:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{9} = 11$$

The **population covariance** for a population of size N is

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

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
30	510	0	0	99

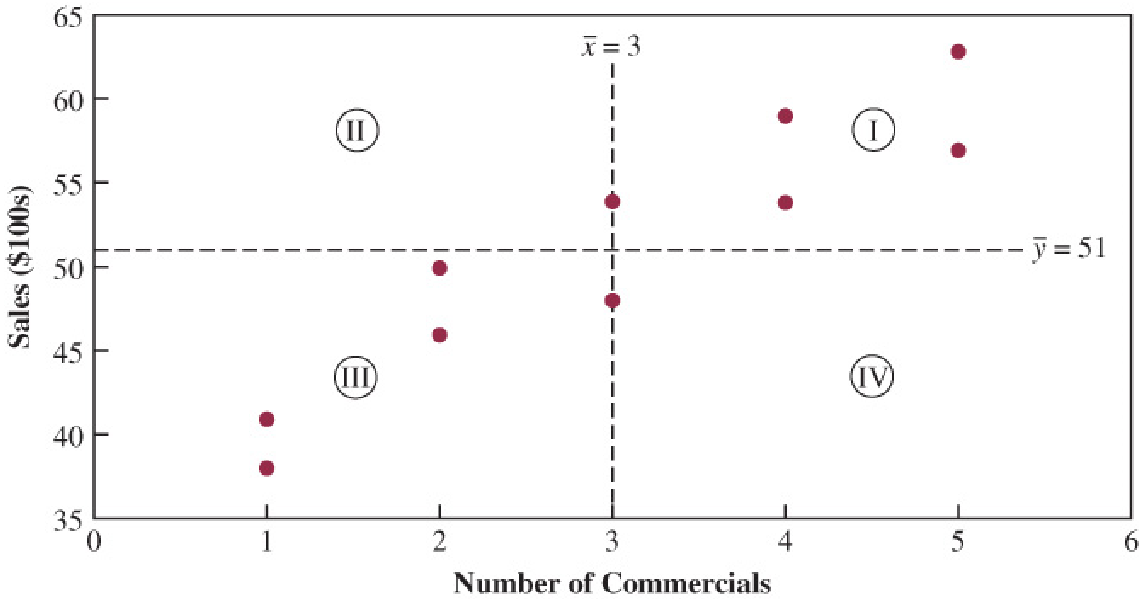
Interpretation of Sample Covariance

The figure shown to the right is the scatter diagram for the electronics store data, in which two dashed lines, a vertical line at $\bar{x} = 3$, and a horizontal line at $\bar{y} = 51$, dissect the diagram into four quadrants, labeled from I to IV.

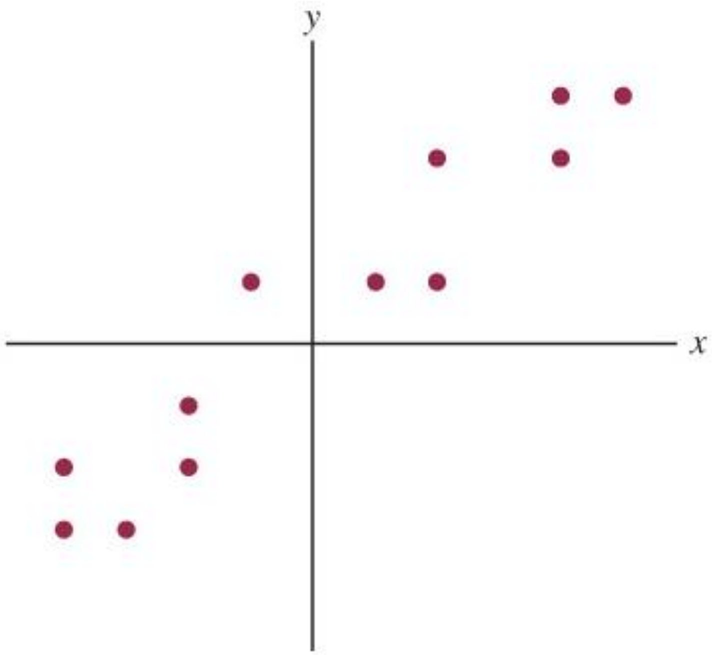
Because the covariance is related to the $(x_i - \bar{x})(y_i - \bar{y})$ term, it follows that quadrants:

- I and III have a *positive* contribution to the covariance, because $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have the same sign.
- II and IV have a *negative* contribution to the covariance, because $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have opposite sign.

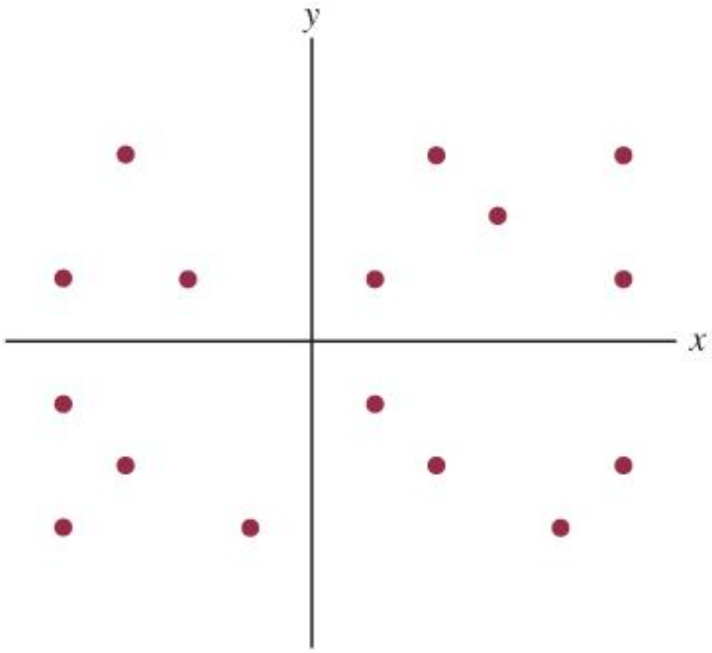
The scatter diagram for the electronics store data shows that none of the data appear in quadrants II or IV. Thus, s_{xy} is positive.



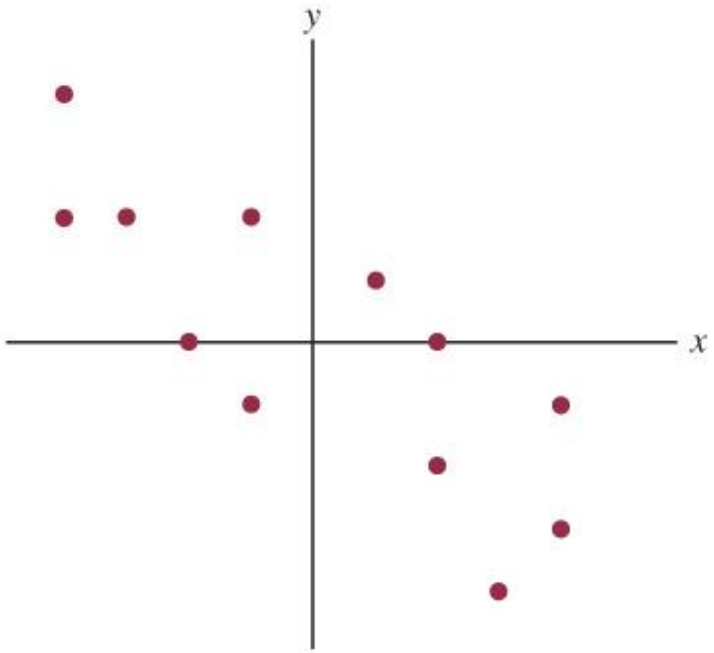
Relationship Between Pattern and Covariance



s_{xy} Positive:
(x and y are positively
linearly related)



s_{xy} Approximately 0:
(x and y are not
linearly related)



s_{xy} Negative:
(x and y are negatively
linearly related)

Correlation Coefficient

The **correlation coefficient** is a measure of the relationship between *x* and *y* that is not affected by the units of measurement.

The *Pearson product moment correlation coefficient* is defined as

Sample Data

$$\boldsymbol{r_{xy} = \frac{s_{xy}}{s_x s_y}}$$

Where

r_{xy} = sample correlation coefficient

s_{xy} = sample covariance

s_x = sample standard deviation of *x*

s_y = sample standard deviation of *y*

Population Data

$$\boldsymbol{\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}}$$

Where

ρ_{xy} = population correlation coefficient

σ_{xy} = population covariance

σ_x = population standard deviation for *x*

σ_y = population standard deviation for *y*

Interpretation of the Correlation Coefficient

Let us now compute the sample correlation coefficient for the San Francisco electronics store.

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49 \qquad S_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

We already know that $s_{xy} = 11$. Thus, the sample correlation coefficient is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = 0.93$$

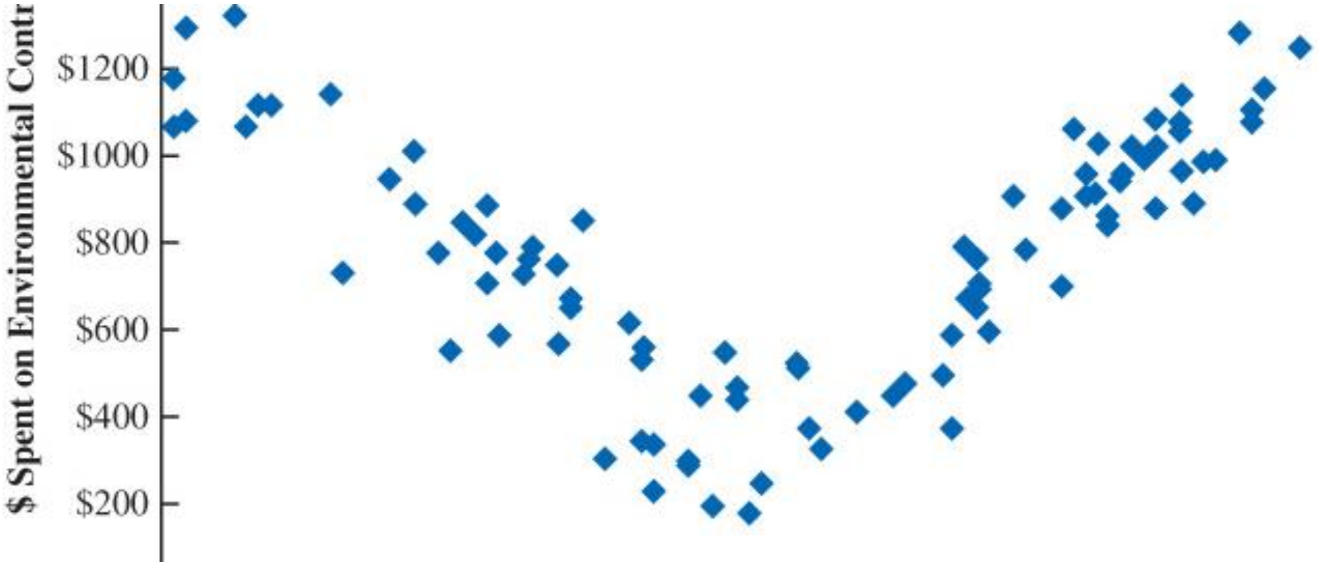
- The correlation coefficient is always between −1 and +1.
- The sign of the correlation coefficient matches the direction of the association.
- The closer to +1, the stronger the positive linear association between *x* and *y*.
- The closer to −1, the stronger the negative linear association between *x* and *y*.
- The closer to 0, the weaker the linear association between *x* and *y*.

A Note on Linear Association

The correlation coefficient measures only the strength of the *linear association* between two quantitative variables.

The sample correlation coefficient for the data shown here is $r_{xy} = -0.007$ and indicates that there is no linear relationship between the two variables.

However, the scatter diagram provides strong visual evidence of a *nonlinear* relationship


$ Spent on Environmental Contr
$1200
$1000
$800
$600
$400
$200