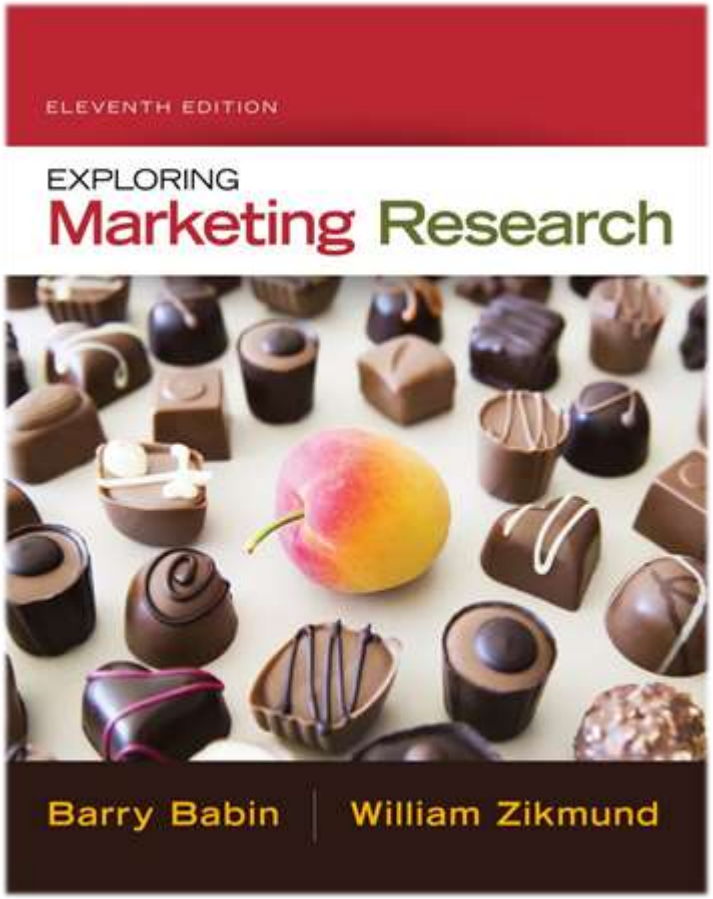


Chapter 13
Big Data Basics: Describing Samples and Populations

LEARNING OUTCOMES

After studying this chapter, you should

1. Use basic descriptive statistics to analyze data and make basic inferences about population metrics
2. Distinguish among the concepts of population, sample, and sampling distributions
3. Explain the central-limit theorem
4. Use confidence intervals to express inferences about population estimates
5. Understand major issues in specifying sample size
6. Know how to assess the potential for nonresponse bias

Introduction

- All the statistics in this chapter are univariate in the sense that only one variable is involved
  - Provides a good background for understanding equations related to sample size requirements
- These equations complement the material in the previous chapter
  - Provides us with a better understanding of the usefulness of a particular sample

Descriptive Statistics and Basic Inferences

- Raw data are simply numbers and words – with little meaning
- The most basic statistical tools for summarizing information from data include:
  - Frequency distributions
  - Proportions
  - Measures of central tendency and dispersion
- Metrics provide a means of comparison

Descriptive Statistics and Basic Inferences (cont'd.)

- The combination of summary metrics from basic statistics and a valid sample proves vital in making effective marketing and business decisions
- Inferential statistics allow inferences about a whole population from a sample
- Two applications of statistics:
  - To describe characteristics of the population or sample and
  - To generalize from a sample to a population

What Are Sample Statistics and Population Parameters?

- The primary purpose of inferential statistics is to make a judgment about the population
- The sample is a subset or relatively small fraction of the total number of elements in the population
- Sample statistics are measures computed from sample data

What Are Sample Statistics and Population Parameters? (cont'd.)

- Population parameters are measured characteristics of a specific population
- We generally use Greek lowercase letters to denote population parameters (e.g., μ or σ) and English letters to denote sample statistics (e.g., X or S)

Frequency Distributions

- Constructing a frequency table or frequency distribution is one of the most common means of summarizing a set of data
  - The frequency of a value is the number of times a particular value of a variable occurs
  - A distribution of relative frequency, or a percentage distribution, is developed by dividing the frequency of each value by the total number of observations, and multiplying the result by 100
  - Probability is the long-run relative frequency with which an event will occur

EXHIBIT 13.1 Frequency Distribution of Deposits

Amount	Frequency (Number of People Who Hold Deposits in Each Range)
Under $3,000	499
$3,000–$4,999	530
$5,000–$9,999	562
$10,000–$14,999	718
$15,000 or more	811
	3,120

Source: © Cengage Learning 2013.

EXHIBIT 13.2 Percentage Distribution of Deposits

Amount	Percent (Percentage of People Who Hold Deposits in Each Range)
Under $3,000	16
$3,000–$4,999	17
$5,000–$9,999	18
$10,000–$14,999	23
$15,000 or more	26
	100

Source: © Cengage Learning 2013.

Proportions

- A proportion indicates the percentage of population elements that successfully meet some standard on the particular characteristic
  - May be expressed as a percentage, a fraction, or a decimal number

EXHIBIT 13.3 Probability Distribution of Deposits

Amount	Probability
Under $3,000	.16
$3,000–$4,999	.17
$5,000–$9,999	.18
$10,000–$14,999	.23
$15,000 or more	.26
	1.00

Source: © Cengage Learning 2013.

Top-Box/Bottom-Box Scores

- A top box score generally refers to the portion of respondents who choose the most favorable response toward a company – the portion that would highly recommend a business to others or the portion expressing the highest likelihood of doing business again
  - The logic is that respondents who choose the most extreme response are really quite unique compared to the others

Top-Box/Bottom-Box Scores (cont'd.)

- Managers should examine the bottom-box score – the portion of respondents who choose the least favorable response to some question about customer opinion
  - More diagnostic of customer problems
  - Often signals a need for some managerial reaction

Central Tendency Metrics: The Mean

- The arithmetic average
- A common measure of central tendency
- The sum of all the observations divided by the number of observations
  - A sample mean, $\bar{X}$ (read as "X bar"), can be calculated when there is not enough data to calculate the population mean, μ
- Can sometimes be misleading, particularly when extreme values or outliers are present

EXHIBIT 13.4 Number of Sales Calls per Day by Salesperson

Index		Salesperson	Variable		Number of Calls
1	=	Mike	X_1	=	4
2	=	Patty	X_2	=	3
3	=	Billie	X_3	=	2
4	=	Bob	X_4	=	5
5	=	John	X_5	=	3
6	=	Frank	X_6	=	3
7	=	Chuck	X_7	=	1
8	=	Samantha	X_8	=	5
		Total		=	26

Source: © Cengage Learning 2013.

Central Tendency Metrics: The Median

- The midpoint of the distribution, or the 50th percentile
- The value below which half the values in the sample fall
- A better measure of central tendency in the presence of extreme values or outliers

Central Tendency Metrics: The Mode

- The measure of central tendency that merely identifies the value that occurs most often
- Determined by listing each possible value and noting the number of times each value occurs
- Used for data that is less than interval, with one large peak

Dispersion Metrics

- Accurate analysis of data also requires knowing the tendency of observations to depart from the central tendency
- Another way to summarize the data is to calculate the dispersion of the data, or how the observations vary from the mean

EXHIBIT 13.5 Sales Levels for Two Products with Identical Average Sales

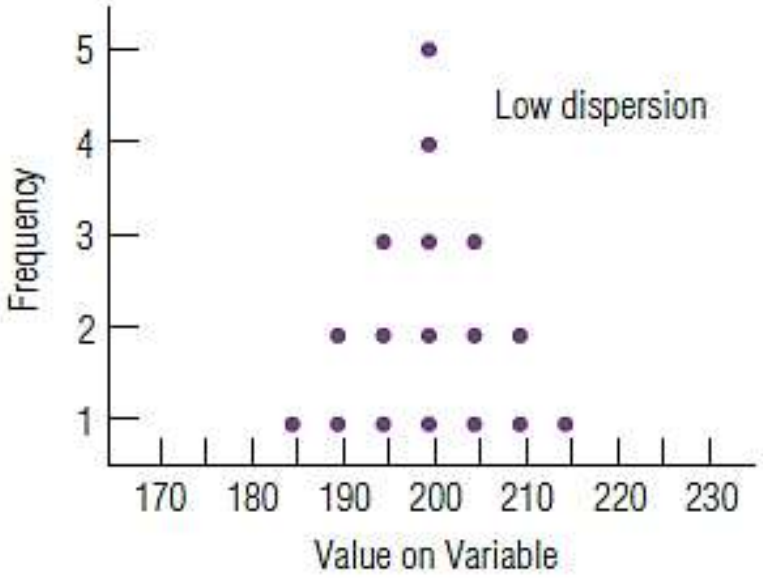
	Units Product A	Units Product B
January	196	150
February	198	160
March	199	176
April	200	181
May	200	192
June	200	200
July	200	201
August	201	202
September	201	213
October	201	224
November	202	240
December	202	261
Average	**200**	**200**

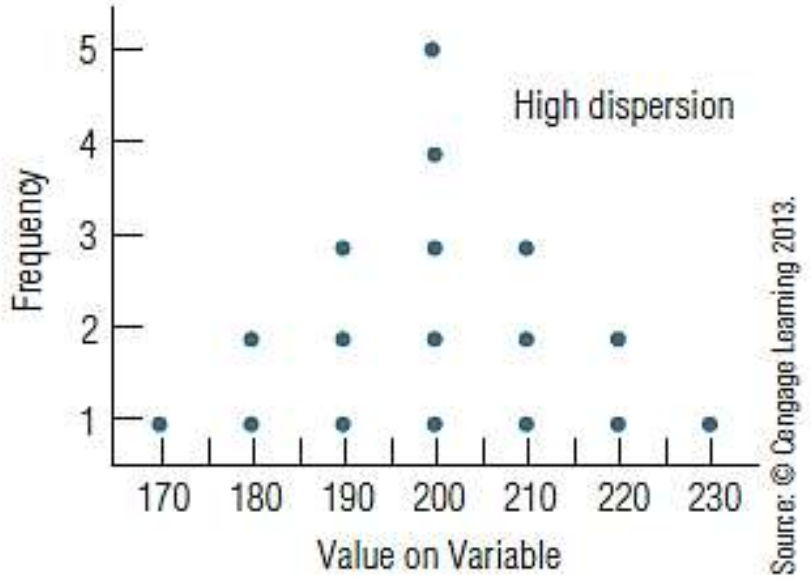
Source: © Cengage Learning 2013.

The Range

- The simplest measure of dispersion – the distance between the smallest and largest values of a frequency distribution
  - Does not take into account all the observations
  - Indicates the extreme values of the distribution
  - In a skinny distribution, values are a short distance from the mean; in a fat distribution values are spread out
- The interquartile range encompasses the middle 50 percent of the observations, i.e., the range between the bottom quartile and the top quartile

EXHIBIT 13.6 Low Dispersion versus High Dispersion





Source: © Cengage Learning 2013.

Deviation Scores

- A method of calculating how far any observation is from the mean is to calculate individual deviation scores
  - A deviation of any observation from the mean can be calculated by subtracting the mean from that observation

Why Use the Standard Deviation?

- It is perhaps the most valuable index of spread, or dispersion
- Other measures of dispersion that may be used:
  - Average deviation – determined by calculating the deviation score of each observation value (i.e., its difference from the mean) and summing these scores; then dividing by the sample size (*n*)
  - Variance – useful for describing the sample variability; will equal to zero if and only if each and every observation in the distribution is the same as the mean

Why Use The Standard Deviation? (cont'd.)

- Standard deviation
  - The square root of the variance for distribution is called the standard deviation
  - *S* is the symbol for the sample standard deviation

$$S = \sqrt{S^2} = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n - 1}}$$

EXHIBIT 13.7 Calculating a Standard Deviation: Number of Sales Calls per Day for Eight Salespeople

X	$(X-\bar{X})^1$	$(X-\bar{X})^2$
4	$(4 - 3.25) = .75$	.5625
3	$(3 - 3.25) = -.25$	.0625
2	$(2 - 3.25) = -1.25$	1.5625
5	$(5 - 3.25) = 1.75$	3.0625
3	$(3 - 3.25) = -.25$	.0625
3	$(3 - 3.25) = -.25$	.0625
1	$(1 - 3.25) = -2.25$	5.0625
5	$(5 - 3.25) = 1.75$	3.0625
$\sum$[a]	0[a]	13.5000

$$n = 8 \qquad \bar{X} = 3.25$$

$$S = \sqrt{\frac{\sum(X - X)^2}{n - 1}} = \sqrt{\frac{13.5}{8 - 1}} = \sqrt{\frac{13.5}{7}} = \sqrt{1.9286} = 1.3887$$

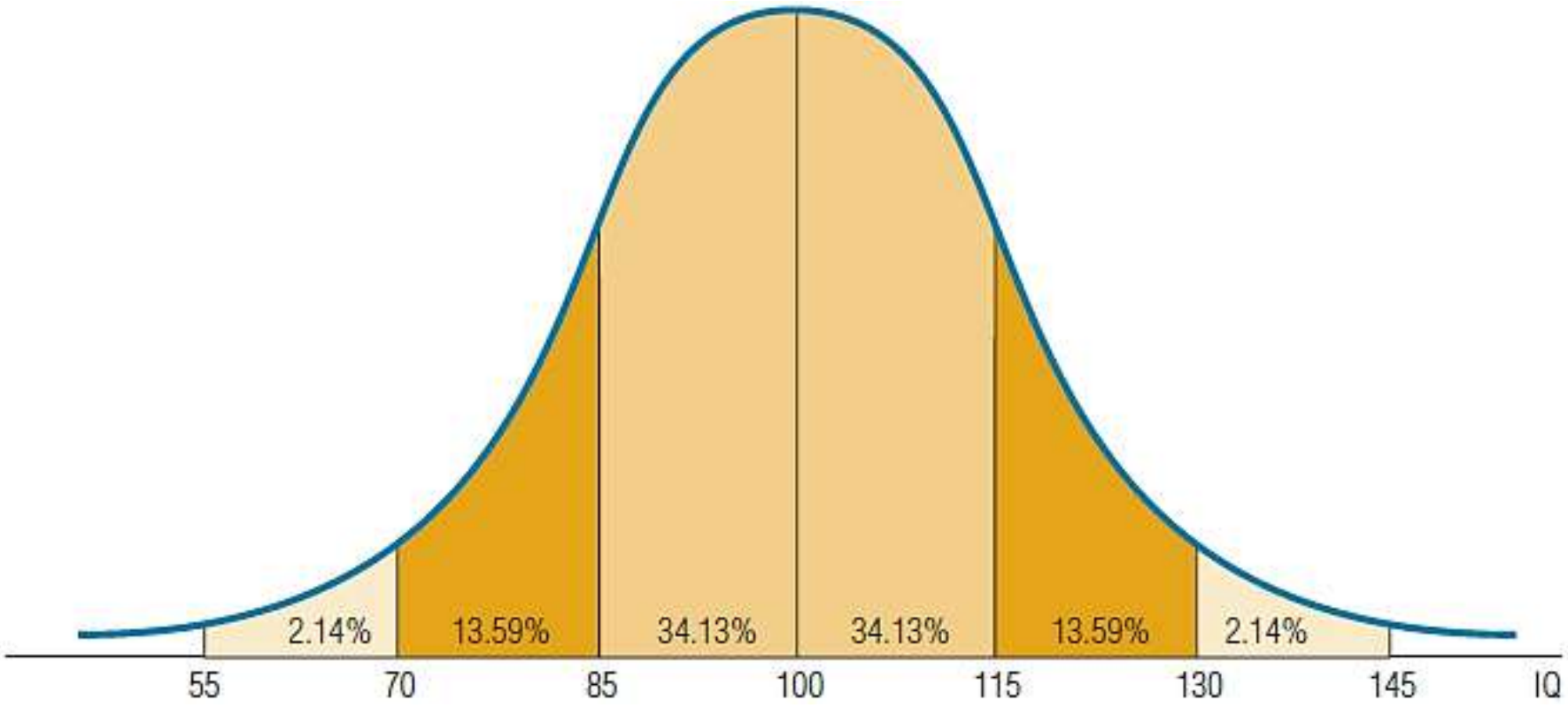
[a]The summation of this column is not used in the calculation of the standard deviation.

Source: © Cengage Learning 2013.

Distinguish between Population, Sample and Sample Distribution

- The normal distribution
  - One of the most common probability distributions in statistics
  - Also known as the normal curve
  - Bell-shaped
  - Almost all (99 percent) of its values are within ±3 standard deviations from its mean

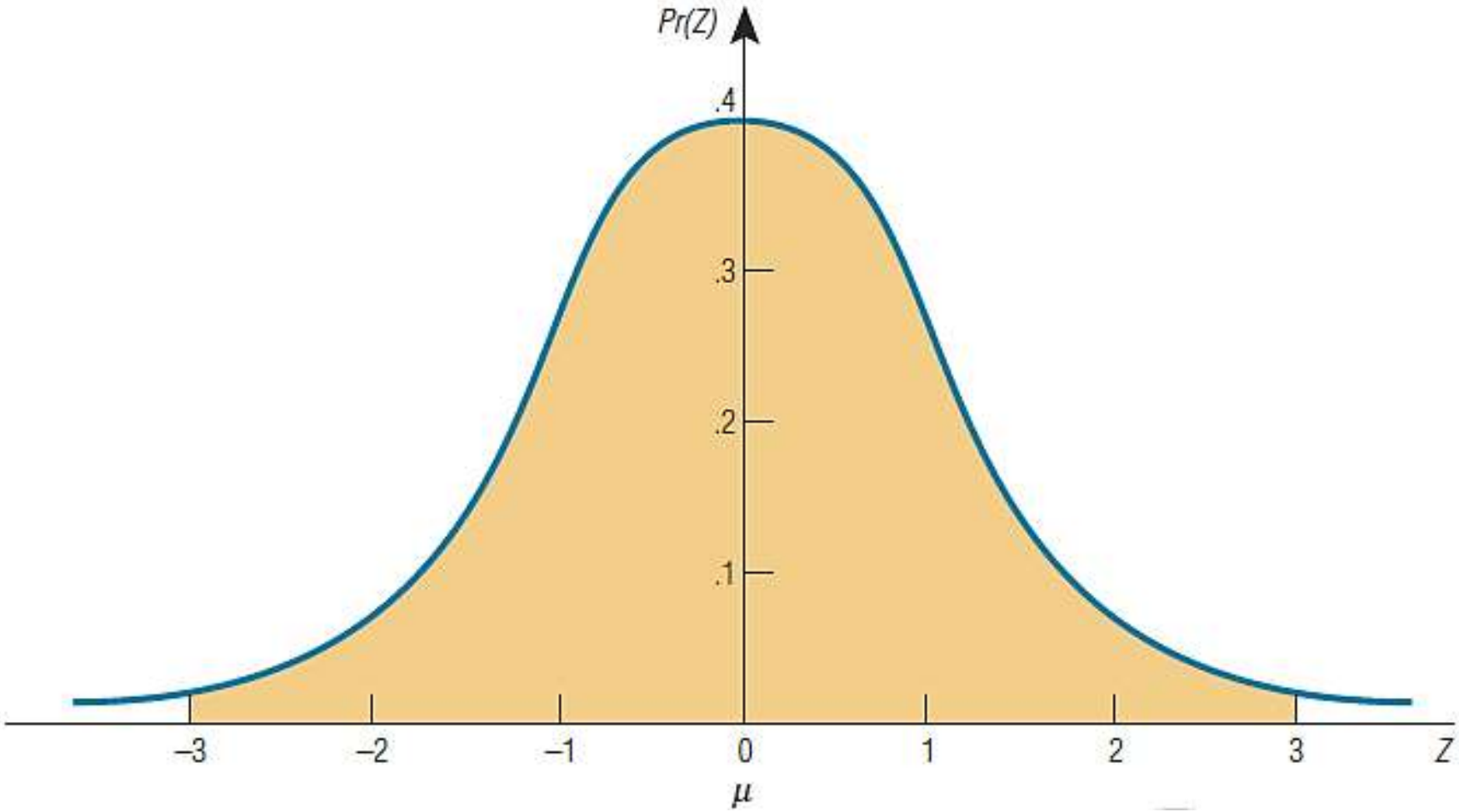
EXHIBIT 13.8 Normal Distribution: Distribution of Intelligence Quotient (IQ) Scores



The Standardized Normal Distribution

- A specific normal curve with several characteristics:
  - It is symmetrical about its mean
  - The mean identifies its highest point (the mode) and vertical line about which this curve is symmetrical
  - The normal curve has an infinite number of cases (it is a continuous distribution), and the area under the curve has a probability density equal to 1.0
  - The standardized normal distribution has a mean of 0 and a standard deviation of 1

EXHIBIT 13.9 Standardized Normal Distribution



The Standardized Normal Distribution and *Z*-Scores

- The standardized normal distribution is extremely valuable because we can translate or transform any normal variable, *X*, into the standardized value, *Z*
  - This has many pragmatic implications for the marketing researcher
  - A typical standardized normal table allows us to evaluate the probability of the occurrence of certain events without any difficulty

Computing *Z* Scores

- Subtract the mean from the value to be transformed, and divide by the standard deviation (all expressed in original units)
- In the formula, note that σ, the population standard deviation, is used for calculation:

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the hypothesized or expected value of the mean

Example: Standardized Values

- Suppose that a toy manufacturer has experienced mean sales, μ, of 9,000 units and a standard deviation, σ, of 500 units during the month of September
  - The production manager wishes to know if wholesalers will demand between 7,500 and 9,625 units during the month of September this year
  - Because there are no tables in the back of our textbook showing the distribution for a mean of 9,000 and a standard deviation of 500, we must transform our distribution of toy sales, X, into the standardized form with our simple formula

Standardized Values Example (cont'd.)

$$Z = \frac{7,500 - 9,000}{500} = -3.00 \qquad Z = \frac{9,625 - 9,000}{500} = 1.25$$

- When Z = 3.00, the area under the curve (probability) equals .499
- When Z = 1.25, the area under the curve (probability) equals .394
- Thus, the total area under the curve is .499 + .394 = .893
- The area under the curve portraying this computation is the shaded area in Exhibit 13.12
- Thus, the sales manager knows there is a .893 probability that sales will be between 7,500 and 9,625

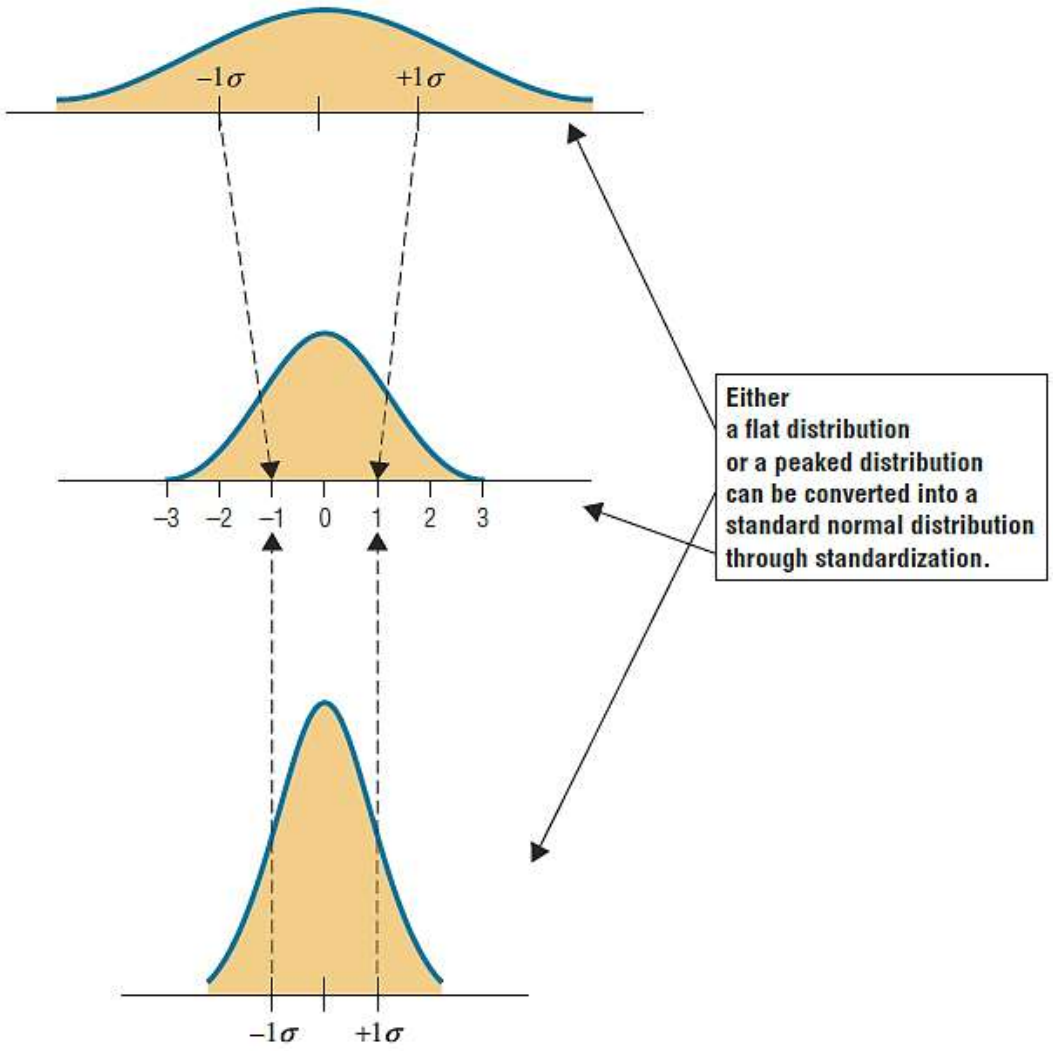
EXHIBIT 13.10 Standardized Normal Table: Area under Half of the Normal Curve

Z Standard Deviations from the Mean (Units)	*Z* Standard Deviations from the Mean (Tenths of Units)[a]									
	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.0	.000	.040	.080	.118	.155	.192	.226	.258	.288	.315
1.0	.341	.364	.385	.403	.419	.433	.445	.455	.464	.471
2.0	.477	.482	.486	.489	.492	.494	.495	.496	.497	.498
3.0	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499

[a] Area under the segment of the normal curve extending (in one direction) from the mean to the point indicated by each row-column combination. For example, about 68 percent of normally distributed events can be expected to fall within 1.0 standard deviation on either side of the mean (0.341 × 2). An interval of almost 2.0 standard deviations around the mean will include 95 percent of all cases (.477 + .477).

Source: © Cengage Learning 2013.

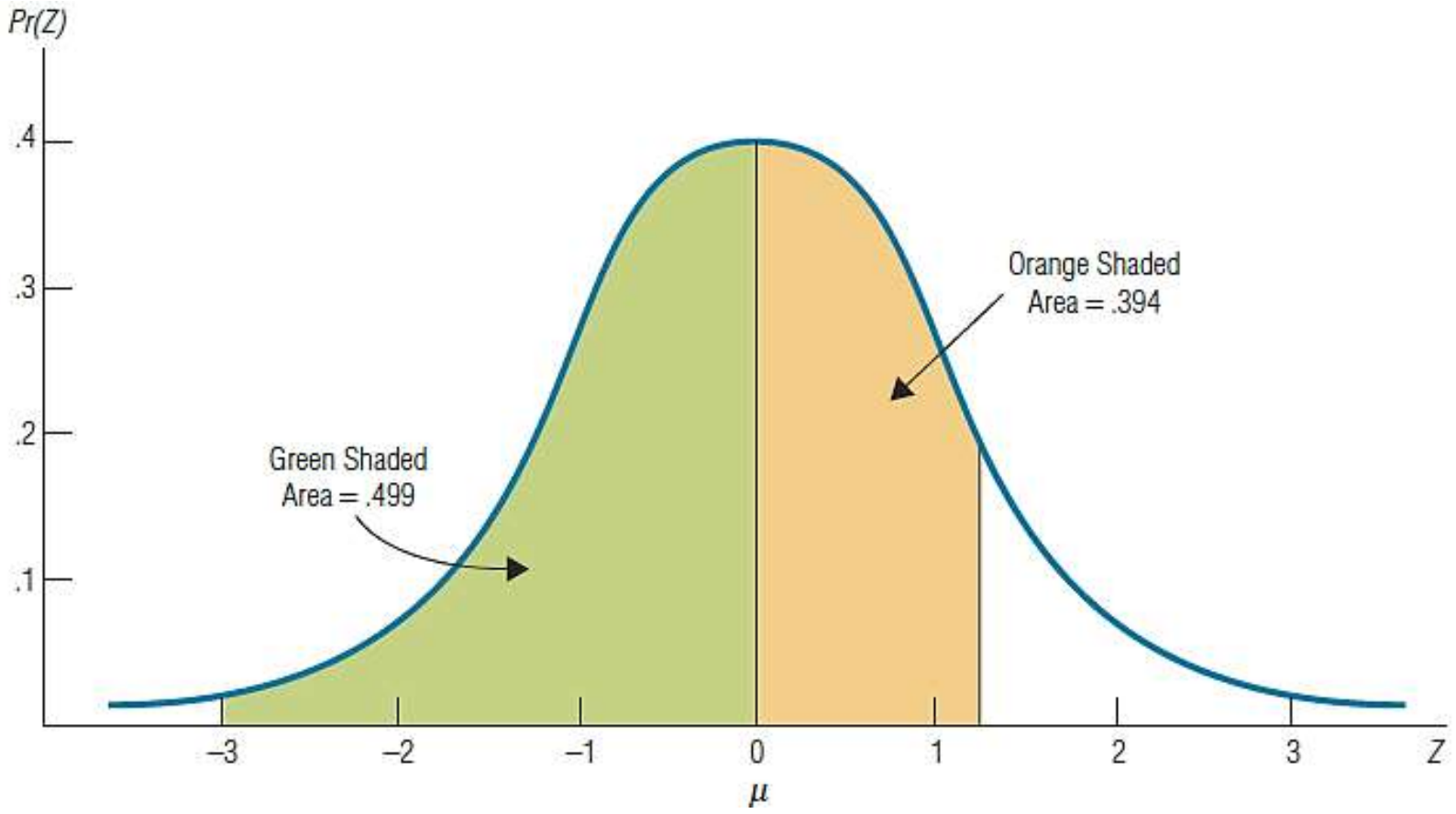
EXHIBIT 13.11 Standardized Values Can Be Computed from Flat or Peaked Distribution Resulting in a Standardized Normal Curves



Population Distribution and Sample Distribution

- Population distribution – a frequency distribution of the population elements
  - Mean and standard deviation represented by the Greek letters μ and σ
- Sample distribution – a frequency distribution of a sample is called the
  - The sample mean is designated with $\bar{X}$ and the sample standard deviation is S

EXHIBIT 13.12 Standardized Distribution Curve



Sampling Distribution

- Illustrates the functional relation between the possible values of some characteristic of *n* cases drawn at random and the probability associated with each value over all possible samples of size *n*

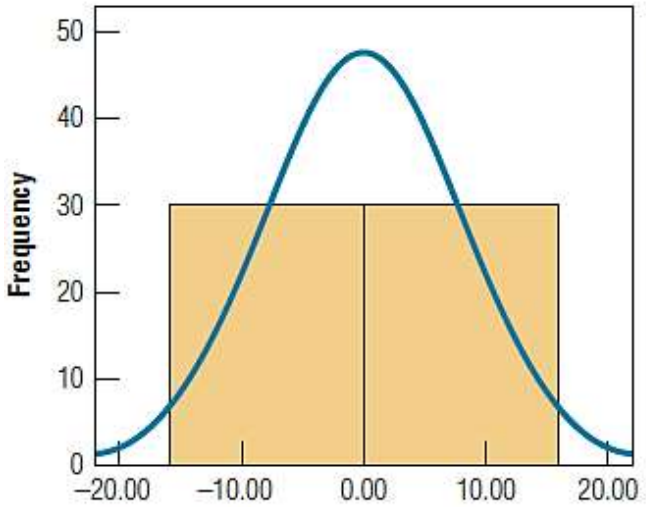
Sampling Distribution (cont'd.)

- The sampling distribution's mean is called the expected value of the statistic
  - The expected value of the mean of the sampling distribution is equal to μ
  - The standard deviation of the sampling distribution is called the standard error of the mean ($S_{\bar{X}}$) and is approximately equal to $\sigma/\sqrt{n}$
  - As sample size increases, the spread of the sample mean around μ decreases

Central Limit Theorem

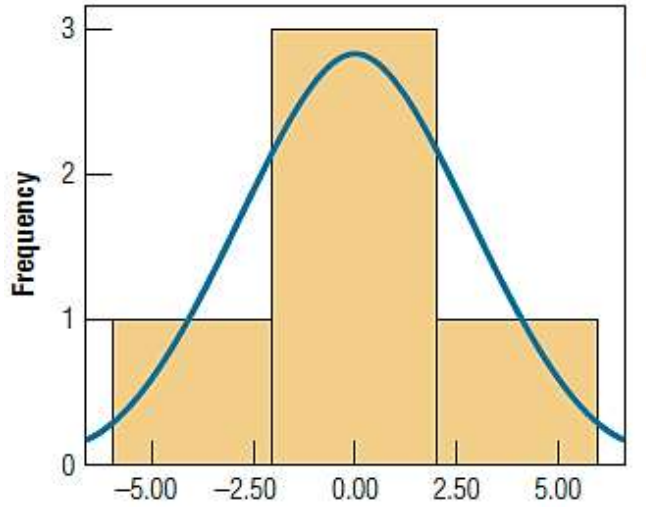
- As the sample size, *n*, increases, the distribution of the mean, $\bar{X}$, of a random sample approaches a normal distribution, with a mean μ and a standard deviation, $\sigma/\sqrt{n}$
- The central-limit theorem works regardless of the shape of the original population distribution
- Theoretical knowledge about distributions can help solve practical research problems
  - Estimating parameters
  - Determining sample size

EXHIBIT 13.13 The Mean Distribution of Any Distribution Approaches Normal as *n* Increases



Mean = 0.00
Std. Dev. = 8.068
n = 60

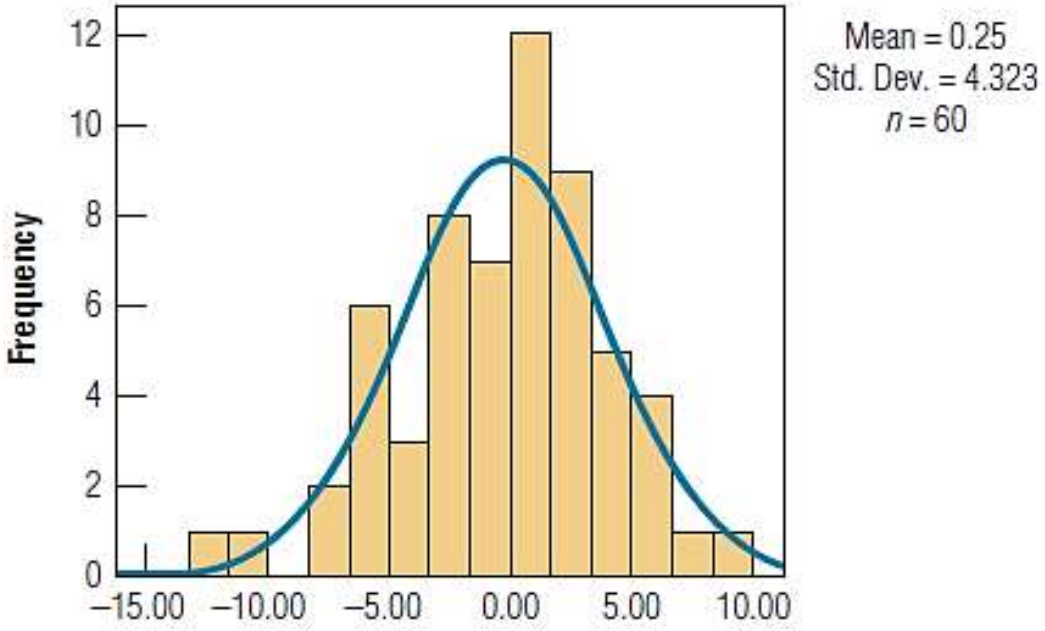
In this frame, an actual distribution of bimodal observations is shown (30 responses of -10 and 30 of +10). This is clearly not a normal distribution. The deviation from normality is shown by the difference between the normal curve and the bars.



Mean = 0.00
Std. Dev. = 2.828
n = 5

Here, a small sample of 5 means from the distribution above is plotted. Three of the observations are 0 while one is -4 and another is 4.

EXHIBIT 13.13 The Mean Distribution of Any Distribution Approaches Normal as *n* Increases (cont'd.)



As the number of samples of means increases, the total area of the bars will increasingly correspond to the normal curve.

EXHIBIT 13.14 Population Distribution: Hypothetical Toy Expenditures

Child	Toy Expenditures
Alice	$1.00
Becky	2.00
Noah	3.00
Tobin	4.00
George	5.00
Freddy	6.00

Source: © Cengage Learning 2013.

EXHIBIT 13.15 Calculation of Population Mean

X
\$1.00
2.00
3.00
4.00
5.00
6.00
$\sum$ \$21.00

Calculations: $\mu = \frac{\sum X}{n} = \frac{21}{6} = 3.5 = \mu_{\bar{x}}$

Estimation of Parameters and Confidence Intervals

- Point estimates
  - Our goal in utilizing statistics is to make an estimate about population parameters
  - The population mean μ, and the standard deviation σ, are constants, but usually unknown
  - Point estimate: an estimate of the population mean in the form of a single value, usually the sample mean
  - One would be extremely lucky if the sample estimate were exactly the same as the population value

Confidence Intervals

- A confidence interval estimate is based on the knowledge that $\mu = \bar{X} \pm$ a small sampling error
- After calculating an interval estimate, we can determine how probable it is that the population mean will fall within this range of statistical values
- The confidence level is a percentage or decimal that indicates the long-run probability that the results will be correct
  - Traditionally, researchers have utilized the 95 percent confidence level

Step By Step Calculation of the Confidence Interval

1. Calculate $\bar{X}$ from the sample
2. Assuming μ is unknown, estimate the population standard deviation by finding S, the sample deviation
3. Estimate the standard error of the mean, using the following formula $S_{\bar{X}} = \frac{S}{\sqrt{n}}$

Step By Step Calculation of the Confidence Interval (cont'd.)

4. Determine the *Z*-value associated with the confidence level desired, and then, divide the confidence level by 2
5. Calculate the confidence interval

Sample Size

- Random error and sample size
  - Random sampling error varies with samples of different sizes
  - Increasing the sample size decreases the width of the confidence interval at a given confidence level
- When the standard deviation of the population is unknown, a confidence interval is calculated by using the following formula:
  - Confidence interval = $\overline{X} \pm Z\frac{S}{\sqrt{n}}$

Factors in Determining Sample Size for Questions Involving Means

- Three factors required to specify sample size
  - The variance, or heterogeneity, of the population
  - The magnitude of acceptable error
  - The confidence level
- The variance, or heterogeneity, of the population in statistical terms refers to the standard deviation of the population parameter
  - Only a small sample is required if the population is homogeneous
  - As heterogeneity increases, so must sample size

Factors in Determining Sample Size for Questions Involving Means (cont'd.)

- The magnitude of error, or the confidence interval, is defined in statistical terms as *E*
  - Indicates a certain precision level
  - From a managerial perspective, the importance of the decision in terms of profitability will influence the researcher's specifications of the range of error
  - The third factor of concern is the confidence level (typically 95 percent)

EXHIBIT 13.16 Statistical Information Needed to Determine Sample Size for Questions Involving Means

Variable	Symbol	Typical Source of Information
Standard deviation	S	Pilot study or rule of thumb
Magnitude of error	E	Managerial judgment or calculation ($ZS_{\bar{x}}$)
Confidence level	$Z_{c.l.}$	Managerial judgment

Source: © Cengage Learning 2013.

Estimating Sample Size for Questions Involving Means

- The researcher must follow three steps:
  1. Estimate the standard deviation of the population
  2. Make a judgment about the desired magnitude of error
  3. Determine a confidence level

Estimating Sample Size for Questions Involving Means (cont'd.)

- The only problem is estimating the standard deviation of the population
  - Ideally, similar studies conducted in the past will be used as a basis for judging the standard deviation
  - In practice, researchers without prior information conduct a pilot study to estimate the population parameters so that another, larger sample, of the appropriate sample size, may be drawn

Estimating Sample Size for Questions Involving Means (cont'd.)

- Using sequential sampling, researchers take an initial look at the pilot study results before deciding on a larger sample to provide more precise information
- A rule of thumb for estimating the value of the standard deviation is to expect it to be one-sixth of the range
- In a general sense, doubling sample size will reduce error by only approximately one-quarter

Population Size and Required Sample Size

- In most cases the size of the population does not have a major effect on the sample size
  - The variance of the population has the largest effect on sample size
  - A finite correction factor may be needed to adjust the sample size if that size is more than 5 percent of a finite population
  - If the sample is large relative to the population, the above procedures may overestimate sample size, and there may be a need to adjust sample size

Determining Sample Size for Proportions

- When a question involves the estimation of a proportion, the researcher requires some knowledge of the logic for determining a confidence interval around a sample proportion (p) of the population proportion (π)
- For a confidence interval to be constructed around the sample proportion (p), estimate the standard error of the proportion (S_p) and specify a confidence coefficient

Determining Sample Size for Proportions (cont'd.)

- Confidence interval: $p \pm Z_{c.l.} S_p$
  - $S_p = \frac{\sqrt{pq}}{n}$

    where

    p = proportion of successes

    q = 1 – p, or proportion of failures
  - To determine sample size for a proportion, make a judgment about the confidence level and the maximum allowance for random sampling error
  - The size of the proportion influences sampling error:

$$n = \frac{Z_{c.l.}^2 \, pq}{E^2}$$

Determining Sample Size on the Basis of Judgment

- Sample size may also be determined on the basis of managerial judgments
  - Using a sample size similar to those used in previous studies provides the inexperienced researcher with a comparison of other researchers' judgments
  - Another judgmental factor is the selection of the appropriate item, question, or characteristics to be used for the sample size calculations
  - Often the item that will produce the largest sample size will be used to determine the ultimate sample size

Determining Sample Size on the Basis of Judgment (cont'd.)

- However, the cost of data collection becomes a major consideration
  - Judgment must be exercised regarding the importance of such information
- Another consideration stems from most researchers' need to analyze the various subgroups within the sample
  - Rule of thumb for selecting minimum subgroup sample size: each subgroup to be separately analyzed should have a minimum of 100 or more units in each category of the major breakdowns

Assess the Potential For Nonresponse Bias

- Nonresponse bias, in particular the bias caused when sample units provide no response, can significantly damage generalizability
  - The reason that nonresponders must be considered routinely a threat to external validity is that there could be some systematic reason that members selected for inclusion from a sampling frame did not respond
  - Any systematic connection between sample unit characteristics and their likelihood to respond is a potential source for bias

Assess the Potential For Nonresponse Bias (cont'd.)

- Factors to consider to help increase confidence in the generalizability of the sample
  - A well-managed sampling frame that contains accurate information provides the greatest potential for generalizability
  - Auxiliary variables provide a useful means of detecting potential systematic reasons for nonresponse (or conversely response)
  - A high response rate in and of itself does not guarantee freedom from bias
  - Post-hoc sampling procedures can be employed