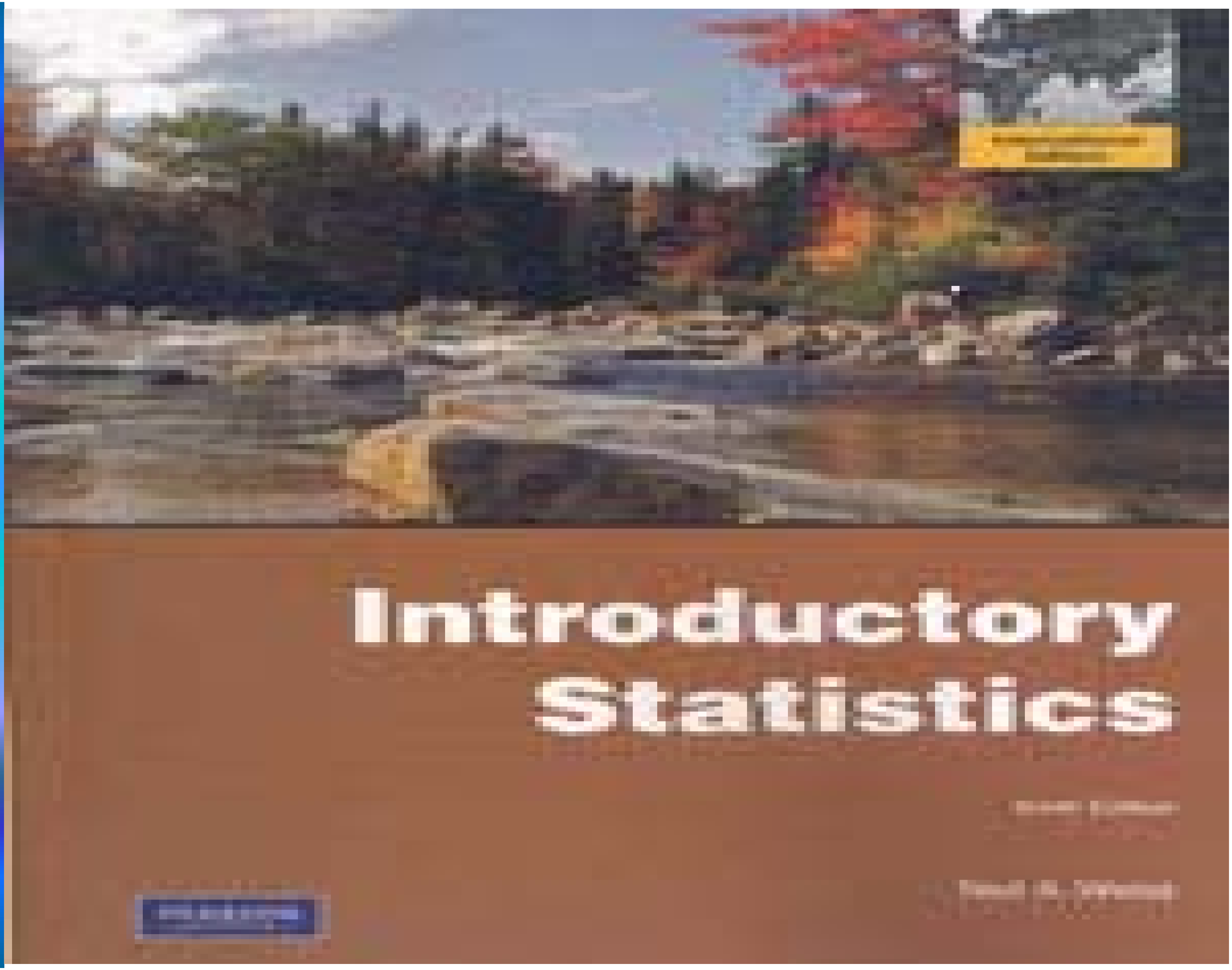

Introductory
Statistics

Chapter 7

The Sampling Distribution of the Sample Mean

Chapter Goals

- Investigate the variability in sample statistics from sample to sample
- Find measures of central tendency for sample statistics
- Find measures of dispersion for sample statistics.
- Find the pattern of variability for sample statistics

The Engineering Method and Statistical Thinking

- Statistical techniques are useful for describing and understanding variability.
- By variability, we mean successive observations of a system or phenomenon do *not* produce exactly the same result.
- Statistics gives us a framework for describing this variability and for learning about potential sources of variability.

Reasons for Sampling

- Sampling can save money.
- Sampling can save time.
- For given resources, sampling can broaden the scope of the data set.
- Because the research process is sometimes destructive, the sample can save product.
- If accessing the population is impossible; sampling is the only option.

Reasons for Taking a Census

- Eliminate the possibility that a random sample is not representative of the population.
- The person authorizing the study is uncomfortable with sample information.

Section 7.1

Sampling Error; the Need for Sampling Distributions

7.1 Sampling Error and the Need for Sampling Distributions

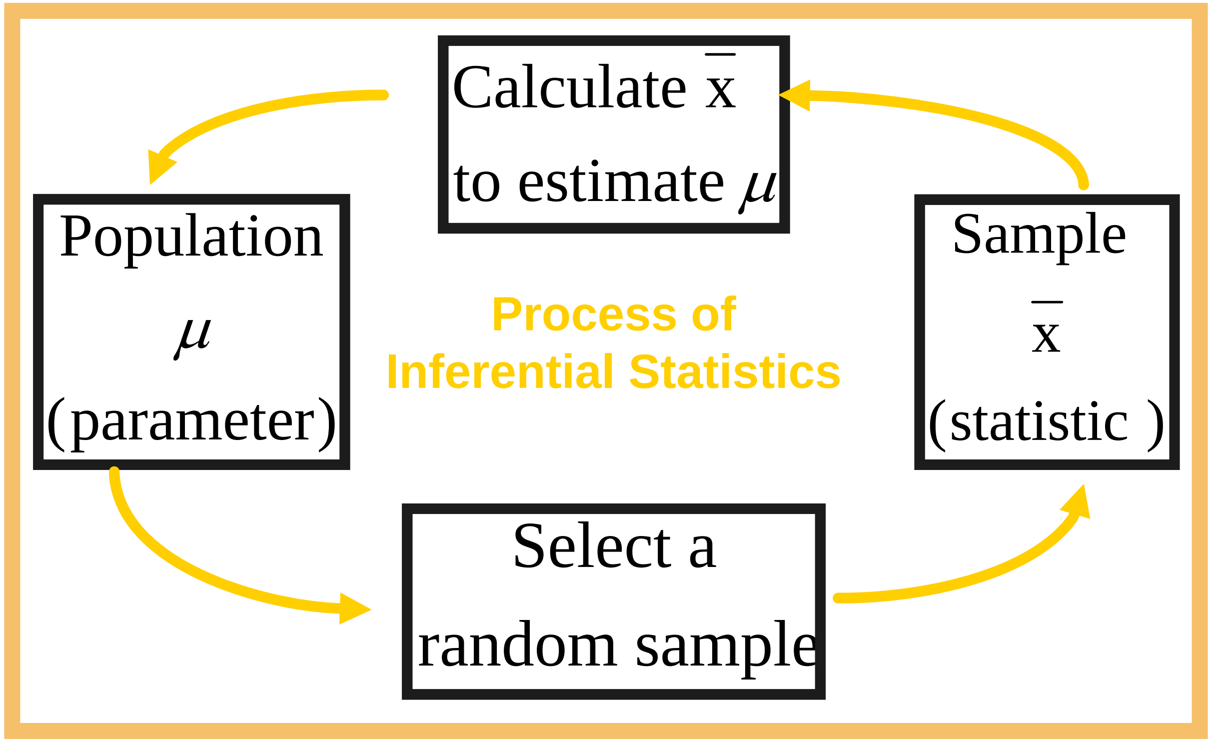
- To make inferences about a population, we need to understand sampling

- The sample mean varies from sample to sample

- The sample mean has a distribution; we need to understand how the sample mean varies and the pattern (if any) in the distribution

Sampling Distribution of $\bar{x}$

Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.



Definition 7.1

Sampling Error

Sampling error is the error resulting from using a sample to estimate a population characteristic.

Definition 7.2

Sampling Distribution of the Sample Mean

For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the **sampling distribution of the sample mean.**

Table 7.2

Possible samples and sample means for samples of size 2

Player	Height
A	76
B	78
C	79
D	81
E	86

Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

Important Notes & Random Sample

1. $\overline{\overline{x}}$: the mean of the sample means
2. $s_{\overline{x}}$: the standard deviation of the sample means
3. The theory involved with sampling distributions described in the remainder of this chapter requires random sampling

Random Sample: A sample obtained in such a way that each possible sample of a fixed size n has an equal probability of being selected

- (*Every possible handful* of size n has the same probability of being selected)

Figure 7.1

Dotplot for the sampling distribution of the sample mean for samples of size 2 (n = 2)

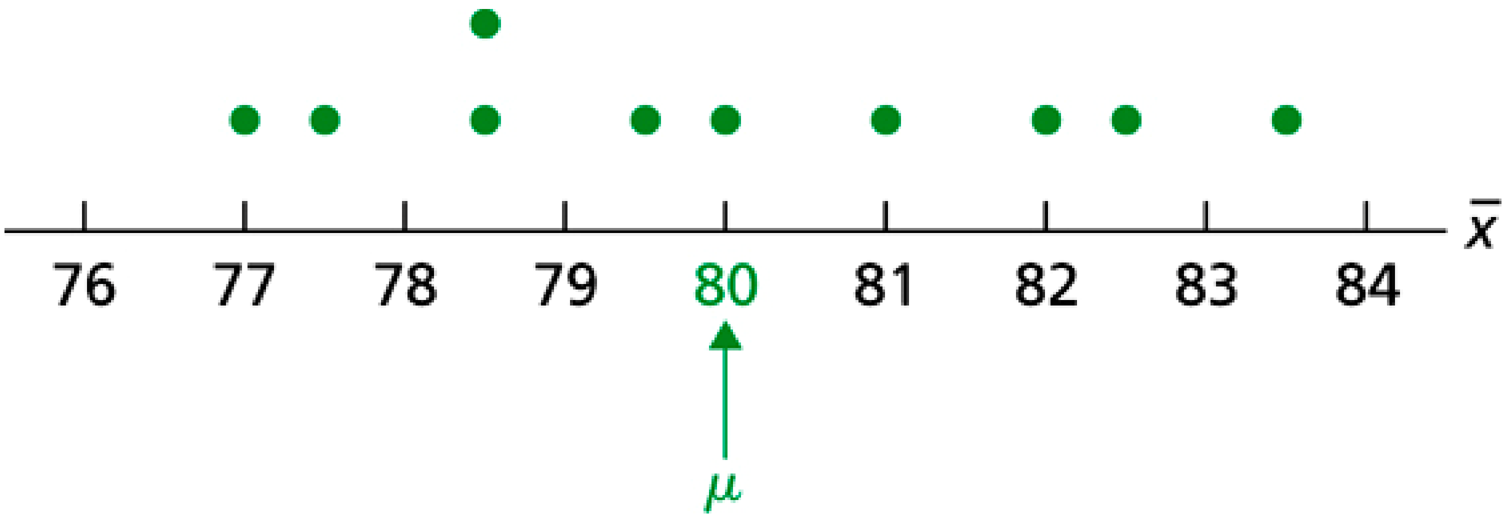


Figure 7.3

Dotplots for the sampling distributions of the sample mean for the heights of the five starting players for samples of sizes 1, 2, 3, 4, and 5

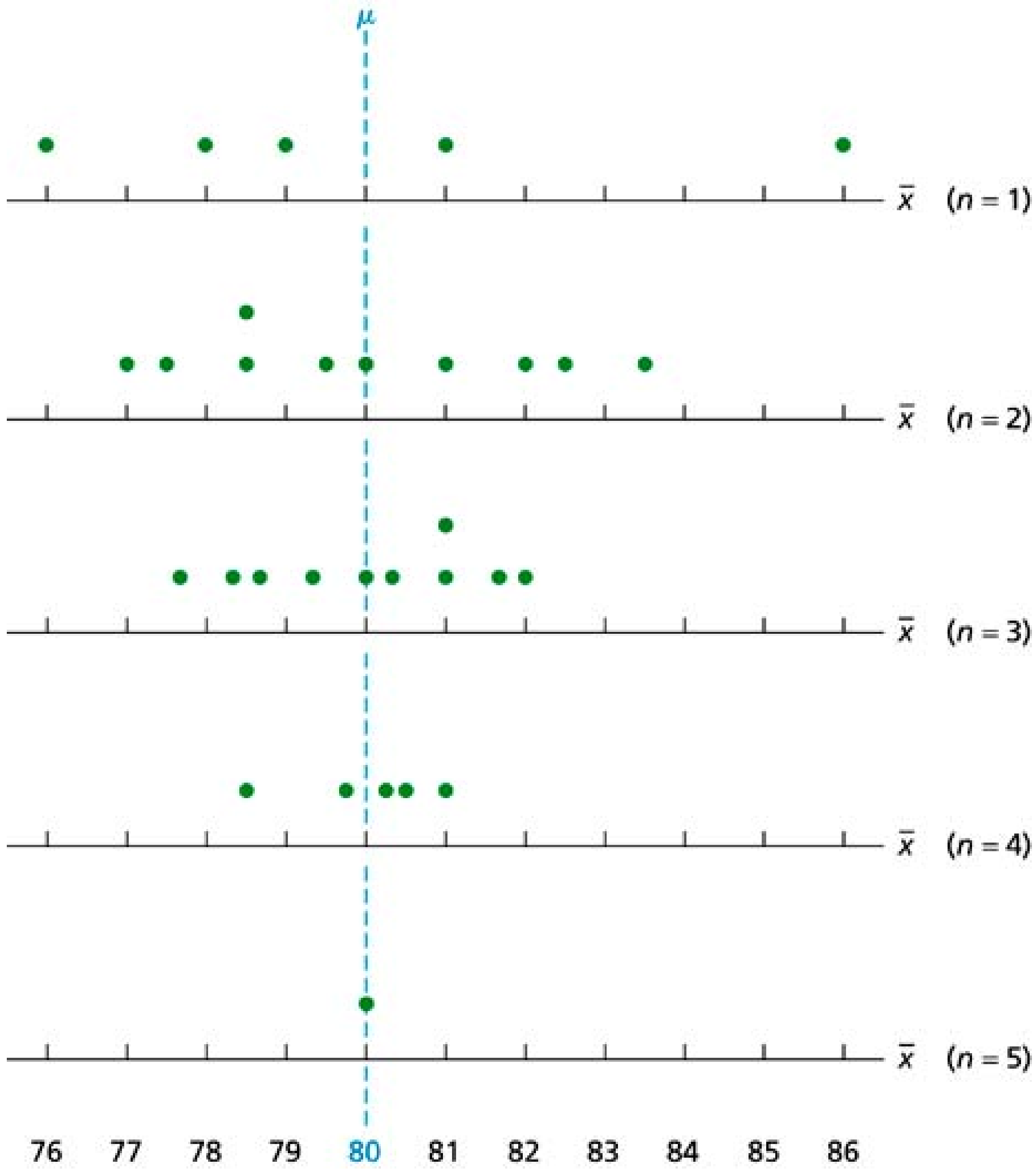


Table 7.4

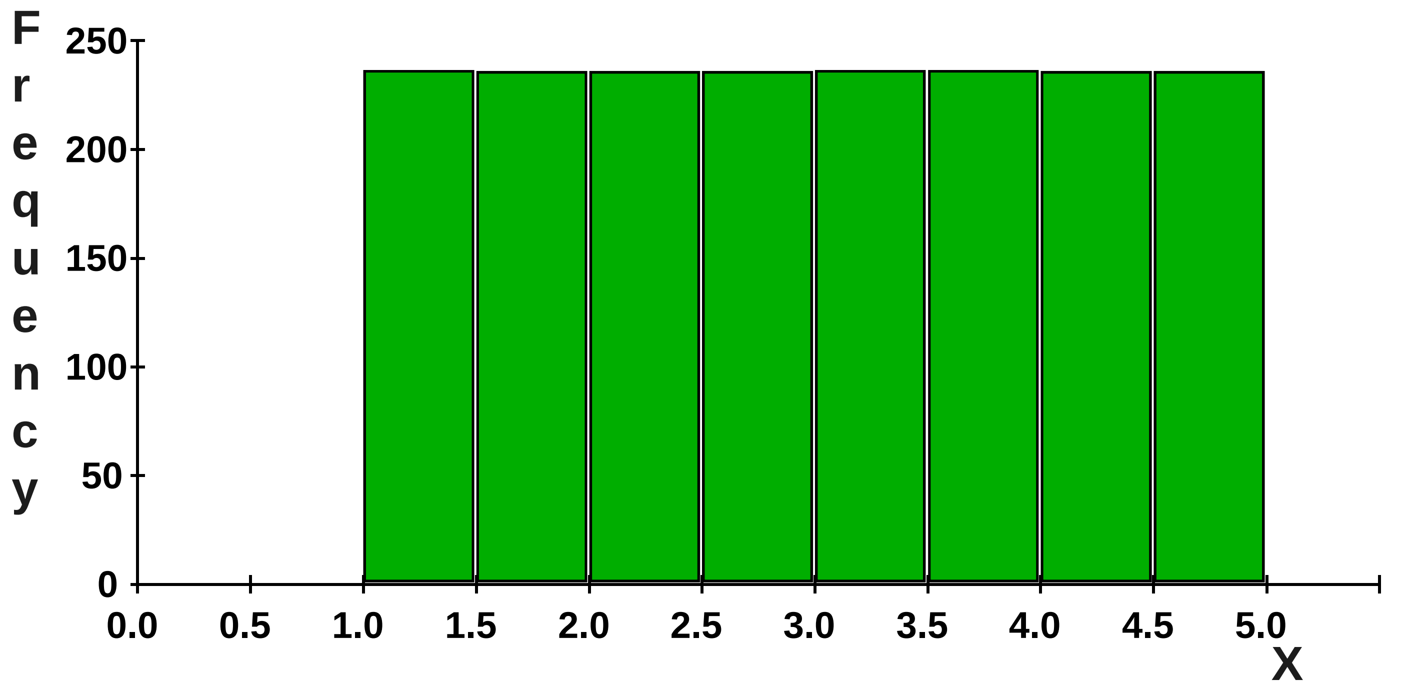
Sample size and sampling error illustrations for the heights of the basketball players (“No.” is an abbreviation of “Number”)

Sample size n	No. possible samples	No. within $1''$ of μ	% within $1''$ of μ	No. within $0.5''$ of μ	% within $0.5''$ of μ
1	5	2	40%	0	0%
2	10	3	30%	2	20%
3	10	5	50%	2	20%
4	5	4	80%	3	60%
5	1	1	100%	1	100%

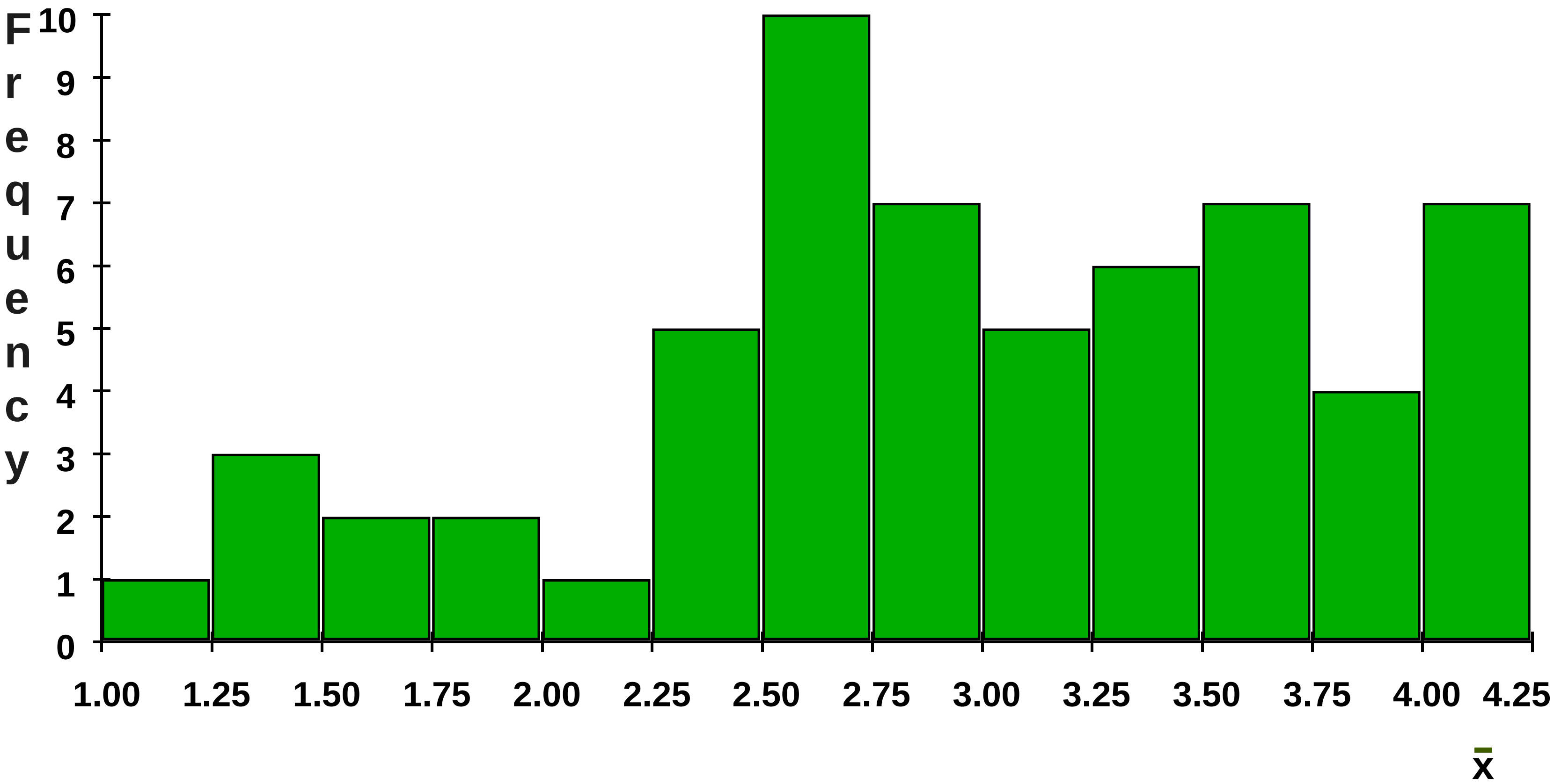
Section 7.2

The Mean and Standard Deviation of the Sample Mean

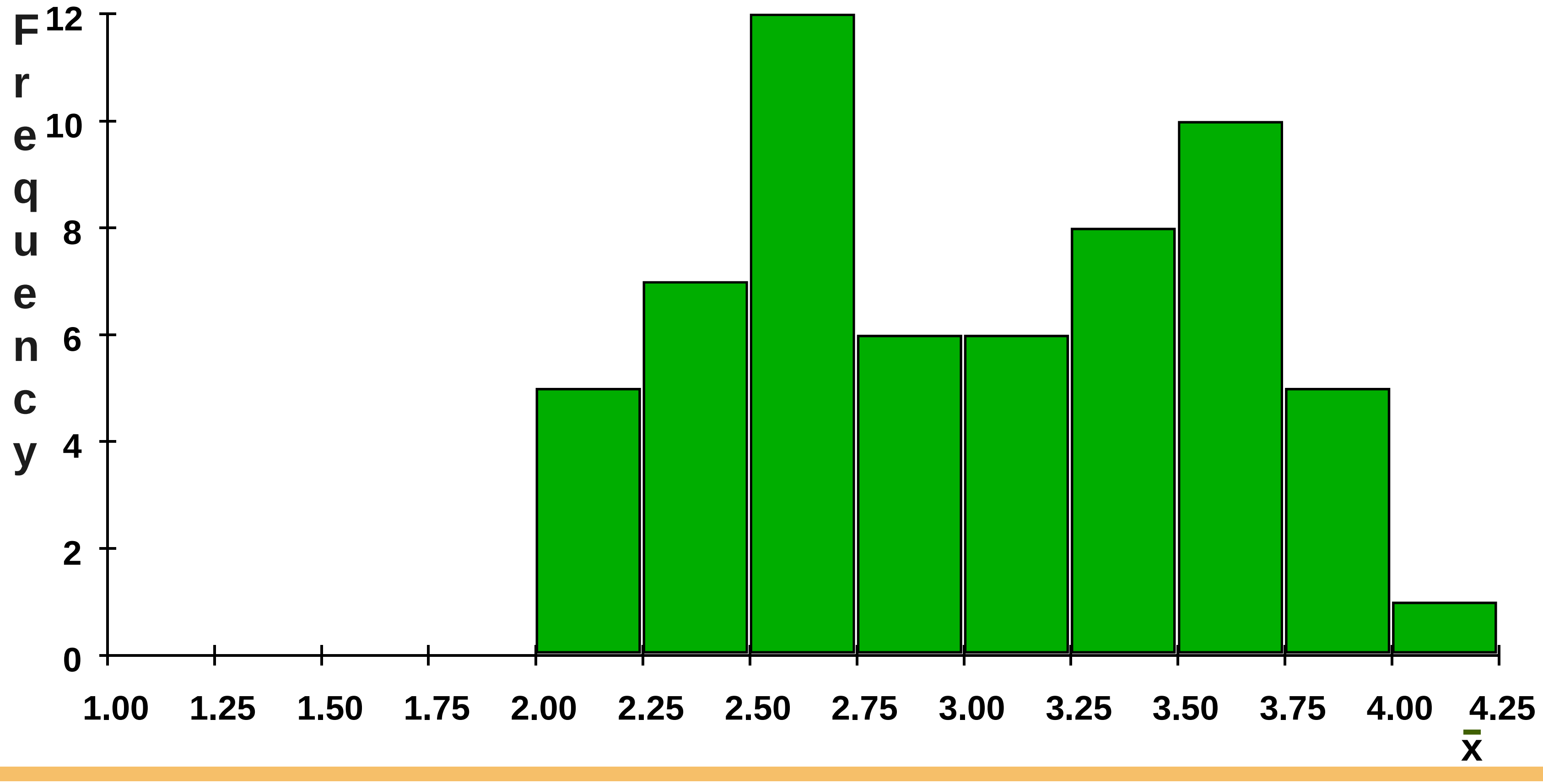
1,800 Randomly Selected Values from a Uniform Distribution



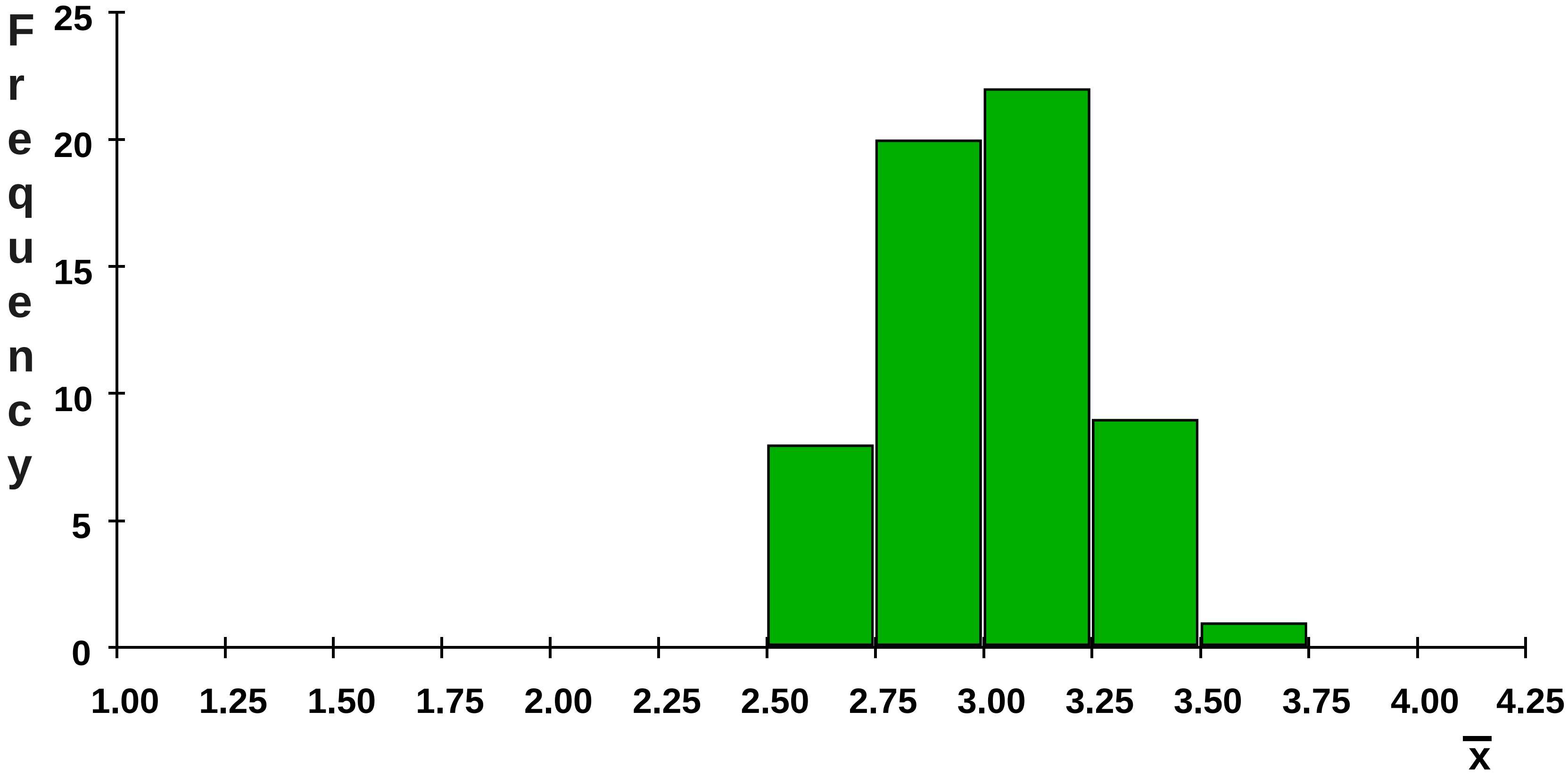
Means of 60 Samples (n = 2) from a Uniform Distribution



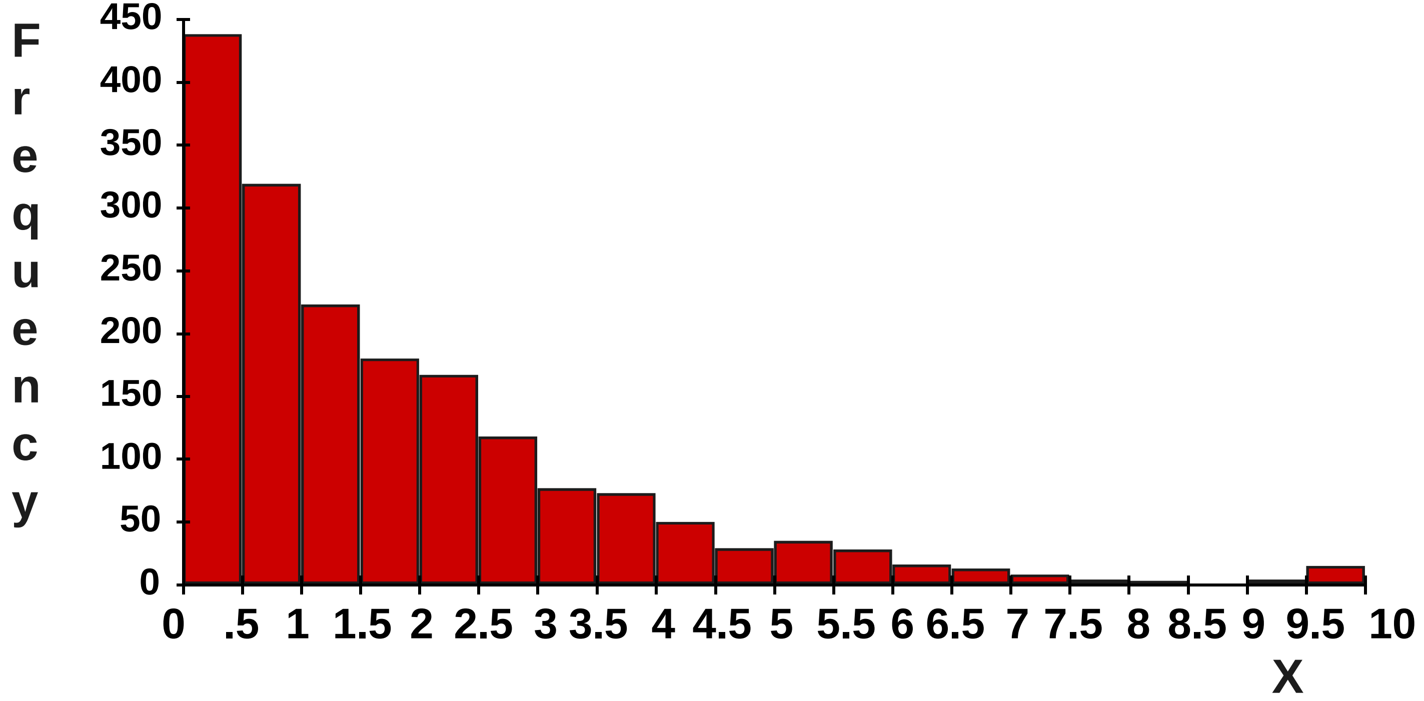
Means of 60 Samples (n = 5) from a Uniform Distribution



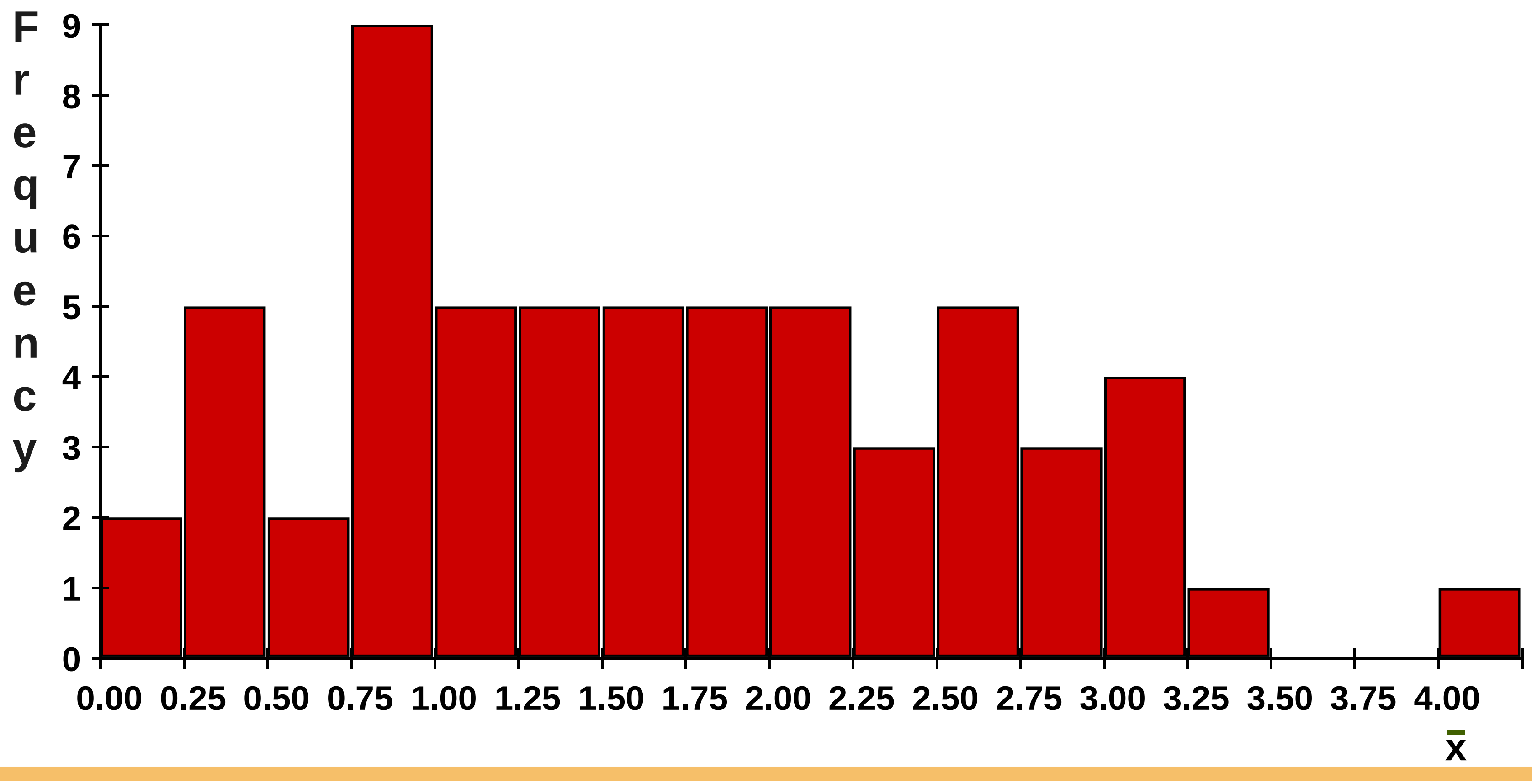
Means of 60 Samples (*n* = 30) from a Uniform Distribution



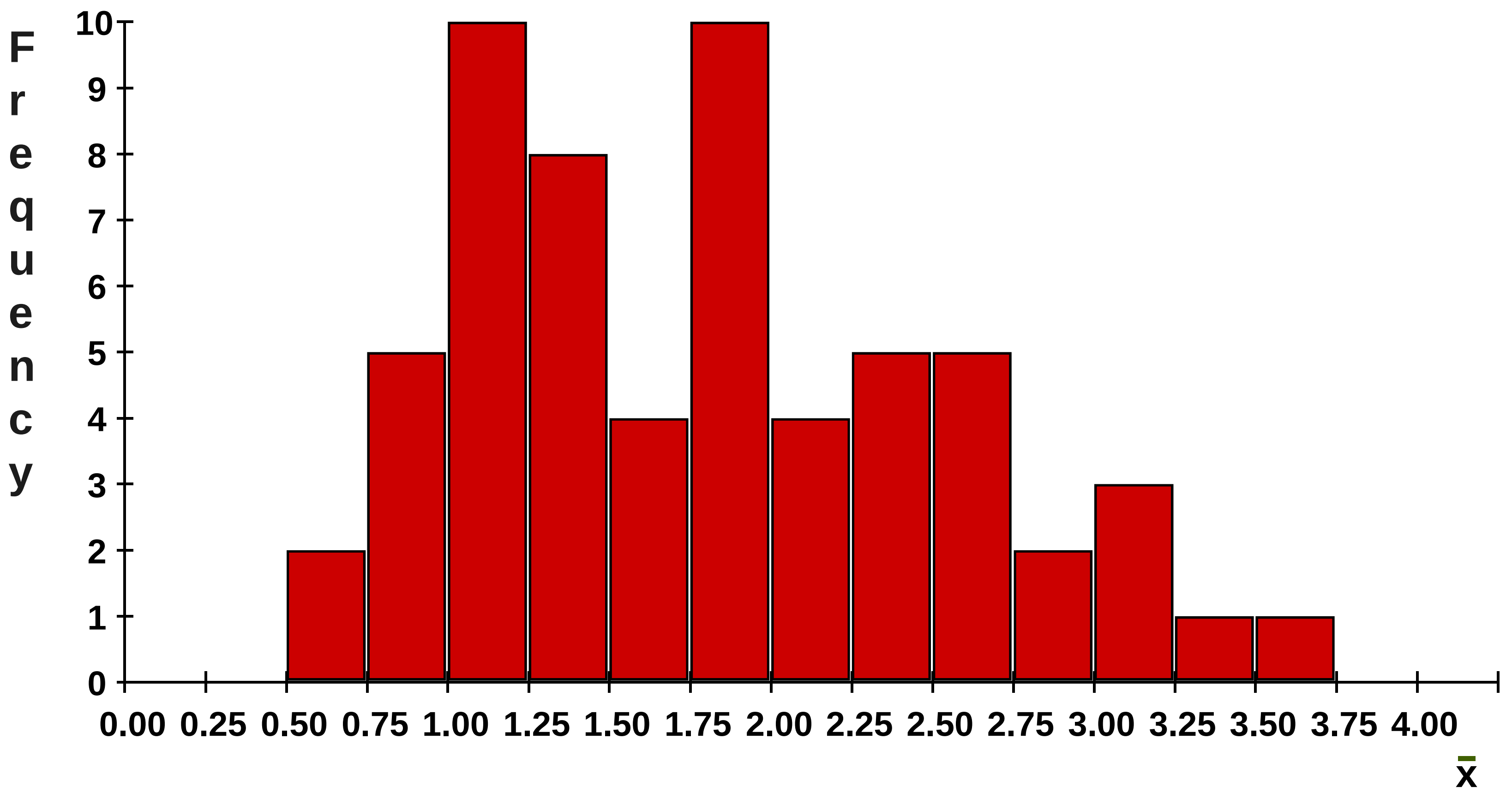
1,800 Randomly Selected Values from an Exponential Distribution



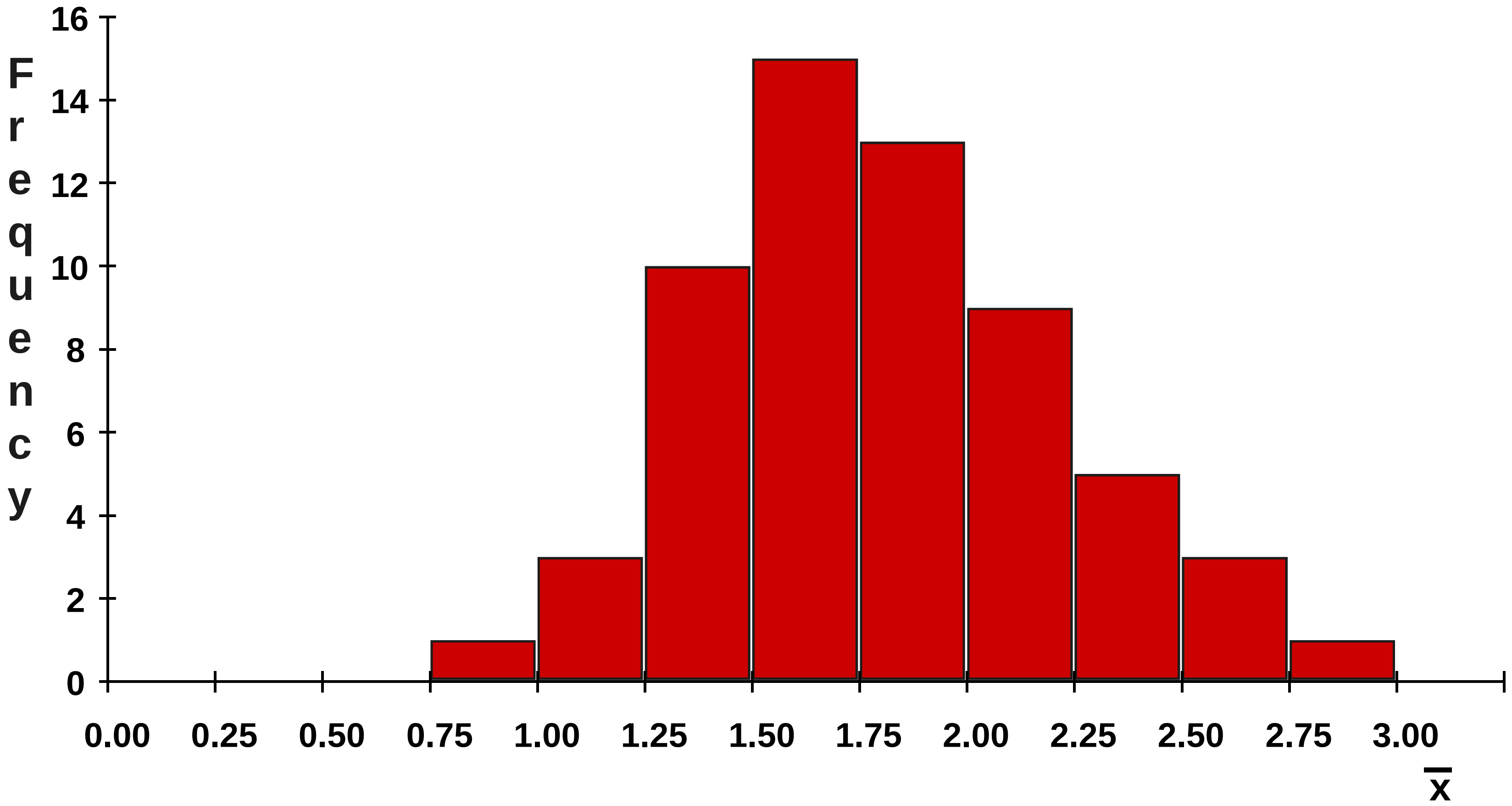
Means of 60 Samples (n = 2) from an Exponential Distribution



Means of 60 Samples (*n* = 5) from an Exponential Distribution



Means of 60 Samples (n = 30) from an Exponential Distribution



Where Does This Lead Us?

- Describing the most important idea in all of statistics
- Describes the sampling distribution of the sample mean
- Examples suggest: the sample mean (and sample total) tend to be normally distributed

Important Definition & Theorem

Sampling Distribution of Sample Means

If all possible random samples, each of size n, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of sample means will:

1. have a mean $\mu_{\bar{x}}$ equal to μ
2. have a standard deviation $\sigma_{\bar{x}}$ equal to $\sigma / \sqrt{n}$

Further, if the sampled population has a normal distribution, then the sampling distribution of $\bar{x}$ will also be normal for samples of all sizes

Central Limit Theorem

The sampling distribution of sample means will become normal as the sample size increases.

Formula 7.1

Mean of the Sample Mean

For samples of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. In symbols,

$$\mu_{\bar{x}} = \mu.$$

Formula 7.2

Standard Deviation of the Sample Mean

For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size. In symbols,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Central Limit Theorem

- **For sufficiently large sample sizes ($n \geq 30$),**
- **the distribution of sample means $\overline{X}$, is approximately normal;**
- **the mean of this distribution is equal to μ, the population mean; and**
- **its standard deviation is $\dfrac{\sigma}{\sqrt{n}}$,**

regardless of the shape of the population distribution.

Section 7.3

The Sampling Distribution of the Sample Mean

Output 7.1

Histogram of the sample means for 1000 samples of four IQs with superimposed normal curve

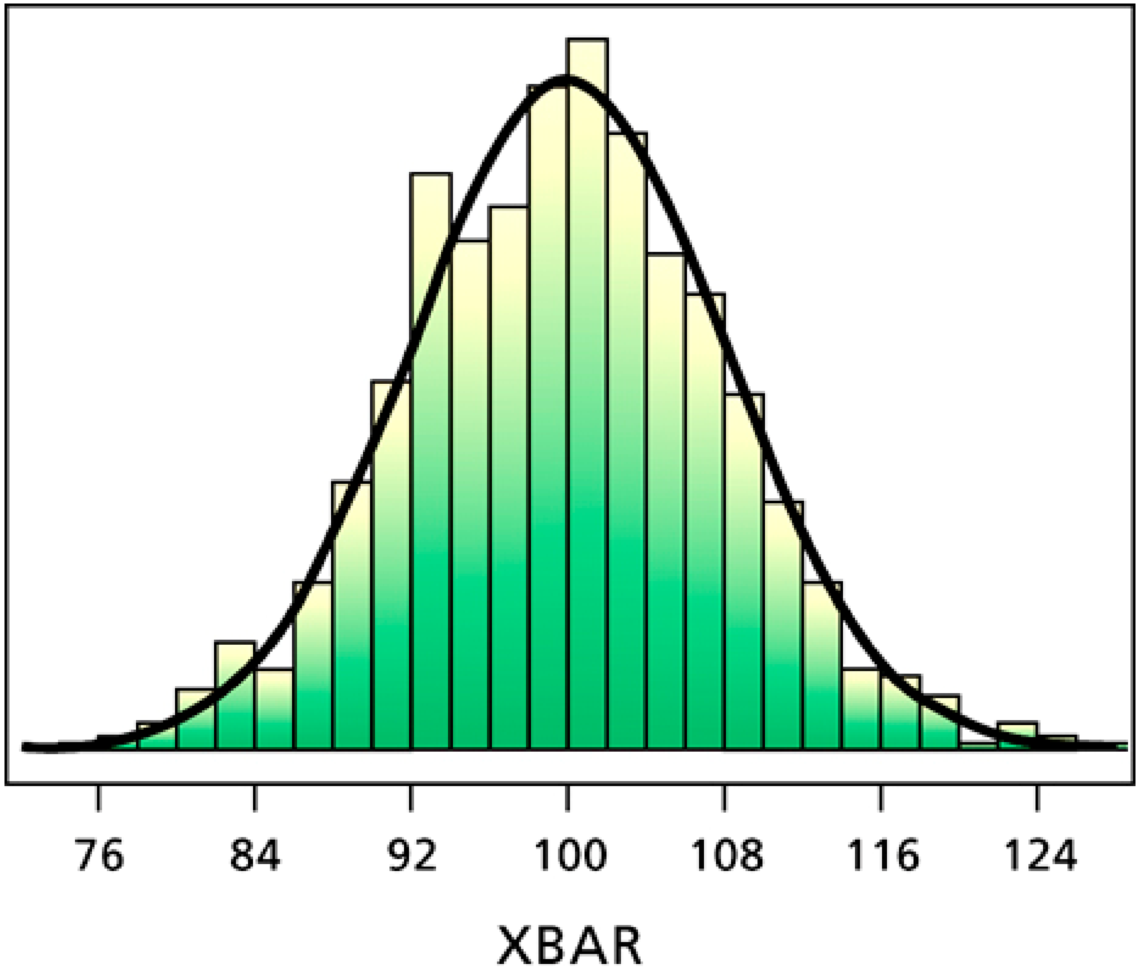


Figure 7.4

(a) Normal distribution for IQs; (b) sampling distribution of the sample mean for n = 4; (c) sampling distribution of the sample mean for n = 16

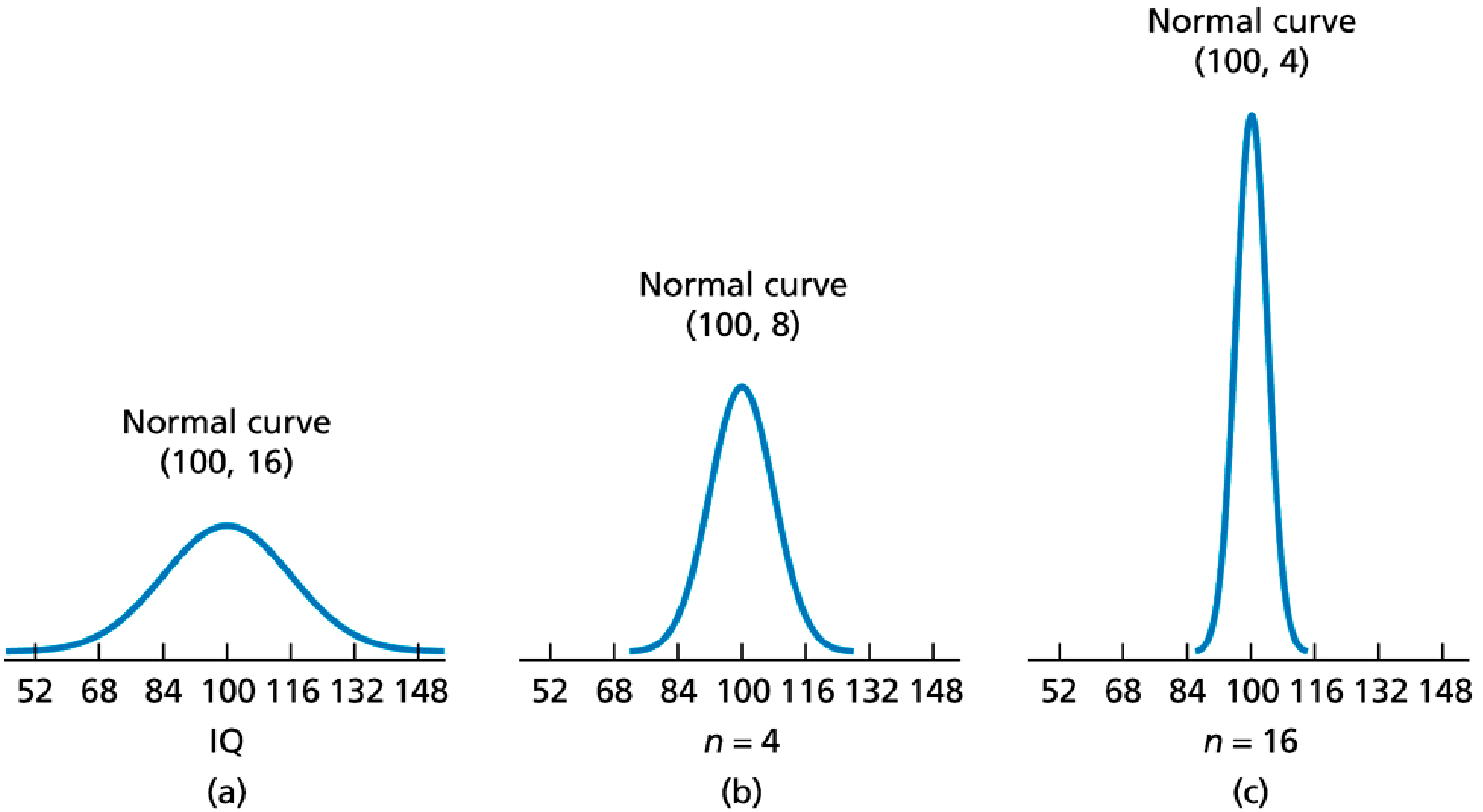
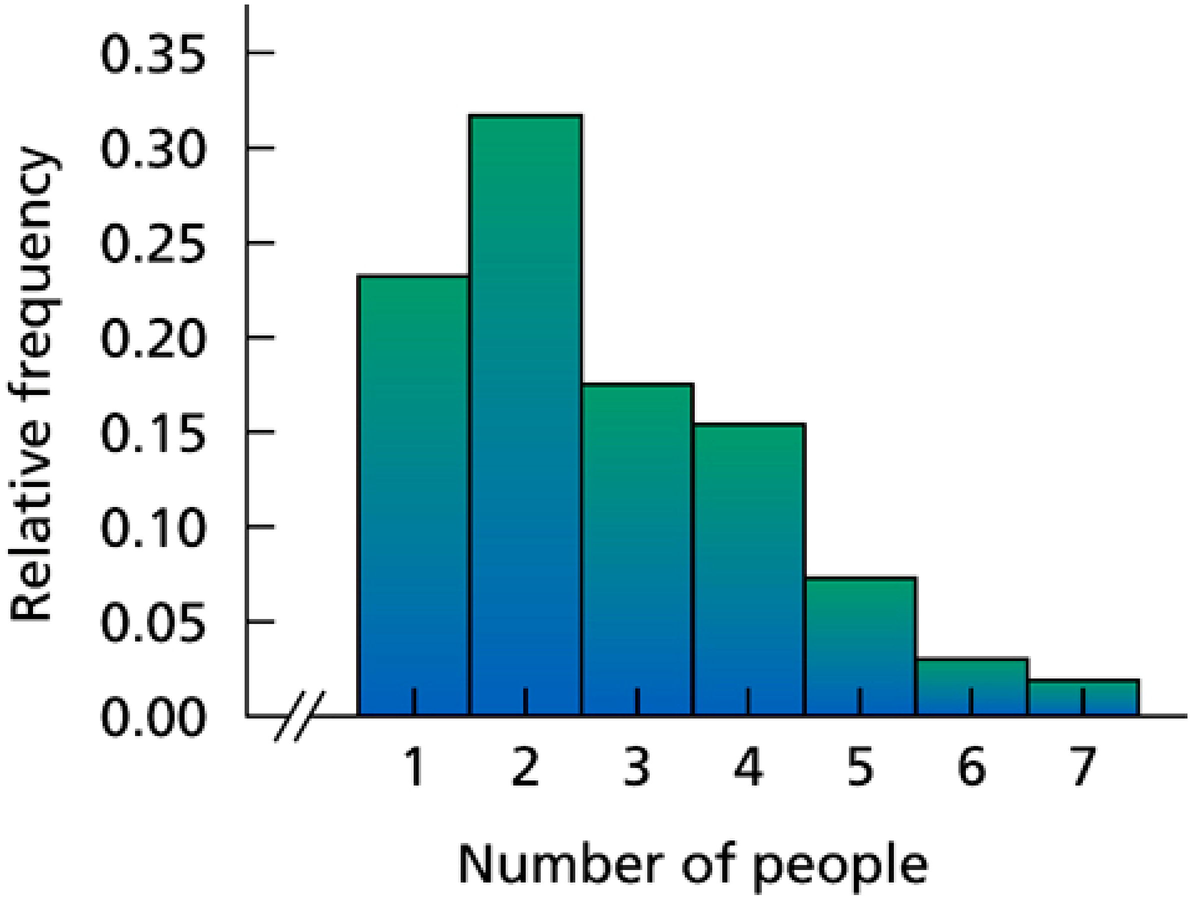


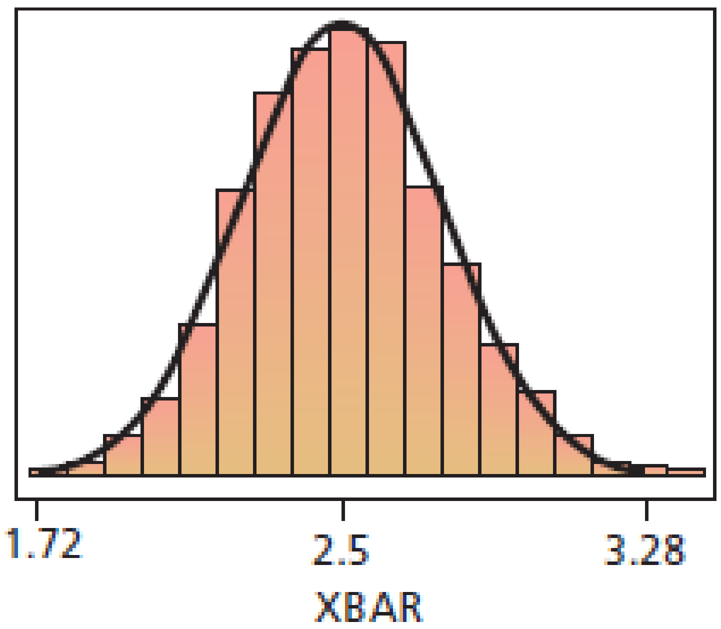
Figure 7.5

Relative-frequency histogram for household size



Output 7.2

Histogram of the sample means for 1000 samples of 30 household sizes with superimposed normal curve



Key Fact 7.4

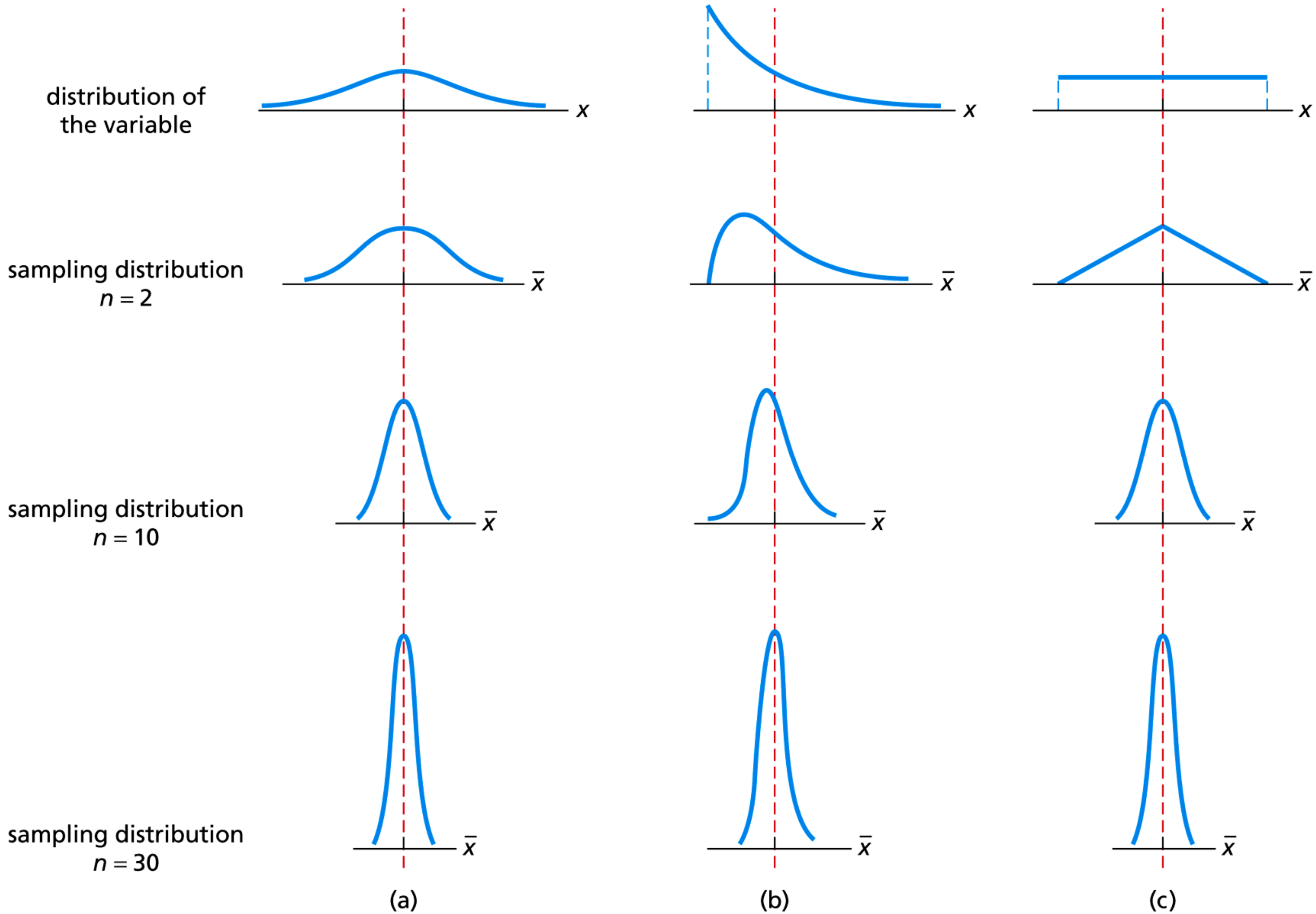
Sampling Distribution of the Sample Mean

Suppose that a variable x of a population has mean μ and standard deviation σ. Then, for samples of size n,

- the mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$;
- the standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \sigma/\sqrt{n}$;
- if x is normally distributed, so is $\bar{x}$, regardless of sample size; and
- if the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

Figure 7.6

Sampling distributions of the sample mean for (a) normal, (b) reverse-J-shaped, and (c) uniform variables



Point Estimates

Examples of point estimates are the *sample mean*, the *sample standard deviation*, the *sample variance*, the *sample proportion*.

A point estimate is one value (a single point) that is used to estimate a population parameter.

Point Estimates

- If a population follows the normal distribution, the sampling distribution of the sample mean will also follow the normal distribution.
- To determine the probability a sample mean falls within a particular region, use:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Standard Error of the Mean

<u>**Standard Error of the Mean**</u>: The standard deviation of the sampling distribution of sample means: $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

Notes:

- The *n* in the formula for the standard error of the mean is the size of the sample
- The proof of the Central Limit Theorem is beyond the scope of this course
- The following example illustrates the results of the Central Limit Theorem

Statistical Estimation

Chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run:

$$\mathrm{E}\left[\mu\right] = \bar{X}.$$

- Point estimate -- the single value of a statistic calculated from a sample,

- Interval Estimate -- a range of values calculated from a sample statistic(s) and standardized statistics, such as the *Z*.
  - Selection of the standardized statistic is determined by the sampling distribution.
  - Selection of critical values of the standardized statistic is determined by the desired level of confidence.

Interval Estimates

- An Interval Estimate states the range within which a population parameter probably lies.
  - The interval within which a population parameter is expected to occur is called a confidence interval.
  - The two confidence intervals that are used extensively are the 95% and the 99%.

Applications of the Central Limit Theorem

- When the sampling distribution of the sample mean is (exactly) normally distributed, or approximately normally distributed (by the CLT), we can answer probability questions using the standard normal distribution, use Table.

Example 1

Example: Consider a normal population with $\mu = 50$ and $\sigma = 15$. Suppose a sample of size 9 is selected at random. Find:

$$(1) \qquad P\left(45 \le \bar{x} \le 60\right)$$

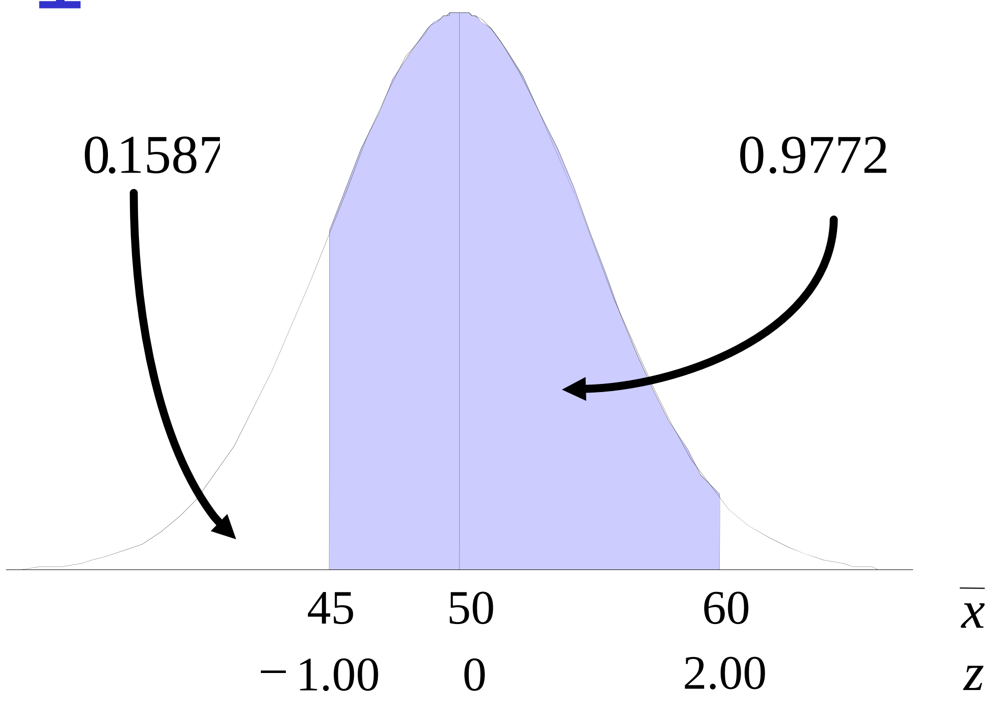
$$(2) \qquad P\left(\bar{x} \le 47.5\right)$$

Solutions: Since the original population is normal, the distribution of the sample mean is also (exactly) normal

1) $\mu_{\bar{x}} = \mu = 50$

2) $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 15 / \sqrt{9} = 15/3 = 5$

Example 1



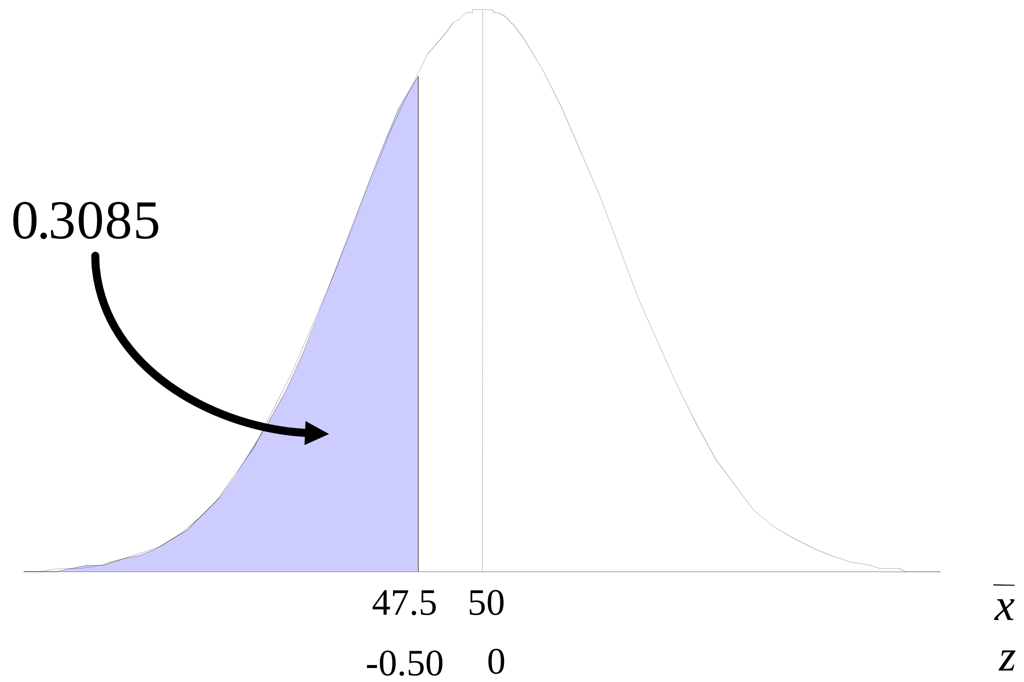
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}};$$

$$P(45 \le \bar{x} \le 60) = P\left(\frac{45-50}{5} \le z \le \frac{60-50}{5}\right)$$

$$= P(-1.00 \le z \le 2.00)$$

$$= 0.9772 - 0.1587 = 0.8185$$

Example 1



$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}};$$

$$P(\bar{x} \leq 47.5) = P\left(\frac{\bar{x} - 50}{5} \leq \frac{47.5 - 50}{5}\right)$$

$$= P(z \leq -.5)$$

$$= \ 0.3085$$

Example 3

Example: A recent report stated that the day-care cost per week in Boston is \$109. Suppose this figure is taken as the mean cost per week and that the standard deviation is known to be \$20.

1) Find the probability that a sample of 50 day-care centers would show a mean cost of \$105 or less per week.
2) Suppose the actual sample mean cost for the sample of 50 day-care centers is \$120. Is there any evidence to refute the claim of \$109 presented in the report?

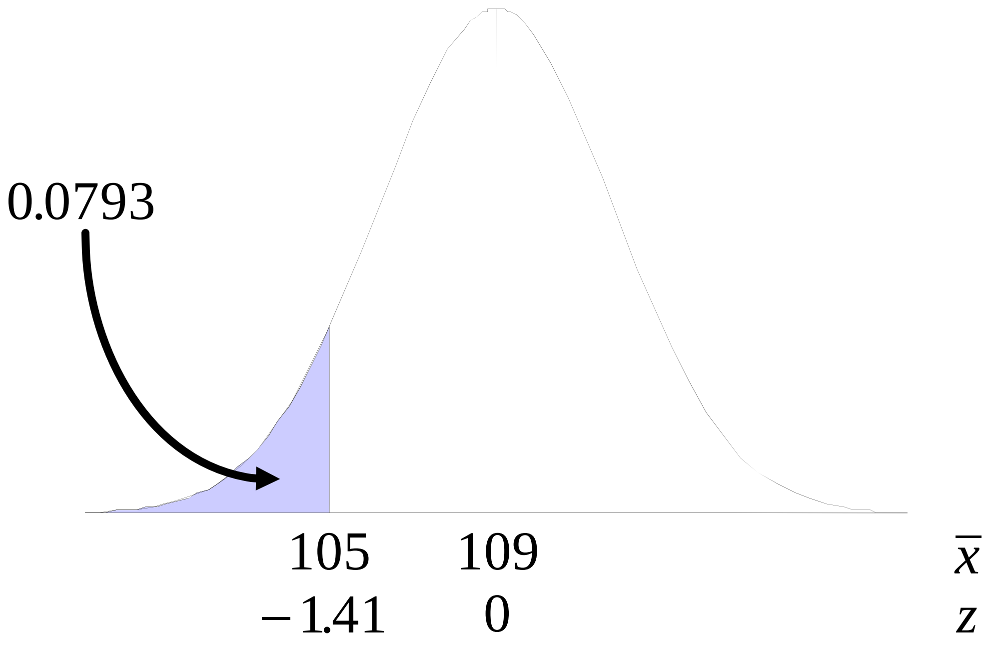
Solutions:

- The shape of the original distribution is unknown, but the sample size, *n*, is large. The CLT applies.
- The distribution of $\bar{x}$ is approximately normal

$$\mu_{\bar{x}} = \mu = 109 \qquad \sigma_{\bar{x}} = \sigma / \sqrt{n} = 20 / \sqrt{50} \approx 2.83$$

Example 3

1)



$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}};$$

$$P(\bar{x} \le 105) = P\left(z \le \frac{105 - 109}{2.83}\right)$$

$$= P(z \le -1.41)$$

$$= 0.0793$$

Example 2

2) To investigate the claim, we need to examine how *likely* an observation is the sample mean of \$120

- Consider how far out in the tail of the distribution of the sample mean is \$120

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}; \quad P(\bar{x} \geq 120) = P\left(z \geq \frac{120 - 109}{2.83}\right)$$

$$= P(z \geq 3.89)$$

$$= 1 - 0.9999 = 0.0001$$

- Since the probability is so small, this suggests the observation of \$120 is very rare (if the mean cost is really \$109)

- There is evidence (the sample) to suggest the claim of μ = \$109 is likely wrong

Sampling from a Finite Population without Replacement

- In this case, the standard deviation of the distribution of sample means is smaller than when sampling from an infinite population (or from a finite population with replacement).
- The correct value of this standard deviation is computed by applying a finite correction factor to the standard deviation for sampling from a infinite population.
- If the sample size is less than 5% of the population size, the adjustment is unnecessary.

Sampling from a Finite Population

- Finite Correction Factor

$$\boxed{\sqrt{\frac{N-n}{N-1}}}$$

- Modified Z Formula

$$\boxed{Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}}$$