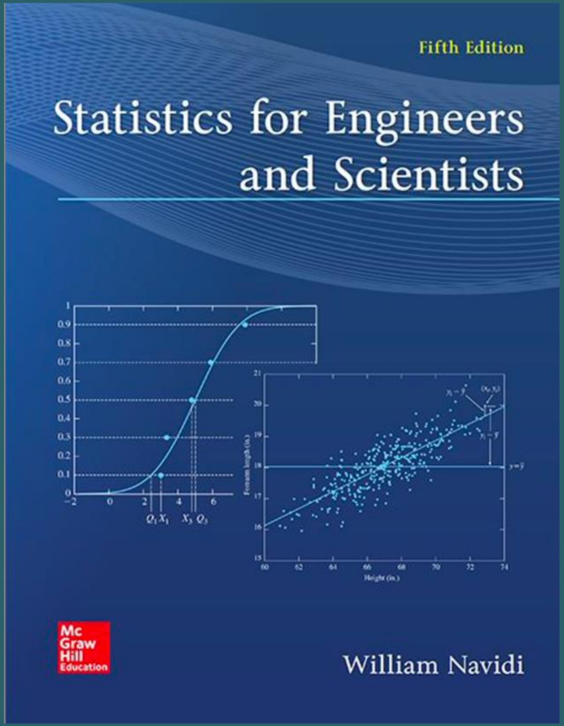


Ch 4: Commonly Used Distributions

(not all sections are required)

Ch 4: Overview (Required Sections)

- 4-1 The Bernoulli Distribution
- 4-2 The Binomial Distribution
- 4-3 The Poisson Distribution
- 4-5 The Normal Distribution
- ~~*4-9 Some Principles of Point Estimation*~~
- 4-10 Probability Plots
- **4-11 Central Limit Theorem**

The Central Limit Theorem

In addition to knowing how individual data values vary about the mean for a population, statisticians are interested in knowing how the means of samples of the same size taken from the same population vary about the population mean.

Distribution of Sample Means

- A **sampling distribution of sample means** is a distribution obtained by using the means computed from random samples of a specific size taken from a population.
- **Sampling error** is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

Properties of the Distribution of Sample Means

- The mean of the sample means will be the <u>same</u> as the population mean.
- The standard deviation of the sample means will be <u>smaller</u> than the standard deviation of the population, and will be equal to the population standard deviation divided by the square root of the sample size.

Example

Suppose a professor gave an 8-point quiz to a small class of four students.

The results of the quiz were 2, 6, 4, and 8.

For the sake of discussion, assume that <u>the four students constitute the population</u>.

The population mean is:

$$\mu = \frac{2+6+4+8}{4} = 5$$

Example...

The population mean is:

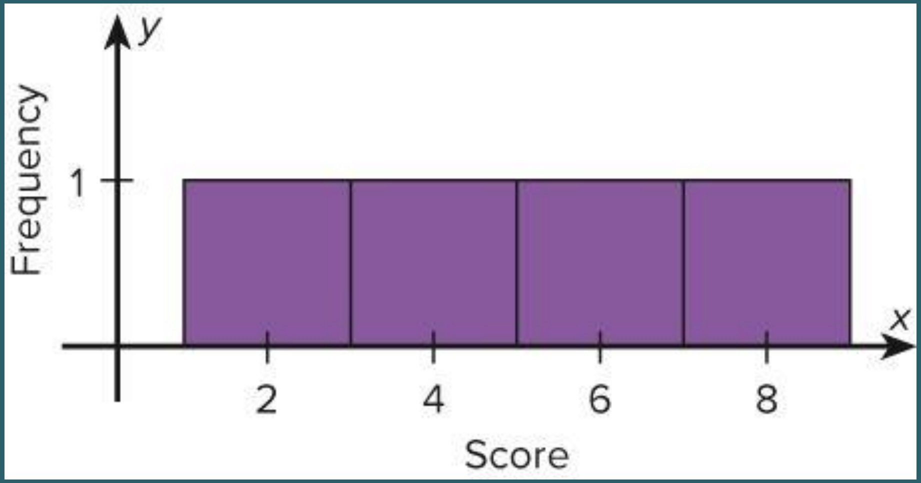
$$\mu = \frac{2 + 6 + 4 + 8}{4} = 5$$

The standard deviation of the population is:

$$\sigma = \frac{\sqrt{(2 - 5)^2 + (6 - 5)^2 + (4 - 5)^2 + (8 - 5)^2}}{4} \approx 2.236$$

Example…

The distribution of Quiz Scores is uniform



Example…

If all samples of size 2 are taken with replacement and the mean of each sample is found, the distribution is:

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

Example…

A frequency distribution of sample means is as follows.

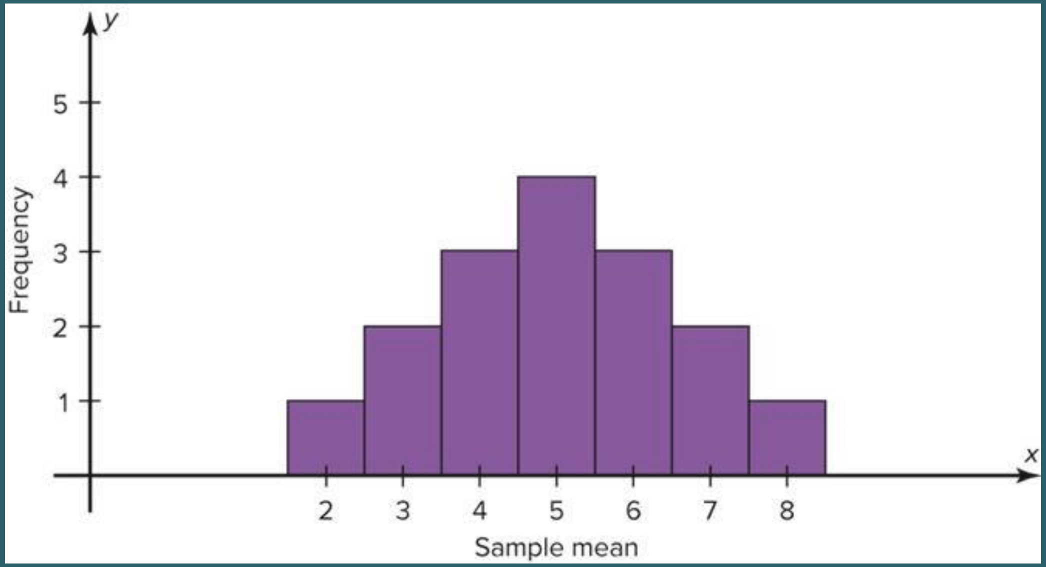
Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

$\bar{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1

Example…

The graph of the sample means appears to be normal:

$\bar{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1



Example…

$\bar{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1

The mean of the sample means is:

$$\mu_{\bar{X}} = \frac{2 + 3 + \cdots + 8}{16} = \frac{80}{16} = 5$$

which is the same as the population mean $\mu_{\bar{X}} = \mu$

The standard deviation of sample means is:

$$\sigma_{\bar{X}} = \frac{\sqrt{(2-5)^2 + (3-5)^2 + \cdots + (8-5)^2}}{16} \approx 1.581$$

which is equal to the population mean divided by $\sqrt{2}$

$$\sigma_{\bar{X}} = \frac{2.236}{\sqrt{2}} \approx 1.581$$

The Central Limit Theorem

- As the sample size *n* increases, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution.
- The mean of the sample means equals the population mean. $\mu_{\bar{X}} = \mu$.
- The standard deviation of the sample means is called the **standard error of the mean** (S.E.)

$$\sigma_{\bar{X}} = \sigma/\sqrt{n}.$$

The Central Limit Theorem…

- The central limit theorem can be used to answer questions about sample means in the same manner that the normal distribution can be used to answer questions about individual values.
- A new formula must be used for the *z* values:

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The Central Limit Theorem (p.291)

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ^2.

Let $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + \cdots + X_n$ be the sum of the sample observations.

Then if n is sufficiently large,

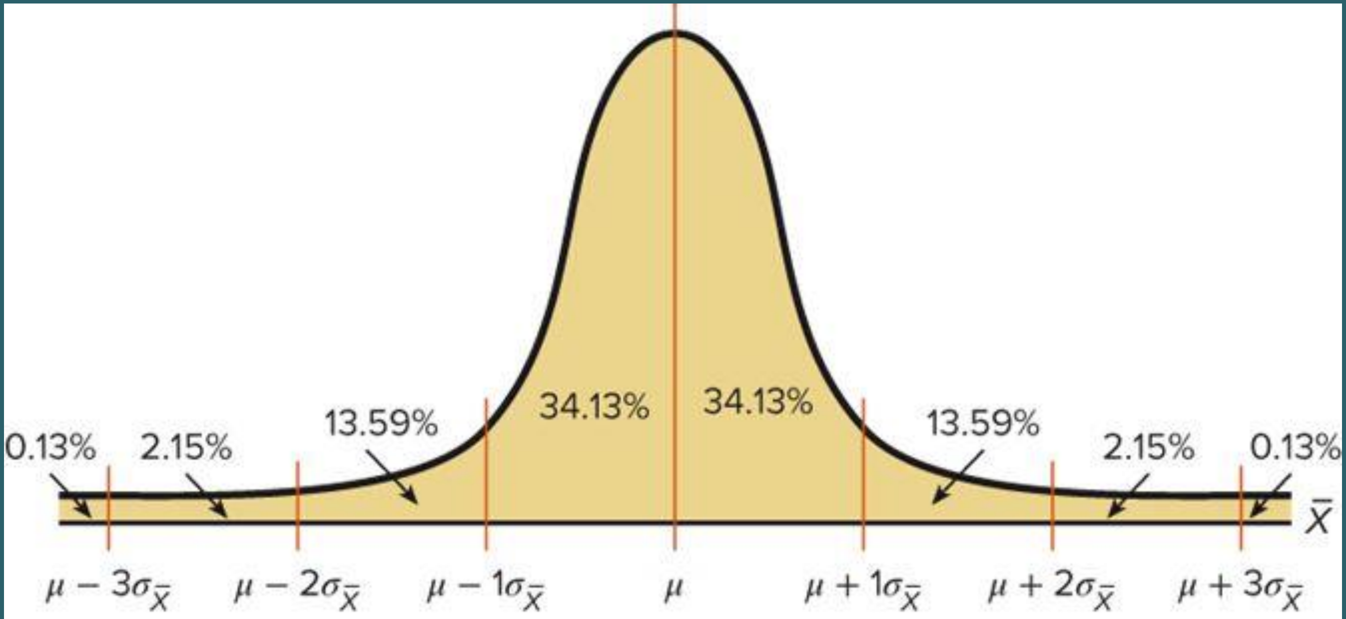
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approximately} \tag{4.55}$$

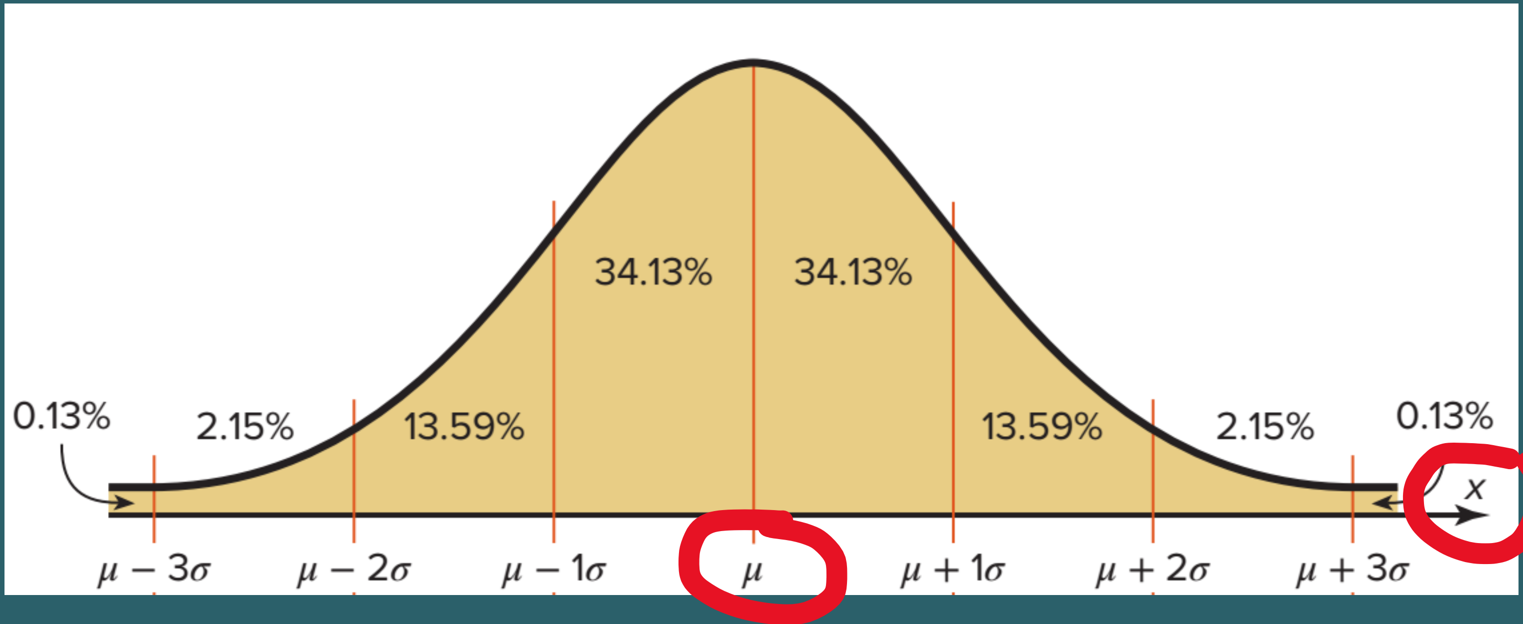
and

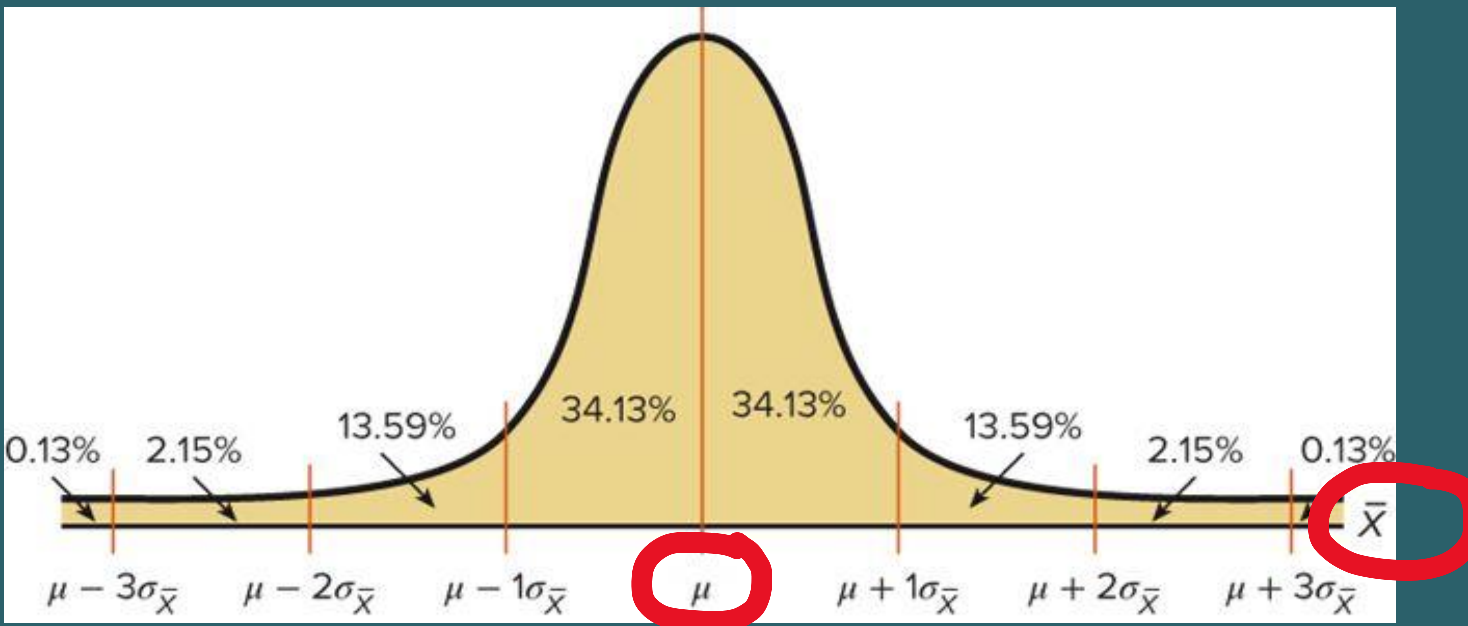
$$S_n \sim N(n\mu, n\sigma^2) \quad \text{approximately} \tag{4.56}$$

The Central Limit Theorem

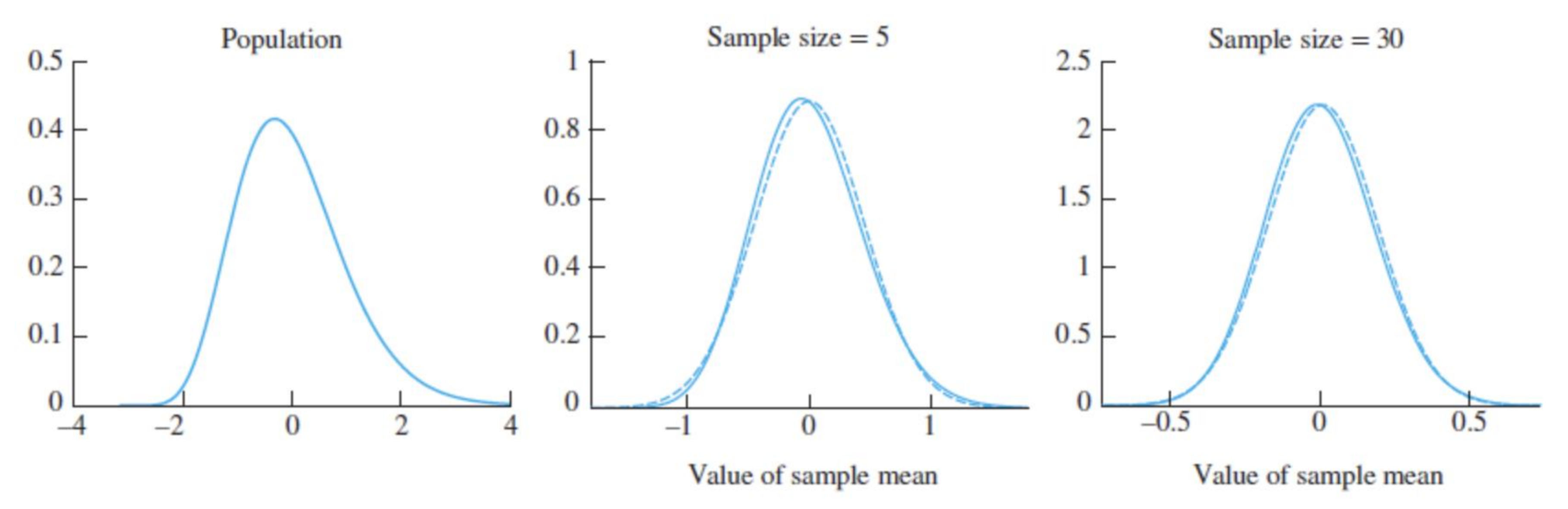
If (1) a large number of samples of a given size are selected from a normally distributed population, or (2) if a large number of samples of a given size that is greater than or equal to 30 are selected from a population that is not normally distributed, and the sample means are computed, then the distribution of sample means will look like the one shown below.




34.13%
34.13%
0.13%
2.15%
13.59%
13.59%
2.15%
0.13%
x
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - 1\sigma$
μ
$\mu + 1\sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

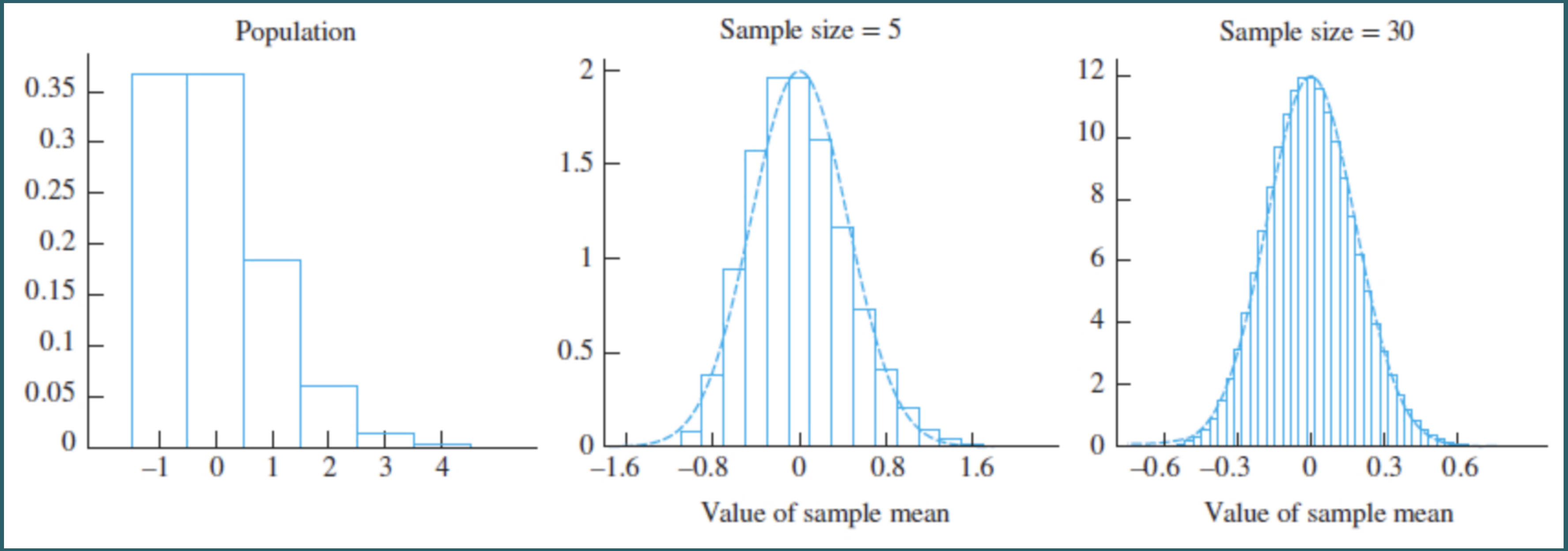

34.13%
34.13%
13.59%
13.59%
0.13%
2.15%
2.15%
0.13%
$\bar{X}$
$\mu - 3\sigma_{\bar{X}}$
$\mu - 2\sigma_{\bar{X}}$
$\mu - 1\sigma_{\bar{X}}$
μ
$\mu + 1\sigma_{\bar{X}}$
$\mu + 2\sigma_{\bar{X}}$
$\mu + 3\sigma_{\bar{X}}$

Example 1 (p.293)



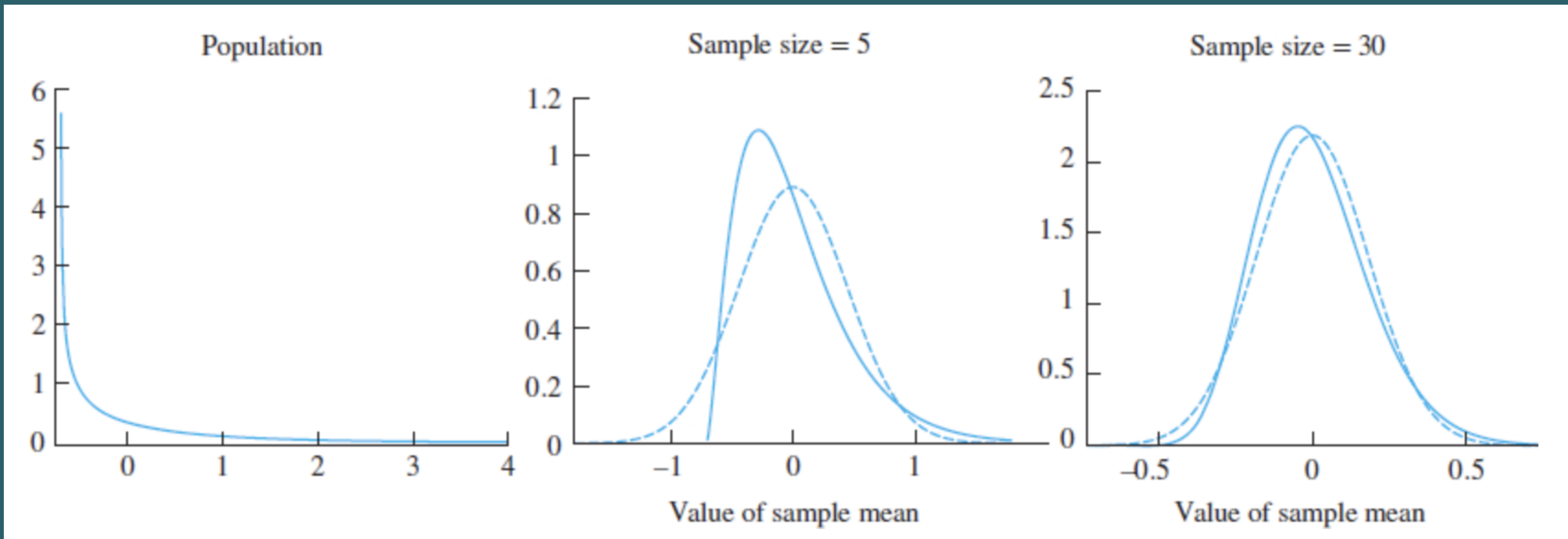
Since the original distribution is nearly symmetric, the normal approximation is good even for a sample size as small as 5.

Example 2 (p.293)



- The original distribution is somewhat skewed.
- Even so, the normal approximation is reasonably close even for a sample of size 5, and very good for a sample of size 30.

Example 3 (p.293)

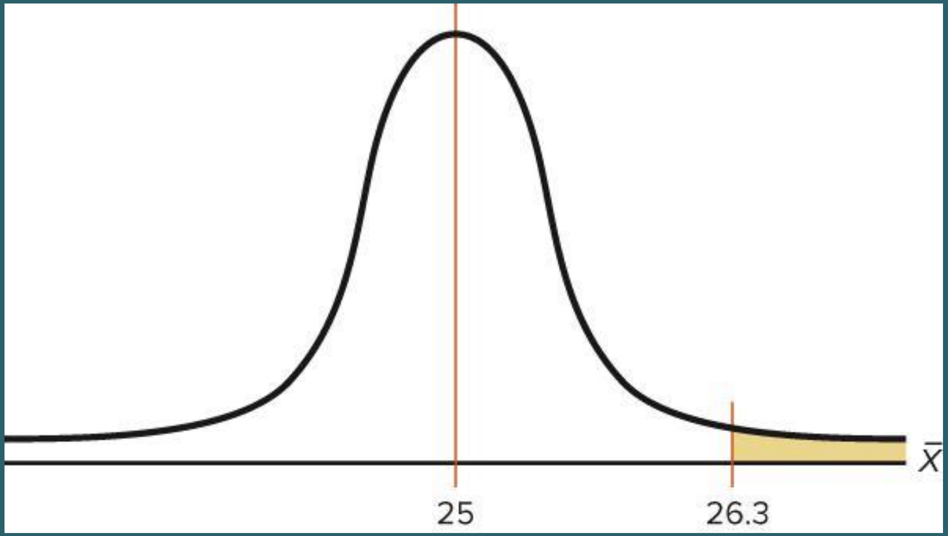


- The original distribution is highly skewed.
- The normal approximation is not good for a sample size of 5, but is reasonably good for a sample of size 30.

Example: Hours of Television

A. C. Nielsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.

Example: Hours of Television…



Since we are calculating probability for a sample mean, we need the Central Limit Theorem formula

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = 1.94$$

The area is 1.0000 – 0.9738 = 0.0262. The probability of obtaining a sample mean larger than 26.3 hours is 2.62%.

Area Corresponding to z=1.94 (=0.9738)

TABLE E *(continued)*

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	$\boxed{.9738}$	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890

Students sometimes have difficulty deciding whether to use

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \qquad \text{or} \qquad z = \frac{X - \mu}{\sigma}$$

The formula

$$z = \frac{\boxed{\overline{X}} - \mu}{\sigma / \sqrt{n}}$$

should be used to gain information about a sample mean, as shown in this section. The formula

$$z = \frac{\boxed{X} - \mu}{\sigma}$$

is used to gain information about an individual data value obtained from the population. Notice that the first formula contains $\overline{X}$, the symbol for the sample mean, while the second formula contains X, the symbol for an individual data value. Example 6–15 illustrates the uses of the two formulas.

Example: Working Weekends

The average time spent by construction workers who work on weekends is 7.93 hours (over 2 days). Assume the distribution is approximately normal and has a standard deviation of 0.8 hour.

a. Find the probability that an individual who works at that trade works fewer than 8 hours on the weekend.

b. If a sample of 40 construction workers is randomly selected, find the probability that the mean of the sample will be less than 8 hours.

Source: Bureau of Labor Statistics.

Example: Working Weekends…

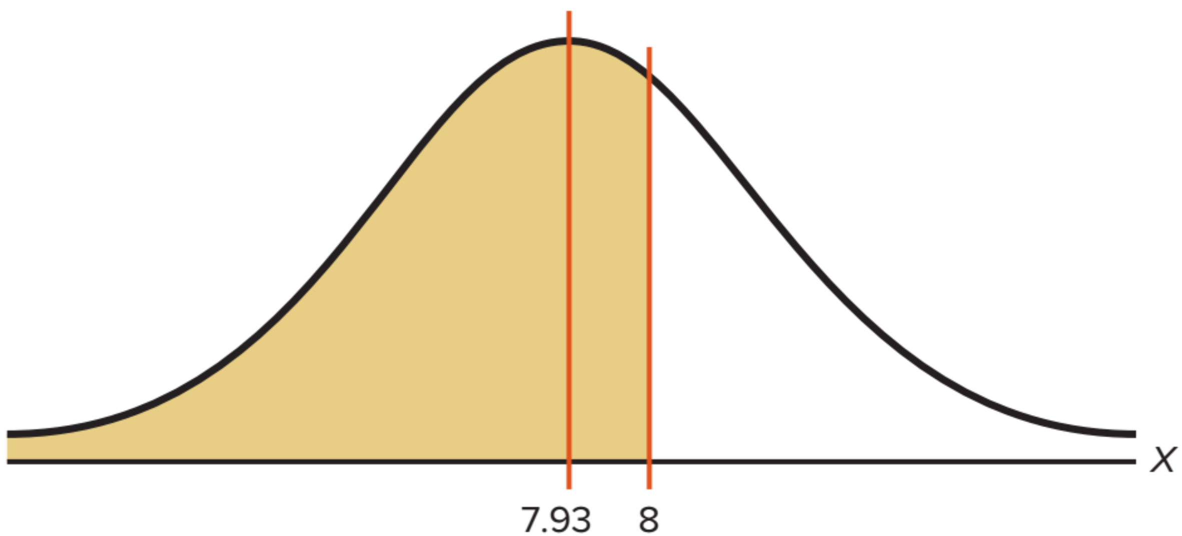
SOLUTION *a*

Step 1 Draw a normal distribution and shade the desired area.

Since the question concerns an individual person, the formula $z = (X - \mu)/\sigma$ is used. The distribution is shown in Figure 6–36.

FIGURE 6–36

Area Under a Normal Curve for Part *a* of Example 6–15



Example: Working Weekends…

Step 2 Find the z value.

$$Z = \frac{X - \mu}{\sigma} = \frac{8 - 7.93}{0.8} \approx 0.09$$

Step 3 Find the area to the left of $z = 0.09$.

It is 0.5359.

TABLE E *(continued)*

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141

Hence, the probability of selecting a construction worker who works less than 8 hours on a weekend is 0.5359, or 53.59%.

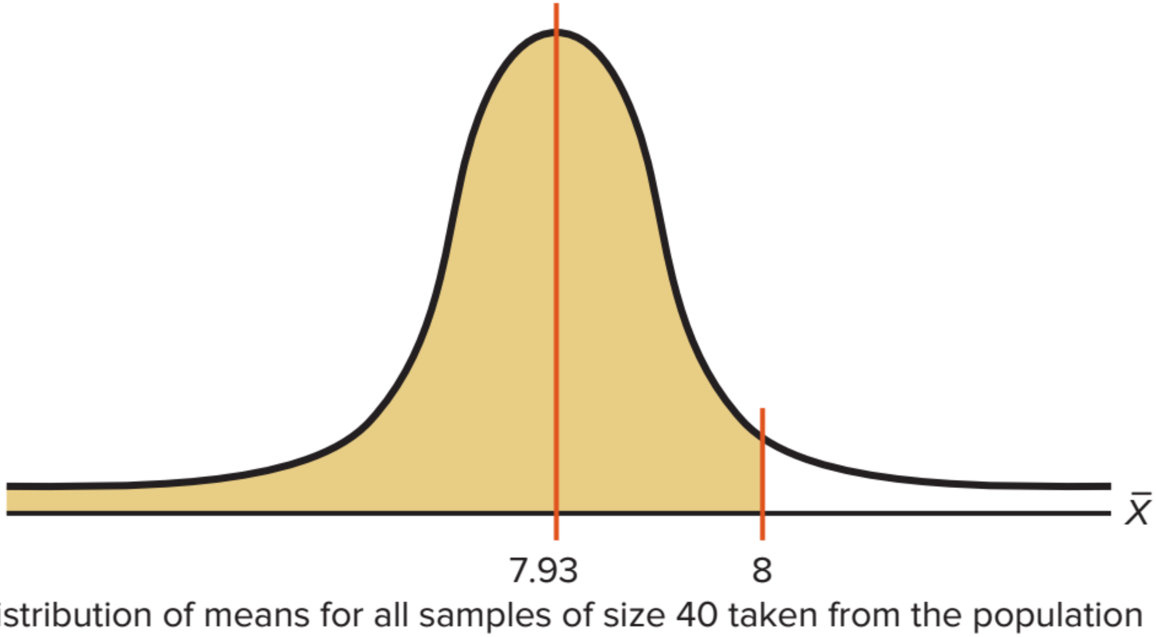
Example: Working Weekends…

SOLUTION *b*

Step 1 Draw a normal curve and shade the desired area.

Since the question concerns the mean of a sample with a size of 40, the central limit theorem formula $z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is used. The area is shown in Figure 6–37.

FIGURE 6–37
Area Under a Normal Curve for Part *b* of Example 6–15



Example: Working Weekends…

Step 2 Find the z value for a mean of 8 hours and a sample size of 40.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{8 - 7.93}{0.8/\sqrt{40}} \approx 0.55$$

Step 3 Find the area corresponding to $z = 0.55$. The area is 0.7088.

Hence, the probability of getting a sample mean of less than 8 hours when the sample size is 40 is 0.7088, or 70.88%.

TABLE E *(continued)*

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224

Continuity Correction

- The Central Limit Theorem holds for both continuous and discrete distributions.
- For example, the binomial distribution is discrete, while the normal distribution is continuous.
- The continuity correction is an adjustment, made when approximating a discrete distribution with a continuous one, that can improve the accuracy of the approximation.

Continuity Correction…

- A fair coin is tossed 100 times.
- Let X represent the number of heads. Then **X ~ Bin(100, 0.5)**.
- Imagine that we wish to compute the probability that X is between 45 and 55.
- This probability will differ depending on whether the endpoints, 45 and 55, are included or excluded.

Continuity Correction…

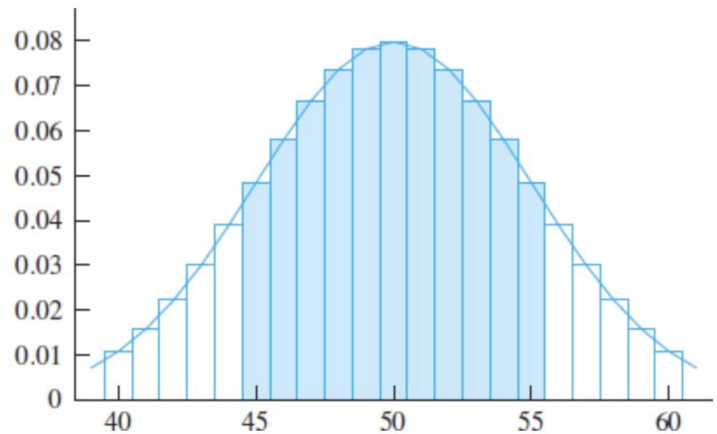
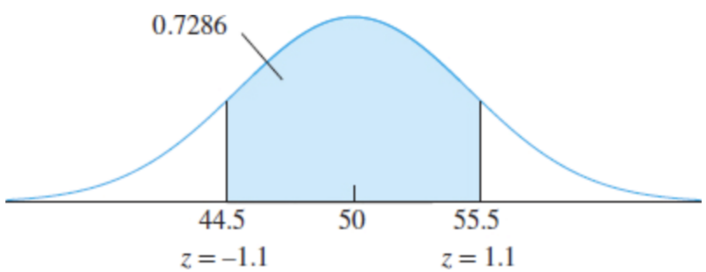


FIGURE 4.28 To compute $P(45 \le X \le 55)$, the areas of the rectangles corresponding to 45 and to 55 should be included. To approximate this probability with the normal curve, compute the area under the curve between 44.5 and 55.5.

$$z = \frac{44.5 - 50}{5} = -1.1, \qquad z = \frac{55.5 - 50}{5} = 1.1$$

From the *z* table we find that the probability is 0.7286. See Figure 4.30 (page 298).



Continuity Correction…

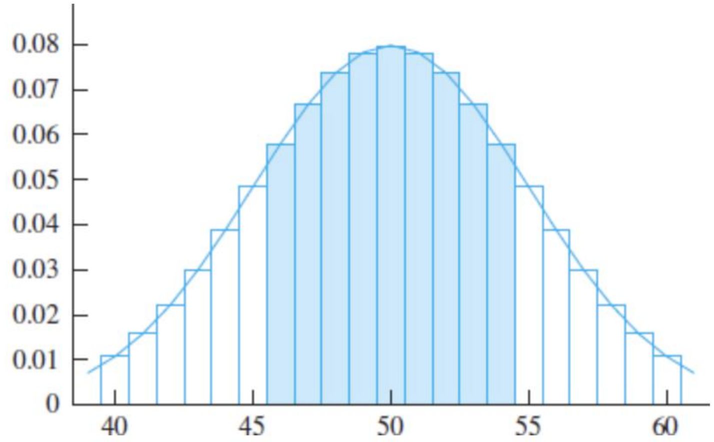
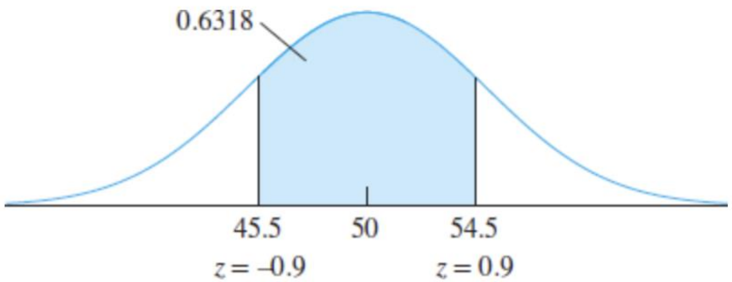


FIGURE 4.29 To compute $P(45 < X < 55)$, the areas of the rectangles corresponding to 45 and to 55 should be excluded. To approximate this probability with the normal curve, compute the area under the curve between 45.5 and 54.5.

$$z = \frac{45.5 - 50}{5} = -0.9, \qquad z = \frac{54.5 - 50}{5} = 0.9$$

From the z table we find that the probability is 0.6318. See Figure 4.31.



End of Section Ch.4