Chapter 6
Normal Probability Distributions

6-1 Review and Preview

6-2 The Standard Normal Distribution

6-3 Applications of Normal Distributions

6-4 Sampling Distributions and Estimators

6-5 The Central Limit Theorem

6-6 Normal as Approximation to Binomial

6-7 Assessing Normality

Preview

Chapter focus is on:

- Continuous random variables
- Normal distributions

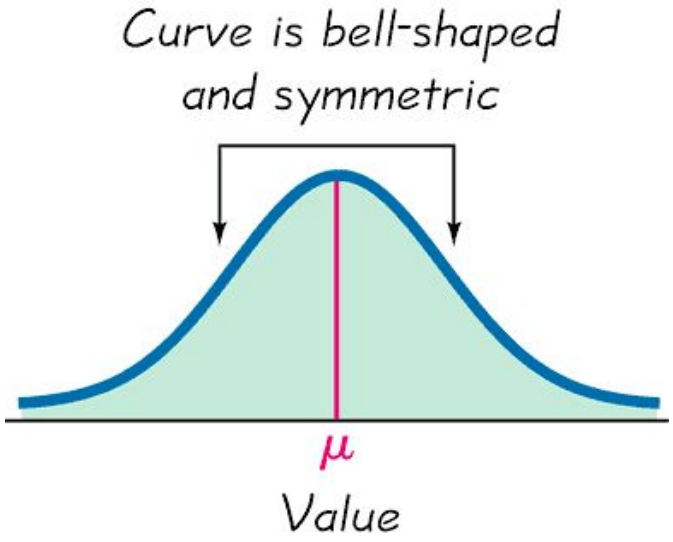


Figure 6-1

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Formula 6-1

Distribution determined by fixed values of mean and standard deviation

Key Concept

This section presents the *standard normal distribution* which has three properties:

1. **It's graph is bell-shaped.**
2. **It's mean is equal to 0 ($\mu = 0$).**
3. **It's standard deviation is equal to 1 ($\sigma = 1$).**

Develop the skill to find areas (or probabilities or relative frequencies) corresponding to various regions under the graph of the standard normal distribution. Find *z*-scores that correspond to area under the graph.

Uniform Distribution

A continuous random variable has a uniform distribution if its values are spread evenly over the range of probabilities. The graph of a uniform distribution results in a rectangular shape.

Density Curve

A **density curve** is the graph of a continuous probability distribution. It must satisfy the following properties:

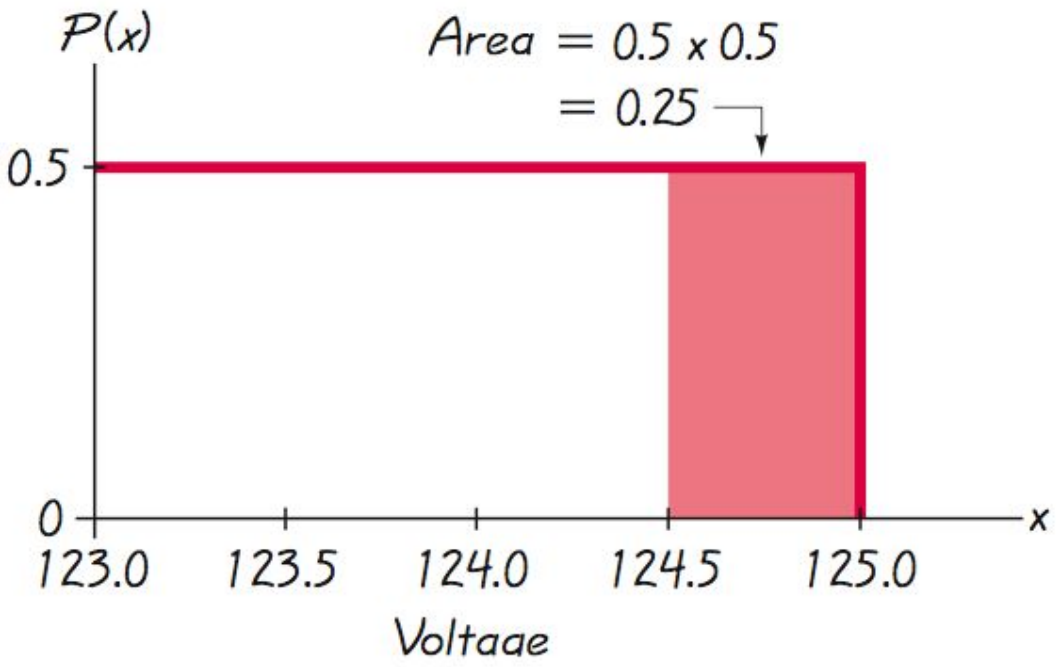
1. The total area under the curve must equal 1.
2. Every point on the curve must have a vertical height that is 0 or greater. (That is, the curve cannot fall below the *x*-axis.)

Area and Probability

Because the total area under the density curve is equal to 1, there is a correspondence between *area* and *probability*.

Using Area to Find Probability

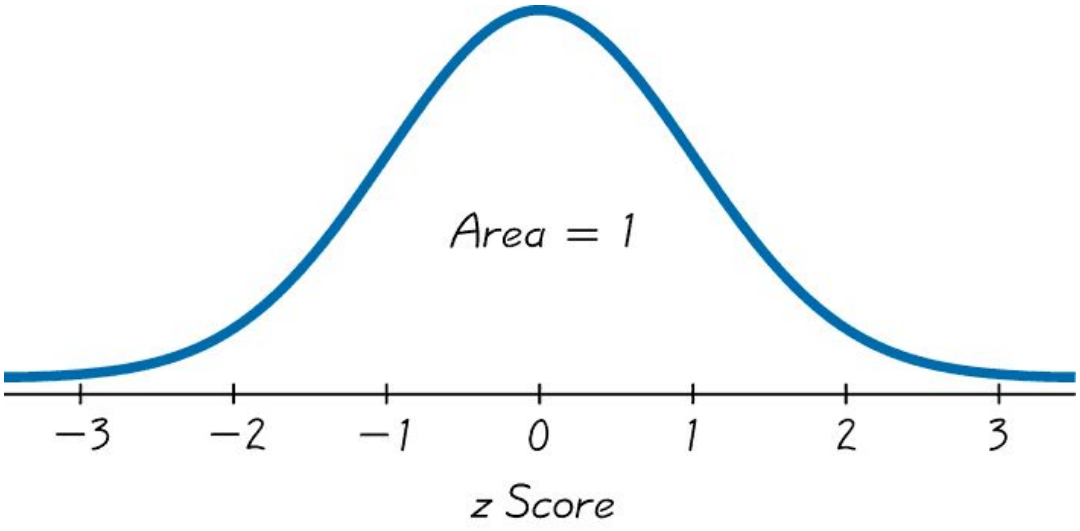
Given the uniform distribution illustrated, find the probability that a randomly selected voltage level is greater than 124.5 volts.



Shaded area represents voltage levels greater than 124.5 volts. Correspondence between area and probability: 0.25.

Standard Normal Distribution

The standard normal distribution is a normal probability distribution with $\mu = 0$ and $\sigma = 1$. The total area under its density curve is equal to 1.



Finding Probabilities When Given *z*-scores

- Table A-2 (in Appendix A)
- Formulas and Tables insert card
- Find areas for many different regions

Table A-2

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

Using Table A-2

1. It is designed only for the *standard* normal distribution, which has a mean of 0 and a standard deviation of 1.
2. It is on two pages, with one page for *negative* *z*-scores and the other page for *positive* *z*-scores.
3. Each value in the body of the table is a *cumulative area from the left* up to a vertical boundary above a specific *z*-score.

Using Table A-2

4. When working with a graph, avoid confusion between *z*-scores and areas.

   ***z* Score**

   Distance along horizontal scale of the standard normal distribution; refer to the leftmost column and top row of Table A-2.

   Area

   Region under the curve; refer to the values in the body of Table A-2.

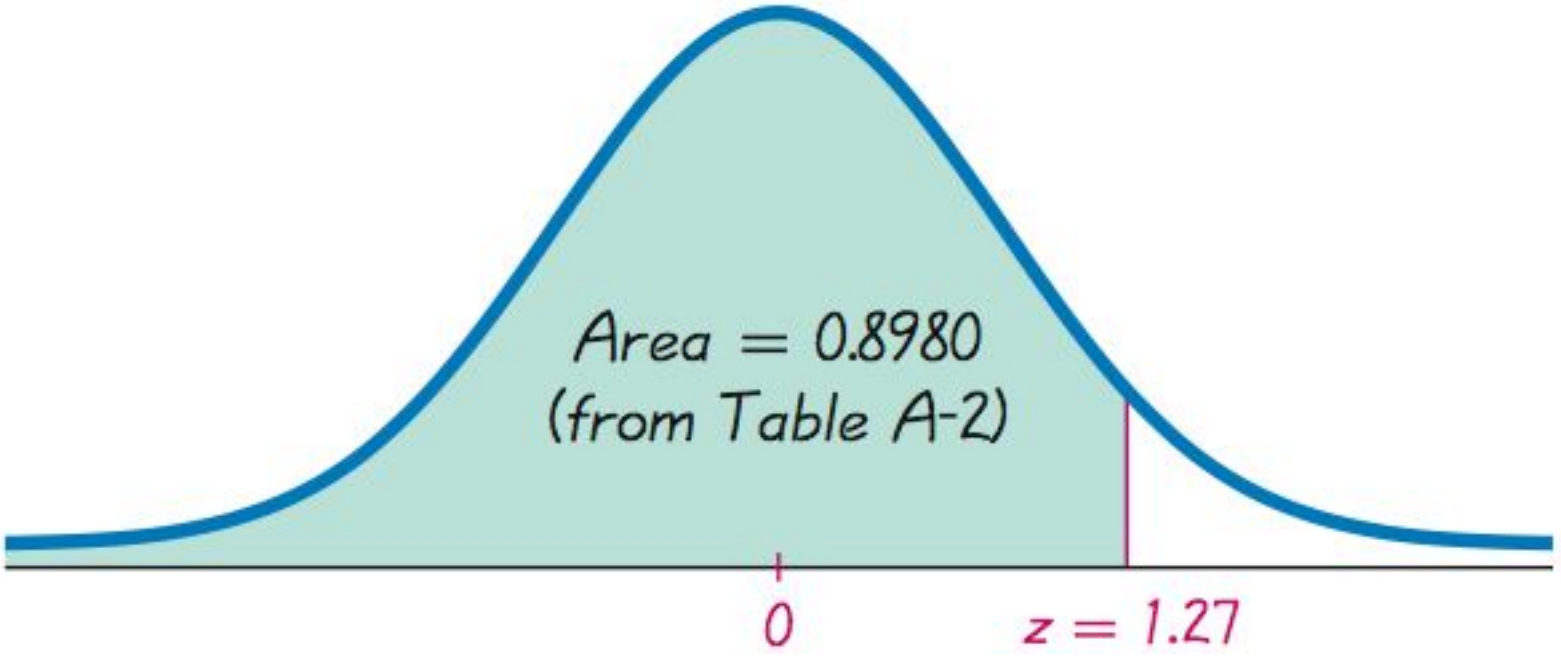
5. The part of the z-score denoting hundredths is found across the top.

Example - Thermometers

The Precision Scientific Instrument Company manufactures thermometers that are supposed to give readings of 0ºC at the freezing point of water. Tests on a large sample of these instruments reveal that at the freezing point of water, some thermometers give readings below 0º (denoted by negative numbers) and some give readings above 0º (denoted by positive numbers). Assume that the mean reading is 0ºC and the standard deviation of the readings is 1.00ºC. Also assume that the readings are normally distributed. If one thermometer is randomly selected, find the probability that, at the freezing point of water, the reading is less than 1.27º.

Example - (Continued)

P(z < 1.27) =




```
> pnorm(1.27, mean=0, sd=1)
[1] 0.8979577
```

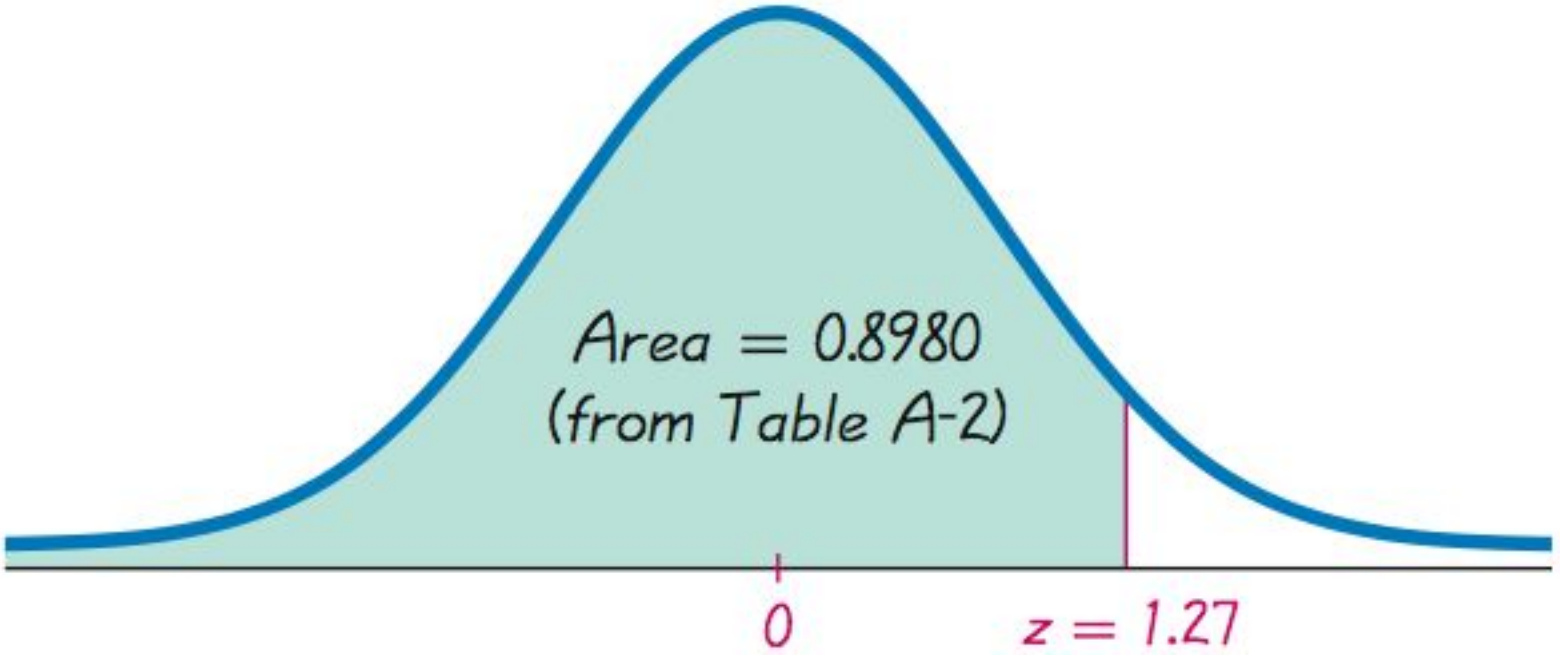
Look at Table A-2

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292

Example - cont

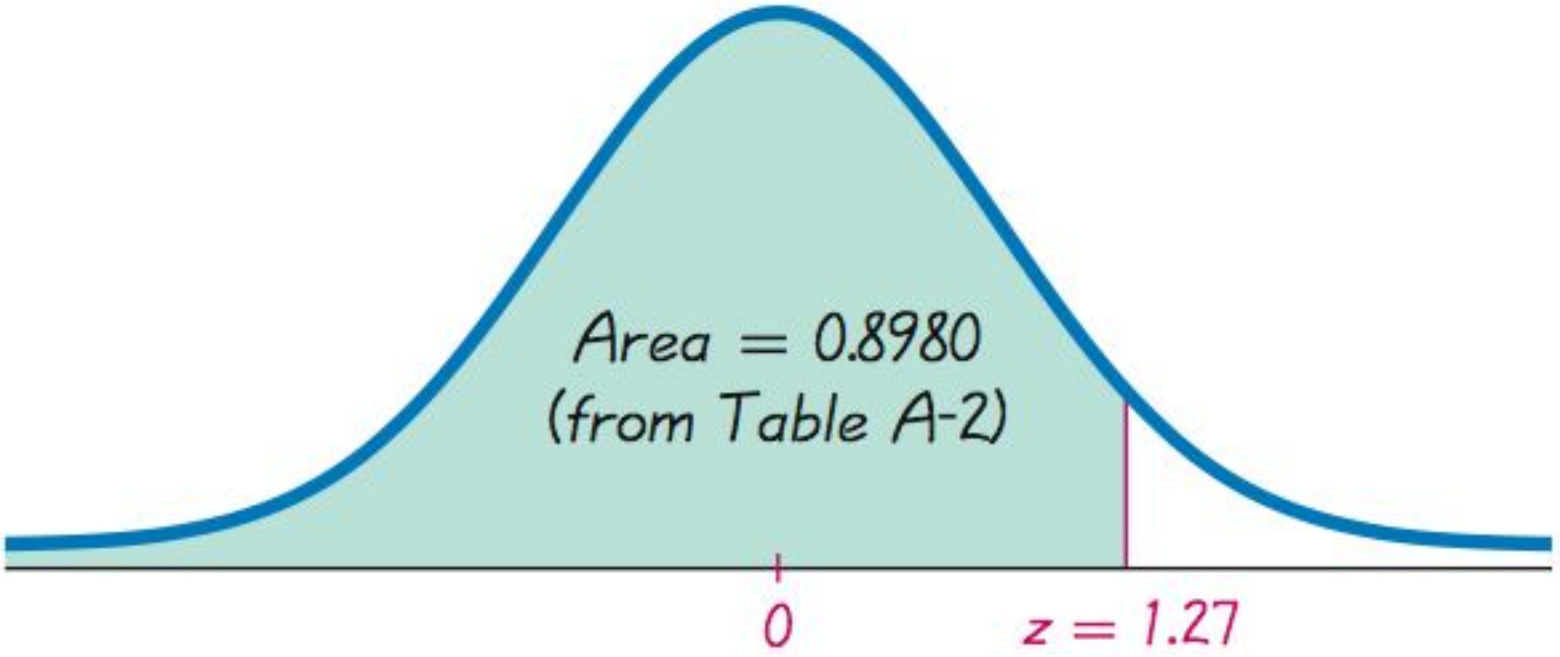
$P(z < 1.27) = 0.8980$



The *probability* of randomly selecting a thermometer with a reading less than 1.27° is 0.8980.

Example - cont

$P(z < 1.27) = 0.8980$

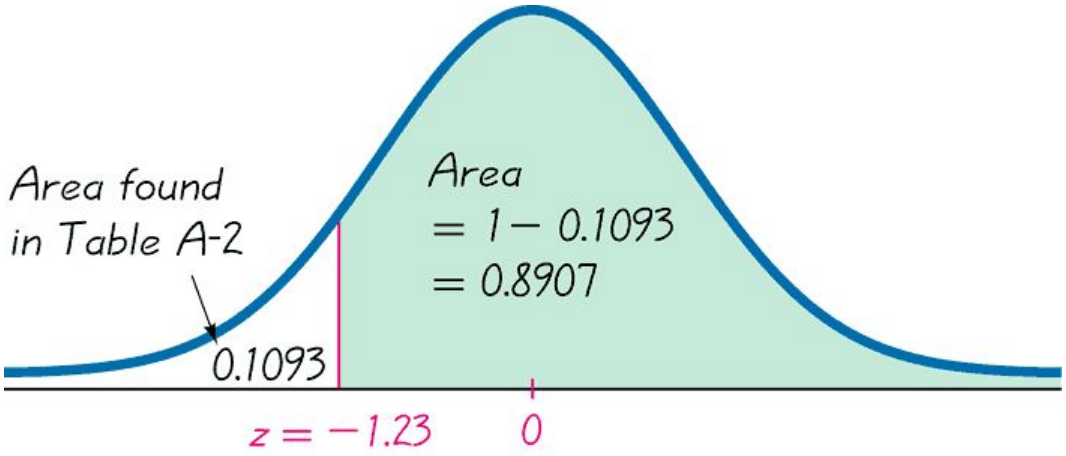


Or 89.80% will have readings below 1.27º.

Example - Thermometers Again

If thermometers have an average (mean) reading of 0 degrees and a standard deviation of 1 degree for freezing water, and if one thermometer is randomly selected, find the probability that it reads (at the freezing point of water) above –1.23 degrees.

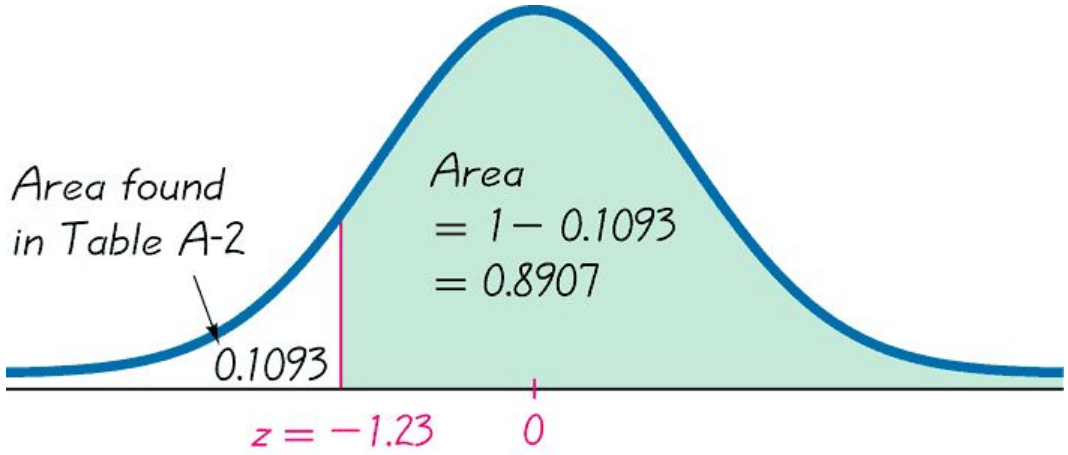
$$P(z > -1.23) = 0.8907$$



Probability of randomly selecting a thermometer with a reading above –1.23º is 0.8907.

Example - cont

P (z > –1.23) = 0.8907

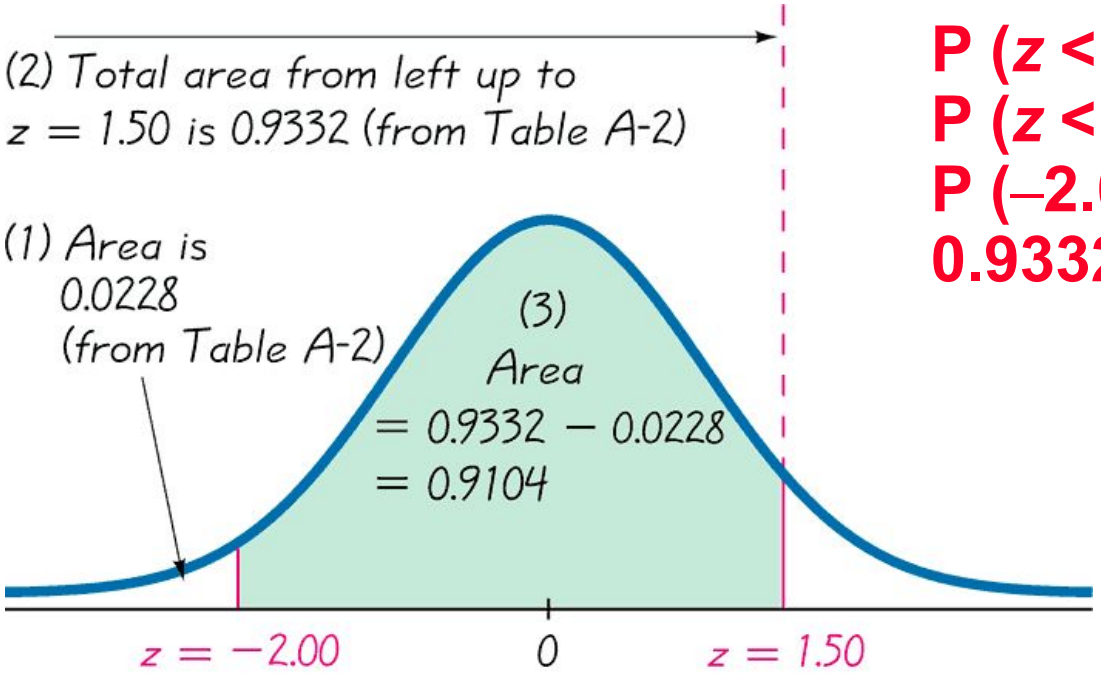


89.07% of the thermometers have readings above –1.23 degrees.

```
> pnorm(-1.23, mean=0, sd=1, lower.tail = FALSE)
[1] 0.8906514
```

Example - Thermometers III

A thermometer is randomly selected. Find the probability that it reads (at the freezing point of water) between –2.00 and 1.50 degrees.

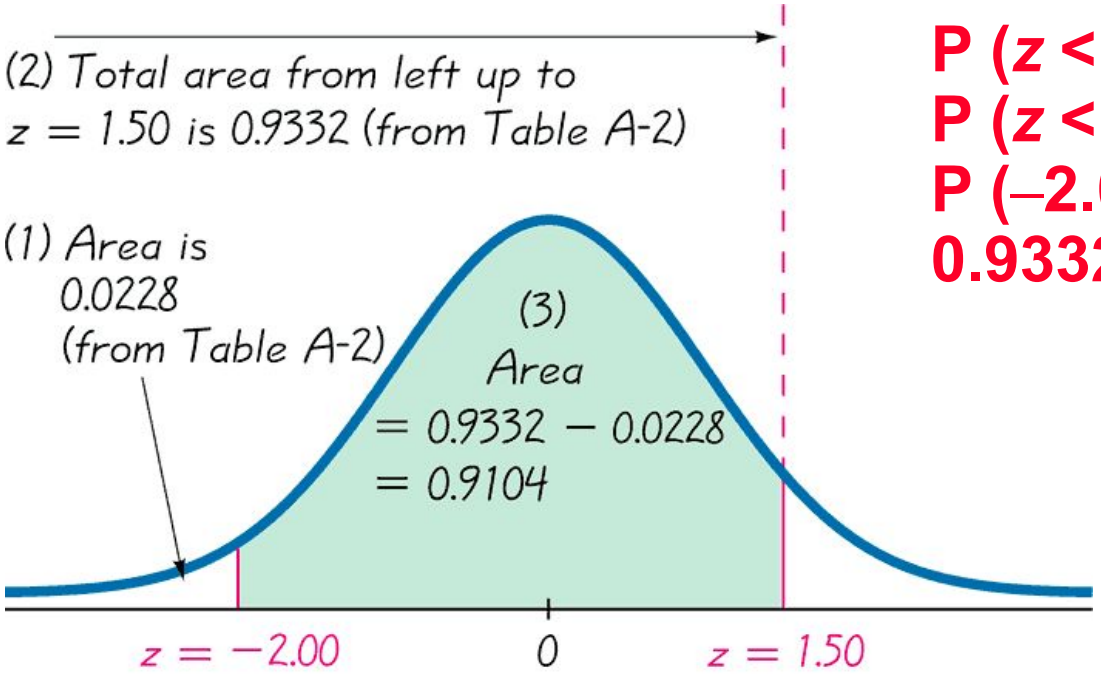


P (z < –2.00) = 0.0228
P (z < 1.50) = 0.9332
P (–2.00 < z < 1.50) =
0.9332 – 0.0228 = 0.9104

The probability that the chosen thermometer has a reading between – 2.00 and 1.50 degrees is 0.9104.

Example - cont

A thermometer is randomly selected. Find the probability that it reads (at the freezing point of water) between –2.00 and 1.50 degrees.



P (z < –2.00) = 0.0228
P (z < 1.50) = 0.9332
P (–2.00 < z < 1.50) =
0.9332 – 0.0228 = 0.9104

If many thermometers are selected and tested at the freezing point of water, then 91.04% of them will read between –2.00 and 1.50 degrees.

Notation

$P(a < z < b)$

denotes the probability that the *z* score is between *a* and *b*.

$P(z > a)$

denotes the probability that the *z* score is greater than *a*.

$P(z < a)$

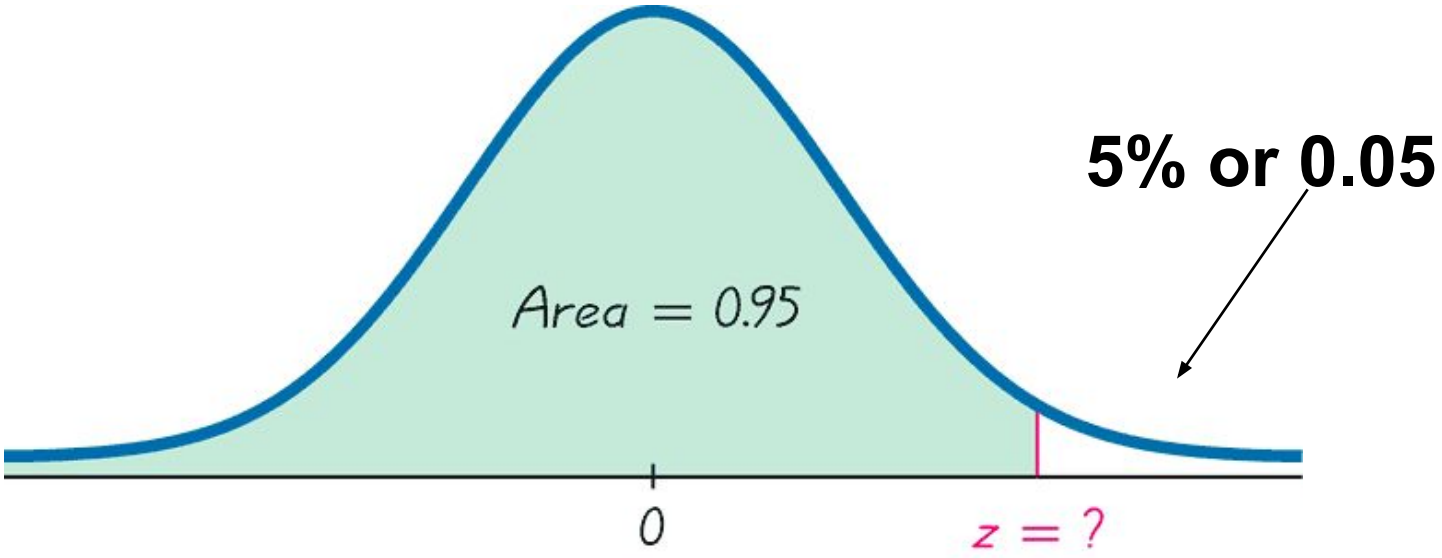
denotes the probability that the *z* score is less than *a*.

Finding a *z* Score When Given a Probability Using Table A-2

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given probability. If that region is not a cumulative region from the left, work instead with a known region that is a cumulative region from the left.

2. Using the cumulative area from the left, locate the closest probability in the body of Table A-2 and identify the corresponding *z* score.

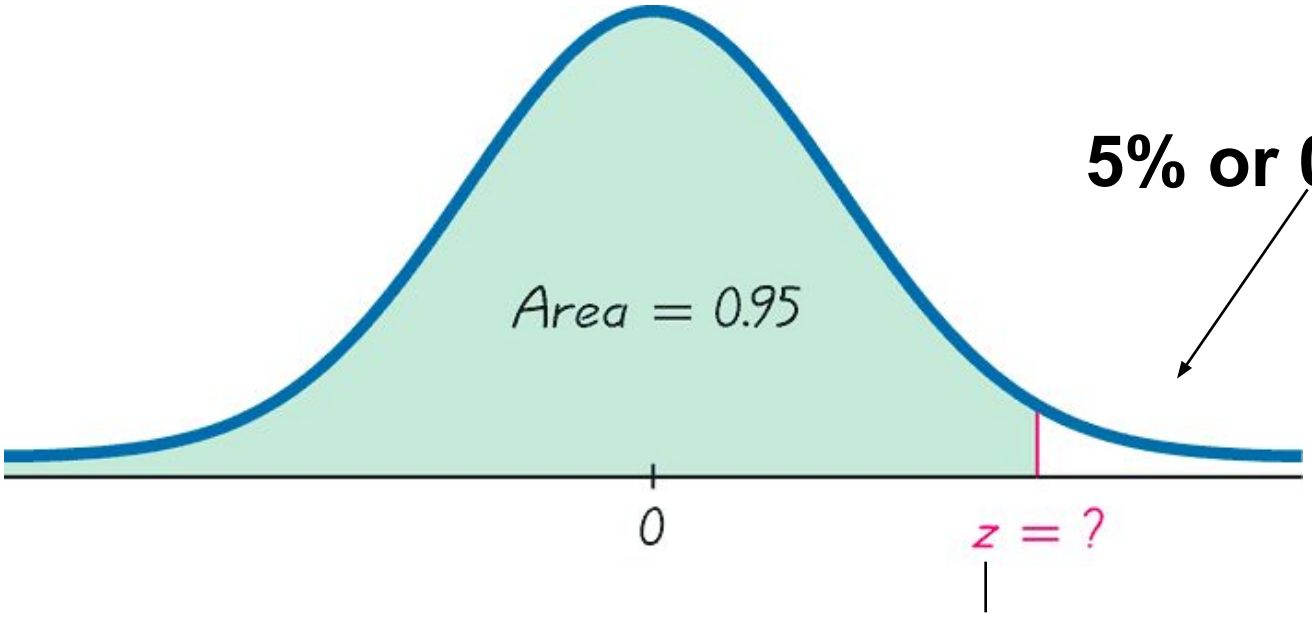
Finding *z* Scores When Given Probabilities



(*z* score will be positive)

Finding the 95th Percentile

Finding *z* Scores When Given Probabilities - cont



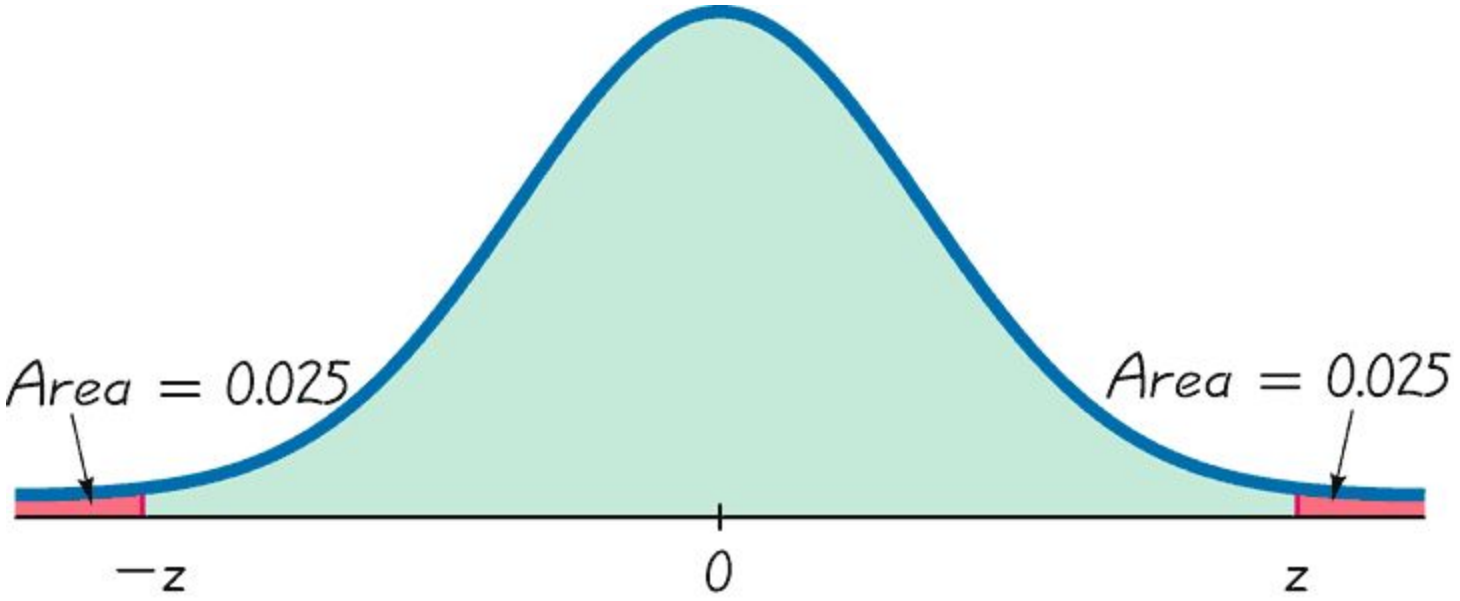
(*z* score will be positive)

1.645

Finding the 95th Percentile

```
> qnorm(.95, mean=0, sd=1)
[1] 1.644854
```

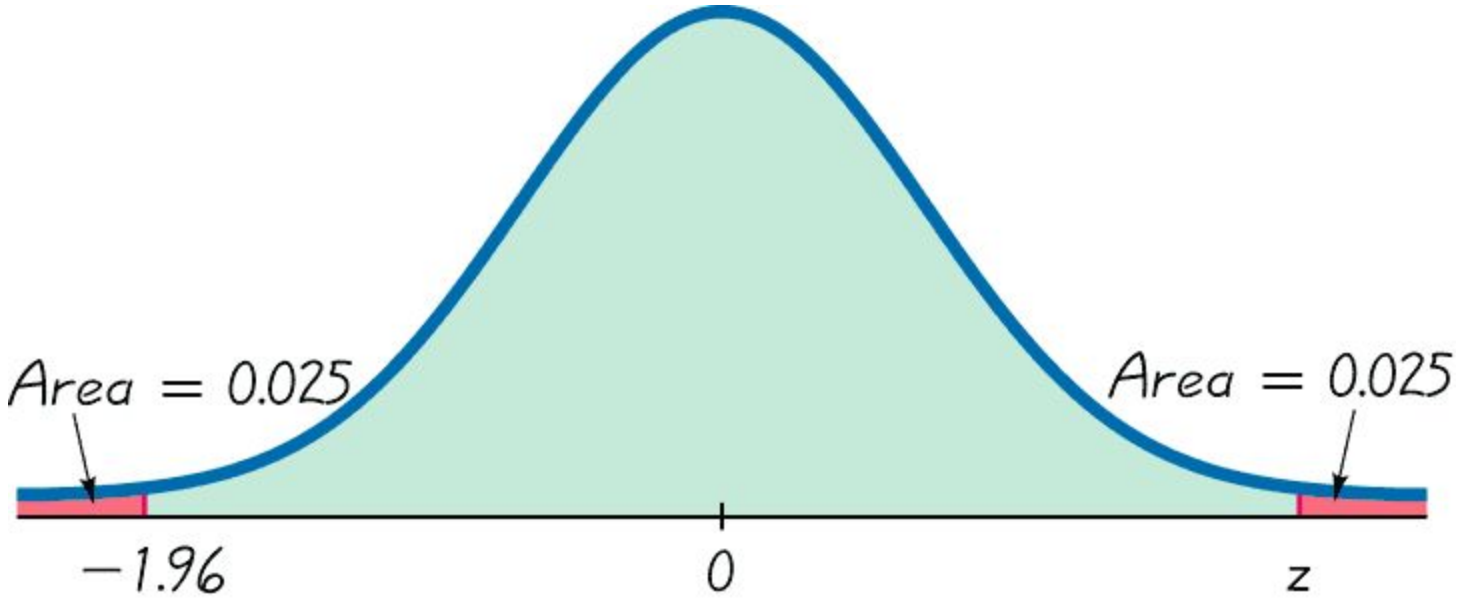
Finding *z* Scores When Given Probabilities - cont



(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

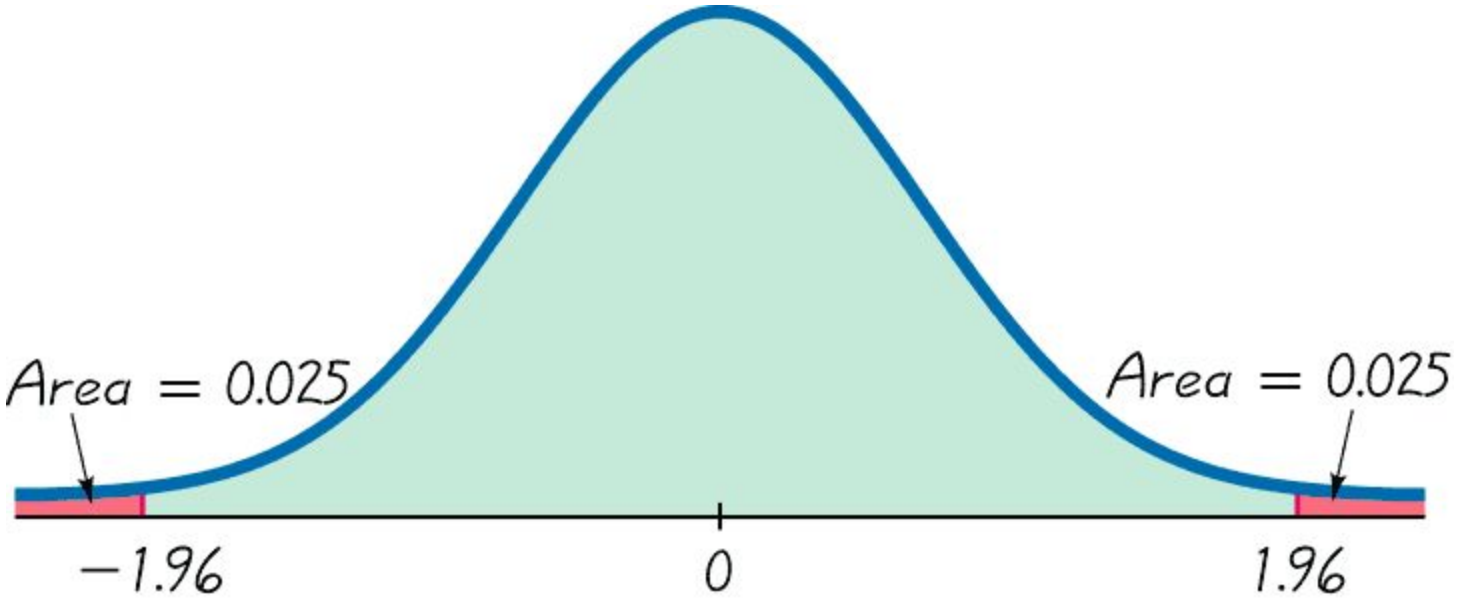
Finding *z* Scores When Given Probabilities - cont



(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

Finding *z* Scores When Given Probabilities - cont



(One *z* score will be negative and the other positive)

Finding the Bottom 2.5% and Upper 2.5%

```
> qnorm(.025, mean=0, sd=1)
[1] -1.959964
> qnorm(.025, mean=0, sd=1, lower.tail = FALSE)
[1] 1.959964
```

Key Concept

This section presents methods for working with normal distributions that are not standard. That is, the mean is not 0 or the standard deviation is not 1, or both.

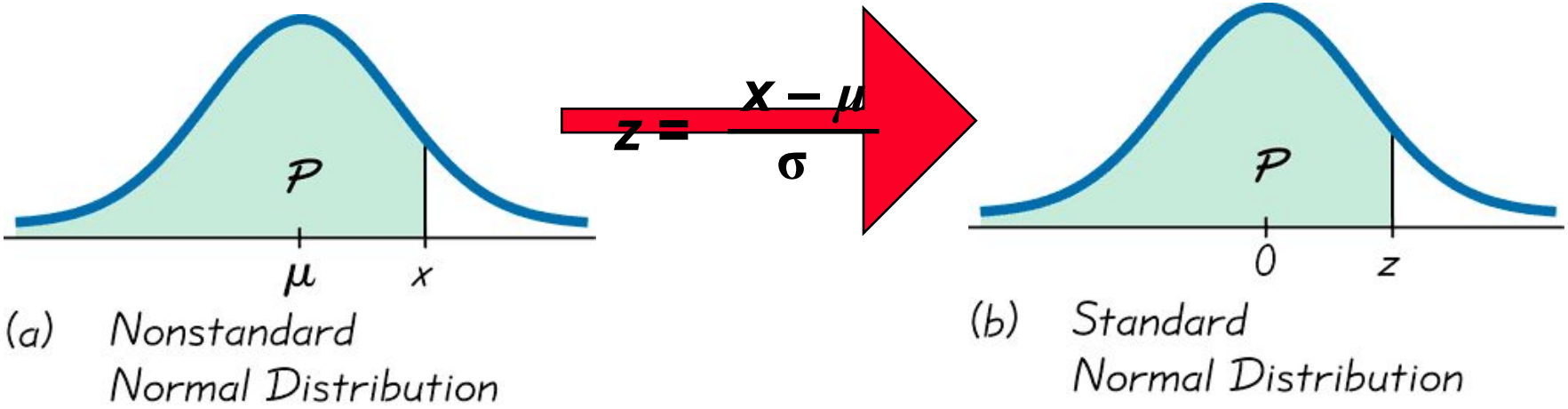
The key concept is that we can use a simple conversion that allows us to standardize any normal distribution so that the same methods of the previous section can be used.

Conversion Formula

$$z = \frac{x - \mu}{\sigma}$$

Round *z* scores to 2 decimal places

Converting to a Standard Normal Distribution



(a) Nonstandard Normal Distribution

(b) Standard Normal Distribution

Example – Weights of Water Taxi Passengers

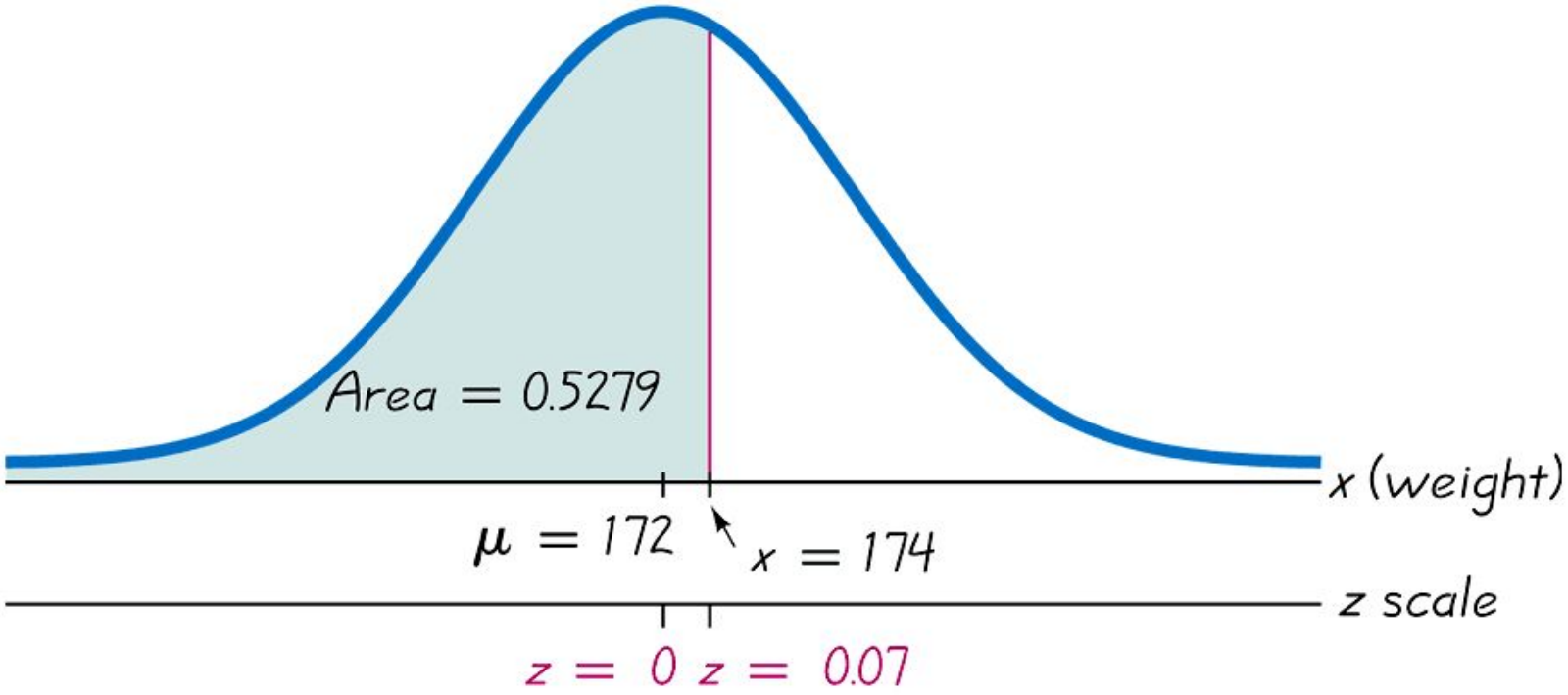
In the Chapter Problem, we noted that the safe load for a water taxi was found to be 3500 pounds. We also noted that the mean weight of a passenger was assumed to be 140 pounds. Assume the worst case that all passengers are men. Assume also that the weights of the men are normally distributed with a mean of 172 pounds and standard deviation of 29 pounds. If one man is randomly selected, what is the probability he weighs less than 174 pounds?

Example - cont

$\mu = 172$

$\sigma =$

$$z = \frac{174 - 172}{29} = 0.07$$

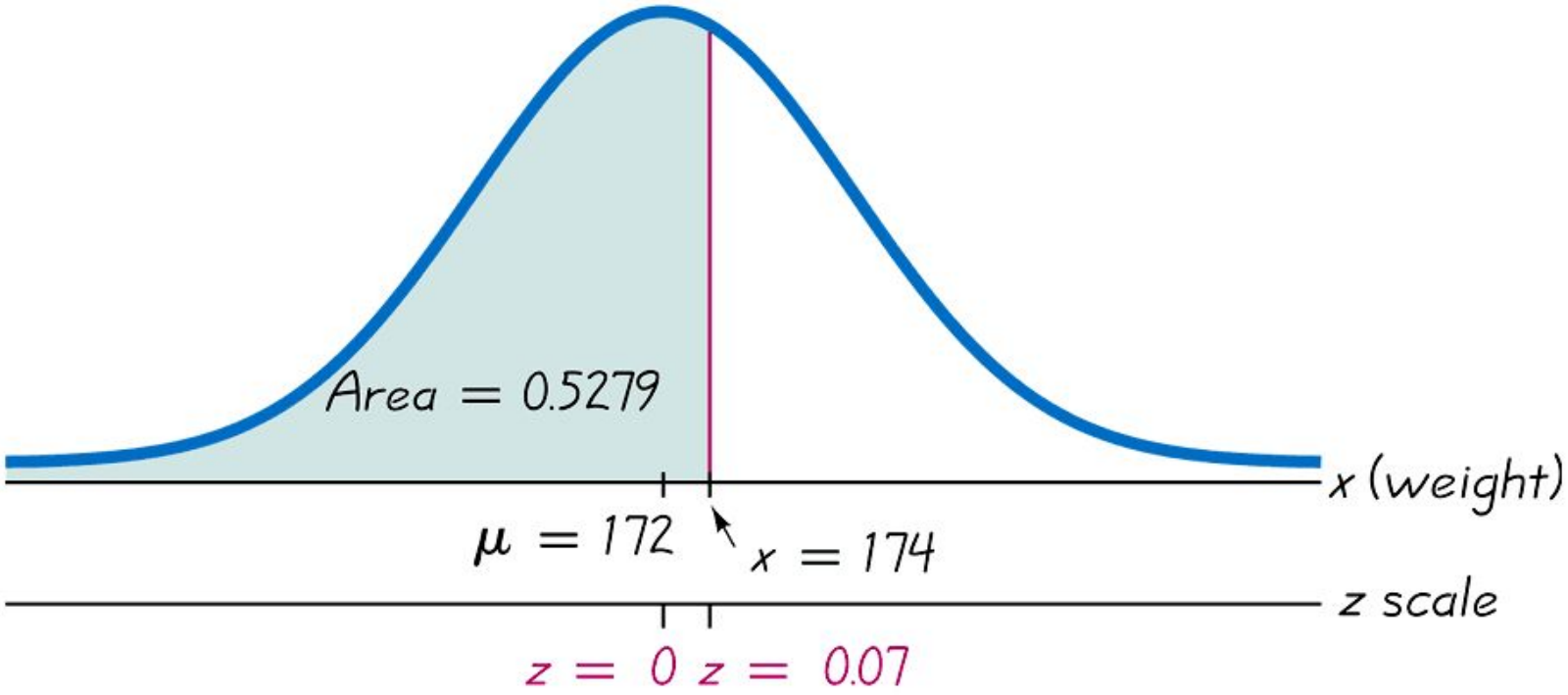


Example - cont

μ = 172

σ =

P (x < 174 lb.) = P(z < 0.07)



Helpful Hints

1. Don't confuse *z* scores and areas. *z* scores are distances along the horizontal scale, but areas are regions under the normal curve. Table A-2 lists *z* scores in the left column and across the top row, but areas are found in the body of the table.

2. Choose the correct (right/left) side of the graph.

3. A z score must be negative whenever it is located in the left half of the normal distribution.

4. Areas (or probabilities) are positive or zero values, but they are never negative.

Procedure for Finding Values Using Table A-2 and Formula 6-2

1. Sketch a normal distribution curve, enter the given probability or percentage in the appropriate region of the graph, and identify the *x* value(s) being sought.

2. Use Table A-2 to find the *z* score corresponding to the cumulative left area bounded by *x*. Refer to the body of Table A-2 to find the closest area, then identify the corresponding *z* score.

3. Using Formula 6-2, enter the values for μ, σ, and the *z* score found in step 2, then solve for *x*.

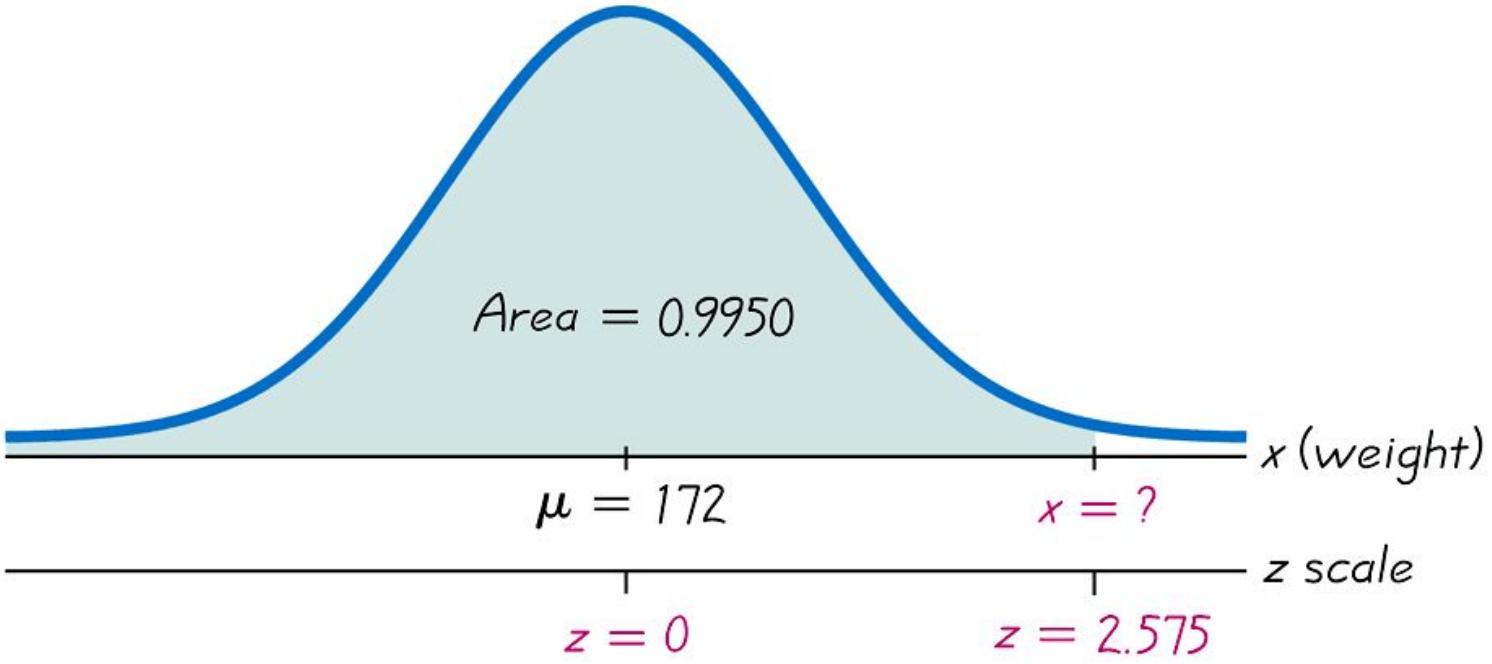
$$x = \mu + (z \cdot \sigma)$$ (Another form of Formula 6-2)

(If *z* is located to the left of the mean, be sure that it is a negative number.)

4. Refer to the sketch of the curve to verify that the solution makes sense in the context of the graph and the context of the problem.

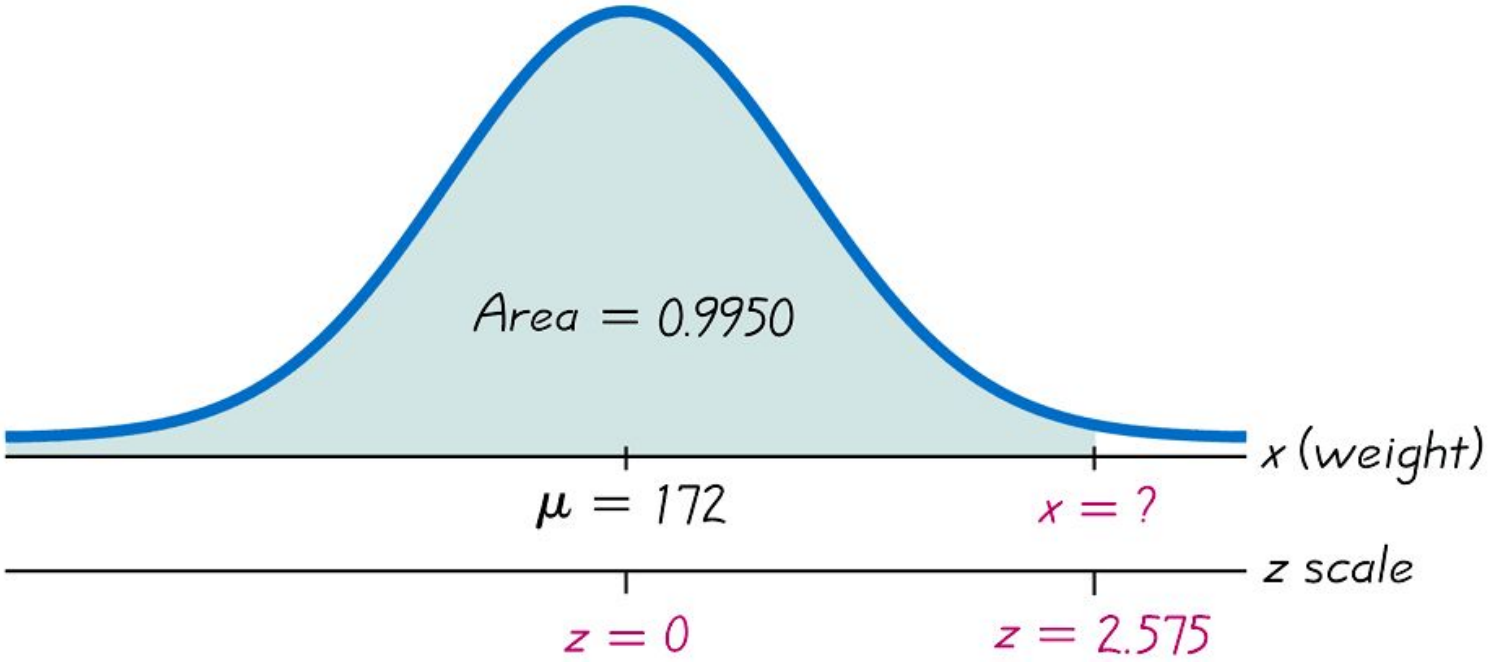
Example – Lightest and Heaviest

Use the data from the previous example to determine what weight separates the lightest 99.5% from the heaviest 0.5%?



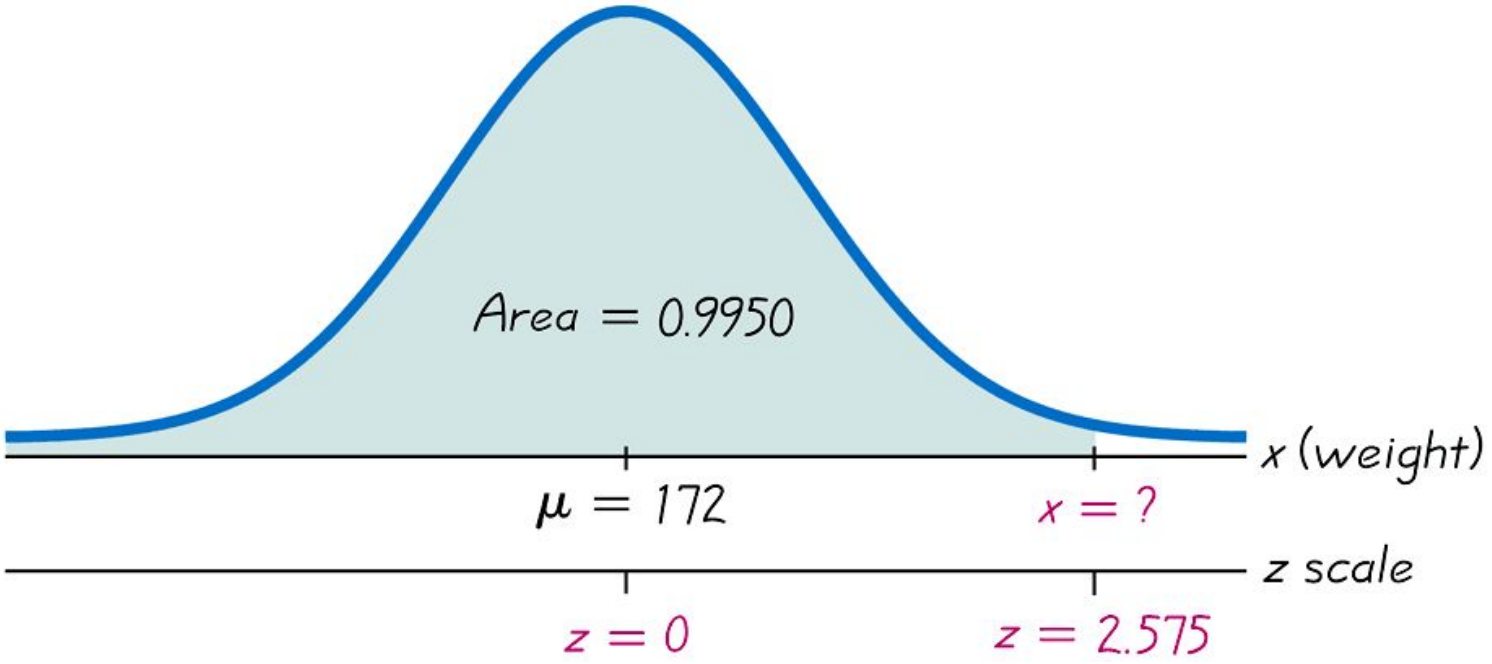
Example – Lightest and Heaviest - cont

$$x = \mu + (z \bullet \sigma)$$
$$x = 172 + (2.575 \cdot 29)$$
$$x = 246.675 \text{ (247 rounded)}$$



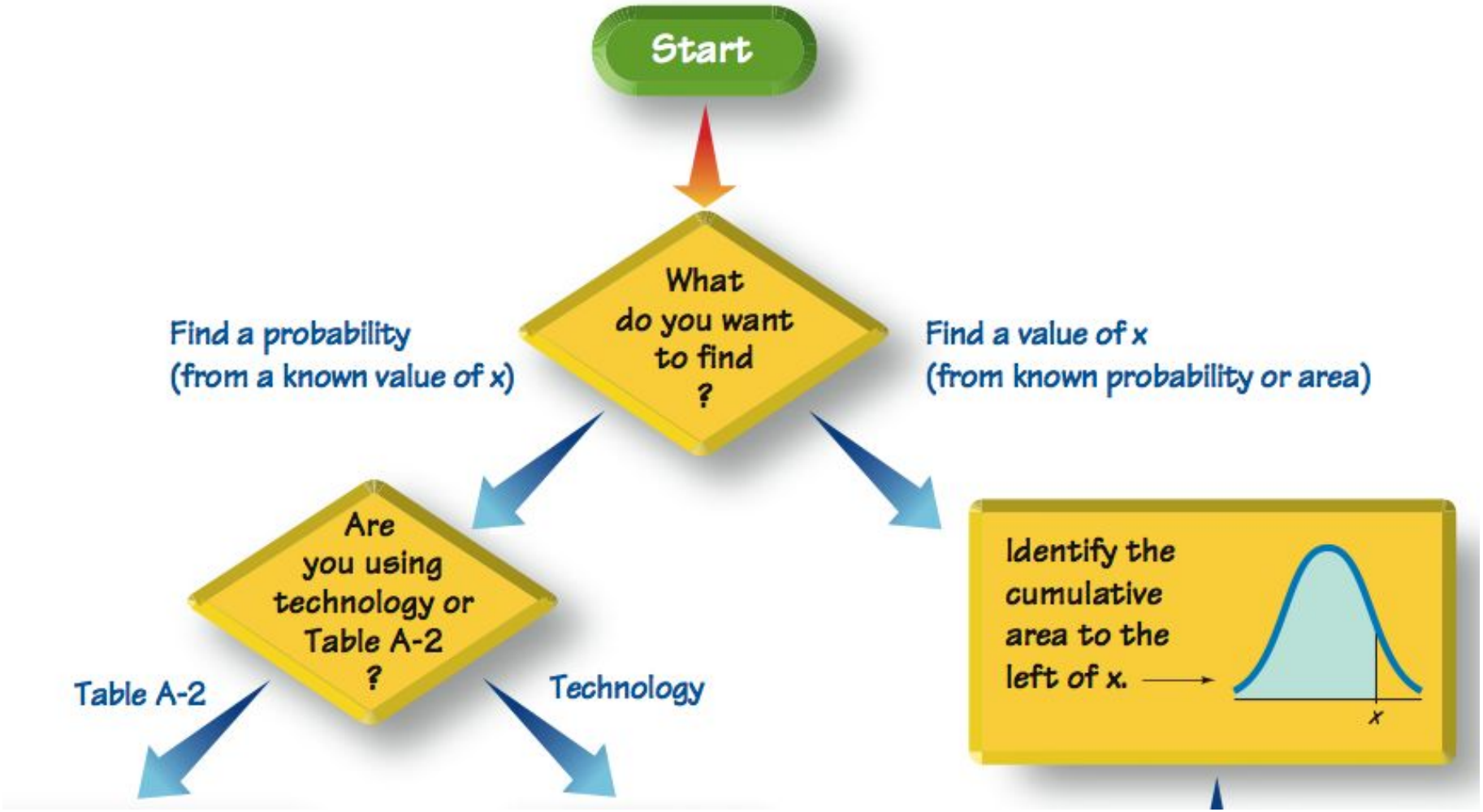
Example – Lightest and Heaviest - cont

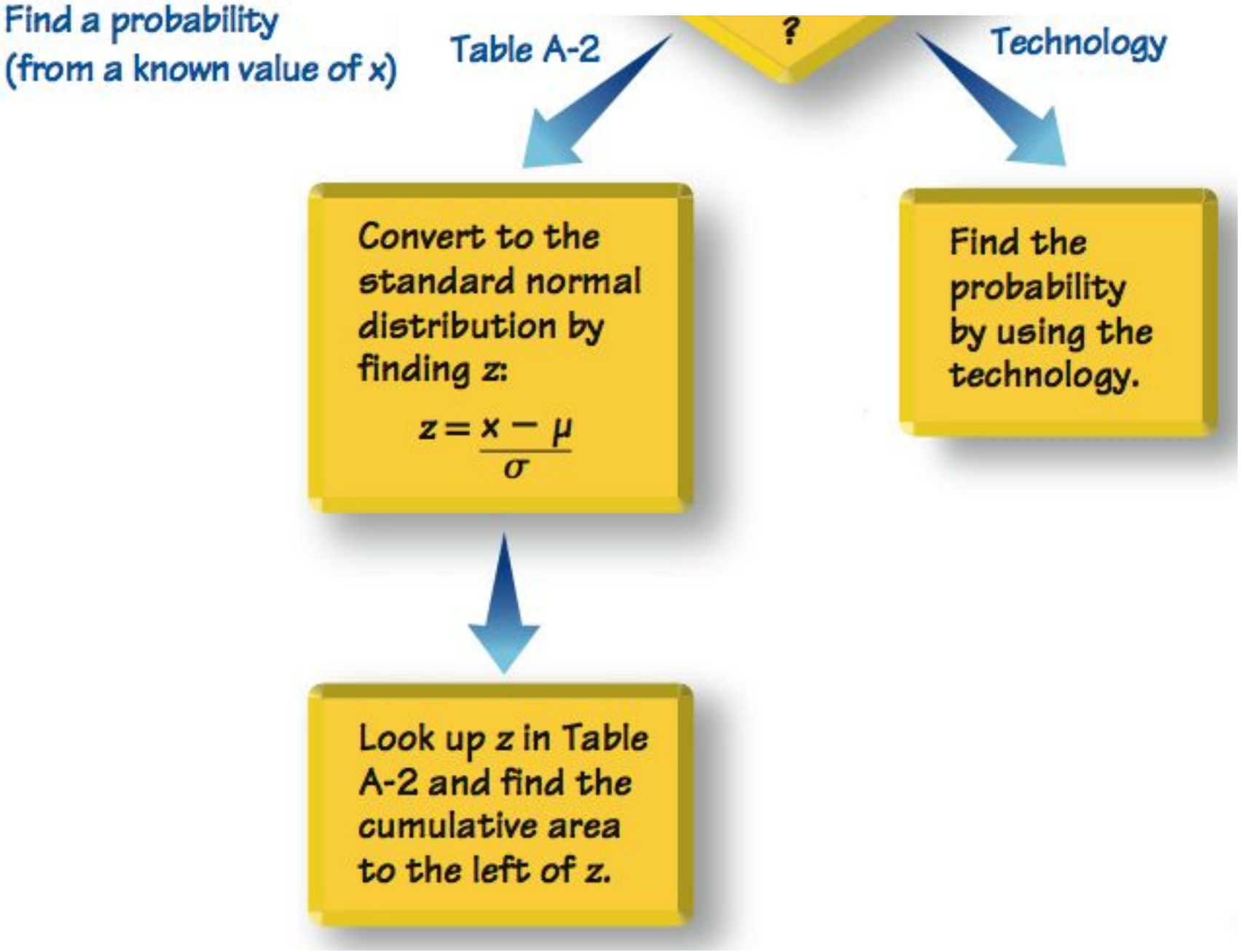
The weight of 247 pounds separates the lightest 99.5% from the heaviest 0.5%

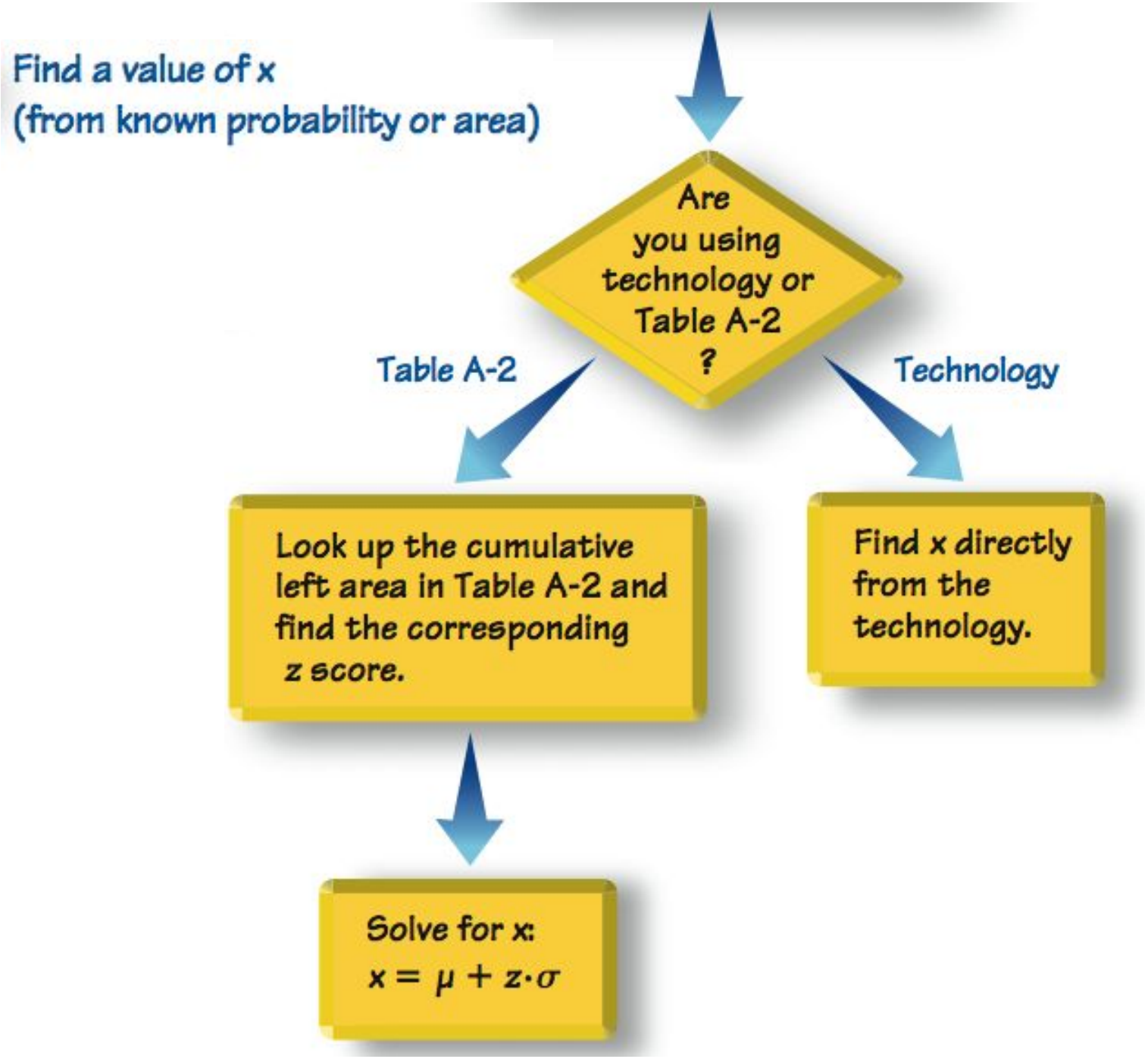



```
> qnorm(0.995, mean=172, sd=29)
[1] 246.699
```

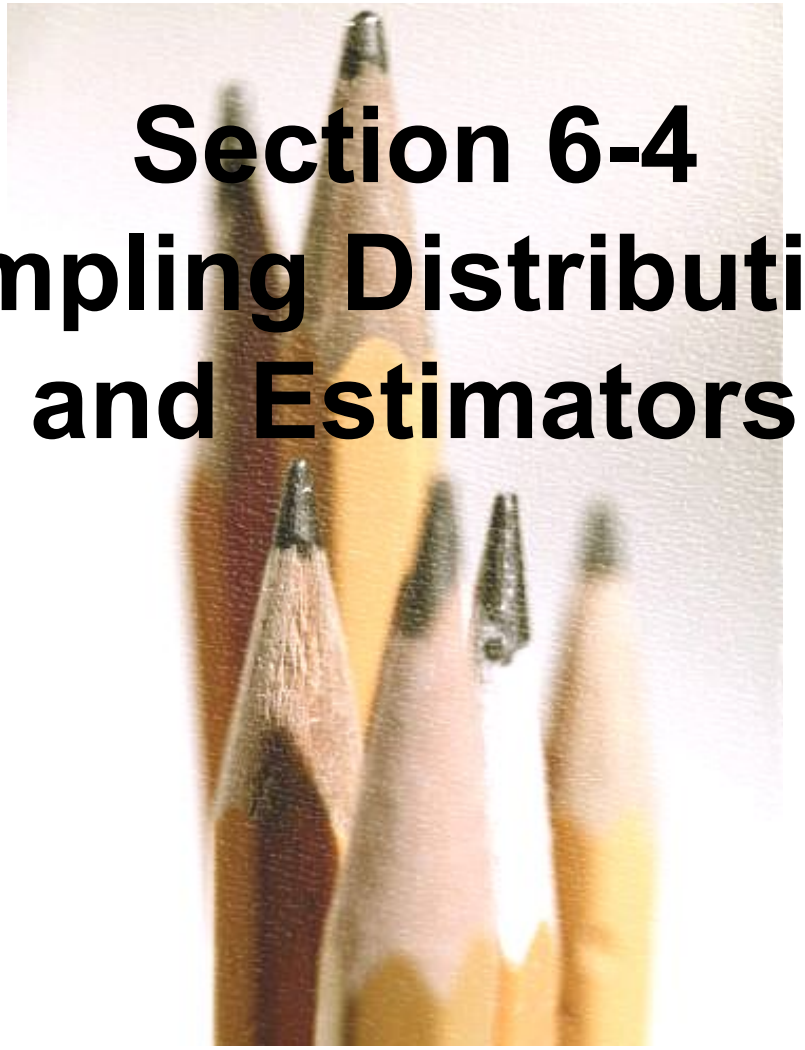
Applications with Normal Distributions







Section 6-4
Sampling Distributions and Estimators



Key Concept

The main objective of this section is to understand the concept of a sampling distribution of a statistic, which is the distribution of all values of that statistic when all possible samples of the same size are taken from the same population.

We will also see that some statistics are better than others for estimating population parameters.

Definition

The sampling distribution of a statistic (such as the sample mean or sample proportion) is the distribution of all values of the statistic when all possible samples of the same size *n* are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Definition

The sampling distribution of the mean is the distribution of sample means, with all samples having the same sample size *n* taken from the same population. (The sampling distribution of the mean is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Properties

- Sample means target the value of the population mean. (That is, the mean of the sample means is the population mean. The expected value of the sample mean is equal to the population mean.)
- The distribution of the sample means tends to be a normal distribution.

Definition

The sampling distribution of the variance is the distribution of sample variances, with all samples having the same sample size *n* taken from the same population. (The sampling distribution of the variance is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Properties

- Sample variances target the value of the population variance. (That is, the mean of the sample variances is the population variance. The expected value of the sample variance is equal to the population variance.)
- The distribution of the sample variances tends to be a distribution skewed to the right.

Definition

The sampling distribution of the proportion is the distribution of sample proportions, with all samples having the same sample size *n* taken from the same population.

Let there be x successes out of n Bernoulli trials. The sample proportion is the fraction of samples which were successes, so

$$\hat{p} = \frac{x}{n}.$$

Definition

We need to distinguish between a population proportion *p* and some sample proportion:

***p* = population proportion**

$\hat{p}$ **= sample proportion**

Properties

- Sample proportions target the value of the population proportion. (That is, the mean of the sample proportions is the population proportion. The expected value of the sample proportion is equal to the population proportion.)
- The distribution of the sample proportion tends to be a normal distribution.

Unbiased Estimators

Sample means, variances and proportions are unbiased estimators.

That is they target the population parameter.

These statistics are better in estimating the population parameter.

Biased Estimators

Sample medians, ranges and standard deviations are biased estimators.

That is they do NOT target the population parameter.

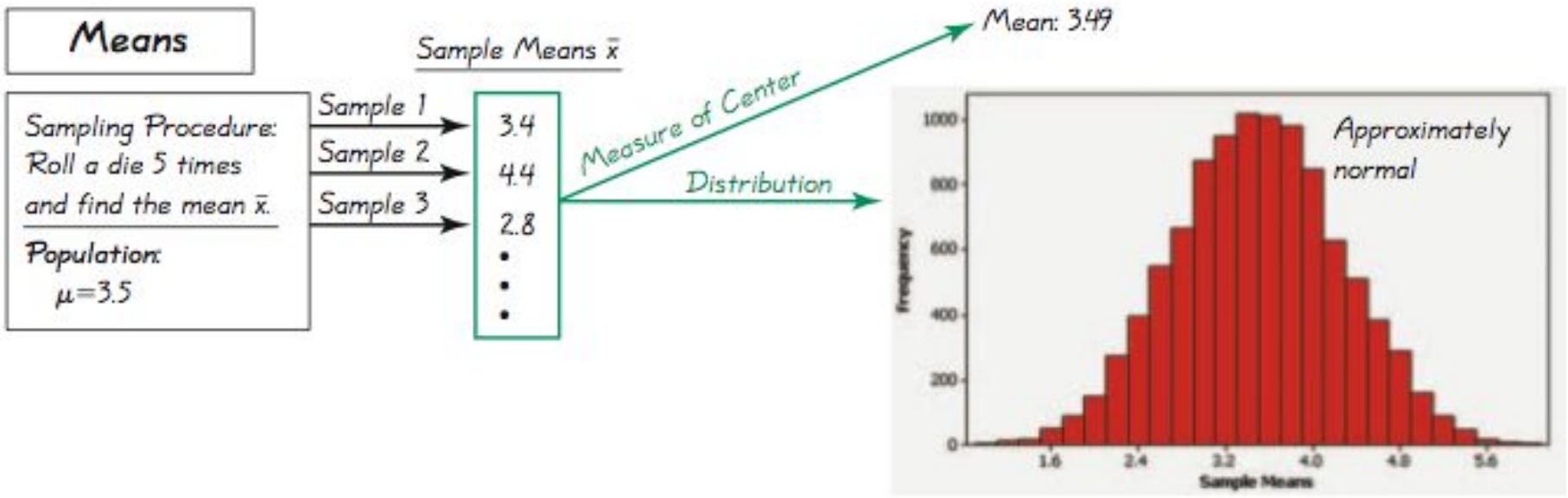
Note: the bias with the standard deviation is relatively small in large samples so *s* is often used to estimate.

Example - Sampling Distributions

Consider repeating this process: Roll a die 5 times, find the mean $\bar{x}$, variance s^2, and the proportion of *odd* numbers of the results. What do we know about the behavior of all sample means that are generated as this process continues indefinitely?

Example - Sampling Distributions

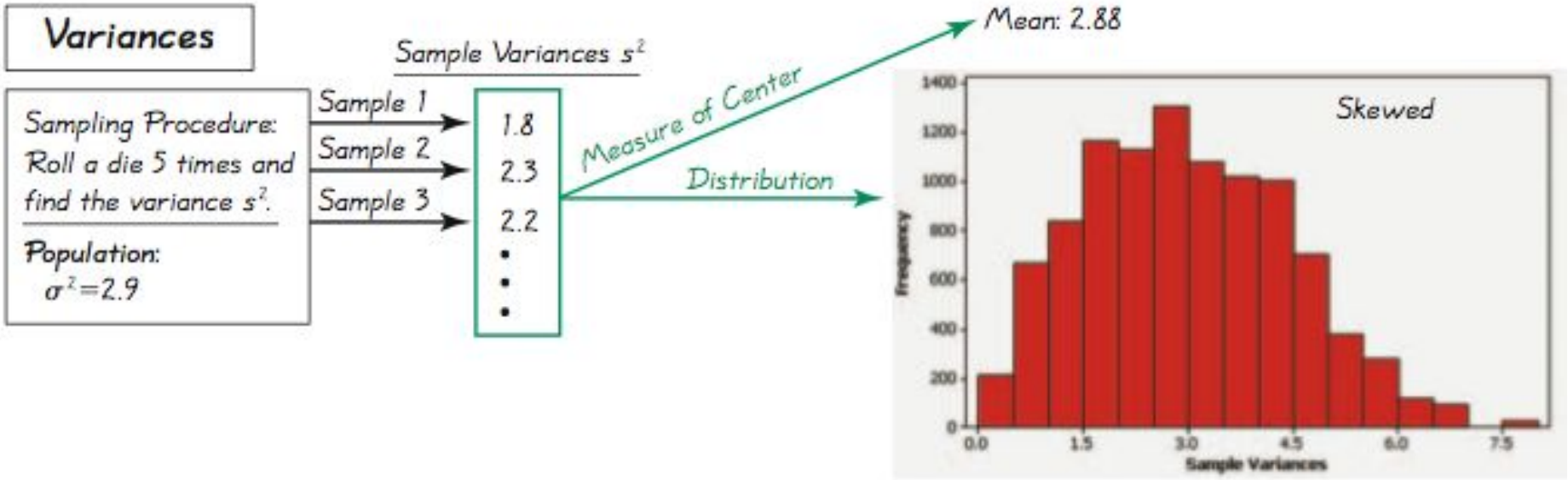
Specific results from 10,000 trials



All outcomes are equally likely so the population mean is 3.5; the mean of the 10,000 trials is 3.49. If continued indefinitely, the sample mean will be 3.5. Also, notice the distribution is "normal."

Example - Sampling Distributions

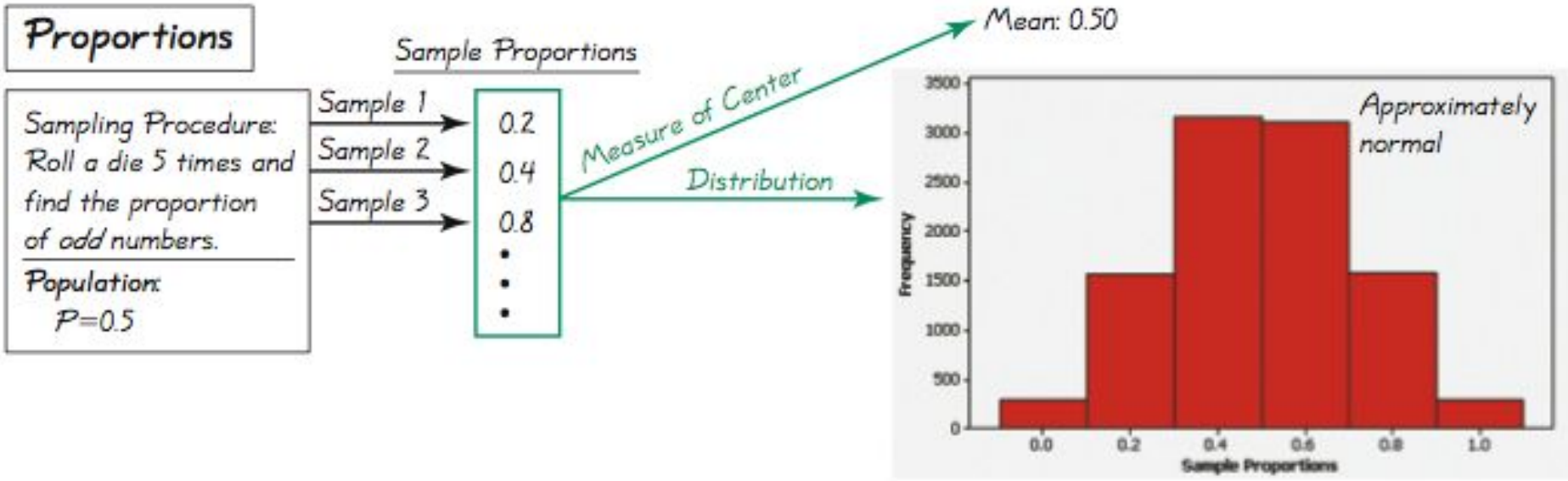
Specific results from 10,000 trials



All outcomes are equally likely so the population variance is 2.9; the mean of the 10,000 trials is 2.88. If continued indefinitely, the sample variance will be 2.9. Also, notice the distribution is "skewed to the right."

Example - Sampling Distributions

Specific results from 10,000 trials



All outcomes are equally likely so the population proportion of odd numbers is 0.50; the proportion of the 10,000 trials is 0.50. If continued indefinitely, the mean of sample proportions will be 0.50. Also, notice the distribution is "approximately normal."

Why Sample with Replacement?

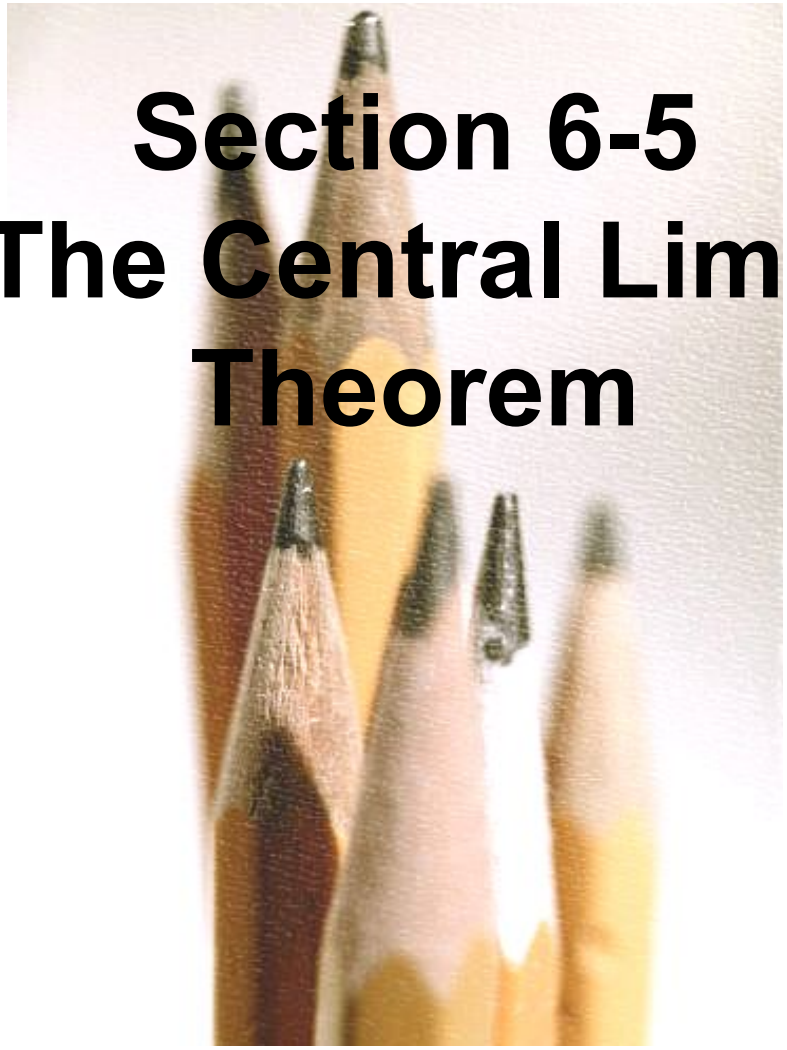
Sampling *without replacement* would have the very practical advantage of avoiding wasteful duplication whenever the same item is selected more than once. However, we are interested in sampling *with replacement* for these two reasons:

1. **When selecting a relatively small sample form a large population, it makes no significant difference whether we sample with replacement or without replacement.**
2. **Sampling with replacement results in independent events that are unaffected by previous outcomes, and independent events are easier to analyze and result in simpler calculations and formulas.**

Caution

Many methods of statistics require a *simple random sample*. Some samples, such as voluntary response samples or convenience samples, could easily result in very wrong results.

Section 6-5
The Central Limit Theorem



Key Concept

The *Central Limit Theorem* tells us that for a population with *any* distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

The procedure in this section form the foundation for estimating population parameters and hypothesis testing.

Central Limit Theorem

Given:

1. The random variable *x* has a distribution (which may or may not be normal) with mean μ and standard deviation σ.
2. Simple random samples all of size *n* are selected from the population. (The samples are selected so that all possible samples of the same size *n* have the same chance of being selected.)

Central Limit Theorem – cont.

Conclusions:

1. The distribution of sample $\bar{x}$ will, as the sample size increases, approach a normal distribution.
2. The mean of the sample means is the population mean μ.
3. The standard deviation of all sample means is $\sigma/\sqrt{n}$.

Practical Rules Commonly Used

1. For samples of size *n* larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets closer to a normal distribution as the sample size *n* becomes larger.
2. If the original population is *normally distributed*, then for any sample size *n*, the sample means will be normally distributed (not just the values of *n* larger than 30).

Notation

the mean of the sample means

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

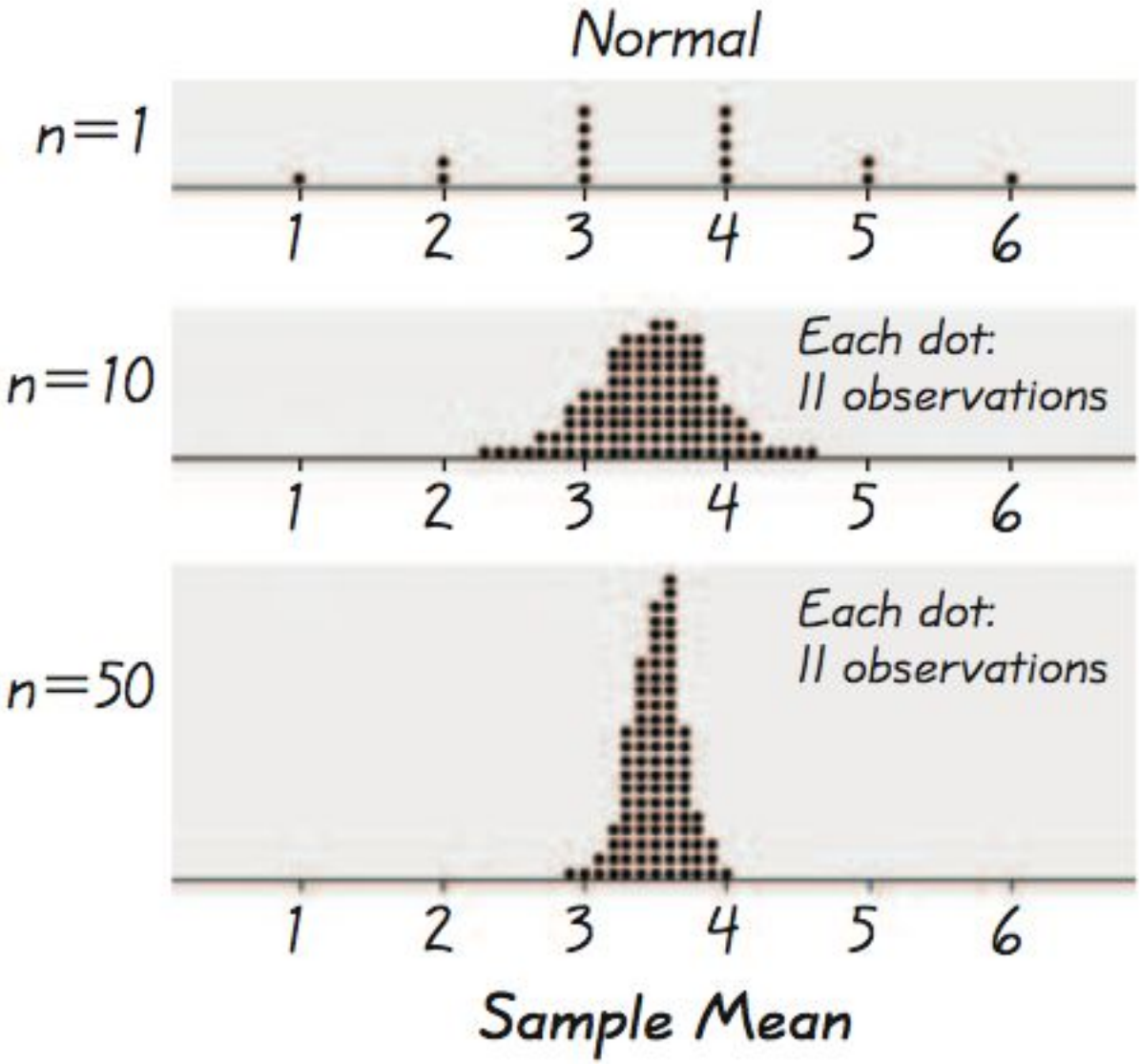
the standard deviation of sample mean

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

(often called the standard error of the mean)

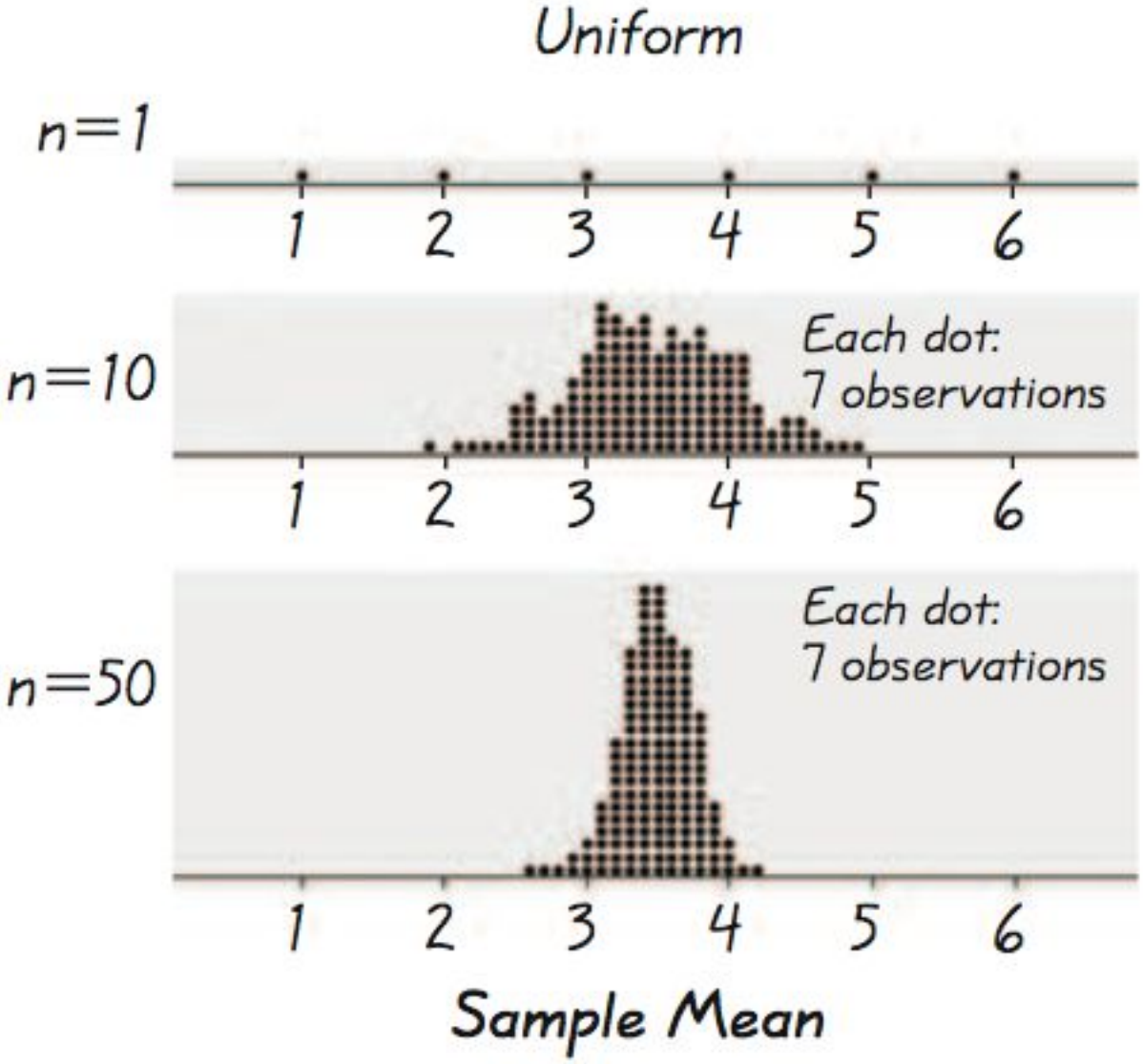
Example - Normal Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



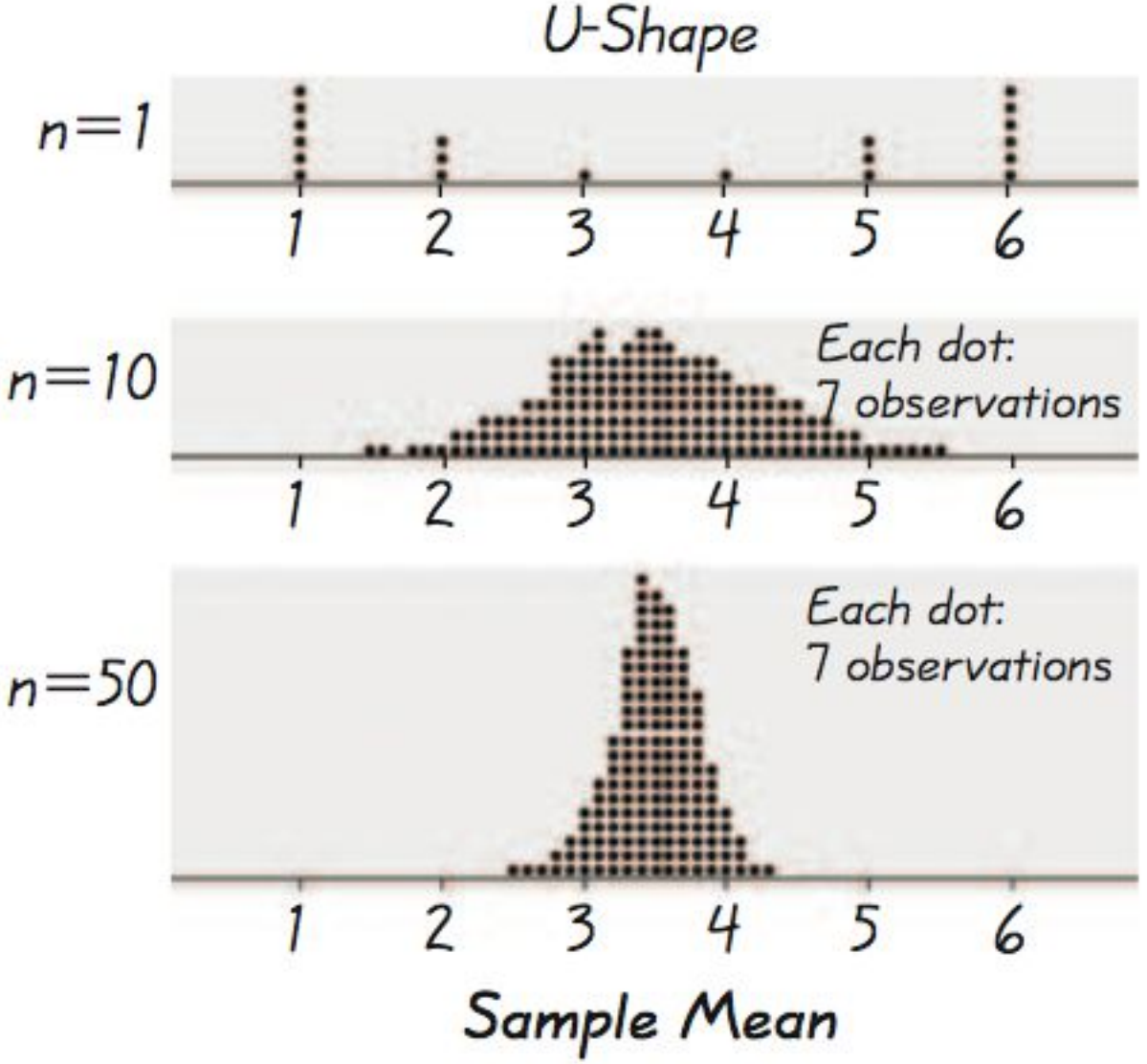
Example - Uniform Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



Example - U-Shaped Distribution

As we proceed from $n = 1$ to $n = 50$, we see that the distribution of sample means is approaching the shape of a normal distribution.



Important Point

As the sample size increases, the sampling distribution of sample means approaches a normal distribution.

Example – Water Taxi Safety

Use the Chapter Problem. Assume the population of weights of men is normally distributed with a mean of 172 lb and a standard deviation of 29 lb.

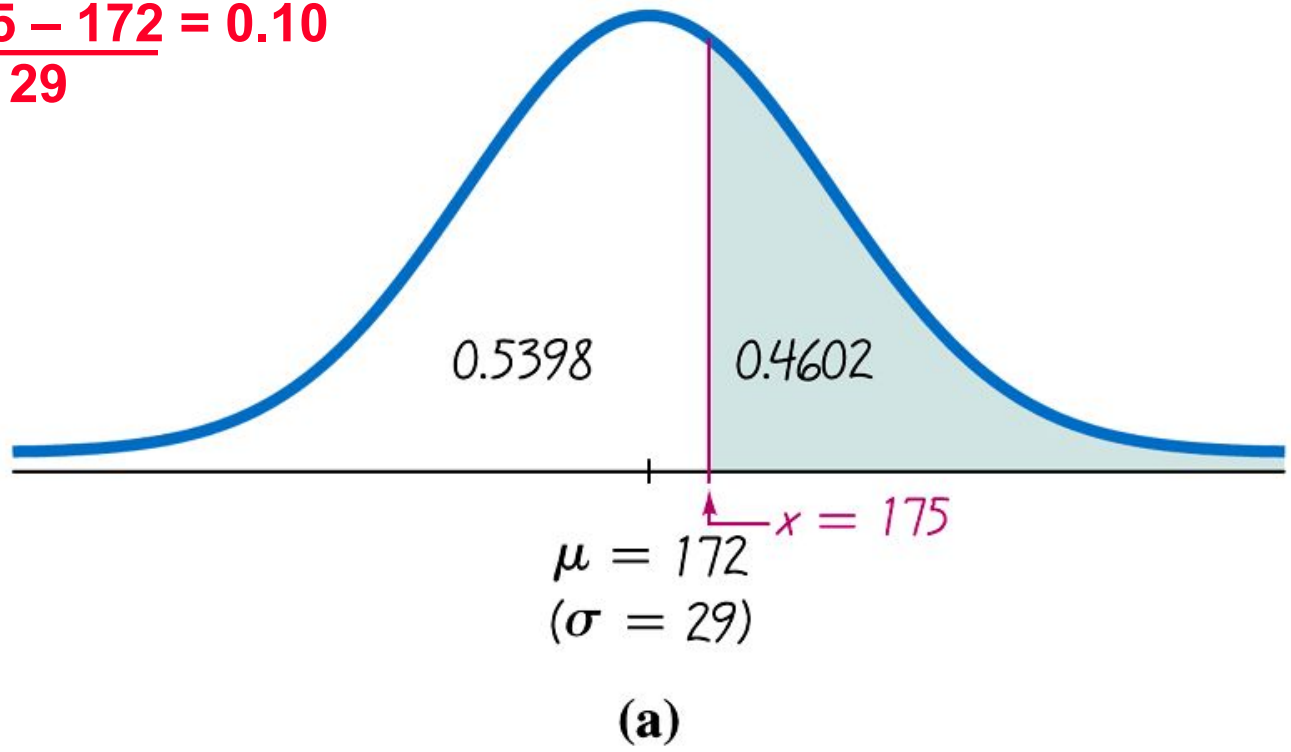
a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

b) b) Find the probability that *20 randomly selected men* will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

Example – cont

a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

$$z = \frac{175 - 172}{29} = 0.10$$

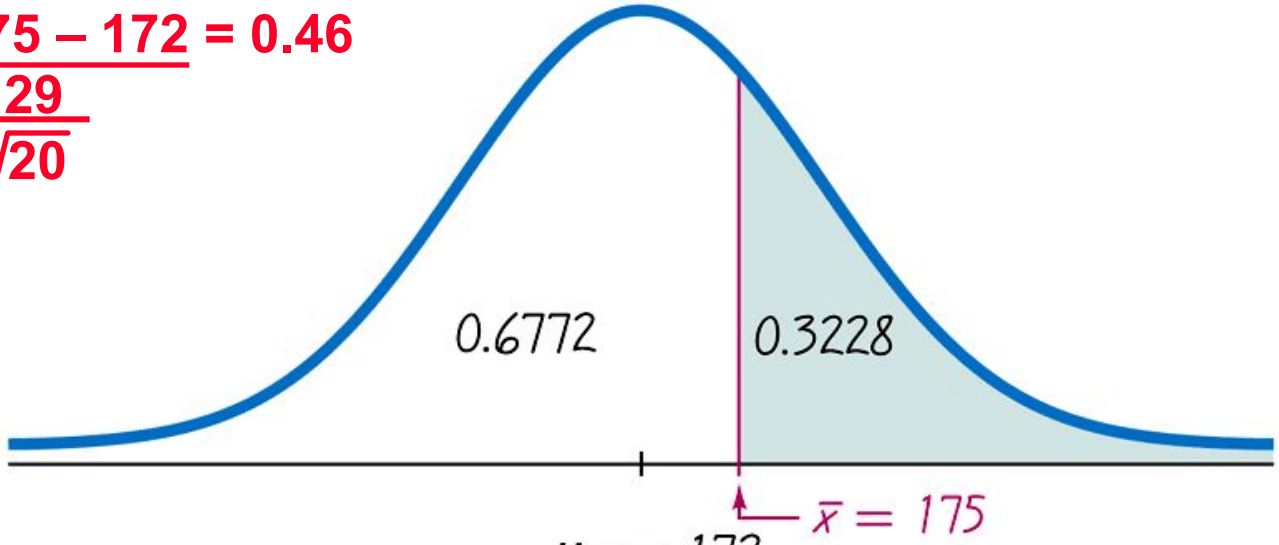


(a)

Example – cont

b) Find the probability that *20 randomly selected men* will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

$$z = \frac{175 - 172}{\frac{29}{\sqrt{20}}} = 0.46$$



(b)

Example - cont

a) Find the probability that if an *individual* man is randomly selected, his weight is greater than 175 lb.

$$P(x > 175) = 0.4602$$

b) Find the probability that *20 randomly selected men* will have a mean weight that is greater than 175 lb (so that their total weight exceeds the safe capacity of 3500 pounds).

$$P(\bar{x} > 175) = 0.3228$$

It is much easier for an individual to deviate from the mean than it is for a group of 20 to deviate from the mean.

Interpretation of Results

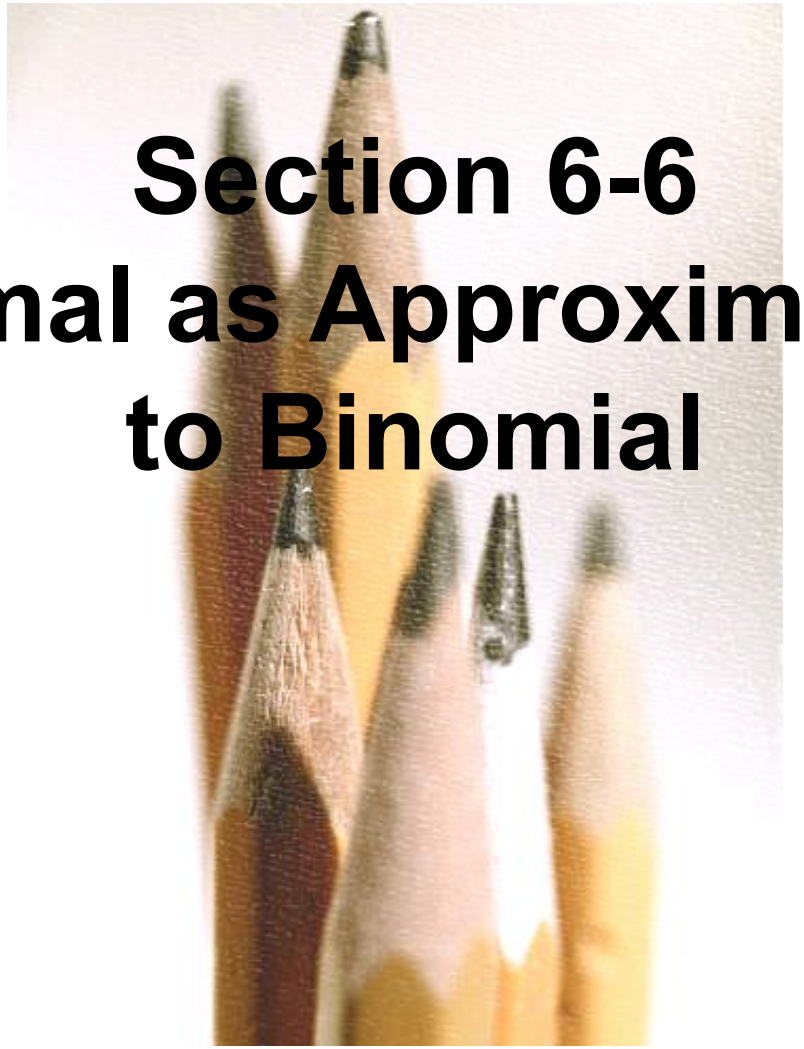
Given that the safe capacity of the water taxi is 3500 pounds, there is a fairly good chance (with probability 0.3228) that it will be overloaded with 20 randomly selected men.

Correction for a Finite Population

When sampling without replacement and the sample size *n* is greater than 5% of the finite population of size *N* (that is, *n* > 0.05*N*), adjust the standard deviation of sample means by multiplying it by the *finite population correction factor*:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

finite population correction factor



Section 6-6
Normal as Approximation to Binomial

Key Concept

This section presents a method for using a normal distribution as an approximation to the binomial probability distribution.

If the conditions of $np \geq 5$ and $nq \geq 5$ are both satisfied, then probabilities from a binomial probability distribution can be approximated well by using a normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$.

Review

Binomial Probability Distribution

1. The procedure must have a fixed number of trials.
2. The trials must be independent.
3. Each trial must have all outcomes classified into two categories (commonly, success and failure).
4. The probability of success remains the same in all trials.

Solve by binomial probability formula, Table A-1, or technology.

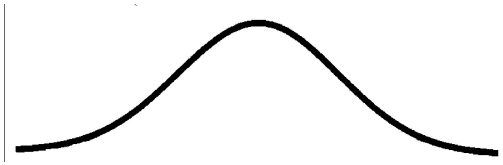
Approximation of a Binomial Distribution with a Normal Distribution

$np \geq 5$

$nq \geq 5$

then $\mu = np$ and $\sigma = \sqrt{npq}$

and the random variable has

a  distribution.

(normal)

Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

1. Verify that both $np \geq 5$ and $nq \geq 5$. If not, you must use software, a calculator, a table or calculations using the binomial probability formula.

2. Find the values of the parameters μ and σ by calculating $\mu = np$ and $\sigma = \sqrt{npq}$.

3. Identify the discrete whole number x that is relevant to the binomial probability problem. Focus on this value temporarily.

Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

4. Draw a normal distribution centered about μ, then draw a *vertical strip area* centered over x. Mark the left side of the strip with the number equal to $x - 0.5$, and mark the right side with the number equal to $x + 0.5$. *Consider the entire area of the entire strip to represent the probability of the discrete whole number itself.*

5. Determine whether the value of x itself is included in the probability. Determine whether you want the probability of at least x, at most x, more than x, fewer than x, or exactly x. Shade the area to the right or left of the strip; also shade the interior of the strip *if and only if x itself* is to be included. This total shaded region corresponds to the probability being sought.

Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

6. Using $x - 0.5$ or $x + 0.5$ in place of x, find the area of the shaded region: find the z score; use that z score to find the area to the left of the adjusted value of x; use that cumulative area to identify the shaded area corresponding to the desired probability.

Example – Number of Men Among Passengers

Finding the Probability of "At Least 122 Men" Among 213 Passengers

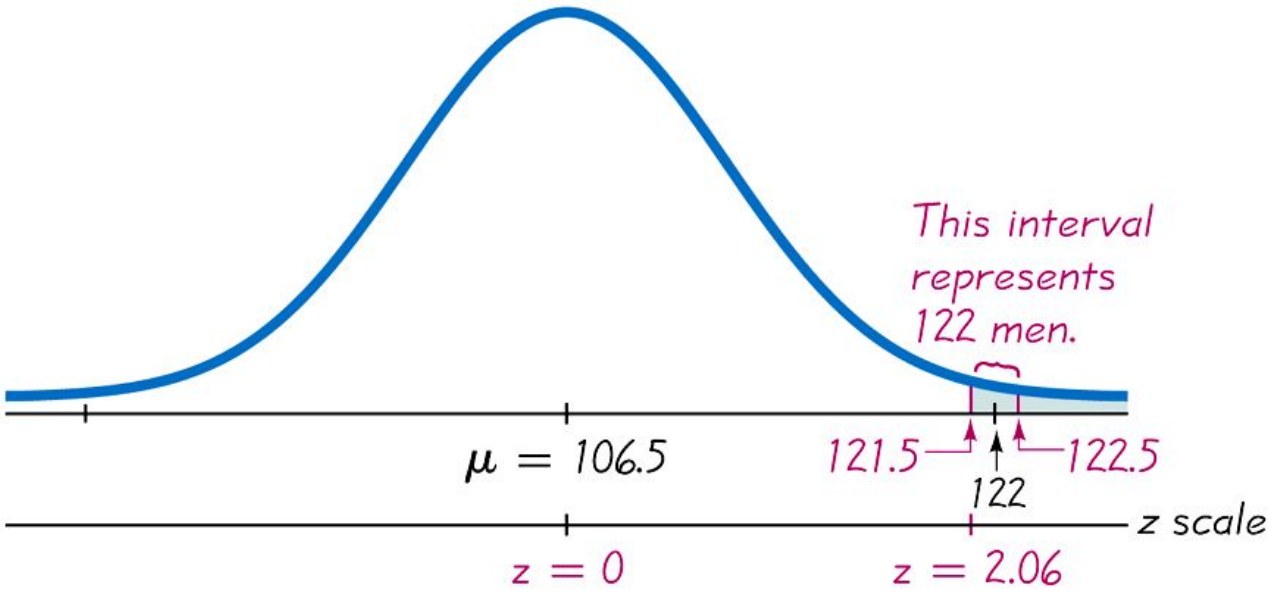


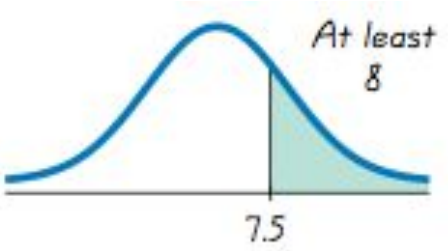
Figure 6-21

Definition

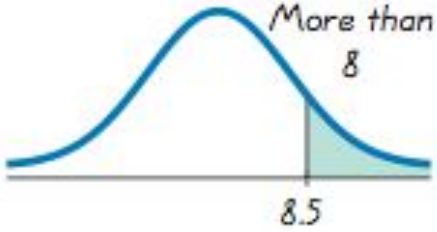
When we use the normal distribution (which is a continuous probability distribution) as an approximation to the binomial distribution (which is discrete), a continuity correction is made to a discrete whole number *x* in the binomial distribution by representing the discrete whole number *x* by the interval from

$$x - 0.5 \text{ to } x + 0.5$$

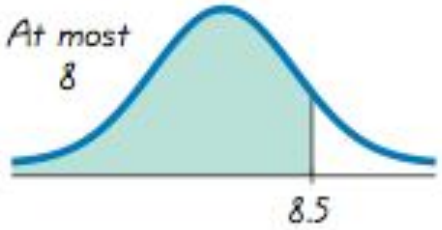
(that is, adding and subtracting 0.5).



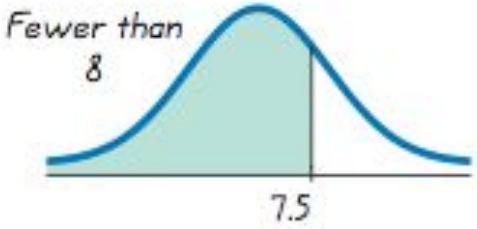
X = **at least** 8
(includes 8 and above)



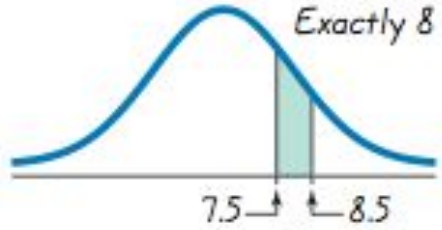
X = **more than** 8
(doesn't include 8)



X = **at most** 8
(includes 8 and below)

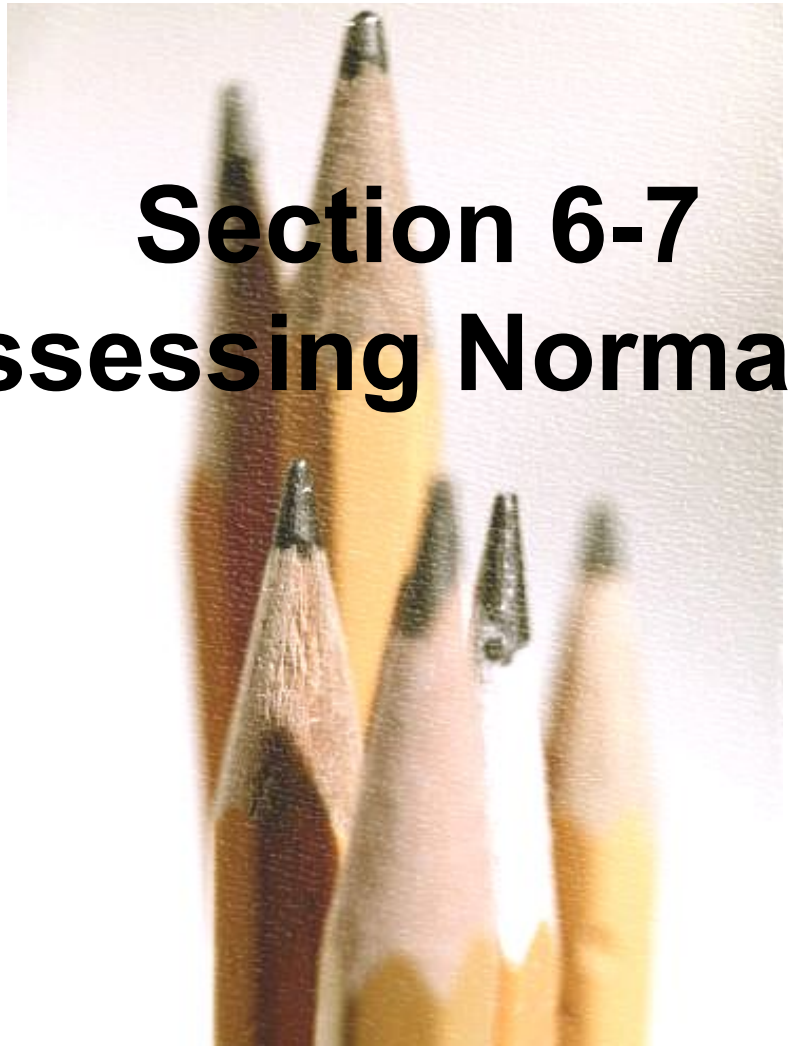


X = **fewer than** 8
(doesn't include 8)



X = **exactly** 8

Section 6-7
Assessing Normality



Key Concept

This section presents criteria for determining whether the requirement of a normal distribution is satisfied.

The criteria involve visual inspection of a histogram to see if it is roughly bell shaped, identifying any outliers, and constructing a graph called a normal quantile plot.

Definition

A normal quantile plot (or normal probability plot) is a graph of points (*x*,*y*), where each *x* value is from the original set of sample data, and each *y* value is the corresponding *z* score that is a quantile value expected from the standard normal distribution.

Procedure for Determining Whether It Is Reasonable to Assume that Sample Data are From a Normally Distributed Population

1. **Histogram:** Construct a histogram. Reject normality if the histogram departs dramatically from a bell shape.

2. **Outliers:** Identify outliers. Reject normality if there is more than one outlier present.

3. **Normal Quantile Plot:** If the histogram is basically symmetric and there is at most one outlier, use technology to generate a **normal quantile plot**.

https://www.calvin.edu/~rpruim/courses/m243/F03/handouts/normquant.pdf

Procedure for Determining Whether It Is Reasonable to Assume that Sample Data are From a Normally Distributed Population

3. Continued

Use the following criteria to determine whether or not the distribution is normal.

Normal Distribution: The population distribution is normal if the pattern of the points is reasonably close to a straight line and the points do not show some systematic pattern that is not a straight-line pattern.

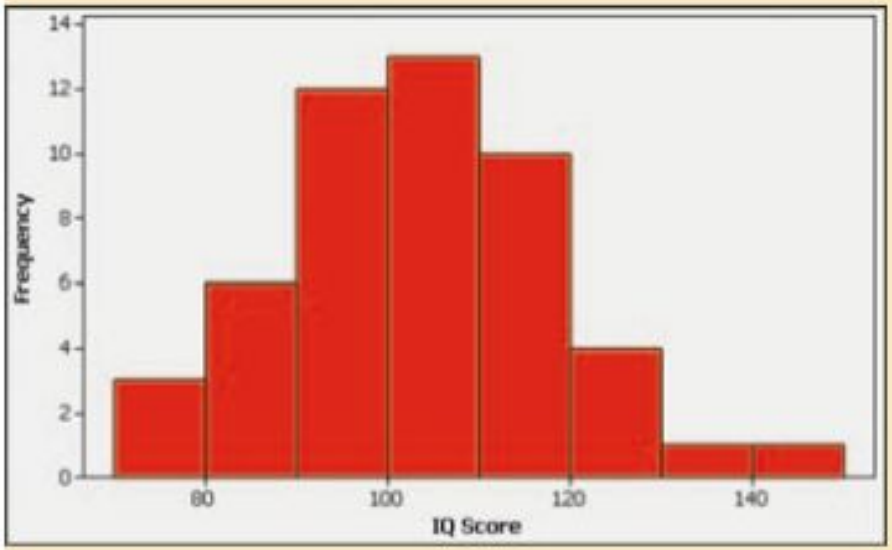
Procedure for Determining Whether It Is Reasonable to Assume that Sample Data are From a Normally Distributed Population

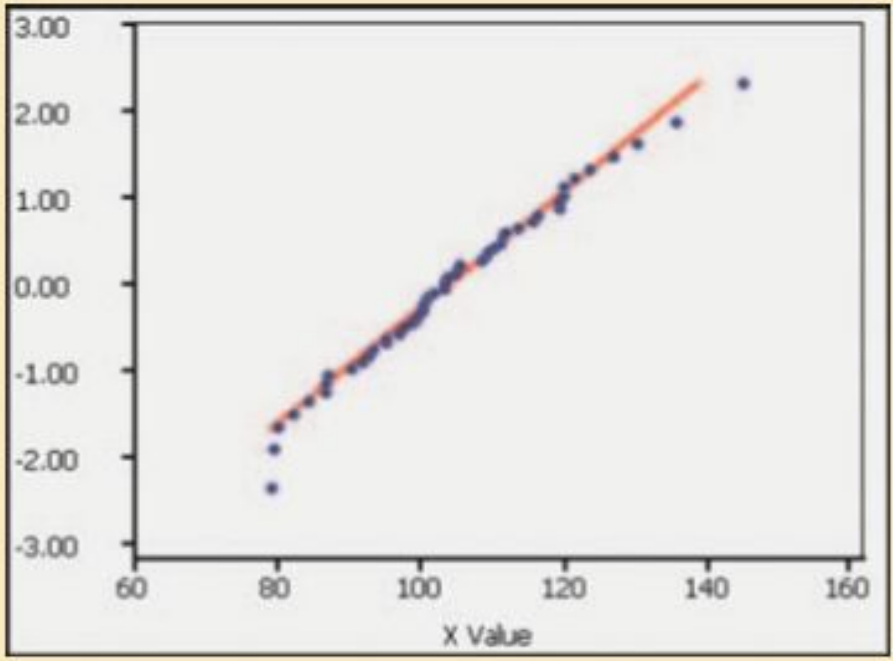
3. Continued

Not a Normal Distribution: **The population distribution is not normal if either or both of these two conditions applies:**

- **The points do not lie reasonably close to a straight line.**
- **The points show some systematic pattern that is not a straight-line pattern.**

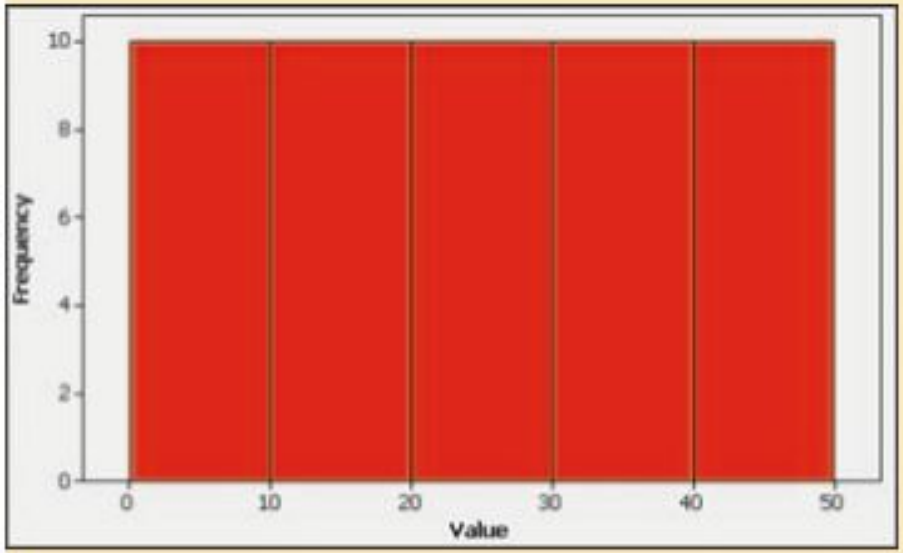
Example

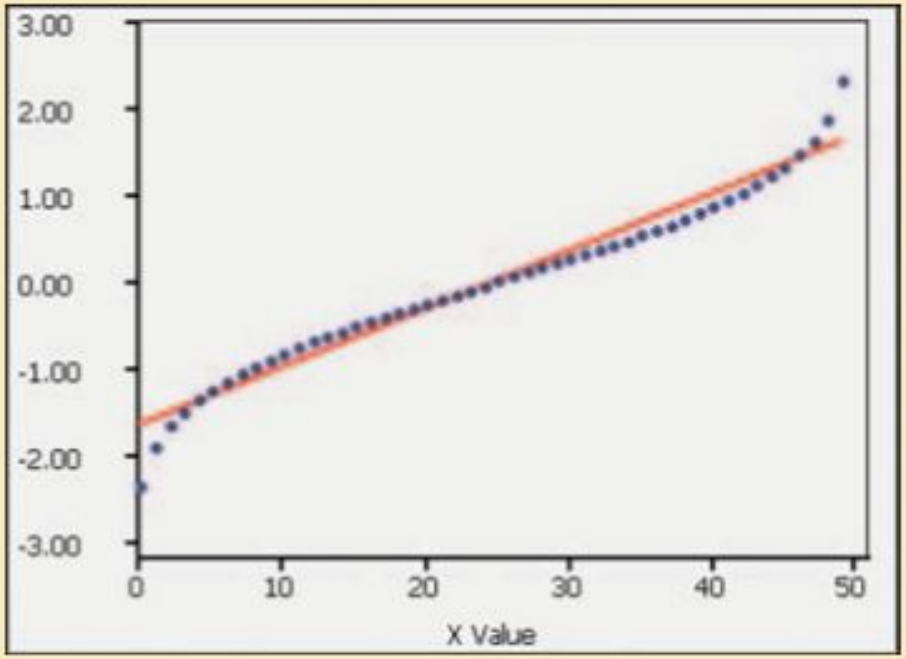




Normal: **Histogram of IQ scores is close to being bell-shaped, suggests that the IQ scores are from a normal distribution. The normal quantile plot shows points that are reasonably close to a straight-line pattern. It is safe to assume that these IQ scores are from a normally distributed population.**

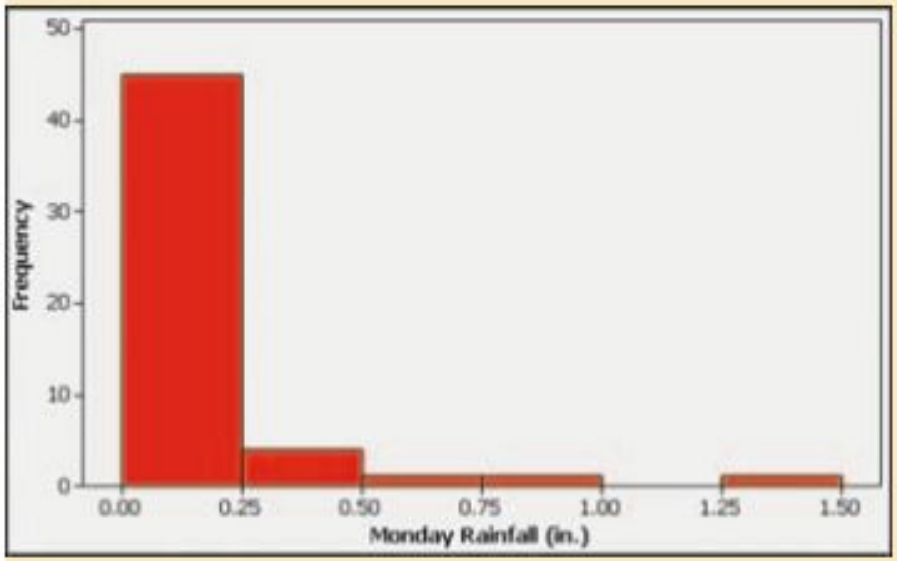
Example

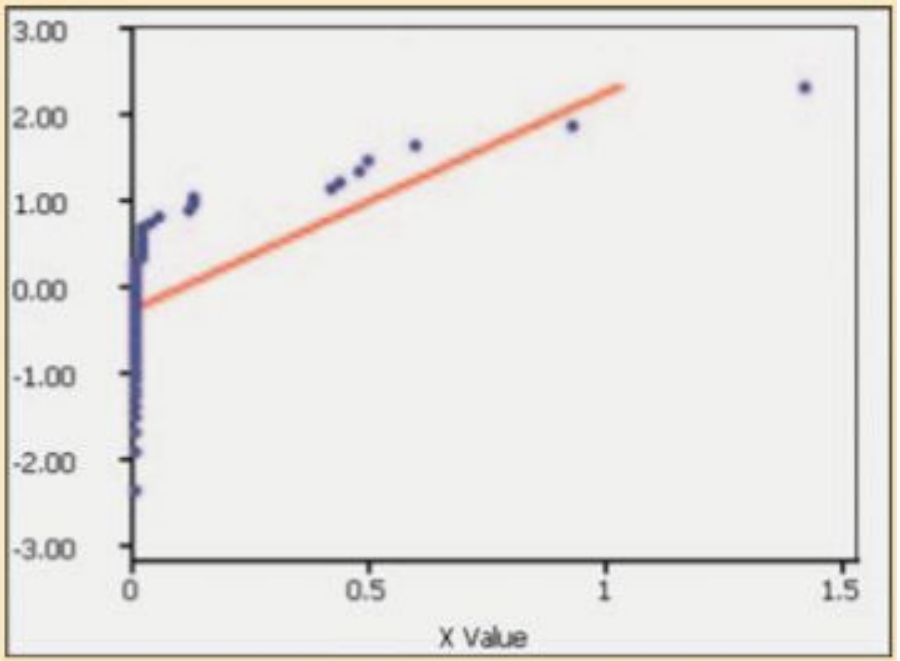




Uniform: **Histogram of data having a uniform distribution. The corresponding normal quantile plot suggests that the points are not normally distributed because the points show a systematic pattern that is not a straight-line pattern. These sample values are not from a population having a normal distribution.**

Example





Skewed: **Histogram of the amounts of rainfall in Boston for every Monday during one year. The shape of the histogram is skewed, not bell-shaped. The corresponding normal quantile plot shows points that are not at all close to a straight-line pattern. These rainfall amounts are not from a population having a normal distribution.**

Manual Construction of a Normal Quantile Plot

Step 1. First sort the data by arranging the values in order from lowest to highest.

Step 2. With a sample of size n, each value represents a proportion of $1/n$ of the sample. Using the known sample size n, identify the areas of $1/2n$, $3/2n$, and so on. These are the cumulative areas to the left of the corresponding sample values.

Step 3. Use the standard normal distribution (Table A-2 or software or a calculator) to find the z scores corresponding to the cumulative left areas found in Step 2. (These are the z scores that are expected from a normally distributed sample.)

Manual Construction of a Normal Quantile Plot

Step 4. Match the original sorted data values with their corresponding *z* scores found in Step 3, then plot the points (*x*, *y*), where each *x* is an original sample value and *y* is the corresponding *z* score.

Step 5. Examine the normal quantile plot and determine whether or not the distribution is normal.

Ryan-Joiner Test

The Ryan-Joiner test is one of several formal tests of normality, each having their own advantages and disadvantages. STATDISK has a feature of Normality Assessment that displays a histogram, normal quantile plot, the number of potential outliers, and results from the Ryan-Joiner test. Information about the Ryan-Joiner test is readily available on the Internet.

Data Transformations

Many data sets have a distribution that is not normal, but we can transform the data so that the modified values have a normal distribution. One common transformation is to replace each value of x with log (x + 1). If the distribution of the log (x + 1) values is a normal distribution, the distribution of the x values is referred to as a lognormal distribution.

Other Data Transformations

In addition to replacing each x value with the log (x + 1), there are other transformations, such as replacing each x value with $\sqrt{x}$, or 1/x, or x^2. In addition to getting a required normal distribution when the original data values are not normally distributed, such transformations can be used to correct other deficiencies, such as a requirement (found in later chapters) that different data sets have the same variance.

Recap

In this section we have discussed:

- **Normal quantile plot.**
- **Procedure to determine if data have a normal distribution.**