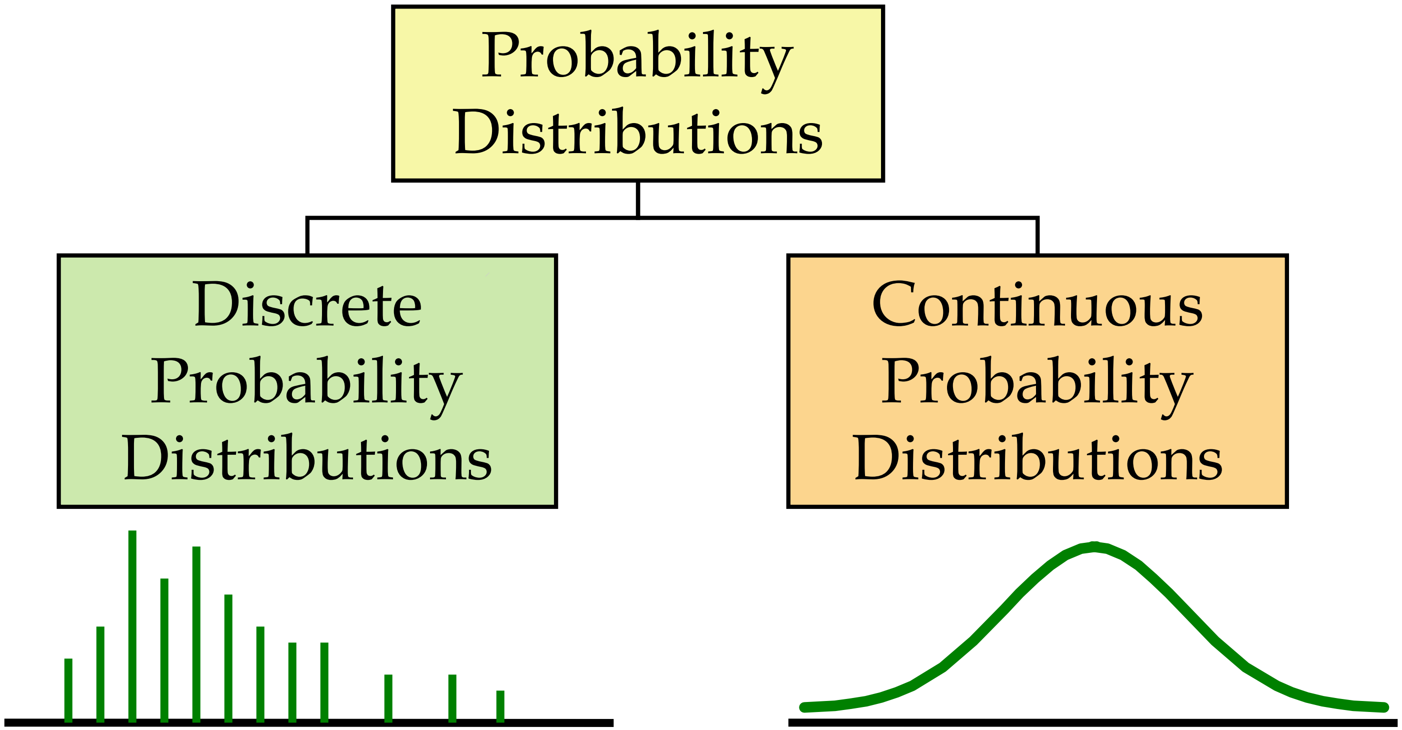
Continuous Probability Distributions

- Continuous Random Variables
- Normal Probability Distribution
- Uniform Probability Distribution
- Exponential Probability Distribution

Won't discuss

- Reading: Chapter 6 (Sections 6.1-6.2)

Probability Distributions



Continuous Probability Distributions

- A ***continuous random variable*** can assume any value in an interval on the real line or in a collection of intervals
- It is not possible to talk about the probability of the random variable assuming a particular value
  - Because there is an infinite number of possible values, the probability of one specific value occurring is theoretically equal to zero

$$P(x = x_0) = 0$$

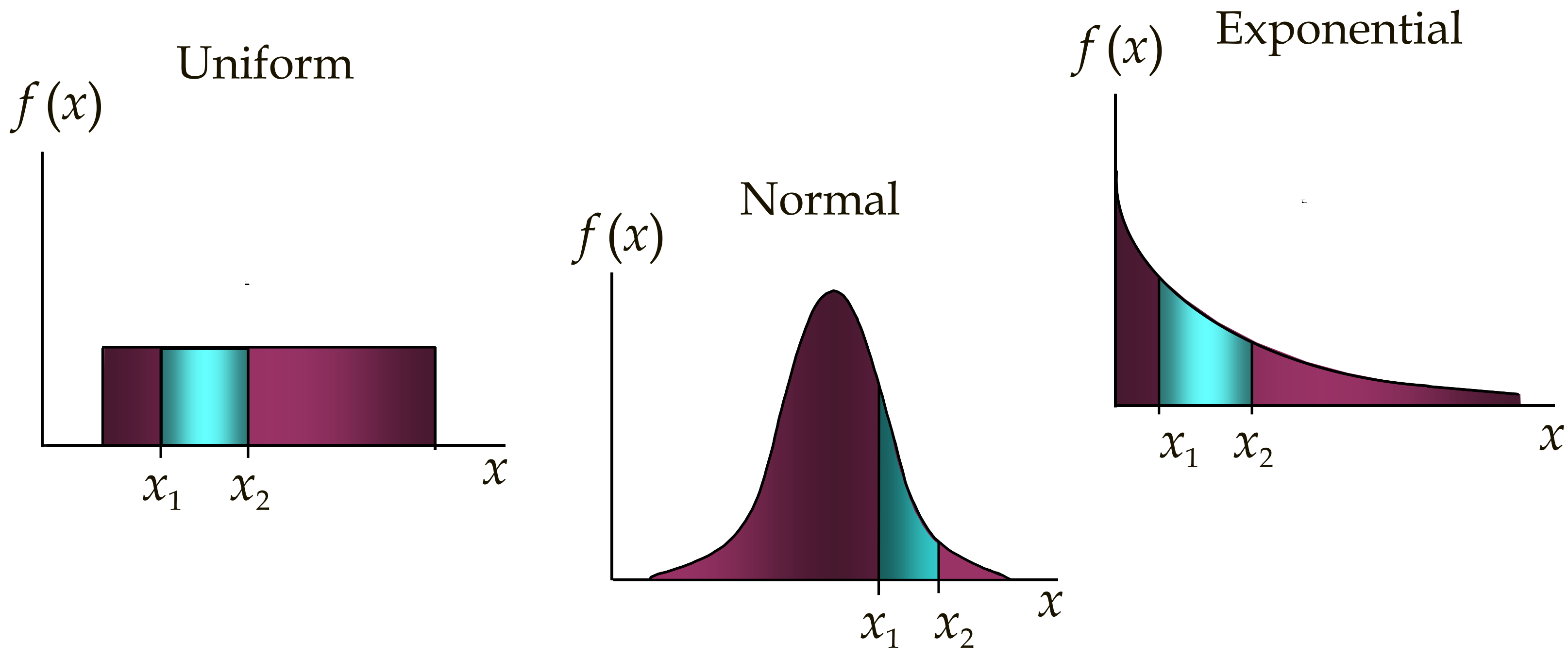
- Instead, we talk about the probability of the random variable assuming a value within a given interval
  - $P(x > x_1), P(x \geq x_1), P(x < x_2), P(x \leq x_2)$
  - $P\left(x_1 \begin{Bmatrix} < \\ \leq \end{Bmatrix} x \begin{Bmatrix} < \\ \leq \end{Bmatrix} x_2\right)$

Probability Density Function

- **Probability density function** $f(x)$ (people also call it pdf) is used to specify the probability of the random variable falling *within a particular range of values*, as opposed to taking on any one value
  - I.e. $f(x)$ is used to calculate probabilities BUT
  - the value of $f(x)$ is not a probability per se
- The probability that x takes on a value between some lower value x_1 and some higher value x_2 can be found by computing the integral of the probability density function $f(x)$ over the interval from x_1 to x_2
- Graphically, it is equivalent to computing the area under the graph of $f(x)$ over the interval from x_1 to x_2

Probability Density Function

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the <u>area under the graph</u> of the <u>probability density function</u> $f(x)$ between x_1 and x_2



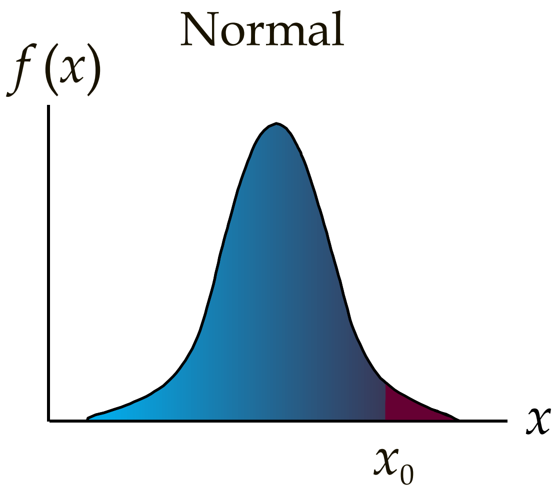
Note: All area under the graph is equal to 1

Cumulative Probability

Cumulative probability function $F(x)$ is used to calculate cumulative probability at any point x_0 or probability that random variable x takes a value less than or equal to x_0

$$F(x_0) = P(x \leq x_0)$$

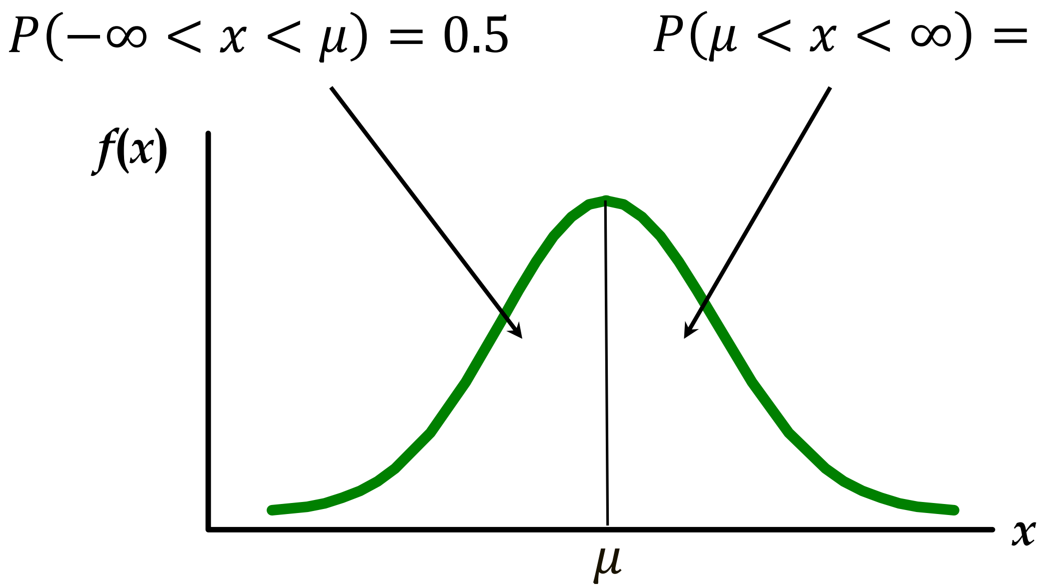
Graphically, cumulative probability is the area under the graph of $f(x)$ on the left of x_0:



Some Useful Notes

- The probability of a random variable falling within a particular range of values can always be expressed in terms of cumulative probabilities
  - E.g., $P(x > x_1)$ can be expressed as a function of $P(x \leq x_1)$
  - It is useful because most often a software allows compute only cumulative probabilities and values of the probability density function $f(x)$
- For the continuous random variable, $P(x \leq x_1) = P(x < x_1)$
  - $P(x \leq x_1) = P(x = x_1) + P(x < x_1) = 0 + P(x < x_1)$
  - Same holds for other intervals with inclusive boundaries

Normal Probability Distribution



$$P(-\infty < x < \infty) = 1$$

Normal Probability Distribution

1. Bell-shaped
2. Symmetric
3. The entire family of normal probability distributions is defined by its mean μ and standard deviation σ
4. The highest point on the normal curve is at the mean, which is also the median and the mode
5. The mean can be any numerical value
6. The standard deviation determines the width of the curve: larger values result in wider and flatter curves
7. Probabilities for the normal random variable are given by areas under the curve. The total area under the curve is 1 (0.5 to the left of the mean and 0.5 to the right)
8. Values near the mean, where the curve is the tallest, have a higher likelihood of occurring than values far from the mean, where the curve is shorter

Normal Probability Density Function

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

where: μ = mean

σ = standard deviation

π = 3.14159

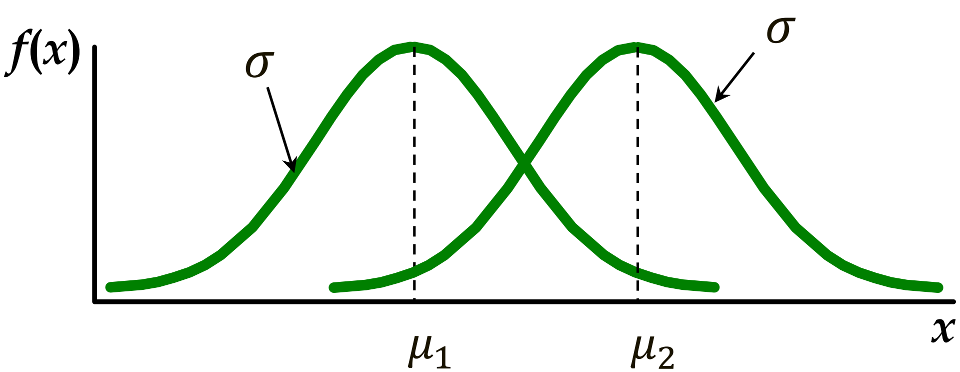
e = 2.71828

Normal Probability Distributions

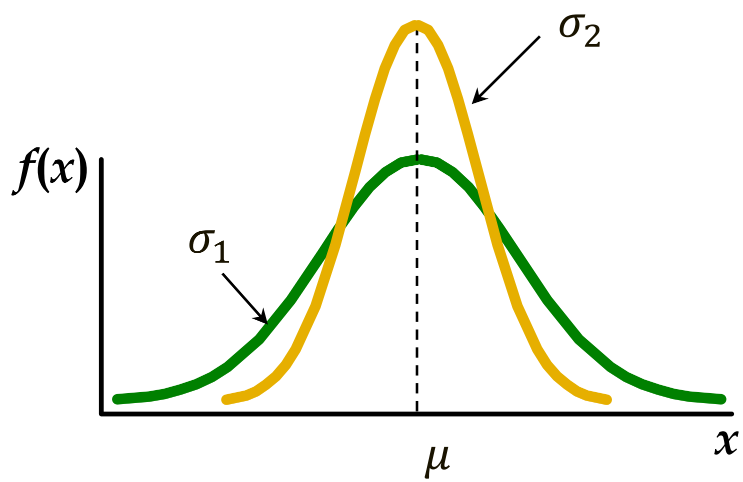
- A distribution's mean (μ) and standard deviation (σ) completely describe its shape

Changing μ shifts the distribution left or right

Changing σ increases or decreases the spread



$$\mu_2 > \mu_1$$



$$\sigma_2 < \sigma_1$$

Standard Normal Probability Distribution

Characteristics

A random variable having a normal distribution with a *mean of* 0 *and a standard deviation of* 1 is said to have a **standard normal probability distribution**

The letter z is designated for the standard normal random variable

Relationship with Normal Distribution

Any normal distribution (with any mean and standard deviation) can be transformed into the **standard normal distribution**

- Need to transform all values of x into z-score

Converting to Standard Normal Distribution

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

where:

x = data value of interest

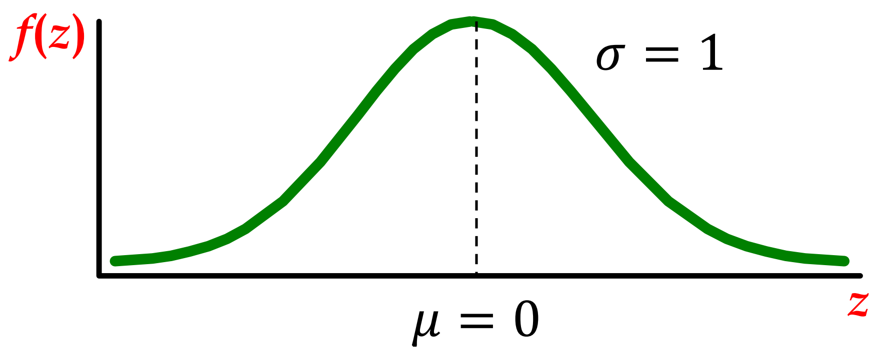
μ = the distribution's mean

σ = the distribution's standard deviation

Recall: We can think of z as a measure of the number of standard deviations x is from μ

The Standard Normal Distribution

When the original random variable, x, follows the normal distribution → z-scores also follow a normal distribution with $\boldsymbol{\mu = 0}$ **and** $\boldsymbol{\sigma = 1}$ or, in other words, follow the standard normal distribution



Example

The time customers spend on the phone for service follows the normal distribution with a mean of 12 minutes and a standard deviation of 3 minutes.

What is the probability that the next customer who calls will spend 14 minutes **or more** on the phone?

Solution Steps:

1. Determine what's given
2. Determine what probability you need to find
3. Find z-scores for the numerical boundaries of the interval of interest and rewrite probability in terms of z-score
4. **Express probability as a function of cumulative probability**
5. Calculate the answer

Example

Known: $\mu = 12$ and $\sigma = 3$

Question: $P(x > 14)$?

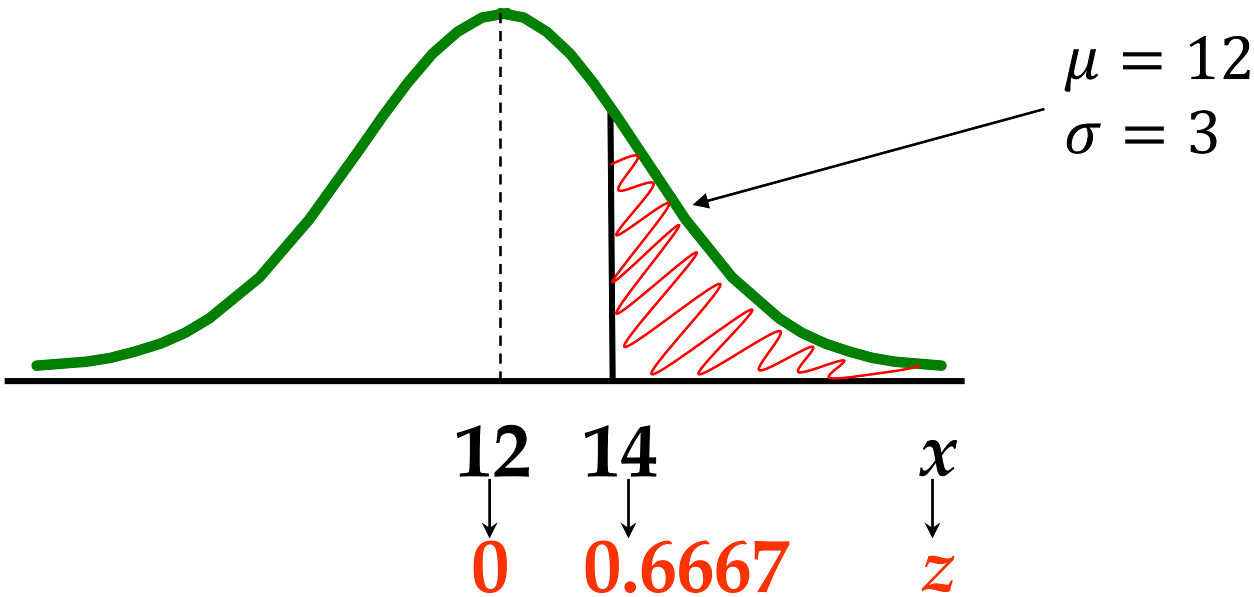
z-transformation: Find the *z*-score for $x = 14$:

$$z = \frac{x - \mu}{\sigma} = \frac{14 - 12}{3} = 0.6667$$

This says that $x = 14$ is 0.6667 standard deviations above the mean of 12

Note: the step of calculating *z*-scores for the boundary can be skipped if you have Excel to calculate probabilities

Example



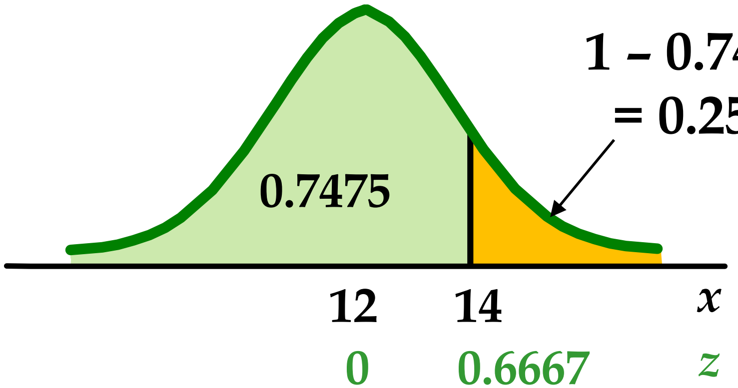
- Note that the shape of distribution is the same, only the scale has changed
- Instead of original units (x), we expressed the problem in standardized units (z)

Example

Upper tail probabilities

The area under the normal curve equals 1.0, so

$$
\begin{aligned}
P(x > 14) &= P(z > 0.6667) \\
&= 1 - P(z \le 0.6667) \\
&= 1 - 0.7475 \\
&= 0.2525
\end{aligned}
$$



Note: Analytically,

$P(A) + P(A') = 1$

$\rightarrow P(A) = 1 - P(A')$

$\rightarrow P(z > 0.6667) = 1 - P(z \le 0.6667)$

Example

Lower tail probabilities

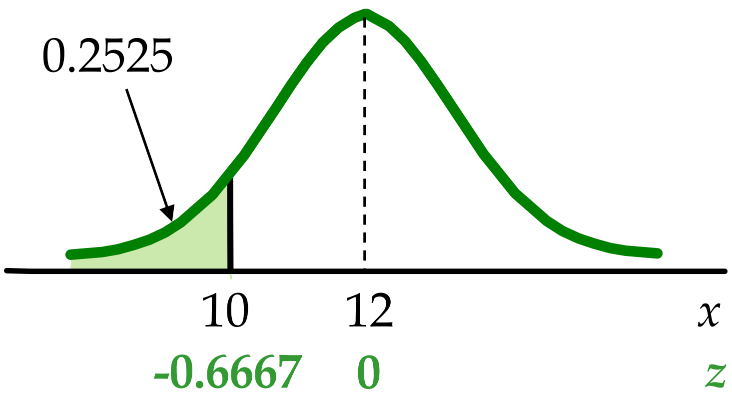
Let's go back to the original example and ask:

What is the probability that the next customer who calls will spend 10 minutes **or less** on the phone?

$$P(x \leq 10) = P\left(z \leq \frac{10-12}{3}\right) =$$

$$= P(z \leq -0.6667)$$

$$= 0.2525$$



Note: the distribution is symmetric around 0. Therefore, if we know $P(z \geq 0.6667)$ or area on the right of 0.6667, we also know $P(z \leq -0.6667)$ or area on the left of -0.6667

Finding the z or x value

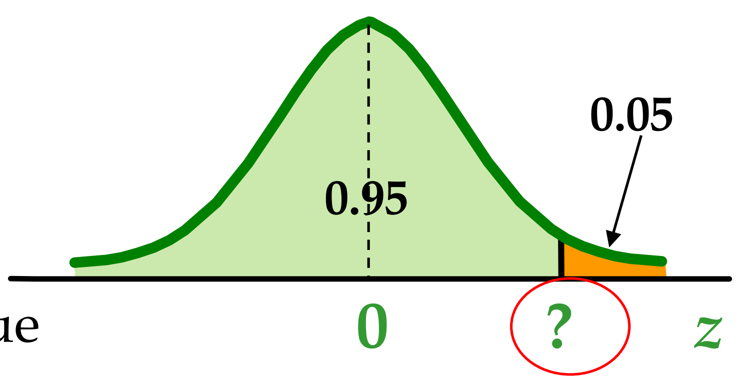
A reverse exercise: Finding z or x value

In our example, the time on the phone follows the normal distribution with $\mu = 12$ and $\sigma = 3$. What is the wait time so that 95% of calls have a shorter wait time?

Find x_0 value so that $P(x \leq x_0) = 0.95$?

1. Find the necessary z-score: the z value needed to include 95% of the area under the curve to the left of the z-score

Note 1: This step requires Excel.
In this example, $z = 1.645$.
Note 2: With Excel, this and the next steps may be combined:
x can be found without z value



Finding z or x value

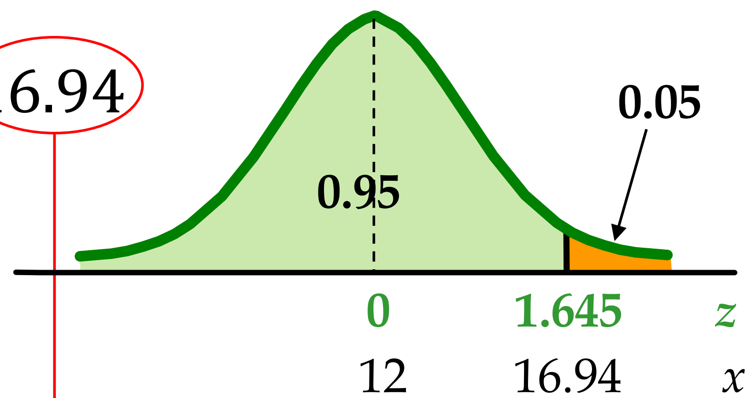
A reverse exercise: Finding z or x value

2. Find x value that is 1.645 (or z) standard deviations above the mean:

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

$$x = \mu + z\sigma$$

$$x = 12 + 1.645 \times 3 = 16.94$$



In words, 95% of calls have a wait time less than 16.94 min

Other Normal Probability Intervals

Probability between two values

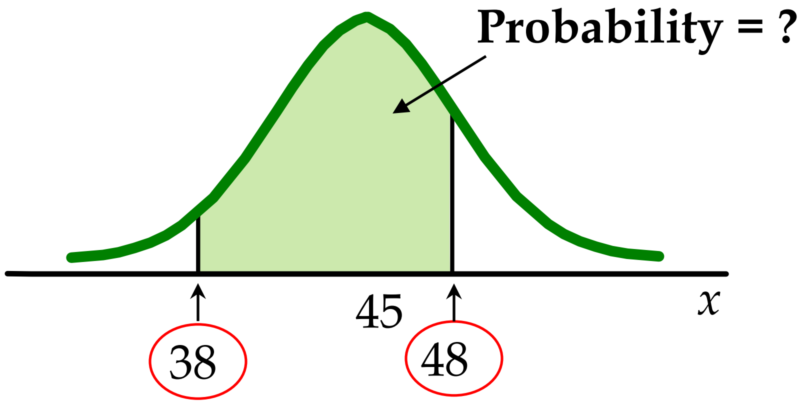
Suppose income is normally distributed for a group of workers, with $\mu = \$45{,}000$ and $\sigma = \$5{,}000$.

Find the probability that a randomly selected worker from this group has an income between \$38,000 and \$48,000.

Known: $\mu = 45$ and $\sigma = 5$

Question: $P(38 \le x \le 48)$?

(Convert all values to 1000's for simplicity)



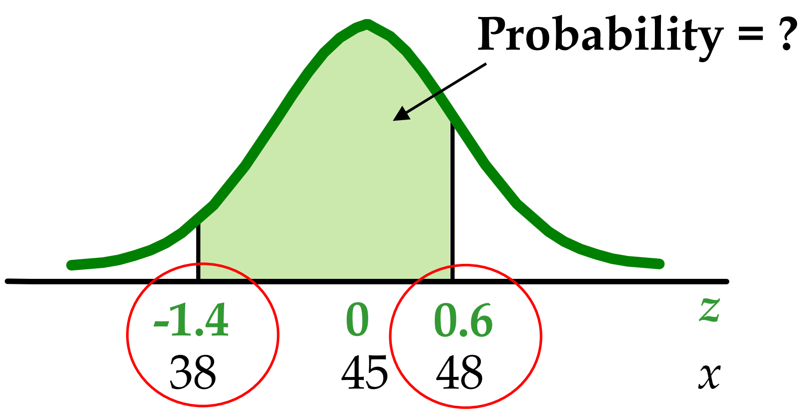
Other Normal Probability Intervals

$P(38 \le x \le 48)$?

Convert x = 38 and x = 48 to z-scores:

$$z_{38} = \frac{x - \mu}{\sigma} = \frac{38 - 45}{5} = -1.4$$

$$z_{48} = \frac{x - \mu}{\sigma} = \frac{48 - 45}{5} = 0.6$$



Other Normal Probability Intervals

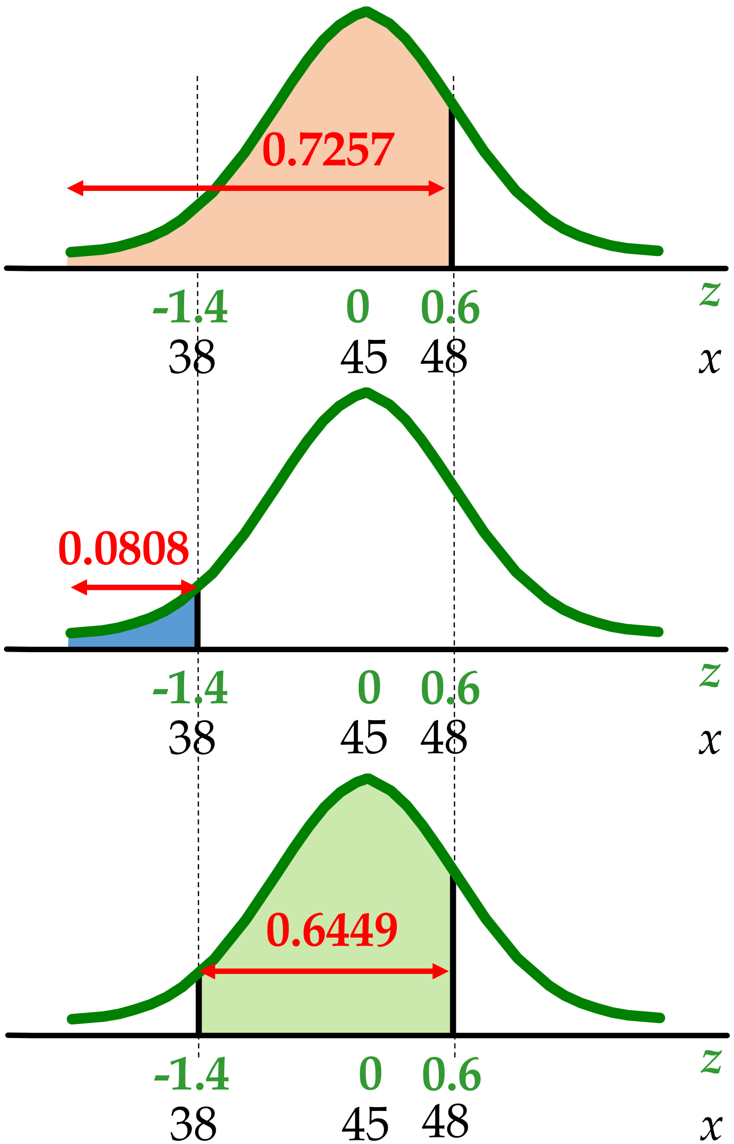
$P(38 \leq x \leq 48)$

$= P(-1.4 \leq z \leq 0.6)$

$= P(z \leq 0.6) - P(z \leq -1.4)$

$= 0.7257 – 0.0808$

$= 0.6449$



Can you confirm analytically that $P(x_1 \leq x \leq x_2) = P(x \leq x_2) - P(x \leq x_1)$?
Hint: express $x_1 \leq x \leq x_2$ as a union or intersection of two simpler events. Then, use your knowledge on the probability of a union/intersection of two events

Excel Exercise >>

Binomial Distribution Approximation

The normal distribution can be used as an approximation to the binomial distribution

- The normal distribution approximation can be used when the sample size is large enough so that

$$np \geq 5 \text{ and } nq \geq 5$$

- When using the normal approximation to the binomial, for the normal distribution we set

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

Binomial Distribution Approximation

Example: Suppose that 15% of people are left-handed. What is the probability of finding exactly 9 left-handed people in a random sample of 50?

We know how to find the binomial probability using Excel. It is equal P(x = 9, n = 50) = 0.1230.

To approximate this probability with the normal distribution, use the mean and the SD of the binomial distribution:

$$\mu = np = 50 \times 0.15 = 7.5$$

$$\sigma = \sqrt{npq} = \sqrt{50 \times 0.15 \times 0.85} = 2.525$$

Binomial Distribution Approximation

Approximation steps (using normal distribution):

- Even though we calculate $P(x = 9)$, we need an **interval** instead of a single value
  - $P(x = 9) = 0$ for the normal (any continuous) distribution
- The interval is found using the **continuity correction factor**:
  - Apply continuity correction factor to make the interval **bigger**
    1. Subtract 0.5 from the left side of the boundary
    2. Add 0.5 to the right side of the boundary
    3. No correction factor is needed for the boundaries $-\infty$ and $+\infty$
- In our example, the interval is from $9-0.5 = 8.5$ to $9+0.5 = 9.5$
- Find $P(8.5 \leq x \leq 9.5)$ **using the normal distribution**

Binomial Distribution Approximation

$P(8.5 \leq x \leq 9.5)$

Boundaries adjusted using the continuity correction factor

$$= P\left(\frac{8.5-7.5}{2.525} \leq z \leq \frac{9.5-7.5}{2.525}\right)$$

Mean of the binomial distribution

SD of the binomial distribution

$$= P(z \leq 0.7921) - P(z \leq 0.3960)$$

$$= 0.7859 – 0.6540$$

$$= 0.1319$$

$\Rightarrow$ Approximated probability is 0.1319 whereas the exact binomial probability is 0.1230

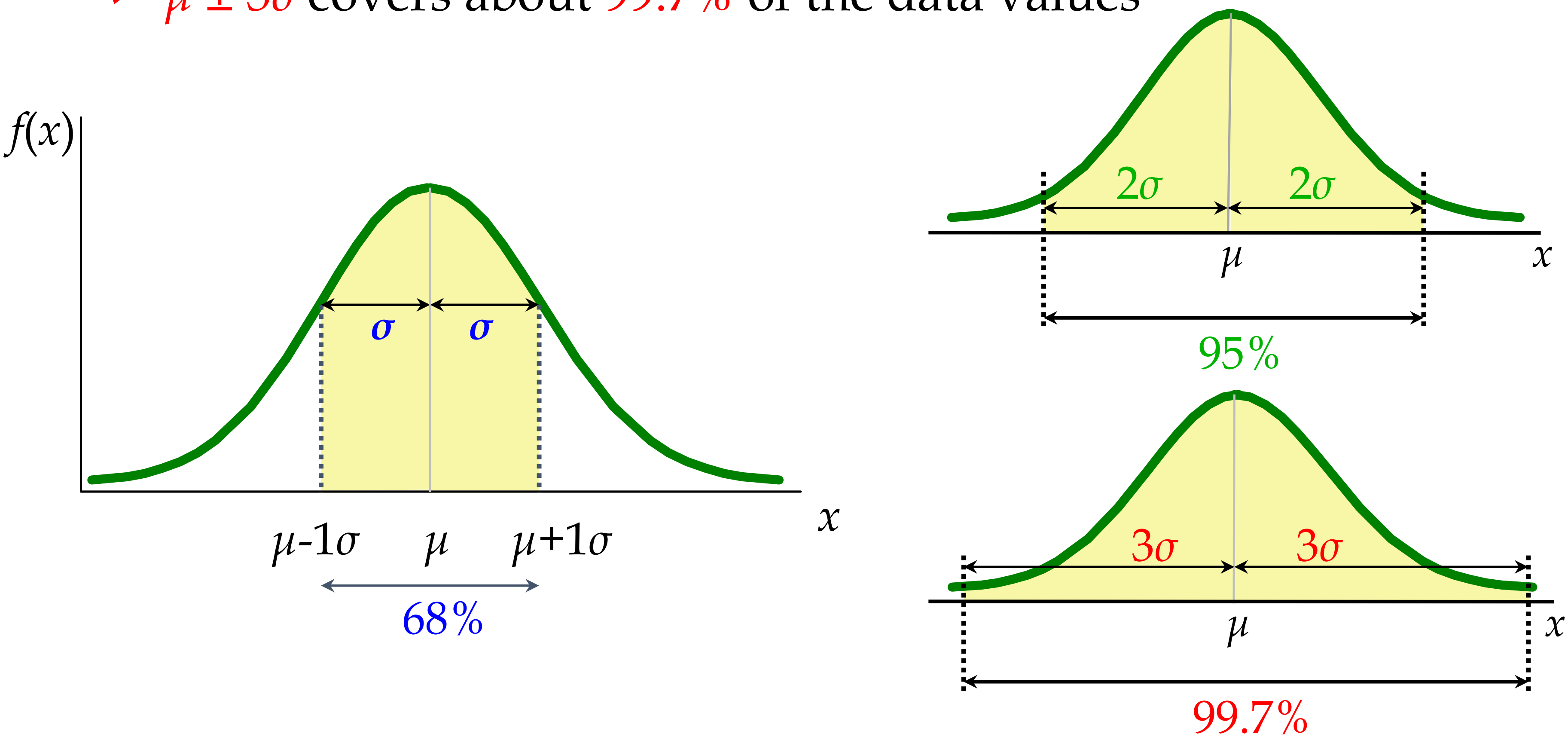
Note: approximation results depend on whether $np \geq 5$ and $nq \geq 5$ are satisfied

Excel Exercise >>

Revisiting the Empirical Rule

For a normal distribution:

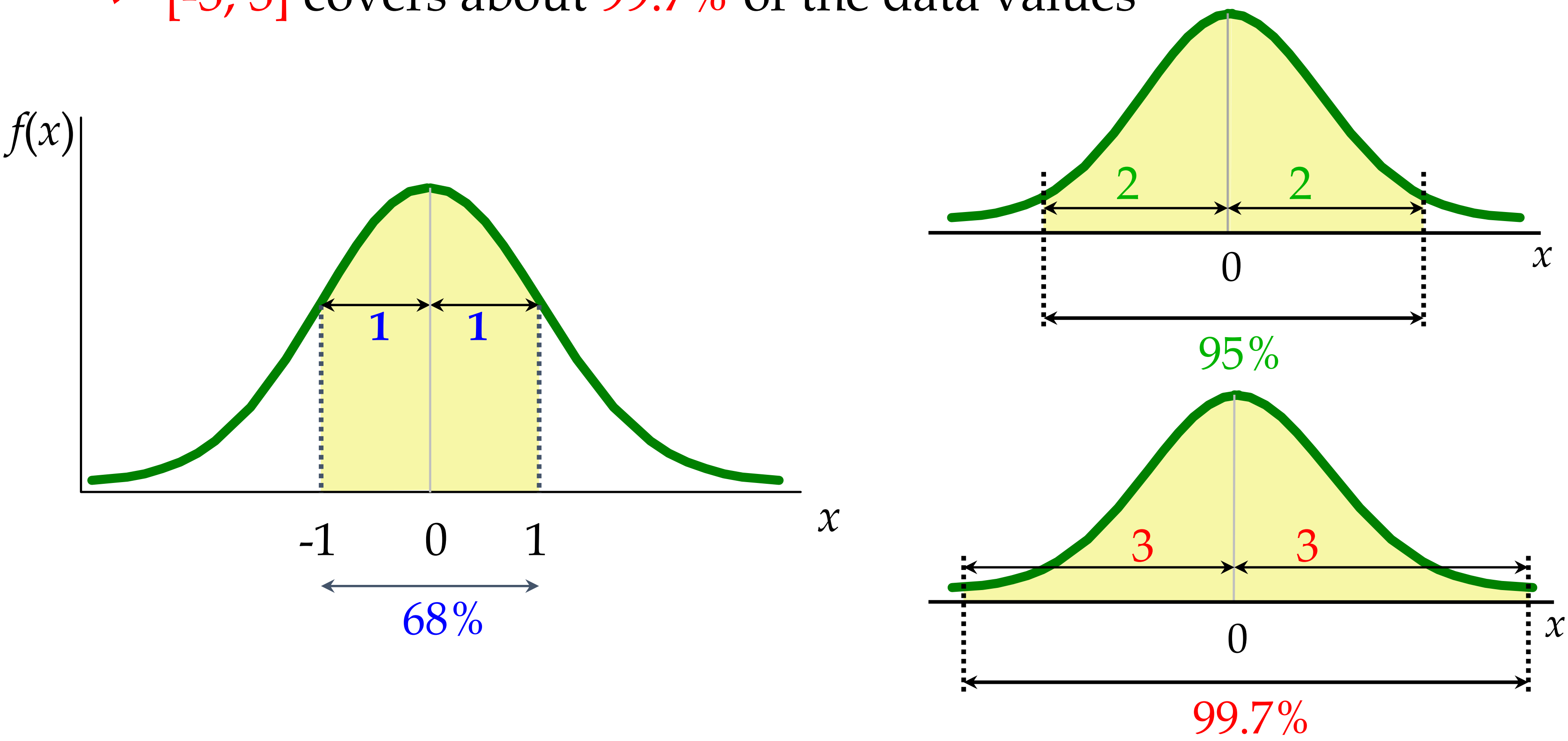
- ✓ $\mu \pm 1\sigma$ encloses about 68% of the data values
- ✓ $\mu \pm 2\sigma$ covers about 95% of the data values
- ✓ $\mu \pm 3\sigma$ covers about 99.7% of the data values



Revisiting the Empirical Rule

For a standard normal distribution:

- ✓ [-1; 1] encloses about 68% of the data values
- ✓ [-2; 2] covers about 95% of the data values
- ✓ [-3; 3] covers about 99.7% of the data values



Excel Time: Exercise 6.8, I

According to a recent survey by Smith Travel Research, the average daily rate for a luxury hotel in the United States is $237.22. Assume the daily rate follows a normal probability distribution with a standard deviation of $21.45.

What is the probability that **a randomly selected luxury hotel's daily rate** will be

a. Less than $250 per night?

b. More than $260?

c. Between $210 and $240?

Back <<

Excel Time: Exercise 6.8, II

d. The managers of a local luxury hotel would like to set the hotel's average daily rate at the 80th percentile, which is the rate below which 80% of hotels' rates are set. What rate should they choose for their hotel?

e. The managers of a local luxury hotel consider a price as a signal of a prestige of the hotel and would like to set the hotel's average daily rate not lower than 10% of the most expensive hotels. What rate should they choose for their hotel?

Back <<

Excel Time: Normal Probabilities in Excel

Excel's **NORM.DIST()** function can be used to find normal probabilities. Format for the NORM.DIST():

= NORM.DIST(*x*, mean, standard_dev, cumulative)

where:

cumulative = FALSE (or 0) if you need the probability density function

cumulative = TRUE (or 1) if you need the *cumulative* probability

In our class, we use cumulative = TRUE for calculating normal probabilities

Excel Time: Normal Probabilities in Excel

NORM.S.DIST() function can also be used to find probabilities from the *standard normal distribution:*

- utilize this function **when the *z*-score is known**
- format for the NORM.S.DIST() function:

$$= \text{NORM.S.DIST}(z, \text{cumulative})$$

where:

cumulative = FALSE if you need the probability density function

cumulative = TRUE if you need the *cumulative* probability

For this function, we don't need the mean and the SD:

- Standard normal distribution has mean = 0, SD = 1
- These values are already built-in into the function

Excel Time: Inverse Probabilities in Excel

NORM.INV() returns the inverse of the normal cumulative distribution for the specified mean and standard deviation:

- calculates the value of *x* that corresponds to a given value of the cumulative probability

= NORM.INV(probability, mean, standard_dev)

where probability = a given cumulative probability for the normal distribution

NORM.S.INV() calculates *z* value which corresponds to a given cumulative probability of the **standard normal distribution:**

NORM.S.INV(probability)

- Mean = 0 and SD = 1 are already built-in

Excel Time: Calculation Example

Given:

	A	B
1	Mean	237.22
2	SD	21.45

Solution (no *z*-transformation):

1. Lower tail probability:

D	E	F
$P(x \leq 250)$	0.7243	=NORM.DIST(250,B1,B2,TRUE)

2. Upper tail probability:

D	E	F
$P(x > 260)$	0.1441	=1-NORM.DIST(260,B1,B2,TRUE)

3. Probability between two values:

	D	E	F
1	$P(210 < x < 240)$	0.4493	=NORM.DIST(240,B1,B2,TRUE)-NORM.DIST(210,B1,B2,TRUE)

4. Inverse exercise:

- Lower tail probability is given:

D	E	F
x	255.2728	=NORM.INV(0.8,B1,B2)

- Upper tail probability is given:

D	E	F
x	264.7093	=NORM.INV(1-0.1,B1,B2)

Excel Time: Calculation Example

Given:

	A	B
1	Mean	237.22
2	SD	21.45

Solution (with a *z*-transformation):

1. Lower tail probability:

D	E	F
Value, *x*	250	
z-score	0.5958	=(E1-B1)/B2
$P(x < 250) = P(z < 0.5958)$	0.7243	=NORM.S.DIST(E2, TRUE)

2. Upper tail probability:

D	E	F
Value, *x*	260	
z-score	1.0620	=(E1-B1)/B2
$P(x > 260) = P(z > 1.0620) =$	0.1441	=1-NORM.S.DIST(E2, TRUE)

3. Probability between two values:

D	E	F
Upper *x*	240	
Lower *x*	210	
z-score (upper *x*)	0.1296	=(E1-B1)/B2
z-score (lower *x*)	-1.2690	=(E2-B1)/B2
$P(210 < x < 240) =$	0.4493	=NORM.S.DIST(E3, TRUE)-NORM.S.DIST(E4,TRUE)

4. Inverse exercise:
   - Lower tail probability is given:

D	E	F
z-score	0.8416	=NORM.S.INV(0.8)
x value	255.2728	=B1+E1*B2

   - Upper tail probability is given:

D	E	F
z-score	1.2816	=NORM.S.INV(1-0.1)
x value	264.7093	=B1+E1*B2

Excel Time: Exercise 6.18

According to the Pew Research Center, 53% of U.S. adults owned a tablet in January 2018. Calculate the probability that, from a sample of 24 U.S. adults:

a. Less than 10 adults own a tablet;
b. 6, 7, 8, 9, or 10 adults own a tablet;
c. Exactly 7 own a tablet.

For each part, find the probabilities:

- Using the binomial distribution
- Approximating the binomial distribution with the normal distribution

Note: See additional instructions for step-by-step procedures in Excel

Back <<

Excel Time: Exercise 6.9 (Extra Practice)

Major League Baseball teams have become concerned about the length of games. During a recent season, games averaged 2 hours and 52 minutes (172 minutes) to complete. Assume the length of games follows the normal distribution with a standard deviation of 16 minutes.

What is the probability that a randomly selected game will be completed in

a. Less than 200 minutes? More than 200 minutes?
b. Less than 150 minutes? More than 150 minutes? Exactly 150 minutes?
c. What is the completion time in which 90% of the games will be finished?

Excel Time: Exercise 6.10 (Extra Practice)

A study conducted by Hershey's discovered that Americans consumed an average of 11.4 pounds of chocolate per year. Let's assume that the annual chocolate consumption follows the normal distribution with standard deviation of 3.6 pounds. What is the probability that a randomly selected individual consumes

a. Less than 7 lbs per year?
b. More than 9 lbs per year?
c. Between 8 and 12 lbs per year?
d. What is the annual consumption of chocolate that represents 60^{th} percentile?

Excel Time: Exercise 6.19 (Extra Practice)

According to the Bureau of Transportation Statistics, 81.9% of American Airlines flights were on time during 2017. Assume this percentage still holds true for the American Airlines. Calculate probability that, from a sample of 30 flights:

a. Less than 22 will arrive on time;
b. Exactly 26 flights will arrive on time;
c. Between 24 and 28 will arrive on time.

For each part, find the probabilities:

- Using the binomial distribution
- Approximating the binomial distribution with the normal distribution