Central Limit Theorem: Introduction*

Susan Dean
Barbara Illowsky, Ph.D.

Abstract

This module provides a brief introduction to the Central Limit Theorem.

1 Student Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Averages.
- Apply and interpret the Central Limit Theorem for Sums.

2 Introduction

What does it mean to be average? Why are we so concerned with averages? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study averages and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n and n is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means (averages) and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, n, that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more

*Version 1.13: Mar 9, 2010 8:00 am US/Central
†http://creativecommons.org/licenses/by/2.0/

observations are needed for the sample averages or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times. (The 8, 7, 9, and 11 were randomly chosen.)

Each time a person rolls more than one die, he/she calculates the **average** of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The average is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one average when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more averages for a total of 10 averages.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the average. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "average" when you roll one die, is just the face on the die, what distribution do these "averages" appear to be representing?

Draw the graph for the averages using 2 dice. Do the averages show any kind of pattern?

Draw the graph for the averages using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the averages using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The average of the averages remains approximately the same.
2. The spread of the averages (the standard deviation of the averages) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means (averages) tend toward a normal distribution (the sampling distribution).**

Glossary

Definition 1: Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Definition 2: Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(\mathrm{n}\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Central Limit Theorem: Central Limit Theorem for Sample Means (Averages)*

Susan Dean
Barbara Illowsky, Ph.D.

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$ which consists of sample means, tends to be **normally distributed** and
$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$

The Central Limit Theorem for Sample Means (Averages) says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means (averages) form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n, the sample size. n is the number of values that are averaged together not the number of times the experiment is done.

The random variable $\overline{X}$ has a different z-score associated with it than the random variable X. $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \quad (1)$$

μ_X is both the average of X and of $\overline{X}$.

$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$ = standard deviation of $\overline{X}$ and is called the **standard error of the mean.**

Example 1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population.

Problem 1

Find the probability that the **sample mean** is between 85 and 92.

*Version 1.16: Mar 9, 2010 8:06 am US/Central

Solution

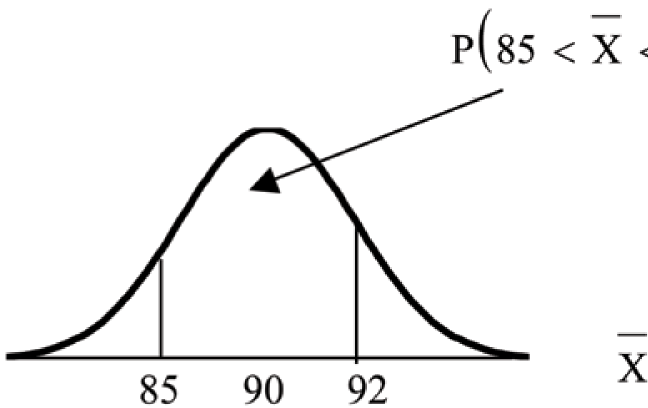
Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean (or average)**.

Let $\overline{X}$ = the mean or average of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$; then $\overline{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$

Find $P\left(85 < \overline{X} < 92\right)$ Draw a graph.

$P\left(85 < \overline{X} < 92\right) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.



TI-83 or 84: `normalcdf`(lower value, upper value, mean for averages, `stdev` for averages)

`stdev` = standard deviation

The parameter list is abbreviated (lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf` $\left(85, 92, 90, \frac{15}{\sqrt{25}}\right) = 0.6997$

Problem 2

Find the average value that is 2 standard deviations above the the mean of the averages.

Solution

To find the average value that is 2 standard deviations above the mean of the averages, use the formula

value $= \mu_X + (\#\text{ofSTDEVs})\left(\frac{\sigma_X}{\sqrt{n}}\right)$

value $= 90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$

So, the average value that is 2 standard deviations above the mean of the averages is 96.

Example 2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size** $n = $ **50** is drawn randomly from the population.

Problem 1

Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean or average time, in hours**, it takes to play one soccer match.

Let $\overline{X}$ = the **average** time, in hours, it takes to play one soccer match.

Problem 2 *(Solution on p. 4.)*

If μ_X = __________, σ_X = ___________, and n = ____________, then $\overline{X} \sim N($______, ______$)$ by the Central Limit Theorem for Averages of Sample Means.

Find $P\left(1.8 < \overline{X} < 2.3\right)$. Draw a graph.

$P\left(1.8 < \overline{X} < 2.3\right) = 0.9977$

normalcdf $\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the sample mean is between 1.8 hours and 2.3 hours is ______.

Solutions to Exercises in this Module

Solution to Example 2, Problem 2 (p. 3)

μ_X =**2**, σ_X =**0.5**, n =**50**, and $X \sim N\left(2\ ,\ \frac{0.5}{\sqrt{50}}\right)$

Glossary

Definition 1: Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, and geometric mean.

Definition 2: Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\overline{X}\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(\text{n}\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Definition 3: Normal Distribution

A continuous random variable (RV) with pdf f(x) $=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)2}/2\sigma^2$, where μ is the mean of the distribution and σ is its standard deviation. The notation is as follows: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

Definition 4: Standard Error of the Mean

The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem: Central Limit Theorem for Sums*

Susan Dean
Barbara Illowsky, Ph.D.

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

$\Sigma X \sim N\left(n \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right)$

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution). **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.
b. $z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$

a. $n \cdot \mu_X$ = the mean of ΣX
b. $\sqrt{n} \cdot \sigma_X$ = standard deviation of ΣX

Example 1

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

Problem

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
b. Find the sum that is 1.5 standard deviations below the mean of the sums.

*Version 1.12: Feb 12, 2011 10:57 am US/Central
†http://creativecommons.org/licenses/by/3.0/

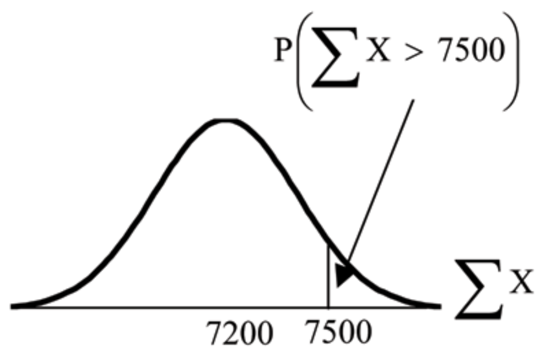
Solution

Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values.**

ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then
$\Sigma X \sim N\left(80 \cdot 90, \sqrt{80} \cdot 15\right)$

a. mean of the sums $= n \cdot \mu_X = (80)(90) = 7200$
b. standard deviation of the sums $= \sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
c. sum of 80 values $= \Sigma \text{x} = 7500$

Find $P(\Sigma \text{X} > 7500)$ Draw a graph.
$P(\Sigma \text{X} > 7500) = 0.0127$



`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)
The parameter list is abbreviated (lower, upper, $n \cdot \mu_X, \sqrt{n} \cdot \sigma_X$)
`normalcdf`(7500,1E99, $80 \cdot 90, \sqrt{80} \cdot 15 = 0.0127$
Reminder:1E99 $= 10^{99}$. Press the `EE` key for E.

Glossary

Definition 1: Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(\text{n}\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Definition 2: Normal Distribution

A continuous random variable (RV) with pdf f(x) $= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)2}/2\sigma^2$, where μ is the mean of the distribution and σ is its standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Central Limit Theorem: Using the Central Limit Theorem*

Susan Dean
Barbara Illowsky, Ph.D.

It is important for you to understand when to use the **CLT**. If you are being asked to find the probability of an average or mean, use the CLT for means or averages. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for averages and sums.

> NOTE: If you are being asked to find the probability of an **individual** value, do **not** use the CLT. **Use the distribution of its random variable.**

1 Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample gets closer and closer to μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample averages follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ must be close to the population mean μ. We can say that μ is the value that the sample averages approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean (Average) and Sum Examples

Example 1

A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the **average stress score** for the 75 students is less than 2.
2. The 90th percentile for the **average stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.

Problems 1. and 2. ask you to find a probability or a percentile for an **average** or **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.

*Version 1.15: Mar 8, 2010 2:48 pm US/Central
†http://creativecommons.org/licenses/by/2.0/

Since the individual stress scores follow a uniform distribution, $X \sim U(1,5)$ where $a = 1$ and $b = 5$ (See Continuous Random Variables[1] for the uniform).

$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$

$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$

For problems 1. and 2., let $\overline{X}$ = the average stress score for the 75 students. Then,

$\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where n = 75.

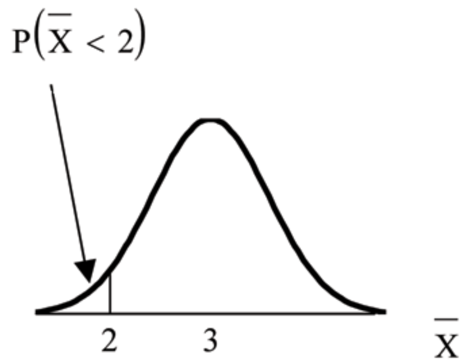
Problem 1

Find $P\left(\overline{X} < 2\right)$. Draw the graph.

Solution

$P\left(\overline{X} < 2\right) = 0$

The probability that the average stress score is less than 2 is about 0.



`normalcdf` $\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

Reminder: The smallest stress score is 1. Therefore, the smallest average for 75 stress scores is 1.

Problem 2

Find the 90th percentile for the average of 75 stress scores. Draw a graph.

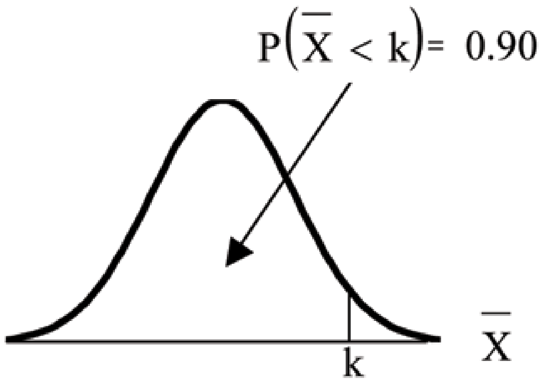
Solution

Let k = the 90th precentile.

Find k where $P\left(\overline{X} < k\right) = 0.90$.

$k = 3.2$

[1] "Continuous Random Variables: Introduction" <http://cnx.org/content/m16808/latest/>



The 90th percentile for the average of 75 scores is about 3.2. This means that 90% of all the averages of 75 stress scores are at most 3.2 and 10% are at least 3.2.

`invNorm` $\left(.90, 3, \frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma\text{X} \sim N\left[(75)\cdot(3), \sqrt{75}\cdot 1.15\right]$

Problem 3

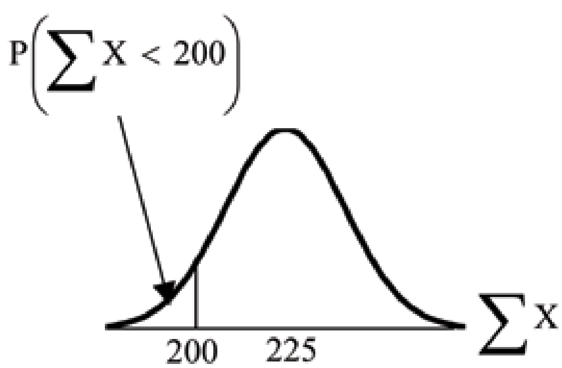
Find $P(\Sigma\text{X} < 200)$. Draw the graph.

Solution

The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75}\cdot 1.15 = 9.96$

$P(\Sigma\text{X} < 200) = 0$



The probability that the total of 75 scores is less than 200 is about 0.

`normalcdf`$\left(75, 200, 75\cdot 3, \sqrt{75}\cdot 1.15\right) = 0.$

REMINDER: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Problem 4

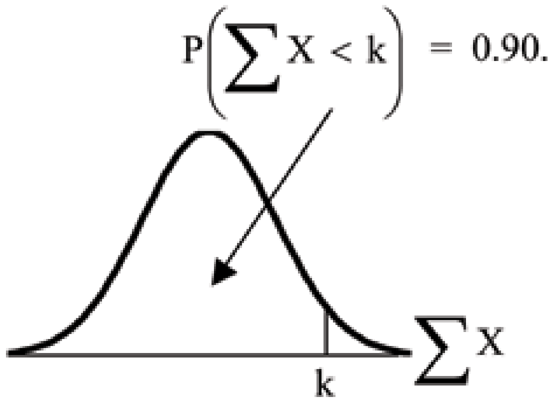
Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution

Let k = the 90th percentile.

Find k where $P(\Sigma\text{X} < k) = 0.90$.

$k = 237.8$



The 90th percentile for the sum of 75 scores is about 237.8. This means that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm` $\left(.90, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 237.8$

Example 2

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract. The analyst finds that for those customers who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes. Consider a random sample of 80 customers. Find

1. The probability that the **average excess time** used by the 80 customers in the sample is longer than 20 minutes. Draw a graph.
2. The 95th percentile for the **average excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Let X = the excess time used by one individual cell phone customer who exceeds his contracted time allowance. Then $X \sim \text{Exp}\left(\frac{1}{22}\right)$ (see Continuous Random Variables[2] for the exponential). Because X is exponential, $\mu = 22$ and $\sigma = 22$. The sample size is $n = 80$.

Let $\overline{X}$ = the **average** excess time used by a sample of $n = 80$ customers who exceed their contracted time allowances. Then

$\overline{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means or Averages

Problem 1

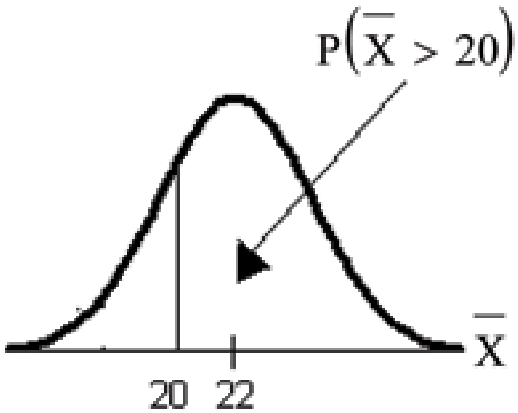
Find $P\left(\overline{X} > 20\right)$. Draw the graph.

Solution

$P\left(\overline{X} > 20\right) = 0.7919$

The probability that the average excess time used by a sample of 80 customers is longer than 20 minutes is 0.7919.

[2]"Continuous Random Variables: Introduction" <http://cnx.org/content/m16808/latest/>



`normalcdf` $\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$

REMINDER: $1E99 = 10^{99}$ and$-1E99 = -10^{99}$. Press the `EE` key for E.

Problem 2

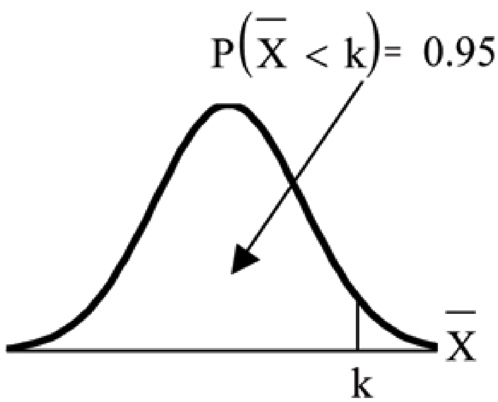
Find the 95th percentile for the **average excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution

Let k = the 95th percentile for the average excess time.

Find k where $P\left(\overline{X} < k\right) = 0.95$

$k = 26.0$



The 95th percentile for the average excess time for samples of 80 customers who exceed their basic contract time allowances is about 26 minutes. This means that 95% of the average excess times are at most 26 minutes and 10% are at least 26 minutes.

`invNorm` $\left(.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$

NOTE: **(HISTORICAL): Normal Approximation to the Binomial**

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for n (say, 20). To calculate the probabilities with large values of n, you had to use the binomial formula which could be very complicated. Using the **Normal Approximation to the Binomial** simplified the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X{\sim}B\left(n,p\right)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to X or subtract 0.5 from X(use $X + 0.5$ or $X - 0.5$. The number 0.5 is called the **continuity correction factor**.

Example 3

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K - 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let $X =$ the number that favor a charter school for grades K - 5. $X{\sim}B\left(n,p\right)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N\left(159, 8.6447\right)$. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so $P\left(X \geq 150\right)$ has normal approximation $P\left(Y \geq 149.5\right) = 0.8641$.

`normalcdf`$(149.5, 10\hat{}99, 159, 8.6447) = 0.8641$.

For Problem 2., you **include 160** so $P\left(X \leq 160\right)$ has normal approximation $P\left(Y \leq 160.5\right) = 0.5689$.

`normalcdf` $(0, 160.5, 159, 8.6447) = 0.5689$

For Problem 3., you **exclude 155** so $P\left(X > 155\right)$ has normal approximation $P\left(Y > 155.5\right) = 0.6572$.

`normalcdf` $(155.5, 10\hat{}99, 159, 8.6447) = 0.6572$

For Problem 4., you **exclude 147** so $P\left(X < 147\right)$ has normal approximation $P\left(Y < 146.5\right) = 0.0741$.

`normalcdf` $(0, 146.5, 159, 8.6447) = 0.0741$

For Problem 5., $P\left(X = 175\right)$ has normal approximation $P\left(174.5 < Y < 175.5\right) = 0.0083$.

`normalcdf` $(174.5, 175.5, 159, 8.6447) = 0.0083$

Because of calculators and computer software that easily let you calculate binomial probabilities for large values of n, it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial.

In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.

For **Example 3**, the probabilities are calculated using the binomial ($n = 300$ and $p = 0.53$) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.

$P(X \geq 150)$: `1 - binomialcdf`$(300, 0.53, 149) = 0.8641$
$P(X \leq 160)$: `binomialcdf`$(300, 0.53, 160) = 0.5684$
$P(X > 155)$: `1 - binomialcdf`$(300, 0.53, 155) = 0.6576$
$P(X < 147)$: `binomialcdf`$(300, 0.53, 146) = 0.0742$
$P(X = 175)$: (You use the binomial pdf.) `binomialpdf`$(175, 0.53, 146) = 0.0083$

Glossary

Definition 1: Average
A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Definition 2: Central Limit Theorem
Given a random variable (RV) with known mean μ and known standard deviation σ, we are sampling with size n and we are interested in two new RVs - called the sample mean, $\overline{X}$,and called the sample sum,ΣX. If the size n of the sample is sufficiently large, then $\overline{X}\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(\mathrm{n}\mu, \sqrt{n}\sigma)$. In words, if the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. And even more, the mean of the sampling distribution will equal the population mean and the mean of sampling sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Definition 3: Exponential Distribution
Continuous random variable (RV) that appears when we are interested in intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: $X\tilde{~}\mathrm{Exp}(m)$; the mean is $\mu = \frac{1}{m}$, and the standard deviation is $\sigma = \frac{1}{m}$, the probability density function is $f(x) = \mathrm{me}^{-\mathrm{mx}}, x \geq 0$ and the cumulative distribution is $P(X \leq x) = 1 - e^{-\mathrm{mx}}$.

Definition 4: Mean
A number to measure the central tendency (average), shortened from the arithmetic mean. By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by u) is $u = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Central Limit Theorem: Homework*

Susan Dean

Barbara Illowsky, Ph.D.

Exercise 1 *(Solution on p. 8.)*

$X \sim N(60, 9)$. Suppose that you form random samples of 25 from this distribution. Let $\overline{X}$ be the random variable of averages. Let ΣX be the random variable of sums. For **c - f**, sketch the graph, shade the region, label and scale the horizontal axis for $\overline{X}$, and find the probability.

a. Sketch the distributions of X and $\overline{X}$ on the same graph.
b. $\overline{X} \sim$
c. $P\left(\overline{X} < 60\right) =$
d. Find the 30th percentile.
e. $P\left(56 < \overline{X} < 62\right) =$
f. $P\left(18 < \overline{X} < 58\right) =$
g. $\Sigma X \sim$
h. Find the minimum value for the upper quartile.
i. $P(1400 < \Sigma X < 1550) =$

Exercise 2

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of $\overline{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\overline{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\overline{X}$ is approximately the same as the standard deviation of X.

Exercise 3 *(Solution on p. 8.)*

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about 10. Suppose that 16 individuals are randomly chosen.

Let $\overline{X}$ =average percent of fat calories.

a. $\overline{X}$~______ (______ , ______)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than 5. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

*Version 1.20: Mar 8, 2010 2:57 pm US/Central

Exercise 4

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of \$0.88. Suppose that we randomly pick 25 daytime statistics students.

a. In words, $X =$
b. $X\tilde{}$
c. In words, $\overline{X} =$
d. $\overline{X}\tilde{}$ ______ (______ , ______)
e. Find the probability that an individual had between \$0.80 and \$1.00. Graph the situation and shade in the area to be determined.
f. Find the probability that the average of the 25 students was between \$0.80 and \$1.00. Graph the situation and shade in the area to be determined.
g. Explain the why there is a difference in (e) and (f).

Exercise 5 *(Solution on p. 8.)*

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X}\tilde{}$ _______ (_______ , _______)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\overline{X}$. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of the average of 49 fly balls.

Exercise 6

Suppose that the weight of open boxes of cereal in a home with children is uniformly distributed from 2 to 6 pounds. We randomly survey 64 homes with children.

a. In words, $X =$
b. $X\tilde{}$
c. $\mu_X =$
d. $\sigma_X =$
e. In words, $\Sigma \mathrm{X} =$
f. $\Sigma \mathrm{X}\tilde{}$
g. Find the probability that the total weight of open boxes is less than 250 pounds.
h. Find the 35th percentile for the total weight of open boxes of cereal.

Exercise 7 *(Solution on p. 8.)*

Suppose that the duration of a particular type of criminal trial is known to have a mean of 21 days and a standard deviation of 7 days. We randomly sample 9 trials.

a. In words, $\Sigma \mathrm{X} =$
b. $\Sigma \mathrm{X}\tilde{}$
c. Find the probability that the total length of the 9 trials is at least 225 days.
d. 90 percent of the total of 9 of these types of trials will last at least how long?

Exercise 8

According to the Internal Revenue Service, the average length of time for an individual to complete (record keep, learn, prepare, copy, assemble and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is 2 hours. Suppose we randomly sample 36 taxpayers.

a. In words, $X =$
b. In words, $\overline{X} =$
c. $\overline{X}$~
d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
e. Would you be surprised if one taxpayer finished his Form 1040 in more than 12 hours? In a complete sentence, explain why.

Exercise 9 ***(Solution on p. 8.)***

Suppose that a category of world class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races.

Let $\overline{X}$ = the average of the 49 races.

a. $\overline{X}$~
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80th percentile for the average of these 49 marathons.
d. Find the median of the average running times.

Exercise 10

The attention span of a two year-old is exponentially distributed with a mean of about 8 minutes. Suppose we randomly survey 60 two year-olds.

a. In words, $X =$
b. X~
c. In words, $\overline{X} =$
d. $\overline{X}$~
e. Before doing any calculations, which do you think will be higher? Explain why.

  i. the probability that an individual attention span is less than 10 minutes; or
  ii. the probability that the average attention span for the 60 children is less than 10 minutes? Why?

f. Calculate the probabilities in part (e).
g. Explain why the distribution for $\overline{X}$ is not exponential.

Exercise 11 ***(Solution on p. 8.)***

Suppose that the length of research papers is uniformly distributed from 10 to 25 pages. We survey a class in which 55 research papers were turned in to a professor. We are interested in the average length of the research papers.

a. In words, $X =$
b. X~
c. $\mu_X =$
d. $\sigma_X =$
e. In words, $\overline{X} =$
f. $\overline{X}$~
g. In words, ΣX =
h. ΣX~
i. Without doing any calculations, do you think that it's likely that the professor will need to read a total of more than 1050 pages? Why?
j. Calculate the probability that the professor will need to read a total of more than 1050 pages.
k. Why is it so unlikely that the average length of the papers will be less than 12 pages?

Exercise 12

The length of songs in a collector's CD collection is uniformly distributed from 2 to 3.5 minutes. Suppose we randomly pick 5 CDs from the collection. There is a total of 43 songs on the 5 CDs.

a. In words, $X =$
b. $X^{\sim}$
c. In words, $\overline{X}=$
d. $\overline{X}^{\sim}$
e. Find the first quartile for the average song length.
f. The IQR (interquartile range) for the average song length is from _______ to _______.

Exercise 13 *(Solution on p. 8.)*

Salaries for teachers in a particular elementary school district are normally distributed with a mean of \$44,000 and a standard deviation of \$6500. We randomly survey 10 teachers from that district.

a. In words, $X =$
b. In words, $\overline{X} =$
c. $\overline{X}^{\sim}$
d. In words, $\Sigma X =$
e. $\Sigma X^{\sim}$
f. Find the probability that the teachers earn a total of over \$400,000.
g. Find the 90th percentile for an individual teacher's salary.
h. Find the 90th percentile for the average teachers' salary.
i. If we surveyed 70 teachers instead of 10, graphically, how would that change the distribution for $\overline{X}$?
j. If each of the 70 teachers received a \$3000 raise, graphically, how would that change the distribution for $\overline{X}$?

Exercise 14

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and few to many wealthy people). Suppose we pick a country with a wedge distribution. Let the average salary be \$2000 per year with a standard deviation of \$8000. We randomly survey 1000 residents of that country.

a. In words, $X =$
b. In words, $\overline{X} =$
c. $\overline{X}^{\sim}$
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average of the 1000 residents will be from \$2000 to \$2100 than from \$2100 to \$2200?

Exercise 15 *(Solution on p. 9.)*

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

a. In words, $X =$
b. In words, $\overline{X} =$
c. $\overline{X}^{\sim}$
d. In words, $\Sigma X =$
e. $\Sigma X^{\sim}$
f. Is it likely that an individual stayed more than 5 days in the hospital? Why or why not?

g. Is it likely that the average stay for the 80 women was more than 5 days? Why or why not?
h. Which is more likely:

- **i.** an individual stayed more than 5 days; or
- **ii.** the average stay of 80 women was more than 5 days?

i. If we were to sum up the women's stays, is it likely that, collectively they spent more than a year in the hospital? Why or why not?

Exercise 16

In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940. (Source: U.S. Dept. of Agriculture)

a. In words, $X =$
b. In words, $\overline{X} =$
c. $\overline{X}$~
d. The IQR for $\overline{X}$ is from _______ acres to _______ acres.

Exercise 17 *(Solution on p. 9.)*

The stock closing prices of 35 U.S. semiconductor manufacturers are given below. (Source: **Wall Street Journal**)

8.625; 30.25; 27.625; 46.75; 32.875; 18.25; 5; 0.125; 2.9375; 6.875; 28.25; 24.25; 21; 1.5; 30.25; 71; 43.5; 49.25; 2.5625; 31; 16.5; 9.5; 18.5; 18; 9; 10.5; 16.625; 1.25; 18; 12.875; 7; 12.875; 2.875; 60.25; 29.25

a. In words, $X =$
b. **i.** $\overline{x} =$
- **ii.** $s_x =$
- **iii.** $n =$

c. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
d. In words, describe the distribution of stock prices.
e. Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 10 averages. List those 10 averages.
f. Use the 10 averages from (e) to calculate:

- **i.** $\overline{x} =$
- **ii.** $\overline{s_x} =$

g. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
h. Does this histogram look like the graph in (c)?
i. In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
j. Based upon the theory of the Central Limit Theorem, $\overline{X}$~

Exercise 18

Use the Initial Public Offering data[1] (see "Table of Contents) to do this problem.

a. In words, $X =$
b. **i.** $\mu_X =$
- **ii.** $\sigma_X =$
- **iii.** $n =$

c. Construct a histogram of the distribution. Start at $x = -0.50$. Make bar widths of \$5.

[1] "Collaborative Statistics: Data Sets": Section Stock Prices <http://cnx.org/content/m17132/latest/#element-949>

d. In words, describe the distribution of stock prices.
e. Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 15 averages. List those 15 averages.
f. Use the 15 averages from (e) to calculate the following:

i. $\overline{x} =$
ii. $\overline{s_x} =$

g. Construct a histogram of the distribution of the averages. Start at $x = -0.50$. Make bar widths of \$5.
h. Does this histogram look like the graph in (c)? Explain any differences.
i. In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
j. Based upon the theory of the Central Limit Theorem, $\overline{X}$~

1 Try these multiple choice questions (Exercises19 - 23).

The next two questions refer to the following information: The time to wait for a particular rural bus is distributed uniformly from 0 to 75 minutes. 100 riders are randomly sampled to learn how long they waited.

Exercise 19 *(Solution on p. 9.)*

The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

A. 315.0
B. 40.3
C. 38.5
D. 65.2

Exercise 20 *(Solution on p. 9.)*

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

A. Yes
B. No
C. There is not enough information.

Exercise 21 *(Solution on p. 9.)*

Which of the following is NOT TRUE about the distribution for averages?

A. The mean, median and mode are equal
B. The area under the curve is one
C. The curve never touches the x-axis
D. The curve is skewed to the right

The next two questions refer to the following information: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$2.59 and a standard deviation of \$0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

Exercise 22 *(Solution on p. 9.)*

The distribution to use for the average cost of gasoline for the 16 gas stations is

A. $\overline{X} \sim N(2.59, 0.10)$

B. $\overline{X} \sim N\left(2.59, \frac{0.10}{\sqrt{16}}\right)$
C. $\overline{X} \sim N\left(2.59, \frac{0.10}{16}\right)$
D. $\overline{X} \sim N\left(2.59, \frac{16}{0.10}\right)$

Exercise 23 *(Solution on p. 9.)*

What is the probability that the average price for 16 gas stations is over $2.69?

A. Almost zero
B. 0.1587
C. 0.0943
D. Unknown

Exercise 24 *(Solution on p. 9.)*

For the Charter School Problem (Example 3) in **Central Limit Theorem: Using the Central Limit Theorem**, calculate the following using the normal approximation to the binomial.

A. Find the probability that less than 100 favor a charter school for grades K - 5.
B. Find the probability that 170 or more favor a charter school for grades K - 5.
C. Find the probability that no more than 140 favor a charter school for grades K - 5.
D. Find the probability that there are fewer than 130 that favor a charter school for grades K - 5.
E. Find the probability that exactly 150 favor a charter school for grades K - 5.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Exercise 25 *(Solution on p. 9.)*

Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities.

A. Find the probability that Janice is the driver at most 20 days.
B. Find the probability that Roberta is the driver more than 16 days.
C. Find the probability that Barbara drives exactly 24 of those 96 days.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Solutions to Exercises in this Module

Solution to Exercise 1 (p. 1)

b. Xbar~$N\left(60, \frac{9}{\sqrt{25}}\right)$
c. 0.5000
d. 59.06
e. 0.8536
f. 0.1333
h. 1530.35
i. 0.8536

Solution to Exercise 3 (p. 1)

a. $N\left(36, \frac{10}{\sqrt{16}}\right)$
b. 1
c. 34.31

Solution to Exercise 5 (p. 2)

a. $N\left(250, \frac{50}{\sqrt{49}}\right)$
b. 0.0808
c. 256.01 feet

Solution to Exercise 7 (p. 2)

a. The total length of time for 9 criminal trials
b. $N(189, 21)$
c. 0.0432
d. 162.09

Solution to Exercise 9 (p. 3)

a. $N\left(145, \frac{14}{\sqrt{49}}\right)$
b. 0.6247
c. 146.68
d. 145 minutes

Solution to Exercise 11 (p. 3)

b. $U(10, 25)$
c. 17.5
d. $\sqrt{\frac{225}{12}} = 4.3301$
f. $N(17.5, 0.5839)$
h. $N(962.5, 32.11)$
j. 0.0032

Solution to Exercise 13 (p. 4)

c. $N\left(44,000, \frac{6500}{\sqrt{10}}\right)$
e. $N\left(440{,}000, \left(\sqrt{10}\right)(6500)\right)$
f. 0.9742
g. $52,330

h. \$46,634

Solution to Exercise 15 (p. 4)

c. $N\left(2.4, \frac{0.9}{\sqrt{80}}\right)$
e. $N\left(192{,}8.05\right)$
h. Individual

Solution to Exercise 17 (p. 5)

b. \$20.71; \$17.31; 35
d. Exponential distribution, $X \sim \text{Exp}\left(1/20.71\right)$
f. \$20.71; \$11.14
j. $N\left(20.71, \frac{17.31}{\sqrt{5}}\right)$

Solution to Exercise 19 (p. 6)
B
Solution to Exercise 20 (p. 6)
A
Solution to Exercise 21 (p. 6)
D
Solution to Exercise 22 (p. 6)
B
Solution to Exercise 23 (p. 7)
A
Solution to Exercise 24 (p. 7)

C. 0.0162
E. 0.0268

Solution to Exercise 25 (p. 7)

A. 0.2047
B. 0.9615
C. 0.0938

Central Limit Theorem: Practice*

Susan Dean
Barbara Illowsky, Ph.D.

1 Student Learning Outcomes

- The student will explore the properties of data through the Central Limit Theorem.

2 Given

Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately 4 hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\overline{X}$ be the random variable representing the average time to complete the 16 reviews. Let ΣX be the total time it takes Yoonie to complete all of the month's reviews.

3 Distribution

Complete the distributions.

1. $X \sim$
2. $\overline{X} \sim$
3. $\Sigma X \sim$

4 Graphing Probability

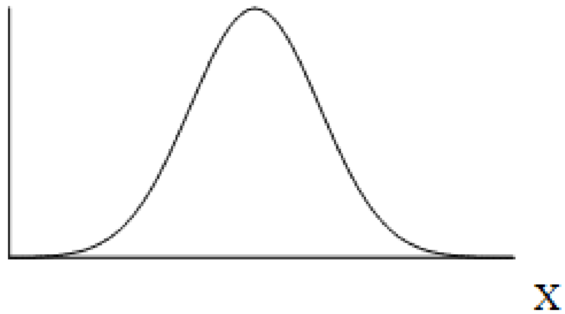
For each problem below:

a. Sketch the graph. Label and scale the horizontal axis. Shade the region corresponding to the probability.
b. Calculate the value.

Exercise 1 ***(Solution on p. 4.)***

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.

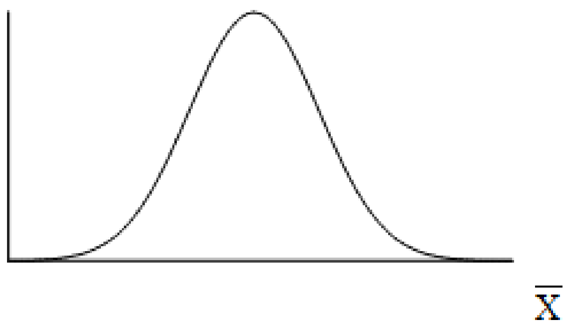
*Version 1.10: Feb 18, 2009 7:59 pm US/Central
†http://creativecommons.org/licenses/by/2.0/



a.
b. $P($ ________ $<X<$ ________ $=$ _______

Exercise 2 *(Solution on p. 4.)*

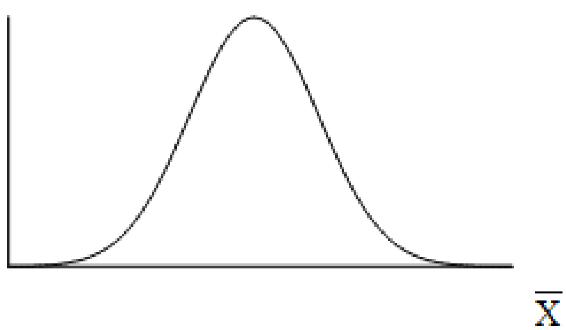
Find the probability that the **average** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs.



a.
b. $P(\ ) =$_______

Exercise 3 *(Solution on p. 4.)*

Find the 95th percentile for the **average** time to complete one month's reviews.

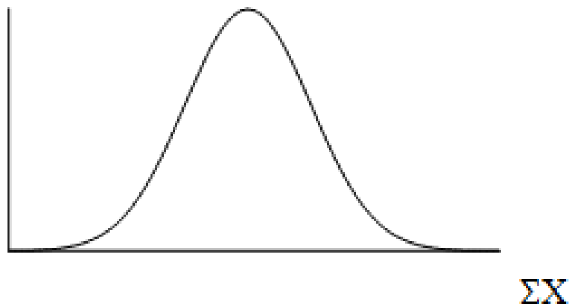


a.
b. The 95th Percentile=

Exercise 4 *(Solution on p. 4.)*

Find the probability that the **sum** of the month's reviews takes Yoonie from 60 to 65 hours.

a.

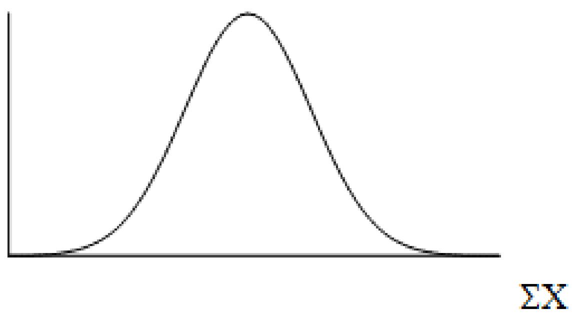


b. The Probability=

Exercise 5 *(Solution on p. 4.)*

Find the 95th percentile for the **sum** of the month's reviews.

a.



b. The 95th percentile=

5 Discussion Question

Exercise 6

What causes the probabilities in Exercise 1 and Exercise 2 to differ?

Solutions to Exercises in this Module

Solution to Exercise 1 (p. 1)

b. 3.5, 4.25, 0.2441

Solution to Exercise 2 (p. 2)

b. 0.7499

Solution to Exercise 3 (p. 2)

b. 4.49 hours

Solution to Exercise 4 (p. 2)

b. 0.3802

Solution to Exercise 5 (p. 3)

b: 71.90