



# The Central Limit Theorem for Sample Means (Averages)

By:

OpenStaxCollege

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:

1.  $\mu_X =$  the mean of  $X$
2.  $\sigma_X =$  the standard deviation of  $X$

If you draw random samples of size  $n$ , then as  $n$  increases, the random variable  $\bar{X}$  which consists of sample means, tends to be normally distributed and

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right).$$

The central limit theorem for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and **calculating their means**, the sample means 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, the sample size. The variable  $n$  is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size  $n$ , the distribution of the random variable  $\bar{X}$ , which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as  $n$ , the sample size, increases.

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

## The Central Limit Theorem for Sample Means (Averages)

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

$\mu_x$  is the average of both  $X$  and  $\bar{X}$ .

$\sigma_x = \frac{\sigma_x}{\sqrt{n}}$  = standard deviation of  $\bar{X}$  and is called the standard error of the mean.

To find probabilities for means on the calculator, follow these steps.

2nd DISTR  
2:normalcdf

*normalcdf(lower value of the area, upper value of the area, mean,  $\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$ )*

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* =  $n$

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.

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

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

Let  $\bar{X}$  = the mean of a sample of size 25. Since  $\mu_x = 90$ ,  $\sigma_x = 15$ , and  $n = 25$ ,

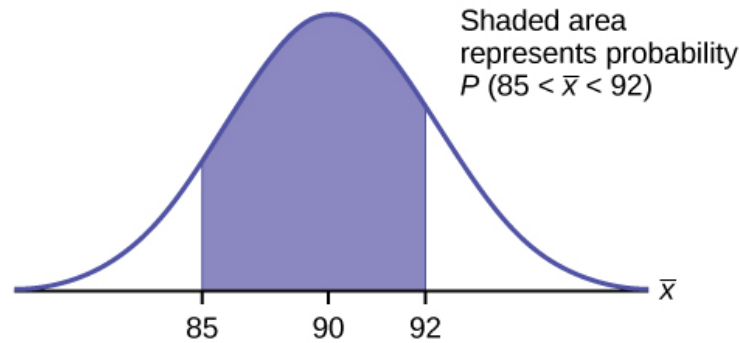
$$\bar{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right).$$

Find  $P(85 < \bar{x} < 92)$ . Draw a graph.

$$P(85 < \bar{x} < 92) = 0.6997$$

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

## The Central Limit Theorem for Sample Means (Averages)



`normalcdf(lower value, upper value, mean, standard error of the mean)`

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

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

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

b. To find the value that is two standard deviations above the expected value 90, use the formula:

$$\text{value} = \mu_x + (\text{\#ofTSDEVs})\left(\frac{\sigma_x}{\sqrt{n}}\right)$$

$$\text{value} = 90 + 2\left(\frac{15}{\sqrt{25}}\right) = 96$$

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is  $\frac{\sigma_x}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$ . Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size  $n$ .

Try It

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size  $n = 30$  are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

$$P(42 < \bar{x} < 50) = \text{normalcdf}(42,50,45,\frac{8}{\sqrt{30}}) = 0.9797$$

## The Central Limit Theorem for Sample Means (Averages)

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 two hours** and a **standard deviation of 0.5 hours**. A **sample of size  $n = 50$**  is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

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 time, in hours**, it takes to play one soccer match.

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

If  $\mu_{\bar{X}} = \underline{\hspace{2cm}}$ ,  $\sigma_{\bar{X}} = \underline{\hspace{2cm}}$ , and  $n = \underline{\hspace{2cm}}$ , then  $\bar{X} \sim N(\underline{\hspace{2cm}}, \underline{\hspace{2cm}})$  by the central limit theorem for means.

$$\mu_{\bar{X}} = 2, \sigma_{\bar{X}} = 0.5, n = 50, \text{ and } \bar{X} \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$$

Find  $P(1.8 < \bar{x} < 2.3)$ . Draw a graph.

$$P(1.8 < \bar{x} < 2.3) = 0.9977$$

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

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Try It

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of  $n = 60$  is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

$$P(2 < \bar{x} < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$$

To find percentiles for means on the calculator, follow these steps.

2<sup>nd</sup> DIST  
3:invNorm

$$k = \text{invNorm}\left(\text{area to the left of } k, \text{ mean}, \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$$

## The Central Limit Theorem for Sample Means (Averages)

where:

- $k$  = the  $k^{\text{th}}$  percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* =  $n$

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size  $n = 100$ .

1. What are the mean and standard deviation for the sample mean ages of tablet users?
  2. What does the distribution look like?
  3. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
  4. Find the 95<sup>th</sup> percentile for the sample mean age (to one decimal place).
1. Since the sample mean tends to target the population mean, we have  $\mu_{\bar{X}} = \mu = 34$ . The sample standard deviation is given by  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$
  2. The central limit theorem states that for large sample sizes( $n$ ), the sampling distribution will be approximately normal.
  3. The probability that the sample mean age is more than 30 is given by  $P(X > 30) = \text{normalcdf}(30, E99, 34, 1.5) = 0.9962$
  4. Let  $k$  = the 95<sup>th</sup> percentile.  
 $k = \text{invNorm}(0.95, 34, \frac{15}{\sqrt{100}}) = 36.5$

Try It

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$$P(29 < \bar{x} < 35) = \text{normalcdf}(29, 35, 28, \frac{4.8}{\sqrt{100}}) = 0.0186$$

You can conclude there is approximately a 19% chance that your game will be played by men whose mean age is between 29 and 35.

## The Central Limit Theorem for Sample Means (Averages)

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

1. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
2. What is the standard error of the mean?
3. Find the 90<sup>th</sup> percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
4. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

1.  $\mu_x = \mu = 8.2$   $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$

2. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.

3. Let  $k$  = the 90<sup>th</sup> percentile

$k = \text{invNorm}(0.90, 8.2, \frac{1}{\sqrt{60}}) = 8.37$ . This value indicates that 90 percent of the average app engagement time for tablet users is less than 8.37 minutes.

4.  $P(8 < x < 8.5) = \text{normalcdf}(8, 8.5, 8.2, \frac{1}{\sqrt{60}}) = 0.9293$

Try It

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are  $n = 34$ ,  $x = 16.01$  ounces. If the cans are filled so that  $\mu = 16.00$  ounces (as labeled) and  $\sigma = 0.143$  ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

We have  $P(x > 16.01) = \text{normalcdf}(16.01, E99, 16, \frac{0.143}{\sqrt{34}}) = 0.3417$ . Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

## References

Baran, Daya. "20 Percent of Americans Have Never Used Email." WebGuild, 2010. Available online at <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email> (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

## The Central Limit Theorem for Sample Means (Averages)

Data from the United States Department of Agriculture.

### Chapter Review

In a population whose distribution may be known or unknown, if the size ( $n$ ) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size ( $n$ ).

### Formula Review

The Central Limit Theorem for Sample Means:  $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

The Mean  $\bar{X}$ :  $\mu_x$

Central Limit Theorem for Sample Means z-score and standard error of the mean:

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

Standard Error of the Mean (Standard Deviation ( $\bar{X}$ )):  $\frac{\sigma_x}{\sqrt{n}}$

*Use the following information to answer the next six exercises:* 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 four 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  $\bar{X}$  be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

What is the mean, standard deviation, and sample size?

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

Complete the distributions.

1.  $\bar{X} \sim \underline{\hspace{1cm}}(\underline{\hspace{1cm}}, \underline{\hspace{1cm}})$
2.  $X \sim \underline{\hspace{1cm}}(\underline{\hspace{1cm}}, \underline{\hspace{1cm}})$

## The Central Limit Theorem for Sample Means (Averages)

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



- 1.
2.  $P(\text{_____} < x < \text{_____}) = \text{_____}$

a. Check student's solution.

b. 3.5, 4.25, 0.2441

Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



- 1.
2.  $P(\text{_____}) = \text{_____}$

What causes the probabilities in [\[link\]](#) and [\[link\]](#) to be different?

The fact that the two distributions are different accounts for the different probabilities.

Find the 95<sup>th</sup> percentile for the mean time to complete one month's reviews. Sketch the graph.



## The Central Limit Theorem for Sample Means (Averages)



- 1.
2. The 95<sup>th</sup> Percentile = \_\_\_\_\_

### Homework

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.

1. In words,  $X =$  \_\_\_\_\_
2.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
3. In words,  $X =$  \_\_\_\_\_
4.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
5. Find the probability that an individual had between \$0.80 and \$1.00. Graph the situation, and shade in the area to be determined.
6. 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.
7. Explain why there is a difference in part e and part f.
  1.  $X =$  amount of change students carry
  2.  $X \sim E(0.88, 0.88)$
  3.  $\bar{X} =$  average amount of change carried by a sample of 25 students.
  4.  $\bar{X} \sim N(0.88, 0.176)$
  5. 0.0819
  6. 0.1882
  7. The distributions are different. Part a is exponential and part b is normal.

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.

1. If  $\bar{X} =$  average distance in feet for 49 fly balls, then  $\bar{X} \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)

## The Central Limit Theorem for Sample Means (Averages)

2. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for  $X$ . Shade the region corresponding to the probability. Find the probability.
3. Find the 80<sup>th</sup> percentile of the distribution of the average of 49 fly balls.

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, 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 two hours. Suppose we randomly sample 36 taxpayers.

1. In words,  $\underline{X}$  = \_\_\_\_\_
2. In words,  $\underline{X}$  = \_\_\_\_\_
3.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
4. 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.
5. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.
  1. length of time for an individual to complete IRS form 1040, in hours.
  2. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
  3.  $N(10.53, \frac{1}{3})$
  4. Yes. I would be surprised, because the probability is almost 0.
  5. No. I would not be totally surprised because the probability is 0.2312

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  $X$  the average of the 49 races.

1.  $\underline{X} \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
2. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
3. Find the 80<sup>th</sup> percentile for the average of these 49 marathons.
4. Find the median of the average running times.

The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

1. In words,  $X$  = \_\_\_\_\_

The Central Limit Theorem for Sample Means (Averages)

2.  $X \sim$  \_\_\_\_\_
  3. In words,  $X =$  \_\_\_\_\_
  4.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
  5. Find the first quartile for the average song length.
  6. The IQR(interquartile range) for the average song length is from \_\_\_\_\_.
1. the length of a song, in minutes, in the collection
  2.  $U(2, 3.5)$
  3. the average length, in minutes, of the songs from a sample of five albums from the collection
  4.  $N(2.75, 0.0220)$
  5. 2.74 minutes
  6. 0.03 minutes

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.

1. In words,  $\bar{X} =$  \_\_\_\_\_
2. In words,  $X =$  \_\_\_\_\_
3.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
4. The IQR for  $\bar{X}$  is from \_\_\_\_\_ acres to \_\_\_\_\_ acres.

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

1. When the sample size is large, the mean of  $\bar{X}$  is approximately equal to the mean of  $X$ .
  2. When the sample size is large,  $\bar{X}$  is approximately normally distributed.
  3. When the sample size is large, the standard deviation of  $\bar{X}$  is approximately the same as the standard deviation of  $X$ .
1. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
  2. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
  3. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of  $X$  as the sample size increases.

## The Central Limit Theorem for Sample Means (Averages)

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 ten. Suppose that 16 individuals are randomly chosen. Let  $\bar{X}$  = average percent of fat calories.

1.  $\bar{X} \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
2. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
3. Find the first quartile for the average percent of fat calories.

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be \$2,000 per year with a standard deviation of \$8,000. We randomly survey 1,000 residents of that country.

1. In words,  $\bar{X}$  = \_\_\_\_\_
  2. In words,  $X$  = \_\_\_\_\_
  3.  $X \sim$  \_\_\_\_\_ (\_\_\_\_\_, \_\_\_\_\_)
  4. How is it possible for the standard deviation to be greater than the average?
  5. Why is it more likely that the average of the 1,000 residents will be from \$2,000 to \$2,100 than from \$2,100 to \$2,200?
    1.  $X$  = the yearly income of someone in a third world country
    2. the average salary from samples of 1,000 residents of a third world country
    3.  $X \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
    4. Very wide differences in data values can have averages smaller than standard deviations.
    5. The distribution of the sample mean will have higher probabilities closer to the population mean.
- $$P(2000 < \bar{X} < 2100) = 0.1537$$
- $$P(2100 < \bar{X} < 2200) = 0.1317$$

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

1. The mean, median, and mode are equal.
2. The area under the curve is one.
3. The curve never touches the  $x$ -axis.
4. The curve is skewed to the right.

## The Central Limit Theorem for Sample Means (Averages)

The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$4.59 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. The distribution to use for the average cost of gasoline for the 16 gas stations is:

$$1. \bar{X} \sim N(4.59, 0.10)$$

$$2. \bar{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$$

$$3. \bar{X} \sim N\left(4.59, \frac{16}{0.10}\right)$$

$$4. \bar{X} \sim N\left(4.59, \frac{16}{0.10}\right)$$

b