

# Math 111, section 08.x supplement: The Central Limit Theorem

notes by Tim Pilachowski

## Table of Contents

1. Background	1
2. Theoretical	2
3. Practical	3
4. The Central Limit Theorem	4
5. Homework Exercises	7

## 1. Background

*Gathering Data.* Information is being collected and analyzed all the time by various groups for a vast variety of purposes. For example, manufacturers test products for reliability and safety, news organizations want to know which candidate is ahead in a political campaign, advertisers want to know whether their wares appeal to potential customers, and the government needs to know the rate of inflation and the percent of unemployment to predict expenditures.

The items included in the analysis and the people actually answering a survey are the **sample**, while the whole group of things or people we want information about is the **population**. How well does a sample reflect the status or opinions of the population? The answer depends on many variables, some easier to control than others.

*Sources of Bias.* One source of bias is carelessly prepared questions. For example, if a pollster asks, “Do you think our overworked, underpaid teachers deserve a raise?” the answers received are likely to reflect the slanted way in which the question was asked! Another danger is a poorly selected sample. If the person conducting the survey stands outside a shopping center at 2 pm on a weekday, the sample may not be representative of the entire population because people at work or people who shop elsewhere have no chance of being selected.

*Random Sampling.* To eliminate bias due to poor sample selection, statisticians insist on **random sampling**, in which *every member of the population is equally likely to be included in the sample*. For a small population, simple random sampling can be achieved by putting everyone’s name on a slip of paper, placing all the slips in a large bin, mixing well, and selecting the sample.

Technically, each slip should be replaced before the next slip is drawn, so the probability doesn’t change for subsequent draws. That is, the events should be independent, and the probability not conditional. Of course, the slips should be mixed before each draw, as well. In practice, most populations being tested are large enough that the conditional probabilities don’t significantly alter the results.

*Representative Samples.* It is important to note that saying the sample is free of bias does not guarantee that it will perfectly represent the population. A random sample may, by chance variation, be very different from the population. However, the mathematics underlying statistics allows us to specify how much a random sample is likely to vary and how often it will be extremely non-representative. Bias, on the other hand, affects the results in unknown ways that cannot be quantified or specified in any such manner.

### Example 1

Suppose that a club has 30 members, and five are to be randomly selected to represent the club at a regional conference. Names of female members are in *italics*. Non-italicized names belong to male members.

Abby	Greg	Mark	Sam	Zelda
Ben	Helen	Norris	Tom	Art
Charlie	Ivan	Oscar	Ursula	Bob
David	Jamal	Patricia	Vic	Curt
Eric	Kevin	Quentin	Walt	Dan
Frank	Larry	Rob	Yvette	Evan

Suppose you were to select ten different random samples of five club members from this population of 30. Even though exactly 1/5 of the population is women, it is very likely your samples would not all include exactly 1/5 female. Some samples might have no women, two women, or maybe even all five picks would be women. This doesn't mean the sampling process was biased—it simply illustrates sampling variability.

### Example 2

Similarly, if you flip a coin ten times, you would not be surprised if it came up heads only four times or six times rather than exactly 1/2, or five, of the times. If, on the other hand, you flipped a coin ten times and got ten heads, you probably would be surprised. We know intuitively that it is unlikely we will get such an extremely unrepresentative sample. It is not impossible, just very rare. We may even decide the coin must be weighted in some way so that heads are more likely to appear.

Sampling variability is also affected by the number of observations we include. If we flip a coin ten times and get only 3 heads, or 30%, we may not be very surprised. If we flip the same coin 1000 times and only get 300 heads, however, this is more surprising. Again, we may begin to wonder whether the coin is fair. Over a larger number of trials, the percent of heads should be closer to the 50% heads we expect.

## 2. Theoretical

If you were able to construct every possible sample of a specified size from the same population, you would create what statisticians call a **sampling distribution**. The Central Limit Theorem tells us, in short, that a sampling distribution (based on the theoretical possibility of creating every possible sample of size  $n$ ) is close to a normal distribution.

Furthermore, the Central Limit Theorem tells us that the peak of a sampling distribution is the population mean (symbolized by  $\mu$ ). The technical statistical term for this “true” mean is the “population parameter” or just “parameter”. In contrast, the mean calculated from a given sample (symbolized by  $\bar{x}$ ) is a “sample statistic” or just “statistic”. The mean of the sampling distribution *is* the mean of the population [ $E(\bar{x}) = \mu$ ]. Calculated sample means (statistics) will vary—while they are most likely to occur near the “true” population mean, they may also be far away from the population mean, depending on the particular characteristics of the particular sample selected.

Also, we know some facts about normal distributions that will be helpful in determining how likely it is that a sample statistic occurs near the population parameter. One characteristic of normal probability distributions:

- 68% of the outcomes are within 1 standard deviation of the mean
- 95% of the outcomes are within 2 standard deviations of the mean
- 99.7% of the outcomes are within 3 standard deviations of the mean.

What does this mean for random sampling? It tells us that 68% of the time, a random sample will give us a result—a statistic—within 1 standard deviation of the “true” parameter. We would expect that 95% of the time, a random sample will give a statistic within 2 standard deviations of the population parameter, and 99.7% of the time, a random sample will give a statistic within 3 standard deviations of the population parameter. In a beginning level statistics course, you would be introduced to confidence intervals and hypothesis tests, each of which makes use of the percents listed above.

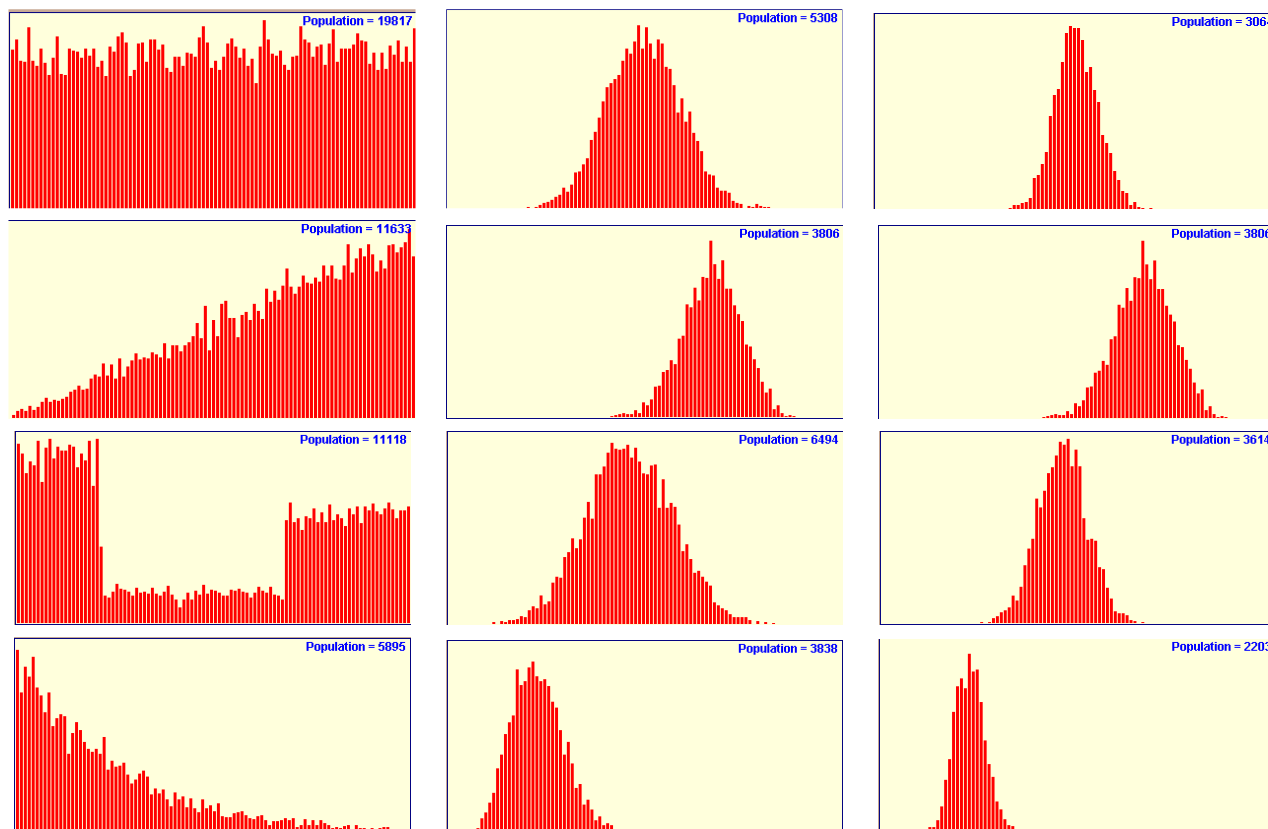
### 3. Practical

Constructing every possible sample of a specified size from the population is a theoretical construct. In practical terms, it's never going to happen. Time and cost would prevent researchers from even trying. Likewise, an infinite number of trials is an impossibility. However, the Central Limit Theorem also helps us in this regard.

A sample of size  $n = 1$  is not likely to be representative of the entire population. Neither is a sample of size  $n = 2$ . It makes sense intuitively that a larger sample size would be more likely to be representative of the population. (See Example 2 above).

The Central Limit Theorem not only verifies this intuition, but also tells us that, as sample size increases, the shape of the sampling distribution grows closer to the shape of a normal distribution. The following series of graphics illustrate this. They were drawn by an applet found on the website [http://www.chem.uoa.gr/applets/AppletCentralLimit/Applet\\_Centrallimit2.html](http://www.chem.uoa.gr/applets/AppletCentralLimit/Applet_Centrallimit2.html). In each case, the graph on the left illustrates a sampling distribution for a set of samples of size  $n = 1$ , and gives a rough idea of the shape of the population distribution. For the middle graph the sample size has been increased to 10. In the graph on the right, the sample size is  $n = 30$ .

Notice that, *no matter what the shape of the population distribution*, as the sample size increases, the shape of the sampling distribution becomes very much like the shape of a normal distribution. Close enough, in fact, that we can use the normal distribution table to determine probabilities.



#### 4. The Central Limit Theorem

Given a population with mean  $\mu$  and standard deviation  $\sigma$ :

- 1) As the sample size  $n$  increases, or as the number of trials  $n$  approaches infinite, the shape of a sampling distribution becomes increasingly like a normal distribution.
- 2) The mean of sampling distribution,  $E(\bar{x}) = \mu$ .
- 3) The standard deviation of sampling distribution, called the standard error,  $= \frac{\sigma}{\sqrt{n}}$ .

For statistics, a sample size of 30 is usually large enough to use the normal distribution probability table for hypothesis tests and confidence intervals. For the homework exercises below, you'll use the normal distribution table to find various probabilities involving sample means ( $\bar{x}$ ).

##### Example 3

The population of fish in a particular pond is known to have a mean length  $\mu = 15$  cm with standard deviation  $\sigma = 5$ .

- a) You catch one fish from the pond. What is its expected length?

The lengths of the fish in the pond exhibit some variability—we know this because the standard deviation does not equal 0. So, while we cannot say with absolute certainty that the fish we catch *will be* any specific length, our best (educated) guess is that we *expect* the fish to be 15 cm in length. (This is why the mean is also called “expected value”.)

- b) You catch one fish from the pond. What is the probability that its length is less than 14 cm?

We cannot answer this question because we do not know the shape of the probability distribution. Specifically, since it is not explicitly stated that the distribution is normal, we cannot use the normal distribution table.

- c) You catch fifty fish from the pond and measure their lengths. What are the expected value of the sample mean ( $\bar{x}$ ) and standard error for the sampling distribution?

While the length of any single fish from the pond is not likely to be representative of the whole population, a sample of size 50 is large enough for the Central Limit Theorem to apply, and we can say how likely it is that this sample will be representative of the population.

By the Central Limit Theorem, for sample size  $n = 50$ ,

$$\text{expected value } E(\bar{x}) = \mu = 15 \text{ and standard error } = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{50}} \approx 0.7071.$$

- d) You catch fifty fish from the pond and measure their lengths. What is the probability that the average length of those fish (i.e. sample mean  $\bar{x}$ ) is less than 14 cm?

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{14 - 15}{5/\sqrt{50}} \approx -1.41, P(\bar{x} < 14) = P(Z < -1.41) = 0.0793$$

- e) You catch fifty fish from the pond and measure their lengths. What is the probability that the average length of those fish (i.e. sample mean  $\bar{x}$ ) is between 14 cm and 16.5 cm?

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{16.5 - 15}{5/\sqrt{50}} \approx 2.12, P(\bar{x} < 16.5) = P(Z < 2.12) = 0.9830$$

$$\begin{aligned} P(14 < \bar{x} < 16.5) &= P(-1.41 < Z < 2.12) \\ &= P(-\infty < Z < 2.12) - P(-\infty < Z < -1.41) \\ &= 0.9830 - 0.0793 = 0.9037 \end{aligned}$$

Example 4

A population exhibits the following probability distribution:

$X$	10	20	30	40	50
$P(X = x)$	0.4	0.2	0.2	0.0	0.2

- a) What are the expected value and standard deviation for a single randomly-chosen subject from this population?

$$E(X) = \mu = 10(0.4) + 20(0.2) + 30(0.2) + 40(0.0) + 50(0.2) = 4 + 4 + 6 + 0 + 10 = 24$$

$$\text{Var}(X) = (10 - 24)^2 (0.4) + (20 - 24)^2 (0.2) + (30 - 24)^2 (0.2) + (40 - 24)^2 (0.0) + (50 - 24)^2 (0.2) = 224$$

$$\sigma = \sqrt{224} = 4\sqrt{14} \approx 14.9666$$

- b) What is the probability that a single randomly-chosen subject from this population will exhibit a value of at most 30?

In contrast to Example 3, we *do know* the shape of this probability distribution, and can calculate an answer directly from the probability distribution.

$$P(X \leq 30) = P(X = 10) + P(X = 20) + P(X = 30) = 0.4 + 0.2 + 0.2 = 0.8$$

- c) You select a sample of 45 from this population. (sample size  $n = 45$ ). What are the expected value and standard error for the sampling distribution?

By the Central Limit Theorem, for  $n = 45$ ,

$$\text{expected value} = \mu = 24 \text{ and standard error} = \frac{\sigma}{\sqrt{n}} = \frac{4\sqrt{14}}{\sqrt{45}} \approx 2.2311$$

- d) You select a sample of 45 from this population. (sample size  $n = 45$ ). What is the probability that the sample mean is at most 30?

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{30 - 24}{\frac{4\sqrt{14}}{\sqrt{45}}} \approx 2.69, P(\bar{x} \leq 30) = P(Z \leq 2.69) = 0.9964$$

- e) You select a sample of 45 from this population. (sample size  $n = 45$ ). What is the probability that the sample mean is at least 30?

This is the complement of the answer to part d).  $P(\bar{x} \geq 30) = P(Z \geq 2.69) = 1 - 0.9964 = 0.0036$

**Note the difference between the questions asked in parts a) and b) as opposed to the questions asked in parts c), d) and e).**

Parts a) and b) asked about one single subject, randomly chosen from within the population. The probabilities given in the probability distribution applied.

Parts c), d) and e) asked about 45 subjects, randomly chosen from within the population. The sample thus created can be viewed as one of the many possible samples of size  $n = 45$  that *could have been* created from the population. That is, the sample we ended up with is part of the sampling distribution, and we were able to make use of the Central Limit Theorem.

The calculations in parts a) and b) involved  $x$  and  $P(X = x)$ , that is, they focused on one specific value.

The calculations in parts c), d) and e) involved  $\bar{x}$  and standard error, that is, they focused on the mean of a sample of size  $n$ .

Example 2 revisited – a special case

You flip a fair coin. Define a random variable  $Y = 0$  for tails and  $Y = 1$  for heads.

a) What is the expected value for a single toss?

$$E(Y) = \mu = 0(0.5) + 1(0.5) = 0.5$$

b) What are the variance and standard deviation of  $Y$  for a single toss?

$$\text{Var}(Y) = (0 - 0.5)^2(0.5) + (1 - 0.5)^2(0.5) = 0.25$$

$$\sigma = \sqrt{0.25} = 0.5$$

c) You flip a fair coin forty times (sample size  $n = 40$ ). What are the expected value and standard error for the sampling distribution?

By the Central Limit Theorem, for  $n = 40$ ,

$$\text{expected value} = \mu = 0.5 \text{ and standard error} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{40}} \approx 0.0791$$

d) You flip a fair coin forty times (sample size  $n = 40$ ). What is the probability that the average value for  $Y$  (i.e. the sample mean  $\bar{y}$ ) is less than 0.4?

$$z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.4 - 0.5}{\frac{0.5}{\sqrt{40}}} \approx -1.26, P(\bar{y} < 0.4) = P(Z < -1.26) = 0.1038$$

This Example is a special case, not quite the same as Example 3 and Example 4.

Tossing a coin, for which there are two possible outcomes, is a binomial experiment. For this type of experiment, we have usually defined our random variable as  $X =$  number of heads (successes) in  $n$  tosses of the coin. For Example 2 we have defined our random variable as a *proportion*, i.e.

$$Y = \frac{\text{number of heads}}{\text{number of tosses}} = \frac{X}{n} = p \text{ (probability of success).}$$

From the perspective of proportions, the answers above would be interpreted as follows.

a) The expected value for a single toss =  $E(Y) = p = 0.5$ .

translation: If we toss a single coin, the probability of heads is 50%.

c) You flip a fair coin forty times (sample size  $n = 40$ ).

Expected value of the sampling distribution =  $E(\bar{y}) = 0.5$ .

translation: We would expect 50% of the tosses to be heads.

**important note for the standard error:**

If we generalized the mathematics used above, we would find that the standard error for the sampling

distribution for a proportion will always equal  $\sqrt{\frac{pq}{n}}$ , in this case  $\sqrt{\frac{(0.5)(0.5)}{40}} = \sqrt{\frac{1}{160}} \approx 0.0791$ .

d) You flip a fair coin forty times (sample size  $n = 40$ ). The probability that the average value for  $Y$  (i.e. the sample mean  $\bar{y}$ ) is less than 0.4 = 0.1038.

translation: If we toss a coin forty times, there is a 10.38% chance that less than 40% of the tosses will be heads.

**important note for the z-score calculation:**

In terms of proportions, the z-score formula and calculation would be  $z = \frac{y - p}{\sqrt{\frac{pq}{n}}} = \frac{0.4 - 0.5}{\sqrt{\frac{0.5 * 0.5}{40}}} \approx -1.26$ .

For more on proportions and the corresponding sampling distribution and standard error, see Appendix E, pages 15-19 in your text.

## 5. Homework Exercises

1. In 2010, the mean score for the three sections of the SAT was  $\mu = 1509$ , with a standard deviation  $\sigma = 339$ . (source: *College Entrance Examination Board, College-Bound Seniors: Total Group Profile (National) Report, 1966-67 through 2009-10*)

- One student takes the SAT. What is her expected score on the three parts?
- One hundred students take the SAT. What is the expected value of their average score (i.e. expected value of the sample mean,  $\bar{x}$ )? What is the standard error for the sampling distribution?
- One hundred students take the SAT. What is the probability that their average score is less than 1450?
- One hundred students take the SAT. What is the probability that their average score is between 1450 and 1600?
- One hundred students take the SAT. What is the probability that their average score is above 1600?

2. An experiment has the following probability distribution:

$X$	1	2	3	4
$P(X)$	0.4	0.1	0.3	0.2

- What are the expected value and standard deviation for a single randomly-chosen value of  $X$ ?
  - You randomly select a sample of size  $n = 50$ . What is the expected value for the sample mean,  $\bar{x}$ ? What is the standard error for the sampling distribution?
  - You randomly select a sample of size  $n = 50$ . What is the probability that the sample mean is less than 2.2?
  - You randomly select a sample of size  $n = 50$ . What is the probability that the sample mean is between 2.2 and 2.5?
3. You pay \$1 to play a game in which you roll one standard six-sided die. You lose your dollar if the die is 1, 2, 3 or 4. You get your dollar back if the die is a 5, and if the die is a 6 you get your dollar back plus \$2 more (total of \$3).
- Calculate expected value and standard deviation for a single toss. (Be sure to include the dollar you pay to play the game.)
  - If you play the game 100 times, what are the expected value and standard error for the sampling distribution?
  - If you play the game 100 times, what is the probability that your average outcome will be positive? (That is, you walk away with more money than what you had before the game.)
  - If you play the game 100 times, your average winnings have a 90% probability of being below what value?
4. You flip a fair coin 100 times. Define random variable  $Y =$  proportion of heads.
- What is the expected proportion of heads for this experiment?
  - What is the standard error for the sampling distribution?
  - What is the probability that the proportion of heads is between 0.45 and 0.60?
  - What is the probability that the proportion of heads is greater than 0.60?
5. Lab results indicate that a certain drug is effective 75% of the time (success) and ineffective 25% of the time (failure). As a trial, the drug is administered to a sample of 1000 patients.
- What is the expected proportion of successes for this experiment?
  - What is the standard error for the sampling distribution?
  - What is the probability that between 710 and 770 of the patients are helped by the drug? (Remember to think in terms of proportion: i.e. 710 is what proportion of the 1000 tested? 770 is what proportion of the 1000 tested?)
  - What is the probability that at least 770 of the patients are helped by the drug?