Stat 400, section 5.4 supplement: The Central Limit Theorem

notes by Tim Pilachowski

Table of Contents

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 to test the coin to determine whether it has been 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 $E(X)$ or μ_X). The technical statistical term for this "true" mean is the "population parameter" or just "parameter". In contrast, the mean calculated from a given sample (random variable $\overline{X}$, calculated value = $\overline{x}$) is a "sample statistic" or just "statistic". The mean of the sampling distribution ***is*** the mean of the population [that is, $\mathrm{E}(\overline{X}) = \mu_X$]. 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.

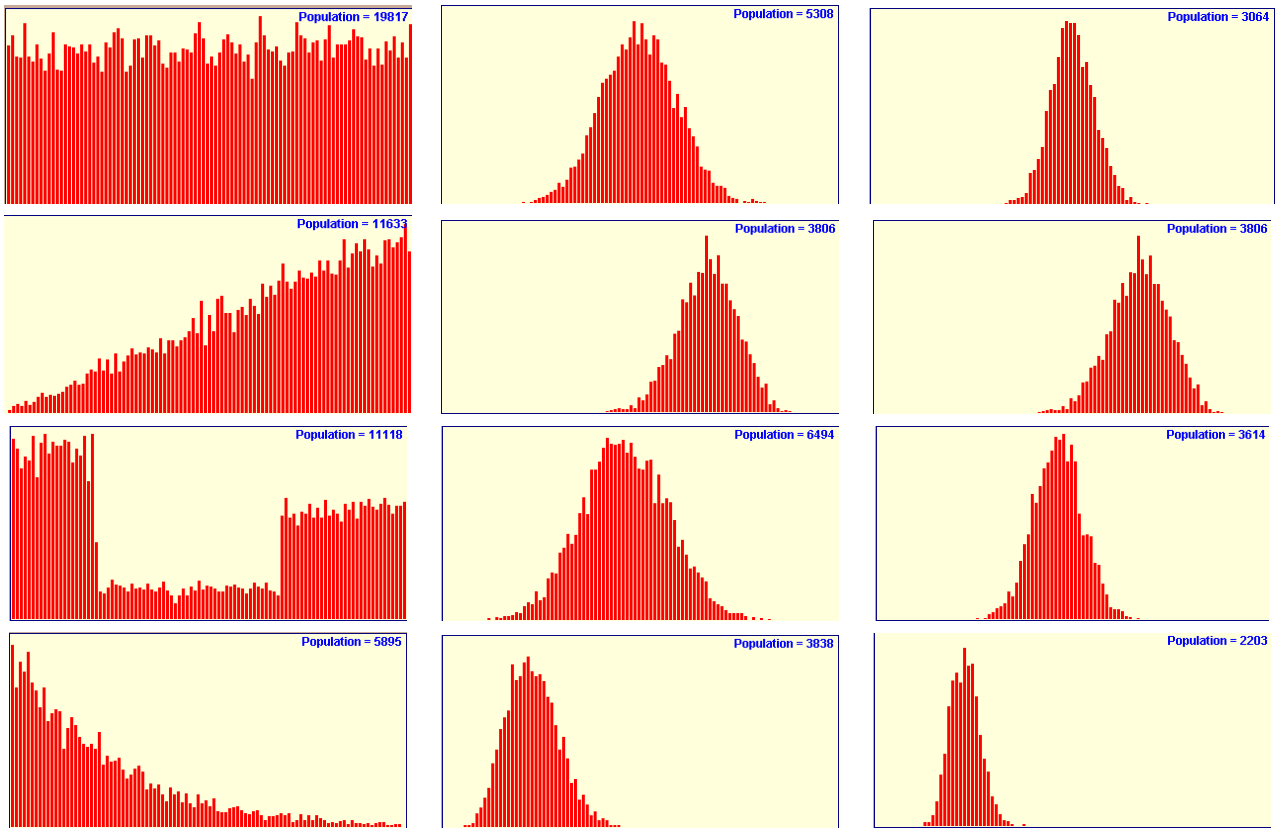
3. Practical

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 appelet found on the website http://www.chem.uoa.gr/applets/AppletCentralLimit/Appl_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 μ_X and standard deviation σ_X:

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 is the population mean, $E(\overline{X}) = \mu_X$.

3) The standard deviation of sampling distribution, called the standard error, is $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}}$.

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 (random variable $\overline{X}$).

Example 3

The population of fish in a particular pond is known to have a mean length μ = 15 cm with standard deviation σ = 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 [i.e. $E(\overline{X})$] and standard error [i.e. $\sigma_{\overline{X}}$] 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,

expected value $E(\overline{X}) = \mu_X = 15$ and standard error $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{5}{\sqrt{50}} \approx 0.7071.$

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 $\overline{X}$) is less than 14 cm?

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \quad \Rightarrow \quad \frac{14-15}{5/\sqrt{50}} \approx -1.41,\ P(\overline{X} < 14) = P(Z < -1.41) = 0.0793$$

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 $\overline{X}$) is between 14 cm and 16.5 cm?

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \quad \Rightarrow \quad \frac{16.5-15}{5/\sqrt{50}} \approx 2.12,\ P(\overline{X} < 16.5) = P(Z < 2.12) = 0.9830$$

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

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?

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

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

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

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.

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

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 $n = 45$,

expected value $E(\overline{X}) = \mu_X = 24$ and standard error $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{4\sqrt{14}}{\sqrt{45}} \approx 2.2311$

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?

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \quad \Rightarrow \quad \frac{30-24}{4\sqrt{14} / \sqrt{45}} \approx 2.69 \text{ , } P(\overline{X} \le 30) = P(Z \le 2.69) = 0.9964$$

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). $P(\overline{X} \ge 30) = P(Z \ge 2.69) = 1 - 0.9964 = 0.0036$

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 random variable X and $P(X = x)$, that is, they focused on one specific value.

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

Example 1 revisited

A club has 6 female members and 24 male members. Five club members are to be randomly selected to represent the club at a regional conference. Define a random variable X = number of women in the delegation.

The probability distribution (binomial probability) looks like this:

X	0	1	2	3	4	5
$P(X = x) \approx$	0.3277	0.4096	0.2048	0.0512	0.0064	0.0003

a) What is the expected number of women in the delegation?

$E(X) = \mu_X = 0(0.3277) + 1(0.4096) + 2(0.2048) + 3(0.0512) + 4(0.0064) + 5(0.0003) = 1$

b) What are the variance and standard deviation for this probability distribution?

$\text{Var}(X) = (0 - 1)^2(0.3277) + (1 - 1)^2(0.4096) + (2 - 1)^2(0.2048) + (3 - 1)^2(0.0512)$
$+ (4 - 1)^2(0.0064) + (5 - 1)^2(0.0003) = 0.8$

$\sigma_X = \sqrt{0.8} \approx 0.8944$

c) Why wouldn't we apply the Central Limit Theorem to Example 1?

The sample size is not large enough. We need $n > 30$. The shape of the sampling distribution is not close enough to normal to justify using the normal distribution table. Rather, we would need to recognize this as a hypergeometric distribution and use the appropriate formula.

Example 2 revisited

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

a) What is the expected value for a single toss (sample size $n = 1$)?

$E(X) = \mu_X = 0(0.5) + 1(0.5) = 0.5 = 50\%$

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

$V(X) = (0 - 0.5)^2(0.5) + (1 - 0.5)^2(0.5) = 0.25$

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

c) You flip a 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 $n = 40$,

expected value $E(\overline{X}) = \mu_X = 0.5$ and standard error $\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{0.5}{\sqrt{40}} \approx 0.0791$

d) What is the probability that the average value for X is less than 0.4?

$Z = \dfrac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \quad \Rightarrow \quad \dfrac{0.4 - 0.5}{0.5 / \sqrt{40}} \approx -1.26$, $P(\overline{X} < 0.4) = P(Z < -1.26) = 0.1038$

Example 5

A continuous random variable X has probability density function $f(x) = \dfrac{3}{8}x^2$ on the interval [0, 2].

a) What is the probability that a single randomly-chosen value of X will be greater than 1.8?

(You can think of this as a sample size $n = 1$.)

$$P(X > 1.8) = \int_{1.8}^{2} \frac{3}{8}x^2 \, dx = \left[\frac{3}{8} * \frac{x^3}{3}\right]_{1.8}^{2} = \frac{2^3}{8} - \frac{1.8^3}{8} = 0.271$$

b) What are the mean and standard deviation for this probability distribution?

$$\mu_X = \int_0^2 x * \frac{3}{8}x^2 \, dx = \int_0^2 \frac{3}{8}x^3 \, dx = \left[\frac{3}{8} * \frac{x^4}{4}\right]_0^2 = \frac{48}{32} - 0 = \frac{3}{2} = 1.5$$

$$V(X) = \int_0^2 x^2 * \frac{3}{8}x^2 \, dx - \left(\frac{3}{2}\right)^2 = \left[\frac{3}{8} * \frac{x^5}{5}\right]_0^2 - \frac{9}{4} = \frac{12}{5} - 0 - \frac{9}{4} = \frac{3}{4} = 0.75$$

$$\sigma_X = \sqrt{\frac{3}{4}} \approx 0.8660$$

c) Given a sample of size n = 40, what are the expected value and standard error for the sampling distribution?

By the Central Limit Theorem, for n = 40,

expected value $E(\overline{X}) = \mu_X = 1.5$ and standard error $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{\sqrt{\frac{3}{4}}}{\sqrt{40}} = \sqrt{\frac{3}{160}} \approx 0.1369$

d) Given a sample of size n = 40, what is the probability that the sample mean is greater than 1.8?

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} = \frac{1.8 - 1.5}{\sqrt{3/160}} \approx 2.19$$

$P(\overline{X} > 1.8) = P(Z > 2.19) = 1 - P(Z < 2.19) = 1 - 0.9857 = 0.0143$

e) There is an 80% probability that the sample mean will be below what value?

From the normal distribution table, the closest we can get to 0.8000 is 0.7995 = $P(Z < 0.84)$.

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}} \quad \Rightarrow \quad 0.84 = \frac{\overline{x} - 1.5}{\sqrt{3/160}} \quad \Rightarrow \quad \overline{x} = 0.84\left(\sqrt{3/160}\right) + 1.5 \approx 1.615$$

5. Homework Exercises

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

a) One student takes the SAT. What is her expected score (on the three parts)?

b) 100 students take the SAT. What is the expected value of their average score (i.e. expected value of the sample mean)? What is the standard error for the sampling distribution?

c) 100 students take the SAT. What is the probability that their average score is less than 1450?

d) 100 students take the SAT. What is the probability that their average score is between 1450 and 1600?

e) 100 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

a) What are the expected value and standard deviation for a single randomly-chosen value of X?

b) You randomly select a sample of size n = 50. What is the expected value for the sample mean? What is the standard error for the sampling distribution?

c) You randomly select a sample of size n = 50. What is the probability that the sample mean is less than 2.2?

d) 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 flip a coin 100 times. Define a random variable $X = 0$ for tails and $X = 1$ for heads.
   a) What are the expected value and standard deviation of X for a single toss?
   b) What are the expected value and standard error for the sampling distribution?
   c) What is the probability that the average (mean) value for X is between 0.45 and 0.60?
   d) What is the probability that the average (mean) value for X is greater than 0.60?

4. 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).
   a) Calculate expected value and standard deviation for a single toss. (Be sure to include the dollar you pay to play the game.)
   b) If you play the game 100 times, what are the expected value and standard error for the sampling distribution?
   c) 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.)
   d) If you play the game 100 times, your average winnings have a 90% probability of being below what value?

5.* A random variable X has probability density function $f(x) = \frac{1}{x}, \quad 1 \le x \le e$.
   a) Find the expected value and standard deviation for X.
   b) What is the probability that a single randomly-chosen value of X is less than 1.72?
   c) Given a sample size of 900, calculate the expected value and standard error for the sampling distribution.
   d) Given a sample size of 900, what is the probability that the sample mean is less than 1.72?
   e) Given a sample size of 900, there is a 5% probability that the sample mean is above what value?

6.* A certain drug is to be rated either effective or ineffective. Suppose lab results indicate that 75% of the time the drug increases the lifespan of a patient by 5 years (effective) and 25% of the time the drug causes a complication which decreases the lifespan of a patient by 1 year (ineffective). As part of a study you administer the drug to 10000 patients.
   a) Find $E(X)$ and σ_X for a single patient.
   b) Find the expected value and standard error for the sampling distribution.
   c) What is the probability that the lifespans of those in the study will be increased by an average of 3.55 years or more?

7.* A honeybee drone has an expected lifespan which is exponentially distributed with mean 45 days. You collect one thousand honeybee drones.
   a) What is the standard deviation for the lifespan of a honeybee drone?
   b) For an individual bee what is the probability that its lifespan will be above 47 days?
   c) What is the probability that the average lifespan for all 1000 honeybee drones will be above 47 days?
   d) There is a 10% probability that the average lifespan for all 1000 honeybee drones will be less than what age?

*Exercises "borrowed" (and in some cases slightly revised) from Justin Wyss-Gallifent.