Chapter 6: Sampling Distributions

Nathan Lutz

2022-07-14

Guiding Scenario

As part of your internship at the county jail, you have conducted observations and recorded the amount of time detainees spend in various parts of the facility, including the fitness center, the library, and the yard. The superintendent, who has now come to rely on your statistical knowledge and analytical savvy, wants to know how each of these areas are being utilized and how much time inmates tend to spend at each. You observe and record the duration in minutes of 200 different inmates' visits to the three locations, and your results are shown in Table 6.1

Table 6.1

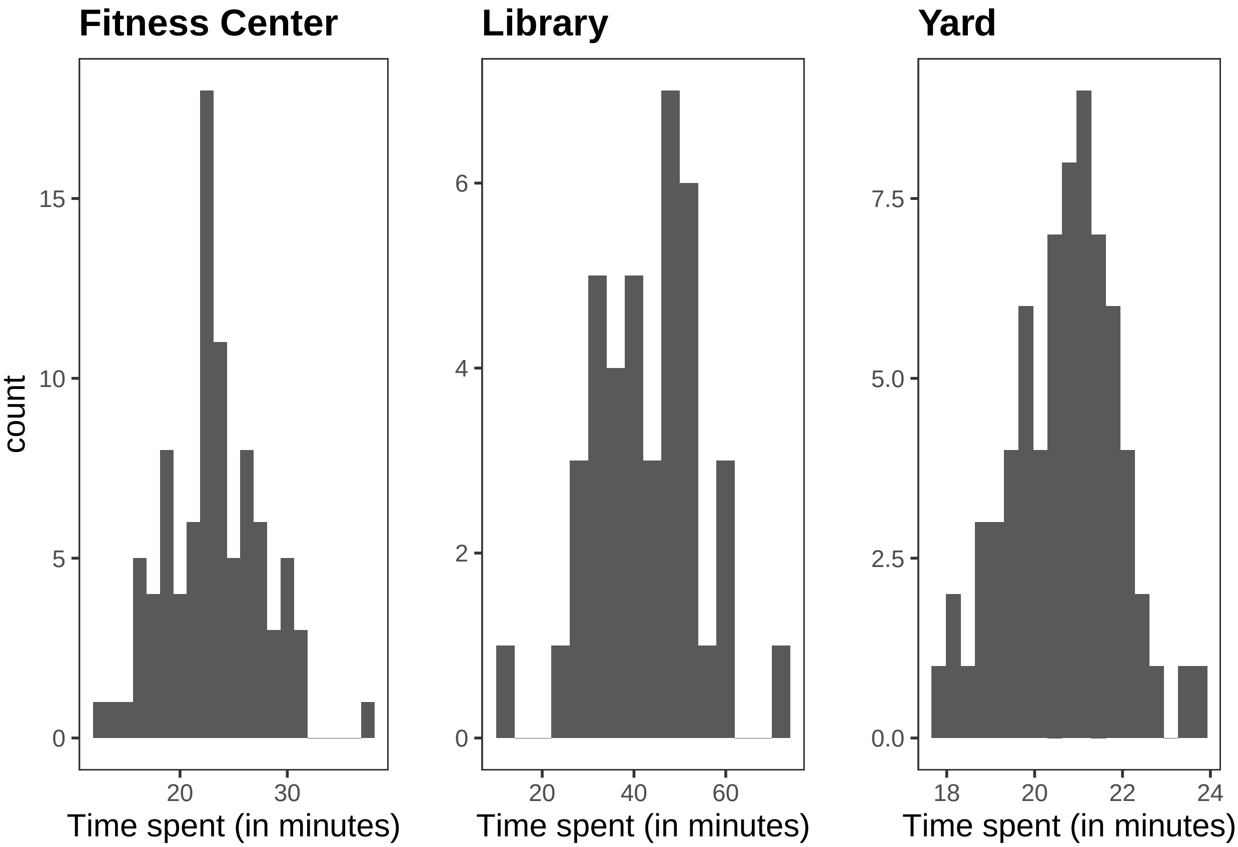
Observations of time spent in jail facilities

Location	N	Mean (minutes)	Standard Deviation
Fitness Center	90	23.6	4.3
Library	40	41.8	9.7
Yard	70	20.7	1.2

Additionally, you create histograms of your observations in each of the three locations to show the shape of the distributions of time spent in each facility. You are pleased to see that these observations follow a roughly normal distribution, and there don't seem to be any major outliers in your data. These histograms are shown in Figure 6.1 .

Figure 6.1

Sample distributions of time spent in jail facilities



The superintendent is impressed by your work, but she is unsure about how much she can trust the values you have provided. She asks you whether the mean values you recorded are a good representation of how much time all detainees spend in these areas on average, not just the 200 you happened to observe when you were there. You are stumped at first, but then you remember this exact topic was covered in your favorite undergraduate textbook. You enthusiastically open the book and turn to the chapter on sampling distributions.

Linking to Previous Chapter

The previous chapter discussed various distributions that data tend to follow. It ended by claiming the normal distribution is arguably one of the most important distributions in statistics. This chapter will introduce you to the ways we can use what we know about the properties of the normal distribution (outlined on page ___) to draw robust and powerful inferences about our data. We have already seen how the normal curve can be used to draw inferences about single values using the mean and standard deviation of the population and how we can estimate where that value would fall in the distribution of all values in the population. However, the normal curve's value goes far beyond z scores. By the end of this chapter, you will see why this bell-shaped curve seems to be everywhere you look in statistics.

Sampling Distributions

You are already familiar with the concept of a statistic drawn from a single sample and how it differs from a parameter drawn from the population. In in the example in the introduction to this chapter, you were interested in how much time people spent in the fitness center, library, and yard, but you only knew the mean amount of time spent in each of the areas by the people you observed. While this provides a good estimate of the mean amount of time spent by the entire population of detainees at the jail, you don't actually know what the true population value is. In fact, the likelihood that the mean of your sample is the exact same as the population mean is extremely small when measuring on a continuous scale. This is illustrated in the following dice rolling example.

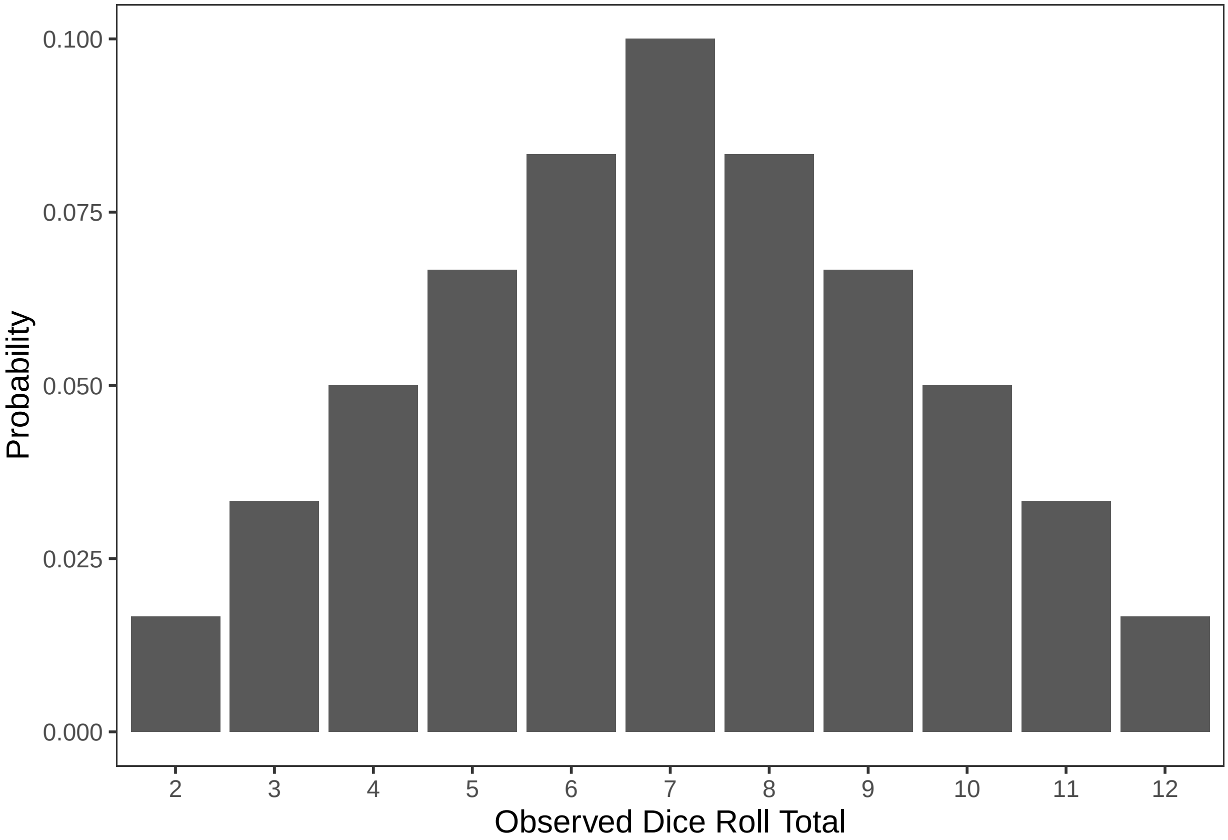
Imprecision of Samples

[INSERT STOCK IMAGE OF TWO DICE?]

Imagine you have two six-sided dice. You know that the probability of each number, one through six, is $\frac{1}{6}$ or 0.167 for each of the die. When you roll both dice, the values can range from 2 (i.e., two ones) to 12 (i.e., two sixes). The possible sums of your dice rolls have probabilities displayed in Figure 6.2 below.

Figure 6.2

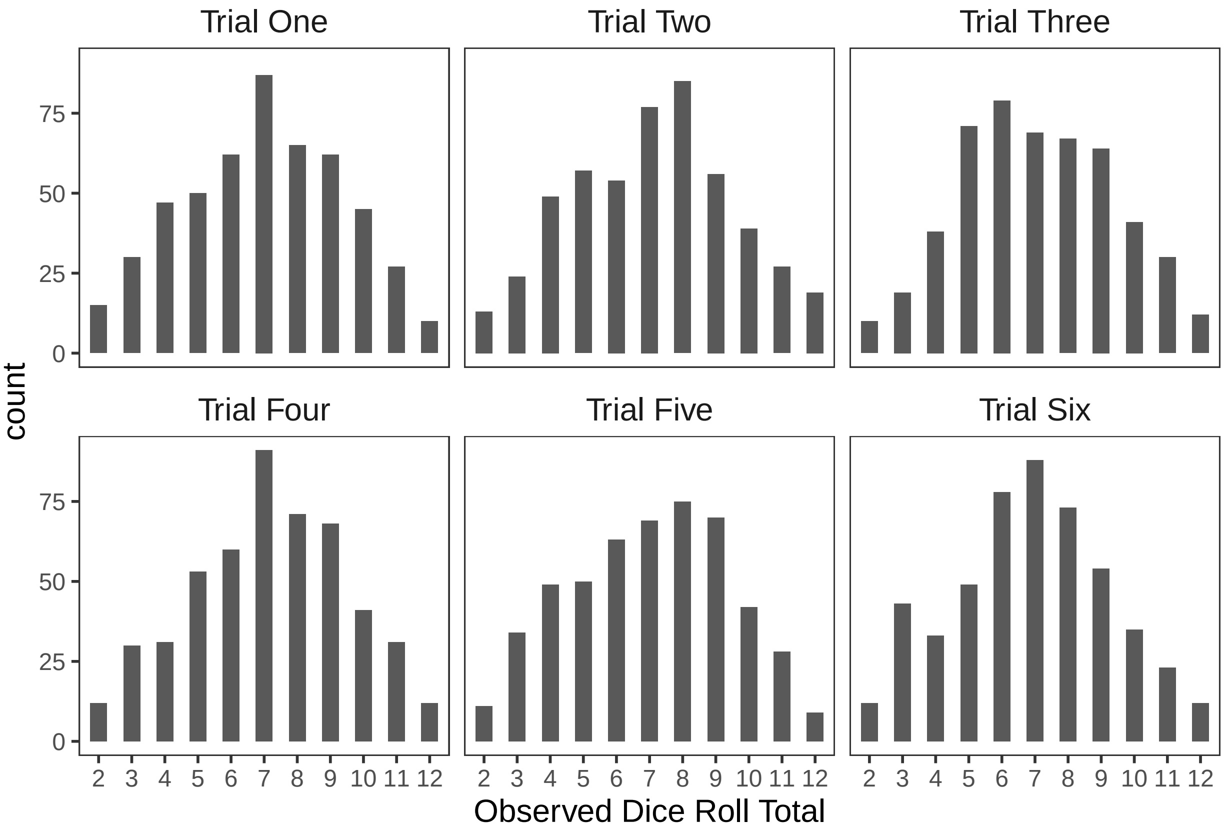
Probabilities of values observed when rolling two dice



The number seven, which can be observed with six different combinations of the two dice (i.e., one and six, two and five, three and four, and the same three combinations on opposite dice), is the most likely value. Since the distribution is perfectly symmetric around seven, it also represents the "expected value" or the average if we were to roll the two dice an infinite number of times. To save the time and energy it would take to roll two dice over and over again, we can use a computer simulation to show what might happen if we rolled these dice a large number of times. Figure 6.3 shows the results of six different "trials" of 500 dice rolls (about 4 hours of dice rolling done by a computer in a matter of seconds!).

Figure 6.3

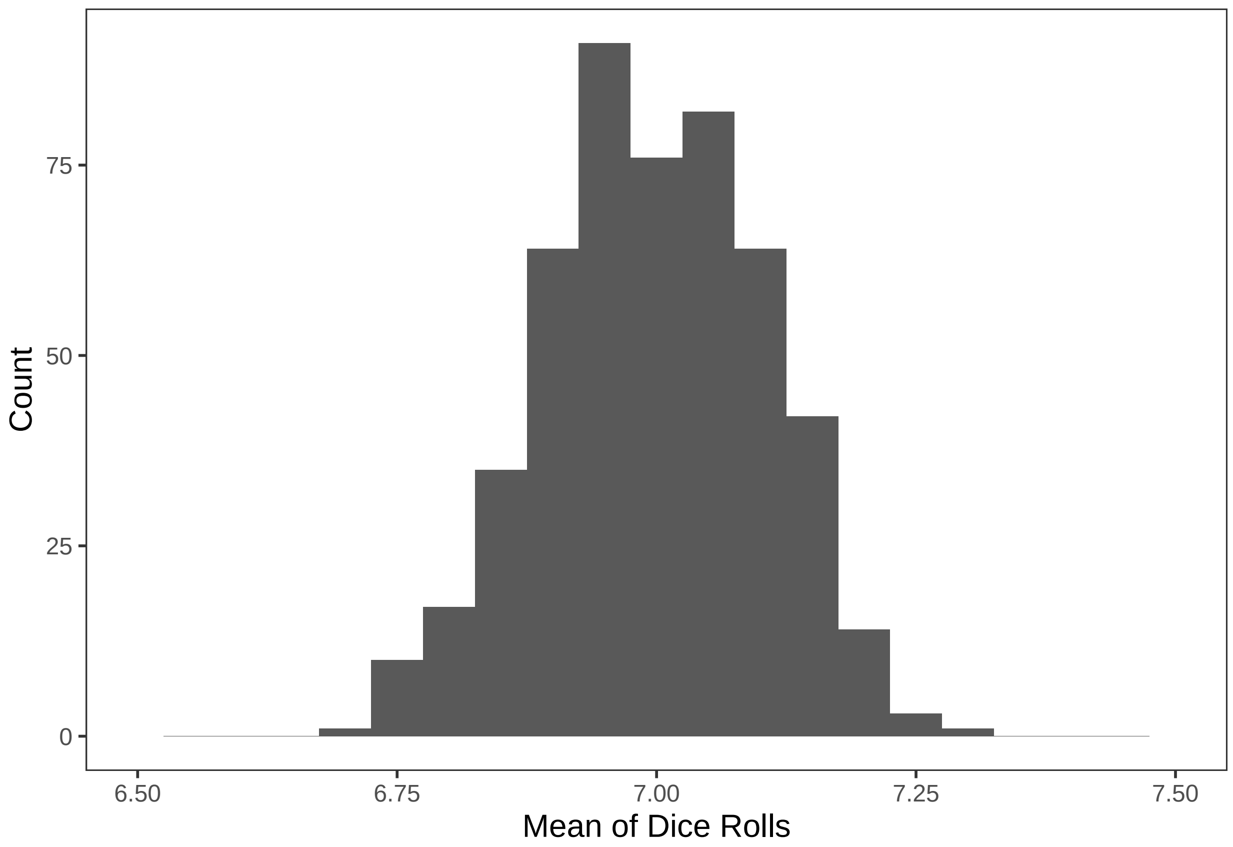
Sample distributions of six trials of 500 dice rolls



At first glance, these trials all seem to have distributions similar to what we would expect with an infinite number of dice rolls, but none of them are a perfect match with Figure 6.2. The means of these dice rolls are 6.97, 7.08, 7.07, 7.15, 6.99, and 6.86. While they are all close to our known population mean of 7.00, none of them are exactly equal to the mean. If we were to repeat this with 500 trials instead of six (about 14 straight days of dice rolling), compute the mean from each trial, and plot those means, we might get a distribution like the one in Figure 6.4.

Figure 6.4

Sampling distribution of the means of 500 trials of 500 dice rolls



Across 500 trials, each with quite large sample sizes also equal to 500, we are seeing some variability in the mean. In this simulation, the mean ranges from 6.68 to 7.3, with most of the values close to the true population mean of 7.00. The mean of the means of the 500 trials is 6.996, and the median is 6.999. What may be surprising to hear is that only 3 of the 500, i.e., 0.6%, of the trials produced a mean exactly equal to our known population mean. It is actually more likely to be wrong about the mean than it is to be exactly right! This distribution of means is known as a sampling distribution.

Distinguishing Sample Distributions and Sampling Distributions

One essential distinction in this chapter is the difference between sample distributions and sampling distributions. A sample distribution is simply the shape of some set of data that you have collected. Sample distributions are shown in Figures 6.1, 6.2, and 6.3. They represent actual observed data, and they are commonly reported in research manuscripts. Various sample distributions and their potential shapes were discussed in Chapter 5.

Sampling distributions, however, are a bit less common in practice and are typically more of a hypothetical construct. Rarely in a basic statistics course do we repeat the procedures represented by Figure 6.4. We instead use an important statistical theorem (more on that later) to create a hypothetical sampling distribution from what we observed in our single sample. A sampling distribution represents the amount we would expect our estimate of a population parameter (i.e., a statistic) to differ if we were continue to draw samples over and over again. It is important to know that we do not usually perform this re-sampling when using basic statistical methods, and we often work with single samples.

Note about Imprecision

An important note about sampling distributions is that the sampling variability we observe above does not come from mistakes made in measurement, bias, or someone trying to avoid rolling 7's (as one might on casino night). The dice rolling trials were completely random with fixed equal probabilities for each number on each die, yet we still failed to achieve the exact population mean in 99.4% of the trials. The variability we observe comes from the fact that we are only drawing a single sample from the population, and that sample may randomly deviate from the mean. These deviations are shown in sampling distributions like the one in Figure 6.4. This is important to keep in mind when reading the upcoming section on standard error, which uses the somewhat misleading term "error." This term does not imply that the person collecting the data made a mistake or that there is something wrong with the data themselves. Instead, this is an "error" that is inherent in drawing samples from a population.

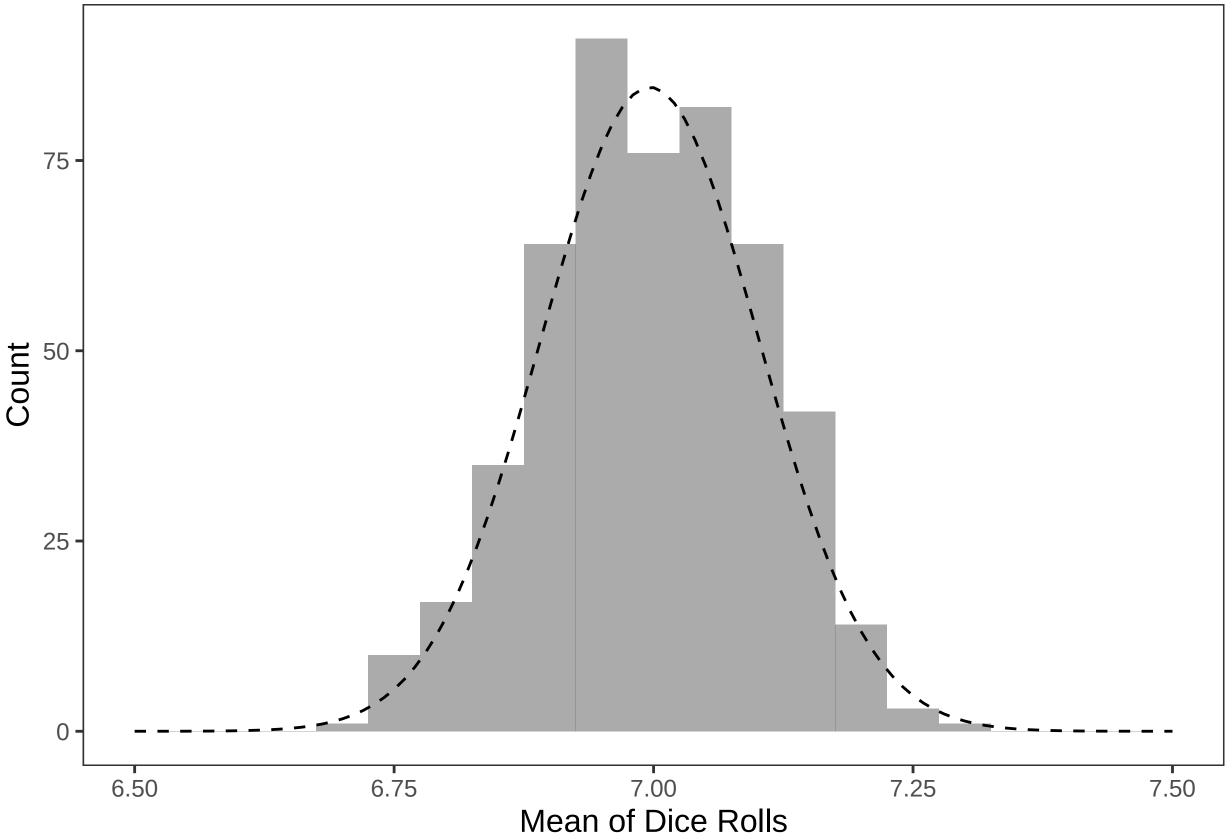
The Central Limit Theorem

If you were paying close attention, you may have noticed that the sampling distribution in Figure 6.4 has a quite recognizable shape. Having diligently paid attention to the previous chapter, your professor's lecture, and this chapter so far, you may in fact have immediately recognized the approximately normal curve that the means drawn from the 500 trials formed. This is no coincidence, and it is at the heart of virtually all statistical methods that fall into the category of "frequentist" statistics. These methods get their name from the idea that we can make estimations about the population by imagining what would happen if we drew an infinite number of samples and observed the "frequency" of the statistics we draw from each sample. These frequencies are represented by sampling distributions. In our dice example, we were able to simulate 500 trials relatively quickly and efficiently, but it would certainly be less quick and efficient to repeat most scientific studies 500 times to approximate the population distribution. For example, repeating your observations of the common spaces at the jail 500 times would take years, if not multiple lifetimes.

This is where the Central Limit Theorem saves the day. In… [CITE SOMETHING?]. We see in Figure 6.5 that a normal curve with a mean and standard deviation equal to the mean and standard deviation of the means of all of our dice roll trials fits almost perfectly over our histogram.

Figure 6.5

The means of 500 trials of 500 dice rolls with the normal curve overlaid



It should be noted that the normal curve above is not the standard normal curve that we discussed in the previous chapter but is instead just a normal curve. If we were to standardize the means from our dice roll trials, i.e., subtract the mean from each value and divide by the standard deviation, the standard normal curve could be overlaid in the same way.

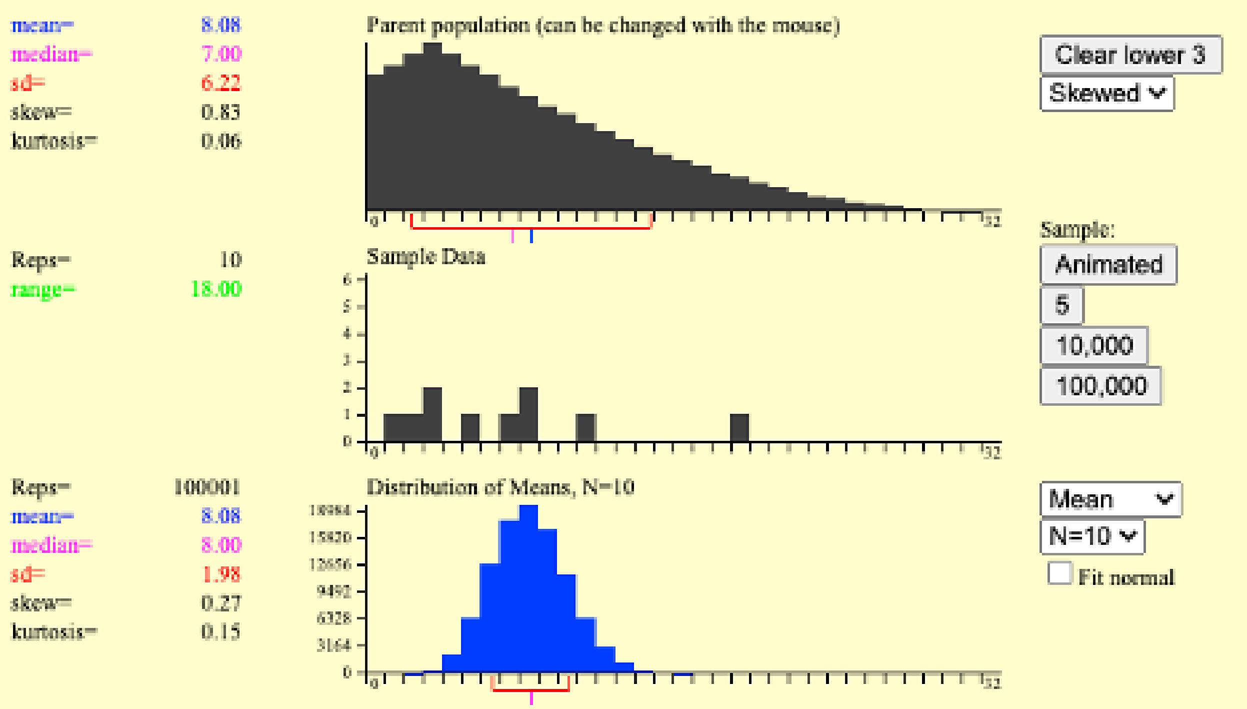
As you learned in Chapter 5, there are many advantages to knowing that something is normally distributed. In our dice roll example, we can use our mean of 6.996 and our standard deviation of 0.105 from our sampling distribution to estimate the probability of observing various means when drawing similar 500-roll samples. For example, we would expect that about 68.3% of the means would fall between 6.89 and 7.1, or within one standard deviation of the mean. We would also expect that about 99.7% of the means would fall between 6.68 and 7.31, or within three standard deviations of the mean. Based on just two values, the mean and standard deviation, we can make statements about how likely any number of values for the true mean are.

For our dice rolling example, we did not actually need to create our own sampling distribution, because we know the mean and standard deviation of the population of dice rolls. In cases when the mean and standard deviation are unknown, knowing about how the mean might be distributed is extremely helpful, and the beauty of the Central Limit Theorem is that we only need a single sample to make guesses about this distribution (more on that later).

One of the most important properties of the Central Limit Theorem is that it holds true no matter what the shape of the drawn sample is. This means you do not need to have normally distributed sample data for the Central Limit Theorem to apply. Further, with a large enough sample, even the underlying population distribution can be non-normal. If you are having a hard time believing this, the applet at this website may help convince you. This interactive applet allows you to draw from various parent population distributions then draw means from those samples one at a time, five at a time, or even 100,000 at a time. Figure 6.6 shows an example of a positively skewed parent population, a sample draw of ten values from that population, and a distribution of 100,001 means from similarly drawn samples. You can see that, although the single sample with N = 10 does not appear to be normally distributed, the sampling distribution (or Distribution of Means) is much closer to a normal distribution. This applet allows you to simulate in the same style as our dice roll example on other distributions and see the power of the Central Limit Theorem.

Figure 6.6

Screenshot from onlinestatbook sampling distribution applet



mean= 8.08
median= 7.00
sd= 6.22
skew= 0.83
kurtosis= 0.06

Reps= 10
range= 18.00

Reps= 100001
mean= 8.08
median= 8.00
sd= 1.98
skew= 0.27
kurtosis= 0.15

If you are starting to become concerned about the rigorous process of developing sampling distributions by re-sampling and re-calculating statistics over and over again, you should know these processes are only being done as an example. This concept, while complex conceptually, is actually quite simple to apply in practice. The next section will provide you with the final piece of the puzzle that will allow you to create sampling distributions using just a single sample.

Standard Error

In the previous chapter, we learned that any normal distribution can be characterized by just two values, the mean and the standard deviation. If you know these two values, you can approximate the shape of the distribution using the normal curve. The mean represents the central point of the normal distribution. In our dice rolling example, we saw that the mean of our distribution of means (6.996) was almost identical to the true population mean of 7.00. When something is perfectly normally distributed, the mean, median, and mode are all equal. We also saw this in our example, as the median observed value was 6.999. Calculating our best estimate of the population mean based on the sample is easy; we simply use the sample mean to approximate the mean in the population. If we did not know that our population mean was 7.00, any of the sample means, which ranged from 6.68 to 7.3, would have been reasonable estimates of the true population mean. In most real-world applications, the population mean will be unknown, and you will use the sample mean to estimate this value. Similarly, you will not know how far your sample estimate is from the true population mean. What we have not yet covered is how to estimate the standard deviation of the normal curve characterized by the Central Limit Theorem (i.e., the sampling distribution).

The standard deviation represents the “spread” of the values and how much they tend to differ from the mean on average. We know that the population standard deviation for all dice rolls is 2.42, but our sampling distribution had a standard deviation of 0.105. Is this a flaw in the way we sampled? This sampling standard deviation is clearly not a good estimate of the population standard deviation, but should it be?

Unfortunately, the standard deviation of the sampling distribution cannot simply be estimated using the standard deviation of a single sample, but it actually would not make much sense for these values to be the same. The population standard deviation tells us how much all values in a population deviate from the population mean, but we are interested in how much the sample means differ from the true population mean. Only in cases in which we draw samples with N = 1 would we expect the standard deviation from the population and the standard error to be the same. When we draw a mean from a sample with any other size, however, we are reducing the information from all the values in the sample into a single value, which is one advantage of calculating the mean in the first place. This single value estimation makes it less likely that we will see values in the tails of the population distribution as our sample size increases. In the dice rolling example, getting a mean of 2.0 with a sample size of even something as low as N = 5 would require rolling two "ones" five times in a row, an extremely rare outcome with a probability of .000000016. We need to alter the standard deviation to achieve an estimate of the standard error.

We can manipulate the population standard deviation or even the standard deviation we draw from a single sample to formulate an estimate of this value at the "sampling" level. Before we can discuss how to calculate this value, we need to cover a bit of statistics history and talk about a county fair, a weight-guessing contest, and the importance of large samples.

The Law of Large Numbers

[STOCK IMAGE OF A BULL?]

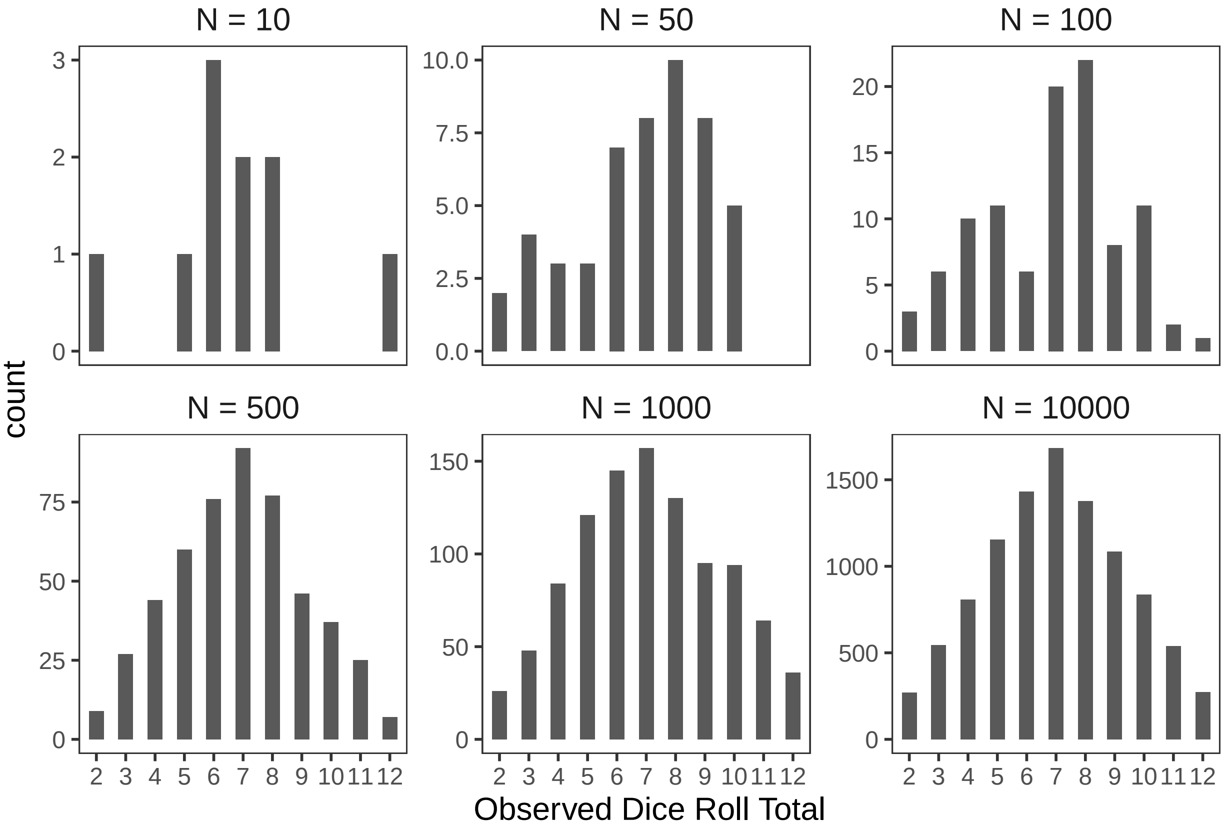
In 1907, Francis Galton wrote a one-page summary of a study titled Vox Populi, Latin for "the voice of the people," that was published in Nature (Galton, 1907). He had been given 787 tickets entered by farmers, butchers, and non-experts that included the guess of the weight of a bull at a county fair. The entrants paid a sixpenny fee and were promised a prize if they guessed correctly. While the guesses ranged in their accuracy, and some were quite far from the true weight of 1198 pounds, the "middle-most estimate" was 1207 pounds, only off by 9 pounds, or less than 1% of the weight of the bull. Galton observed that the estimates were nearly normally distributed around the true weight of the bull, i.e., the "population" mean, and that most of the guesses were quite close to the true weight of the bull. Galton's "future directions" section simply called for better record-keeping at cattle shows, but the actual implications of his findings transcend county fair policies and represent an important principle in statistics.

While Galton's bull-weighing anecdote is entertaining, it was not the first time someone realized that the "voice of the people" may be stronger than the "voice of a person." In a somewhat less entertaining, equation-filled proof, Jacob Bernoulli (1713) showed how an experiment's accuracy improves as the number of trials increases. This somewhat intuitive concept, titled "Bernoulli's theorem," was expanded upon by the French mathematician, S. D. Poisson (1837), whose "la loi des grands nombres" or "law of large numbers" continues to be cited today. The mathematics behind these theorems are beyond the scope of this textbook, but their conclusions are relatively easy to understand.

The law of large numbers states that the statistics drawn from a sample approach the true population parameters as the sample size approaches infinity. You probably entered this course with the understanding that a scientific experiment with 500 participants gives us more useful information than an experiment with 5 participants, and this law formalizes that intuition. If we return to our example of dice rolling, we can see how larger sample sizes lead to estimates that better represent the population. Figure 6.7 shows an example of six different simulated trials of dice rolls with various sample sizes. As we increase our sample size to $10,000$, we have a distribution that almost perfectly replicates what we would expect the population to look like.

Figure 6.7

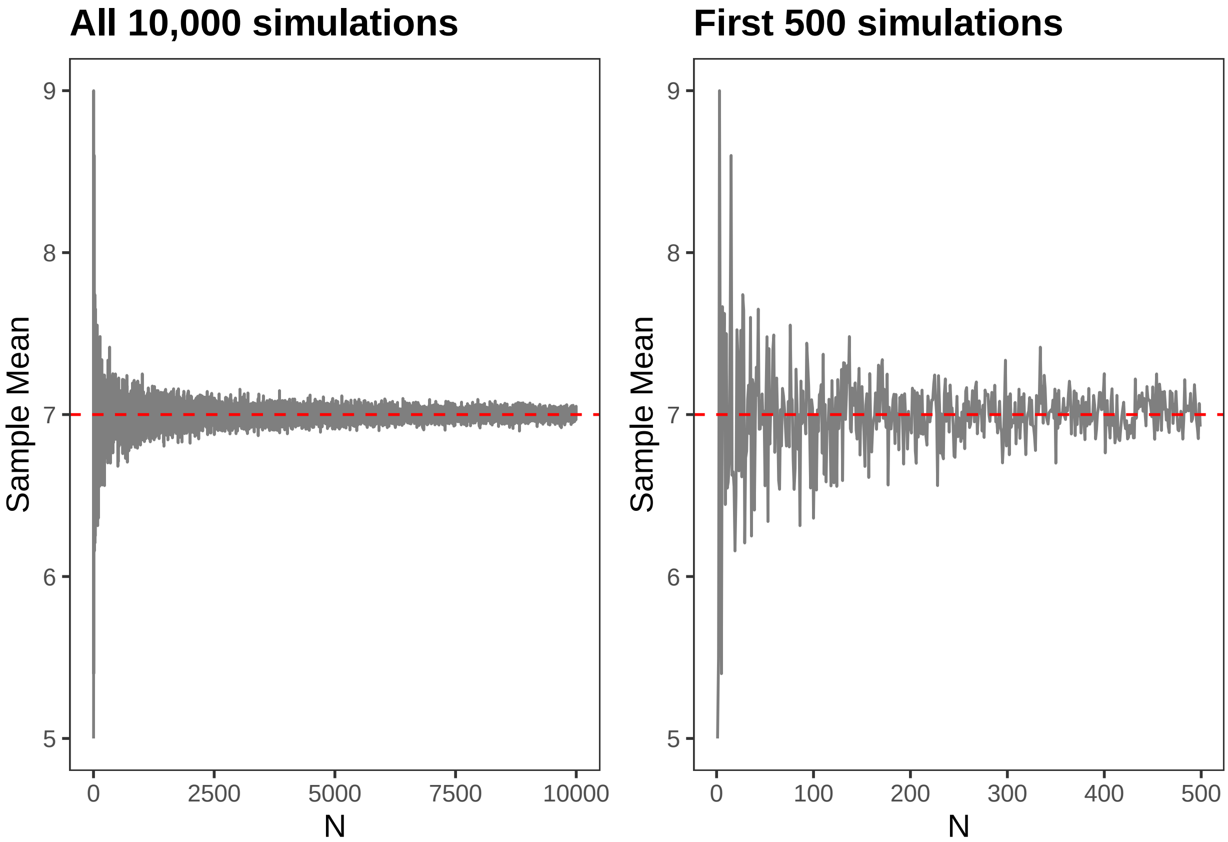
Sample distributions of dice rolls with various sample sizes



It follows that the statistics drawn from these samples also improve in their accuracy as we increase the sample size, which is supported by the law of large numbers. In the samples shown in Figure 6.7 , the computed means are 6.7, 6.86, 6.87, 6.87, 7.08, and 6.99 respectively. As sample size increases, we approach the true expected mean of 7.00. We can see this concept in Figure 6.8 , which draws random samples of dice rolls of increasing size and displays the average. This average converges to the true population average of 7.00 as the sample size increases. The plot on the left shows all 10000 simulations with N ranging from 1 to 10000, while the plot on the right shows the first 500 simulations.

Figure 6.8

Sample means converge to population mean with increasing sample size



Although the shapes of these distributions may not perfectly replicate the population distribution, you can see we do not need an extremely large sample to get a good estimate of the true population mean. In this example, we are close to a mean of 7.00 even with samples as small as 100. After a certain point, the accuracy of our estimates appear to improve only at a marginal rate, but they are indeed improving. We will use the understanding that sample size affects the accuracy of our estimates of sample means in the calculation of the standard error.

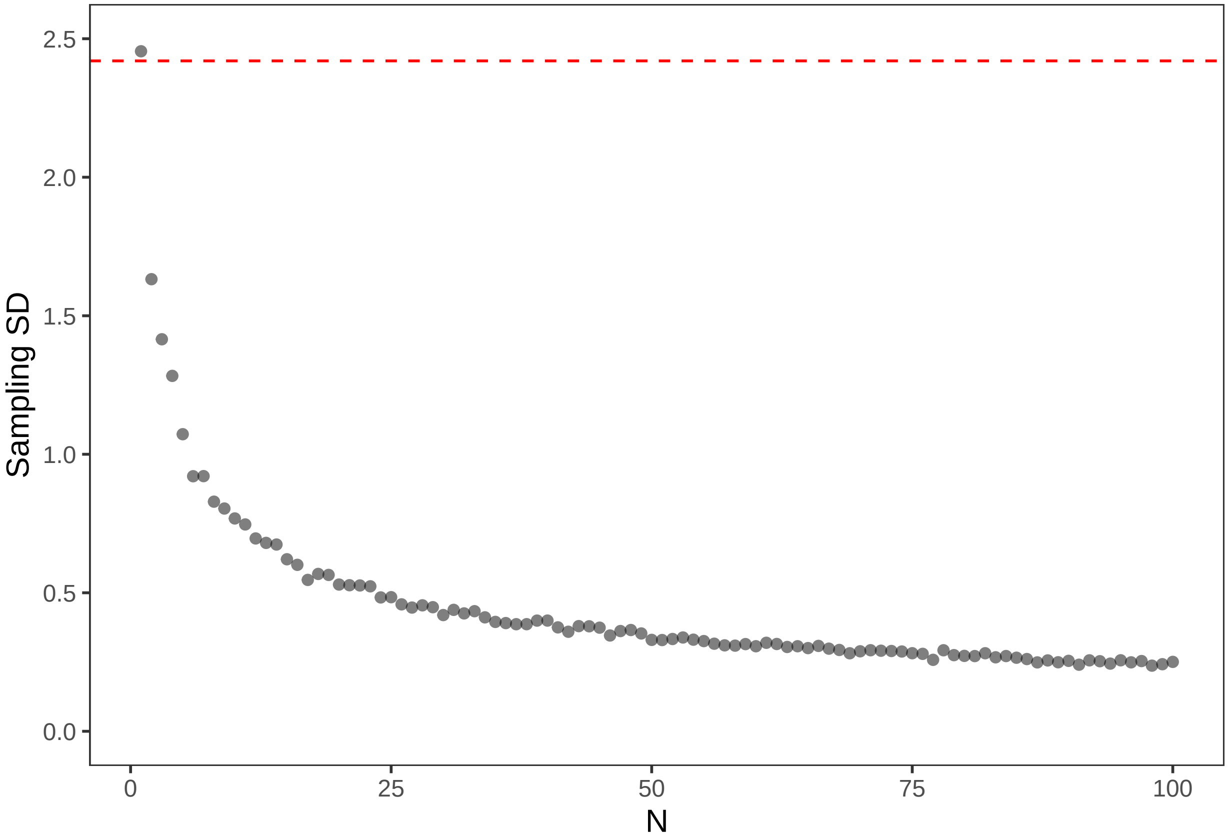
Calculating the Standard Error

We can use the dice rolling example, a case in which the population mean and standard deviation are known, to illustrate the sampling distribution that has the most straightforward calculation. Across the population of all possible dice rolls, we know the mean is 7.00 and the standard deviation is 2.42. As stated before, if our sample size was $N = 1$, these numbers would also characterize the sampling distribution's mean and standard deviation, i.e., the standard error of the mean. Larger samples will require additional information. To calculate the standard error for a sample with $N > 1$, we will need to use our understanding of how the sample size affects the standard deviation.

In Figure 6.8 , we were able to show that increasing sample sizes lead to better overall estimates of the mean. Figure 6.9 confirms that the same pattern does not hold for the standard deviation. With increasing sample size, we observe a quick drop-off in the variability of the means computed in 500 trials of varying sample sizes.

Figure 6.9

Sampling standard deviations do not converge to population standard deviation with increasing sample size



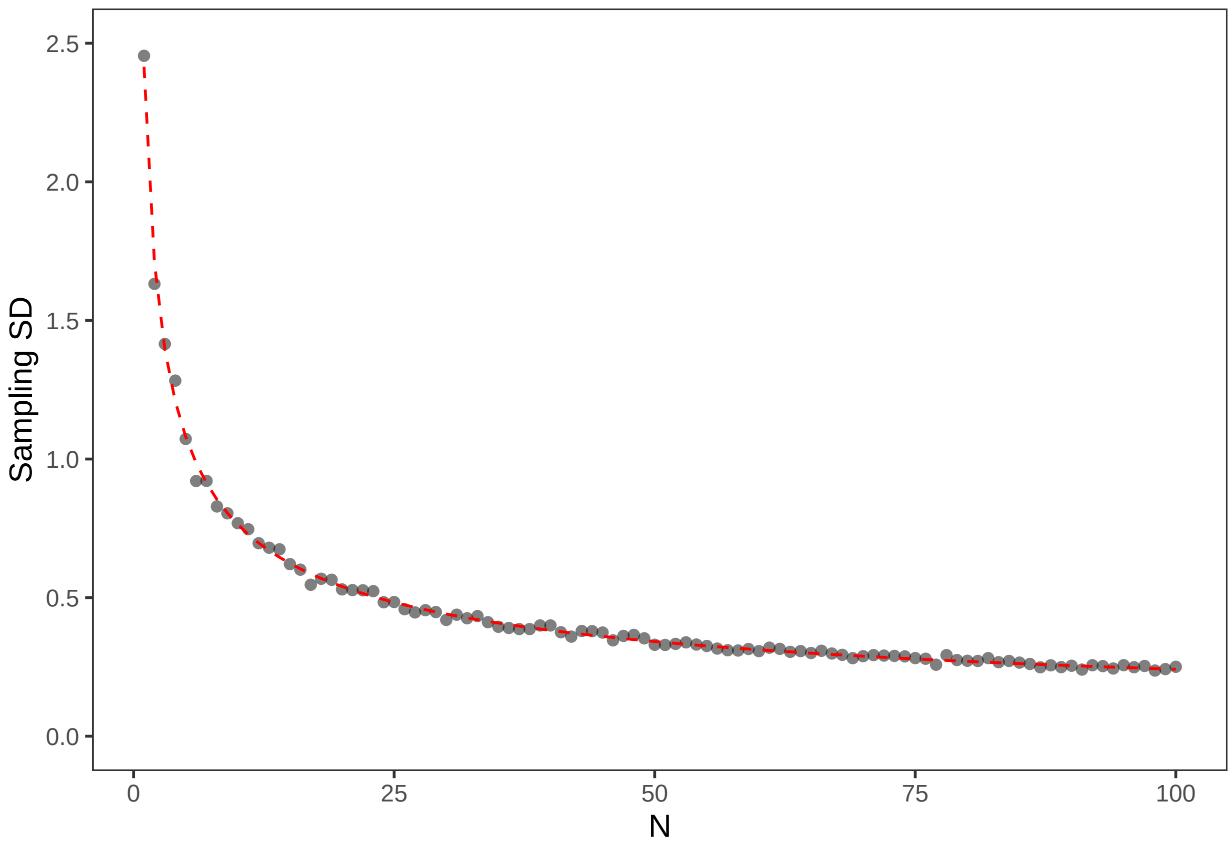
The decrease in the variability of these estimates is directly related to sample size, and it can be calculated simply by dividing the population standard deviation by the square root of the sample size. Equivalently, you can take the square root of the population variance divided by the sample size. This creates an estimate such that the variability in the estimates of various statistics approaches zero as N approaches infinity, or the estimates of these statistics will exactly equal their related population parameter as the sample size approaches the size of the entire population.

[MATHEMATICAL PROOF HERE?]

Sure enough, when we use the equation above to compute the standard error of our sample of 500 dice rolls by dividing the population standard deviation 2.42 by the square root of 500, we get 0.108, which is close to our simulated standard error of 0.105. We can apply this understanding to Figure 6.10 and see that the curve created by computing expected standard errors almost perfectly fits the curve of our simulated standard errors.

Figure 6.10

Sampling standard deviations with calculated standard errors overlaid



The understanding that this computation results in almost identical estimates to what you would get if you simulated samples over and over again shows why these computed estimates are preferred to gathering information from the entire population or to drawing repeated samples. A single sample can give us a reasonable estimate of what would happen if we repeated our sampling methods an infinite number of times.

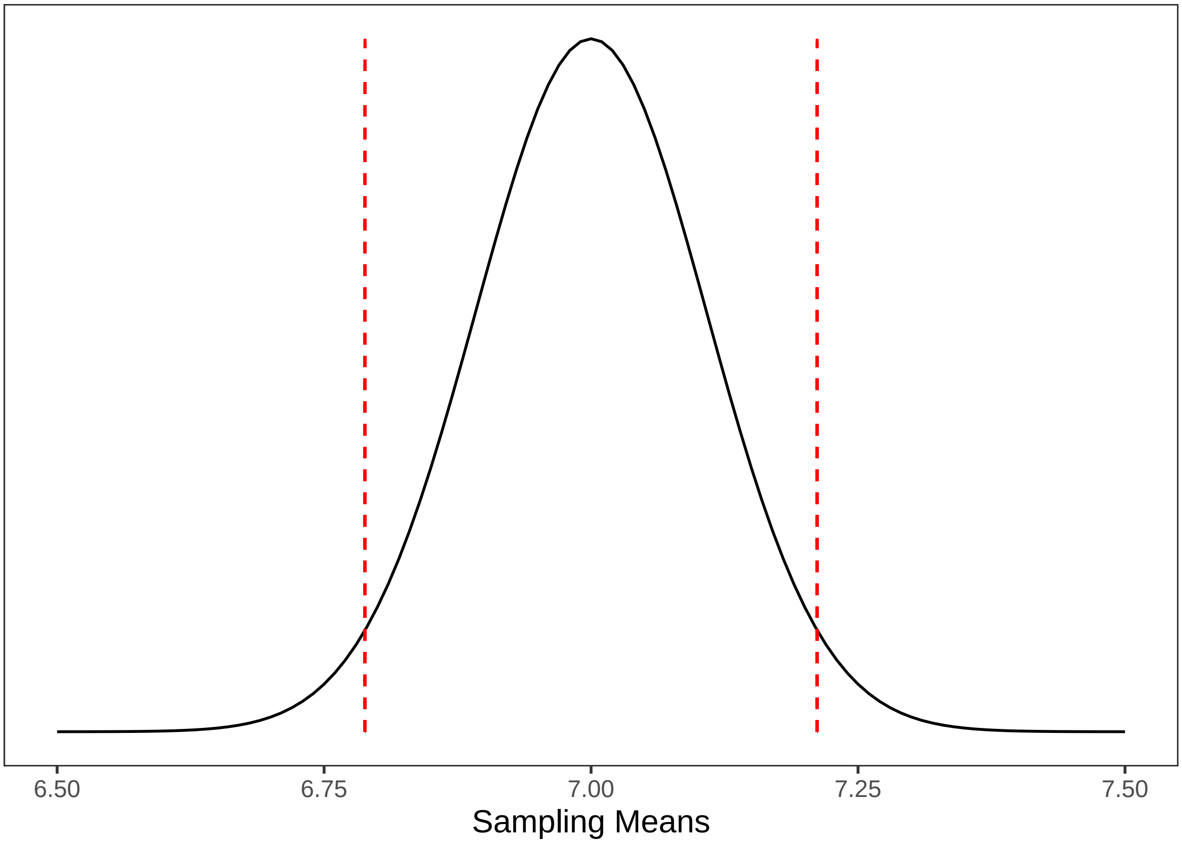
When we do not know the population values and are working with unknown population parameters, we need to take a slightly different approach. This approach will be covered in detail in Chapter 7. Instead of z distributions, we need to use the slightly modified t distribution when population parameters are unknown.

Confidence Intervals

As mentioned previously, we can leverage the power of sampling distributions to make statements about how we would expect the mean to deviate upon infinite re-sampling. One common way this is done in practice is to compute "confidence" intervals that tell us the lower and upper bounds that would create a distribution that captures a given proportion of means from the sampling distribution. In our dice rolling example with a sample size of 500, we can use our mean of 7.00, our standard error of 0.108, and our knowledge of the normal curve to determine an example of what these bounds may be. Like we did in Chapter 5, we can multiply 0.108 by ±1.96, the values at the 2.5^{th} and 97.5^{th} percentiles of the normal curve, to get our 95% confidence interval. We add these values to 7.00 to get our confidence interval bounds of 6.79 and 7.21. These bounds are shown along with the sampling distribution for 500 dice rolls in Figure 6.11 .

Figure 6.11

95% confidence interval for the means of 500 dice rolls



Computing a “confidence” interval in this scenario is only here as an example. Obviously, this interval is not useful to us in this case, since our “100% confidence interval” when the mean is known is $[7.00, 7.00]$ with a given population mean of 7.00. There is a slight difference in the way confidence intervals are interpreted when the population mean and standard deviation are unknown, but the computation is similar. This distinction will be discussed in depth in Chapter 7.

Revisiting the Opening Example

Equipped with the information from above, you are now prepared to answer the superintendent’s question. We will get into the exact calculations of these values in the next chapter, but the interpretations of the standard errors and confidence intervals remain the same. These values have been added to Table 6.1 and are shown in Table 6.2.

Table 6.2

Observations of time spent in jail facilities (with standard error)

Location	N	Mean (minutes)	Standard Deviation	Standard Error of the Mean	95% CI of the Mean
Fitness Center	90	23.6	4.3	0.46	[22.71, 24.49]
Library	40	41.8	9.7	1.55	[38.76, 44.84]
Yard	70	20.7	1.2	0.14	[20.42, 20.98]

You explain to the superintendent that the mean amount of time spent in each area is about 24, 42, and 21 minutes, but that if we were to repeat these observations over and over, those means would range from 22.7 to 24.5, 38.8 to 44.8, and 20.4 to 21.0 in the vast majority of the repetitions.

Summary

[ADDED WHEN CONTENT IS FINALIZED]

References

Bernoulli, Jakob (1713). Ars conjectandi: Usum & applicationem praecedentis doctrinae in civilibus, moralibus & oeconomicis (in Latin). Translated by Sheynin, Oscar.

Galton, F. (1907), Vox populi, Nature, 75, 450-451. doi: 10.1038/075450a0

Poisson, S. D. (1837). Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés (in French). Paris, France: Bachelier, pp. 139–143 and pp. 277