SDI: A Program to Illustrate the Concept of a "Sampling Distribution"

Jeff Miller

July 5, 2008

Contents

Overview

This document accompanies the Sampling Distribution Illustrator (SDI), a computer program designed to help illustrate the statistical concept of a "Sampling Distribution". Section 1 of this document—"The Tutorial"—presents the same information shown in the program's on-line tutorial. Section 2 presents supplementary information and demonstrations.

1 The Tutorial

This section contains the same information shown in SDI's tutorial. In fact, it is intended that you read this section while running the program, so that you can see the program's graphs, interact with the program to run computer simulations, and see the results produced by the program.

1.1 Introduction

The concept of a "sampling distribution" is one of the most important concepts in statistics, but it is also one of the most difficult.

This program was designed to help you learn about this concept.

The program uses computer simulation to show you how a sampling distribution could be constructed.

This tutorial will start by presenting some background information about populations, random samples, and summary statistics, all of which is needed to understand what a sampling distribution is.

1.2 Populations and Samples

Researchers take random samples in order to learn about populations.

The population is the set of all individuals who potentially *might be* included in a study.

The sample consists of those individuals who actually *are* included in the study.

Researchers study samples rather than whole populations because it is too much work to study populations.

1.3 Population Probability Distributions

A population can be represented with a graph like the one shown in SDI's upper left panel.

The "Dependent Variable" axis shows the different possible values that the researcher might obtain when measuring individuals on some characteristic of interest (e.g., height, measured in cm).

The "Probability" axis shows how often each value occurs within the population.

The probability curve is highest over the most common values of the dependent variable; lots of individuals in the population have those values.

For example, in this population the most common height is around 150 cm.

The probability curve is lower over less common values of the dependent variable, meaning that the population has fewer individuals with those values.

Many populations have the bell-shaped distribution shown here, which is called the "normal" distribution,. With this distribution, most individuals have values near the middle (where the probability curve is relatively high), whereas fewer individuals have especially higher or lower values.

Some populations have shapes quite different from this normal shape, though.

1.4 Demo: Random Sampling

This program simulates the process of taking random samples. To see that, click on the "One Sample" button in the window to the left. (Do it now and watch what happens.)

Individuals are sampled randomly from the population, one at a time. Each sampled individual is shown briefly as a small red square on the population distribution.

Then, this randomly selected individual is added into a tabulation of the sample shown in the bottom window.

Sampling stops when SDI has sampled the required number of individuals, called "N".

Click the "One Sample" button several more times to see the program take more random samples.

Notice that randomly sampled individuals mostly tend to come from the middle of the distribution, where the probability is higher. This is because most individuals have "medium" values on the dependent variable with this normal distribution.

1.5 Sample Frequency Distributions

The lower window on the left depicts the sample results, using the same type of graph that was used to represent the population.

Again, the horizontal axis shows the different values of the dependent variable, and the vertical axis shows how often each one was observed, now called its "frequency".

Because of the randomness of random sampling, the frequencies in the sample do not usually match exactly with the true probabilities in the population.

1.6 Statistical Inference

Researchers use "statistical inference" to draw conclusions about the whole population based on a random sample.

These conclusions rely on the idea that the sample distribution looks a lot like the population distribution.

But such conclusions are not 100% guaranteed.

The problem is that a random sample might look very different from the population, just by chance.

The key question for statistical inference is: "How much might the picture for our sample differ from the true picture for the whole population?"

1.7 Demo: Sampling Variability

Within this program, you can use computer simulation to get an idea of how different samples tend to be from their populations.

Just look at a lot of randomly generated samples, and see how well they tend to reflect the population.

Click again on SDI's "One Sample" button to see a new random sample from the population distribution shown in the upper left window.

Visually, compare the sample with the population and try to form an intuitive overall impression of how similar or different they are. For example, do the measurements cover more or less the same range of values? Are they centered at about the same point? How similar are their shapes?

Now click on the "One Sample" button again to look at another sample. Overall, how accurately does this new sample reflect the population? For that matter, how well does it match up with the previous sample?

If you repeat this process and look at 10–20 samples, you will begin to see how different samples can be from their population.

With $N = 10$, you will probably agree that they can sometimes be quite different. These population versus sample differences are inherent in the randomness of random sampling.

The term "sampling variability" is often used to refer to the fact that samples vary randomly from the population and from one another.

1.8 Demo: Effect of Sample Size

Large samples tend to give a better description of the population than small samples.

This is just because there is less sampling variability with larger samples.

You can easily see this using computer simulation.

You have already looked at some samples of size $N = 10$,

Now, use SDI's "Choose Sample Size" menu option to change to larger samples, like $N = 100$.

After you change the sample size, click on the "One Sample" button several times to look at some larger samples.

(You can use the speed bar below this window to speed up the sampling.)

If you compare each of these new (and larger) samples with the population, you will probably see that samples of $N = 100$ tend to be much more like the population than were the samples of $N = 10$.

If you have enough patience, look also at some samples of $N = 1000$. You will see that they tend to give an even better picture of the population.

If you have even more patience, look at some samples of $N = 10,000$ or $N = 100,000$. You will see that they tend to give almost perfect pictures of the population.

Statisticians sometimes express this by saying that the sample frequency distribution “converges to” the true population distribution as the sample size increases.

1.9 Sample Statistics

In addition to looking at sample frequency distributions as pictures of their results, researchers also generally summarize their results with one or two overall values, each of which is called a “statistic”.

SDI’s upper right panel illustrates the computation of the summary statistic for each sample. The most common summary statistic is the average or “mean” value. Click on the “One Sample” button now, and SDI will take another random sample and show you the computation of its mean.

Naturally enough, researchers look at the sample mean to make inferences about the population mean. Again, though, the inferences are not 100% guaranteed, because the mean of the random sample may not exactly match the population mean.

Many other summary statistics besides the mean are also used for various purposes, but they all suffer from this same problem: The sample value may not match the population value.

1.10 Demo: Sampling Variability of the Mean

Because random samples differ from one another, they also tend to produce different values of the mean (and any other summary statistic).

For example, you can easily see how sample means differ from one another across many samples.

Again use the “One Sample” button to generate a series of random samples.

Look at each sample’s mean, which is computed in the upper right panel.

You will see that the sample mean changes from sample to sample, just as you previously saw that the sample frequency distribution picture (bottom left window) changed from sample to sample.

These random changes in the sample means are called “sampling variability of the mean”.

Again, this sampling variability is an inescapable part of random sampling, and it would be present for any sample statistic (not just the mean).

1.11 Sampling Distributions (finally!)

This is where the concept of a sampling distribution comes in.

A sampling distribution is a tabulation showing how the values of a sample statistic (like the mean) vary across different samples.

SDI illustrates this idea by showing how the sampling distribution could be built up by taking sample after sample from the population.

The summary statistic (e.g., mean) is computed for each sample, and each sample’s statistic is added in to a tabulation in the lower right panel.

This tabulation is a “sampling distribution”.

Click the “Repeating Samples” button now, and you can watch the program build up the “sampling distribution for the mean” using computer simulation.

Across lots and lots of samples, you can see the graph converging toward the true shape of the sampling distribution. (Click the “Stop” button when you have seen enough.)

Of course you don’t get a good idea of the true sampling distribution from just a few samples; you would need a lot of samples for that.

Try ticking the “Warp Speed” box to see the sampling distribution build up much faster.

2 Demonstrations and Further Information

You can use SDI for many different demonstrations that will help you to understand what goes on with random samples and to see what sampling distributions look like. This part describes some of those demonstrations.

2.1 Demo: The Central Limit Theorem

One of the most important facts in statistics is the "central limit theorem" (CLT). Somewhat informally, the CLT says that the sampling distribution of the mean always approaches the shape of a normal distribution as the sample size increases, regardless of the shape of the population of individuals. In other words, the *means* of many large samples always vary approximately normally, regardless of the original distribution of the individual scores.

With SDI, you can do lots of demos to convince yourself that this is true.

From the population menu, select a population that looks really different from the normal distribution—maybe the U-shaped one or one of the "messy" ones is a good choice for this demonstration.

Set the sample size to 10, click the "Repeating Samples" button, tick the "Warp Speed" box, and watch SDI build up the sampling distribution of the mean.

For samples of only 10, the sampling distribution of the mean may not look all that normal. It may be asymmetric, for example, or a bit wider and flatter than the usual normal shape.

Now increase the sample size to 50 and try it again. Unless you have a very unusual population distribution, the sampling distribution of the mean will probably now have a very normal-looking shape.

If 50 wasn't enough to get the normal shape, try 100 or 200. Eventually, as the sample size gets large enough, the sampling distribution of the mean is guaranteed to approach a normal shape.

This predictability of the behavior of sample means—regardless of the population shape—is one of the things that makes means so convenient for statistical inference, because sample means can always be depended upon to follow the normal curve. This kind of convergence to the normal is not guaranteed for other measures of central tendency (e.g., the median or mode), or indeed for any other summary statistic. You can easily verify this by repeating this demonstration with some other summary statistic.

2.2 Demo: Standard Error of the Mean

Like every statistical distribution, the sampling distribution of the mean has a standard deviation that summarizes the amount of variation within the distribution. In fact, the standard deviation of the sampling distribution of the mean is so important that it has a special name: "the standard error of the mean".

Another important property of the sampling distribution of the mean is that *its* standard deviation (i.e., the standard error of the mean) is completely predictable from the population standard deviation and the sample size.

We will work through an example to develop the formula. Start with any population that you like, and set its standard deviation to 12.

First, set the sample size to 4. Have SDI generate lots of samples with $N = 4$, and watch the σ ("sigma") value in the sampling distribution window, which estimates the standard error of the mean. You should see that σ converges toward 6—half the value of the original population distribution's $\sigma = 12$.

Second, change to a sample size of 9 (keeping the same population) and generate some new samples. With $N = 9$, you should see the standard error of the mean converge toward 4—one-third of the original population distribution's $\sigma = 12$.

Third, try it again with a sample size of $N = 16$ (still with the same population). Now you should see the standard error of the mean converge toward 3—one-fourth of the original population distribution's $\sigma = 12$.

If the original population has a standard deviation of 12, what is the standard error of the mean?

Sample size	Standard Error of the Mean
4	6
9	4
16	3

The above table summarizes the results. Can you guess the formula? The standard error of the mean is always $1/\sqrt{N}$ times the original population distribution's standard deviation. This formula is virtually always valid with random sampling, regardless of the shape of the original population distribution, and it is another part of what makes the sample mean such a powerful summary statistic for mathematical purposes.[1]

[1]Technically, there actually are a few bizarre population distributions for which this rule is violated, but they are not

2.3 Demo: Sample Proportions

So far, we have only examined sampling distributions in cases where the measured variable is a *numerical quantity*, like height. The concept of a sampling distribution applies just as well when the measured variable is categorical rather than numerical.

For example, consider political polling. The pollster asks "Do you plan to vote for Joe Blogg or Jane Doe?", and each prospective voter answers one way or the other. Each individual in the random sample is simply categorized into the group favoring one candidate or the other.

With categorical measured variables, the usual summary statistic is the *proportion* of individuals in a given category (say, the proportion who plan to vote for Joe Blogg), and the concept of a sampling distribution applies perfectly well to such sample proportions. Imagine 100 separate pollsters, each taking the same poll on a different sample of (say) 500 people. Each pollster would get a slightly different proportion in favor of Joe Blogg, just due to random sampling. The "sampling distribution of the sample proportion" describes this sample-to-sample variation in proportions, just as the sampling distribution of the mean described sample-to-sample variations in means. Thus, the "sampling distribution of the sample proportion" is the frequency distribution *across samples* of the proportion in a certain category (e.g., favoring Joe).

You can visualize sampling distributions of proportions within SDI using the "Binary" distribution. This distribution has just two distinct outcomes, labelled "0" and "1", corresponding to two possible categories of responses (e.g., favoring Joe versus favoring Jane). You can imagine the 0 and 1 corresponding to the two categories however you like—it's just arbitrary—so let's say "0" means favoring Joe for this example.

When you select the Binary distribution, SDI asks you to specify the "probability of the 1". A binary population is completely described by this one parameter, because there is some proportion of 1's and the rest are 0's. To illustrate, let's look at a distribution with a small majority in one category by setting this probability equal to 0.55. That means 55% of the population belongs to category 1, and the other 45% belongs to category 0.

Now look at a single sample of size $N = 10$ from this distribution. You will not be surprised to see that you get only 0's and 1's as scores in the sample—those were the only possibilities. Since the population has more 1's than 0's (55% vs. 45%), the sample is more likely to have a majority of 1's too, but this is by no means certain. Take several samples, and you are quite likely to get at least one with a majority of 0's, despite the majority of 1's in the population.

Now if you want to look at the sampling distribution of proportions with SDI, you need to know a little trick, because "sample proportion" is *not* included among the summary statistics that SDI offers to compute. Here is the trick: With the 0/1 binary distribution, the sample mean is exactly the same as the proportion of ones in the sample. For example, a sample of six 0's and four 1's gives a mean of $4/10 = 0.4$, which is the same as the proportion of 1's. (This is true for any sample size.) So, with this 0/1 binary distribution, you can just look at the sampling distribution of the mean, and that is the same as looking at the sampling distribution of the proportion.

Now use the "Repeating Samples" button to take lots of samples, and watch SDI build up the "Sampling Distribution of the Mean" (but we know it is really the proportion). Perhaps the first thing you notice is that this sampling distribution consists of a number of discrete spikes. With $N = 10$ the sample proportion can only be one of these discrete values: 0.0, 0.1, 0.2, $\ldots$, 0.8, 0.9, 1.0—so there are spikes at those points. There is no way to get an intermediate value, such as 0.65, as the proportion of 1's in a sample of 10, because this would require an intermediate score like 0.5, whereas only 0 and 1 are possible. Of course, these different sample proportions are not equally likely, so some of the spikes are taller than others. Since we set the true proportion of 1's in the population to 0.55, sample proportions tend to be close to that value (0.5 or 0.6 are the closest possibilities with $N = 10$).

Note also that the sample proportion is occasionally *quite far* from the true population proportion. Even though the true population proportion is 0.55, some samples have proportions as low as 0.1 or 0.2, and others have proportions as high as 0.9. These sample values are quite far from the true value of 0.55. The possibility of such large discrepancies tells us that a sample of $N = 10$ isn't really large enough to provide very accurate information about the population proportion.

Larger samples give better information, so let's look next at what happens with $N = 100$. The distribution

population distributions that you would ever encounter in the real world. In fact, they all have $\sigma = \infty$, which is really just a mathematical curiousity.

is again made up of discrete spikes, but now the spikes are much closer together, separated by steps of only 1/100 rather than 1/10. Furthermore, the sample proportions are now much closer to the true population proportion of 0.55. With $N = 100$, almost all of the samples have proportions within the range of about 0.40 to 0.70—a much narrower range than we found with $N = 10$. There is still clearly some error, though. Obviously, a sample of $N = 100$ is not really large enough to say which candidate is leading (i.e., to say whether the true proportion is less than 1/2 or more than 1/2), because you can easily get sample proportions on either side of 1/2 with $N = 100$, even though the true proportion is known to be larger than 1/2 (i.e., 0.55).

Now try $N = 1000$. Of course, the sample proportions stay even closer to the true population proportion now, mostly ranging from about 0.50 to 0.60. Importantly, virtually all of the samples have proportions greater than 0.5. With $N = 1000$, then, your sample would almost always lead you to identify the correct candidate as the leader. So, $N = 1000$ is a large enough sample to identify the leading candidate in a 0.55/0.45 race.

What if the race is closer—say 0.52/0.48—would $N = 1000$ be large enough in that case too? Try it (set the binary population parameter to 0.52). With $N = 1000$, the range of sample proportions is around 0.48–0.56. That is, you might get a sample proportion of 0.48, and conclude that Joe is leading, even though Jane is really ahead with 52% of the whole population. That tells us that either candidate might get a majority in samples of 1000, so with this closer race $N = 1000$ is not really enough to be sure of identifying the correct leader. With a race this tight, you need a sample of around $N = 5,000$. You don't see too many political polls with samples that large, but then the news organizations paying for the polls generally don't care so much about whether the polls are right, as long as they have a number to report.

2.4 Demo: Sampling Distributions For Other Statistics

We have looked at the sampling distributions of the mean and the proportion. To truly understand the concept of a sampling distribution, you should realise that the concept applies equally well to *any* statistic that could be used to summarize a sample.

For a different example, let's look at the sampling distribution of the minimum.

First, use the population menu to select a normal population, if you don't already have one. Also, select the sample size of $N = 10$ for a good illustration.

Then, click on SDI's "Statistic" main menu item and choose the option "Make sampling distribution for minimum". Now click "One Sample" to take a single sample, and SDI computes the minimum (smallest) value for that sample.

Click on "Repeating Samples" to see what different minimum values are obtained across lots of samples. You will probably see that the sample minimum values generally tend to come from the low end of the population distribution, as you would expect.

Due to sampling variability, though, the minimum varies quite a bit from one sample to the next. As you look at lots of different samples, you will see that the sampling distribution of the minimum spreads out farther to the left (low end) than to the right (high end). In other words, the sampling distribution of the minimum is not symmetric.

The most important point to realize, though, is the conceptual similarity between the sampling distribution of the minimum and the sampling distribution of the mean that you looked at previously. They are both tabulations, across many different samples, of the value of some sample statistic.

You could make such a sampling distribution for absolutely any sample statistic that you could compute from a sample. That sampling distribution would simply show the results you would get if you took lots and lots of samples, computed the statistic for each one, and tabulated the results.

The sample maximum and sample standard deviation are two other examples that you can look at in SDI, and they have their own somewhat distinctive shapes. Furthermore, the shapes of sampling distributions for most statistics other than the mean depend on the shape of the population. For example, if you look at the shape of the sampling distribution of the maximum with different messy distributions, you will get different sampling distribution shapes.

2.5 Summary of Sampling Distributions

You have probably now seen enough to get the general idea: For any statistic, the "sampling distribution" describes the variation in that statistic's values, across random samples of a given size. The sampling distribution of the mean is the best understood of all sampling distributions, because it is always approximately normal ("central limit theorem") and its standard deviation is always $1/\sqrt{N}$ times the population's standard deviation. The sampling distribution of any other statistic can always be studied quite easily through computer simulation, though, because we can always have the computer generate lots of samples compute the statistic for each one, and tabulate the results.

2.6 Demo: Hypothesis Testing

Now that you have seen what sampling distributions are, you can begin to get an idea of how they are used. One of the major uses of sampling distributions is in "hypothesis testing" procedures. Although SDI was not specifically designed to illustrate hypothesis testing, you can see the general idea behind these procedures using SDI.

To set the scene, imagine that your company is trying to decide whether to market a particular new product. Your accountants say you will make money if your market share is at least 10%, but you will lose money if your market share is less than that. A market research firm tests your product on a sample of 240 people, but only 17 say they would buy it—that's only 7%. Should you abandon your plans for the new product, based on the fact that 7% is less than 10%? Or is it possible that your true share of the whole market really is 10%, and you just got unlucky in that your sample happened to include (by chance) especially many people who didn't like your product?

This is an example of a hypothesis testing situation. The value of 10% is a "hypothesized" value, and you want to know whether your observed value (7%) is consistent with it or not. To decide, you need to look at the sampling distribution of sample percentages.

To do that with SDI, choose the binary population distribution. Using this distribution as the model amounts to classifying every person in the population as either a zero (someone who will not buy your product) or a one (someone who will buy it). For this distribution, set the "probability of the 1" to be 0.1 (that's 10%) to match the hypothesized value. Finally, set the sample size to 240 to match the actual size of the market research sample.

Now have SDI generate samples. As we saw in section 2.3, a handy fact about this binary 0/1 variable is that the sample mean is equal to the the proportion of ones in the sample; this proportion is the value we want to determine. Use the "Warp Speed" option to generate lots of samples and look at the sampling distribution of the mean (proportions) that emerges.

In the market research scenario, our initial question was whether our true market share might really be 10%, even though we got only 7% in a sample of 240. That is, might we get only 7% in a sample of 240 even if the true proportion were 10%?

We can see the answer by looking at the sampling distribution. Look along the horizontal axis and find 0.07, corresponding to 7%. Is that a possible sample value even with a true population proportion of 0.1? The sampling distribution tells you it is. In fact, the sampling distribution clearly includes proportions at least as low as 0.05, maybe even a little lower. That means that we could easily get 7% in a sample of 240 even though the true proportion was 10%. It follows that this sample is *not* convincing evidence that the true market share is less than 10%. Of course, the true market share might well be less than 10%, but based on this evidence we have to conclude that it *still could be* 10%.

Now consider a more extreme example. Suppose that the market research firm had found only 4% of the sample wanted to buy our product. Looking again at the sampling distribution we just constructed, you see there are no samples with proportions of 4% or lower. Evidently, the 4% figure is too low to be consistent with a 10% share for the population. In essence, the sampling distribution tells us that we would virtually never get a proportion as low as 4% in a sample of 240 if the true proportion were 10%. Therefore, when actually do observe a sample value of 4%, we must conclude that the true share cannot be as high as 10%.

To summarize this example in the language of hypothesis testing:

- We are interested in testing whether the true population proportion could be 10%. Technically, we have the "null hypothesis" that the true proportion is 10%.

- We use SDI or some other mathematical procedures to see what sample proportions might be observed if the null hypothesis were correct (i.e., if 10% were the true value).
- If the sample value we actually observe is within the range of what is commonly observed when the null hypothesis is true (e.g., if the sample proportion is 7%), then we conclude that the null hypothesis might be true.
- If the sample value we actually observe is outside the range of what is commonly observed when the null hypothesis is true (e.g., if the sample proportion is 4%), then we conclude that the null hypothesis is surely false.

2.7 Sampling Distributions Mini-FAQ

2.7.1 Why do we care about the distributions of all possible samples if we only actually observe one sample?

Knowing about all the samples tells us how much confidence you can have in the accuracy of any single one. If we know that samples like ours (say, from the same population and with the same N) are usually pretty accurate, then we can be more trusting of the results we got from it. If we know that such samples can be quite far off, then we will be less trusting.

2.7.2 If a sampling distribution shows what would happen with all possible samples, how can we ever know what it looks like without actually taking all possible samples?

Two ways. First, sometimes it is possible to prove mathematically (from certain assumptions) what a sampling distribution must be, at least approximately. In that case, having the proof obviates the need to actually look at all of the different samples. The "central limit theorem" (see section 2.1) is one example. Second, even when no mathematical proof is possible, we can always program a computer to generate a huge number of samples, and tabulate the results across those samples. With enough simulated samples, this tabulation gives an approximation of the true sampling distribution that is adequate for any practical purposes.

2.7.3 What is the "standard error" of a statistic?

For any statistic, the "standard error" is the standard deviation of its sampling distribution. For example, the standard error of the mean is the standard deviation of the sampling distribution of the mean. Likewise, the standard error of a proportion is the standard deviation of the sampling distribution of the proportion. And so on. As we saw in several demos, the standard error tends to decrease as the sample size increases.

2.7.4 What is the difference between "frequency" and "probability"?

Statisticians find it very useful to distinguish between numbers that describe the population as a whole and numbers that just describe a sample. Generally, for example, the mean for the population as a whole is called μ ("mu"), whereas the mean for a sample is called $\bar{X}$ ("X bar"). Similarly, "frequency" indicates how often a value was observed in a sample, whereas "probability" indicates how often it is observed in the whole population. This is why SDI uses the label "Probability" for the vertical axis of the population distribution (upper left panel) but uses the label "Frequency" for the vertical axis of the sample distribution (lower left panel).

2.7.5 What are the similarities and differences between a sampling distribution and a population distribution?

Population distributions and sampling distributions are similar in that both are theoretical probability distributions. Each describes the set of all possible outcomes from random sampling, and the relative probability of each outcome. There two main differences, though. One difference concerns the "individual" being sampled. For a population distribution, the individual is a single person, place, or thing, selected randomly from some larger set. For a sampling distribution, the "individual" is a single *random sample*, selected randomly from the set of *all possible samples* from that population. A second difference concerns the measured variable. For a population distribution, the measured variable is some property of each individual. For a sampling distribution, the measured variable is some property of the sample.

2.8 SDI: The Fine Print

Jeff Miller
Department of Psychology
University of Otago
Dunedin, New Zealand
miller@psy.otago.ac.nz

If you use this software, please register it by sending me your email address. There is no charge to register, and your email address will be used only to notify you of program bugs or updates. *License and warranty:*