A Student's Guide to the Conceptual Side of Inferential Statistics

by Dr. Stefano A. DeCaro

Table of Contents

Inferential Statistics: A Preview

With descriptive statistics we condense a set of known numbers into a few simple values (either numerically or graphically) to simplify an understanding of those data. This is analogous to writing up a summary of a lengthy book. The book summary is a tool for conveying the gist of a story to others, and the mean and standard deviation of a set of numbers is a tool for conveying the gist of the individual numbers (without having to specify each and every one). Inferential statistics, on the other hand, is used to make claims about the populations that give rise to the data we collect. This requires that we go beyond the data available to us. Consequently, the claims we make about populations are always subject to error; hence the term "inferential statistics" and not deductive statistics.

Inferential statistics encompasses a variety of procedures to ensure that the inferences are sound and rational, even though they may not always be correct. In short, inferential statistics enables us to make confident decisions in the face of uncertainty.

At best, we can only be confident in our statistical assertions, but never certain of their accuracy.

Trying to Understand the True State of Affairs

The world just happens to be a certain way, regardless of how we view it. The phrase "true state of affairs" refers to the real nature of any phenomenon of interest. In statistics, the true state of affairs refers to some quantitative property of a population. Numeric properties of populations (such as their means, standard deviations, and sizes) are called parameters. Samples (or subsets) of populations also have numeric properties, but we call them statistics. Thus, for the scientist using inferential statistics, population parameters represent the true state of affairs.

We seldom know the true state of affairs. The process of inferential statistics consists of making use of the data we *do* have (observed data) to make inferences about population parameters. Unfortunately, the true state of affairs is also dependent on all of the data we *don't* have (unobserved data). Nevertheless, an important aspect of sample data is that they are actual elements from an underlying population. In this way, sample data are 'representatives' of the population that gave rise to them. This implies that sample data can be used to estimate population parameters.

However, as sample data are only representatives, they are not expected to be perfect estimators. Consider that we necessarily lose information about a book when we only read a book review. Similarly, we lack information about a population when we only have access to a subset of that population. Remember that the parameters of a population (say, its mean and standard deviation) *are based on each and every element in that population*. It would be useful to have some measure of how reliable (or representative) our sample data really are. To this end, we must first consider the sampling process itself, and it is in this context that the importance of probability theory and random and independent sampling begin to emerge.

In the absence of prior knowledge about the details of some population of interest, sample data serve as our best estimate of that population.

True State of Affairs + Chance = Sample Data

Some elements (say, 'heights') in a population are more frequent than others. These more frequent elements are thus over-represented in the population compared to less common elements (e.g., the heights of very short and very tall individuals). The laws of chance tell us that it is always possible to randomly select *any* element in a population, no matter how rare (or under-represented) that element may be in the population. If the element exists, then it can be sampled, plain and simple. However, the laws of probability tell us that rare elements are not expected to be sampled often, given that there are more numerous elements in that same population. It is the more numerous (or more frequent) elements that tend to be sampled each time a random and independent sample is obtained from the population.

A sample is random if all elements in the population are equally eligible to be sampled, meaning that chance, and chance alone, determines which elements are included in the sample. A sample is independent if the chances of being sampled are not affected by which elements have already been sampled. To illustrate these two ideas, imagine that you are interested in the average age of all university students in the United States. For convenience sake, you decide to randomly select one student from each class offered at your university this term. With respect to the original population of interest (all university students in the U.S.), your sample is *not* random, because only students at your university are eligible to be sampled. Your sample is also *not* independent, because once you select a student from a class, no other student in that class has a chance of being sampled. In this case, any claims you make

based on your sample cannot be applied to the population you are really interested in. At best, you are only investigating the population of students at one particular university.

When the sampling process is truly random and independent, samples are expected to reflect the most representative elements of the underlying population. But rare outcomes do occur (every now and then). A rare sample occurs when, just by chance, a relatively large number of the extreme (high *or* low) elements in the population end up in the sample. In other words, the percentage of extreme values in the sample is higher than the actual percentage in the population, as might be the case if you measured the heights of everyone present in the basketball locker room. Although the heights of basketball players are part of the overall population, they are likely to be over-represented in the sample, in which case the sample mean would not accurately reflect the true state of affairs. Specifically, the sample mean would be biased by the presence of too many heights from "tall" people.

An important consequence of random and independent sampling is that chance factors virtually guarantee that sampled data will vary in their degree of representativeness from sample to sample. Most samples will tend to be good approximations of the underlying population, and a minority of samples will provide misleading accounts of the true state of affairs--just by chance selection. The problem, of course, is that we can never know whether our particular sample is biased by the presence of too many extreme (i.e., rare) elements. But just as you probably don't expect to win the lottery, you should also not expect to be the rare individual who just happens to obtain a rare sample. It is not rational to expect an outcome that has a low probability associated with it. Hence, the logic is to assume that any particular sample mean is typical of the underlying population. This assumption is reasonable only when the sampling process is random and independent; otherwise, rare samples might artificially occur too often.

To summarize thus far, the underlying population represents the true state of affairs, which naturally affects the outcome of any particular sample. For instance, if the shortest person in the population is 4' and the tallest person is 8', then it must be the case that the mean of any sample taken from the population will fall within the range of 4 to 8 feet. There are also chance factors operating on the sampling process, which makes it very unlikely that *exactly* the same elements will be sampled each time. Thus, sample data are expected to vary across repeated sampling. This "sampling error" must be taken into account when making inferences about a population from sample data.

Sampling error refers to discrepancies between the statistics of random samples and the true population values; but this "error" is simply due to which elements in the population end up in the sample. In other words, sampling error refers to *natural chance factors*, not to errors of measurement or errors due to poorly designed and poorly executed experiments. We have control over the latter, but nature imparts a certain degree of unavoidable error.

To illustrate the idea of sampling error, imagine that we toss a fair coin six times and obtain {HHHHHH}. We expect a fair coin to land heads 50% of the time, so what went wrong? To answer this

question, we have to think about the *population* of outcomes when a fair coin is tossed six times (see Figure 1).

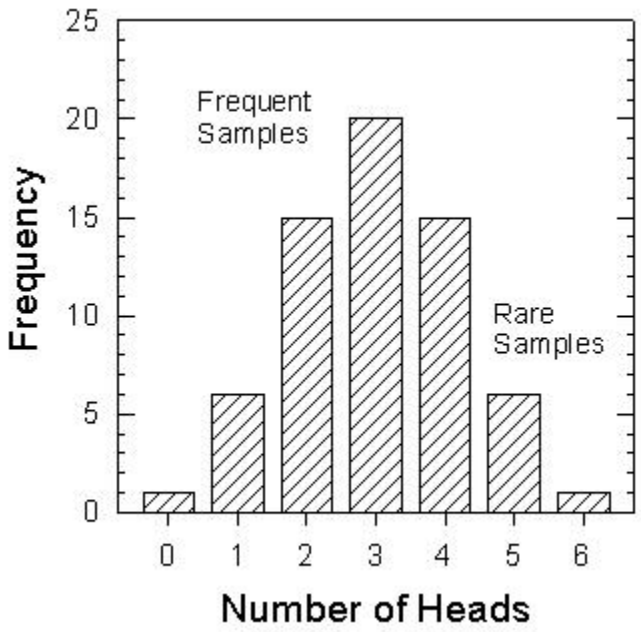


Figure 1. Sampling distribution of heads when a fair coin is tossed six times.

It turns out there are ***N*** = 64 possibilities, but only 20 contain exactly three heads and three tails. Nonetheless, three heads (in any order) is the most frequent element in this population; it is also the mean. In contrast, there is only one outcome containing exactly six heads, which makes it a rare (but not impossible) event. In fact, Figure 1 allows us to easily calculate the exact probability of {HHHHHH}; it is 1/64 (or .016). Likewise, the probability of three heads is 20/64 (or .313), meaning that we expect to get three heads about 1/3 of the time we toss a fair coin six times. It was because of random sampling that we failed to observe one of these more representative samples, such as {HTHHTT}, *not* because the mean of the population isn't really 3. Thus, {HHHHHH} is an example of sampling error. It is "error" in the sense that the true population mean is 3 heads, but the sample (i.e., the six tosses) yielded 6 heads, just by chance. If our sampling (coin tossing) process is fair, then we expect this rare event to occur about once every 64 times, on average.

The laws of chance combined with the true state of affairs create a natural force that is always operating on the sampling process. Consequently, the means of different samples taken from the same population are expected to vary around the 'true' mean just by chance.

Sampling Distributions

A population is the collection of all possible elements that fit into some category of interest, such as "all adults living in the United States." Once we've defined a population, we need to specify *with respect to*

what? For instance, all adults living in the United States *with respect to their height*. Now the population of interest has shifted from a collection of people to a collection of numbers (heights, in this case). When the elements in the population have been measured or scored in some way, it is possible to talk about distributions. We can generate a distribution of anything, as long as the elements can take on values. This is precisely what we did in the coin-tossing example. First we obtained a sample of six tosses, and then we scored the sample with respect to the *number of heads*. If we had done this for all 64 possible samples and then counted the number of times each value (0 through 6) occurred, we would have ended up with the frequency distribution in Figure 1. We could also have calculated the *mean* number of heads for each sample, in which case the x-axis would have consisted of seven means ranging from 0 (0/6) to 1 (6/6), with 0.5 (3/6) in the middle. This would show more clearly that the probability of heads is 0.5 (or 50%) in the population, regardless of the number of tosses.

When the distribution of interest consists of all the unique samples of size ***n*** that can be drawn from a population, the resulting distribution of sample *means* is called the sampling distribution of the mean. Thus, a "sampling distribution" in general is a distribution of sampling outcomes, like the one depicted in Figure 1. A sampling distribution of the mean is one particular kind of a sampling distribution, one that is based on sample means. There are also sampling distributions of medians, standard deviations, and any other statistic you can think of.

Populations, which are distributions of individual elements, give rise to sampling distributions, which describe how collections of elements are distributed in the population.

It may be helpful to think of populations has *having* their own sampling distributions, because we are now making a distinction between two distributions: (a) the distribution of individual elements (the population) and (b) the distribution of all unique samples of a particular size *from* that population (the sampling distribution). [A sample is unique if no other sample in the distribution contains exactly the same elements.] Before reading on, make certain that you are comfortable with the idea that a *sample* of elements can represent a single, unique element in a distribution consisting of many other unique samples (see Table 1).

Table 1. Basic Properties of Populations, Samples, and Sampling Distributions

Level	Collection	Elements
Population	All individuals (***N*** = size of population)	The scores each individual receives on some attribute.
Sample	Subset of individuals from the population. (***n*** = size of sample)	The scores each individual in the sample receives on some attribute.
Sampling Distribution	All unique samples of size ***n*** from the population.	The values of a statistic applied to each sample.

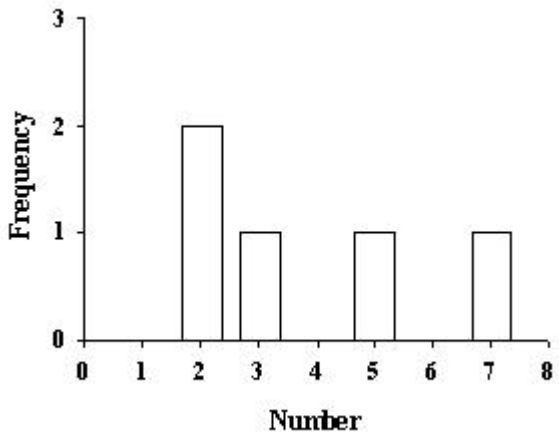
Why are sampling distributions important in inferential statistics? The answer is simple: because we obtain *samples* of data when we conduct studies. If we are going to make inferences about populations based on sample data, then we need to understand the sampling properties of those samples. In inferential statistics we make use of two important properties of sampling distributions, better known as the central limit theorem:

1. The mean of all unique samples of size ***n*** (i.e., the average of all the means) is identical to the mean of the population from which those samples are drawn. This is equivalent to saying that the mean of the sampling distribution equals the mean of the original population. Thus, any claims about the mean of the sampling distribution apply to the population mean.
2. The shape of the sampling distribution increasingly approximates a normal curve as sample size (***n***) is increased, even if the original population is not normally distributed. [**Note**--If the original population is itself normally distributed, then the sampling distribution will be normally distributed even when the sample size is only one. *Why?*]

Confused? Perhaps if you *see* these properties you'll understand just how simple they really are. First let's create a small, hypothetical population of numbers:

$$Pop = \{2, 5, 7, 3, 2\}$$

The distribution for our hypothetical population looks like this:



In this case ***N*** = 5 (because there are five elements in the population), and **μ** = 3.8 (the mean of the population). Property #1 says that if we gather all the unique samples of a particular size, and then calculate means for each sample, the average of those means will equal the population mean. We'll do this twice, once using ***n*** = 3, and again using ***n*** = 4. Table 2.1 lists all of the unique samples (and their means) that are possible when three elements are sampled at a time.

Table 2.1. All unique samples from the hypothetical population when ***n*** = 3.

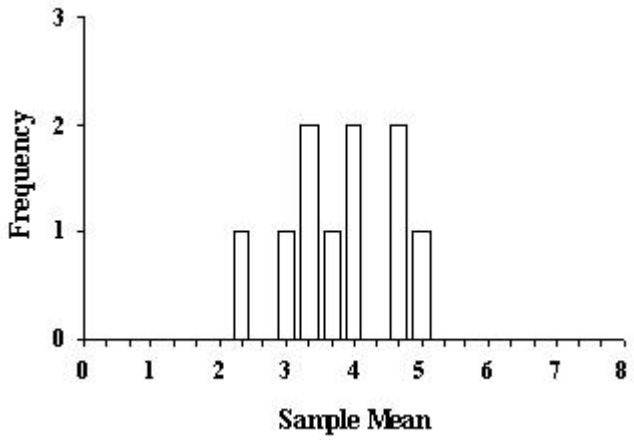
Sample	Sample Mean
{2 2 3}	2.33
{2 2 5}	3.00
{2 3 5}	3.33
{2 3 5}	3.33
{2 2 7}	3.67
{2 3 7}	4.00
{2 3 7}	4.00
{2 5 7}	4.67
{2 5 7}	4.67
{3 5 7}	5.00
Grand Mean	**3.80**

At first glance it may appear that the samples in Table 2.1 are not unique because, for example, {2 3 5} has been listed twice. However, remember that there are two 2s in the population; they are different elements that simply share the same value. Thus, Table 2.1 indicates there are 10 unique samples in the sampling distribution when ***n*** = 3. Notice also that the mean of the 10 sample means is 3.8. This is the same value we obtained when we calculated the mean of the five elements in the population (**μ**). Now consider Table 2.2, which lists all of the unique samples that are possible when sample size is increased to four. The first thing to notice is that the range of sampling outcomes is smaller (3.00 to 4.25 instead of 2.33 to 5.00)—there is less variability. Nonetheless, the mean of the sample means is still 3.8.

Table 2.2. All unique samples from the hypothetical population when ***n*** = 4.

Sample	Sample Mean
{2 2 3 5}	3.00
{2 2 3 7}	3.50
(2 2 5 7}	4.00
{2 3 5 7}	4.25
(2 3 5 7}	4.25
Grand Mean	**3.80**

The central limit theorem also states that the sampling distribution will approximate a normal distribution if sample size is sufficiently large, even if the underlying population is not normally distributed. It is clear that the hypothetical population in our example is not normally distributed, primarily because it is so small. For example, the distribution is not symmetrical around its mean, which is the most salient feature of normal distributions. But compare the shape of the population with the shape of the sampling distribution corresponding to Table 2.1:



These 10 sample means are far from being normally distributed, but we can see hints of a bell curve: The distribution is peaked near the center and shorter at the tails. This distribution is also more symmetrical around its mean compared with the underlying population. If our hypothetical population were somewhat larger (so that more samples could be generated), the sampling distribution would be more normal. Nonetheless, we can still see the effects of the central limit theorem even with this overly-simplified example. Most real-world populations are very large, and so their sampling distributions contain millions of sample combinations and therefore many possible values of a statistic.

The Standard Error of the Mean: A Measure of Sampling Error

Sampling distributions have a standard deviation, which describes the variability of sample means from *their* mean (which, remember, equals the population mean). There is a different sampling distribution for each value of ***n***, for two reasons. First, as illustrated above, the *number* of unique samples that can be drawn from a population depends on the size of those samples. In other words, sample size determines how many elements (sample means) are in the sampling distribution to begin with. Second, as sample size increases, the *variability* among all possible sample means decreases. This must be the case, because if all the elements in the original population are sampled (i.e., if ***n*** = ***N***), then there is only one possible sample that can be obtained (the sample *is* the population) and the variability of a single number is zero. Thus, sample size determines both the size and the variability of a sampling distribution (compare Tables 2.1 and 2.2).

The standard deviation of a sampling distribution of means is given a special name: standard error of the mean (abbreviated as SEM). It may not be obvious, but the SEM is a measure of sampling error because it describes the variability among all possible means that could be sampled in an experiment. [*Recall that the elements of interest are now sample means, not the individual scores within a sample or population*.] Simply put, the degree of variability in the sampling distribution bears directly on the degree to which observed results (sample means) are expected to vary just by chance. If there is a lot of

variability in the sampling distribution (as is the case when the distribution consists of *small samples*), then sample means can vary greatly. On the other hand, if there is little variability in the sampling distribution (as is the case when the distribution consists of *large samples*), then sample means will tend to be very similar, and very close to the true population mean.

At this point we can begin to address the question raised earlier, namely *How can we know whether our sample is representative of the underlying population?* Obviously it is important to avoid small samples, as there are more extreme (i.e., rare) sample means in the sampling distribution--and we are more likely to get one of them in an experiment. Thus, we can increase our confidence in a particular sample (as being representative of the population) by increasing the number of elements included in the sample. The means of large samples tend to cluster tightly around the true population mean. Consequently, rare samples (whose means are very different from the true population mean) are less common in the sampling distribution and therefore less likely to arise just by chance. Notice that by choosing a sample size we are also determining which sampling distribution our sample will come from. Ideally, we always want to sample from the distribution with the least variability, because less variability translates into *more reliability!*

We have some control over sampling error because sample size determines the standard error (variability) in a sampling distribution.

Theoretical Sampling Distributions as Statistical Models of the True State of Affairs

Unless the details of a population are known in advance, it is not possible to describe any of its sampling distributions. We would have to first measure all the elements in the population, in which case we could simply calculate the desired parameter, and then there would be no point in collecting samples. For this reason, a variety of idealized, theoretical sampling distributions have been described mathematically. The Student-t distribution, for instance, is a standardized version of a theoretical sampling distribution, meaning that it can be used as a statistical *model* for many of the real sampling distributions of interest to behavioral scientists. The reason for using theoretical sampling distributions is to obtain the likelihood (or probability) of sampling a particular mean if the mean of the sampling distribution (and hence the mean of the original population) is some particular value. In practice, the population parameter must first be hypothesized, as the true state of affairs is generally unknown. This is called the null hypothesis.

In the coin-tossing example we were able to deduce the sampling distribution shown in Figure 1. It too is theoretical because we constructed it without tossing a single coin! This underscores an important point, namely that many of the populations and sampling distributions addressed in statistics are abstract; they exist in a mathematical sense.

Theoretical sampling distributions have been generated so that researchers can estimate the probability of obtaining various sample means from a pre-specified population (real or hypothetical).

Making Formal Inferences about Populations: Preview to Hypothesis Testing

When there are many elements in the sampling distribution, it is always possible to obtain a rare sample (i.e., one whose mean is very different from the true population mean). The probability of such an outcome occurring just by chance is determined by the particular sampling distribution specified in the null hypothesis (in much the same way that Figure 1 provided us with the probability of tossing 6 heads). When the probability (*P*) of the observed sample mean occurring by chance is really low (typically less than one in 20, e.g., $P < .05$), the researcher has an important decision to make regarding the hypothesized true state of affairs. One of two inferences can be made:

1. The hypothesized value of the population mean is correct and a rare outcome has occurred just by chance (as in the coin-tossing example).
2. The true population mean is probably some other value that is more consistent with the observed data. Reject the null hypothesis in favor of some alternative hypothesis.

The rational decision is to assume #2, because the observed data (which represent direct, albeit partial, evidence of the true state of affairs), are just too unlikely if the hypothesized population is true. Thus, rather than accept the possibility that a rare event has taken place, the statistician chooses the more likely possibility that the hypothesized sampling distribution is wrong. However, rare samples do occur, which is why statistical inference is always subject to error. Indeed, even when observed data are consistent with a hypothesized population, they are also consistent with many other hypothesized populations. It is for this reason that the hypothesized value of a population parameter can never be proved nor disproved from sample data. We use inferential statistics to make tentative assertions about population parameters that are most consistent with the observed data. Actually, inferential statistics only helps us to *rule out* values; it doesn't tell us what the population parameters are. We have to infer the values, based on what they are likely not to be.

Only in the natural sciences does evidence contrary to a hypothesis lead to rejection of that hypothesis without error. In statistical reasoning there is also rejection (inference #2), but with the possibility that a rare sample has occurred by chance (sampling error). This is the nature of making inferences based on random sampling.

The proper APA citation for this article is:
DeCaro, S. A. (2003). A student's guide to the conceptual side of inferential statistics. Retrieved [Month Day, Year], from http://psychology.sdecnet.com/stathelp.htm.