Sampling distribution of the sample mean

We already know how to find parameters that describe a population, like mean, variance, and standard deviation. But we also know that finding these values for a population can be difficult or impossible, because it's not usually easy to collect data for every single subject in a large population.

So, instead of collecting data for the entire population, we choose a subset of the population and call it a "sample." We say that the larger population has N subjects, but the smaller sample has n subjects.

In the same way that we'd find parameters for the population, we can find statistics for the sample. Then, based on the statistic for the sample, we can infer that the population parameter might be similar to its corresponding sample statistic.

Sampling distribution of the sample mean

Consider the fact though that pulling one sample from a population could produce a statistic that isn't a good estimator of the corresponding population parameter.

For example, maybe the mean height of girls in our statistics class is 65 inches. Let's say there are 30 girls in the class, and we take a sample of 3 of them. If we happened to pick the three tallest girls, then the mean of that sample wouldn't be a good estimate of the population mean, because the mean height from the sample would be significantly higher than the mean height of the population. Similarly, if we instead happened to choose the three shortest girls, the sample mean would be much lower than the actual population mean.

So how do we correct for this? How do we adjust for the fact that individual samples might produce sample statistics that are bad estimates of their corresponding population parameters?

Well, instead of taking just one sample from the population, think about what happens when we take every possible sample of 3 girls from the population of girls in our class. The total number all of possible samples is N^n, where N is the total population from which we take our samples, and n is the sample size.

So if we take a sample of 3 girls from a population of 30 girls, the total number of possible samples is

$$N^n = 30^3 = 27{,}000$$

Keep in mind that, in this scenario, we're **sampling with replacement**, which means we randomly choose one girl, then "put her back" into the population and pick another girl randomly, then put the second girl back into the population and pick another girl randomly. Those three choices generate our random sample of three girls, and the replacement after each selection means the same girl can appear multiple times within one 3 -girl sample. We'd keep doing this over and over until we've taken every unique 3-girl sample.

What we want to realize is that every one of these samples has its own mean, so now we have a data set of $27{,}000$ sample means. Here's the key point: This set of sample means actually forms its own distribution around the real population mean. In other words, if we look at these means as a probability distribution, it turns out that this probability distribution of sample means is always normal (as long as we're taking large enough samples, more on this later), and this normal distribution is called the **sampling distribution of the sample mean (SDSM)**.

Just think about the sampling distribution of the sample mean as the probability distribution of all possible values of the sample mean $\bar{x}$. Because the sampling distribution of the sample mean is normal (assuming the original population was normal, and/or we used a large enough sample size $n \geq 30$), we can of course find a mean and standard deviation for the distribution, and therefore answer probability questions about it.

Central Limit Theorem

We just said that the sampling distribution of the sample mean is normal, but let's clarify. In actuality,

- If the original population is normally distributed, then the SDSM will also be normally distributed, regardless of the sample size n that we use.

- If the original population is not normally distributed, or if we don't know the shape of the population distribution, then the SDSM is only guaranteed to be normally distributed when we use a sample size of at least $n = 30$.

This conclusion about the normality of the SDSM is the Central Limit Theorem. In reality, many populations don't follow a normal distribution, meaning that they don't approximate the bell-shaped-curve of a normal distribution. Real-life distributions are all over the place, because real-life phenomena don't always follow a perfectly normal distribution.

The **Central Limit Theorem (CLT)** is how we turn a non-normal population into a normal SDSM. It tells us that, even if a population distribution is non-normal, as long as we use a large enough sample ($n \geq 30$), that we can make inferences about our sample statistics, because of the fact that the SDSM will be a normal distribution.

So the Central Limit Theorem is useful because it lets us apply what we know about normal distributions, like the properties of mean, variance, and standard deviation, to non-normal populations.

Mean, variance, and standard deviation

The Central Limit Theorem also states that the mean of the sampling distribution of the sample mean will always be the same as the mean of the original distribution.

$$\mu_{\bar{x}} = \mu$$

If the population is infinite, or if we're sampling with replacement, then the variance of the SDSM is equal to the population variance divided by the sample size.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

If population variance is unknown (which will almost always be the case), then we can use sample variance as an estimate of population variance.

$$s_{\bar{x}}^2 \approx \frac{s^2}{n}$$

The standard deviation of the sampling distribution, also called the **standard error**, the standard deviation of sample means, or the standard error of the mean, is therefore given by

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is population standard deviation and n is sample size. When we don't know population standard deviation, we can use sample standard deviation as an estimate of population standard deviation.

$$SE = s_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

The standard error tells us how accurate the mean of any given sample is likely to be as an estimate of the actual population mean. "Error" doesn't mean there's been a mistake, it just refers to the distance between any particular sample mean and the mean of the population.

When the standard error is larger, it indicates that the sample means in the SDSM are more spread out, so it's less likely that any given sample mean is an accurate representation of the true population mean. But when the standard error is smaller, the sample means are less spread out, so it's more likely that any given sample mean is an accurate representation of the true population mean.

- The standard error will be larger when population standard deviation is larger, and/or when the sample size is smaller.

- The standard error will be smaller when population standard deviation is smaller, and/or when the sample size is larger.

Finite population correction factor

We've already said that the standard deviation of the sampling distribution of the sample mean (standard error) is given by $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. In fact, this formula leaves something out. The complete formula is actually

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

This extra $\sqrt{(N-n)/(N-1)}$ is what we call the **finite population correction factor (FPC)**. When the population we're sampling from is

- infinite, or

- when we're sampling with replacement, or

- when the population is finite but large in comparison to a smaller sample (the sample size is less than or equal to $5\,\%$ of the population, $n/N \leq 0.05$),

then the value of the FPC is 1 or very close to 1. And when the FPC's value is 1 or very close to 1, it'll have no or little impact on the value of the standard error, which is when we can simplify the standard error formula to just $SE = \sigma_{\bar{x}} = \sigma/\sqrt{n}$.

In other words, as long as the population is infinite, or we're sampling with replacement, or we're sampling from no more than 5% of a finite population, then we can use the simplified formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ for standard error. Otherwise, if we're sampling without replacement or sampling from more than 5% of a finite population, we should use

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

And of course, given this formula for the standard error, we know that the variance of the SDSM is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Keep in mind that there's some debate among statisticians and statistics textbooks about whether the FPC should be applied only when $n/N > 0.10$, instead of using the stricter threshold $n/N > 0.05$. For our purposes, we'll stick to using the $n/N > 0.05$ rule.

Let's do an example so that we can see how these formulas work.

Example

A group of 4 people have the following weights in pounds: 150, 156, 158, 164. Find all possible random samples of size 2 if we're sampling with replacement. Then find the sample mean for every sample. Determine the probability distribution of the sample mean, the mean of the SDSM $\mu_{\bar{x}}$, and the standard error $\sigma_{\bar{x}}$.

Let’s first determine the total number of possible samples, using N^n, given $N = 4$ and $n = 2$.

$$N^n = 4^2 = 16$$

The complete sample space, and the mean for each sample, is

Sample	Sample mean
150, 150	150
150, 156	153
150, 158	154
150, 164	157
156, 150	153
156, 156	156
156, 158	157
156, 164	160
158, 150	154
158, 156	157
158, 158	158
158, 164	161
164, 150	157
164, 156	160
164, 158	161
164, 164	164

Build a table for the probability distribution of the sample mean. Because there are 16 total samples, the probability of each sample mean will be given by the number of times that sample mean occurs, divided by the total number of possible samples, so "count$/16$."

Sample mean	P(x$_i$)
150	1/16
153	2/16
154	2/16
156	1/16
157	4/16
158	1/16
160	2/16
161	2/16
164	1/16

Now we can calculate the mean of the sampling distribution of the sample mean, $\mu_{\bar{x}}$, where $\bar{x}_i$ is a given sample mean, $P(\bar{x}_i)$ is the probability of that particular sample mean occurring, and N is the number of samples.

$$\mu_{\bar{x}} = \sum_{i=1}^{N} \bar{x}_i P(\bar{x}_i)$$

$$\mu_{\bar{x}} = 150\left(\frac{1}{16}\right) + 153\left(\frac{2}{16}\right) + 154\left(\frac{2}{16}\right) + 156\left(\frac{1}{16}\right) + 157\left(\frac{4}{16}\right)$$

$$+158\left(\frac{1}{16}\right) + 160\left(\frac{2}{16}\right) + 161\left(\frac{2}{16}\right) + 164\left(\frac{1}{16}\right)$$

$$\mu_{\bar{x}} = \frac{2{,}512}{16}$$

$$\mu_{\bar{x}} = 157$$

Because we're sampling with replacement, we would expect this mean of the SDSM to be equivalent to the mean of the population, $\mu_{\bar{x}} = \mu$, and we can see that it is if we calculate the mean of the population.

$$\mu = \frac{150 + 156 + 158 + 164}{4} = \frac{628}{4} = 157$$

Both means are $\mu_{\bar{x}} = \mu = 157$. The population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{(150 - 157)^2 + (156 - 157)^2 + (158 - 157)^2 + (164 - 157)^2}{4}$$

$$\sigma^2 = \frac{(-7)^2 + (-1)^2 + 1^2 + 7^2}{4}$$

$$\sigma^2 = \frac{49 + 1 + 1 + 49}{4}$$

$$\sigma^2 = \frac{100}{4}$$

$$\sigma^2 = 25$$

which means that the population standard deviation is

$$\sigma = \sqrt{25}$$

$$\sigma = 5$$

Because we're sampling with replacement, we can use the simplified formula for standard error (the one *without* the FPC).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{5}{\sqrt{2}}$$

$$\sigma_{\bar{x}} \approx 3.54$$

Instead of calculating population variance and standard deviation, and using those values to find standard error, we could have calculated the variance of the SDSM directly,

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{N} (\bar{x}_i - \mu)^2 P(\bar{x}_i)$$

$$\sigma_{\bar{x}}^2 = (150 - 157)^2\left(\frac{1}{16}\right) + (153 - 157)^2\left(\frac{2}{16}\right) + (154 - 157)^2\left(\frac{2}{16}\right)$$

$$+(156 - 157)^2\left(\frac{1}{16}\right) + (157 - 157)^2\left(\frac{4}{16}\right) + (158 - 157)^2\left(\frac{1}{16}\right)$$

$$+(160 - 157)^2\left(\frac{2}{16}\right) + (161 - 157)^2\left(\frac{2}{16}\right) + (164 - 157)^2\left(\frac{1}{16}\right)$$

$$\sigma_{\bar{x}}^2 = (-7)^2\left(\frac{1}{16}\right) + (-4)^2\left(\frac{2}{16}\right) + (-3)^2\left(\frac{2}{16}\right) + (-1)^2\left(\frac{1}{16}\right)$$

$$+1^2\left(\frac{1}{16}\right)+3^2\left(\frac{2}{16}\right)+4^2\left(\frac{2}{16}\right)+7^2\left(\frac{1}{16}\right)$$

$$\sigma_{\bar{x}}^2=\frac{49}{16}+\frac{32}{16}+\frac{18}{16}+\frac{1}{16}+\frac{1}{16}+\frac{18}{16}+\frac{32}{16}+\frac{49}{16}$$

$$\sigma_{\bar{x}}^2=\frac{200}{16}$$

$$\sigma_{\bar{x}}^2=\frac{25}{2}$$

$$\sigma_{\bar{x}}^2=12.5$$

and then found standard error directly.

$$\sigma_{\bar{x}}=\sqrt{12.5}$$

$$\sigma_{\bar{x}}\approx 3.54$$