Conditions for inference with the SDSM

Now that we've defined the sampling distribution for the sample mean and the Central Limit Theorem, we want to be able to use it to make inferences about the population. After all, the whole point of statistics is being able to use samples to make educated guesses about the population.

So, even when we don't know the population mean, our goal now will be to take a sample, find the mean of the sample, and then compare the sample mean to the mean of the SDSM. That comparison will allow us to draw conclusions about the population mean.

But, in order for this process to be valid, there are always three conditions we need to meet when we're sampling: the sample needs to be random, it needs to be large enough for the Central Limit Theorem to kick in and ensure normality, and our samples need to be independent.

Random

Any sample we take needs to be a simple random sample. Often we'll be told in the problem that sampling was random.

Normal (large counts)

We need to know that our sample size is large enough. In the case of the sampling distribution of the sample mean, 30 is a magic number for the sample size we need to use to make the sampling distribution normal. In other words, the sample size needs to be at least 30 in order for the CLT to create a normal SDSM for a non-normally distributed population (unless we're sampling with replacement, in which case we can get away with a sample size smaller than 30).

If the original population is normally distributed, then this rule doesn't apply because the sampling distribution will also be normal, even if our sample size is less than 30.

If our population is finite, and we're sampling without replacement or taking a sample larger than 5% of the population, then we just have to remember to use the finite population correction factor that we talked about earlier.

Independent (10% rule)

If we're sampling with replacement, then the 10% rule tells us that we can assume the independence of our samples. But if we're sampling without replacement (we're not "putting our subjects back" into the population every time we take a new sample), then we need to keep our sample size below 10% of the total population.

For example, if the original population is $2{,}000$ subjects, we need to make sure that the sample size is no more than 200 subjects so that we stay under the $200/2{,}000 = 1/10 = 10\%$ threshold.

In other words, as long as we keep the sample size to 10% or less of the total population, we can "get away with" a sample that isn't truly independent (we can get away with sampling without replacement), because this 10% threshold is small enough to approximate an independent sample.

If our sample meets these conditions, then we can use the sampling distribution of the sample mean to answer questions about the probability that any given sample mean will fall within some distance of the population mean.

For these types of problems, we'll need to use

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

to determine z-scores that apply to the sampling distribution of the sample mean. Let's do an example.

Example

A sports equipment company produces approximately $1{,}000{,}000$ soccer balls per year, and the pressure in the soccer balls is normally distributed with a mean of 8.7 PSI (pounds per square inch), and a standard deviation of 0.4 PSI. They randomly select 25 soccer balls (without replacement) to check their pressure. Find the probability that the sample mean $\bar{x}$ is within 0.2 PSI of the population mean.

Before we can answer this probability question, we need to check our conditions for inference. We were told in the problem that the sample was taken randomly, so we can assume we've met the "random" condition.

We were also told that the PSI in the population of soccer balls is normally distributed, so our SDSM will also be normal, even though our sample size is smaller than 30, and we've therefore met the "normal" condition.

We're sampling without replacement, which means our sample needs to be at most 10% of the population, but 25 soccer balls is a significantly smaller sample than 10% of the population, so we've met the "independent" condition.

To answer the probability question, we'll start by finding the mean of the SDSM. The Central Limit Theorem tells us that it'll be equal to the population mean, so $\mu_{\bar{x}} = 8.7$. The standard error will be

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

$$\sigma_{\bar{x}} = \frac{0.4}{\sqrt{25}}$$

$$\sigma_{\bar{x}} = \frac{0.4}{5}$$

$$\sigma_{\bar{x}} = 0.08$$

Remember here that we're sampling without replacement from a finite population, but our sample size is significantly smaller than 5% of the population, so we didn't need to apply the finite population correction factor when we found the standard error.

We want to know the probability that the sample mean $\bar{x}$ is within 0.2 PSI of the population mean, 8.7. A 0.2 interval around 8.7 gives us the interval 8.5 to 8.9, so

$$P(8.5 < \bar{x} < 8.9) = P\left(\frac{8.5 - 8.7}{0.08} < z_{\bar{x}} < \frac{8.9 - 8.7}{0.08}\right)$$

$$P(8.5 < \bar{x} < 8.9) = P\left(\frac{-0.2}{0.08} < z_{\bar{x}} < \frac{0.2}{0.08}\right)$$

$$P(8.5 < \bar{x} < 8.9) = P(-2.50 < z_{\bar{x}} < 2.50)$$

In the z-table, a z-value of 2.50 gives 0.9938,

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964

and a z-value of -2.50 gives 0.0062.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

Which means the probability under the normal curve between these z-scores is

$$P(-2.50 < z_{\bar{x}} < 2.50) = 0.9938 - 0.0062$$

$$P(-2.50 < z_{\bar{x}} < 2.50) = 0.9876$$

$$P(-2.50 < z_{\bar{x}} < 2.50) \approx 98.8\,\%$$

So there's an approximately $98.8\,\%$ chance that the mean $\bar{x}$ of the 25-ball sample the company takes will fall within 0.2 PSI of the population mean of $\mu = 8.7$ PSI.
