Sampling Distribution of the Mean

Connecting sampling distributions with Standard Error, Confidence Intervals, and Hypothesis Testing

Central Limit Theorem

The central limit theorem (CLT) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with a "large enough n," then the resulting distribution can be approximated by the normal distribution.

Applying the law of large numbers here, we could say that if you take larger and larger samples from a population, then the mean {x} of the sample tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for {x} is .) This means that the sample mean {x} must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. Sampling is done with replacement.

The CLT means says that if you keep drawing larger and larger samples and calculating their means, **the sample means form their own normal distribution** (the sampling distribution).

The sampling distribution of the mean is generated by repeated sampling from the same population and recording the sample mean per sample. This forms a distribution of different means, and this distribution has its own **mean and variance**.

The normal distribution has **the same mean as the original distribution and a variance that equals the original variance divided by the sample size.**

Drawing samples of people's weight from the NHANES dataset.

We have established that different samples yield different statistics due to sampling variability . These statistics have their own distributions, called sampling distributions, that reflect this as a random variable. The sampling distribution of a sample statistic is the distribution of the point estimates based on samples of a fixed size, n, from a certain population. It is useful to think of a particular point estimate as being drawn from a sampling distribution.

Recall the sample mean weight calculated from a previous sample of 173.3 lbs. Suppose another random sample of 60 participants might produce a different value of x, such as 169.5 lbs. Repeated random sampling could result in additional different values, perhaps 172.1 lbs, 168.5 lbs, and so on. Each sample mean can be thought of as a single observation from a random variable X. The

distribution of X is called the sampling distribution of the sample mean, and has its own mean and standard deviation like the random variables discussed previously. We will simulate the concept of a sampling distribution using technology to repeatedly sample, calculate statistics, and graph them. However, the actual sampling distribution would only be attainable if we could theoretically take an infinite amount of samples.

Each of the point estimates in the table above have their own unique sampling distributions which we will look at in the future

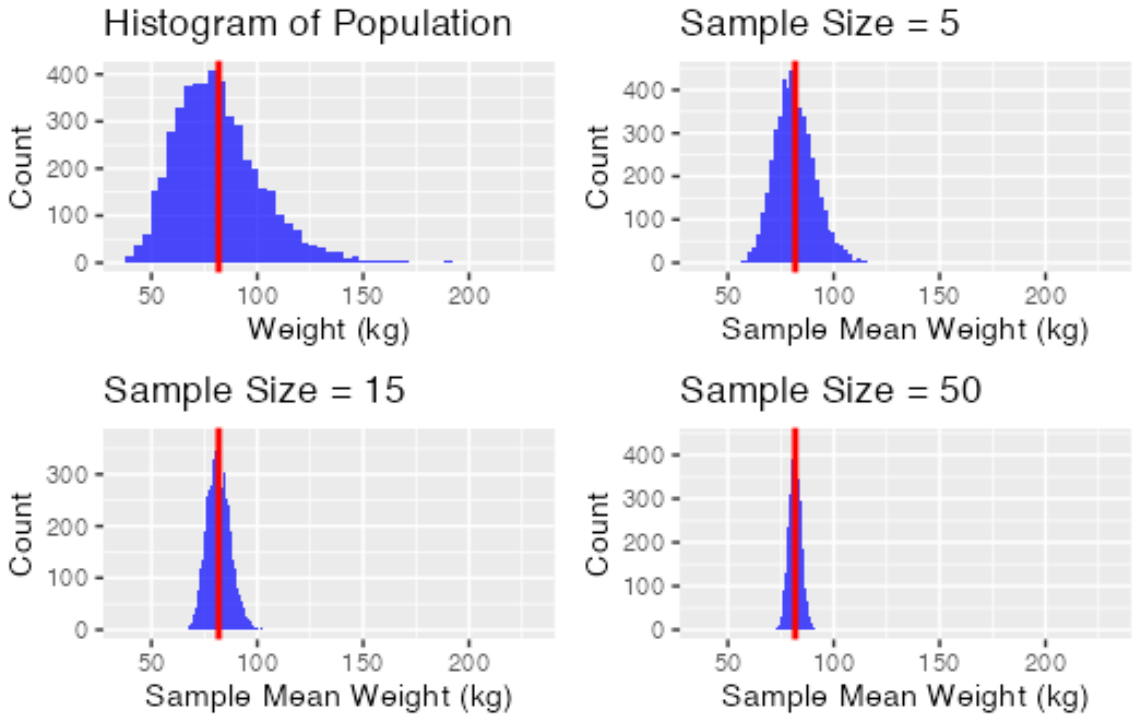

Figure 1: We are drawing a random sample of people from the dataset and calculating the mean weight for that sample. *Sample size* is the number of data points we pull. We then repeat this 5000 times (`n_samples`) to build up the sampling distribution.

Standard Error

A sampling distribution is what we get by simulating multiple samples (of sample size n) from a population.

Recall: The Standard Error is the standard deviation of the sampling distribution.

$$SEM = \sigma_{\bar{x}\ (means)}$$

Standard Error

A sampling distribution is a probability distribution of a statistic at a given sample size.

Recall: The Standard Error is the standard deviation of the sampling distribution. This is also equal to the standard deviation σ of the population divided by the square root of the sample size.

$$SEM = \sigma_{\bar{x}\ (means)} = \frac{\sigma}{\sqrt{n}} \approx \frac{\sigma_x}{\sqrt{n}} \left[i.e. \frac{\text{Est. Std Dev of the sample}}{\sqrt{\text{Sample size}}} \right]$$

In other words:

> If you draw random samples of size n, the distribution of the random variable $\bar{X}$, which consists of sample means, is called the sampling distribution of the sample mean. The sampling distribution of the mean approaches a normal distribution as n, the sample size, increases.

In the SEM formula, remember the *sampling distribution* is the distribution of multiple means - not the distribution of our sample.

Quote from https://pressbooks.lib.vt.edu/introstatistics/chapter/the-central-limit-theorem-for-sample-means-averages/

Standard Error

Key Takeaways

- A sampling distribution is what we get by simulating multiple samples from a population.
- The Standard Error is the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution.
- The SE decreases as the sample size n increases.
- Because of this relationship - we can *estimate* the SE from a single sample $\frac{\sigma_x}{\sqrt{n}}$

$$SEM = \sigma_{\bar{x}\ (means)} = \frac{\sigma}{\sqrt{n}} \approx \frac{\sigma_x}{\sqrt{n}} \left[i.e. \frac{\text{Est. Std Dev of the sample}}{\sqrt{\text{Sample size}}} \right]$$

Bibliography