Example: Blood Pressure in Massachusetts

We picked a random sample of 100 people from Massachusetts and took their blood pressure. The average diastolic blood pressure in our sample was 75. We don't know the true underlying population mean and want to use our sample estimate to make inferences about it. For now assume that even though we don't know the population mean, we do know that the population variance is $\sigma^2 = 552.25$.

In addition, we know that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$. In this example:

$$Var(\bar{X}) = \frac{552.25}{100} = 5.52$$

and the standard deviation of $\bar{X}$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{23.5}{10} = 2.35$.

The sample mean is useful but not so much without its standard deviation. Every time we pick a random sample, it is one out of many samples we could have picked. In which case, the sample mean of 75 we got above is one of the many sample means we could have got. We know that despite this sampling variation, the sample mean is an unbiased estimator of μ. So we could get numbers larger or smaller than the true μ but on average, they will cancel out.

The standard deviation tells us how much further from μ our realizations could be. In other words, the standard deviation tells us how good or precise our estimate is. The standard deviation of the sample mean is lower if the underlying population is more similar (lower σ) or if we have a larger sample size (higher n) since both of these reduce the scope of sampling variation affecting our estimate.

Going a step further, if we know the distribution of the sample mean, we can attach probabilities to certain intervals of outcomes. In our example, since $n \geq 100$ we can invoke the Central Limit Theorem (CLT) and conclude that $\bar{X}$ is normally distributed. So,

$$\bar{X} \sim N(\mu, 5.52)$$

Knowing the distribution of the sample mean helps us calculate areas under the curve that correspond to probabilities of outcomes in certain intervals.

90% Confidence Interval

Say we are interested in constructing a 90% confidence interval (CI) around the mean. A $1 - \alpha$ CI is given by

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{n}$$

where $z_{\alpha/2}$ leaves area $\alpha/2$ in the upper tail of the normal distribution. Here $1 - \alpha = 0.90$, so $\alpha = 0.10$. We need to find $z_{0.5}$ which is equal to 1.645 (from the standard normal table). So our 90% CI in this example is given by

$$75 \pm 1.645 \times 2.35$$

Interpretation: We are 90% confident that the true mean diastolic blood pressure in Massachusetts is between 71.13 and 78.86. Although the true mean may or may not be in this interval, 90% of intervals formed in this manner will contain the true mean.

A 95% CI would be wider as the margin of error has to be larger when we want to be 95% confident vs 90% confident. The margin of error will also be larger if the standard deviation of the sample mean is larger.