7.1: The Central Limit Theorem for Sample Means

The **sampling distribution** is a theoretical distribution. Specifically, the sampling distribution is the probability distribution of the mean of a random sample. Since the sample is obtained randomly, the value of the sample mean is a random variable. The sampling distribution is created by taking many samples of size n from a population. Each sample mean is then treated like a single observation of this new distribution, the sampling distribution. The genius of thinking this way is that it recognizes that when we sample we are creating an observation and that observation must come from some particular distribution. The Central Limit Theorem answers the question: from what distribution did a sample mean come? If this is discovered, then we can treat a sample mean just like any other random variable and calculate probabilities about what values it might take on. We have effectively moved from the world of statistics where we know only what we have from the sample, to the world of probability where we know the distribution from which the sample mean came and the parameters of that distribution.

The reasons that one samples a population are obvious. The time and expense of checking every invoice to determine its validity or every shipment to see if it contains all the items may well exceed the cost of errors in billing or shipping. For some products, sampling would require destroying them, called destructive sampling. One such example is measuring the ability of a metal to withstand saltwater corrosion for parts on ocean going vessels.

Sampling thus raises an important question; just which sample was drawn. Even if the sample were randomly drawn, there are theoretically an almost infinite number of samples. With just 100 items, there are more than 75 million unique samples of size 5 that can be drawn. If 6 are in the sample, the number of possible samples increases to just more than one billion. Of the 75 million possible samples, then, which one did you get? If there is variation in the items to be sampled, there will be variation in the samples. One could draw an "unlucky" sample and make very wrong conclusions concerning the population. This recognition that any sample we draw is really only one from a distribution of samples provides us with what is probably the single most important theorem is statistics: **the Central Limit Theorem**. Without the Central Limit Theorem it would be impossible to proceed to inferential statistics from simple probability theory. In its most basic form, the Central Limit Theorem states that **regardless** of the underlying probability density function of the population data, the theoretical distribution of the means of samples from the population will be normally distributed. In essence, this says that the mean of a sample should be treated like an observation drawn from a normal distribution. The Central Limit Theorem only holds if the sample size is "large enough" which has been shown to be only 30 or more.

Figure 7.1.1 graphically displays this very important proposition.

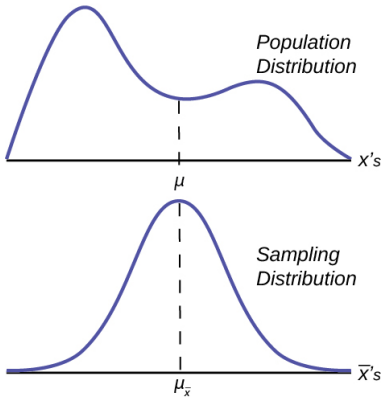

Figure 7.1.1

Notice that the horizontal axis in the top panel is labeled x. These are the individual observations of the population. This is the **unknown** distribution of the population values. The graph is purposefully drawn all squiggly to show that it does not matter just how odd ball it really is. Remember, we will never know what this distribution looks like, or its mean or standard deviation for that matter.

The horizontal axis in the bottom panel is labeled $\overline{x}$'s. This is the theoretical distribution called the sampling distribution of the sample mean. Each observation on this distribution is a sample mean. All these sample means were calculated from individual samples with the same sample size. The theoretical sampling distribution contains all of the sample mean values from all the possible samples that could have been taken from the population. Of course, no one would ever actually take all of these samples, but if they did this is how they would look. And the Central Limit Theorem says that they will be normally distributed.

The Central Limit Theorem goes even further and tells us the mean and standard deviation of this theoretical distribution, as detailed in Table 7.1.1.

Table 7.1.1

Parameter	Population distribution	Sample	Sampling distribution of $\overline{X}$
Mean	μ	$\overline{X}$	$\mu_{\overline{X}} = \mathrm{E}\left[\overline{X}\right] = \mu$
Standard deviation	σ	s	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

The practical significance of The Central Limit Theorem is that now we can compute probabilities for drawing a sample mean, $\overline{X}$, in just the same way as we did for drawing specific observations, X's, when we knew the population mean and standard deviation and that the population data were normally distributed. The standardizing formula has to be amended to recognize that the mean and standard deviation of the sampling distribution, sometimes, called the standard error of the mean, are different from those of the population distribution, but otherwise nothing has changed. The new standardizing formula is

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Notice that $\mu_{\overline{X}}$ in the first formula has been changed to simply μ in the second version. The reason is that mathematically it can be shown that the expected value of $\overline{X}$, the mean of $\overline{X}$, is equal to μ. This was stated in Table 7.1.1 above. This formula will be used in the next chapter to provide interval estimates of the **unknown** population parameter μ.