Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population **with replacement**, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually $n \geq 30$). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that $\min(np, n(1-p)) \geq 5$, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

For the random samples we take from the population, we can compute the mean of the sample means:

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

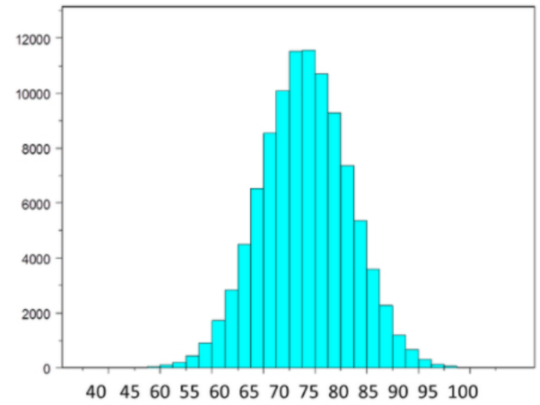
and the standard deviation of the sample means:

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

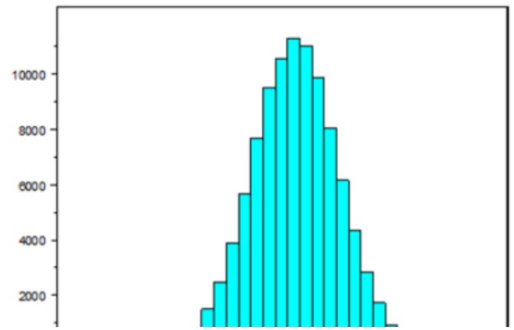
Before illustrating the use of the Central Limit Theorem (CLT) we will first illustrate the result. In order for the result of the CLT to hold, the sample must be sufficiently large ($n \geq 30$). Again, there are two exceptions to this. If the population is normal, then the result holds for samples of any size (i..e, the sampling distribution of the sample means will be approximately normal even for samples of size less than 30).

Central Limit Theorem with a Normal Population

The figure below illustrates a normally distributed characteristic, X, in a population in which the population mean is 75 with a standard deviation of 8.



If we take simple random samples (**with replacement**) of size n=10 from the population and compute the mean for each of the samples, the distribution of sample means should be approximately normal according to the Central Limit Theorem. Note that the sample size (n=10) is less than 30, but the source population is normally distributed, so this is not a problem. The distribution of the sample means is illustrated below. Note that the horizontal axis is different from the previous illustration, and that the range is narrower.



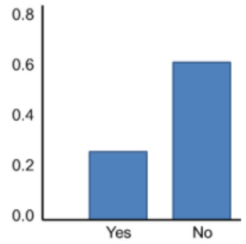
The mean of the sample means is 75 and the standard deviation of the sample means is 2.5, with the standard deviation of the sample means computed as follows:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{10}} = 2.5$$

If we were to take samples of n=5 instead of n=10, we would get a similar distribution, but the variation among the sample means would be larger. In fact, when we did this we got a sample mean = 75 and a sample standard deviation = 3.6.

Central Limit Theorem with a Dichotomous Outcome

Now suppose we measure a characteristic, X, in a population and that this characteristic is dichotomous (e.g., success of a medical procedure: yes or no) with 30% of the population classified as a success (i.e., p=0.30) as shown below.



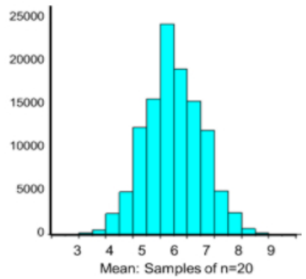
The Central Limit Theorem applies even to binomial populations like this provided that the minimum of np and n(1-p) is at least 5, where "n" refers to the sample size, and "p" is the probability of "success" on any given trial. In this case, we will take samples of n=20 with replacement, so min(np, n(1-p)) = min(20(0.3), 20(0.7)) = min(6, 14) = 6. Therefore, the criterion is met.

We saw previously that the ***population*** mean and standard deviation for a binomial distribution are:

Mean binomial probability: $\mu = np$

Standard deviation: $\sigma = \sqrt{n(p)(1-p)}$

The distribution of sample means based on samples of size n=20 is shown below.



The mean of the ***sample means*** is

$$\bar{X} = np = 20(0.3) = 6$$

and the standard deviation of the sample means is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{n(p)(1-p)}}{\sqrt{n}}$$

$$\sigma_{\bar{X}} = \frac{\sqrt{20(0.3)(0.7)}}{\sqrt{20}} = 0.46$$

Now, instead of taking samples of n=20, suppose we take simple random samples (with replacement) of size n=10. Note that in this scenario we do not meet the sample size requirement for the Central Limit Theorem (i.e., min(np, n(1-p)) = min(10(0.3), 10(0.7)) = min(3, 7) = 3).The distribution of sample means based on samples of size n=10 is shown on the right, and you can see that it is not quite normally distributed. The sample size must be larger in order for the distribution to approach normality.

Central Limit Theorem with a Skewed Distribution

Central Limit Theorem with a Skewed Distribution

The Poisson distribution is another probability model that is useful for modeling discrete variables such as the number of events occurring during a given time interval. For example, suppose you typically receive about 4 spam emails per day, but the number varies from day to day. Today you happened to receive 5 spam emails. What is the probability of that happening, given that the typical rate is 4 per day? The Poisson probability is:

$$P(x;\mu) = \frac{(e^{-\mu})(\mu^{x})}{x!}$$

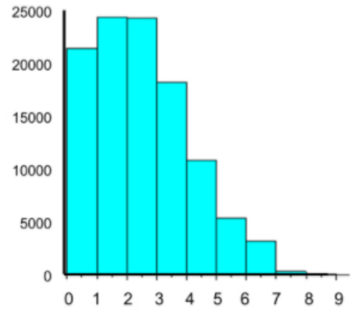
Mean = μ

Standard deviation = $\sigma = \sqrt{\mu}$

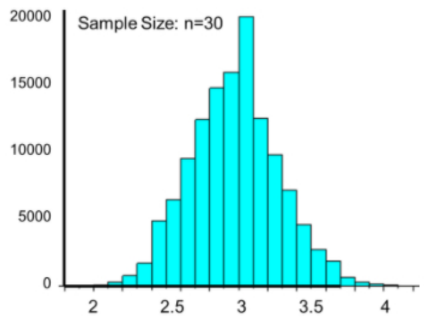
The mean for the distribution is μ (the average or typical rate), "X" is the actual number of events that occur ("successes"), and "e" is the constant approximately equal to 2.71828. So, in the example above

$$P(5;4) = \frac{(2.71828^{-\mu})(\mu^{x})}{X!} = 0.15829 = 15.8\%$$

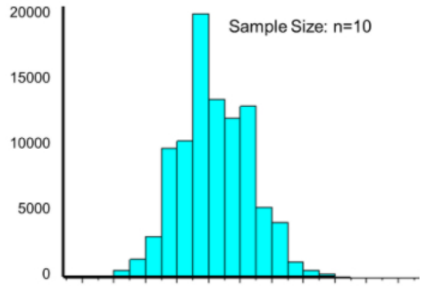
Now let's consider another Poisson distribution. with μ=3 and σ=1.73. The distribution is shown in the figure below.



This population is not normally distributed, but the Central Limit Theorem will apply if n ≥ 30. In fact, if we take samples of size n=30, we obtain samples distributed as shown in the first graph below with a mean of 3 and standard deviation = 0.32. In contrast, with small samples of n=10, we obtain samples distributed as shown in the lower graph. Note that n=10 does not meet the criterion for the Central Limit Theorem, and the small samples on the right give a distribution that is not quite normal. Also note that the sample standard deviation (also called the "**standard error**") is larger with smaller samples, because it is obtained by dividing the population standard deviation by the square root of the sample size. Another way of thinking about this is that extreme values will have less impact on the sample mean when the sample size is large.



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.73}{\sqrt{30}} = 0.32$$



1.5 2 2.5 3 3.5 4 4.5

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.73}{\sqrt{10}} = 0.55$$

return to top | previous page | next page

Date last modified: July 24, 2016.
Wayne W. LaMorte, MD, PhD, MPH

Boston University School of Public Health