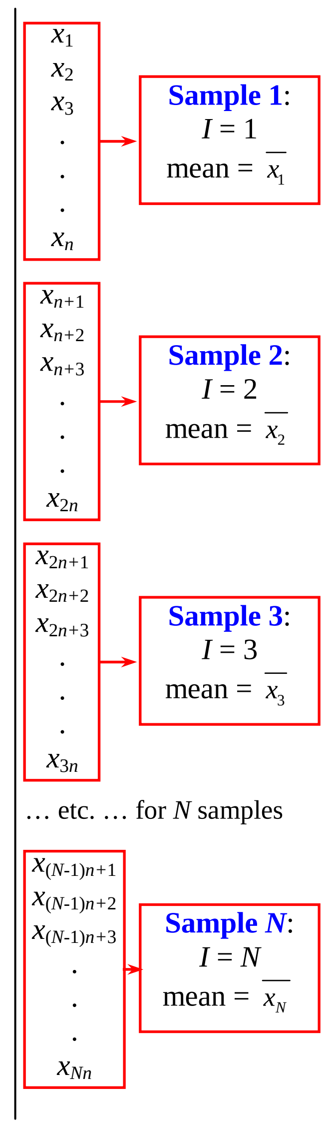
The Central Limit Theorem (CLT)

Author: John M. Cimbala, Penn State University
Latest revision: 18 January 2013

Introduction



- It is rare that anyone can measure something for an entire *population* – instead, a *sample* (or several samples) of the population is measured, and the population statistics are estimated from the sample.
- The ***Central Limit Theorem*** (***CLT***) is an extremely useful tool when dealing with multiple samples.

Multiple samples and the Central Limit Theorem

- Consider a population of random variable x (we assume that variations in x are purely random – in other words, if we would plot a PDF of variable x, it would look Gaussian or normal).
- The population mean μ and the population standard deviation σ are not known, but are instead estimated by taking several samples.
- We take N samples, each of which contains n measurements of variable x, as indicated in the sketch to the right.
- We define the ***sample mean*** for sample I as $\overline{x_I} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where index I = 1, 2, 3, … N (one for each sample). In other words, we calculate a sample mean in the usual fashion – an average value – *for each sample* 1 through N.
- We collect all N values of sample mean $\overline{x_I}$, and treat this collection as a sample itself (we call it the sample of the sample means). The ***sample of the sample means*** consists of N data points: $\overline{x_1}, \overline{x_2}, \overline{x_3}, \ldots \overline{x_N}$.
- We perform standard statistical analyses on this sample of the sample means:
  - The ***sample mean of the sample means*** is simply the average of all the sample means, $\left(\overline{\overline{x}}\right) = \frac{1}{N}\sum_{I=1}^{N} \overline{x_I}$. It should be obvious that the sample mean of the sample means is identically equal to the *overall* sample mean of *all* the data points combined (all N samples of n data points each): $\left(\overline{\overline{x}}\right) = \left(\overline{x}\right)_{\text{overall}}$.
  - Our best estimate of the population mean is thus $\mu \approx \left(\overline{\overline{x}}\right) = \left(\overline{x}\right)_{\text{overall}}$.
  - The ***sample standard deviation of the sample means*** (also called the ***standard error of the mean***) is calculated by the usual definition of sample standard deviation, but applied to the N sample means, $S_{\overline{x}} = \sqrt{\frac{\sum_{I=1}^{N}\left(\overline{x_I} - \left(\overline{\overline{x}}\right)\right)^2}{N-1}}$.
- The ***Central Limit Theorem*** (***CLT***) is stated as follows: **As *n* approaches infinity, the sample standard deviation of the sample means approaches the overall sample standard deviation divided by the square root of *n***. Mathematically, $S_{\overline{x}} \approx \frac{S_{\text{overall}}}{\sqrt{n}}$.
- Excel calculates this value in the *Descriptive Statistics* macro for a sample, and calls it the ***Standard Error***.
- However, as n gets large, the *sample* standard deviation approaches the *population* standard deviation, and thus we write the Central Limit Theorem in its more popular forms, $\sigma_{\overline{x}} \approx \frac{\sigma}{\sqrt{n}}$ or $S_{\overline{x}} \approx \frac{\sigma}{\sqrt{n}}$ or $\sigma_{\overline{x}} \approx \frac{S}{\sqrt{n}}$.
- In general, n is considered "large" for $n > 30$. N must also be "large" for best results. Of course, the larger n is, the better the CLT works. Likewise, the larger N is, the better the CLT works.
- Here is why the Central Limit Theorem is so useful in statistics: **As *n* and *N* get large, the PDF of the sample of the sample means is Gaussian even if the underlying population is *not* Gaussian.**
- The Central Limit Theorem is the foundation that enables us to use the student's t PDF to estimate the confidence interval of the population mean based on a sample.