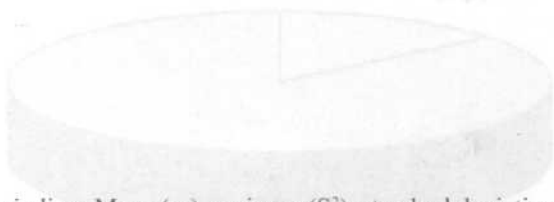


# Statistics for Clinicians

## The Four Basic Indices



Ahmed A. Hassouna, MD

**T**he four basic indices Mean ( $m$ ), variance ( $S^2$ ), standard deviation (SD) and standard error of mean (SEM) are the usual indices of numerical measurements or observations and are the basis of most of the statistical tests used to compare those measurements as well.

The mean value ( $m$ ) is calculated by dividing the sum ( $\sum$ ) of all values ( $x_i$ ) by the number of those values ( $n$ ); or  $m = \sum (x_i) / n$ . Let's take a simple example and imagine 2 groups of patients: group (a) consisting of 3 men aging 49, 50 and 51 years, while group (b) is formed of a 99 years very old man, a 50 years old man and a very young 1 year old child. Although the mean age of both groups is 50 years [ $m = (49 + 50 + 51)/3 = (99 + 50 + 1)/3 = 50$ ], yet no one can ever consider that both groups are comparable as regards the studied variable which is age. The mean alone is not only meaningless but can shadow very important information, which is in our case: the close proximity of individual values of the first group and their wide variability in the second group. The variance ( $S^2$ ), the standard deviation (SD) and the standard error of means (SEM) are 3 measures that were created to uncover this information. In other words, to show the extent of deviation of individual values from the mean: the more the individual values are dispersed away from the mean; the larger these measures of deviation will be.

The variance ( $S^2$ ) was designed to represent the mean of those deviations by summing ( $\sum$ ) the differences between each of the individual values and the mean ( $x_i - m$ ) and then dividing this sum by the number of values ( $n$ ). However, statisticians were faced by 2 technical drawbacks: First, simple deviations may be either positive or negative and, when summed, positive values may be annulated by negative ones. In our example, the sum of deviations =  $(49-50) + (50-50) + (51-50) = (1) + 0 + (-1) = 0$  and  $(1-50) + (50-50) + (100-50) = (49) + 0 + (-49) = 0$ . Of course this extreme example where the sum of deviations equals zero is not the rule and was mending to show a possible effect of such annulations. The best solution was to square those deviations  $(x_i - m)^2$  to get ride of the signs (+ or -) before summing them. In other words, the sum of deviations was replaced by the sum of squared deviations. However, this was not all as follows.

Secondly, the mean of such deviations was thought to be obtained as usual by dividing the sum of squared deviations  $\sum (x_i - m)^2$  by the sample size ( $n$ ). However, the sum of squared deviations is not divided by ( $n$ ) but by ( $n-1$ ) and this also needs some explanation: Let's take group (a) and calculate the sum of squared deviations or  $\sum (x_i - m)^2$ :  $(49-50)^2 + (50-50)^2 + (51-50)^2 = 1 + 0 + 1$ . It is obvious that only 2 variations were taken into consideration  $(49-50)^2$  and  $(51-50)^2$  and that the middle one  $(50-50)^2$  equaled zero because its value was

Accepted for publication Jan 10 ,2005  
 Address reprint request to Dr. A. Hassouna  
 Department of Cardio- thoracic surgery  
 Ain Shams University  
 980 El Mokattam ,  
 Cairo, Egypt .  
 Email : Ahmedhassouna@hotmail.com  
 Codex :03/02/edct/0506

the mean itself. In other words, only 2 variables were free to vary around the mean while the third which is the mean itself was not free to do so. As the variance was designed to describe the variations around the mean, and as in any sample one of the individual values may be the mean itself, it was more appropriate to subtract this 1 (that do not vary around the mean because it is the mean itself) from the number of values in the denominator. This correction was suggested not only on this theoretical basis, but after computing the variance with repeated samples with the use of (n). The values obtained did not have the desired property of averaging around the population variance and subtracting 1 from the number of values appeared to straighten such bias. The value n-1 is called the degree of freedom (df) and variance is better defined as sum of squared deviations divided by the (df) =  $\sum (x_i - m)^2 / n - 1$ .

The Standard deviation (SD) is the square root of variance (=  $\sqrt{S^2}$ ) and the Standard error of mean (SEM) is the square root of the variance after being divided by the number of values =  $\sqrt{(S^2/n)}$ .

In group (a):

$$S^2 = [ (49-50)^2 + (50-50)^2 + (51-50)^2 ] / (3-1) = 2 / 2 = 1.$$

$$SD = \sqrt{1} = 1 \text{ and } SEM = \sqrt{(1 / 3)} = 0.58$$

In group (b):

$$S^2 = [ (99-50)^2 + (50-50)^2 + (1-50)^2 ] / (3-1) = [2401 + 0 + 2401] / 2 = 2401.$$

$$SD = \sqrt{2401} = 49 \text{ and } SEM = \sqrt{(2401 / 3)} = 28.3$$

As we can easily notice, the age of group (a) patients was homogenous as evidenced by the small variance, SD and SEM in comparison to the calculated mean, while group (b) patients were far from being so. Variance is mostly used in mathematical equations of statistical tests; SD and SEM are commonly used for data pre-

sentation and the question rises: which of them should one use and when? While the SD measures the deviation of the individual values ( $x_i$ ) from the calculated mean (m), the SEM measures the deviation of the calculated mean (m) from the real mean (M) of the population from which our values ( $x_i$ ) were selected.

In the words of our example, the mean age of our group (a) patients is 50 years and the calculated SD of 1 is very small compared to the calculated mean and reflects the homogeneity of the individual values and their near proximity to the mean. The calculated SEM (0.58) reflects the deviation of the calculated mean itself (50 years) from the true mean or grand mean or the mean age of the population from which our sample study was issued. Both SD and SEM are deductible the one from the other ( $SEM = SD/\sqrt{n}$ ) but the latter is of more sense for studies carried out on large groups of patients and hence, have more chance to approach the true mean values of the population, compared to small samples. This is part of the central limit theorem: the more we add values, the more our values have a tendency to accumulate around the mean

In other words, the larger is our sample size (n), the more our calculated mean (m) approaches the true mean (M) of the population and, consequently, the smaller SEM will be. This is the theoretical basis for taking into consideration the sample size in calculating such deviation (i.e. SEM) by dividing the variance by n before taking its square root.

The mathematical equations:

$$m = \sum (x_i) / n$$

$$S^2 = \sum (x_i - m)^2 / n - 1$$

$$SD = \sqrt{S^2}$$

$$SEM = \sqrt{(S^2 / n)} = SD / \sqrt{n}.$$