Standard Deviation and Variance

The *standard deviation* is the most commonly used measure for variability. This measure is related to the distance between the observations and the mean. For example, suppose we have the following range of numbers: 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The mean is 55 ((10 + 20 + 30 …. + 100) / 10). How can the variability around the mean be best defined? Taking all distances from the mean together is inappropriate as this would result in the range: -45 (= 10 – 55), -35, -25, -15, -5, 5, 15, 25, 35 and 45. The sum of this range is *always* 0, which of course is not informative of the variability. It is more appropriate to turn all distances into *absolute* distances (that is, multiplying the negative numbers by -1). The sum then amounts to 250 (45 + 35 + 25 + 15 + 5 + 5 + 15 + 25 + 35 + 45). This sum, divided by the number of observations, yields the mean distance: 250 / 10 = 25. However, this absolute measure is not often used because it does not relate well to inferential statistics (see chapter 3).

Another strategy is to sum the *squared* distances (a negative score turns positive when squared). This results in a sum of 8250 (= $-45^2 + -35^2 + -25^2 + -15^2 + -5^2 + 5^2 + 15^2 + 25^2 + 35^2 + 45^2$ = 2025 + 1225 + 625 + 225 + 25 + 25 + 225 + 625 + 1225 + 2025). By dividing this sum by the number of observations (10), the average squared distance to the mean equals 825. In statistics, this number is known as the *variance*. The variance can be compared to the area of a square (see Figure 2.20)

Sides = 28.72

Area = 28.72 * 28.72 = 825

Figure 2.20 *Variance Compared to the Area of a Square*

In statistics, the measure of variability is preferably indicated as a distance instead of a squared distance (i.e., a square). The square root of 825 (= 28.72) is taken (this value equals the length of the sides in Figure 2.20) and the resulting measure is called the *standard deviation*.[i] Roughly, the standard deviation can be interpreted as the *average distance from the mean*, although mathematically this is not correct.[ii]

---

[i] In our example on p. 38 we divided by the total number of observations (n): $\sigma^2 = \frac{\sum_i^n (x_i - \mu)^2}{n}$. The standard deviation is obtained by taking the square root: $\sqrt{\sigma^2}$. These formulas are correct when dealing with a population. In a random sample the best (unbiased) estimates for both variance and standard error in the population are obtained with n-1 as the devisor:

$s^2 = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}$, the standard deviation (s) in the random sample is $\sqrt{s^2}$.
With n – 1 the variance within a random sample (s^2) is an unbiased approximation of the variance within the population (σ^2). Division by n would result in a systematic underestimation of σ^2. The reason for this bias is interdependency: when all observations are summed, the mean is no longer unknown (dividing the total sum by n yields the mean). In statistics this interdependency is expressed as the *degrees of freedom* (df). In this case df equals n – 1. This simply means that n – 1 observations and the mean suffice to calculate the exact value of the last observation. In statistical packages, including SPSS, s^2 and s are calculated.

[ii] The variance is the average squared distance to the mean. However, it is erroneous to think that the standard deviation (the square root of the variance) equals the average distance to the mean. If one takes the square root from a sum of numbers, then the outcome is not equal to the square root taken from the sum of the same but squared numbers. Example: $\sqrt{(2^2 + 3^2)} = \sqrt{(4 + 9)} = 3{,}6$ but $2 + 3 = 5$. However, we believe that a concept like the standard deviation is better understood if one defines it roughly as 'the average distance to the mean'.