DEFINITION

The **sample variance,** denoted by s^2, is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation,** denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

The unit for s is the same as the unit for each of the x_i's. If, for example, the observations are fuel efficiencies in miles per gallon, then we might have $s = 2.0$ mpg. A rough interpretation of the sample standard deviation is that it is the size of a typical or representative deviation from the sample mean within the given sample. Thus if $s = 2.0$ mpg, then some x_i's in the sample are closer than 2.0 to $\bar{x}$, whereas others are farther away; 2.0 is a representative (or "standard") deviation from the mean fuel efficiency. If $s = 3.0$ for a second sample of cars of another type, a typical deviation in this sample is roughly one and one half times what it is in the first sample, an indication of more variability in the second sample.

Motivation for s^2

To explain why s^2 rather than the average squared deviation is used to measure variability, note first that whereas s^2 measures sample variability, there is a measure of variability in the population called the population variance. We will use σ^2 (the square of the lowercase Greek letter sigma) to denote the population variance and σ to denote the population standard deviation (the square root of σ^2). When the population is finite and consists of N values,

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2/N$$

which is the average of all squared deviations from the population mean (for the population, the divisor is N and not $N - 1$). More general definitions of σ^2 appear in Chapters 3 and 4.

Just as $\bar{x}$ will be used to make inferences about the population mean μ, we should define the sample variance so that it can be used to make inferences about σ^2. Now note that σ^2 involves squared deviations about the population mean μ. If we actually knew the value of μ, then we could define the sample variance as the average squared deviation of the sample x_i's about μ. However, the value of μ is almost never known, so the sum of squared deviations about $\bar{x}$ must be used. But *the x_i's tend to be closer to their average $\bar{x}$ than to the population average μ, so to compensate for this the divisor $n - 1$ is used rather than n.* In other words, if we used a divisor n in the sample variance, then the resulting quantity would tend to underestimate σ^2 (produce estimated values that are too small on the average), whereas dividing by the slightly smaller $n - 1$ corrects this underestimating.

It is customary to refer to s^2 as being based on $n - 1$ **degrees of freedom** (df). This terminology results from the fact that although s^2 is based on the n quantities $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, these sum to 0, so specifying the values of any $n - 1$ of the quantities determines the remaining value. For example, if $n = 4$ and $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = -6$, and $x_4 - \bar{x} = -4$, then automatically we have $x_3 - \bar{x} = 2$, so only three of the four values of $x_i - \bar{x}$ are freely determined (3 df).