Computing the Sample Variance

Suppose that we have a data set $X_1, \ldots, X_n$. The usual sample variance is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$  \hspace{1cm} (1)

where $\overline{X}$ is the usual sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

For $n \geq 4$, a more efficient computing formula for $s^2$ is

$$s^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \right].$$  \hspace{1cm} (2)

Note that (1) requires $n$ differences, $n$ squarings, $n$ additions, and 1 division, for a total of $3n + 1$ arithmetical operations. On the other hand, (2) requires 1 difference, $n + 1$ squarings, $n$ additions, 2 divisions, and 1 multiplication, for a total of $2n + 5$ arithmetical operations. Now $3n + 1 \geq 2n + 5$ if and only if $n \geq 4$. Hence, for $n \geq 4$, (2) is computationally more efficient than (1), even though both formulas are mathematically equivalent.

**Proof of the Equivalence.** Beginning with (1), we have

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - 2\overline{X}X_i + \overline{X}^2)$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - 2\overline{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \overline{X}^2 \right]$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - 2\overline{X} n \overline{X} + n \overline{X}^2 \right]$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - 2n \overline{X}^2 + n \overline{X}^2 \right]$$

$$= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - n \overline{X}^2 \right]$$

$$= \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \right].$$

**Example.** For $n = 25$, (1) requires $3 \times 25 + 1 = 76$ arithmetical operations, whereas (2) requires $2 \times 25 + 5 = 55$ arithmetical operations.

-- RJS, 1/17/2011