Next Up Previous

Next: Frequency Distribution Revisited **Up:** 10.001: Data Visualization and **Previous:** Quantitative Description of the

Variance, Standard Deviation and Coefficient of Variation

The most commonly used measure of variation (dispersion) is the sample standard deviation, σ. The square of the sample standard deviation is called the sample variance, defined as[2]

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2. \qquad (2)$$

However,

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\mu\left(\sum_{i=1}^{n} x_i\right) + n\mu^2$$

$$= \left(\sum_{i=1}^{n} x_i^2\right) - 2n\mu^2 + n\mu^2$$

$$= \left(\sum_{i=1}^{n} x_i^2\right) - n\mu^2 . \qquad (3)$$

So an alternate equation for computing the variance is given by

$$\sigma^2 = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\left(\sum_{i=1}^{n} x_i\right)^2\right] . \qquad (4)$$

The advantage of Eq. 4 over Eq. 2 is that it allows for the computation of $\sum x_i^2$ required for the evaluation of σ and $\sum x_i$ required for the evaluation of μ in one loop, whereas Eq. 2 requires the precomputed value of μ before we can compute σ. For this reason, Eq. 4 is used often in the computations of the mean and variance.

However, if you closely examine Eq. 2 and Eq. 4, one important difference can be pointed out: Eq. 2 guarantees a non-negative variance because variance is given there as the sum of squares. This is not necessarily true of Eq. 4 where we subtract $n(\sum_{i=1}^{n} x_i)^2$ from $\sum_{i=1}^{n} x_i^2$. From a computational perspective, we know that this can cause difficulties for large samples prone to potential roundoff errors. So we are interested in developing an algorithm which computes (a). both the mean and the variance in the same loop and (b). variance as a sum of squares. How can this be accomplished? Well, we can resort to developing recursive relations. Applying Eq. 1 for the the first p - 1 and p data and subtracting one from the other, we get

$$p\,\mu_p = (p - 1)\,\mu_{p-1} + x_p, \tag{5}$$

where μ_p denotes the mean value of the first p data of the sample. We can now compute the sample mean

recursively by letting $\mu_1 = x_1$ and subsequently applying Eq. 5 for $p = 2, 3, \cdots, n$. We can also construct a simple recursion relation for computing σ^2 by applying Eq. 4 for the first p - 1 and p data in the sample. This gives the two equations

$$(p - 2)\,\sigma_{p-1}^2 = x_1{}^2 + x_2{}^2 + \cdots + x_{p-1}{}^2 - (p - 1)\,\mu_{p-1}^2$$

$$(p - 1)\,\sigma_p^2 = x_1{}^2 + x_2{}^2 + \cdots + x_p{}^2 - p\,\mu_p^2 . \qquad (6)$$

subtracting the first of Eq. 6 from the second one gives

$$(p - 1)\,\sigma_p^2 = (p - 2)\,\sigma_{p-1}^2 + (p - 1)\,\mu_{p-1}^2 + x_p{}^2 - p\,\mu_p^2, \qquad (7)$$

which can be rewritten (to get rid of subtractions) by substituting for μ_{p-1}^2 from Eq. 5 as follows:

$$(p - 1)\,\sigma_p^2 = (p - 2)\,\sigma_{p-1}^2 + p(x_p - \mu_p)^2/(p - 1),\ \ p = 2, 3, \cdots, n. \tag{8}$$

Now, once we initialize $\mu_1 = x_1$ and $\sigma_1 = 0$, we can compute the sample mean and variance using Eq. 5 and 8 for $p = 2, 3, \cdots, n$ within the same loop. Note that the variance thus computed is guaranteed to be non-negative.

The coefficient of variation of the sample data, denoted by CV is defined as

$$\mathrm{CV} = \frac{\sigma}{\mu}. \tag{9}$$

Note that CV is independent of the units of measurement.

Next Up Previous

Next: Frequency Distribution Revisited **Up:** 10.001: Data Visualization and **Previous:** Quantitative Description of the

Michael Zeltkevic

1998-04-15