A6 Estimation of the variance

Why is the best estimate for the variance larger than the mean squared deviation of the average?

Assume x_i are independent samples from a distribution $f(\mu, \sigma)$ with mean μ and s.d. σ. In order to find out the relation between $\langle(\Delta x)^2\rangle$ and σ it is necessary to compute the expectation of $\langle(\Delta x)^2\rangle$. After realizing that

$$\begin{aligned}\langle(\Delta x)^2\rangle &= \langle(x - \langle x\rangle)^2\rangle \\ &= \langle[x - \mu - (\langle x\rangle - \mu)]^2\rangle, \\ &= \langle(x-\mu)^2\rangle - (\langle x\rangle - \mu)^2, \end{aligned} \tag{A6.1}$$

we see that

$$\begin{aligned} E[\langle(\Delta x)^2\rangle] &= \sigma^2 - E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right)^2\right] \\ &= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[(x_i-\mu)(x_j-\mu)] \end{aligned} \tag{A6.2}$$

Uncorrelated data points

When all samples are independent of each other (and therefore uncorrelated),[1] the double sum reduces to a single sum because x_i and x_j are

[1] The terms *independent* and *uncorrelated* mean different things. Two random variables x and y are statistically independent when the random processes selecting either of them are independent of each other; x and y are statistically uncorrelated when $E[(x-\mu_x)(y-\mu_y)] = 0$. Independent samples are also uncorrelated, but uncorrelated samples need not be independent. For example, the random variable x sampled from $N(0, 1)$ and x^2 are uncorrelated (because $E[x^3] = 0$), but they are not independent!

independent and only the term $j = i$ in the second sum survives:

$$\begin{aligned} E[\langle(\Delta x)^2\rangle] &= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n} E[(x_i - \mu)^2] \\ &= \sigma^2\left(1 - \frac{1}{n}\right). \end{aligned} \tag{A6.3}$$

Thus it follows that the best estimate for σ^2 equals $n/(n-1)$ times the average of the squared deviations from the average. Note that this is true for any kind of distribution with finite variance.

Correlated data points

In the derivation of (A6.3) explicit use has been made of the assumption that the deviations from the mean are uncorrelated. In practice subsequent data points are often correlated, i.e., $E[(x_i - \mu)(x_j - \mu)] \neq 0$ for $j \neq i$. If the latter is the case, more terms will remain in the double sum of (A6.2) and a larger term will be subtracted from σ^2. The best estimate for the variance will be larger.

Here we shall give the equation for the cases that a known correlation between *successive* data points exists (Straatsma *et al.*, 1986).[2] The assumption is made that the ordered series $x_1, \ldots, x_n$ is a *stationary stochastic variable*, i.e. it has a constant variance and correlation coefficients between x_i and x_j that only depend on the distance $|j - i|$.

The term with the double sum in (A6.2) is:

$$\frac{1}{n^2}\sum_i \sum_j E[(x_i - \mu)(x_j - \mu)] = \sigma^2 \frac{n_c}{n}, \tag{A6.4}$$

where n_c is a kind of *correlation length*:

$$n_c = 1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k. \tag{A6.5}$$

Here ρ_k is the correlation coefficient between x_i and x_{i+k}:

$$E[(x_i - \mu)(x_{i+k} - \mu)] = \rho_k \sigma^2. \tag{A6.6}$$

[2] See reference list on page 124.

$$\begin{array}{c|c|c|c|c|c|}
 & i \rightarrow & & & & \\
\hline
j \downarrow & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\
\hline
 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\
\hline
 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\
\hline
 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\
\hline
 & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \\
\hline
\end{array}$$

Figure A6.1 The correlation matrix ρ_{ij} for the example $n = 5$. If all elements are summed by adding diagonally, one obtains $5 + 2(4\rho_1 + 3\rho_2 + 2\rho_3 + \rho_4)$.

For series that are much longer than the correlation length ($k \ll n$), (A6.5) can be reduced to

$$n_c = 1 + 2\sum_{k=1}^{\infty} \rho_k. \tag{A6.7}$$

Equations (A6.4) and (A6.5) follow simply from counting the number of occurrences in the double sum of (A6.4). Figure A6.1 elucidates how (A6.5) is obtained by summing all matrix elements.

Instead of (A6.3) we now obtain as a result:

$$E[\langle(\Delta x)^2\rangle] = \sigma^2 \left(1 - \frac{n_c}{n}\right). \tag{A6.8}$$

The effect of correlation in the data series on the estimate of σ is not very large and can generally be neglected. However, the effect on the estimate for the standard inaccuracy of the average is quite large and not negligible. This is treated in Appendix A7.