A7 *Standard deviation of the mean*

Why is the variance of the mean ⟨*x*⟩ of *n* independent data equal to the variance of *x* itself divided by *n*?

We investigate the following quantity:

$$\mathrm{var}\,(\langle x\rangle) = E[(\langle x\rangle - \mu)^2] = \frac{1}{n^2}E\left[\left\{\sum_i (x_i - \mu)\right\}^2\right] \tag{A7.1}$$

$$= \frac{1}{n^2}\sum_i \sum_j E[(x_i - \mu)(x_j - \mu)]. \tag{A7.2}$$

For uncorrelated data $E[(x_i - \mu)(x_j - \mu) = \sigma^2 \delta_{ij}$. Therefore

$$\mathrm{var}\,(\langle x\rangle) = \sigma^2/n \tag{A7.3}$$

and

$$\sigma_{\langle x\rangle} = \frac{1}{\sqrt{n}}\sigma. \tag{A7.4}$$

How is this result influenced when the data are correlated?

For this we need to work out the double sum in (A7.2). We have already done that in Appendix A6, see (A6.4), for the case of an ordered sequence in which correlations depend only on the distance $k = |j - i|$. It was found that

$$\mathrm{var}\,(\langle x\rangle) = \sigma^2 \frac{n_c}{n}, \tag{A7.5}$$

where n_c is the correlation length, defined in Appendix A6 in (A6.7). So you see that correlations in the data tend to increase the uncertainty of the mean. It is as if the effective number of data points is less than the number you actually have. In order to make a reliable estimate of the uncertainty, you need to know the correlation length n_c, or deduce n_c from the data. This is in general not a simple task because the correlation between data points is difficult to evaluate, especially for large intervals. The correlation length is an integral over the correlation function, which is notoriously difficult to determine from noisy data.[1]

A practical alternative to the summing of correlation coefficients is the *block average* procedure:[2] group blocks of sequential data together and consider the average of each block as a new data point. If most of the sequential correlation is located within a block, the block averages are mutually almost uncorrelated and can be treated by standard methods. For example, if you have 1000 data points and you expect the correlation to stretch over some 10 or 20 points, then choose 10 blocks of 100 points each. Much better is to vary the block length and check if the results have a reliable limit. This "block average" procedure is not exact because there is always some correlation left between successive blocks, but it is very practical.

Example

Time series generated by Monte Carlo or molecular dynamics simulations often contain significant sequential correlations that complicate the determination of the inaccuracies of averages. In a dynamic simulation of a molecular system a time series of 20 000 data points (t, T) is generated with the "temperature" T (derived from the total kinetic energy) at times t in steps of 0.009 ps. Applying the rules for uncorrelated samples, the average temperature appears to be 309.967 ± 0.022 K. This inaccuracy is likely to be far too low in view of the expected sequential correlation of the data points T. What is the true standard inaccuracy? Figure A7.1 plots the first 500 points versus time: it is apparent that correlation persists over times of the order of picoseconds and includes some oscillatory behavior. Figure A7.2 plots the standard inaccuracy estimated from a series of block averages, with block sizes varying between 1 and 400 points, or 0.009 to 3.6 ps. A plateau of 0.11 K is reached after about 2 ps. This plateau value is 5 times larger than the "uncorrelated value," indicating that the statistical correlation length n_c is some 25 time steps or 0.22 ps. The block length must be taken several times larger than the correlation length for the blocks to be statistically independent. The final result for the average temperature is 309.97 ± 0.11 K.

[1] A discussion of alternative methods to determine the accuracy of the mean of correlated data can be found in Hess (2002), see reference list on page 124.

[2] The method is similar, but not identical to the "jackknife procedure," which estimates the mean and variance by averaging over data sets from which subgroups of data have been omitted. See Wolter (2007), Chapter 4, in the reference list on page 124.

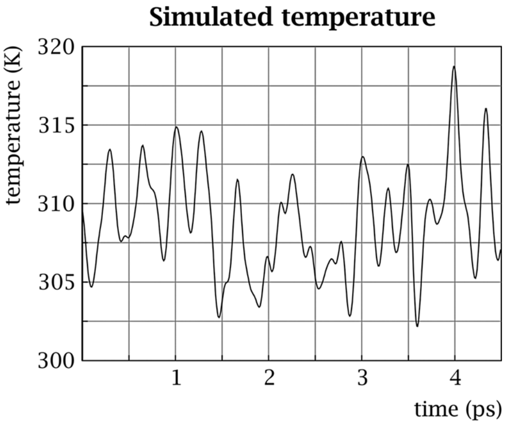

Figure A7.1 The first 500 points of a 20 000 point data set with temperatures derived from the kinetic energy as a function of time of a molecular dynamics simulation of a molecular system. The time interval between points is 0.009 ps.

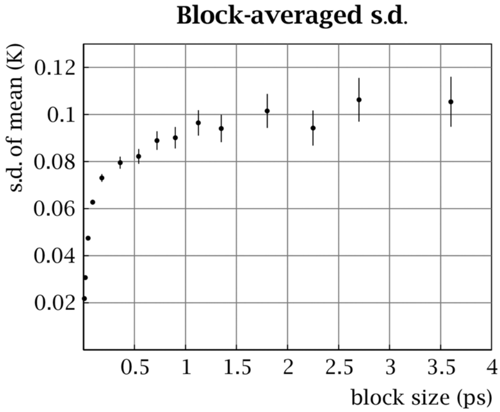

Figure A7.2 Estimates of the inaccuracy (standard deviation) in the mean using block averages of 20 000 data points with temperature data (from a molecular dynamics simulation) as function of time. The block averages are assumed to be uncorrelated. The block size varies from 1 point (0.009 ps) to 400 points (3.6 ps). The error bars indicate the uncertainties in the s.d. based on the limited number n_b of block averages, which amounts to a relative error of $1/\sqrt{2(n_b - 1)}$.

Python code 7.1 on page 195 shows how the standard inaccuracy of averages can be estimated from a set of block averages.

How accurate is the estimated standard deviation?

Because the variance of a distribution is estimated from the sum of squared deviations from the average (divided by $n - 1$), the statistics of the variance satisfies the statistics of a sum of squares of random samples. For normally distributed samples, this sum follows the "chi-squared distribution" (see Section 7.4 and data sheet CHI-SQUARED DISTRIBUTION on page 199). A chi-squared distribution has a mean ν and a variance 2ν; hence it has a relative s.d. of $\sqrt{2/\nu}$, where ν is the number of degrees of freedom: $\nu = n - 1$. So the relative s.d. of the variance is $\sqrt{2/(n-1)}$ and the relative s.d. of the standard deviation itself is $1/\sqrt{2(n-1)}$. This result is valid for normally distributed independent samples. Any sequential correlation will increase the inaccuracy.