T.T. Rantala, 1996

# MISC. THOUGHTS ABOUT ERROR ESTIMATION

## A. DEFINITIONS

Let $X$ be a random variable that is assumed to have *normal distribution* (or close enough) where essential. Let $\mu$, $\sigma^2$ and $\sigma$ denote the *mean*, *variance* and *standard deviation* of the distribution.

Consider a sample $\{X_i\}$, where $i = 1, 2, ..., n$, with the *sample average*

$$\bar{X} = \frac{\sum_i X_i}{n} \tag{A1}$$

and *standard deviation (of the sample)* $S = \sqrt{S^2}$ obtained from the *sample variance*

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}. \tag{A2}$$

These are unbiased estimates of the distribution statistics,

$$\langle \bar{X} \rangle = \mu \tag{A3}$$

and

$$\langle \bar{S^2} \rangle = \sigma^2. \tag{A4}$$

*Root mean square* (RMS) of the sample distribution is defined as

$$RMS = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n}} \approx S. \tag{A5}$$

Sample average $\bar{X}$ is (at least closely) normally distributed with the standard deviation called *standard error of mean*

$$SEM(n) = \frac{\sigma}{\sqrt{n}}. \tag{A6}$$

It is essential to note that $SEM(n)$ is sample size dependent.

If the data of the sample do not have (internal) correlations $\bar{X}$ is within the limits $\mu \pm SEM$, $\mu \pm 2\,SEM$ or $\mu \pm 3\,SEM$ with the probability 68%, 95% or more than 99%, respectively. Therefore, error bars of size $SEM$, $2\,SEM$ or $3\,SEM$ have well-defined meanings and are useful in defining statistical error limits.

## B. STANDARD DEVIATION OF GROUP AVERAGES

Often, a useful way to estimate $SEM(n)$ is the "method of group averages". Let the (uncorrelated) data be divided into $k$ subgroups of $m$ data, i.e.,

$$n = k\,m. \tag{B1}$$

Now, the standard deviation $S(m)$, from

$$S^2(m) = \frac{\sum_{l=1}^{k}(\bar{X}_l - \bar{X})^2}{k-1}, \tag{B2}$$

of the $k$ group averages $(l = 1, 2, ..., k)$

$$\bar{X}_l = \frac{\sum_{i=1}^{m} X_{i,l}}{m} \tag{B3}$$

yields

$$SEM(m) = S(m). \tag{B4}$$

According to (A6) we finally obtain

$$SEM(n) = \sqrt{\frac{m}{n}}S(m) = \frac{S(m)}{\sqrt{k}}. \tag{B5}$$

Note that $S(1) = S$ by (A2) and $S(n) = SEM(n)$ by (A6).

## C. SUMMING INDEPENDENT ERRORS

Suppose that we deal with a random variable

$$X = X^a + X^b, \tag{C1}$$

where $X^a$ and $X^b$ are normally distributed (or close enough). As the distribution of $X$ is a convolution of the distributions of $X^a$ and $X^b$, for the variances it holds

$$S^2 = {S^a}^2 + {S^b}^2. \tag{C2}$$

Furthermore, because for the sample averages obviously

$$\bar{X} = \bar{X^a} + \bar{X^b}, \tag{C3}$$

we obtain for the standard error of mean

$$SEM^2 = {SEM^a}^2 + {SEM^b}^2. \tag{C4}$$

## D. CORRELATED DATA

Now, let us consider data with the type of ”internal correlations” that may follow, e.g., from recording the successive values (in time) of a physical observable $X(t)$. This is the case of the dynamical variable along the trajectory (or path), which is a solution of the equations of motion in molecular dynamics, for example. It is obvious that increasing the sampling frequency beyond some limit,

$$f_c = \frac{1}{\Delta t_c}, \tag{$D1$}$$

where $\Delta t$ is the time step, the recorded data values start approaching interpolates between the adjacent ones rather than new independent information of the observable.

In such a case (A6) cannot be used to estimate the standard deviation of the sample average $\bar{X}$, i.e., $SEM$. However, a straightforward but careful use of (B4) can be used to evaluate at least an upper limit for it, as shown in sec. E, below. But here, we will first consider a more advanced method, based on the direct evaluation of the ”correlations” in the data.

Let us first define some concepts for stationary (or close enough) time series $\{X_i\} = \{X(t)\}$, where $t = t_i = t_0 + i\ \Delta t$ and $i = 1, 2, ..., n$. The covariance of $\{X_i\}$ and $\{Y_i\}$ is defined as

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \langle (X - \mu_X)(Y - \mu_Y) \rangle = \langle XY \rangle - \mu_X\ \mu_Y. \tag{$D2$}$$

Thus, we have $\sigma_X^2 = \sigma_{XX} = \mathrm{Var}(X)$, which is the variance (A4). A related quantity is the correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\ \sigma_Y^2}}. \tag{$D3$}$$

Thus, $\sigma_{XY} = \rho_{XY}\ \sigma_X\ \sigma_Y$. Next, we define the autocovariance of $\{X_i\}$ or $\{X(t)\}$ as

$$\gamma_X(\tau) = \mathrm{Cov}(X(t), X(t \pm \tau)) = \langle X(t)\ X(t \pm \tau) \rangle_t - \bar{X}^2 \tag{$D4$}$$

and autocorrelation function as

$$\rho_X(\tau) = \frac{\gamma_X(\tau)}{\gamma_X(0)}. \tag{$D5$}$$

Note that $\gamma_X(0) = \sigma_X^2$, the variance of $X(t)$.

Now consider error estimation of the sample average $\bar{X}$ of the "correlated" data $\{X(t)\}$ defined above. It can be shown (Wayne A. Fuller, *Introduction to Statistical Time Series*, Wiley 1976) that

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_X(t_i - t_j) = \frac{1}{n} \sum_{i=1-n}^{n-1} \left(1 - \frac{|i|}{n}\right) \gamma_X(t_i), \tag{$D6$}$$

which implies

$$SEM = \frac{1}{n} \sqrt{\sum_{i=-(n-1)}^{n-1} \big(n - |i|\big)\, \gamma_X(t_i)}, \tag{$D7$}$$

as $\mathrm{Var}(\bar{X}) = SEM^2$. In case of uncorrelated data, for which $\gamma(\tau) \equiv 0$, if $\tau \neq 0$, (D7) reduces to $\frac{1}{n}\sqrt{n\ \gamma_X(0)} \longrightarrow \sigma_X/\sqrt{n}$, which is (A6), as expected.

For evaluation of $SEM$ from (D7) one needs to compute $\gamma_X$ from (D4) first. In practice, the fluctuations in $\gamma_X(\tau)$ become large for large $\tau$, which has shown to introduce uncertainty in what (D7) yields for $SEM$. Therefore, the autocovarience function has to be cut or smoothened at large $\tau$.

## E. CORRELATED DATA, A SIMPLE APPROACH

Here, we suggest a staightforward approach to estimate the statistical error limit $SEM$ of the sample average $\bar{X}$ of correlated data, based on concepts given in section B. Hence again, let us divide the trajectory into $k$ pieces of equal length, and consequently, the recorded data to $k$ subgroups of $m$ successive values of the observable,

$$n = k\, m. \tag{$E1$}$$

Next consider the function $S(m)$ as defined by (B2). For a **sufficiently large** $m$, with a good approximation,

$$S(m) \propto \frac{1}{\sqrt{m}}, \tag{$E2$}$$

if the "correlation length" is shorter than $m$. With a constant $k$ the correlation length $m_c$ and the sampling frequency $f$ are proportional and

$$m_c \approx \frac{f}{f_c}. \tag{$E3$}$$

For small $m$, $m < m_c$, on the other hand, $S(m)$ can be expected to be constant. This can be concluded from the fact that increasing $f$ (or $m$) with constant $k$ should not have any effect on the group averages $\bar{X}_l$, and thus, on their standard deviation. Therefore,

$$S(m_c) \approx \sigma. \tag{E4}$$

Note that from the above it follows that $\sqrt{m}S(m)$ is constant for large $m$.

Finally, we estimate the $SEM$ (with the confidence limits given above). If $m_c$ is found (by inspection, for example), from

$$\sqrt{m_c}\ S(m_c) = \sqrt{n}\ S(n)$$

we obtain

$$S(n) = \sqrt{\frac{m_c}{n}}\ S(m_c), \tag{E5}$$

which is the standard deviation of the group averages for the sample size $n$, i.e., it has the confidence limits of $SEM$. By defining the "number of uncorrelated" (or independent) data values

$$n_c = \frac{n}{m_c} \tag{E6}$$

of the sample and using (E4) and (E5), we can finally write

$$SEM(n_c) = \frac{\sigma}{\sqrt{n_c}}. \tag{E7}$$

Also, it is worth of noticing, that the correlations in the time series and the limiting frequency $f_c$ may be inspected and estimated using *Fourier transformations* and the *sampling theorem.*

## APPENDIX a: APPLICATION OF THE THEORY OF SEC. E TO THE ORIENTATION PARAMETERS OF *PLPC* FATTY ACID CHAIN

Consider a system of 36 ($= n_M$) PLPC molecules each containing several carbon chain segments, whose orientation and statistical errors we wish to evaluate. For simplicity, in what follows we talk about one segment or one orientation parameter $X$ as an example, and as a numerical value take a rough average over segments. The molecular dynamics runs over 7200 ($= n_t$) time steps, and thus, we have a collection of

$$n = n_M \, n_t \qquad (a1)$$

data points $X_{ij}$ (for each segment). Defining the indices as $i = 1, 2, ..., n_t$ and $j = 1, 2, ..., n_M$, we can expect strong "internal correlation", as defined in section D, for a certain $j$ as $i$ runs over the successive data in time domain. On the other hand, for a certain $i$ there may be structural correlations through molecular interactions for different $j$.

The computed figures are as follows. In parentheses the limits over the different segments are given, followed by "an average" (median) value as an example. The standard deviation of the whole data

$$S = (0.40 \; ... \; 0.46) = 0.43 \qquad (a2)$$

and according to the (E7)

$$\sqrt{\frac{n_c}{n}} SEM(n_c) = (0.0008 \; ... \; 0.0009) = 0.001, \qquad (a3)$$

but $n_c$ is unknown.

On the other hand, the standard deviation of the $n_M$ time averaged (over $n_t$ data values) group averages $\bar{X}_j^t$ is

$$S^M(n_t) = (0.18 \; ... \; 0.42) = 0.3, \qquad (a4)$$

which implies according to (B5)

$$SEM(n_c) = \frac{S^M(n_t)}{\sqrt{n_M}} = (0.03 \; ... \; 0.07) = 0.05. \qquad (a5)$$

This is a good measure of the statistical error and can be used as an error bar, too.

Now, from (a3) and (a5) we obtain $n_c \approx 100$, or a period of 0.1ps ($10^{-13}$s), for the correlation length. This rough estimate is of the order of typical molecular vibration period (as the vibration of hydrogen atoms is hindered, now). This is actually what one would expect.

The standard deviations attached to each of the group averages $\bar{X}_j^t$ are

$$S_j^t = (0.24 \ ... \ 0.36) = 0.3. \qquad (a6)$$

Now, for the orientation parameter let us suppose that

$$X = X^M + X^t, \qquad (a7)$$

where $X^M$ is sc. *structural distribution* among the PLPC molecules and $X^t$ is a *dynamic distribution* due to the thermal motion in time domain. For a "long enough" time period this separation becomes meaningless as the structural distribution decays to the dynamic one. For shorter periods, however, we can try the following analysis, though its relevance and limitations should be carefully considered.

First note that ($\bar{S}^t$ or)

$$S_j^t \approx 0.3 < S \approx 0.43 \qquad (a8)$$

and

$$S^M(n_t) \approx 0.3 < S \approx 0.43. \qquad (a9)$$

The former inequality suggests that some contribution from the structural distribution beyond $S_j^t$ remains in $S$ and the latter implies that the structural distribution has not decayed to the dynamic one, yet. Furthermore, in the spirit of section E we see that the (E2) holds very well

$$S^{M^2} + S^{j^2} = 0.42 \approx S. \qquad (a10)$$

Hence, the model of two independent contributions to the orientation parameter $X$ works well. However, as limitations in the model we should note that all the involved distributions are not normal. Even the range of the orientation parameter as well as its separate components ($X^M$ and $X^t$) is limited to $-0.5$ ... 1.