# DATA REDUCTION BY AVERAGING

**Grant Foster**
AAVSO
25 Birch Street
Cambridge, MA 02138

*Presented at the AAVSO Annual Meeting, October 28, 1995*

**Abstract**

In many cases, a time series with very many observations can, by averaging over an appropriate time span, be reduced to a manageable number of data points with very little loss of information. I investigate the errors inherent in this process.

## 1. Introduction

The quality of visual observations of variable stars is often called into question because they exhibit a relatively high inherent scatter, usually about 0.2–0.4 magnitude. However, for stars with many observers, or a few prolific observers, the lack of "quality" of a single visual observation is more than compensated by the large quantity of data available. Yet we must pay a price for this bounty: any mathematical analyses performed on such a very large number of data points (in some cases exceeding 50,000 for a single star) will consume considerable computer time and memory. It is therefore often advantageous to reduce the data, by some smoothing technique, to a smaller number of more precise values. One of the most basic and most common methods is *averaging*. We split the observed time span into small bins, usually of equal duration $T$, averaging all the data in a particular bin to generate a single reduced datum. We also obtain thereby an estimate of the scatter $\sigma$ within each bin and an estimated error of the average itself $\sigma/\sqrt{N}$, where $N$ is the number of data in the bin.

Often this is done without much consideration of the appropriate bin size. It should be chosen large enough to encompass many data points, so that the reduced values are as precise as possible. Yet it must be chosen small enough that the underlying signal does not evolve significantly during the entire duration of the bin. We shall consider the effect of the bin size $T$ on the probable error associated with data averaging. It is sometimes the practice to choose the midpoint of the bin as the "reduced time," i.e., the time coordinate of the reduced datum, and sometimes the practice of taking the average time as the reduced time. We shall consider the errors associated with both choices.

## 2. Modeling the underlying signal

We assume that the data (which for illustrative purposes we will treat as variable star magnitudes) are the sum of an underlying physical signal $f(t)$ and errors $\varepsilon$, as

$$x_n = x(t_n) = f(t_n) + \varepsilon_n \,. \qquad (1)$$

The errors are assumed to be independent random variables with mean zero and variance $\sigma^2$, i.e.,

$$<\varepsilon_n> = 0 \qquad <(\varepsilon_n)^2> = \sigma^2 \,, \qquad (2)$$

where enclosing any quantity in brackets "<>" denotes the expected value of a random variable, or the average value of an observed variable. We further assume that the bin size $T$ is small enough that for the duration of the bin, the signal $f(t)$ can be described

by a low-order polynomial (a Taylor's series) about some time $t_0$ within the bin

$$f(t) = m_0 + \alpha(t - t_0) + \beta(t - t_0)^2 + \gamma(t - t_0)^3 + \ldots \quad (3)$$

We shall seek to estimate the value of the signal at the time $t_0$.

Since we are only interested in the lowest-order non-zero error due to signal variation, most of the time we can reliably model the signal over a small time span by a simple linear approximation

$$f(t) = m_0 + \alpha(t - t_0). \quad (4)$$

However, this approximation is insufficient for two reasons. First, whenever the signal reaches an extremum, its time derivative is zero and our simple approximation (4) reduces to a constant. It therefore fails to model any variation of the signal at extrema; we must include higher-order terms. Second, as we shall soon see, if the reduced time $t_0$ is chosen as the average time within the bin, then the linear term (4) introduces no error at all; we must include higher order terms to get to the first nonzero error term.

Including linear and quadratic terms is also insufficient. If we choose $t_0$ as the average time, then the linear term produces no error, and at any time at which the second time derivative of the signal is zero, the quadratic term vanishes; the lowest-order term actually contributing to the error will be the cubic term. We shall therefore adopt a cubic polynomial in time

$$f(t) = m_0 + \alpha(t - t_0) + \beta(t - t_0)^2 + \gamma(t - t_0)^3, \quad (5)$$

as the approximate form of the physical signal for the duration of the bin.

### 3. Extreme values of powers of time

We can now compute the expected value of the average of all the data in the bin. We have (using the fact that $<\varepsilon> = 0$)

$$<x> = m_0 + \alpha<(t - t_0)> + \beta<(t - t_0)^2> + \gamma<(t - t_0)^3>. \quad (6)$$

The true value of the signal at time $t_0$ is $m_0$. We therefore have three error terms, due to 1st, 2nd, and 3rd-order signal variations

$$E_1 = |\alpha<(t - t_0)>|, \quad (7)$$

$$E_2 = |\beta<(t - t_0)^2>|, \quad (8)$$

$$E_3 = |\gamma<(t - t_0)^3>|. \quad (9)$$

Note that we have taken absolute values, to consider the size of the separate errors, not their signs. I emphasize that (7), (8), and (9) are *not* errors associated with measurement or observation; they refer only to the bias introduced by the averaging process itself.

First consider the 1st-order term. For data in a bin of width $T$, centered at time 0, all data points have

$$|t_n| \leq \tfrac{1}{2}T. \quad (10)$$

Therefore, if the reduced time is the bin midpoint ($t_0 = 0$) then

$$|<(t - t_0)>| = |<t>| \leq \tfrac{1}{2}T \qquad (t_0 = 0), \quad (11)$$

whereas if the reduced time is the average time, then by definition

$$t_0 = <t>, \quad (12)$$

so that

$$|\langle (t - t_0)\rangle| = 0 \quad (t_0 = \textit{average time}). \tag{13}$$

For the 2nd-order term, we clearly have for all data points

$$t_n^2 \le (\tfrac{1}{2}T)^2 , \tag{14}$$

so that

$$\langle (t - t_0)^2\rangle \le (\tfrac{1}{2}T)^2 . \tag{15}$$

This limit applies to both cases, when $t_0 = 0$ and when $t_0$ = average time.

Finally, for the 3rd-order term we have

$$|t_n^3| \le (\tfrac{1}{2}T)^3 , \tag{16}$$

so that

$$|\langle (t - t_0)^3\rangle| \le (\tfrac{1}{2}T)^3 \qquad (t_0 = 0). \tag{17}$$

For the case $t_0$ = average time, we can compute the maximum possible $|\langle (t - t_0)^3\rangle|$ by considering the extreme cases of $n_1$ data points at $t = -\frac{1}{2}T$ and $n_2$ data points at $t = +\frac{1}{2}T$. We find that in all cases,

$$|\langle (t - t_0)^3\rangle| \le (\tfrac{1}{2}T)^3 /10 \qquad (t_0 = \langle t\rangle). \tag{18}$$

These are the extreme-case limits on absolute values of powers of time.

## 4. Extremes of polynomial coefficients

It remains to estimate upper limits for the polynomial expansion coefficients $\alpha$, $\beta$, and $\gamma$. Of course, they cannot be estimated in general, except by actually fitting a smooth curve to the data. We can, however, compute the coefficients for a pure sinusoid of (semi-) amplitude $A$ and frequency $\nu$ (period $P = 1/\nu$). Without loss of generality, we may take the signal as

$$f(t) = A\cos(2\pi\nu t), \tag{19}$$

from which it is straightforward to compute the limits

$$|\alpha| \le 2\pi A\nu, \tag{20}$$

$$|\beta| \le 2\,\pi^2 A\nu^2, \tag{21}$$

$$|\gamma| \le \tfrac{4}{3}\pi^3 A\nu^3. \tag{22}$$

To be conservative in error estimation, we shall adopt *twice* these values as upper limit estimates for absolute values of the coefficients, for a variable star of period $P = 1/\nu$. Therefore we shall use the limits

$$|\alpha| \le 4\pi A\nu, \tag{23}$$

$$|\beta| \le 4\pi^2 A\nu^2, \tag{24}$$

$$|\gamma| \le \tfrac{8}{3}\pi^3 A\nu^3. \tag{25}$$

This leads to the following error limits:

for $t_0 = 0$ (bin midpoint):

$$E_1 \le 4\pi A\nu(\tfrac{1}{2}T) = 2A(\pi\nu T), \tag{26}$$

$$E_2 \le 4\pi^2 A\nu^2(\tfrac{1}{2}T)^2 = A(\pi\nu T)^2, \tag{27}$$

$$E_3 \le \tfrac{8}{3}\pi^3 A\nu^3(\tfrac{1}{2}T)^3 = \tfrac{1}{3}A(\pi\nu T)^3. \tag{28}$$

For $t_0$ = average time:

$$E_1 = 0, \tag{29}$$

$$E_2 \leq 4\pi^2 A\nu^2(\tfrac{1}{2}T)^2 = A(\pi\nu T)^2, \tag{30}$$

$$E_3 \leq \tfrac{8}{3}\pi^3 A\nu^3(\tfrac{1}{2}T)^3/10 = A(\pi\nu T)^3/30. \tag{31}$$

The first thing to notice is that the errors are quite different in the two cases. The second-order error is the same, but the third-order error is ten times smaller when we take the time coordinate as the average time rather than the bin midpoint. Most important, if we take the bin midpoint as our reduced time, the first-order error can be sizeable, but by taking the average time we reduce it to zero. I therefore draw the following general rule:

**The time coordinate of the averaged datum should *always* be chosen as the average time, *never* as the bin midpoint.**

Taking this advice, our errors are given by (29) (which is conveniently zero), (30), and (31). Unless $\pi\nu T > 30$ (which we needn't really worry about), the second-order error is larger than the third-order. Let the tolerance $E_T$ be the maximum tolerable error due to the biasing introduced by the averaging procedure. Then we require that

$$E_2 \leq E_T. \tag{32}$$

This amounts to

$$A(\pi\nu T)^2 \leq E_T, \tag{33}$$

or simply

$$T \leq \frac{1}{\pi\nu}\sqrt{\frac{E_T}{A}}. \tag{34}$$

## 5. Example

Surely an example will illuminate the procedure. Consider the much-analyzed variable star o Ceti (Mira). It has a period of about 330 days (Kholopov *et al.* 1985) and (semi-)amplitude of 3 magnitudes. Because there are so many data for Mira, it is usual to analyze not the raw data but 10-day averages ($T = 10$). Furthermore, a maximum error level of 0.05 magnitude in the reduced values is acceptable for most analyses.

Straightforward application of (34) indicates that we can reduce the data by averages, as long as the bin size $T \leq 13.6$ days. For $T = 10$, from equations (26)–(31) we have the following errors due to averaging: for $t_0 = 0$ (bin midpoint):

$$E_1 \leq 2A\pi\nu T = 0.57 \text{ magnitude}, \tag{35}$$

$$E_2 \leq A(\pi\nu T)^2 = 0.027 \text{ magnitude}, \tag{36}$$

$$E_3 \leq A(\pi\nu T)^3/3 = 0.00086 \text{ magnitude}. \tag{37}$$

For $t_0$ = average time:

$$E_1 = 0, \tag{38}$$

$$E_2 \leq A(\pi\nu T)^2 = 0.027 \text{ magnitude}, \tag{39}$$

$$E_3 \leq A(\pi\nu T)^3/30 = 0.000086 \text{ magnitude}. \tag{40}$$

Two points become clear. First, the 3rd-order error is so much smaller than the 1st- and 2nd-order errors that we needn't worry about it. Clearly these errors are within the acceptable 0.05 magnitude range for Mira. Therefore, using 10-day averages is an

acceptable procedure for Mira. Second, the 1st-order error, which is always zero for $t_0 = < t >$, can be *huge* when we take the reduced time as the bin midpoint ($t_0 = 0$). This validates our requirement that the time coordinate of the averaged datum should be the average time.

## 6. Conclusion

Averaging is an excellent way to reduce the sheer size of a large data set, with very little loss of information. Yet we must be aware of two things: first, we must always record the average time, as well as the average magnitude; and second, if the time span over which we average is too large, then a bias is introduced by the averaging process itself.

## 7. Acknowledgement

I wish to express my sincere thanks to the multitude of AAVSO observers, who have given us a treasure trove of data to analyze.

## Reference

Kholopov, P. N. *et al.* 1985, *General Catalogue of Variable Stars*, 4th ed., Moscow.