

Confidence Limits on Simulation Averages

In this section we consider the means taken to estimate the quality of a simulation average. A simulation result can be compromised in many ways. Some, such as conceptual mistakes and programming errors, are entirely avoidable. Other sources of error are more difficult to eliminate. It may be that the simulation algorithm is simply incapable of generating a good sample of configurations in a reasonable amount of time. This section does not deal with these issues. We assume the programmer is competent, and that the simulation yields a good sample. Further, we do not consider “errors” in the molecular model, errors that cause the simulation results to differ from the true behavior of the system it is meant to describe. At present there is very little rigorous means to gauge *a priori* the quality of a result that is meant to reproduce or predict quantitative experimental measurements; this is indeed a very difficult problem.

At the end of a simulation, we have a value for some property, and the number that the computer reports to us is expressed to perhaps 16 digits of precision. We would like to know how many of these digits are good, and how much of it is noise. In this regard, error analysis applied to simulation is little different than that commonly employed in experimental measurement.

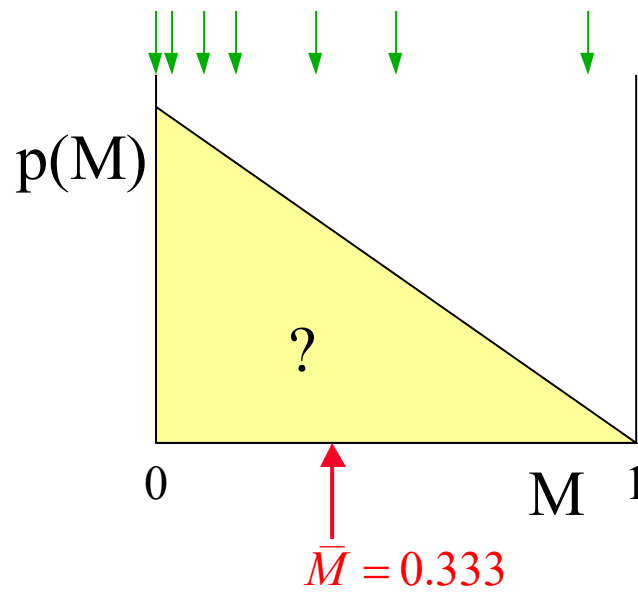
Let us speak generally then, without regard to whether we are conducting an experiment or a simulation. In both cases the aim is to take a set of independent measurements $\{m_i\}$ of some property M . For the sake of example, let us say that M naturally ranges from zero to unity, and that we have taken 7 measurements of M ; the measured values are

$\{0.01, 0.1, 0.9, 0.06, 0.5, 0.3, 0.02\}$

This is all we know about the system we are studying. In fact, there is some underlying probability distribution that governs (or at least characterizes) the measurement process, but we have no specific details information about the nature of this distribution. Let us for this example say that it is the triangular distribution depicted in Illustration 1. In this case, the probability of observing a value in the range M to $M+dM$ is $p(M) = 2(1-M)$. We emphasize that this detail about the distribution is completely unknown to us, and it is not even the aim of the experiment to uncover this detail. Instead, it is desired to know only the mean of the distribution, \bar{M} . In this example, the true mean is $1/3$, and it is the aim of the experiments to reveal this fact.

We would like to obtain the best possible estimate of \bar{M} from the measurements. Not surprisingly, this is given by the mean of the measurements $\bar{M} \approx \frac{1}{n} \sum_{i=1}^n m_i \equiv \langle m \rangle$. Note we designate \bar{M} as the true mean, and $\langle m \rangle$ as the average of our measurements. For our example, $\langle m \rangle = 0.2700$. This value differs from the correct result of 0.3333 , but from the experimental information available to us, we yet have no way to know the magnitude of our error. Looking at our data, we like to know if there's a good chance that the correct

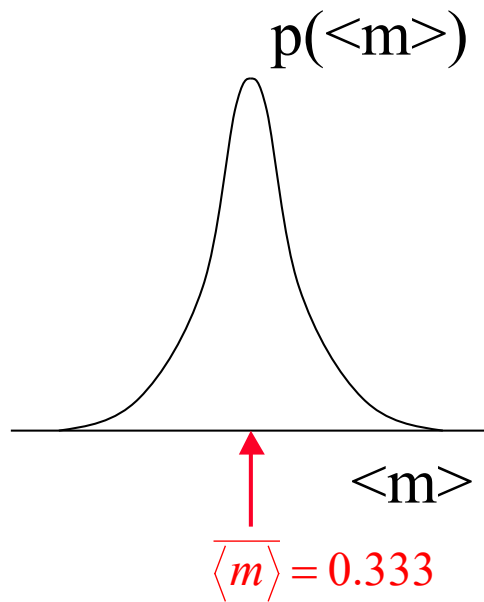
$\{0.01, 0.1, 0.9, 0.06, 0.5, 0.3, 0.02\}$



result is 0.5 or 0.1, or is the correct result very likely to be no greater than 0.2705. We want a confidence limit on our result.

Imaging repeating this experiment many (infinity) times, each time taking a sample of size n (7, in our example), and each time recording the $\langle m \rangle$ of our n sample points.

Consider now the distribution of mean values observed in this (infinite) process. It might look as shown in Illustration 2.



Again, without knowing anything about the true underlying distribution or the true mean, it certainly must be true that there is a number σ , such that 67% (say) of our sample means lie within σ of the true mean. Of course, we do not repeat this process an infinity of times, we get our n measurements only once. If we knew the value of σ , it would give a good measure of the confidence limit of our single realization of \bar{M} , we could say that there is a 67% probability that our estimate is within σ of the true mean \bar{M} . It would be helpful to know if σ is 0.001 or 0.1, for example.

The Central Limit Theorem tells us that the distribution of means discussed above follows a gaussian distribution, as suggested by Illustration 2. Moreover, the mean of this distribution of means, $\langle \bar{m} \rangle$, coincides with the mean of the underlying distribution \bar{M} and, most interesting now, the variance of this gaussian $\sigma_{\langle m \rangle}^2$ is given in terms of the (unknown) variance of the underlying distribution σ_M^2 , thus

$$\sigma_{\langle m \rangle}^2 = \frac{1}{n} \sigma_M^2$$

This indicates that the variance of the distribution of sample means decreases if each of the means is taken from a larger sample (n is increased). So if each of our (infinite number of) hypothetical 7-point samples had instead 14 points, then the variance of the distribution of sample means would be cut in half (the distribution in Illustration 2 would be narrower).

We have our n sample points available to estimate σ_M^2 , and again the most reasonable estimate is given by the same statistic applied to the sample. We take the variance of the data to construct our confidence limit, which goes as the square-root of the variance

$$\sigma_{\langle m \rangle} = \frac{1}{\sqrt{n}} \sigma_M \approx \frac{1}{\sqrt{n-1}} \left[\frac{1}{n} \sum m_i^2 - \left(\frac{1}{n} \sum m_i \right)^2 \right]^{1/2}$$

For our example data set of 7 points, this gives us a confidence limit of 0.13. Note that we replace n by $n-1$ under the radical. This is a technical point related to the fact that the error estimate is based on the same data used to estimate the mean.

Several points remain to be made in connection with this discussion. First, all of the above assumes that the n data points in our sample represent independent measurements. One must take some care to ensure that this condition is met. Successive configurations generated in a simulation tend to differ little from one another, and consequently “measurements” taken from each are likely to be very similar (*i.e.*, they probably differ from the true mean by a similar amount). The need to generate independent “measurements” leads to the introduction of block averaging as an integral part of the structure of a simulation program. Finally, we note that the “67% confidence limit” criterion that leads to the use of a (single) standard deviation (because 67% of the area under a gaussian curve lies within one standard deviation of the mean) is an arbitrary

choice and not universally used. Some researchers prefer to report their results with “95% confidence limits”, which correspond to two standard deviations. It is good practice to state the criterion used when reporting confidence limit. This discussion should emphasize the point that the confidence limits are not absolute, and should not be interpreted as a guarantee that the true value is within the given error bars. Confidence limits are meant primarily to convey the order of magnitude of the precision of the result, and are not some limit signifying where the true result *must* lie.