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, i.e., incaccuracies 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 capability to gauge, *a priori* and rigorously, 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 meaningful, and how much of it is noise—which digits would repeatedly change if we ran the calculation again and again but with a different random seeds. In other words, we'd like to know the *uncertainty* in the average. Here is a brief outline of how this is done.

- The simulation yields an ensemble average. This may or may not give directly the property being determined, but it is at least the starting point for determining the property and its uncertainty.
- We express the uncertainty in the average in terms of a confidence interval.
- The average is viewed as a random deviate, a sample from a probability distribution. The confidence interval is constructed based on knowledge of this probability distribution.
- We collect data in a way that assures that this distribution is Gaussian (or more generally, at t-distribution). We do this by ensuring that the data being averaged are themselves samples from a Gaussian distribution.
- We use block averages to generage Gaussian-distributed contributions to the larger simulation average.
- We give attention to correlations between the blocks to ensure that the analysis isn't compromised by such issues.

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 about the nature of this distribution. Let us for this example say that it is the triangular distribution depicted in Fig. 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, μ_M. In this example, the true mean is 1/3, and it is the aim of the experiments to reveal this fact.

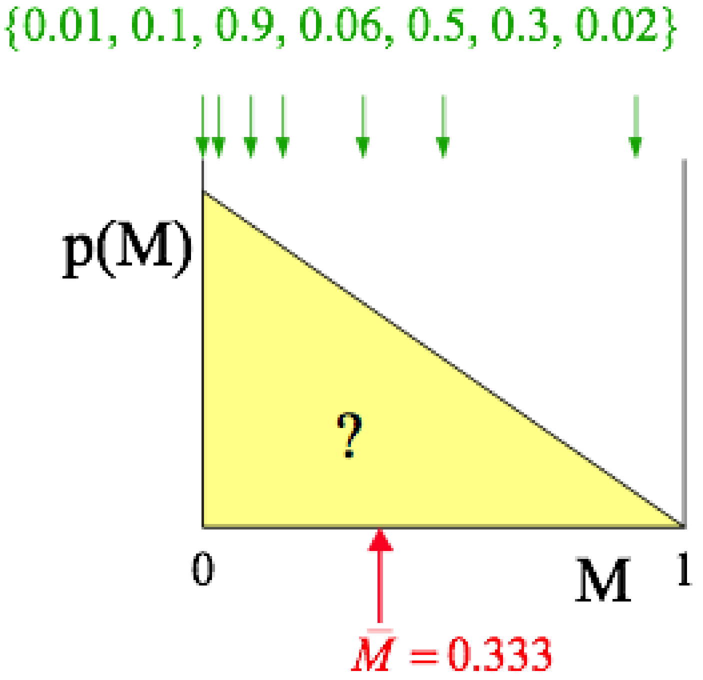

Figure 1: Hypothetical distribution governing sampling for example in text. (note: mean μ_M is labeled $\bar{M}$ here – figure needs to be updated)

We would like to obtain the best possible estimate of μ_M from the measurements. Not surprisingly, lacking specific knowledge of the sampled distribution, this is given by the mean of the measurements, the “sample mean”:

$$\mu_M \approx \frac{1}{n}\sum_{i=1}^{n} m_i \equiv \bar{M}. \tag{1}$$

Note we designate μ_M as the true mean, and $\bar{M}$ as the average of our measurements. For our example, $\bar{M} = 0.2700$. This value differs from the correct result of 0.3333, but from the experimental information available to us, we as yet have no way to know the magnitude of our error. Looking at our data, we would like to know if there’s a good chance that the correct result is 0.5 or 0.1, or instead is the correct result very likely to be no greater than 0.2705. We want a confidence limit on our result.

Imagine repeating this experiment many (infinity) times, each time taking a sample of size n (7, in our example), and each time recording the of our n sample points. Consider now the distribution of mean values observed in this (infinite) process. It might look as shown in Fig. 2.

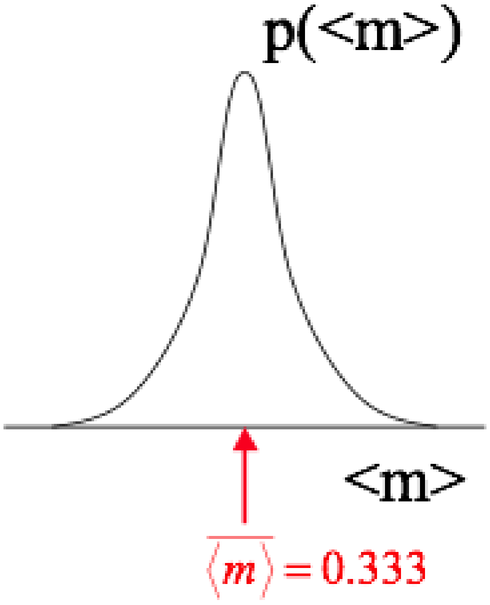

Figure 2: Hypothetical distribution of 7-sample averages. (average should be labeled as $\mu_{\bar{M}}$; also $\langle m\rangle \equiv \bar{M}$).)

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 68% (say) of our sample means lie within σ of the true mean $\mu_{\bar{M}}$. 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 68% probability that our estimate is within σ of the true mean $\mu_{\bar{M}}$. 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 Fig. 2. Moreover, the mean of this distribution of means, $\mu_{\bar{M}}$, coincides with the mean of the underlying distribution μ_M and, most interesting now, the variance of this Gaussian $\sigma^2_{\bar{M}}$ is given in terms of the (unknown) variance of the underlying distribution σ^2_M, according to:

$$\sigma^2_{\bar{M}} = \frac{1}{n}\sigma^2_M$$

This indicates that the variance of the distribution of sample means ("the variance of the mean") decreases if each of the means is taken from a larger sample (i.e., 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 Fig. 2 would be narrower).

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

$$\sigma_{\bar{M}} = \frac{1}{\sqrt{n}}\sigma_M \approx \frac{1}{\sqrt{n}}\left[\frac{1}{n-1}\sum_i \left(m_i - \bar{M}\right)^2\right]^{1/2}$$

where $\bar{M}$ is given from the data according to Eq. (1).

For our example data set of 7 points, this gives us a confidence limit of ±0.13. Note that we replace n by $n-1$ in the calculation of the sample variance. 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.

Second, we note that the "68% confidence limit" criterion that leads to the use of a (single) standard deviation (because 68% 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 a meaningful measure of the precision of the result, and are not some limit signifying where the true result *must* lie.

Finally, note that we use the sample data itself to estimate the variance of the sampled distribution. This estimate is subject to uncertainty of its own, which is not formally accounted for in the procedure outlined above. In principle, we should work with a t-distribution rather than a Gaussian when determining the uncertainty for a given confidence level. For $n > 20$ or so, the difference is not significant, and the Gaussian-derived confidence limit is fine. However, for the $n = 7$ example developed above, the 68% uncertainty should include a t-distribution multiplier to given the correct confidence limit. For a Gaussian, the multiplier for 68% uncertainty is 1.0; for a t-distribution with 7-1 = 6 degrees of freedom, the multiplier is 1.08, yielding a confidence limit of $\pm(1.08)(0.13) = \pm 0.14$. On the other hand, the t-distribution is more strict with respect to the shape of the sampled distribution, which in principle must be nearly Gaussian, so in this respect the example is problematic. The best practice is to generate enough samples so that these issues are not a concern; if that is one possible, one may resort to more sophisticated analyses than described in this brief overview.