LA-UR-18-29274

Title: Evaluating Equivalent Monte Carlo Calculations

Author(s): Picard, Richard Roy
Zukaitis, Anthony J.
Forster Iii, Robert Arthur

Intended for: project report

Issued: 2018-10-01

Evaluating Equivalent Monte Carlo Calculations

Rick Picard, Tony Zukaitis, and Art Forster

Abstract

We address several basic issues, primarily whether two sets of MCNP tallies are stochastically equivalent (i.e., that the sets of particle history scores follow the same statistical distribution) and, if different, quantify the significance level of the difference. The underlying topics here are of longstanding interest to the statistical community, and no "new" theoretical work needs to be carried out to address them. Related comparisons are useful to code developers in assessing code modifications and to code users in assessing other simulations and related physical measurements. However, these statistical methods must be applied with care to avoid obtaining misleading results, as we show in several examples. The text is written for MCNP practitioners.

1. Introduction

1.1 Background

MCNP is a large general purpose Monte Carlo transport code (Werner et al. 2017), with earlier versions used for more than 50 years at LANL and world-wide. Briefly, MCNP uses random numbers to sample appropriate physical probability density functions for simulating the movement or transport of neutral and charged particles through physical phase space: space, energy, angle, and time. A particle history is defined as all the particle movements from birth (source) to termination (escape, absorption, etc.), including all subsequent progeny; each such history is independent of all other histories.

The results from each particle history (or random walk) are scored depending on the tallies defined by the user in an MCNP problem. Tallies can summarize almost any aspect of a particle history. To keep the bookkeeping manageable for millions/billions of particle histories, results are often binned (tally binning is completely arbitrary in physical phase space), in each bin storing only the totals of the history scores and squared history scores. A single particle history can contribute to many tally bins, so that different tally bins can depend on some of the same particle histories, and need not be statistically independent.

The estimated mean result for each tally bin is the average of all the scores to each tally bin from all the particle histories run in the calculation. The estimated statistical standard deviation of the mean for each tally bin uses the universal definition involving the first and second moments of the history scores for each tally bin. Statistical errors can also be estimated by using “batch” statistics, i.e., the results from multiple runs or massively parallel calculations where each process constitutes one batch.

Over the years, versions of MCNP have undergone several V&V (verification and validation) efforts as per longstanding practice (e.g., Kliejnen 1995 among numerous related efforts), many of those efforts detailed in reports that can be found upon searches of the LANL library website. Described herein is a different kind of V&V effort, one that is highly statistical in nature. As such, it is intended to complement other V&V work.

In what follows, we focus on two separate areas of interest. The first involves the “one-sample problem” where the user has obtained a single set of MCNP particle history scores, and is interested in issues such as

a) obtaining a statistically valid confidence interval for the mean history score, or

b) comparing the observed score distribution to a postulated model (as one example, assessing whether batch means conform to a Gaussian distribution).

The second, and more important, context involves the “two-sample problem,” where the user has obtained two sets of nominally identical tallies for the same physical application. These two sets could arise, for example, upon running an MCNP problem on two different computing platforms, or from running one set of analog tallies and another set of biased tallies. Here, the user could be interested in issues such as

a) whether the mean for one set of scores is the same as the mean for a another set, or

b) whether two sets of history scores conform to the same probability distribution.

The two-sample problem is more important to this work, in that MCNP verification test problems generally do not track exactly on different machines, compilers, compiler options, and multiprocessing methods. There are also other MCNP comparisons that are important to make, such as verifying variance reduction methods (e.g., if a variance reduction technique has been properly implemented, the analog and reduced variance scores will have the same mean but different variances), comparing MCNP runs that use different random number generators, and assessing new Monte Carlo computational physics algorithms where exact tracking is impossible.

For both the one- and two-sample problems, the notion of "statistically different" is quantified by a p-value (e.g., Greenland et el. 2016), which measures the likelihood of seeing the observed degree of difference under a hypothesis that the distribution(s) are exactly as postulated. In Gaussian contexts, p-values can be converted to an equivalent number of standard deviations (e.g., "three sigma"). For the two-sample problem, the p-value is useful to quantify differences caused by machines, compilers, etc., new computational physics algorithms, or different nuclear and atomic data sets.

Importantly, there are established statistical methods for addressing the one-sample and two-sample problems, so that no "new" theoretical work needs to be pursued. In the sections to follow, we detail relevant analytical methods for assessing stochastic equivalence and show how to properly apply them in examples.

1.2 MCNP Bookkeeping

The nature of MCNP bookkeeping dictates the statistical testing that can be done. Probabilistically speaking, MCNP particle histories represent the “smallest” independent quantity and are “i.i.d.” (meaning “independent and identically distributed”). The underpinning of the statistical tests to follow are based on an i.i.d. assumption. If MCNP allowed for storing *all* detailed particle histories, very powerful one-sample and two-sample tests could be implemented.

Unfortunately, it is not practical to record/store every aspect of every particle history. Instead, results are accumulated in bins, such as defined by energy, time, physical location, etc. Because the same particle history usually impacts multiple tally bins (because of fissions, etc.), there need not be bin-to-bin independence. As such, the interpretation of bin-to-bin comparisons must account for the correlations induced by the same particle histories on multiple bins. Although confidence intervals can be obtained for the mean score in each bin individually, those intervals are correlated. Implications of this phenomenon are discussed later.

For certain V&V problems, the bookkeeping aspects of MCNP can be overcome by using a large number of excessively fine bins. For example, energy bins could be defined having width 0.01 (say, measured in whatever units are most sensible for the problem), such as an energy bin from 22.055 to 22.065 MeV. Then the number of counts in the bin would reflect the number of particle histories whose tallied energies were equal to 22.06

MeV when rounded to the nearest hundredth of a unit. If this amount of rounding error were negligible relative to other uncertainties of interest, then the actual history scores could be (nearly) recovered.

Recovering the actual scores can be extremely helpful for V&V, as shown in the sections to follow. Examining particle history scores from two nominally equivalent simulations allows for a more precise comparison of the simulations than does just comparing binned counts, especially when the bins are wide enough to partially obscure the actual scores. Comparative procedures based on individual particle history scores, such as the Kolmogorov-Smirnov test, can effectively compare different results using a modest number (a few hundred or less) of independent nonzero particle histories, while procedures based on more coarse binning, such as the chi-square test, require larger sample sizes. Taking advantage of this phenomenon is especially useful in cases where simulating individual particle histories is time consuming.

2. The One-Sample Problem

Suppose that MCNP has produced a set of particle history scores. Two common − and related − issues of interest are:

1) obtaining a statistically valid confidence interval for the mean score, and

2) assessing whether the set of individual scores (or, say, a set of estimated tally bin means) is consistent with a statistical sample from a postulated theoretical distribution, e.g., from a normal ("Gaussian") distribution.

Because understanding many elements of the one-sample problem is essential in understanding aspects of the two-sample problem, we review this subject in some detail.

2.1 Statistically Valid Confidence Intervals for a Mean Tally

The issue of valid confidence intervals has been of longstanding interest to the MCNP community. Most confidence intervals implicitly rely on the central limit theorem, which is at the core of why the Gaussian distribution is widely used in statistics. In the context of MCNP and its independent, identically distributed particle history scores, this theorem says that under mild conditions (i.e., that the probability distribution for particle history scores has a finite variance), the stochastic distribution for mean score converges as a function of sample size to a Gaussian.

Unfortunately, the central limit theorem does not provide across-the-board results on its rate of convergence, i.e., on the number of samples needed to obtain valid central-limit-theorem-based confidence intervals (see, e.g., Hall 1982 for convergence rates in specialized situations). For example, if the distribution of the individual scores were Gaussian to begin with, then the mean score would be Gaussian no matter how small the sample size is. On the other hand, if the distribution of the individual history scores is severely skewed, a comparatively large sample size (many hundreds or thousands) may be required. And lastly, if the distribution of the individual scores has infinite variance, then the central limit theorem doesn't apply at all, and specialized methods (Picard and Booth 2009) are required to obtain valid confidence intervals.

In practice, data are collected with some fixed sample size, and statistical checks are then made as to whether the sample size is "large enough" for the central limit theorem to apply. As has been discovered relative to uncertainty quantification for other LANL applications (Nakhleh, Webster, and Haynes 2015; Picard and Vander Wiel 2016), it is essential to examine all assumptions underlying the statistical methods used and to confirm, to the extent practical, that such assumptions are met. Otherwise, it is easy to reach unjustified conclusions. Relative to confidence intervals for mean history scores when the checks on the underlying assumptions are satisfactorily passed, a valid central-limit-theorem-based confidence interval is obtained. If these checks, such as the RE and VOV metrics to follow, are not satisfactorily passed, then the user has the option of obtaining more scores via additional MCNP runs (thereby increasing the overall sample size) or using confidence interval methods for non-Gaussian means.

A previous interaction between the LANL Monte Carlo group and the stat group examined this specific issue (Pederson, Forster, and Booth 1997) and derived several useful checks to assess the properties of of central-limit-theorem-based confidence intervals. For a similar paper aimed at more general (non-transport) applications, see Boos and Hughes-Oliver (2000). Those interested in the theoretical underpinnings of the checks to follow should consult these references.

Standard statistical checks on central-limit-theorem-based confidence intervals follow from empirical higher-order central moments of the particle history score distribution. Letting $\{x_i\}$ denote an observed set of n particle history scores, the first-order moment is the sample mean

$$\bar{x} \;=\; \hat{\mu}_1 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \; .$$

The second order central moment is the sample variance of the $\{x_i\}$, whose maximum likelihood estimate for Gaussians is

$$s^2 \;=\; \hat{\mu}_2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \; .$$

The third order central moment is

$$\hat{\mu}_3 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \; ,$$

and the fourth-order central moment is

$$\hat{\mu}_4 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \; .$$

The sample skewness, which is a measure of asymmetry for unimodal distributions, is $\hat{\mu}_3/s^3$ and the sample kurtosis, which is a measure of peakedness and tail weight, is $\hat{\mu}_4/s^4$. When the underlying distribution is a standard Gaussian, the theoretical skewness is 0 and the theoretical kurtosis is 3.

Common metrics for checking the applicability of central-limit-theorem-based methods utilize higher order moments, and include the relative error

$$RE \;=\; \frac{s\,/\,\sqrt{n}}{\bar{x}} \;,$$

and the estimated relative variance of the variance,

$$VOV \;=\; \frac{\hat{\mu}_4 \,-\, s^4}{n\; s^4} \;.$$

Note that the computation of VOV requires knowledge of fourth moments; VOV can not be computed from only knowledge of a bin's sums of scores and squared scores alone, so that an entry on the DBCN card is required for this calculation.

Specific threshold values for the statistical checks are calibrated to the desired confidence level. As the level of the confidence interval increases (e.g., from 95% to 99% to 99.9%), the greater the necessary sample size for the confidence interval to be valid. Simulation studies are needed to calibrate threshold values to confidence levels; for example, Pederson, Forster, and Booth (1997, p. 70) found that the threshold values $VOV \leq 0.10$ and $RE \leq 0.05$ are well calibrated to the usual 95% confidence interval; see also the MCNP6 User Manual p. 3-183, and X-5 Monte Carlo Team (2003), p. 2-117 and p. 2-125 for more information on these checks.

The operational metric values here, $VOV \leq 0.10$ and $RE \leq 0.05$, are not "magic numbers." Still better results would be achieved if the number of particle histories were increased to the point where even lower metric values were obtained. Practical constraints preclude making MCNP runs ad infinitum, however, and go-or-no-go decisions must be made as to when statistically valid results have been obtained. In this context, the recommended metric values provide good results at the 95% confidence level.

Although the central limit theorem is the most commonly used approach for developing confidence intervals, it is not the only such approach. In certain contexts, such as with expensive physical experiments, sample sizes are often limited, and if those sizes are not adequate for the central limit theorem to apply, alternative confidence interval methods must be used.

We begin with one of the simplest possible examples, involving neutron leakage. A particle is sourced from the centroid of a cube-shaped material in the direction of one of the cube faces. Of interest is the probability distribution of particle weights for the particles that cross a particular cube surface. (Aside: although the particle weights − in and of themselves − are not physically meaningful in many problems, the particle weight is used as a "particle history score" in this example to show that quantities useful for V&V of code modifications need not be confined to those of physics importance.)

One million particle histories are run, the vast majority of which, some 99.63% of them, do not cross the surface of interest. Of the particles which do cross, some 3663 particles had weights within the range (0.75, 1); the remaining 16 weights fell within the range (1.75, 2).

The structured nature of the particle weight data motivate a structured analysis. Consider developing a confidence interval for the portion of particles that cross the surface of interest. The observed portion of such particles in the simulation is

$$\hat{p} \;=\; \frac{\text{number of particles crossing the surface of interest}}{\text{number of total particle histories}} \;=\; \frac{3679}{10^6} \;=\; 0.003679\ .$$

The estimated standard deviation of the estimated portion $\hat{p}$ is

$$\hat{\sigma}\ /\ \sqrt{n} \;=\; \sqrt{(0.003679)(1-0.003679)}\ /\ \sqrt{10^6} \;=\; 0.000061\ .$$

The relative error here is

$$RE \;=\; \frac{0.000061}{0.003679} \;=\; 0.016\ ,$$

and the relative variance of the variance is

$$\begin{aligned} VOV \;&=\; \frac{\mu^4 \,-\, \hat{\sigma}^4}{n\,\hat{\sigma}^4} \\ &=\; \frac{0.003679(1-0.003679)^4 + (1-0.003679)0.003679^4 - (0.003679(1-0.003679))^2}{10^6 \times (0.003679(1-0.003679))^2} \\ &=\; 0.00027\ . \end{aligned}$$

Because the RE and VOV metrics are well behaved, i.e., they are substantially below the guidelines $RE < 0.05$ and $VOV < 0.10$, the usual central limit theorem provides a good confidence interval for the probability of crossing the surface,

$$\hat{p} \,\pm\, 2\hat{\sigma}/n \;=\; 0.003679 \,\pm\, 0.000122 \;=\; (0.00356, 0.00380)\ .$$

Next consider developing a confidence interval for the probability that a particle crossing the surface of interest has a high particle weight (1.75, 2). Only 16 of the 10^6 particle histories fall into this category. Repeating the above calculations with $\hat{p} = 0.000016$,

$$RE \;=\; \frac{\sqrt{0.000016(1-0.000016)\,/\,10^6}}{16\,/\,10^6} \;=\; 0.2499\ ,$$

indicating that the central-liimit-theorem-based confidence interval is not reliable. If obtaining a reliable estimate of uncertainty is needed, additional particle histories should be obtained.

Finally consider the particles with particle weights in the middle (0.75, 1) range, i.e., in the middle mode of the trimodal distribution. A histogram of these weights is given in Figure 1. The histogram is somewhat skewed to the left, i.e., fewer histories are observed to the left of the midpoint of the range than to the right. Otherwise the distribution is bounded and well behaved. The average particle weight of 3663 particle histories which crossed the cube surface of interest is

$$\bar{x} \;=\; 0.9566$$

and the corresponding confidence interval is

$$\bar{x} \;\pm\; 2\,\hat{\sigma}\,/\,n \;=\; 0.9566 \;\pm\; 2\,(0.0395\,/\,\sqrt{3663}\,) \;=\; (0.9553, 0.9579)\ .$$

The RE and VOV checks are satisfied for the particle history scores in this calculation.

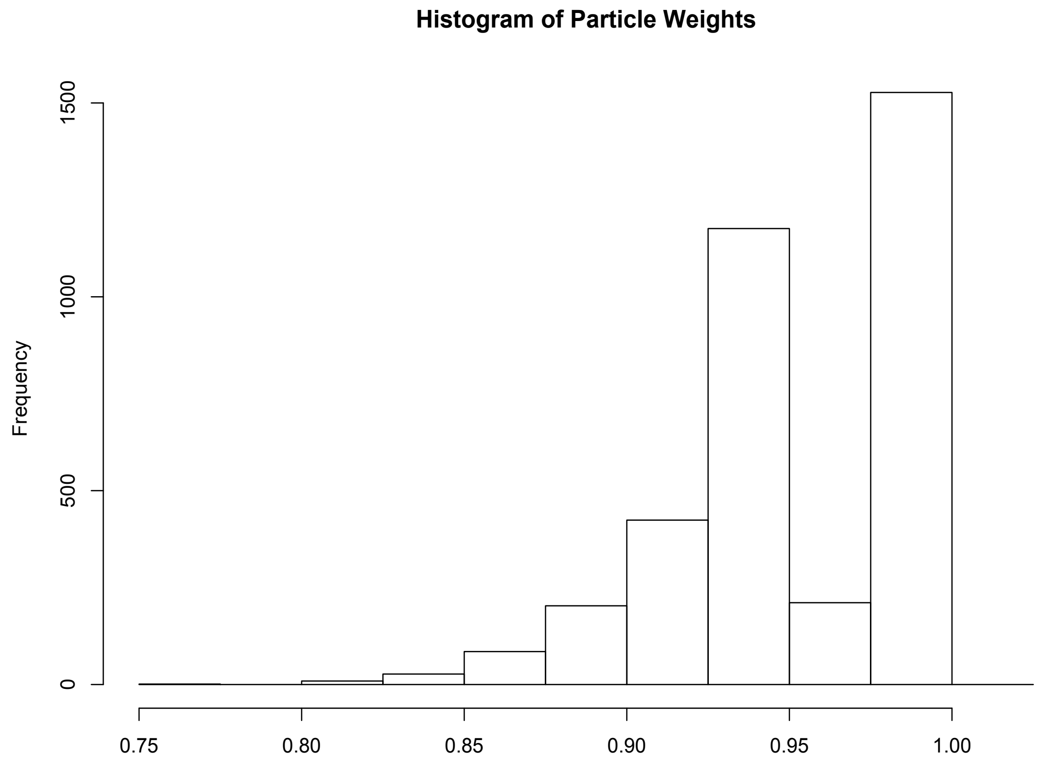

Figure 1: Particle Weights within the $0.75 - 1$ Range.

2.2 One-Sample Tests for a Tally Distribution

We now move to a more general one-sample problem, namely evaluating whether a set $\{x_i\}$ of particle history scores is consistent with a statistical sample from a postulated distribution. This problem is artificial: except in idealized situations, there is no first-principles-physics justification for postulating a specified distribution for a set of particle history scores, and in most cases scores are known to have skewed distributions. The approach is useful, however, in assessing an ensemble $\{\bar{x}_b\}$ of independent batch means, i.e., for B batch runs producing batch-specific means having the same sample size, the collection of means $\{\bar{x}_b\}$ should behave as a sample of size B from a normal distribution if the sample size is large enough for valid confidence intervals to be formed at the batch level. Moreover, understanding this one-sample problem (i.e., assessing the consistency of one set of history scores with a theoretical distribution) establishes a groundwork for later discussion of the more important two-sample problem (i.e., assessing the consistency of one set of scores with another).

Obviously, the one-sample assessment requires that all particle history scores $\{x_i\}$ or individual batch means $\{\bar{x}_b\}$ be known; unlike forming a confidence interval for a single mean (as above), it is not sufficient to have only the mean and variance of the set of scores. This requirement limits the ability to assess history score distributions to situations where MCNP bookkeeping/binning allows for obtaining such individual values.

The problem of assessing an entire score distribution has been analyzed extensively over the years in the context of statistical samples. A common example involves the testing of random number generators, especially in regard to a normal ("Gaussian") distribution and to the uniform distribution. Many users of statistics invoke tests for normality as a precursor to forming confidence intervals (e.g., if the data pass a normality check, then confidence intervals for the mean can be obtained in the usual way without regard to sample size); a test for conformance with a uniform distribution is helpful in interpreting an ensemble of p-values from individual statistical tests, as indicated in Section 3.3.

It is not the purpose of this report to extensively review the considerable literature on goodness-of-fit tests for the one-sample problem; references on more sophisticated test procedures are contained in the bibliography.

Informal tests for normality are based on visualization concepts. One such approach involves seeing if the a histogram of the data is bell-shaped. Another involves visually inspecting a normal probability plot of the scores, which should be linear for normal data. When nonnormality is pronounced, visualization can detect the type of nonnormality, e.g., whether the distribution is skewed, multimodal, contaminated with numerous outliers, etc. Still another approach entails visually examining higher-order moments of the distribution and their convergence properties (e.g., Kiedrowski and Solomon 2010).

Unfortunately, statistical visualization is user-dependent, not unlike a Rorschach test in psychiatry, where different people "see" different things in the same visual display. Also important is that most visualization is uncalibrated, e.g., what is the false positive rate for a visual assessment of nonnormality? Consequently, visualization often works well when practiced by skilled, experienced scientists, but can be problematic in many cases.

The calibrated version of a visual histogram inspection involves the chi-square test, e.g., Bishop, Fienberg, and Holland (2007). The idea is to compare the observed number of counts in each histogram bin with the expected theoretical number of counts for the bin. That is, let the lower and upper limits for the j-th histogram bin be denoted L_j and U_j. For $f(x)$ denoting the probability density function of the postulated normal distribution for an individual particle history score, the expected number of counts for the j-th histogram bin is

$$E_j \;=\; n \,\times\, \int_{L_j}^{U_j} f(x)\, dx \;, \tag{1}$$

where again n denotes the total number of scores. Here, specification of the normal probability density function $f(x)$ requires a postulated mean and variance; for large enough samples, $\bar{x}$ and s^2 can be used.

Comparison of "how close" the observed histogram bin counts $\{O_j\}$ are to their theoretical counterparts $\{E_j\}$ from the postulated model is based on a chi-square test statistic. The most commonly used test statistic is the familiar Pearson chi-square (again see Bishop, Fienberg, and Holland 2007),

$$\chi^2_{\text{Pearson}} \;=\; \sum_{j=1}^{B} \frac{(O_j \;-\; E_j)^2}{E_j} \;,$$

where the histogram has B bins.

Similar to the central limit theorem, as the number of scores increases, the chi-square statistic converges in distribution to a chi-square distribution with B degrees of freedom. Quantifying whether the sample size is "large enough" is often done using the minimum of the expected counts $\{E_j\}$. The most common guideline for deciding when the number of history scores n is "large enough" for use of χ^2_{Pearson} being when all cell expected counts $\{E_j\}$ exceed 5 and, if counts were on that order, that those expected counts are not too unequal (Haberman 1988).

For completeness, the chi-square distribution with B degrees of freedom has probability density function

$$f(x) \quad = \quad \frac{x^{B/2-1}\ e^{-x/2}}{\Gamma(B/2)\ 2^{B/2}} \qquad \text{for } x > 0\ .$$

Regarding intuition as to the shape of this density function, squaring each of B standard normal random variables and summing them yields a quantity that has a chi-squared distribution with B degrees of freedom. A quantity important for chi-square testing is the 95-th percentile value q, for which

$$95\% \quad = \quad \int_0^q f(x)dx\ .$$

Obtaining this critical value requires numerical integration, but is readily available using most statistical software.

To illustrate the chi-square approach, return to the data for Figure 1, consisting of 3663 particle history scores between 0.75 and 1. Unlike a set of batch means, there is no reason to believe that these individual history scores should be normally distributed, and the histogram in Figure 1 does not appear to be bell-shaped, and so this example is largely for illustration.

The histogram bins in Figure 1 are equally spaced, with endpoints 0.75, 0.775, 8.0, 8.025, and so on up to 1.0. The bin counts $\{O_j\}$ for this histogram are provided in Table 1; note that the four leftmost bins have been combined for use in the chi-square test because of the need for certain minimum cell counts for the test's validity.

Table 1. Chi-Square Test for Normality

Bins	< 0.85	0.85−0.875	0.875−0.90	0.90−0.925	0.925−0.95	0.95−0.975	0.975−1	>1
j	1	2	3	4	5	6	7	8
O_j	37	85	203	424	1176	211	1527	0
E_j	12.8	58.5	207.2	497.9	812.2	899.5	676.5	498.4

Expected cell counts $\{E_j\}$ are computed using a normal distribution having the mean $\bar{x} = 0.9566$ and standard deviation $\hat{\sigma} = 0.0395$, as per the 95% confidence interval above. Straightforward calculation gives the Pearson chi-square statistic

$$\chi^2_{\text{Pearson}} = \sum_{j=1}^{B} \frac{(O_j - E_j)^2}{E_j} = 2326.4 \,.$$

This chi-square value, 2326.4, is off-the-chart relative to a chi-square distribution with $B = 8$ degrees of freedom, having a probability of less than 10^{-10}. As would be expected, a histogram of 3663 data points from a normal distribution would much more closely conform to the bell-shaped form of the expected counts than does Figure 1.

[Important aside. The chi-square and other tests to follow are interpreted using a "p-value." A p-value is the probability − assuming that all assumptions related to the hypothesis are correct − of observing a value of the test statistic as extreme as the value computed from the data. As such, a small p-value (say, less than 0.05) indicates an inconsistency between the hypothesis being tested and the data. The value $\chi^2_{\text{Pearson}} = 2326.4$ above is very extreme as a result of a relatively large sample size (3663) together with a huge difference between the shapes of the observed and theoretical distributions.

In Gaussian contexts such as for batch means, there is a strong connection between p-values and confidence intervals. A 95% confidence interval ("mean value plus-or-minus two sigma") corresponds to a p-value of 0.05. That is, only 5% of the time will an observed Gaussian mean be more than two standard deviations from its theoretical value; thus, a p-value of 0.05 is analogous to a "2σ difference." When the degree of such an inconsistency becomes sufficiently great, it raises the strong possibility that some aspect of the postulated Gaussian model is incorrect.

P-values have often been misinterpreted in the scientific literature. In particular, *a p-value is NOT the probability that the hypothesis being tested is true.* There is no guarantee that all modeling-based assumptions are met (aside from the one assumption being tested), nor are any alternative possible explanatory models being considered. The subject of p-values is extensively discussed in the literature (see, e.g., Greenland et al. 2016 for a recent review), to which interested readers are referred.]

The chi-square approach is sometimes criticized by statisticians because it has "knobs," i.e., the user must specify the number B of bins and the bounds $\{L_j, U_j\}$ defining where the bins are located. In the above, there were $B = 8$ bins, equally spaced to allow for visualization of the distribution shape. Effects of the knobs are usually minor unless knob values are cherry-picked to affect a desired result.

Nonetheless, normal probability plots are knob-free and generally preferred to chi-square procedures for that reason. In a normal probability plot, the n particle history scores $\{x_i\}$ or batch means $\{\bar{x}_b\}$ are ordered from low to high. Notationally, let $x_{(1)}$ denote the smallest of the scores (or batch means), $x_{(2)}$ denote the second smallest of the scores (or batch means), and so on, to where $x_{(n)}$ denote the largest of the scores (or batch means). Based on the postulated normal distribution, the expected i-th order statistic, denoted $m_{(i)}$, can be computed. If the postulated normal distribution were correct, the $\{x_{(i)}\}$ should be linearly related to their theoretical counterparts $\{m_{(i)}\}$.

Quantifying the linear correlation is usually done by Wilk's W statistic (Shapiro, Wilk, and Chen 1968; Razali and Wah 2011) and has been found to have excellent properties in testing for normality. Computationally efficient implementations of this approach have been developed (Weisberg and Bingham 1975; Pederson, Forster, and Booth 1997, p. 58) and embedded in widely available software.

Consider applying the W test to the data in Figure 1. As noted above, there is no reason to believe that the particle weights should be normally distributed, so this example is largely for illustration. The normal probability plot for the data is given in Figure 2, and it does not at all resemble a straight line; as is apparent from the histogram (Figure 1) the lower tail of the data appears to increase in roughly Gaussian fashion, but the upper tail does not decrease smoothly at all. The large (3663) sample size and the pronounced nonlinearity in the normal probability plot strongly indicating (p-value less than 10^{-10}) that the data are not normally distributed.

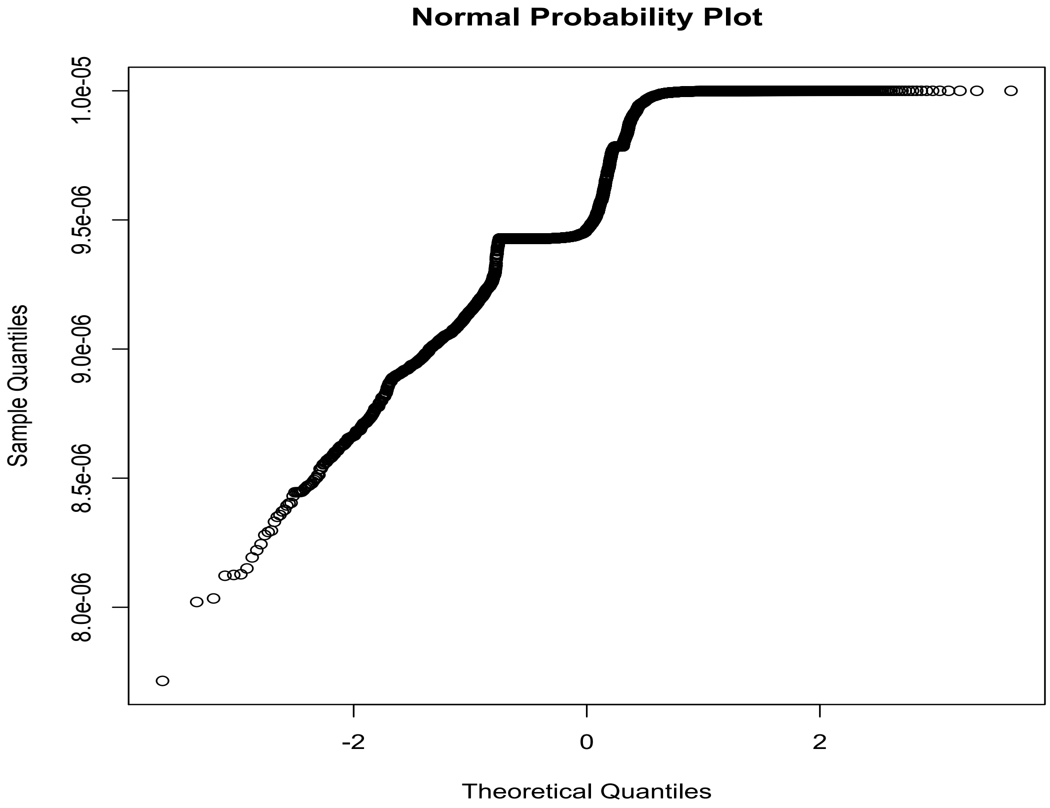

Figure 2: Particle Weights within the $0.75 - 1$ range.

Another approach for assessing normality is to directly compare the empirical cumulative distribution function to its theoretical counterpart. To this end, the most flexible, and knob-free approach, is based on the Kolmogorov-Smirnov ("K-S") test (see, e.g., Conover 1999). For a set of n scores, the empirical distribution function, denoted $\hat{F}_n(x)$ and evaluated at a specified value x, is simply the portion of the scores that are less than or equal to the value of x:

$$\hat{F}_n(x) \quad = \quad \frac{\text{number of } \{x_i\} \ \leq \ x}{n} \,,$$

where n again denotes the total number of scores. The theoretical cumulative distribution function is

$$F(x) \quad = \quad \int_{-\infty}^{x} f(x)dx \ ,$$

where $f(x)$ again denotes the postulated theoretical probability density function for an individual particle history score. As the number of scores $n \to \infty$, $\hat{F}_n(x)$ converges to $F(x)$ for all values of x when the scores in fact have the postulated distribution. A useful visualization here is an overlay plot of $\hat{F}_n(x)$ and $F(x)$, which is capable of displaying many kinds of discrepancies between the observed score distribution and its postulated counterpart, e.g., if the tails of the score distribution are too light or heavy, or if the degrees of skewness are not similar.

This visualization is, of course, subjective and uncalibrated. To formally quantify the discrepancy between the observed and theoretical cumulative distribution functions, the maximized absolute difference over all values of x is computed,

$$max_x \ |\hat{F}_n(x) - F(x)| \ .$$

When this test statistic is "too large," it is evidence that the observed scores are inconsistent with the postulated distribution. Critical values for this test can be found in Miller (1956) or Conover (1999, Table A13). For sample sizes $n > 40$, the asymptotic value

$$\frac{1.36}{\sqrt{n \ + \ \sqrt{n/10}}}$$

corresponds to significance at the 5% level; the asymptotic value $1.63/\sqrt{n + \sqrt{n/10}}$ corresponds to significance at the 1% level.

Returning to the example of the 3663 particle weights in the middle particle weight range, an overlay plot for the empirical distribution function $\hat{F}_n(x)$ and its theoretical Gaussian counterpart $F(x)$ is given in Figure 3. As is apparent from the plot, these distributions do not closely resemble each other.

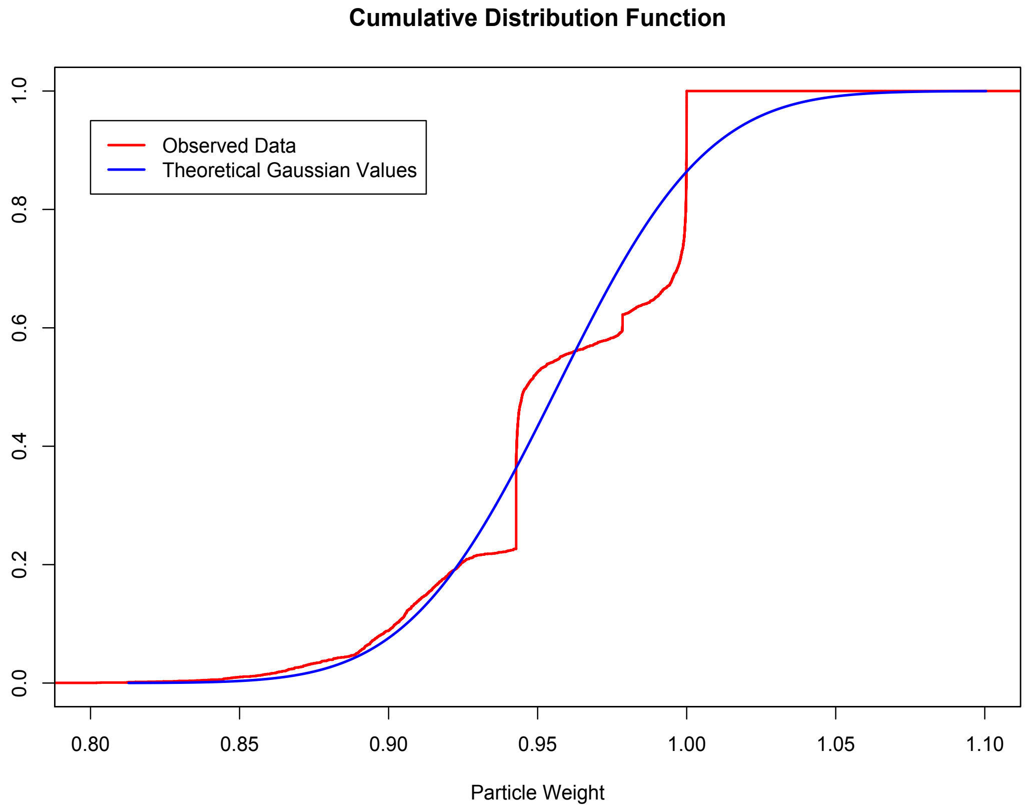

Figure 3: Particle Weights within the $0.75 - 1$ Range.

The K-S test measures the maximum vertical distance between the red and blue curves in Figure 3. This distance is roughly 0.16, well in excess of the K-S critical value

$$\frac{1.36}{\sqrt{n \,+\, \sqrt{n/10}}} \;=\; \frac{1.36}{\sqrt{3663 \,+\, \sqrt{3663/10}}} \;=\; 0.022.$$

As such, the p-value far below the 0.05 significance level provides strong evidence against the notion that these particle history scores are normally distributed.

3. The Two-Sample Problem

3.1 Comparison of Mean History Scores

In the two-sample situation, there are two sets of particle history scores, denoted $\{x_i;\ i = 1,\ldots, n_x\}$, and $\{y_j;\ j = 1,\ldots, n_y\}$. In many cases, it turns out that the sample sizes $n_x = n_y$, although this need not be the case and is not necessary for what follows.

The two most important questions here are whether the actual mean history scores are the same for the two sets, and whether the two particle score distributions themselves are the same. For comparing estimated means, the standard procedure is the two-sample t-test, which for MCNP-like sample sizes is a two-sample z-test. That test involves computing

$$Z \;=\; \frac{\bar{x} \,-\, \bar{y}}{\sqrt{s_x^2 \,/\, n_x \,+\, s_y^2 \,/\, n_y}}$$

Here, Z quantifies how many standard deviations ("sigmas") from zero that the observed mean tally difference $\bar{x} - \bar{y}$ is. As with one-sample procedures, it is important to check the assumptions needed for a valid interpretation of Z. In this regard, the advised approach is to individually check for $\bar{x}$ and $\bar{y}$ being normally distributed, as per the one-sample situation in Section 2.1 with the RE and VOV metrics regarding valid confidence intervals. If both $\bar{x}$ and $\bar{y}$ are normally distributed, then the difference $\bar{x} - \bar{y}$ is as well.

Depending on circumstances, the variances for the two sets of scores may be expected to be the same (e.g., if running the same MCNP problem on two different computing platforms) or may be expected to be different (e.g., comparing a set of analog scores to another set of scores obtained upon using a variance reduction method).

When the variances for the two sets of scores are different, the test for $\bar{x} - \bar{y}$ is called the *Behrens Fisher problem.* Convergence of Z to a Gaussian distribution depends on the two sample sizes n_x and n_y, on the ratio of the two standard deviations, and on the distributions of $\bar{x}$ and $\bar{y}$. For even modest sample sizes (say, more than 20), modest heteroscedasticity (say, a ratio s_x/s_y of standard deviations between 1/5 and 5), and modestly skewed distributions,the Behrens-Fisher problem has been studied by Miao and Chiou (2008) and the normal distribution approximation appears to be very robust at the 5% ("two sigma") level, and only somewhat less so at more extreme levels.

The example of the previous section, involving neutron leakage for an idealized cube, was re-run with a different set of random numbers. As such, results for the second run would be expected to be stochastically equivalent to those of the first run.

When comparing two sets of results, many comparisons can be made. In the first run, for example, some 3679 out of 10^6 particles crossed the surface of interest. In the second run, 3744 out of 10^6 particles crossed the surface. As with the first run, the RE and VOV checks for normality are satisfied. Upon comparing the runs, the difference is

$$\hat{p}_1 - \hat{p}_2 \;=\; .003679 - 0.003744 \;=\; -0.000065 \,.$$

The standard deviation of this difference is estimated as

$$\begin{aligned} \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} &= \sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{10^6} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{10^6}} \\ &= \sqrt{\frac{.003679\,(1-.003679)}{10^6} + \frac{0.003744\,(1-0.003744)}{10^6}} \\ &= 0.000086 \,. \end{aligned}$$

Thus, the difference is

$$Z \;=\; \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}} \;=\; \frac{-0.000065}{0.000086} \;=\; -0.76$$

standard deviations from zero.

For a normal distribution with mean zero, 55% of the distribution lies between −0.76 and +0.76 sigma. Which means that in this case, a 55% confidence interval would barely touch zero:

$$\hat{p}_1 - \hat{p}_2 \,\pm\, 0.76\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \;=\; (-0.000130, 0) \,.$$

Stated differently, the p-value for the test that $p_1 = p_2$ is $1 - 0.55 = 0.45$. A p-value of this magnitude is consistent with the two runs being stochastically equivalent, as compared using their portion of trajectories crossing the cube surface of interest.

Of course, there are other ways to compare the two MCNP runs. Recall that the first run yielded a trimodal distribution with 3663 and 16 trajectories having particle weights in the second and third modes, respectively. For the second run, the corresponding numbers are 3714 and 30 trajectories. In the first run, the mean particle weight of the 3663 trajectories in the second mode was $\bar{x}_1 = 0.9566$; for the second run, the mean particle weight of the 3714 trajectories in the second mode was $\bar{x}_2 = 0.9569$. The difference between these mean values is

$$\bar{x}_1 \; - \; \bar{x}_2 \quad = \quad 0.9566 \; - \; 0.9569 \quad = \quad -0.0003 \; .$$

The standard deviation of the difference is

$$\sqrt{\left(\frac{\hat{\sigma_1}}{\sqrt{n_1}}\right)^2 + \left(\frac{\hat{\sigma_2}}{\sqrt{n_2}}\right)^2} \quad = \quad \sqrt{\left(\frac{0.0395}{3663}\right)^2 + \left(\frac{0.0395}{3714}\right)^2} \quad = \quad 0.0009 \times 10^{-6} \; .$$

Here, the difference between the means is roughly $Z = 1/3$ sigma from zero, and the corresponding p-value is roughly 0.74. Again, there is no evidence of a difference between the runs based on this metric.

3.2 Stochastic Equivalence

For MCNP applications, there is usually no basis for postulating a specific probability distribution for individual particle history scores. As such, the most common approach for comparison is the two-sample chi-square. There is also a two-sample version of the K-S test. When scores for each set of runs are binned, they conform to a so-called product multinomial sampling. Generalizing the chi-square from the one-sample problem, there are now now two sets of counts for the same set of histogram bins. Let $O_{i,j}$ denote the observed count in the i-th histogram ($i = 1, 2$) from the j-th bin ($j = 1, \ldots, B$). In other words, the histogram counts can be displayed in tabular form:

Counts	Bin #1	Bin #2	...	Bin #B	Total # Counts
1st Set	$O_{1,1}$	$O_{1,2}$	...	$O_{1,B}$	$n_1 = \sum_j O_{1,j}$
2nd Set	$O_{2,1}$	$O_{2,2}$	...	$O_{2,B}$	$n_2 = \sum_j O_{2,j}$

When both history score sets conform to the same distribution, the expected count for the i-th set and j-bin is

$$E_{i,j} \quad = \quad \frac{\sum_j O_{i,j} \ \times \ \sum_i O_{i,j}}{n_1 + n_2} \ .$$

The most common chi-square statistic for the two-sample problem is then

$$\chi^2_{\text{Pearson}} \quad = \quad \sum_{i=1}^{2} \sum_{j=1}^{B} \frac{(O_{i,j} \ - \ E_{i,j})^2}{E_{i,j}} \ .$$

The reference distribution is a chi-square with $B - 1$ degrees of freedom. Importantly, the procedure is easily extended from two independent samples to $k > 2$ samples, so as to compare multiple independent sets of particle history scores.

As an example, a histogram for the 3714 particle history scores in the middle mode of the trimodal particle score distribution is given in Figure 4. The binning of this histogram is identical to that for Figure 1.

For the two MCNP runs at hand, the histogram counts are given in Table 2. The corresponding bin counts are, not surprisingly, very similar, and the histograms are also visually very similar.

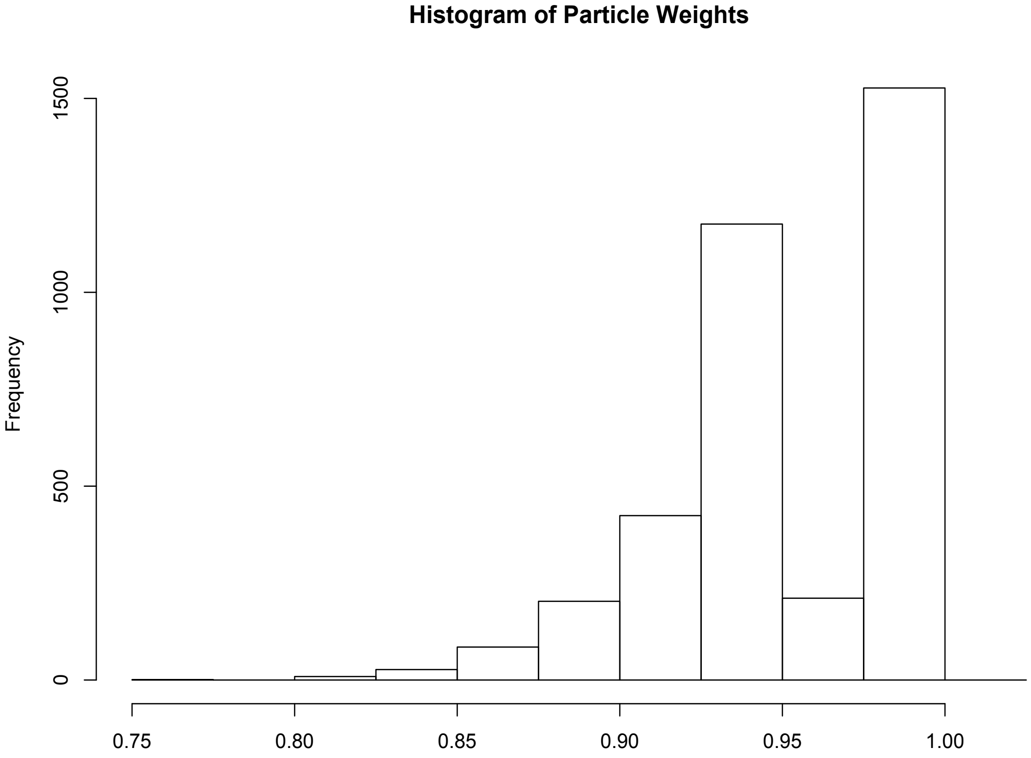

Figure 4: Particle Weights within the $0.75 - 1$ Range for the Second MCNP Run.

Table 2. Chi-Square Test for Stochastic Equivalence

Bins	< 0.85	$0.85-0.875$	$0.875-0.90$	$0.90-0.925$	$0.925-0.95$	$0.95-0.975$	$0.975-1$	# Counts
j	1	2	3	4	5	6	7	
$O_{1,j}$	37	85	203	424	1176	211	1527	$n_1 = 3663$
$O_{2,j}$	31	78	243	413	1151	258	1540	$n_2 = 3714$

The expected cell counts $\{E_{i,j}\}$ can be summarized in a similar table:

Table 2. Expected Bin Counts

Bins	< 0.85	0.85−0.875	0.875−0.90	0.90−0.925	0.925−0.95	0.95−0.975	0.975−1	# Counts
j	1	2	3	4	5	6	7	
$E_{1,j}$	33.76	80.94	221.46	415.61	1155.46	232.88	1522.90	$n_1 = 3663$
$E_{2,j}$	34.24	82.06	224.54	421.39	1171.54	236.12	1544.10	$n_2 = 3714$

The Pearson chi-square value is

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{2}\sum_{j=1}^{B} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 9.24\,,$$

with the reference distribution a chi-square with $B-1=6$ degrees of freedom. The p-value for this test is 0.16. In that 0.16 exceeds the usual "two sigma" p-value of 0.05, there is very little evidence against stochastic equivalence.

Similar to one-sample problems, use of the chi-square approach requires the "knobs" that define the histogram, i.e., choosing the number of histogram bins and the bin locations. As might be expected, there is a two-sample version of the K-S test which is free of knobs and, because the test is based on actual history scores instead of on binned histogram counts, allows for a more precise comparison of the two sets of values. The empirical distribution functions for the sets $\{x_i\}$ and $\{y_j\}$ of history scores are

$$\hat{F}_{n_x}(x) = \frac{\text{number of } \{x_i\} \leq x}{n_x}$$

and

$$\hat{G}_{n_y}(x) = \frac{\text{number of } \{y_j\} \leq x}{n_y}\,.$$

When the two sets of scores are stochastically equivalent, it would be expected that the two empirical distributions would be close to each other. As in the one-sample case, closeness is measured by the absolute value

$$max_x \; |\hat{F}_{n_x}(x) - \hat{G}_{n_y}(x)| \; .$$

Critical values can be found in Table A20 of Conover (1999). For large (n_x, n_y), say, both greater than 40, values of the absolute difference greater than

$$1.36 \; \times \; \sqrt{\frac{n_x + n_y}{n_x \, n_y}}$$

corresponds to significance at the 5% level. More generally, the p-value for this test is a measure of the evidence that the two probability distributions are different.

Using the $n_x = 3663$ particle history scores from the first MCNP run and the $n_y = 3714$ scores from the second run which correspond to the middle bin for the trimodal score distribution, the empirical cumulative distribution functions $\hat{F}_{n_x}(x)$ and $\hat{G}_{n_y}(x)$ can be plotted. This is done in Figure 5, where the two curves are overlaid.

The maximum vertical distance between the red and green curves is only 0.019. The 5% critical value is

$$1.36 \; \times \sqrt{\frac{n_x + n_y}{n_x \, n_y}} \;\; = \;\; 1.36 \; \times \sqrt{\frac{3663 + 3714}{3663 \; \times \; 3714}} \;\; = \;\; 0.032 \; ,$$

and so the comparison does not show a significant difference. Indeed the p-value for the test is roughly 0.49, so that it would be expected that the observed difference 0.019 is typical of what would be expected from two stochastically equivalent simulations.

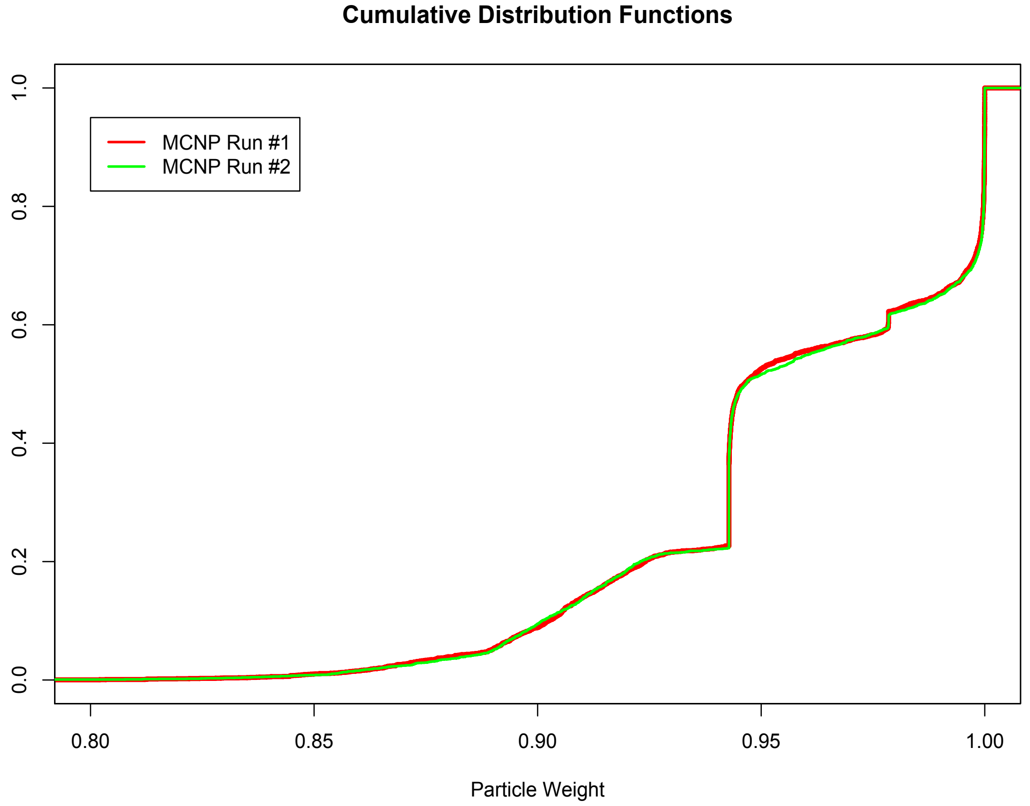

Figure 5: Particle Weights within the 0.75 − 1 Range.

3.3 Multiple Comparisons: A Complicating Factor

For many MCNP applications, there are numerous quantities of interest. MCNP problems may involve many isotopes of many elements, many energy bins, time bins, physical locations, and so on. Still other quantities, such as particle weights, can be useful for V&V even though they may not be of direct physics interest. That *some* quantities from two MCNP runs are stochastically equivalent does not guarantee that the same is true of *all* quantities. Consequently, it is important to make many such two-sample comparisons, and to give careful thought as to which output quantities warrant scrutiny.

When multiple quantities are examined from one (or two) MCNP runs, results must be interpreted from a broader perspective. For example, for a large number of statistically valid 95% confidence intervals, it would be expected that 5% of intervals would fail to contain the actual quantity of interest. The same principle extends to p-values and tests for stochastic equivalence. If a large number of independent chi-square tests were run, for example, some 5% of those tests would be expected to give p-values less than 0.05, thus indicating a nominal significant difference. Similarly, it would be expected that 1% of tests would give p-values less than 0.01, 10% of tests would give p-values less than 0.10, and so on. See, e.g., Urbatsch et al. 1995 for one such application of interest.

It turns out that, *when two MCNP runs are truly stochastically equivalent*, the set of p-values from independent tests should behave as a random sample from a uniform distribution on the interval (0,1). This result is useful in assessing an ensemble of independent test results. That is, the set of p-values can be compared, using the above one-sample K-S test for a uniform distribution, to quantitatively assess consistency with a uniform distribution. Letting $\{p_i\}$ denote the set of n independent p-values and $\hat{F}_n(p)$ denote the empirical distribution function of the $\{p_i\}$, then $F(p) = p$ for $p \in (0, 1)$ denote the theoretical distribution function for the uniform distribution, the K-S statistic has the form

$$max_p \; |\hat{F}_n(p) - p| \; .$$

Alternatively, Fisher's omnibus test, which relies on the test statistic $-2 \sum log(p_i)$ having a chi-square distribution, could be pursued.

If the p-value for the K-S test of multiple uniform p-values exceeds 0.05 and is "not significant," then an observed ensemble of independent tests is consistent with all sample means behaving in accordance with their postulated theoretical counterparts. This doesn't guarantee that every quantity examined is stochastically equivalent from the two runs, just that the collection of p-values is consistent with no genuine differences.

If, on the other hand, the p-value for the K-S test of p-values is "significant" and the value of p which maximizes $|\hat{F}_n(p) - p|$ is near zero, this is an indication that at least some of the p-values are too small. When this occurs, the user faces an awkward decision as to which individual p-values reflect "false positives," sometimes called "statistical positives," and which ones indicate genuine issues to be investigated. Resolving this issue obviously requires subject matter knowledge, most likely coupled with additional MCNP runs.

For completeness, if the p-value for the K-S test of p-values is "significant" and the value of p which maximizes $|\hat{F}_n(p) - p|$ is *not* near zero, this is an indication that some of the underlying assumptions are invalid (Christensen 2003). One invalidating factor would be if the sample means were not truly independent; a set of sample means based on the same particle histories certainly need not be. In that multivariate extensions of the K-S test are problematic in such correlated contexts (e.g., Justel, Pena, and Zamar 1997), careful interpretation of the correlated individual results is required.

4. Other Problems (Requiring Research)

The example in this report was deliberately selected to be simple, so as to emphasize methods for comparing MCNP output from distinct simulations and assigning a significance level to the differences. Many realistic problems are substantially more complex than the one here, such as involving more complicated source terms and score functions. Extending the basic V&V methods described herein to more complex settings is of future value.

An important distinction regarding MCNP involves precision and accuracy (see also X-5 Monte Carlo Team (2003), p. 2-106+). Inputs to MCNP, such as cross sections and source distributions, are only known to a certain degree of accuracy. In some cases, it may be of interest to propagate *all* uncertainties (i.e., those in the inputs as well as those related to statistical precision in history scores). Methods for full-blown uncertainty quantification have been developed for deterministic ("Eulerian") computer codes (e.g., Iman and Helton 1988), and would need to be extended for stochastic ("Lagrangian") codes such as MCNP.

Another problem of interest is the so-called "inverse problem." It involves being given MCNP output and reverse-engineering an estimate of the inputs which produced those outputs, together with the uncertainty in that estimate. As one example, consider using MCNP in conjunction with data from a gamma or neutron detector, with the goal of identifying the location of a point source. In addition, uncertainties for the estimated point source location are also desired. There is work on this problem for deterministic codes, but that work would have to be extended to apply to MCNP.

Appendix: Additional Results

This report has placed great emphasis on obtaining valid confidence intervals via use of the central limit theorem, which is far and away the most widely used approach. In many physical experiments, costs are sufficiently large that it is not practical to increase the sample size, yet it is still desired to form confidence intervals. In cases where the sample size is not adequate for the central limit theorem to apply, other methods have been developed for specialized situations.

One situation involves binary data, e.g., pass/fail or go/no-go. Consider for example, analog simulation for a shielding penetration problem. Here, each particle has probability p of penetration, and it is of interest to obtain a valid confidence interval for this probability. From n independent particle histories, the estimated probability is

$$\hat{p} \;=\; \frac{\#\text{ particle histories penetrating the shielding}}{n} \;.$$

The central limit theorem confidence interval, which goes back to LaPlace in the early 1800s, has the form

$$\hat{p} \;\pm\; z_{\alpha/2}\,\sqrt{\hat{p}(1-\hat{p})\,/\,n}\;,$$

For a 95% confidence interval, the Gaussian quantile $z_{\alpha/2} = 1.96 \approx 2$, and the result is the familiar “plus or minus two standard deviations” interval.

As a simple example, consider a penetration problem where only two out of 50 analog MCNP runs yielded particle penetration (supposing that individual particle histories are very time consuming to generate). The relative error here is

$$RE \;=\; \frac{\sqrt{(2/50)\;(1-2/50)\;/\;50}}{2/50} \;=\; 0.69\;.$$

This relative error is large enough (recall from Section 2.1 that $RE < 0.05$ is advised) for a central-limit-theorem based confidence interval to have good coverage properties.

For the one-sample problem, the score-based confidence interval (Agresti and Coull 1998) can be used for such cases. This interval has the form

$$\left(\hat{p} \,+\, \frac{z^2_{\alpha/2}}{2n} \;\pm z_{\alpha/2}\sqrt{\left[\hat{p}(1-\hat{p}) \,+\, z^2_{\alpha/2}/4n\right]\;/\;n}\right) \;/\; \left(1 \,+\, \frac{z^2_{\alpha/2}}{2n}\right)\;.$$

As the sample size $n \to \infty$, the above interval converges to the central-limit-theorem-based interval, as it must. This interval has good properties unless the estimated proportion is extremely close to zero or one and sample sizes are modest. A two-sample version of this approach (Agresti and Caffo 2000) has also been developed .

A second situation of interest involves particle history score distributions that are visibly skewed to the right, and there is an upper tail to the usual histogram. This situation is typified by a sample skewness $\hat{\mu}_3 \gg 0$, and is common for certain MCNP tallies. In contrast to the usual upper limit for the 95% confidence interval when the sample size is large,

$$Pr\left(\,\mu \;\le\; \bar{x} \,+\, 1.96\,s\,/\sqrt{n}\,\right) \;=\; \Phi(1.96) \;=\; 0.975\;,$$

the upper limit for severely skewed situations is

$$Pr\left(\,\mu \;\le\; \bar{x} \,+\, (s\,/\sqrt{n})\;(1.96 - C)\,\right) \;=\; \Phi(1.96) \;=\; 0.975\;,$$

where the skewness correction (Chen 1995, p. 768) to the usual 95% confidence interval uses the constant

$$C \quad = \quad \frac{\hat{\mu}_3}{s^3} \; \frac{(1 + 2 \times 1.96^2)}{6\,\sqrt{n}} \; .$$

As the sample size $n \to \infty$, the skewness correction $C \to 0$, the central limit theorem sets in, and the Gaussian-based interval is obtained.

In isolated cases, the distribution for a set of particle history scores may have infinite variance. The infinite variance condition can be checked using cumulative standard deviation plots. Such a plot displays the calculated standard deviation of observed scores as a function of sample size; as the sample size increases, the plot should converge to the theoretical standard deviation. If the theoretical standard deviation of the score distribution is infinite, the plotted values will continue to drift upward, not converging at all. Alternatively, tail slope metrics, e.g., Pederson, Forster, and Booth (1997, p. 61), can also be used. The presence of infinite variance invalidates the central limit theorem altogether; when this occurs, specialized methods (Picard and Booth 2009) are required to obtain valid confidence intervals.

Another option for infinite variance situations is to consider variance reduction methods, or to consider different such methods if an initial variance reduction application appears to have infinite variance.

For assessing stochastic equivalence for a set of particle history scores, the chi-square and Kolmogorov-Smirnov approaches are most widely used, as noted in the text. Nonetheless, there are alternative methods that perform slightly better in some cases, although they are more complicated to implement. These alternatives include the Cramer-von-Mises test statistic, which is similar to an integrated squared error criterion,

$$\sum_{i=1}^{n} \left[F\left(x_{(i)}\right) - \frac{2i-1}{2n}\right]^2 ,$$

for $x_{(i)}$ again denoting the i-th order statistic as was used in the W test for normality. The observed order statistic is evaluated relative to the postulated distribution function F, and squared differences are accumulated over the sample.

Along similar lines (e.g., Scholz and Stephens 1987), the one-sample Anderson-Darling test is based on a modification of the Cramer-von-Mises test, and has the form

$$-n - \sum_{i=1}^{n} \frac{(2i-1)}{n} \left[ln\, F(X_{(i)}) + ln\,(1 - F(x_{(n+1-i)})\right] .$$

The two-sample Anderson-Darling test is much more complex than Kolmogorov-Smirnov, but slightly more powerful in small sample size situations (Boyerinas 2016; Scholz and Stephens 1987). Like the chi-square test above, it can also be extended to multiple ($k \geq 2$) independent sets of scores. Though not immediately apparent from the above formula, the Anderson-Darling approach places greater emphasis on tail behavior than does the K-S test. Unlike the chi-square and K-S tests, critical values for the Anderson-Darling test statistic depend on the specific distribution $F(x)$ being tested, which is a complication in practice. Work on this test statistic has shown good performance for specific distributions such as the normal, exponential, and Weibull.

Finally, as a historical aside, there is something called the Lilliefors test for normality (Lilliefors 1967), which is just the K-S test with the theoretical distribution function $F(x)$ above replaced by $\tilde{F}$, i.e., $\tilde{F}$ is the normal cumulative distribution function with mean taken as the observed tally mean $\bar{x}$ and variance taken as s^2. In other words, the test mimics the K-S approach for a situation where the postulated normal distribution has estimated mean and variance. The Lilliefors test statistic is then

$$max_x \; |\hat{F}_n(x) - \tilde{F}(x)| \; .$$

Until the advent of the W test, the Lilliefors test was widely used.

Bibliography

Agresti, A., and Cafffo, B. (2000), "Simple and Effective Confdence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures," *American Statistician*, **54**, 280-288.

Agresti A. and Coull, B. (1998), "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," *American Statistician*, **52**, 119-126.

Baggerly, K., Cox, D., and Picard, R. (2000), "Exponential Convergence of Adaptive Importance Sampling for Markov Chains," *Journal of Applied Probability*, **37** 342-358.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007), *Discrete Multivariate Analysis*, Springer: New York.

Boos, D. D. and Hughes-Oliver, J. M. (2000), "How Large Does n Have to be for Z and t Intervals?," *American Statistician*, **54**, 121-128.

Booth, T. E. (1985), "Exponential Convergence for Monte Carlo Transport," *Transactions: American Nuclear Society*, **50**, 267-268

Booth, T. E. (2004), "Ex Post Facto Monte Carlo Variance Reduction," *Nuclear Science and Engineering*, **148**, 391-402.

Boyerinas, B. M. (2016), "Determining the Statistical Power of the Kolmogorov-Smirnov and Anderson-Darling Goodness-of-Fit Tests via Monte Carlo Simulation," CNA Technical Report DOP-2016-U-014638.

Chan, I. S, and Zhang, Z. (1999), "Test-Based Exact Confidence Intervals for the Difference of Two Binomial Proportions," *Biometrics*, **55**, 1202-1209.

Chen, L. (1995), "Testing the Mean of Skewed Distributions," *Journal of the American Statistical Association*, **90**, 767-772.

Christensen, R. (2003), "Significantly Insignificant F Tests," *American Statistician*, **57**, 27-32.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, Wiley: New York.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, P-Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *American Statistician*, Online Supplement.

Haberman S. J. (1988), "A Warning on the Use of Chi-Squared Statistics with Frequency Tables with Small Expected Cell Counts," *Journal of the American Statistical Association*, **83**, 555-560.

Hall, P. (1982), "Bounds on the Rate of Convergence of Moments in the Central Limit Theorem," *The Annals of Probability*, **10**, 1004-1018.

Iman, R. L., and Helton, J. C. (1988), “An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models,” *Risk Analysis*, **8**, 71-90.

Justel, A., Pena, D., and Zamar, R. (1997), “A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit,” *Statistics & Probability Letters*, **35**, 251-259.

Kiedrowski, B. C. and Solomon, C. J. (2010), “Statistical Assessment of Monte Carlo Distributional Tallies,” Los Alamos National Laboratory Technical Report LA-UR-10-08203.

Kleijnen, J. P. C. (1995), “Verification and Validation of Simulation Models,” *European Journal of Operations Research*, **82**, 145-162.

Kollman, C., Baggerly, K., Cox, D., and Picard, R. (1999), “Adaptive Importance Sampling on Discrete Markov Chains,” *Annals of Applied Probability* **9** 391-412.

H. W. Lilliefors (1967), “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” *Journal of the American Statistical Association*, **62**, 399-402.

Miao, W. and Chiou, P. (2008), “Confidence Intervals for the Difference Between Two Means,” *Computational Statistics & Data Analysis*, **52**, 2238-2248.

L. H. Miller (1956), “Table of Percentage Points of Kolmogorov Statistics,” *Journal of the American Statistical Association*, **51**, 111-121.

Nakhleh, C. W., Webster, R. B., and Haynes, D. A. (2015), "Quantication of Margins and Uncertainties Using Imprecise Probabilities," Los Alamos National Laboratory Technical Report LA-UR-15-20764.

Pederson, S, P., Forster, R. A., and Booth, T. E. (1997), "Confidence Interval Procedures for Monte Carlo Transport Simulations," *Nuclear Science and Engineering*, **127**, 54-77.

Picard, R. and Booth, T. (2009), "Ensuring Finite Moments in Monte Carlo Simulations via Iterated Ex Post Facto Sampling," *Mathematics and Computers in Simulation* **79** 2106-2121.

Picard, R. and Vander Wiel, S. (2016), "Imprecise Probability Methods for Weapons UQ," Los Alamos National Laboratory Report LA-UR-16-23428.

Picard, R. and Williams, B. (2013), "Rare Event Estimation for Computer Models," *American Statistician*, **67** 22-32.

Razali, N. M. and Wah, Y. B, (2011), "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests," *Journal of Statistical Modeling and Analytics*, **2**, 21-33.

Scholz, F. W. and Stephens, M. A. (1987), "K-Sample Anderson-Darling Tests," *Journal* *of the American Statistical Association*, **82**, 918-924.

Shapiro, S. S., Wilk, M. B., Chen, H. J. (1968), "A Comparative Study of Various Tests for Normality," *Journal of the American Statistical Association*, **63**, 1343-1372.

Urbatsch, T. J., Forster, R. A., Prael, R. E., and Beckman, R. J. (1995), "Estimation and Interpretation of k_{eff} Confidence Intervals in MCNP," *Nuclear Technology*, **3**, 169-182.

Warren, W. G. (1984), "The Power of Some Tests of Normality Against Weibull Alternatives," *Communications in Statistics - Simulation and Computation*, **13**, 243-255.

Weisberg, S. and Bingham, C. (1975), "An Approximate Analysis of Variance Test for Non-Normality Suitable for Machine Calculation," *Technometrics*, **17**, 133-134.

Werner, C. J., Armstrong, J. C., Brown, F. B., Bull, J. S., Casswell, L., Cox, L. J., Dixon, D. A., Forster III, R. A., Goorley, J. T., Hughes, H. G., Favorite, J. A., Martz, R. L., Mashnik, S. G., Rising, M. E., Solomon, C. J. Jr., Sood, A., Sweezy, J. E., Zukaitis, A. J., Anderson, C. A., Elson, J. S., Durkee, J. W. Jr., Johns, R. C., McKinney, G. W., McMath, G. E., Hendricks, J. S., Pelowitz, D. B., Prael, R. E., Booth, T. E., James, M. R., Fensin, M. L., Wilcox, T., Kiedrowski, B. C. (2017), "MCNP User's Manual Code Version 6.2," Los Alamos National Laborator Technical Report LA-UR-17-29981.

X-5 Monte Carlo Team (2003), "MCNP – A General Monte Carlo N-Particle Transport

Code, Version 5,: Volume 1: Overview and Theory," Los Alamos National Laboratory Technical Report LA-UR-03-1987.