*Transactions of the Korean Nuclear Society Spring Meeting*
*Jeju, Korea, May 7-8, 2026*

# Robust Confidence Interval Estimate for Rare Radiation-Transport Event in Monte Carlo Simulation

Arief Rahman Hakim, Douglas A. Fynan*
*Department of Nuclear Engineering, Ulsan National Institute of Science & Technology,*
*arief.rahman@unist.ac.kr, dfynan@unist.ac.kr*
**Corresponding author*

***Keywords : robust confidence interval, rare event, monte carlo radiation transport**

## 1. Introduction

Central Limit Theorem (CLT) is the standard statistical model applied in Monte Carlo (MC) radiation transport codes to obtain confidence interval estimates of unknown radiation transport process being studied. Both the attractiveness and pitfall of CLT is the asymptotic property that the distribution of sample means from batches of finite independent and identically distributed (i.i.d.) draws tend to be normally distributed as the batch size becomes large. The pitfall lies in the simple question: what batch size is "large enough" to ensure batch statistics approached the asymptotic property? In the absence of physical-based approach [1], the precise requirement is not known a priori. Repeated studies have shown that insufficient batch size can give wrong results up to several orders of magnitude.

The problem is more prominent when extremely rare (i.e., $p \ll 0.001$) transport events are involved, now the batch size requirement increases by several folds; simulation requires a prohibitively large number of particles to run. Consequently, combinations of variance reduction techniques (VRTs) are almost always mandatory to obtain any meaningful results; there is an additional factor MC analyst has to consider: a complex, yet faithful VRTs strategy. US NRC published a regulatory guide on representative statistical checks which can be used to detect anomalous behavior during tally convergence [2]. In practice, MC analyst relies on the ten (heuristic) statistical checks [3,4] and formal normality tests if available. Both tests are generally computational and memory expensive, can be prohibitive for some clusters noting that memory price has surged multiple folds for the past months.

Nuclear regulatory bodies have prior experience on estimating confidence interval (CI) given severely limited dataset available. In [5], US NRC investigates safety systems reliability during anticipated transient event without scram where there were only two reactor protection system failures in 1627 reactor-year. The US NRC statement "… *a failure-free accumulated operating time of about one and a quarter million reactor-year is needed … This is an impossible requirement …*" raises concern on the reliability quantification method when dataset is severely limited. Component failure is actually analogous to a Bernoulli trial where there are *only* two distinct outcomes: component either fail (True) or not fail (False), and vice versa. Given that Bernoulli trials follow Binomial distribution, it is *actually* possible to use the governing distribution to estimate the component lifetime CI. US NRC later selected Poisson CI formulation which is a good approximation to the Binomial CI given the component failure probability is very low (0.00123 per reactor-year). Later study in [6] confirms that the Poisson formulation converges to the engineering method ("normal approximation") as number of trials increases. Unfortunately, MC radiation transport codes appear lagged behind and remains using the standard CLT statistical model to obtain transport event CI.

This paper proposes a robust CI formulation for rare transport events given with little to no reliable "information" at all, a complete substitute when CLT statistical model often breaks down. Many transport events are actually a Bernoulli trial, so the CI inferred from the governing distribution can be used as a substitute for the CI obtained from the standard CLT procedure. There are several formulations available for Binomial CI, but we purposefully select Poisson formulation which is a good approximation when event probability is very low $p \to 0$ . The Poisson CI formulation is a paradigm shift to the standard CLT statistical procedure, so Section 2 derives the Poisson CI formulation and outlines the key features. Section 3 discusses the key differences to the existing MC estimators and their expected asymptotic behavior. Section 4 confirms that Poisson CI asymptotically converges to analog estimator's CI, while successfully reconstructs the CLT CI fine structures despite relied only on the number of observed reactions. Finally, Sections 5 and 6 confirm the robustness of Poisson CI formulation when the standard CLT often breaks down.

## 2. Poisson Confidence Interval Formulation

There are several variants of Poisson CI formulation, but we will follow the Garwood variant first proposed in [7]. The derivation begins by introducing a Chi-squared distribution cumulative distribution function (CDF)

$$P(Z \geq z) = \frac{\Gamma(\nu/2, z/2)}{\Gamma(\nu/2)}. \tag{1}$$

Equation (1) dictates that the probability of sampling Chi-squared deviates (data) $Z$ bigger than a constant $z$ depends on the complete $\Gamma(x)$ and incomplete $\Gamma(x, y)$ gamma function. Substituting $\nu = 2(r+1)$ and $z = 2\mu$ to Eq. (1) so

$$P\{\chi^2[2(r+1)] \geq 2\mu\} = \frac{\Gamma(r+1,\mu)}{\mathrm{r!}}, \quad (2)$$

noting that $\Gamma(r+1) = r!$.

The Poisson CDF is defined as

$$P(X \leq r) = \frac{\Gamma(r+1,\mu)}{r!} = \frac{\alpha}{2}, \quad (3)$$

meaning probability of sampling *below or equal* to $r$ events depends *solely* on number of observed events $r$. Notice that the right-hand sides (RHS) of Eqs (2) and (3) are identical, therefore

$$P(X \leq r) = P\{\chi^2[2(r+1)] \geq 2\mu\} = \frac{\alpha}{2}. \quad (4)$$

The key point of Eq. (4) is that Poisson CI can be derived solely based on the equivalent Chi-squared distribution. Now, Poisson CI can be constructed by inverting the $\chi^2$ term in Eq. (4) such that $\mu$ equals to the tail probability.

The Poisson upper bound can be defined as

$$\mu_U = \chi^2_{1-\alpha/2}[2(r+1)]\,/\,2\,. \quad (5)$$

Note that the Poisson CI upper bound is strictly a *Chi-squared deviate* calculated from the inverse of Chi-squared CDF given the degree of freedom (dof) and tail probability $\alpha$. The Poisson lower CI bound is

$$P_{\mu_L}(X \geq r) = \alpha/2\,. \quad (6)$$

Notice that Eq. (6) asks different question than Eq. (4): what is the probability of sampling *greater than or equal* to $r$ events. Equation (6) is reconstructed as

$$P_{\mu_L}(X \geq r) = 1 - P_{\mu_L}(X \leq r-1) = \alpha/2\,, \quad (7)$$

that reads the probability of sampling greater than or equal to $r$ events is equivalent to the total probability subtracted by the probability of sampling less or equal to $r-1$. Recall the equivalence in Eq. (4), Eq. (7) is now

$$P_{\mu_L}(X \geq r) = P[\chi^2(2r) \geq 2\mu_L] = 1 - \alpha/2\,. \quad (8)$$

Therefore, the lower bound of Poisson CI is

$$\mu_L = \chi^2_{\alpha/2}(2r)\,/\,2\,, \quad (9)$$

Monte Carlo analysts are generally interested in transport event rate CI, so both Eqs. (5) and (9) are normalized with the number of simulated particles $N$. Finally, the Poisson CI for Bernoulli trial-alike tally is defined as

$$\chi^2_{\alpha/2}(2r)\,/\,2N < r/N < \chi^2_{1-\alpha/2}[2(r+1)]\,/\,2N\,. \quad (10)$$

Many scientific packages adopt [8] to calculate the inverse of Chi-squared CDF.

The direct consequences of using Eq. (10) are:

1. Exact Binomial CI formulation as $p \to 0$,
2. Depends only on $r$, practically eliminate needs of batch mean square accumulator in MC code,
3. Independent to empirical pdf structure and history-wise scores ordering; the final CI is identical even if history-wise scores ordering is permutated,
4. Asymmetric CI bound with significantly larger upper bound at early realization, but converges to normal distribution later when $r$ is sufficiently large (generally $r > 88$),
5. Robust coverage rate even if MC estimators observe no score ($r = 0$).

Attentive code developers may realize the indirect consequences of these key features. Poisson CI implementation is 1) code and implementation agnostic so no significant refactoring is needed to the existing code base, 2) off the shelf, can be calculated after simulations from an existing dataset, and 3) produces *the same final CI* in any truly dynamic load balancing strategy as long as the scheme produces the same mean when the simulation is repeated. This is a truly paradigm shift to the standard CLT statistical model in MC codes.

### 3. Differences to the Existing MC Tally Estimators

The usual CLT statistics procedure is 1) accumulate scores from $N_s$ histories within a batch, 2) after current batch is completed, accumulate batch mean and batch mean squared, 3) after all batches are completed, calculate the final mean and CLT CI. While the underlying statistical procedures are the same, there are three fundamentally different approaches to quantify the score during history transport: 1) analog, 2) track length, and 3) collision estimators. These three estimators will produce a numerically identical mean as $N \to \infty$ but different CI because each estimator is governed by different statistical distribution.

Analog MC estimator is perhaps the direct representation of what Bernoulli experiment is. In analog estimator, MC code records how many *observed* event/reaction occurred $r$ within a simulation; there will be two pile ups in history-wise scores histogram plot: 1) at $x = 0$ representing number of histories simulation does not observe select event and 2) at $x = 1$ when select event occurs—identical to one expects from a binomial distribution. Because the distance from each pile to the sample mean is relatively far away, sample variance obtained from analog estimator is generally higher than the two other estimators. In an extreme case where a very rare transport event is involved, analog estimators require prohibitively a larger $N$ than the two estimators to obtain reliable results. If $N$ is too small, analog estimator will rarely *ever* observe the event so MC code is likely to produce erroneous results. Perhaps this may be the reason why no MC codes use analog estimator by default unless necessary. Either way, CI obtained from analog estimator (governed by Binomial distribution) is expected to be numerically identical to the proposed Poisson CI formulation as $p \to 0$ and $N \to \infty$.

There is no strict limitation on how the number of observed events $r$ is obtained, so it will be beneficial to use other MC estimators known to converge faster such as collision and track length estimators. The collision estimator accumulates ratio of select reaction to the total macroscopic cross section $\Sigma/\Sigma_t$ at every collision while track length estimator accumulates $d \times \Sigma$ where $d$ is distance to the next collision site. After each history or batch, Poisson CI can be calculated given the *expected* number of observed events

$$\hat{r} = \text{batch mean accumulator} = X\,N \approx r, \qquad (11)$$

where $X$ is the current simulation mean. Note that Eq. (11) is strictly valid because the three estimators are theoretically guaranteed to produce the same mean. Cautious readers may question the validity of Poisson formulation for these two estimators because $\hat{r}$ will be a floating number, but investigation to the source code implementation of [8] reveals that the $\chi^2$ inverse CDF function can accept a floating number.

Visual inspection of these two estimators' score histogram reveals that these estimators are generally governed by a continuous, wider distribution, somewhat similar to an exponential distribution. Since simulation mean from Bernoulli experiment is constrained between $0 < X < 1$, the distance of each deviate of these two estimators to the sample mean is generally closer than the analog's due to the nature of having continuous and wider distribution; variance from track length estimator is generally the smallest, followed by collision estimator, and later analog estimator. This is a reasonable behavior because a particle always scores to track length whenever they enter the volume, occasionally contributes to collision estimator if they are inside and have a collision, and only contribute to analog score if they are inside, have a collision, and select reaction is sampled; the batch means fluctuation is the smallest on track length, followed by collision, and later analog estimator. Consequently, Poisson CI width $(\mu_U/N - \mu_L/N)$ can be several factors larger than both collision and track-length CI widths. The observed CI width ratio is defined as

$$R = \frac{(\mu_U - \mu_L)/N}{\sigma^{CLT}}. \qquad (12)$$

Recall that Chi-squared distribution converges to normal distribution $\mathcal{N}(\nu, \nu)$, the approximate ratio is

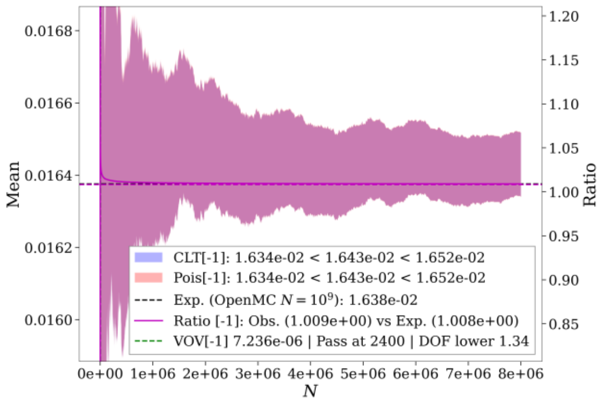
$$R_{exp} = \frac{\sqrt{r/N}}{\sigma} \qquad (13)$$

where $\sigma$ is the current sample standard deviation. Equation (13) is generally valid for $r > 88$ and computationally cheap, particularly useful for later analysis where Poisson and CLT CI (single batch-size method [9]) are calculated at every history, up to one billion histories.

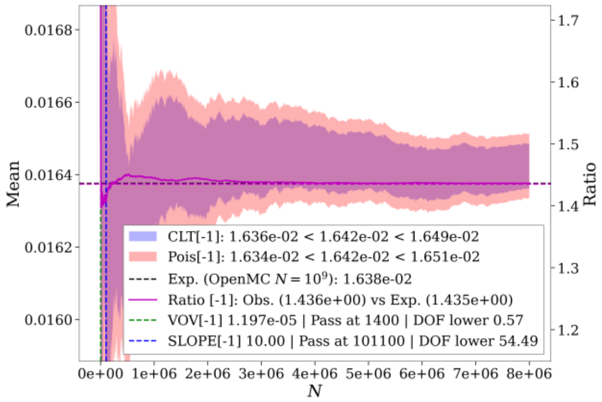
## 4. Asymptotic Behavior of Poisson CI

We introduce a concentric sphere model made in an inhouse MC code which has been verified and validated against the most recent OpenMC version 0.15.3 [1]. The concentric sphere model is divided into 15 spheres, each is one mean free path (mfp) apart. The simulation is a one-group transport with 10% absorbing material and an isotropic point source at the center of cell 1, similar to thermal neutron diffusion in an optically thick, homogeneous moderator.

For the purpose of examining Poisson CI asymptotic behavior, the inhouse MC code exports history-wise absorption scores for a total of five cases: the three estimators in analog simulations and the collision & track length estimators in implicit absorption simulations. Each case is simulated with $N = 8 \times 10^6$ and different random number segment, so these five cases are independent of each other. After simulations were completed, the CLT & Poisson (Eq. 10) CIs, observed (Eq. 12) and approximate (Eq. 13) ratios, variance of variance (VOV) [4], and Pareto slope [4] were calculated for later analysis. For an additional V&V, an independent OpenMC run is simulated with one billion histories and set as the reference solution (Exp.).




(A) Analog estimator




(B) Track-length estimator

Figure 1. Analog run: Cell 10 Poisson and CLT CIs evolution. Poisson CI converges to the same numerical value as the analog CI obtained from CLT.

Figure 1 confirms the applicability of Poisson formulation for estimating reaction rate CI. Both the observed and approximated ratios are approximately unity in Fig. 1A, so these two CIs are practically identical as $N$ is moderately large. Furthermore, Fig. 1B shows the Poisson formulation successfully reconstructs the CLT

CI fine structure despite relied solely on the evolution of $r$ as a function of $N$.

Visual inspection of $R$ in Fig. 1B confirms the discussion in Section 3 that the Poisson CI is expected to be wider than the collision and track length CIs. The same behavior is observed on the two implicit absorption cases, albeit with bigger ratios; given the same $N$, a faithful VRT strategy will produce a smaller variance than an equivalent analog simulation. Most analyst attributes this behavior because "the simulation sees more particles with relatively the same score and weight", but the study in [1] reveals that VRT can increase the probability of small score bins, while extremely reduce the probability of sampling high score bins on the pdf right tail; a faithful nonanalog simulation will produce the same numerical mean, but relatively smaller variance than an analog simulation due to the absence of these right tail bins.

The select cases confirm that Poisson CI formulation converges to one expects from the CLT standard statistical procedure, while successfully reconstructs the CLT CI fine structures despite relied solely on the evolution of $r$. Attentive readers may realize that there is no strong motivation to apply Poisson formulation for the concentric sphere model because the event probability is not extremely low. The ten statistical checks [3,4] further suggest that there is no anomalous behavior during tally convergence after history 2,400 (Fig. 1A) and 101,000 (Fig. 1B). Section 5 will investigate the Poisson CI formulation for a difficult problem where the standard CLT statistical model is expected to break down.

### 5. Robust CI Estimates at Early $N$

We devise a new problem where an extremely rare transport event $p \ll 0.001$ is involved by adopting the point detector problem in [10]. Point detector essentially calculates the probability of particles streaming from current collision site towards a single point in space—A Bernoulli trial-alike tally. Unfortunately, the probability of a particle to *precisely* go into a single point in space is infinitesimal, practically eliminates direct use of analog estimator. Instead, MC code uses a next event estimator where the probability of history $i$ to contribute to a point detector at $\mathbb{R}_p(x, y, z)$ is

$$x_i = \sum_{j}^{C} \omega_{i,j}\, P(\Omega_i \to \Omega')\, P(\mathbb{R}_{i,j} \to \mathbb{R}_p)\,. \tag{14}$$

Equation 14 essentially accumulates the product of particle weight, probability of sampling an outgoing angle $\Omega'$ to precisely $\mathbb{R}_p$, and probability of streaming towards $\mathbb{R}_p$ without collision from current collision site $j$ for a total of $C$ collision sites. Unfortunately, these terms vary significantly depending on the collision site phase space $\mathbb{R}_{i,j}$. Figure 2 shows that the point detector tally density plot has a strong right skewed distribution spanning from $10^{-8}$ to ~0.4. The extremely right-skewed distribution often causes CLT estimate to break down because the simulation failed to sample the

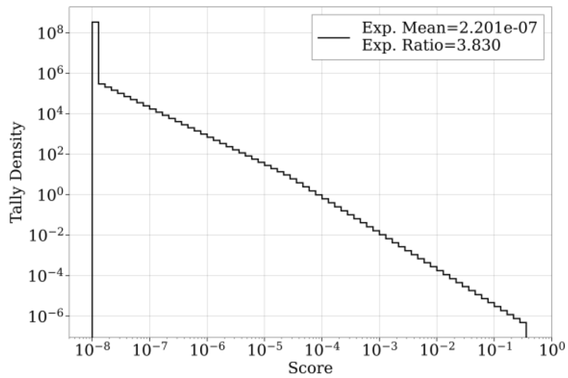



Figure 2. Point detector tally density reconstructed from the original point detector problem in [10].

extremely small probability but important high score bins unless $N$ is sufficiently large (often times, hundreds of millions). The statistical checks in [10] confirm the precarious nature of point detector problem where 600 million particles are needed to pass these checks. These problematic behaviors perhaps explained why some authoritative sources recommend MC analysts to be extremely careful on the use of point detectors.

How do we calculate the Poisson CI if the next event estimator is not an analog estimator? The use of next event estimator is *actually* analogous to how collision and track length estimators substitute analog estimator (see Section 4); despite governed by different statistical distributions, these estimators obtain a valid *expectation value* of select transport event in Bernoulli experiment. Consequently, Eq. (10) can be used to obtain the Poisson CI given the expected number observed events $\hat{r}$.

We reconstruct the underlying point detector pdf (and CDF) in [10] to circumvent the prohibitively large $N$ requirements. Now, history-wise scores can be sampled directly from the reconstructed CDF without the expensive computational requirement of the original problem in [10]. Figure 2 shows the (expected) tally density plot of the reconstructed point detector problem. Please note that the reconstructed pdf is not precisely the same as [10] which explains why the expected mean (and later, the statistical checks results) is different.

We repeat the same procedures in Section 4, but directly calculate these metrics on the fly as the script sample score. Note that these metrics are still calculated at every history but only stored every 1,000 histories which saves ten TBs of storage. The statistical checks in Fig. 3 suggest that the CLT CI should be a reliable estimate after 186 million histories, but CLT CI consistently *barely covers* the expected mean even after five hundred million histories revealing the precarious nature of the point detector problem. In fact, passing statistical checks does not guarantee that the mean and the CLT CI are reliable estimates; these checks only tell analysts that there is *probably no anomalous behavior* on these metrics after select history, but it has no information on the accuracy with respect to the reference solution. Section 6 later confirms that despite passing all

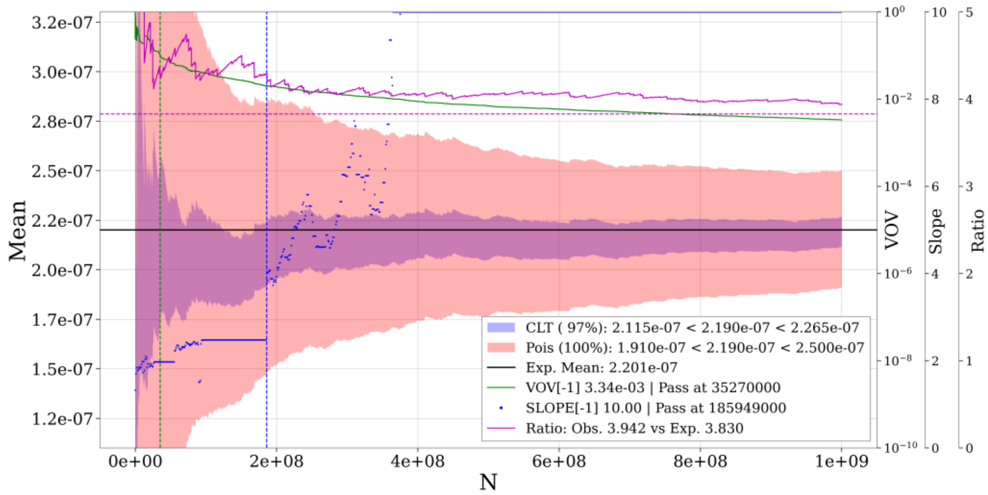



Figure 3. CIs evolution in the point detector problem.

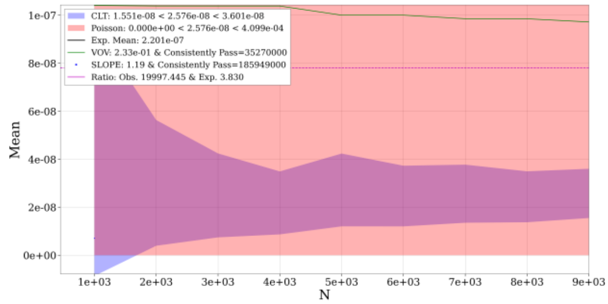
statistical checks, CLT CI often times does not cover the expected mean (lower coverage rate than expected).

The CLT CI behavior appears to be well behaved in Fig. 3, but Fig. 4 shows CLT CI breaks down in the early realization. The CLT CI consistently does not cover the expected mean and the width is up to 20,000 smaller than it should have been. The arrival of single high-score histories consistently causes a spike on the CLT CI as seen in Fig. 4B and later histories. These disturbing behaviors should demonstrate that CLT is not an appropriate statistical model for problems where extremely rare events are involved. On the contrary, Fig. 4 reveals the superior performance of the Poisson CI formulation which consistently maintains a 100% coverage rate during the whole simulation. This is a remarkable performance noting that the simulation mean at early $N$ is an order magnitude smaller than it should have been.

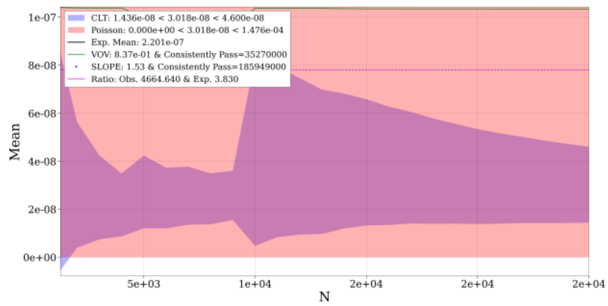
## 6. Robust Coverage Rate on Rare Transport Event

We divide each sphere in the concentric sphere model into 100 equi-volume spheres (a total of 1,500 cells), then rerun the five select cases to inspect the Poisson CI coverage rate. After completion, the same metrics in Section 4 are recalculated from the new history-wise absorption scores; the expected mean for a split cell 15 absorption reaction rate decreased a factor of 100 from $\sim 1.6 \times 10^{-3}$ to $\sim 1.6 \times 10^{-5}$.

Figure 5 shows the coverage rate for the split cells 15 in the new concentric sphere model. In an ideal case, the number of split cells to cover the expected mean should fluctuate precisely around 95 (a 95% target CI), but this is not always the case as seen in Fig. 5A where the collision estimator falls around 93% coverage rate for an extended period of histories (~millions of histories) despite the statistical checks suggesting that the CLT may be a reliable estimate from history 3.4 million. Visual inspection on each split-cell CLT CI evolution reveals the unique feature of current split model where a mean overestimation on certain split cell will




(A) $N = 9 \times 10^3$

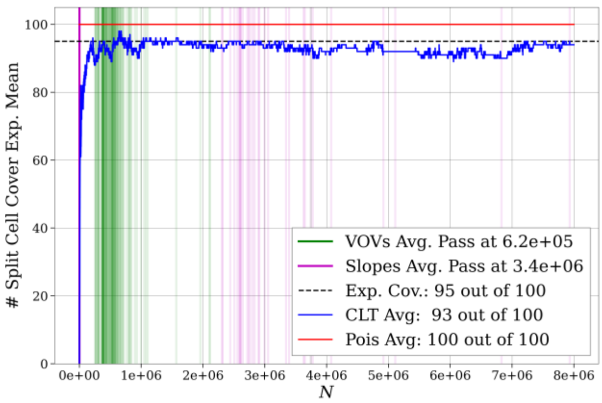



(B) $N = 2.5 \times 10^4$

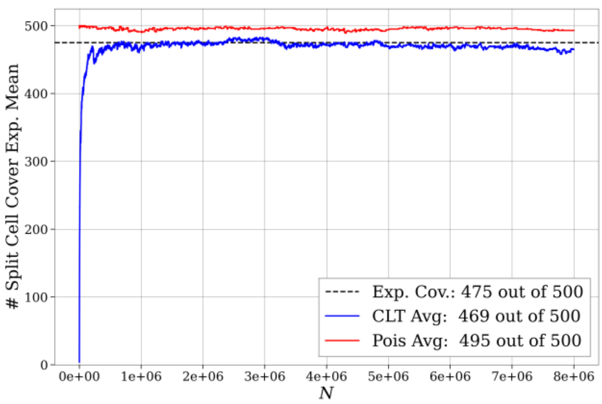
Figure 4. The CLT CI breaks down at early $\boldsymbol{N}$. The CLT CI width is 20,000 smaller than the expected value (Binomial distribution).

consequently introduce an underprediction on the opposing quadrant—a negative autocorrelation. Consequently, the number of split cells does not cover the expected mean is suddenly doubled due to the presence of the under/overestimation behavior. The same behavior is observed on the analog and track length estimators. On the other hand, Fig. 5 reiterates the superior performance of Poisson CI formulation. Poisson CI formulation almost always maintains a 100% coverage during simulation. Occasionally, Poisson CI marginally failed to cover the expected mean due to having slightly higher lower bound than the CLT's.

We repeat the same procedure for the point detector




(A) Analog run: collision estimator




(B) All five cases

Figure 5. Cell 15 coverage rate.

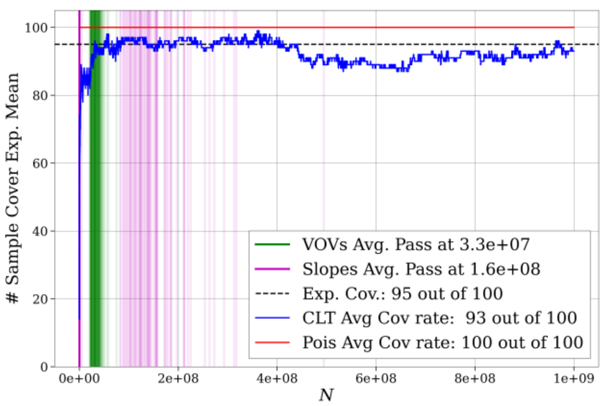



Figure 6. Coverage rate for the point detector problem

problem by collecting an additional 99 samples obtained from independent random number segments. Figure 6 shows how CLT breaks down for rare event estimation. The statistical checks suggest that the CLT CI can be a reliable estimate after history 160 million, but in reality the CLT coverage rate drops to and *consistently* maintains the 87% coverage rate for the last 500 million histories. On the other hand, Poisson CI repeatedly demonstrates a superior performance by consistently maintaining 100% coverage rate during the whole simulation, a remarkable performance noting that there is no information on the right-tail score at early $N$.

Some could argue that the Poisson CI formulation can be characterized as less efficient than the counterparts according to the traditional figure of merit (FOM) metric. This is due to the fact that the Poisson CI width is too conservative (wide), but we remind MC analyst what matters the most is having a good coverage on the true solution than a sufficiently small $\boldsymbol{\sigma}$ but (extremely) poor coverage.

## 6. Ongoing Research

We have established that Poisson CI formulation is applicable to MC radiation transport simulations. The Poisson CI consistently produces a realistic and conservative CI bounds and robust coverage rates of the expected mean despite relied solely on the number of observed events. There are some critical questions on the use of Poisson CI formulation for a source multiplying system. What $N$ should we use in Eq. 10? Would it be number of particles and their daughters scoring to tally volumes? Or their total weights? Preliminary results of the oil-well logging problem (weight window VRT) described in [1] suggests that $N$ should remained as the total number of histories simulated, not number of particles and their daughters, nor sum of weights scoring to tally volumes. We also acknowledge that extensive use of Poisson CI is not always desired especially when the standard CLT procedure is guaranteed to produce reliable estimate (i.e., $p \to 1$), so we actively explore strategies for MC code to switch from Poisson to the standard CLT procedure when certain criteria are met.

## REFERENCES

[1] A R. Hakim and D. Fynan, A., Wrong results when random number stride is exceeded in Monte Carlo radiation transport simulation, Nucl. Eng. and Tech., under review, 2026.

[2] US Nuclear Regulatory Commission, Regulatory Guide 1.190: Calculational and Dosimetry Methods for Determining Pressure Vessel Neutron Fluence, 2001.

[3] R. A. Forster, S. P. Pederson, T. E. Booth, Ten New Checks to Assess the Statistical Quality of Monte Carlo Solutions in MCNP, LA-UR-94-227, LANL, 1994.

[4] M. Lemaire, H. Lee, D. Lee, Implementation of Tally Convergence Tests in UNIST Monte Carlo Code MCS, Trans. of KNS Spring Meeting, 2018.

[5] U.S. Atomic Energy Commission, "Technical Report on Anticipated Transients Without Scram for Water-Cooled Power Reactors," WASH-1270 (1973).

[6] D. A. Fynan, J. C. Lee, "Comparison of Three Methods for Reliability Quantification", Trans. of the Ame. Nucl. Soc., vol. 100, no. 1, p. 435-436, Jun. 2009.

[7] F. Garwood, Fiducial Limits for Poisson Distribution, Biometrika, vol. 28, no. 3/4, Dec. 1936.

[8] A. R. DiDonato and A. H. Morris, Computation of the Incomplete Gamma Function Ratios and Their Inverse, ACM Trans. Math. Softw., vol. 12, no. 4, p. 377-393, Dec. 1986.

[9] A. Dubi, On the analysis of the Variance in Monte Carlo Calculations, Nucl. Sci. Eng. 72 (1979), 108-110.

[10] S. P. Pederson, R.A. Forster, T.E. Booth, Confidence Interval Procedures for Monte Carlo Transport Simulations, Nucl. Sci. Eng. 127 (1997) 54–77.