Verificarlo, Verrou, Interflop: Floating point computing verification on new architecture and large scale systems

Devan Sohier[1], Pablo Oliveira[1,2], François Févotte[3], Bruno Lathuilière[3], Éric Petit[2,4], Olivier Jamond[5]

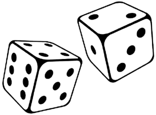
[1]UVSQ; [2]Exascale Computing Research Lab; [3]EDF R&D Pericles; [4]Intel DCG E&G and [5]CEA

September 27, 2019

Section 1

Toward a common theoretical framework

Statistical Analysis of Stochastic Arithmetic: Motivation



- Stochastic Arithmetic
  - Numerical errors modeled by introducing random perturbations.
  - Estimate significance of result by collecting many samples.
- Motivation for statistical analysis
  - How many stochastic samples should be run?
  - What is the probability of over-estimating the number of significant digits?
  - Can we give a sound confidence interval for the number of significant digits?

Example: Kahan 2x2 System

- Ill-conditioned linear system (condition number 2.5×10^8).
- We solve it with the Cramer's formula.

$$\begin{pmatrix} 0.2161 & 0.1441 \\ 1.2969 & 0.8648 \end{pmatrix} x = \begin{pmatrix} 0.1440 \\ 0.8642 \end{pmatrix} \tag{1}$$

$$x_{\text{real}} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \qquad x_{\text{IEEE}} = \begin{pmatrix} 1.9999999958366637 \\ -1.9999999972244424 \end{pmatrix} \tag{2}$$

- The IEEE-754 result has 8 significant decimal digits.
- $x_{\text{IEEE}}[0]$ has 28.8 significant bits.

Kahan 2x2 System – Stott Parker's significant digits

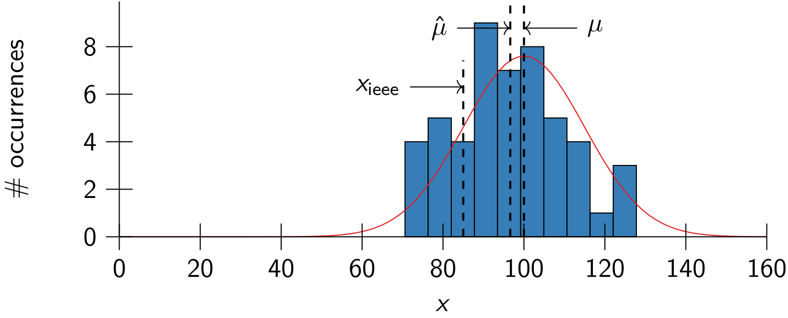
- With Verificarlo, we collect 10000 $t = 52$ FULL MCA samples.

$$s_{\text{PARKER}} = -\log_2 \frac{\hat{\sigma}}{|\hat{\mu}|} \approx 28.5.$$

- But how confident are we that it is a good estimate? Could we have used a smaller number of samples and still get a reliable estimation of the results quality?

Some notations

- x_{IEEE} is the IEEE-754 result
- $X_1, X_2, \ldots, X_n$ are the values returned by n runs of the program using stochastic arithmetic. These are seen as realizations of a random variable X.
- $\hat{\mu}$ and $\hat{\sigma}$ are the empirical average and standard deviation.
- μ and σ are the mean and std. deviation of the random variable X.



Choosing a reference value

- We require a reference value against which accuracy is measured.
- Examples of common reference values,
  - x_{real}, if the exact solution is known.
  - x_{IEEE}, when the program is deterministic.
  - $\hat{\mu}$, if the program is non-deterministic.
  - Y, a random variable, to compare two implementations of an algorithm or measuring significance between runs of the same program.

Modeling the error

- Four kind of scenarios are studied in our paper.
- In each case the error is modeled by a random variable Z.
- For simplicity, in the following we consider the relative precision with scalar reference.

	reference x	reference Y
absolute precision	$Z = X - x$	$Z = X - Y$
relative precision	$Z = X/x - 1$	$Z = X/Y - 1$

- With no error, the expected result of Z is 0.

Significant bits

- Stott Parker defines the significant digits between x and y as the largest s that satisfies $|x/y - 1| \leq 2^{-s}$.
- Or put more simply, the error is less than 2^{-s}.
- We naturally extend this definition to Z the random variable modeling the stochastic error.

Significant bits

The number of significant digits with probability p_s can be defined as the largest number s such that

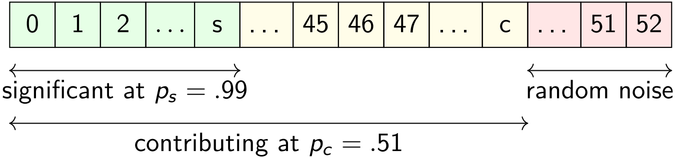
$$\mathbb{P}\left(|Z| \leq 2^{-s}\right) \geq p_s. \tag{3}$$

0	1	2	...	s	...	45	46	47	48	49	50	51	52

s significant — error satisfies $|Z| \leq 2^{-s}$

Contributing bits

- Bits after s still can encode useful information about the result.
  - Even if bits on its left are wrong, they can improve the accuracy...
  - ...if they are correct on average ($p_c > 51\%$).
  - Keeping these bits improves the rounded result on average.
- A bit k after s contributes to the result with probability p_c *iff* the k-th bit of Z is 0 (no error in this bit) with probability p_c.

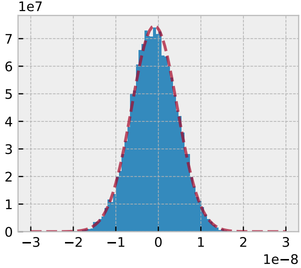


Results

1. Probability for significance and contribution for Normal Centered Distributions.
2. Probability for significance and contribution for General Distributions.

Preprint: *Confidence Intervals for Stochastic Arithmetic*, D. Sohier, P. de Oliveira Castro, F. Févotte, B. Lathuilière, E. Petit, O. Jamond. 2018.

Normality of the Kahan 2x2 System



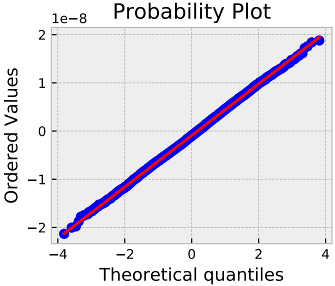
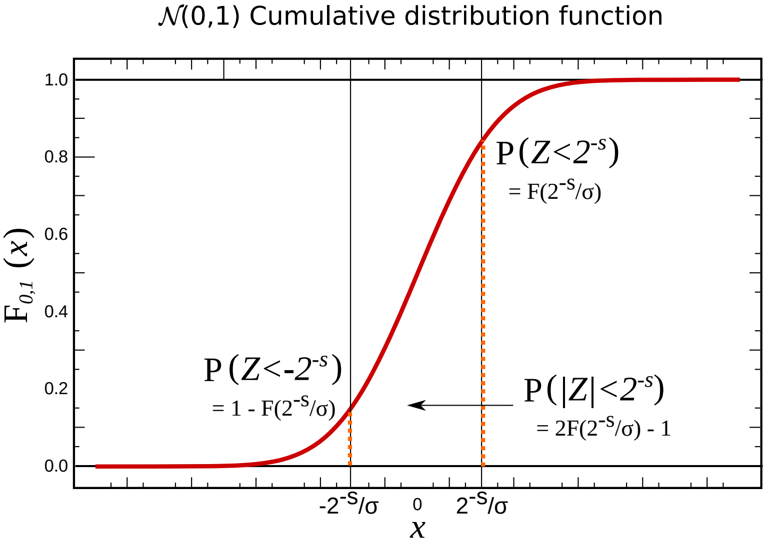


Figure: Normality of 10000 samples of $X[0]$ with $t = 52$ and FULL MCA

- We take as reference the empirical mean $\hat{\mu}$.

Centered Normal Hypothesis: Significant bits



Centered Normal Hypothesis: Significant bits

Theorem

For a normal centered error distribution $Z \sim \mathcal{N}(0, \sigma)$, the s-th bit is significant with probability

$$p_s = 2F\left(\frac{2^{-s}}{\sigma}\right) - 1,$$

with F the cumulative function of the normal distribution with mean 0 and variance 1.

- By inverting this formula, we can provide a formula for the number of significant digits that only depends on σ and p_s,

$$s = -\log_2(\sigma) - \log_2\left(F^{-1}\left(\frac{p_s + 1}{2}\right)\right).$$

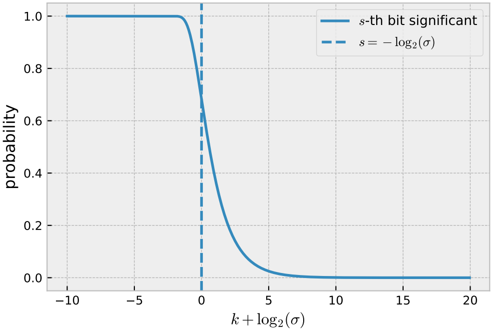


Figure: Profile of the significant bit curve $p_s = 2F\left(\frac{2^{-s}}{\sigma}\right) - 1$

- If we take the empirical average as reference value, we fall back into Stott Parker definition of significant bits assuming a large number of samples $-\log_2(\sigma) = -\log_2(\frac{\sigma_X}{|\hat{\mu}|})$
- The digit of Stott Parker's formula has 68 % chances of being significant. (1-sigma rule)
- If we substract 1.37 bits from Stott Parker's formula, the resulting bit has 99 % chances of being significant.

CNH: Taking into account the estimation bias

$$s = -\log_2(\sigma) - \log_2\left(F^{-1}\left(\frac{p_s+1}{2}\right)\right).$$

- Why is this formula independent of the number of samples n ?
- σ is unknown; we can only estimate it from $\hat{\sigma}$
- For normal distributions, the following confidence interval with confidence $1-\alpha$ based on the χ^2 distribution with $(n-1)$ degrees of freedom is sound [**?**]:

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}. \tag{4}$$

- In the following we choose a confidence of $1-\alpha = 95\%$.

CNH: Significant bits lower bound

- By combination, we produce a sound lower bound on the significant bits,

$$s \geqslant \underbrace{-\log_2(\hat{\sigma}) - \underbrace{\left[\frac{1}{2}\log_2\left(\frac{n-1}{\chi^2_{1-\alpha/2}}\right) + \log_2\left(F^{-1}\left(\frac{p+1}{2}\right)\right)\right]}_{\delta_{\mathrm{CNH}}}}_{\hat{s}_{\mathrm{CNH}}} \qquad (5)$$

- For $n = 30$ samples and $p = 99\%$ $s \geq -log_2\hat{\sigma} - 1.792$
- For $n = 15$ samples and $p = 99\%$ $s \geq -log_2\hat{\sigma} - 2.023$

($log_2\hat{\sigma}$ is Stott Parker's formula when the reference is $\hat{\mu}$)

CNH: Contributing bits

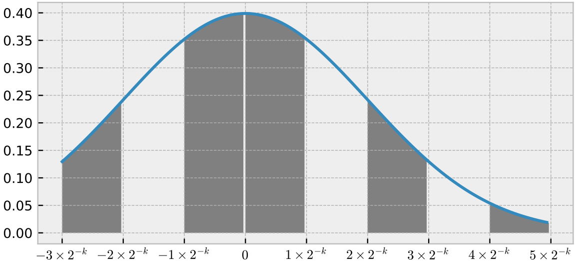


Figure: Normal curve; the gray zones correspond to the area where the k-th bit contributes to make the result closer to 0 (whatever the preceding digits).

$$
\begin{aligned}
& \exists i, \left\lfloor 2^k |Z| \right\rfloor = 2i \\
\Leftrightarrow \quad & 2i \leq 2^k |Z| < 2i + 1 \\
\Leftrightarrow \quad & 2^{-k}(2i) \leq |Z| < 2^{-k}(2i+1). \qquad (6)
\end{aligned}
$$

CNH: Contributing bits

Theorem

For a normal centered error distribution $Z \sim \mathcal{N}(0, \sigma)$, when $\frac{2^{-c}}{\sigma}$ is small, the c-th bit contributes to the result accuracy with probability

$$p_c \sim \frac{2^{-c}}{2\sigma\sqrt{2\pi}} + \frac{1}{2}.$$

If we wish to keep only bits improving the result with a probability greater than p, then we will keep c contributing bits, with

$$c = -\log_2(\sigma) - \log_2(p_c - \frac{1}{2}) - \log_2(2\sqrt{2\pi}). \tag{7}$$

Again, this formula only depends on σ and the probability p_c. As previously, a sound lower or upper bound can be computed with the Chi-2 confidence interval of σ.

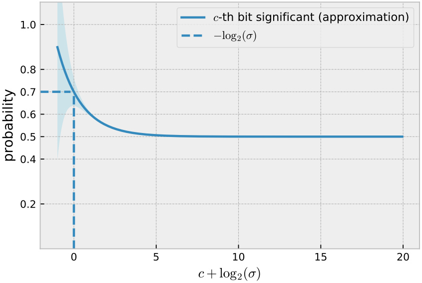


Figure: Profile of the contribution bit curve: The shaded area represents the bound on the error. The approximation is very tight for probabilities less than 70%.

Results: Significant bits

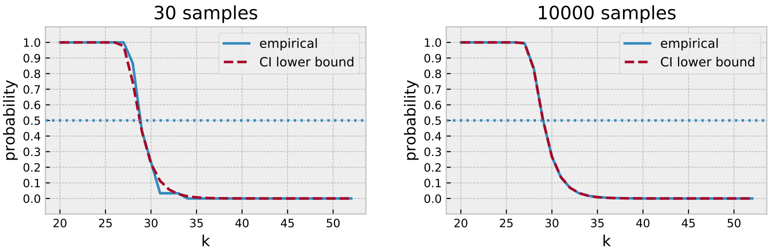


Figure: Significant bits for Cramer $x[0]$ variable computed under the normal hypothesis using 30 and 10000 samples. The Confidence Interval (CI) lower bound is computed by using the probability of theorem 1 and bounding σ with a 95% Chi-2 confidence interval.

Results: Contributing bits

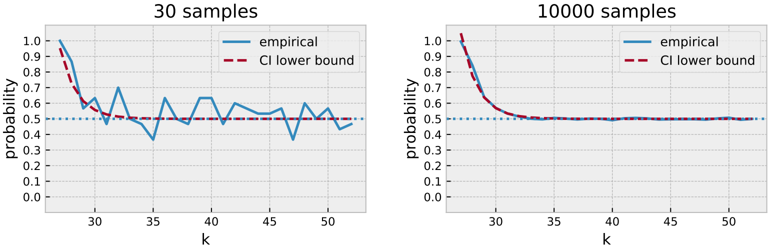
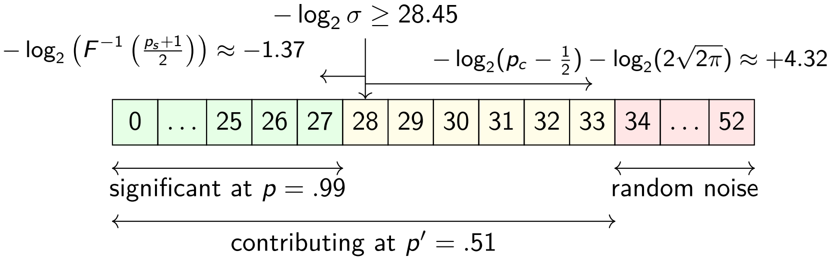


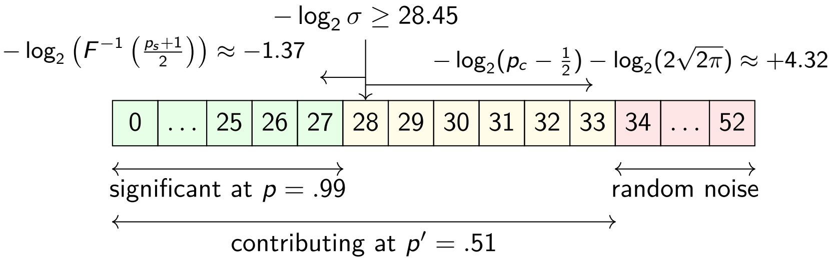
Figure: Contributing bits for Cramer $x[0]$ variable computed under the normal hypothesis using 30 and 10000 samples.

Summary: Significant and Contributing bits in the CNH (1/2)



1. We estimate a lower bound for $-log\sigma \geq 28.45 \approx -\log_2 \hat{\sigma} - \frac{1}{2}\log_2\left(\frac{n-1}{\chi^2_{1-\alpha/2}}\right)$
2. We apply a shift left (computed with $p_s = 99\%$) to get a safe significant bits lower-bound.
3. We apply a shift right (computed with $p_c = 51\%$) to get a safe contributing bits lower-bound.

Summary: Significant and Contributing bits in the CNH (2/2)



- Contributing bits help decide how many digits to print or store during a check-point restart.
- Only keeping contributing bits can help reducing storage and database sizes!

- What if the distribution is not centered normal?

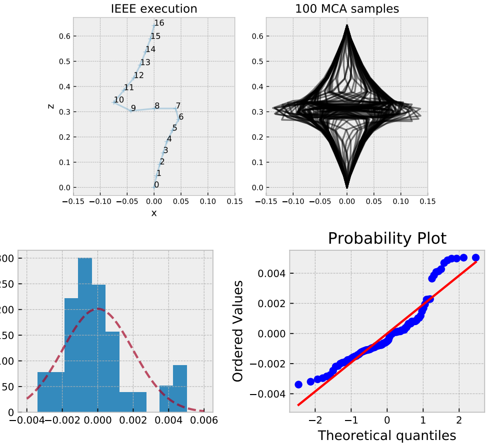
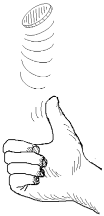


Figure: Non normality of buckling samples on z axis and node 1. Shapiro Wilk rejects the normality hypothesis.

Model by Bernoulli Trials (1/2)

- Let us choose a single k in the mantissa and single sample i among the n samples.
- We can define two binary tests,
  - $S_i^k =$ "$|Z_i| \leq 2^{-k}$", true *iff* for the i-th sample the k-th first bits are significant.
  - $C_i^k =$ "$\lfloor 2^k \, |Z_i| \rfloor$ is even", true *iff* for the i-th sample the k-th bit is contributing.
- With n samples we have n Bernoulli Trials.
- The trials are realizations of two Bernoulli random variables S^k and C^k.



Model by Bernoulli Trials (2/2)

- We choose a given k.

Sample X_1	0	1	2	...	k	...	48	49	50	51	52	S_1^k Success

$$|Z_1| \leq 2^{-k}$$

Sample X_2	0	1	2	...	k	...	48	49	50	51	52	S_2^k Failure

$$|Z_2| > 2^{-k}$$

Sample X_3	0	1	2	...	k	...	48	49	50	51	52	S_3^k Success

$$|Z_3| \leq 2^{-k}$$

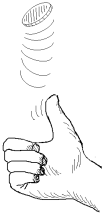
- Out of three samples: 2 success and 1 failure; $n_s = 2$.
- Can we estimate the Bernoulli distribution of S^k ?

Bernoulli Estimator

- [?] gives the following lower-bound for the success probability of a Bernoulli distribution at 95% confidence,

$$\frac{n_s + 2}{n + 4} - 1.65\sqrt{\frac{(n_s + 2)(n - n_s + 2)}{(n + 4)^3}}$$

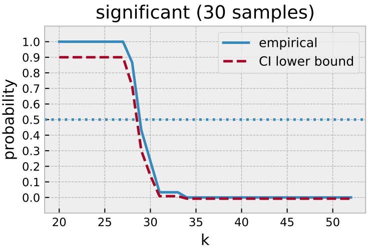
- By counting for S_i^k the number of successes n_s (where the first k digits are significant) we can derive a safe lower-bound probability.

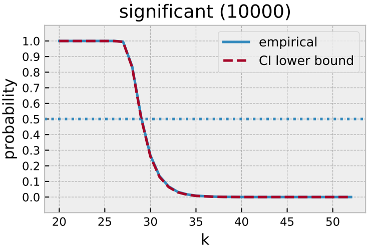


Probability vs. Confidence

- We want to estimate the probability p_s of the s-th bit being significant.
- Suppose $p_s = \frac{1}{3}$ is the true parameter of the Bernoulli distribution.
- We do m different samplings of $n = 10$ values:
  - 1st sampling: $n_s = 3 \rightarrow p_s \in [.15, .57]$
  - 2rd sampling: $n_s = 8 \rightarrow p_s \notin [.52, .91]$
  - 3nd sampling: $n_s = 2 \rightarrow p_s \in [.08, .48]$
  - ...
- The confidence $1 - \alpha$ measures the proportion of samplings that produce an interval containing p_s.
- Increasing the number of samples n reduces the probability of a biased interval and therefore increases the confidence.

Example of Bernoulli Estimator on Kahan's system





Special Case: No failures

- Let us consider the largest k so that S_i^k is true for all i. In other words, k is significant in all the collected samples.
- In that case, [**?**] shows that $\mathbb{P}(S^k) > p$ with confidence $1 - \alpha$ if we have

$$n = n_s \geq \left\lceil \frac{\ln(\alpha)}{\ln(p)} \right\rceil$$

- This formula gives us a simple criterion for choosing a minimal number of samples depending on the required confidence level.
  1. Choose a probability and confidence level that are acceptable for your experiment: eg. $p = 90\%$ and $1 - \alpha = 95\%$
  2. Compute and collect the required number of samples, here $n = 29$.
  3. Find the largest k that is significant for all samples; that k is significant with $p = 90\%$ at confidence level 95%.

How many samples are required?

Confidence level $1-\alpha$	Probability p								
	0.66	0.75	0.8	0.85	0.9	0.95	0.99	0.995	0.999
0.66	3	4	5	7	11	22	108	216	1079
0.75	4	5	7	9	14	28	138	277	1386
0.8	4	6	8	10	16	32	161	322	1609
0.85	5	7	9	12	19	37	189	379	1897
0.9	6	9	11	15	22	45	230	460	2302
0.95	8	11	14	19	29	59	299	598	2995
0.99	12	17	21	29	44	90	459	919	4603
0.995	13	19	24	33	51	104	528	1058	5296
0.999	17	25	31	43	66	135	688	1379	6905

Table: Number of samples necessary to obtain a given confidence interval with probability p, according to the Bernoulli estimator (*i.e.* without any assumption on the probability law).

EuroPlexus Buckling Analysis (1/2)

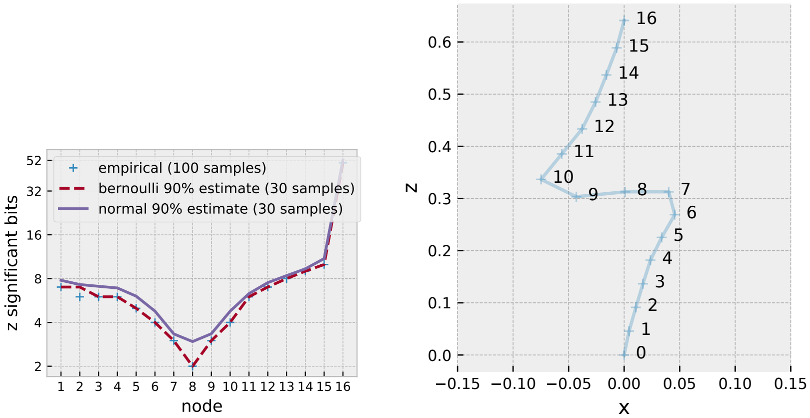


Figure: Significant bits on the z axis distribution. Bernoulli estimation captures precisely the behavior (except for node 2). Normal formula overestimates the number of digits, this is expected since the distribution is strongly non normal.

EuroPlexus Buckling Analysis (2/2)

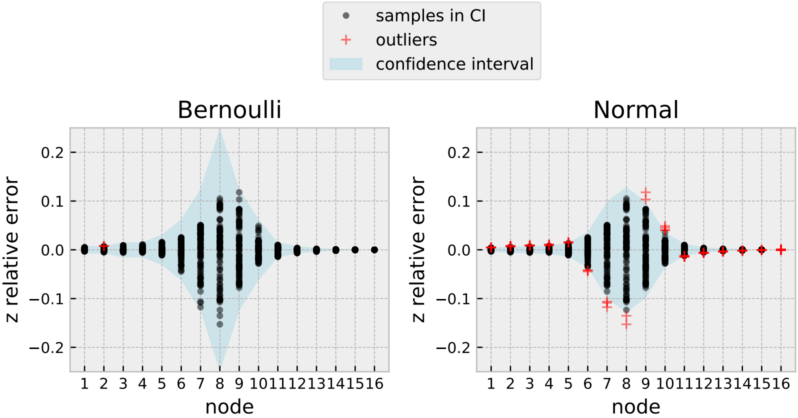


Figure: Relative error between the samples and the mean of the z-axis distribution. The blue envelope corresponds to the computed confidence interval with 30 samples. Black dots are samples that fall inside the CI. Red crosses are outliers that fall outside the CI. In the Bernoulli case, only 3 samples out of 70 fall outside of the interval; which is compatible with the 90% probability threshold.

Limits and Discussion

- These confidence intervals estimate the error of over-estimating s due to sampling errors
  - not enough samples taken or biased sampling
- These confidence intervals do not account for model errors
  - Changes in the dataset
  - Failures of MCA or CESTAC to correctly model FP errors (thread scheduling, model corner-cases, etc.)

Conclusion on Confidence Intervals for Stochastic Arithmetic

- For normal centered distributions:
  - Simple probability formulations for significance and contribution that only depend on $\hat{\sigma}$, n and $1 - \alpha$.
  - Applying a left or right shift to the pivotal $-log_2(\sigma)$ Stott Parker's estimator produces a lower-bound on the number of significant and contributing bits.
- For general distributions:
  - Model each mantissa *bit* as a separate Bernoulli distribution.
  - When only interested in the significant bits, a simple formula computes how many samples are needed to reach a given probability level.
- How can I apply these results to my studies?
  - Tables for the CNH shifts and number of required samples are available in the preprint.
  - A jupyter notebook implemenenting the formulas is also available.

Code_aster : update reference (1/3)

Implementation	$\hat{s}_{\mathrm{MCA}}$		comment
	a	e	
`version0`	Fail	Fail	original version
`version1`	30.89	19.73	fixes an unstable test
`version2`	30.96	19.80	compensated summation
`version3`	32.82	21.65	fully compensated dot product

Table: Summary of the numerical quality assessment of 4 versions of code_aster, using Verrou and the standard MCA estimator with 6 samples.

With `version3` the accuracy seems improved, but we need confidence intervals to update reference value. We choose $p = (1 - \alpha) = 0.995$:

- $N_{sample} = 1058$

Code_aster : update reference (2/3)

Implementation	$\hat{s}_{\mathrm{B}}^{\hat{\mu}}$	$\hat{s}_{\mathrm{B}}^{\mathrm{IEEE}}$	$\hat{s}_{\mathrm{CNH}}$ (normality test p-value)	$\hat{s}_{\mathrm{MCA}}$
`version1`	28	28	29.01 (0.10)	30.59
`version2`	29	29	29.55 (0.89)	31.13
`version3`	30	31	31.22 (0.52)	32.79

(a) quantity a

Implementation	$\hat{s}_{\mathrm{B}}^{\hat{\mu}}$	$\hat{s}_{\mathrm{B}}^{\mathrm{IEEE}}$	$\hat{s}_{\mathrm{CNH}}$ (normality test p-value)	$\hat{s}_{\mathrm{MCA}}$
`version1`	17	17	17.85 (0.10)	19.43
`version2`	18	18	18.39 (0.89)	19.97
`version3`	19	19	20.05 (0.52)	21.63

(b) quantity e

Table: Comparison of stochastic estimators for 3 version of code_aster, with 1058 samples.

Code_aster : update reference (2/3)

Implementation	$\hat{s}_{\mathrm{B}}^{\hat{\mu}}$	$\hat{s}_{\mathrm{B}}^{\mathrm{IEEE}}$	$\hat{s}_{\mathrm{CNH}}$ (normality test p-value)	$\hat{s}_{\mathrm{MCA}}$
`version1`	28.89	28.57	29.01 (0.10)	30.59
`version2`	29.33	29.35	29.55 (0.89)	31.13
`version3`	30.91	31.00	31.22 (0.52)	32.79

(a) quantity a

Implementation	$\hat{s}_{\mathrm{B}}^{\hat{\mu}}$	$\hat{s}_{\mathrm{B}}^{\mathrm{IEEE}}$	$\hat{s}_{\mathrm{CNH}}$ (normality test p-value)	$\hat{s}_{\mathrm{MCA}}$
`version1`	17.73	17.41	17.85 (0.10)	19.43
`version2`	18.16	18.19	18.39 (0.89)	19.97
`version3`	19.75	19.84	20.05 (0.52)	21.63

(b) quantity e

Table: Comparison of stochastic estimators for 3 version of code_aster, with 1058 samples.

Code_aster : update reference (3/3)

`version3` is our new reference.

`version0` analysis

- $\left|\frac{a_{ieee}^{version0} - a_{ieee}^{version3}}{a_{ieee}^{version3}}\right| = 4.29 \times 10^{-10}$
- $\left|\frac{e_{ieee}^{version0} - e_{ieee}^{version3}}{e_{ieee}^{version3}}\right| = 9.84 \times 10^{-7}$

Need to analyze other test cases related to these 2 corrections before integration