A faster estimate with user-specified error for the mean of bounded random variables

Mark Huber
Fletcher Jones Foundation Associate Professor of Mathematics and Statistics and George R. Roberts Fellow
Chair of the Department of Mathematical Sciences
Claremont McKenna College

6 July, 2017

The big picture

The Problem

Find the mean of a stream of bounded random variables

Current Method

Dagum, Karp, Luby, Ross (2000)
About 2.5 to 5 times as slow as CLT

New approach

Asymptotic to CLT without prior knowledge of the variance

The problem

Given a stream of $X_1, X_2, \ldots$ random variables, find their mean to within ϵ relative error with failure probability at most δ.

Prior work

The Central Limit Theorem
de Moivre, Laplace, Gauss, Lyapunov, Lindeberg, Lévy

Sample average

- For finite variance, $\bar{X}$ converges to normality
- Does not say how quickly the convergence occurs
- If convergence is quick (or $X_i \sim \mathrm{N}(\mu, \sigma^2)$,) then need roughly

$$2\frac{\sigma^2}{\mu^2}\ln(2/\delta)$$

How quickly does CLT converge?

The Accuracy of the Gaussian Approximation to the Sum of Independent Variates
Andrew C. Berry,
Trans. Amer. Math. Soc., 49 (1): 122–136, 1941

On the Liapunoff limit of error in the theory of probability
Carl-Gustav Esseen,
Arkiv för matematik, astronomi och fysik. A28: 1–19, 1942

Bounded how far away CLT approximation was from sample average for bounded third central moment

Using Berry-Esseen

Guaranteed conservative fixed width confidence intervals via Monte Carlo sampling
F.J. Hickernell, L. Jiang, Y. Liu, A.B. Owen
Monte Carlo and Quasi Monte Carlo Methods, 105–108, 2012

	X	σ^2/μ^2	kurtosis
known bounds			κ
unknown bounds			
unbounded			

	relative	absolute
error		

	match	greater
CLT	κ small	κ large

Using Berry-Esseen II

Sub-Gaussian mean estimators *L. Devroye, M. Larasle, G. Lugosi, R.I. Oliveira*
Annals of Statistics, 44:2695–2725, 2016

	X	σ^2/μ^2	kurtosis
known bounds			κ
unknown bounds			
unbounded			

	relative	absolute
error		

	match	greater
CLT	κ small	κ large

Catoni M-estimator

Challenging the empirical mean and empirical variance: A deviation study
O. Catoni
Ann. Inst. H. Poincaré Probab. Statist., 48(4):1148–1185, 2012

Using M-estimator requires a rootfinding procedure

- Assume known upper bound on kurtosis...
- ...or known upper bound on σ^2, lower bound on μ^2
- Not an (ϵ, δ)-ras

Extensions to other moments

Input sets for Numerical Integration
R. Kunsch, E. Novak, D. Rudolf
Talk at MCM Montréal 3 July, 2017

Requires known upper bound $M_{p,q}$ on

$$\frac{\mathbb{E}[(X_i-\mu)^p]^{1/p}}{\mathbb{E}[(X_i-\mu)^q]^{1/q}}$$

Light tailed sample averages

An optimal (ϵ, δ)-approximation scheme for the mean of random variables with bounded relative variance
M. Huber
arXiv:1706.01478, 2017

	X	σ^2/μ^2	kurtosis
known bounds		c^2	
unknown bounds			
unbounded			

	relative	absolute
error		

	match	greater
CLT		

What happens when bounds on moments unknown?

Say $B \sim \mathsf{Bern}(p)$

$$\frac{\mathbb{E}[(B-p)^4]}{\mathbb{E}[(B-p)^2]^2} = \frac{p(1-p)^4 + (1-p)(p)^3}{p^2(1-p)^2}$$
$$= \Theta\left(\frac{1}{p^2}\right) \to \infty \text{ as } p \to 0$$

However, B is bounded!

DKLR

An optimal algorithm for Monte Carlo estimation
P. Dagum, R. Karp, M. Luby, and S. Ross
SIAM J. Comput., *Vol 29, No 5, pp. 1484–1496, 2000*

	X	σ^2/μ^2	kurtosis
known bounds	M		
unknown bounds			
unbounded			

	relative	absolute
error		

	match	greater
CLT		

Today

This talk

	X	σ^2/μ^2	kurtosis
known bounds	M		
unknown bounds			
unbounded			

	relative	absolute
error		

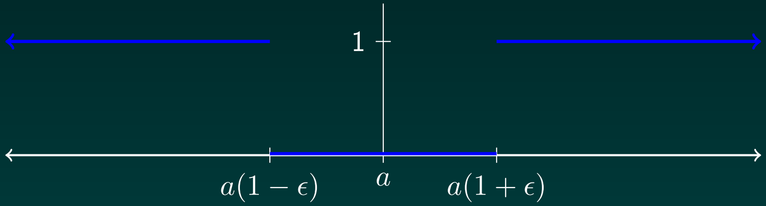
	match	greater
CLT	$\frac{\epsilon M}{\mu} << \frac{\sigma^2}{\mu^2}$	

The new problem

Given a stream of nonnegative $X_1, X_2, \ldots$ *random variables, with known upper bound, but unknown mean, variance, and kurtosis, find the mean to within* ϵ *relative error with failure probability at most* δ*.*

Loss function

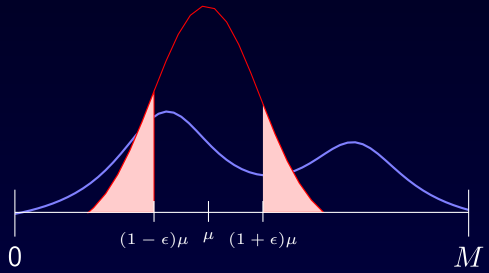
Can view as minimizing the expected loss function "all or nothing"



$$\mathcal{L}(\hat{a}) = \mathbb{1}(|\hat{a} - a| \geq \epsilon a)$$

Finding the mean of $[0, M]$ *random variables*

Given $\epsilon, \delta > 0$



want

$$\mathbb{P}(|\hat{\mu}/\mu - 1| > \epsilon) \leq \delta$$

Call $\hat{\mu}$ an (ϵ, δ)-randomized approximation scheme

When the CLT applies

If sample average was normal, then need (to first order)

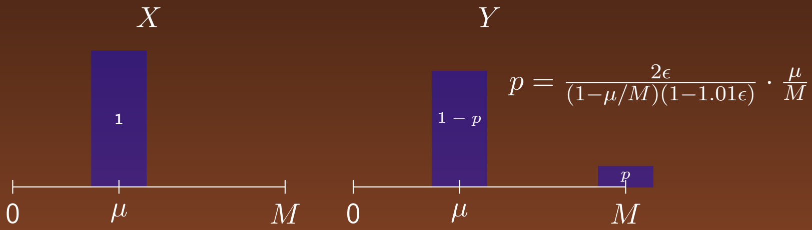
$$2\frac{\sigma^2}{\mu^2}\epsilon^{-2}\ln(2/\delta)$$

samples to get an (ϵ, δ)-ras

Main difficulty

The variance of the X_i is unknown

- In general, the sample variance unreliable
- Consider two models



$\mathbb{E}[X] = \mu$ $\quad\quad$ $\mathbb{E}[Y] = \frac{1+\epsilon}{1-\epsilon}\mu$

- Total variation distance between X and Y is $\Theta(\epsilon\mu/M)$

Telling the difference between X and Y

- Any sample from X will have sample standard deviation of 0
- Any sample from Y will also have sample standard deviation of 0 unless you happen to see a 1
- Because

$$\mathbb{E}[Y](1-\epsilon) > \mathbb{E}[X](1+\epsilon),$$

  have to know whether data comes from X or Y to have at most ϵ relative error.
- Need $\Theta\left(\frac{M\ln(1/\delta)}{\epsilon\mu}\right)$ samples to have at least $1-\delta$ chance of detecting whether data from X or Y

How many samples are needed?

An optimal algorithm for Monte Carlo estimation
P. Dagum, R. Karp, M. Luby, and S. Ross
SIAM J. Comput., *Vol 29, No 5, pp. 1484–1496, 2000*

$$a_1 = \frac{\sigma^2}{\mu^2}, \quad a_2 = \epsilon \frac{M}{\mu}, \quad a_3 = 2\epsilon^{-2} \ln(4/\delta)$$

Theorem (Dagum, Karp, Luby & Ross 2000)

Any (ϵ, δ)*-ras that applies to all* $[0, 1]$ *random variables requires at least (to first order)*

$$(1/32) \max\{a_1, a_2\} a_3$$

samples.

DKLR

An optimal algorithm for Monte Carlo estimation
P. Dagum, R. Karp, M. Luby, and S. Ross
SIAM J. Comput., *Vol 29, No 5, pp. 1484–1496, 2000*

$$a_1 = \frac{\sigma^2}{\mu^2}, \quad a_2 = \epsilon \frac{M}{\mu}, \quad a_3 = 2\epsilon^{-2}\ln(4/\delta)$$

Theorem (Dagum, Karp, Luby & Ross 2000)

There exists an (ϵ, δ)-ras that applies to all $[0,1]$ random variables that uses (to first order)

$$[2.87 \max\{a_1, a_2\} + 5.74 a_2] a_3$$

samples.

New algorithm

$$a_1 = \frac{\sigma^2}{\mu^2}, \quad a_2 = \epsilon \frac{M}{\mu}, \quad a_3 = 2\epsilon^{-2} \ln(4/\delta)$$

Theorem (H. & Jones 2017)

There exists an (ϵ, δ)*-ras that applies to all* $[0, 1]$ *random variables that uses (to first order)*

$$[a_1 + (3/2)a_2 + \sqrt{a_1 a_2 + a_2^2}]a_3.$$

samples

Note: asymptotic to CLT $a_1 a_3$ running time when $a_2 \to 0$

An application

Importance sampling

Goal of IS is to find

$$I = \int_{\mathbb{R}^n} g(x)\ d\mathbb{R}^n$$

For random variable Y with density f_Y, let

$$W = \frac{g(Y)}{f_Y(Y)}$$

so $\mathbb{E}[W] = I$

How many samples?

- Well known that number of samples needed for IS is

$$\Theta(a_1 a_3),$$

  problem is that $a_1 = \sigma_W^2/\mu_W^2$ difficult to find

- Here a_1 is square of coefficient of variation

- Easier to find $\max[W]$ (optimization easier than integration)

A simple 1 dimensional example

Suppose we wish to know

$$I = \int_{-\infty}^{\infty} \exp(-|x|^{2.5})\ dx$$

Can draw from a Cauchy $f_Y(y) = [\pi(1+y^2)]^{-1}$

$$W = \pi(1+Y^2)\exp(-|Y|^{2.5})$$

Here $\max[W] = 3.297\ldots.$

Running time

For IS example it holds that

$$a_1 = 0.6606, \;\; a_2 = 0.1859,$$

Mean number of samples used

	$(\epsilon, \delta) = (0.1, 10^{-6})$	$(\epsilon, \delta) = (0.01, 10^{-6})$
DKLR	10274	$6.3 \cdot 10^5$
New method	3177	$2.2 \cdot 10^5$
$a_1 a_3$	1918	$1.9 \cdot 10^5$

Sampling from the union of sets

Goal: Given k sets $A_1, \ldots, A_k$, where the the size of each A_i is known, estimate size of $\#(A_i)$.

1. Draw random variable I, where probability that $I = i$ is proportional to size(A_i)
2. Draw $Y \leftarrow$ Unif(A_I)
3. Let $W = 1/\#\{i : Y \in A_i\}$

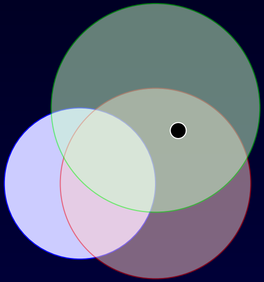
Then $W \in [0, 1]$ satisfies

$$\mathbb{E}[W] = \text{size}(\cup A_i) / \sum_i \text{size}(A_i)$$

$$\text{size}(\cup A_i) = \mathbb{E}[W] \sum_i \text{size}(A_i)$$

Toy example: Circles

Three circles of size 1.2, 1.9 and 2.3



$I = 3$

$W = 1/2$

$$\mathbb{P}(I=1) = \frac{1.2}{C},\ \mathbb{P}(I=2) = \frac{1.9}{C},\ \mathbb{P}(I=3) = \frac{2.3}{C},\ C = 1.2{+}1.9{+}2.3$$

Draw Y uniformly from circle I, set $W = 1/\#$ of circles Y is in

$$W \in \{1, 1/2, 1/3\}$$

Toy example: Circles continued

In this case $W \in \{1, 1/2, 1/3\}$, don't know anything more about

$$\mathbb{E}[W],\ \mathbb{SD}[W]$$

Could be anything consistent with $[0, 1]$ random variable!

Three steps to the estimate

How DKLR works

Scale random variables so in $[0,1]$. Then have three step process:

1. Get $(\epsilon^{1/2}, \delta/3)$ estimate $\hat{\mu}_1$ for μ using Zero-One estimator
2. Use $\hat{\mu}_1$ to get $\hat{a}$ that is an upper bound on $\max\{a_1, a_2\}$
3. Use $\hat{a}$ together with a sample average to get final (ϵ, δ) estimate $\hat{\mu}$

How the new method works

Still a three step process

- The goals of the three steps are almost the same
- The techniques that achieve each goal are quite different

1. Get $(\epsilon^{1/3}, \delta/3)$ estimate $\hat{\mu}_1$ for μ using Gamma Bernoulli Approximation Scheme
2. Use $\hat{\mu}_1$ to get $\hat{a}$ that is an upper bound on a_1 using a Poisson based estimator
3. Use $\hat{a}$ together with a light-tailed sample average estimte to get final (ϵ, δ) estimate $\hat{\mu}$

Step 1: Gamma Bernoulli Approximation Scheme

Lemma

Five elementary facts about distributions:

1. *If* $X_1, X_2, \ldots$ *are* $[0, M]$ *r.v.'s and* $U_1, U_2, \ldots$ *are iid Unif*$([0, 1])$, *then* $\mathbb{1}(MU_1 > X_1), \mathbb{1}(MU_2 > X_2), \ldots$ *are iid Bern*(μ/M).
2. *If* $B_1, B_2, \ldots$ *are iid Bern*(μ/M), *then* $G = \min\{t : B_t = 1\} \sim$ *Geo*(μ/M).
3. *For* $G \sim$ *Geo*(μ/M) *and* $[N|G] \sim$ *Gamma*$(G, 1)$, *it holds that* $N \sim$ *Gamma*$(1, \mu/M)$.
4. *If* $N_1, \ldots, N_k$ *are iid Gamma*$(1, \mu/M)$, *then* $N_1 + \cdots + N_k \sim$ *Gamma*$(k, \mu/M)$.
5. *If* $R \sim$ *Gamma*$(k, \mu/M)$, *then* $R\mu/[M(k+2)] \sim$ *Gamma*$(k, k+2)$.

Putting these ideas together

`GBAS` Input: k, Output $\hat{\mu}$ such that $\mu/\hat{\mu} \sim \text{Gamma}(k, k+2)$

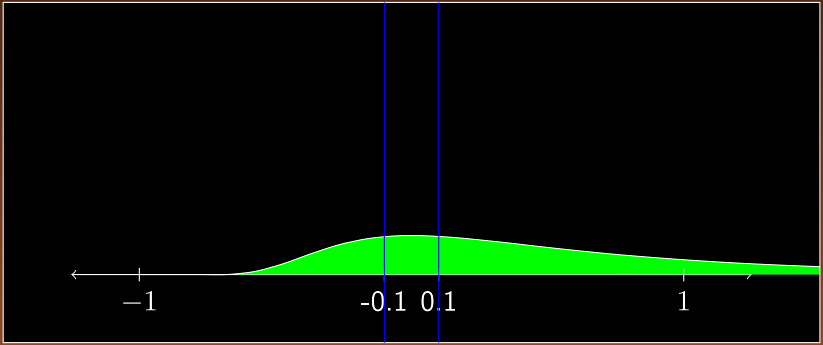
1. $n \leftarrow 0$, $i \leftarrow 0$
2. Repeat
   - 2.1 $i \leftarrow i + 1$, draw X_i from X, draw U_i from $\text{Unif}([0,1])$
   - 2.2 $n \leftarrow n + \mathbb{1}(X_i \leq MU_i)$

   Until $n = k$
3. Draw $R \leftarrow \text{Gamma}(i, 1)$
4. Return$(M(k+2)/R)$

As k increases, Gamma$(k, k+2)$ *concentrates near 1*

User set $k = 5$, $c = 1$

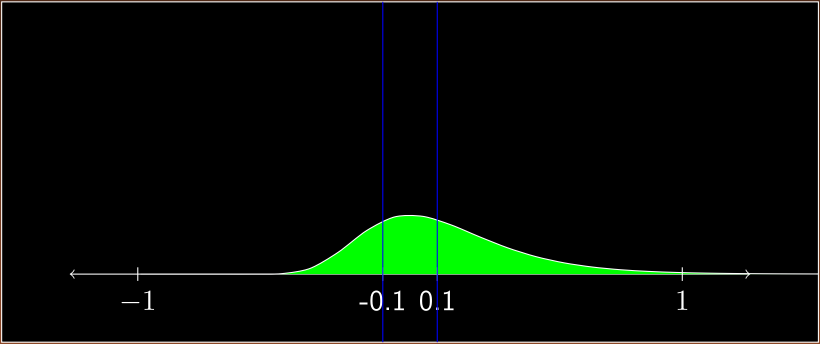
$\mathbb{P}(|\text{rel err}| > 0.1) \approx 92.6\%$



As k increases, Gamma$(k, k+2)$ *concentrates near 1*

User set $k = 20$, $c = 1$

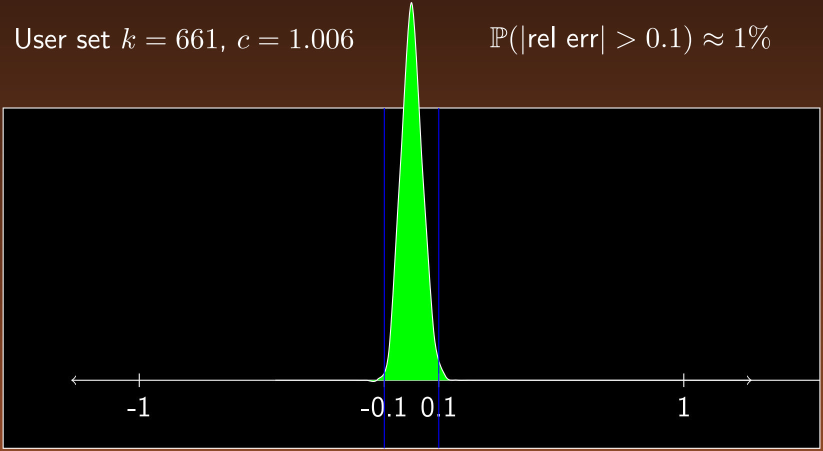
$\mathbb{P}(|\text{rel err}| > 0.1) \approx 66.1\%$



As k increases, Gamma$(k, k+2)$ *concentrates near 1*

User set $k = 661$, $c = 1.006$

$\mathbb{P}(|\text{rel err}| > 0.1) \approx 1\%$



Step 2: Poisson based estimator for σ^2/μ^2

Lemma

Three more elementary facts about distributions:

1. *If* $X_1, X_2, \ldots$ *are* $[0, M]$ *random variables with variance* σ^2 *and* $U_1, U_2, \ldots$ *are iid Unif*$([0,1])$, *then*

$$\mathbb{1}(U_i > 1/2)\mathbb{1}(M^2 U_{i+1} > (X_{i+1} - X_i)^2) \sim \textit{Bern}(\sigma^2).$$

2. *For* $N \sim$ *Pois*(a) *and* $B_1, \ldots, B_N \overset{iid}{\sim}$ *Bern*(σ^2), *then*

$$B_1 + \cdots + B_N \sim \textit{Pois}(a\sigma^2).$$

3. *Let* $c_1 = 2\ln(1/\delta)$. *For* $A \sim$ *Pois*$(a \cdot 2\ln(1/\delta))$,

$$\mathbb{P}(A/c_1 + 1/2 + \sqrt{A/c_1 + 1/4} \leq a) \leq \delta.$$

Using these facts

`PoissonEstimate`

Input: ϵ, c_2, $\hat{\mu}$
Output: c^2 satisfying $\mathbb{P}(\sigma^2/\hat{\mu}^2 \leq c^2) \geq 1 - \exp(-c_2/2)$

1. Draw $N \leftarrow$ Pois$(c_2 M/[\epsilon\hat{\mu}])$
2. Draw $W_1, \ldots, W_N \sim$ Bern(σ^2)
3. Let $A = (W_1 + \cdots + W_N)/c_2$
4. Output $(A + 1/2 + \sqrt{A + 1/4})\epsilon/[M\hat{\mu}]$

Step 3: Light-tailed sample average

- When step 2 a success, we have an upper bound on a_1
- Catoni gave an M-estimator which gave confidence intervals
- Of course, we want an (ϵ, δ)-ras.
- Can convert Catoni to an (ϵ, δ)-ras when σ^2 and μ^2 are each bounded individually
- Here develop a simpler (ϵ, δ)-ras when σ^2/μ^2 bounded

Downweighting samples from from the mean

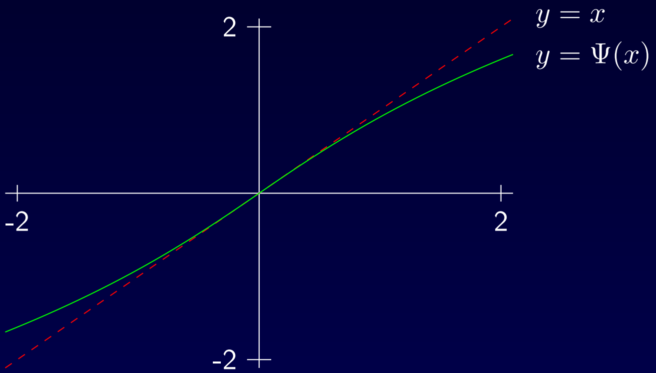
- Idea is to start with initial estimate $\hat{\mu}_1$ of μ
- Downweight samples that are far away from mean
- Given $c^2 > \sigma^2/\mu^2$ and ϵ, far away means

$$\alpha = \frac{\epsilon M}{c^2 \mu}$$

How to get light tails

Start with a function Ψ that is close to $y = x$ for small x, but grows as natural log for large x

$$\Psi(x) = -\ln(1 - x + x^2/2)\mathbb{1}(x \leq 0) + \ln(1 + x + x^2/2)\mathbb{1}(x \geq 0)$$



How to get light tails from Ψ

For X_i, then

$$X_i = \hat{\mu}_1 + \alpha^{-1}(\alpha(X_i - \hat{\mu}_1))$$

So set

$$W_i = \hat{\mu}_1 + \alpha^{-1}\Psi(\alpha(X_i - \hat{\mu}_1))$$

Then W_i always has light tails because of logarithmic growth of Ψ

Step 3: Light-tailed sample average

LTSA

Input: $c^2 > \sigma^2/\mu^2$, ϵ, c_3, initial estimate $\hat{\mu}_1$

Output: Final estimate $\hat{\mu}$

1. Let $n \leftarrow \lceil c^2 \cdot c_3 \rceil$
2. Draw $X_1, \ldots, X_n$
3. Set $\alpha \leftarrow \dfrac{\epsilon M}{c^2 \hat{\mu}_1}$
4. For $i \in 1$ to n,

$$W_i = \hat{\mu}_1 + \alpha^{-1} \Psi(\alpha(X_i - \hat{\mu}_1))$$

5. Output $(W_1 + \cdots + W_n)/n$

Final version

`MainAlgorithm`

Input: ϵ_1, k, c_2, ϵ, c_3

1. $\hat{\mu}_1 \leftarrow \texttt{GBAS}(k)$
2. $c^2 \leftarrow \texttt{PoissonEstimate}(\epsilon_1, c_2\epsilon^2)$
3. $\hat{\mu} \leftarrow \texttt{LTSA}(c^2(1+\epsilon_1)^2, \epsilon, c_3)$

Correctness & expected running time

Theorem

The expected running time of `MainAlgorithm`$(\epsilon_1, k, c_2, \epsilon, c_3)$ *is bounded above by*

$$\frac{kM}{\mu} + c_2 \frac{\epsilon M}{\mu} + 1 + (1+\epsilon_1)^2 c_3 \left[\frac{\sigma^2}{\mu^2} + \frac{\epsilon M}{2\mu} + \sqrt{\frac{\sigma^2}{\mu^2} \cdot \frac{\epsilon M}{\mu} + \frac{\epsilon^2}{4\mu^2}} \right]$$

Theorem

The output $\hat{\mu}$ *of* `MainAlgorithm`$(\epsilon_1, k, c_2, \epsilon, c_3)$ *satisfies*

$$\mathbb{P}\left(\left| \frac{\hat{\mu}}{\mu} - 1 \right| > \epsilon \right) \leq 2 \exp\left(-\frac{(k-1)\epsilon_1^2}{2} \right) + \exp\left(-\frac{c_2 \epsilon^2}{2} \right) + 2 \exp\left(-\frac{c_3 \epsilon^2}{2} \right)$$

One choice of parameters

Theorem

Given

$$\epsilon_1 = \epsilon^{1/3},\ k = \lceil 2\ln(6/\delta)\epsilon_1^{-2}\rceil + 1,\ c_2 = 2\ln(3/\delta), c_3 = 2\epsilon^{-2}\ln(6/\delta),$$

it holds that $\hat{\mu}$ *is an* (ϵ, δ)*-ras.*

Conclusion

New algorithm for estimating $\mu = \mathbb{E}[X]$ when $X \in [0, 1]$

- User specified error tolerance and failure probability
- Asymptotic to CLT as $\epsilon \to 0$
- Does not require prior knowledge of variance
- About 2.5 times as fast as previous approach