

ORIE 4580/5580: Simulation Modeling and Analysis

ORIE 5581: Monte Carlo Simulation

Unit 3: Intro to Monte Carlo Simulation

Sid Banerjee

School of ORIE, Cornell University

expectation and variance of sums of rvs

linearity of expectation

for any rvs X and Y , and any constants $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

note: **no assumptions!** (in particular, does not need independence)

- for general X, Y

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(XY)$$

- when X and Y are independent

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

law of large numbers

let X_1, X_2, \dots be a sequence of independent rvs with $\mathbb{E}[X_i] = \mu$ for all i
then, “almost” always

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \longrightarrow \mu \quad , \quad \text{as } n \rightarrow \infty$$

note: for any finite n , $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is still a random variable

central limit theorem

let X_1, X_2, \dots be a sequence of independent rvs with

$$\mathbb{E}[X_i] = \mu, \text{Var}(X_i) = \sigma^2 < \infty \text{ for all } i$$

then,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \sigma\mathcal{N}(0, 1) = \mathcal{N}(0, \sigma^2) \quad , \quad \text{as } n \rightarrow \infty$$

approximations for large n ,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \stackrel{D}{\approx}$$

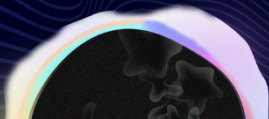
$$S_n = \sum_{i=1}^n X_i \stackrel{D}{\approx}$$

QUESTION OF THE DAY

**How can we tell if we are not already in
a computer simulation?**

 SingularityNET
community.singularitynet.io

#AGICHAT



clicker question: Staffing a Food Bank

a food bank depends on volunteers for its labor pool

on any given day, the number of workers who show up is $Uniform(\{1, 2, \dots, 9\})$,

while the number of donations needed to be collected is $Uniform(\{1, 2, \dots, 29\})$

assuming the work is equally divided among each worker, what is the average load for each worker?

(a) 3

(b) > 3

(c) < 3

(d) can be any of the above on different days

staffing a food bank

a food bank depends on volunteers for its labor pool

on any given day, the number of workers who show up is $Uniform(\{1, 2, \dots, 9\})$,

while the number of donations needed to be collected is $Uniform(\{1, 2, \dots, 29\})$

assuming the work is equally divided among each worker, what is the average load for each worker?

- let X = number of workers, Y = number of donations
- we have $\frac{\mathbb{E}[Y]}{\mathbb{E}[X]} = \frac{30/2}{10/2} = 3$
- on the other hand, $\mathbb{E}[Load] = \mathbb{E}\left[\frac{Y}{X}\right]$; is this also 3?
- let us simulate and check!

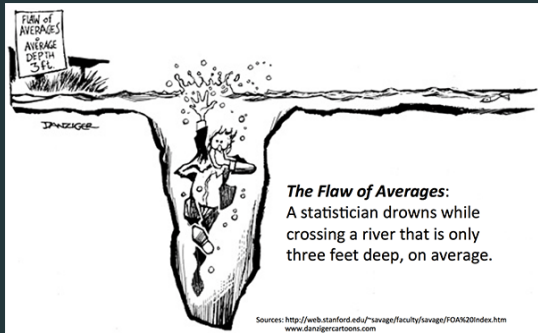
understanding what happened

for random variables X and Y , and function $g(\cdot, \cdot)$,

$$\mathbb{E}[g(X, Y)] \neq g(\mathbb{E}[X], \mathbb{E}[Y])!$$

flaw of averages

moral: in most settings, **average inputs don't give average outputs!**



what can go wrong?

- non-linearities
- correlations between rvs
- 'inspection paradox' (buses take longer to arrive than they should!)
- ...

simulation allows us to avoid such problems!

confidence intervals

how many replications?

simulating food bank for many days (**samples/replications**) = distribution of loads
as number of replications increases, **sample average** \rightarrow **average load** (by LLN)

question: every time we run the simulation model with some fixed number of replications, our estimate of $\mathbb{E}[\text{load}]$ changes.

how 'confident' can we be in our estimate?

- *answer:* use CLT to build a **confidence interval**!

confidence intervals

let X_1, X_2, \dots be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_1] = \mu$ and variance $\text{Var}(X_1) = \sigma^2 < \infty$

want to measure μ from simulations

confidence interval: attempt 1...

an interval $[a, b]$ is called a 95% confidence interval for $\mathbb{E}[X_1]$ if

$$\mathbb{P}[a \leq \mathbb{E}[X_1] \leq b] \geq 0.95$$

Q: what is wrong with this?

confidence intervals: definition

an **random interval** $[A, B]$ (computed from data/experiments) is called a **95% confidence interval** for some (deterministic) quantity μ if

$$\mathbb{P}[A \leq \mu \leq B] \geq 0.95$$

confidence intervals for population mean

X_1, X_2, \dots are i.i.d. rvs with $\mathbb{E}[X_1] = \mu$ and $\text{Var}(X_1) = \sigma^2 < \infty$; $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

- from the central limit theorem:

$$\sqrt{n}(\bar{X}_n - \mu) \stackrel{D}{\approx} \sigma \mathcal{N}(0, 1) = \mathcal{N}(0, \sigma^2)$$

- from the inverse cdf of $\mathcal{N}(0, 1)$, we can compute

$$\mathbb{P}\left[\quad \leq \mathcal{N}(0, 1) \leq \quad \right] \geq 0.95.$$

confidence intervals

want to measure $\mu = \mathbb{E}[X_1]$ from simulations

- from the central limit theorem: $\sqrt{n}(\bar{X}_n - \mu) \stackrel{D}{\approx} \sigma \mathcal{N}(0, 1)$
- from the cdf of $\mathcal{N}(0, 1)$, we have $\mathbb{P}[-1.96 \leq \mathcal{N}(0, 1) \leq 1.96] \geq 0.95$

putting these together, we have:

confidence intervals: problems

- the confidence interval is approximate because
- the confidence interval is 'exact' when
- the confidence interval above requires knowledge of σ^2

confidence intervals: problems

- the confidence interval is approximate because
- the confidence interval is 'exact' when
- the confidence interval above requires knowledge of σ^2

can replace σ^2 with its sample estimator

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

fixed sample-size: recipe for CI

approximate $100(1 - \alpha)\%$ Gaussian CI for $\mathbb{E}X$

1. select a sample size N
2. generate N i.i.d. samples X_1, X_2, \dots, X_N of X
3. compute the estimators \bar{X}_N, s_N^2

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^N X_n, \quad s_N^2 = \frac{1}{N-1} \sum_{n=1}^N (X_n - \bar{X}_N)^2$$

4. look up the value of $z_{\alpha/2}$ such that

$$\mathbb{P}[-z_{\alpha/2} \leq N(0, 1) \leq z_{\alpha/2}] = 1 - \alpha$$

5. the approximate $100(1 - \alpha)\%$ CI for $\mathbb{E}X$ is given by

$$\bar{X}_N \mp z_{\alpha/2} \frac{s_N}{\sqrt{N}}$$

selecting the sample size

how large should N be so that the resulting $100(1 - \alpha)\%$ confidence interval will have a pre-specified width?

- CI $\implies \bar{X}_N \mp z_{\alpha/2} \left(\sigma / \sqrt{N} \right)$
- half-width $\implies z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$
- ℓ be the desired half-width
- Set $N =$

selecting the sample size

- problem: σ^2 is unknown!
- estimating σ^2 through s_N^2 requires simulation!
- solution: 'pilot runs'

perform k simulation runs to get $[X'_n : n = 1, \dots, k]$ as outcomes
compute

$$\tilde{X}_k = \frac{1}{k} \sum_{n=1}^k X'_n, \quad \tilde{S}_k^2 = \frac{1}{k-1} \sum_{n=1}^k (X'_n - \tilde{X}_k)^2$$

use \tilde{S}_k^2 to estimate σ^2

for confidence level α , half-width ℓ , set

$$N = \left\lceil \frac{z_{\alpha/2}^2 \tilde{S}_k^2}{\ell^2} \right\rceil$$

clicker question

we want to estimate $\mathbb{E}[X]$ with ∓ 0.01 accuracy with 95% confidence from 10 replications, $\tilde{s}_{10}^2 = 0.5$

how many replications should we use?

- A 200
- B 2,000
- C 20,000
- D 40,000
- E 200,000

clicker question

we want to estimate $\mathbb{E}[X]$ with ∓ 0.01 accuracy with 95% confidence from 10 replications, $\tilde{s}_{10}^2 = 0.5$

thus we have

$$\left\lceil \frac{z_{\alpha/2}^2 \tilde{s}_k^2}{\ell^2} \right\rceil = \frac{2^2 \times 0.5}{0.01^2} = 20,000$$

how many replications should we use?

- A 200
- B 2,000
- C 20,000
- D 40,000
- E 200,000

basic simulation workflow

- perform **pilot run** of k simulations (sufficient but not large k)
- compute required sample-size N for desired confidence interval
- run N additional simulations \implies **production runs**
- form fixed-sample confidence intervals from these N samples
- **note:** final CI may be different than desired, because it is constructed by using s_N^2 (may be larger/smaller than \tilde{s}_k^2)
- for the final confidence interval, **discard the information from the trial runs** not a problem, since $N \gg k$ usually

clicker question: abusing CI (?)

in a homework assignment in Simulation, students were asked to do a Monte Carlo simulation to compute π up to 2 decimal places
there were 100 homework submissions, and *all* submissions reported 95% CIs which included 3.14
the probability that this happened 'by chance' is approximately

- (a) 0.4
- (b) 0.05
- (c) 0.006
- (d) 0.0007
- (e) 0

confidence intervals as a **social contract**

a random interval $[A, B]$ (computed from data/experiments) is a 95% confidence interval for some unknown μ if **before the experiment is done**

$$\mathbb{P}[A \leq \mu \leq B] \geq 0.95$$

clicker question: (mis)interpreting CI

a 95% confidence interval for the mean annual rainfall in Milford Sound, New Zealand is 26 feet plus or minus 1 foot
therefore, the rainfall next year will be between 25 feet and 27 feet with 95% probability.

(a) true

(b) false

(c) where is New Zealand?

confidence intervals vs quantiles

the CI for the mean is NOT the same as the quantiles of a random variable.

- suppose that X is a rv with probability density function

- we can select q_1 and q_2 so that

$$\mathbb{P}[q_1 \leq X \leq q_2] = 0.95,$$

but $[q_1, q_2]$ is not a 95% confidence interval for $\mathbb{E}X$.

advanced probability tools

tool 1: Jensen's inequality

Q: can we say something about $\mathbb{E}[f(X)]$ vs $f(\mathbb{E}[X])$ (in particular, \geq or \leq) without simulating?

Jensen's inequality

if X is a random variable and f is a **convex function**, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Example – in our food-bank staffing problem, since $f(x) = \frac{1}{x}$, $x > 0$ is convex:

$$\mathbb{E}[Y] \mathbb{E} \left[\frac{1}{X} \right] \geq \frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$$

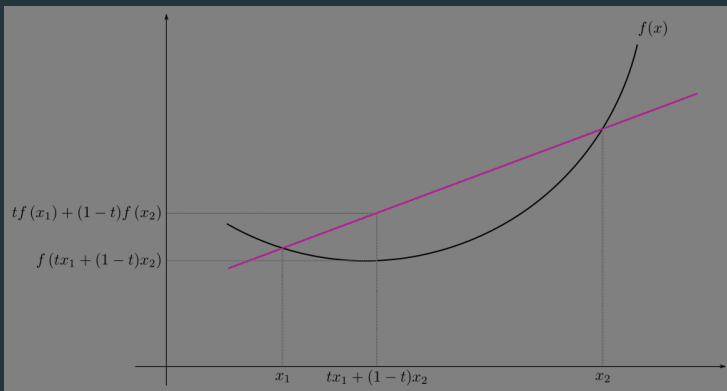
tool 1: Jensen's inequality

Jensen's inequality

if X is a random variable and f is a **convex function**, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

proof sketch (plus way to remember)



tool 2: Chebyshev's Inequality

Q: since the CLT convergence is faster/slower for different rvs, can we be sure that CIs based on variance always make sense?

Chebyshev's inequality

let X be any rv with finite mean μ and finite variance $\sigma^2 > 0$
then for any $k > 0$,

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

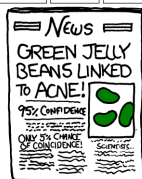
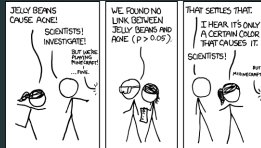
'always-valid' confidence intervals

Chebyshev's inequality

let X be any rv with finite mean μ and finite variance $\sigma^2 > 0$
then for any $k > 0$,

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

worst-case CI: if we choose $k = 2$, then we **always** have 75% confidence intervals



tool 3: the union bound

- Q: how can we get **simultaneous confidence intervals** for multiple hypothesis?
- Eg. I give you five 95% confidence intervals; do they **simultaneously** contain their respective means 95% of the time?

tool 3: the union bound

the union bound

let A_1, A_2, \dots, A_k be events; then

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_k) &= 1 - P(A_1^c \cup A_2^c \cup \dots \cup A_k^c) \\ &\geq 1 - (P(A_1^c) + P(A_2^c) + \dots + P(A_k^c)) \end{aligned}$$

let A_i = event that the i th CI contains its true mean...