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

class poll: 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 ← All CIs were of the form ≈ $[3.13, 3.15]$

the probability that this happened 'by chance' is approximately

$\mathbb{P}[X \in \odot] = \pi/4$

%ages

Conjecture – 'Students are p-hacking'

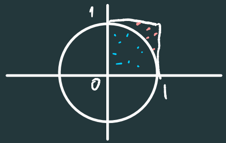
3 (a) 0.4

31 (b) 0.05

35 (c) 0.006 ✓

19 (d) 0.0007

12 (e) 0



$$\Rightarrow \frac{4}{n}\sum_{i=1}^{n} X_i \approx \pi$$

↑ formalized via CI

expectation and variance of sums of rvs

linearity of expectation

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

no assumptions

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

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

- for general X, Y

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(XY)$$

less imp

- when X and Y are independent

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

IMP

law of large numbers

let $X_1, X_2, \ldots$ be a sequence of independent rvs with $\mathbb{E}[X_i] = \mu$ for all i

then, "almost" always

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

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

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mathbb{E}[X_i]$$

by lin of expⁿ

central limit theorem

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

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

then,

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

ie, if $\bar{Z}_n = \frac{\sum (X_i - \mu)}{\sqrt{n}}$, then $\mathbb{E}[\bar{Z}_n] = 0$, $Var(\bar{Z}_n) = \sigma^2$

<u>approximations for large</u> n, (usually, people think sums of rvs are Gaussian)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \overset{D}{\approx} N\left(\mu, \sigma^2/n\right)$$

$$S_n = \sum_{i=1}^{n} X_i \overset{D}{\approx} N\left(n\mu, n\sigma^2\right)$$

Basic simulation workflow

diff for each problem

- If we want to compute μ, code a sim model (ie fn) s.t. output is a rv X with $\mathbb{E}[X] = \mu$ ← what we want to compute

- Run $\boxed{n}$ different independent simulations (replicates)
- Output $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ AND confidence intervals

confidence intervals $[A, B]$ where

A, B are some fn of $\{X_i\}_{i=1}^{n}$

more generally - 'pilot run' to estimate $\mathrm{Var}(X_i)$

- Determine choice of n

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 $\overbrace{\textit{Uniform}(\{1, 2, \ldots, 9\})}^{X}$,

while the number of donations needed to be collected is $\underbrace{\textit{Uniform}(\{1, 2, \ldots, 29\})}_{Y}$

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

i.e, Number of donations they need to pick up

$$\mathbb{E}[X] = \frac{1}{9}\sum_{i=1}^{9} i = \frac{\cancel{9}.10}{\cancel{9}.2}$$

$$\mathbb{E}[Y] = \frac{1}{29}\sum_{i=1}^{29} i = \frac{30}{2}$$

- 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!

'Easy way to see size of real load'

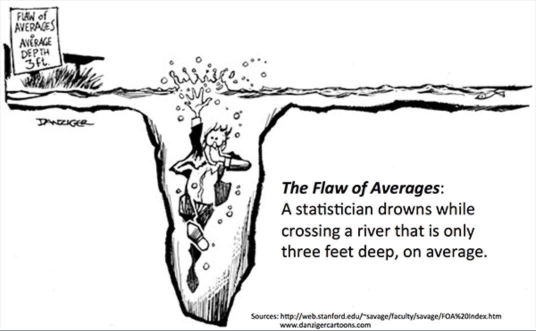
$\mathbb{E}[Y/X]$ must be $>$ $\mathbb{E}[Y]/\mathbb{E}[X]$

flaw of averages

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

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

$$\mathbb{E}[X/Y] \neq \mathbb{E}[X] \cdot \mathbb{E}[1/Y]$$ unless $X \perp\!\!\!\perp Y$



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!

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 $\mathbb{E}[\textit{load}]$ changes.
how 'confident' can we be in our estimate?

- *answer*: use CLT to build a confidence interval!
  ie - want to run enough sims s.t. estimate $\bar{X}_n = \frac{1}{n}\sum X_i$
  'looks like a Gaussian' (ie, $\bar{X}_n \approx N(\mu, \sigma^2/n)$)

confidence intervals

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

want to measure μ from simulations

confidence interval: attempt 1...

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

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

all numbers

what is wrong with this?

The probability of $\mathbb{E}[X] \in$ a FIXED interval is either 0 or 1

· Attempt to fix: $\mathbb{P}[a \leq \bar{X}_n \leq b] \geq 0.95$

confidence intervals: definition

(ideally, smallest)

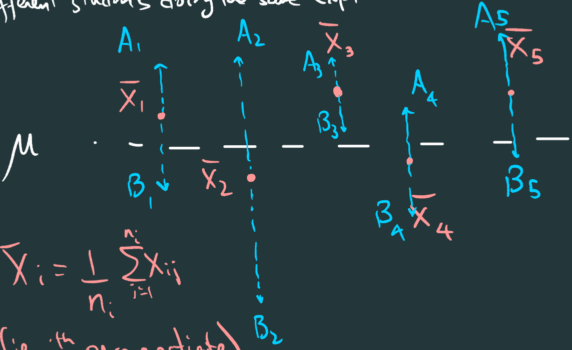
an random interval $[A, B]$ (computed from data/experiments) is called a 95% confidence interval for some (deterministic) quantity μ if

ARE RANDOM

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

IS NOT $\frac{1}{n}\sum x_i$, it is the 'truth'

different students doing the same expt



$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

(ie, i^{th} person's estimate)

We will say $[A,B]$ is a valid $1-\alpha$ CI

if $\mathbb{P}[A \leq \mu \leq B] \geq 1-\alpha$

$\mathbb{E}[X]$

confidence intervals as a social contract

FREQUENTIST CI

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

$$\mathbb{P}[A \le \mu \le B] \ge 0.95$$

Ideal world - Analyst $\boxed{\text{Code Data}} \longrightarrow$ Client runs code ONCE, and sees CIs

p-hacking - run code multiple times until you 'think CIs are correct'

alternate : BAYESIAN Interval.