Monte Carlo: what, why & how

Art B. Owen

Stanford University

Adapted from “Monte Carlo theory, methods and examples”

`http://statweb.stanford.edu/~owen/mc/`

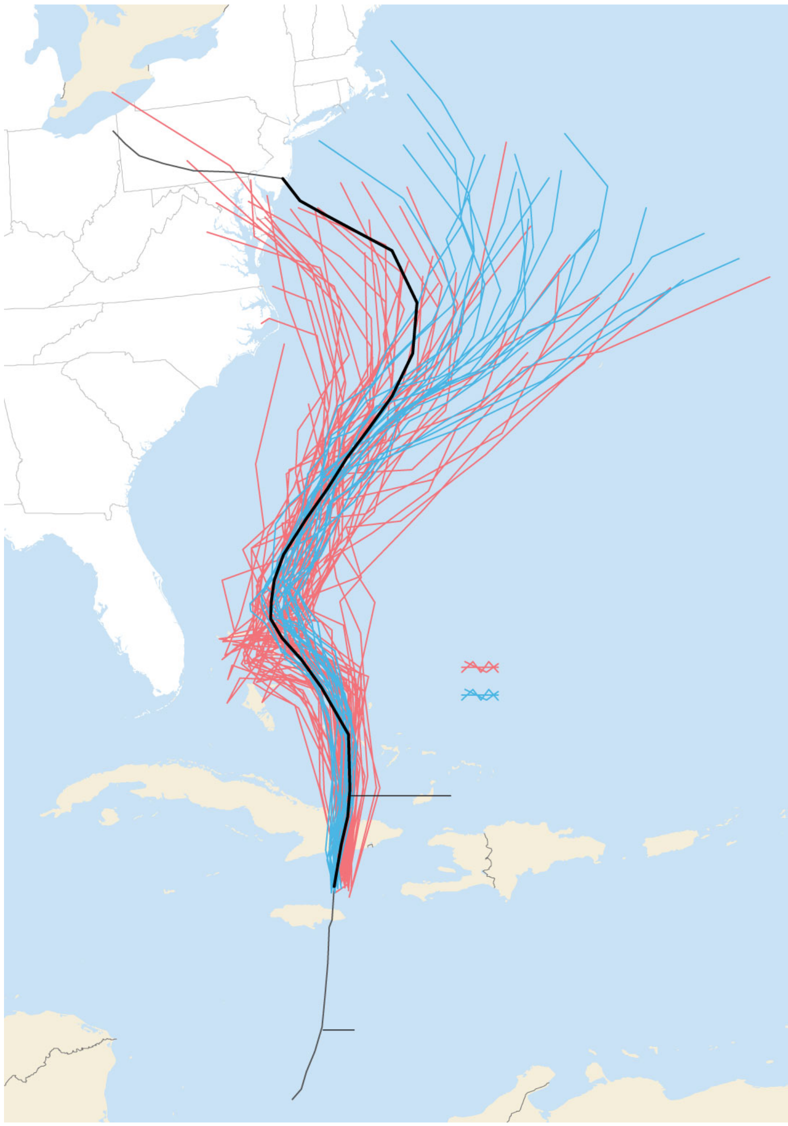
Monte Carlo sampling

1) Take a complicated chance based system.
2) Simulate the outcome multiple times on the computer.
3) Keep track of what happens.

If the system is not chance based

1) Express it as chance based.
2) Go through steps above.

Hurricane Sandy



Simulated trajectories

Oct 25–30, 2012

Black = actual

Red = European sims

Blue = US sims

Source: Wall Street Journal,

Jo Craven McGinty

TIGGE Tropical Cyclone Data

Similar problems

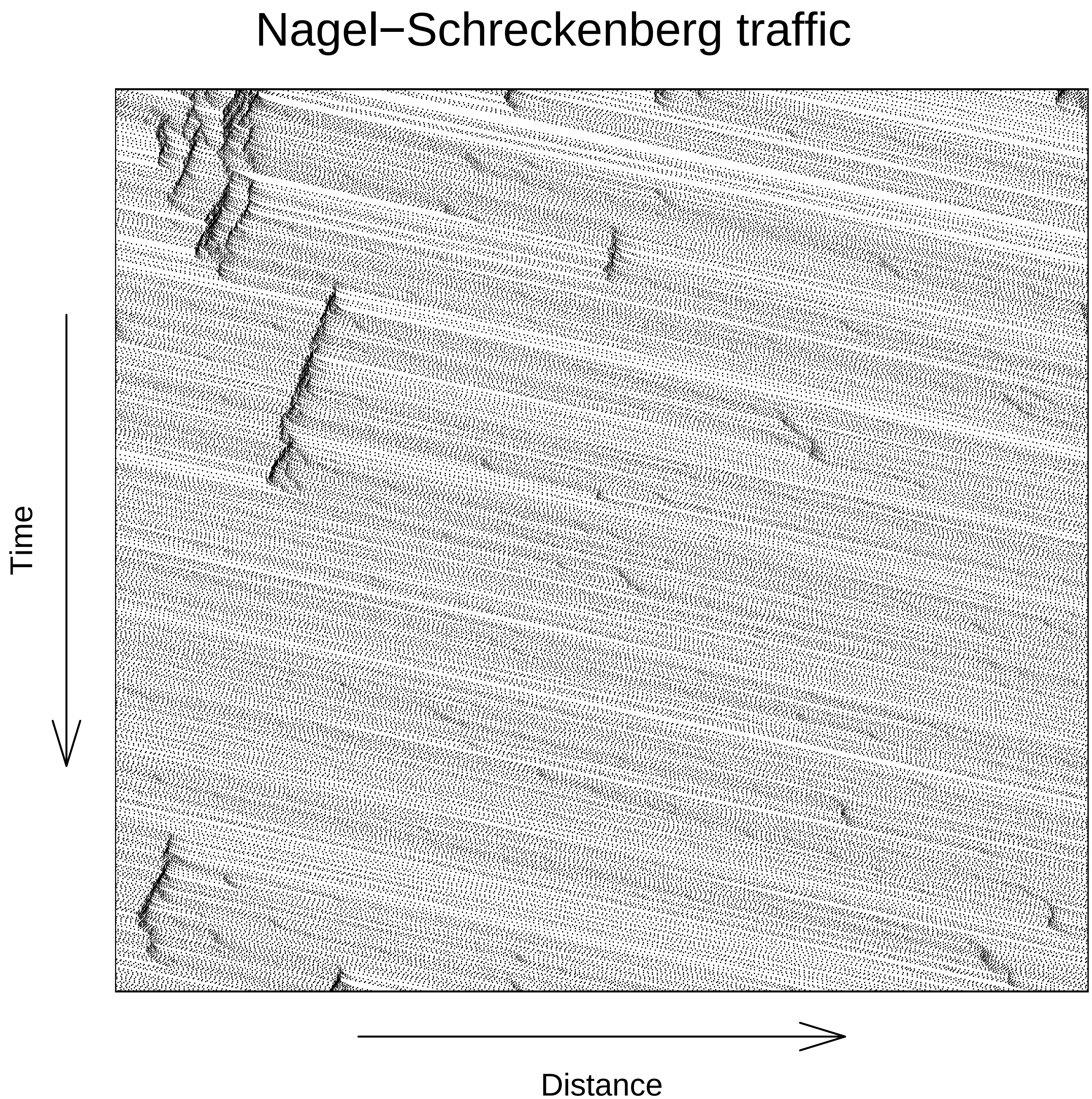
- How a flu epidemic spreads.
- How a stock market evolves.
- How a protein folds.
- How long to wait for your espresso.
- How a scene will look when illuminated.
- How traffic jams appear.

Nagel-Schreckenberg traffic

- N cars on a circular track (roundabout)
- Track has M car sized positions
- Each car has a velocity (# positions per turn)

Rules for the cars

1) Increase your velocity by 1 at each turn
2) but don't go over the speed limit, e.g., 5 spaces
3) also don't plan to hit car in front
4) also reduce speed by 1 with probability p
5) but no negative velocity



100 cars. 1000 slots. $p = 1/3$. Jams emerge.

Traffic continued

A real traffic model can use:

- whole city road map
- multiple lanes
- data on sources and destinations
- data on mixes of vehicle types
- time of day
- proposed alternative roadways and rules

Nagel-Schreckenberg

Interesting patterns as number of cars raised and lowered.

Total distance traveled increases.

Then decreases.

Numerical integration

We want

$$\mu = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Plain / crude Monte Carlo

Sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim \mathbf{U}([0,1]^d)$ and use

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$$

What we did there

We made a problem chance based that was not.

We wanted $\mathbb{E}(f(\boldsymbol{x}))$ and used a sample average.

That turns a numerical problem into a statistical one. (Hooray)

Integrals / expectations do a lot

Suppose $\boldsymbol{x} \in \Omega$ with $\boldsymbol{x} \sim p$ and $f(\boldsymbol{x}) \in \mathbb{R}$

1) Expected value of $f(\boldsymbol{x})$

$$\mu = \int_\Omega f(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

2) Probability of event $\boldsymbol{x} \in A$

$$\int_\Omega 1\{\boldsymbol{x} \in A\} p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

3) Variance of $f(\boldsymbol{x})$

$$\int_\Omega (f(\boldsymbol{x}) - \mu)^2 p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

4) 90'th percentile of $f(\boldsymbol{x})$

$$\int_\Omega 1\{f(\boldsymbol{x}) \leqslant Q^{0.9}\} p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 0.9 \quad \text{solve for } Q^{0.9}$$

Why use Monte Carlo?

Consider $\mu = \int_{[0,1]} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ first, i.e., $d = 1$.
Maybe f is smooth.

Midpoint rule

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f\Bigl(\frac{i-1/2}{n}\Bigr), \quad |\hat{\mu}-\mu| \leqslant \frac{1}{24n^2}\max_{0\leqslant x\leqslant 1}|f''(x)|$$

Simpson's rule

$$\hat{\mu} = \frac{1}{3n}\Bigl(f(0)+4f\Bigl(\frac{1}{n}\Bigr)+2f\Bigl(\frac{2}{n}\Bigr)+4f\Bigl(\frac{3}{n}\Bigr)+\cdots+2f\Bigl(\frac{n-2}{n}\Bigr)+4f\Bigl(\frac{n-1}{n}\Bigr)+f(1)\Bigr)$$

$$|\hat{\mu}-\mu| \leqslant \frac{1}{180n^4}\max_{0\leqslant x\leqslant 1}|f''''(x)|$$

Monte Carlo

$$\sqrt{\mathbb{E}((\hat{\mu}-\mu)^2)} = \frac{\sigma}{\sqrt{n}}, \quad \sigma^2 = \int_0^1 (f(x)-\mu)^2\,\mathrm{d}\boldsymbol{x}$$

Do we use Monte Carlo because 'worse is better'?

For $d \geqslant 1$

$$\mu = \int_0^1 \int_0^1 f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2$$

$$\hat{\mu} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} f\Bigl(\frac{i-1/2}{m}, \frac{j-1/2}{m}\Bigr), \quad n = m^2$$

$$|\hat{\mu} - \mu| = O(m^{-2}) = O(n^{-1})$$

Dimension d and r derivatives

$$|\hat{\mu} - \mu| = O(n^{-r/d}) \qquad \text{Fubini}$$

$$\hat{\mu} - \mu = O_p(n^{-1/2}) \qquad \text{Monte Carlo}$$

Monte Carlo wins for high dimension and / or low smoothness.

A second benefit: we get confidence intervals.

$\hat{\mu} - \mu = O_p(n^{-\alpha})$ will mean $\mathbb{E}((\hat{\mu} - \mu)^2) = O(n^{-2\alpha})$.

What to do?

Use the first of these that suffices:

1) Closed form, e.g., $\int_0^1 x^7 \, \mathrm{d}x = 1/8$

2) Symbolic math: Mathematica, Sage, Maple etc.

3) Quadrature

4) Monte Carlo

So Monte Carlo is a last resort.

Looking ahead

To do Monte Carlo, we need these steps:

1) Get $u \sim \mathbf{U}[0,1]$

2) Get nonunifrom $x \sim p$, $x \in \mathbb{R}$

3) Get random vector $\boldsymbol{x} \sim p$, $\boldsymbol{x} \in \mathbb{R}^d$

4) For $\boldsymbol{x} \in \mathbb{R}^\infty$ (a process) getting some of $\boldsymbol{x} \sim p$

Markov chain Monte Carlo

Sometimes we simply cannot generate $\boldsymbol{x} \overset{\mathrm{iid}}{\sim} p$

Especially in physics and in Bayesian computation

So we can't do (plain) Monte Carlo.

Then we might sample a Markov chain with $\boldsymbol{x}_i \overset{\mathrm{d}}{\to} p$

That is MCMC.

Similarly sequential MC (SMC) is used on hard problems.

Now what to do?

Use the first of these that suffices:

1) Closed form
2) Symbolic math
3) Quadrature
4) Monte Carlo
5) MCMC or SMC
6) Approximate MCMC

MCMC and SMC replace MC as the last resort.

But sometimes we can't do MCMC, hence approximate MCMC

Puzzler:

what will we invent after approximate MCMC?

Quasi-Monte Carlo

Sometimes we can improve on MC by sampling more strategically.

This is quasi-Monte Carlo (QMC) and randomized QMC (RQMC)

Plain QMC can attain errors $O(n^{-1+\epsilon})$

RQMC can attain errors $O_p(n^{-3/2+\epsilon})$

Effect of dimension d

Hidden in ϵ

And also when the rate 'sets in.'

Perfect simulation

Sometimes we can get an MC method with exactly known or bounded error characteristics.

E.g,. perfect sampling. Mark Huber's talks.

Now what to do?

Use the first of these that suffices:

1) Closed form
2) Symbolic math
3) Quadrature
4) QMC or RQMC (A.O. talks)
5) Monte Carlo
6) Perfect sampling (Mark Huber's talks)
7) MCMC Jeffrey Rosenthal's talks or SMC (Nicolas Chopin's talks)
8) Approximate MCMC

Multilevel MC

It is for random processes. It cross-cuts MC, QMC, RQMC, MCMC.

(Michael Giles' talk)

Example

Find the average distance between points $\boldsymbol{x}, \boldsymbol{z}$ in the rectangle $[0, a] \times [0, b]$.

Monte Carlo

$$\frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_{i1} - z_{i1})^2 + (x_{i1} - z_{i1})^2} \quad \boldsymbol{x}_i, \boldsymbol{z}_i \sim \mathbf{U}\big([0, a] \times [0, b]\big)$$

Exact

$$G(a, b) = \frac{1}{15} \left[\frac{a^3}{b^2} + \frac{b^3}{a^2} + \sqrt{a^2 + b^2} \left(3 - \frac{a^2}{b^2} - \frac{b^2}{a^2} \right) \right] + \frac{1}{6} \left[\frac{b^2}{a} \operatorname{arccosh} \left(\frac{\sqrt{a^2 + b^2}}{b} \right) + \frac{a^2}{b} \operatorname{arccosh} \left(\frac{\sqrt{a^2 + b^2}}{a} \right) \right],$$

where $\operatorname{arccosh}(t) = \log\big(t + \sqrt{t^2 - 1}\,\big)$

Ghosh (1951)

Example ctd

For $a = 1$ and $b = 3/5$ Monte Carlo with $n = 10{,}000$ gives

$$\widehat{\mathbb{E}}(\|\boldsymbol{x} - \boldsymbol{z}\|) \doteq 0.4227.$$

Exact formula

$$\mathbb{E}(\|\boldsymbol{x} - \boldsymbol{z}\|) \doteq 0.4239.$$

Relative error

$$\frac{|\hat{\mu} - \mu|}{\mu} \doteq 0.0027$$

Discussion

Monte Carlo was easier than calculus.

Or finding the answer in the literature.

Or implementing it once found.

Exact is better because it is more accurate.

But Monte Carlo easily adapts to changes:

rounded corners, complex geometries

more general distances, e.g., roadways

Error estimation

Get $Y_1,\dots,Y_n$ IID, $\mathbb{E}(Y_i)=\mu$ and $\mathrm{Var}(Y_i)=\sigma^2<\infty$

For $\quad \hat\mu = \frac{1}{n}\sum_{i=1}^n Y_i, \qquad \sqrt{n}(\hat\mu-\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma^2)$ as $n\to\infty$

99% confidence interval

$$\hat\mu \pm \frac{2.58s}{\sqrt{n}} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i-\hat\mu)^2$$

Under favourable but mild conditions (P. Hall)

$$\mathbb{P}\Bigl(|\hat\mu-\mu| \leqslant \frac{2.58s}{\sqrt{n}}\Bigr) = 0.99 + O\Bigl(\frac{1}{n}\Bigr)$$

Corner cases

- n so large s^2 is hard to compute accurately
- n so small that $n-1$ vs n matters, (each Y_i very expensive)
- $\mathbb{E}(Y^2)=\infty$ so no CLT

Random number generators

We start with a pseudo-random number generator.

This simulates $u_1, u_2, u_3, \cdots \overset{\text{iid}}{\sim} \mathbf{U}(0, 1)$

They aren't really uniform random, but good ones are close enough.

Two main rules

1) Pick a good one

2) Set the seed (so you can reproduce your results)

Two good pRNGs

Mersenne Twister Matsumoto & Nishimura (1998)

RngStreams L'Ecuyer, Simard, Chen, Kelton (2002)

Seeds

We do $u \leftarrow \mathrm{rand}()$ (or similar)

Inside $\mathrm{rand}()$

$$\mathrm{state} \leftarrow \mathrm{update}(\mathrm{state})$$

$$\mathrm{return}\ f(\mathrm{state})$$

Finite state space $\implies$ It will repeat. E.g., twister has period

$$P = 2^{19937} - 1 > 10^{6000}$$

Setting a seed lets you control the state.

$\mathrm{setseed}(s)$ does $\mathrm{state} \leftarrow g(s)$

Good for **debugging**, reproducibility and synchronization.

Synchronization

given seed s, parameters $\theta_1, \dots, \theta_J$

for $j = 1, \dots, J$

  $\texttt{setseed}(s)$

  $Y_j \leftarrow \texttt{dosim}(\theta_j)$

end for

return $Y_1, \dots, Y_J$

Now every θ_j gets the exact same stream of u_i.

Differences in Y_j are then due to θ_j.

Streams

Split a RNG into smaller independent ones. RngStreams does that.

Simulate N time series to T steps.

Use N streams.

Later run them all out to $2 \times T$ steps.

Source of pseudo-random number generators

These come from abstract algebra / theory of finite fields.

We owe a huge debt to those mathematicians.

Without them:

no QMC and almost no MC or MCMC

physical sampling is very cumbersome

Their methods (now) just work very reliably.

Compare to

Floating point $\doteq \mathbb{R}$.

Roundoff error requires constant care.

Non-uniform random variables

If your distribution has a name

Normal, exponential, binomial, Poisson, etc.

then it is probably already in

R, Python, Matlab, Julia, Mathematica, etc.

We will look briefly because

- Sometimes a new distribution comes up
- The same ideas get used later
- It's fun

Key concepts

Principled approaches

- inversion
- other transformations
- acceptance-rejection
- mixtures

There are also tricks that perhaps don't generalize but give near ideal solutions when they work.

Inversion

Cumulative distribution function:

$F(x_0) = \mathbb{P}(X \leqslant x_0)$ for $x_0 \in \mathbb{R}$

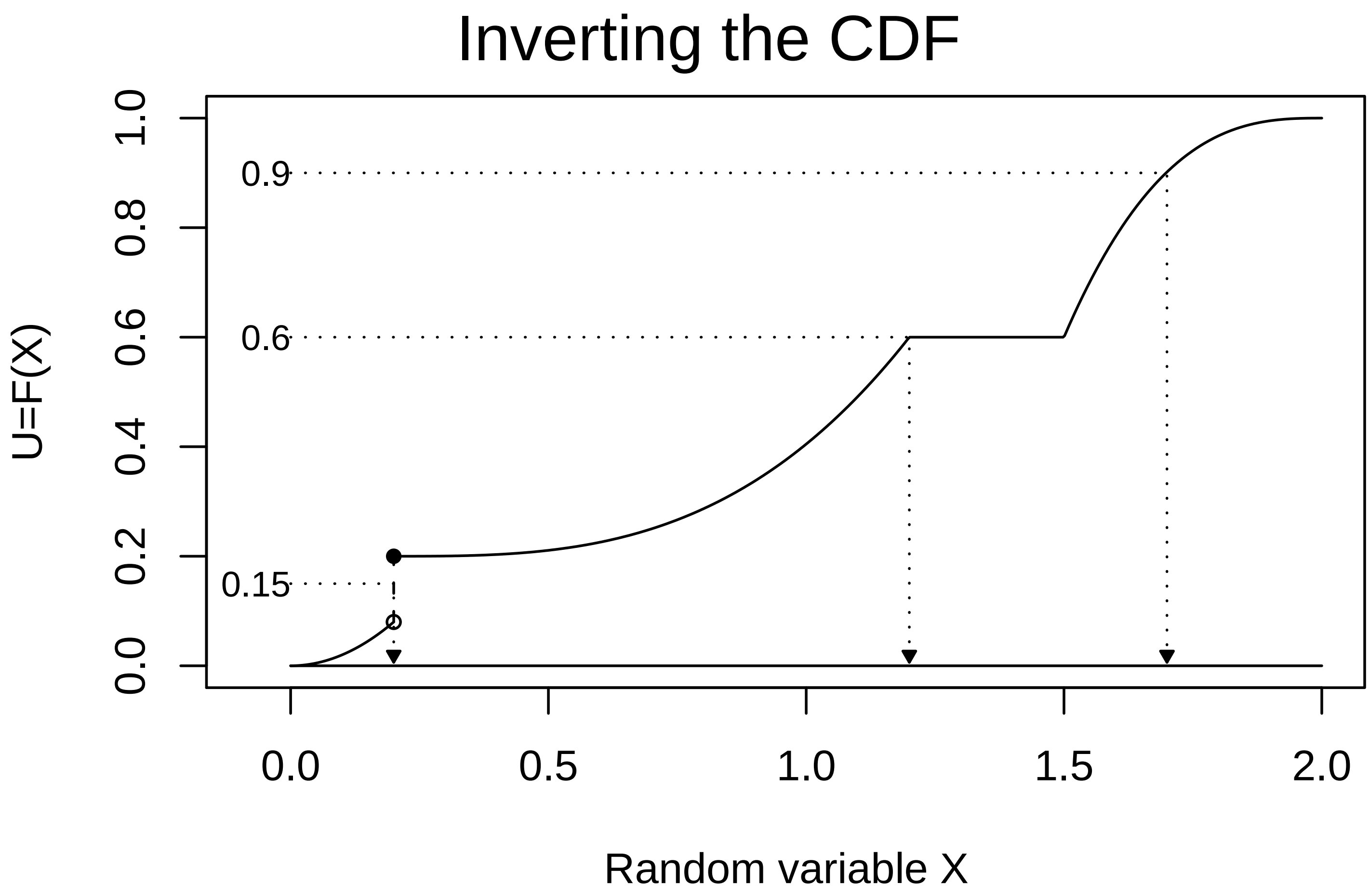
If $F(\cdot)$ is invertible let $X = F^{-1}(U)$ for $U \sim \mathbf{U}(0,1)$.

$$\mathbb{P}(X \leqslant x_0) = \mathbb{P}(F^{-1}(U) \leqslant x_0) = \mathbb{P}(U \leqslant F(x_0)) = F(x_0)$$

For **any** CDF F

$$F^{-1}(u) \equiv \min\{x \in \mathbb{R} \mid F(x) \geqslant u\}.$$

Inversion ctd



Some basic examples

Gaussian

$$Z = \Phi^{-1}(U) \sim \mathcal{N}(0, 1)$$

$$X = \mu + \sigma\Phi^{-1}(U) \sim \mathcal{N}(\mu, \sigma^2)$$

Exponential

$$E = -\log(1 - U) \sim \mathrm{Exp}(1)$$

Or $-\log(U)$, ‘complementary inversion’.

Bernoulli

$$X = 1\{U \geqslant 1 - p\}$$

Or $1\{U \leqslant p\}$

Sharp uses of inversion

It supports various 'fine-motor' style simulation idioms

Sample $X \sim F$ given $a < X \leqslant b$

$$F^{-1}\big(F(a) + U \times (F(b) - F(a))\big), \qquad U \sim \mathbf{U}(0,1)$$

Midpoint rule

$$X_i = F^{-1}\left(\frac{i - 1/2}{n}\right), \quad i = 1, \ldots, n$$

Stratification

$$X_i = F^{-1}\left(\frac{i - U_i}{n}\right), \quad i = 1, \ldots, n$$

Compare distributions F, G, H

$$X = F^{-1}(U) \quad \text{or} \quad G^{-1}(U) \quad \text{or} \quad H^{-1}(U)$$

Transformations

There is a large store-house of transformation rules from probability theory.

See Devroye (1986) (book online).

They connect distributions to each other, e.g.,

- $\min(U_1, U_2) \stackrel{\mathrm{d}}{=}$ Triangle
- $Z_1/Z_2 \stackrel{\mathrm{d}}{=}$ Cauchy
- $\mathrm{Gam}(\alpha) + \mathrm{Gam}(\beta) \stackrel{\mathrm{d}}{=} \mathrm{Gam}(\alpha + \beta)$ (Gamma distn)

Reversals

Sometimes a transformation derived one way can be used in the other direction.

(Examples coming)

Acceptance-rejection

The idea:

we generate $x \sim g(x)$ then

accept them at random with probability $A(x)$.

The result will have density proportional to $g(x) \times A(x)$.

Selection bias

If we **have** g but **want** f,

we take $A(x) \propto f(x)/g(x)$.

Oops

We need $0 \leqslant A(x) \leqslant 1$. Better watch that.

Acceptance-rejection

- we **can sample** $X \sim g$,
- we **know a** $c < \infty$ with $f(x) \leqslant cg(x)$ (always),
- we **can compute** $f(x)/g(x)$.

for density functions f and g.

The algorithm

given c with $f(x) \leqslant cg(x), \forall x \in \mathbb{R}$
repeat
  $Y \sim g$
  $U \sim \mathbf{U}(0,1)$
until $U \leqslant f(Y)/(cg(Y))$
$X \leftarrow Y$
deliver X

The acceptance probability

$$f(x) \leqslant cg(x) \implies A(x) \leqslant 1$$

Proof

This proof is based on Knuth

Probability Y is accepted as X

$$\int_{-\infty}^{\infty} g(y)A(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} g(y)\frac{f(y)}{cg(y)}\,\mathrm{d}y = \frac{1}{c}$$

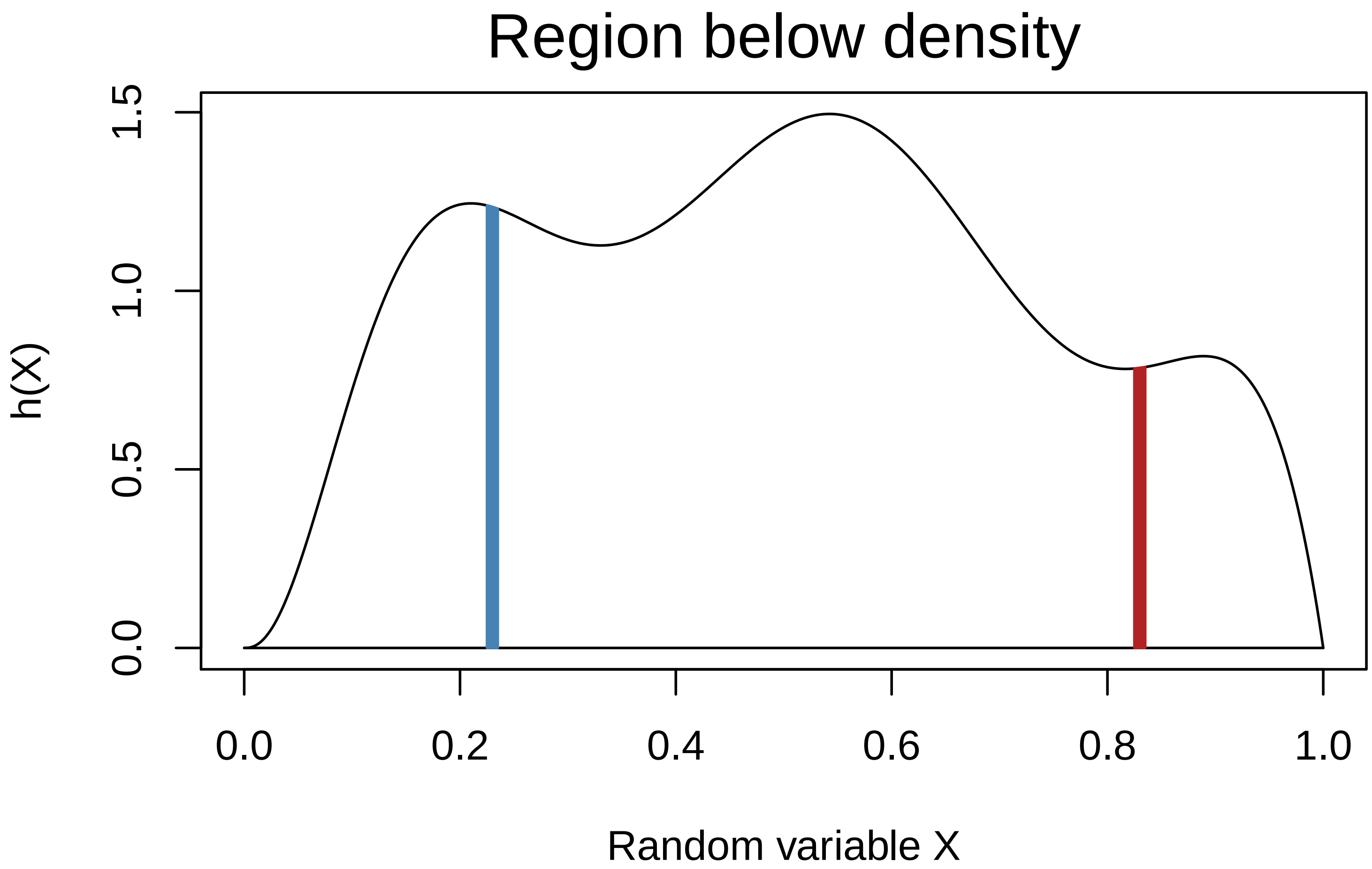
CDF of X

$$\begin{aligned}
\mathbb{P}(X \leqslant x) &= \int_{(-\infty,x]} g(y)A(y)\,\mathrm{d}y + \left(1-\frac{1}{c}\right)\mathbb{P}(X \leqslant x) \\
&= \frac{1}{c}\int_{(-\infty,x]} f(y)\,\mathrm{d}y + \left(1-\frac{1}{c}\right)\mathbb{P}(X \leqslant x) \\
\therefore\quad \mathbb{P}(X \leqslant x) &= \int_{(-\infty,x]} f(y)\,\mathrm{d}y \quad \square
\end{aligned}$$

Geometry of accept / reject

$$M_h = \{(x, y) \mid -\infty < x < \infty,\ 0 \leqslant y \leqslant h(x)\}$$

If $(x, y) \sim \mathbf{U}(M_h)$ then $x \sim h$

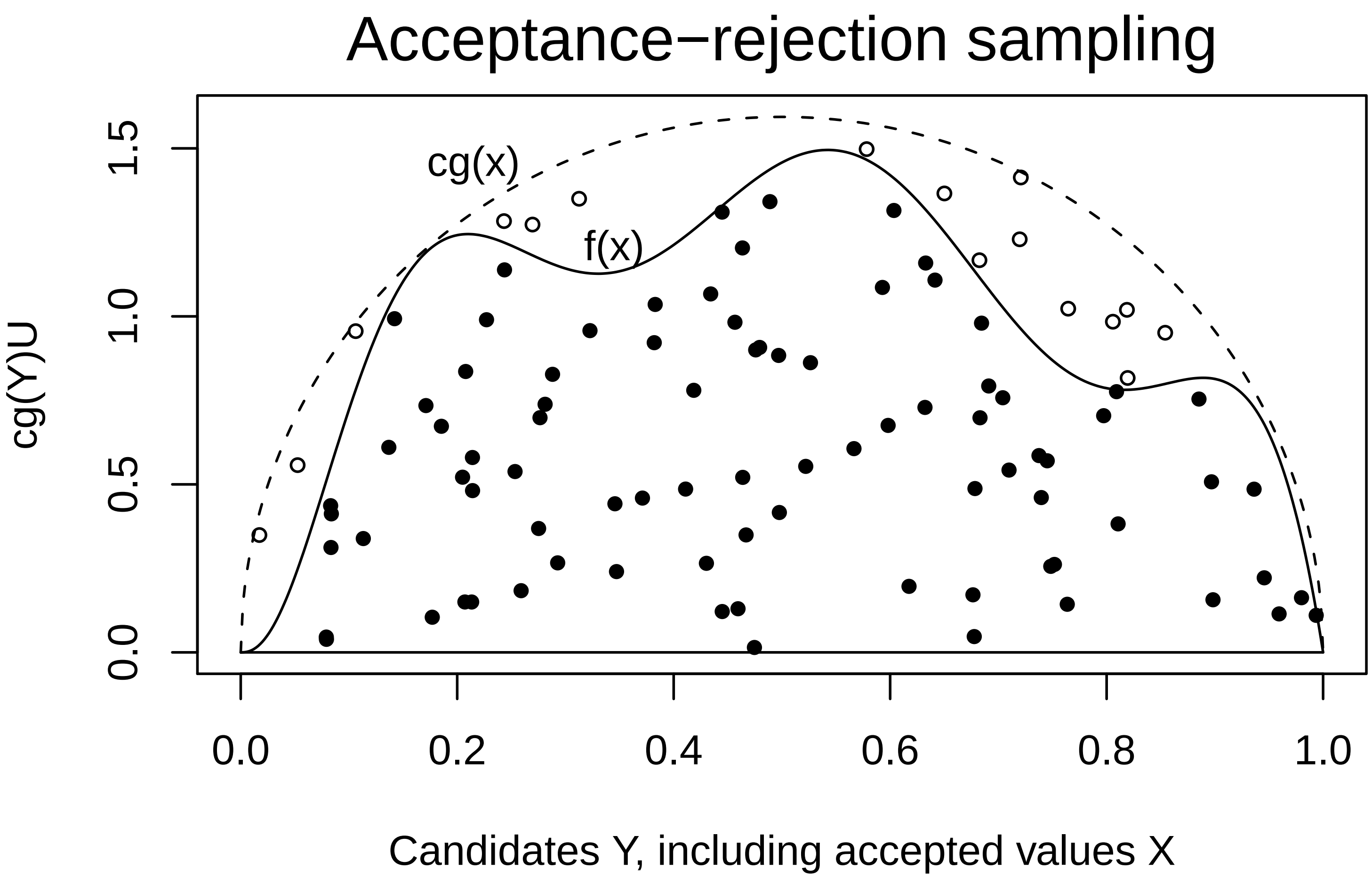


$\mathbb{P}($Blue area$) \doteq h(0.23)\,\mathrm{d}x$

$\mathbb{P}($Red area$) \doteq h(0.83)\,\mathrm{d}x$

Geometry of accept / reject

$$Y \sim g \implies \text{points under } cg \implies \text{points under } f \implies X \sim f$$



Unnormalized f or g

$$g(x) = g_u(x)/c_g \qquad c_g = \int_{-\infty}^{\infty} g_u(x)\,\mathrm{d}x \quad \text{(unknown)}$$

$$f(x) = f_u(x)/c_f \qquad c_f = \int_{-\infty}^{\infty} g_u(x)\,\mathrm{d}x \quad \text{(unknown)}$$

Todo list

1) Sample from g
2) Find $c < \infty$ with $f_u \leqslant c \times g_u$
3) Compute ratio f_u/g_u

Gamma distribution

Important in Bayes and frequentist statistics.

Includes exponential and χ^2.

Usual samplers are acceptance rejection.

Standard Gamma, $\mathrm{Gam}(\alpha)$, shape α

$$f(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}, \quad 0 < x < \infty$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x \qquad \text{(Gamma function)}$$

With rate $\rho > 0$ or scale $\sigma > 0$

$$X = \mathrm{Gam}(\alpha)/\rho \qquad \text{or} \qquad X = \mathrm{Gam}(\alpha) \times \sigma$$

Prize

No good 'closed form' or 'one-line' transformation known.

Despite Devroye offering a prize of Belgian beer for it.

Inverting the CDF doesn't count!

Acceptance-rejection

- Efficiency proportional to acceptance probability $1/c$
- Professional code uses lots of tricks to avoid computing the ratio $f(x)/g(x)$.

Super good use case

We want $X \sim \mathcal{N}(0,1)$ conditionally on $X \geqslant \tau$ for $\tau \geqslant 1$ especially $\tau \gg 1$

Use proposals $Y = \tau + \mathrm{Exp}(1)/\tau$

Mixtures

For a parametric $f(x;\theta)$ that we can sample, use a random θ.

Beta-binomial

$$X \sim \mathrm{Bin}(n,p), \quad p \sim \mathrm{Beta}(\alpha,\beta) \qquad \propto p^{\alpha-1}(1-p)^{\beta-1}$$

Negative binomial

$$X \sim \mathrm{Poi}(\lambda), \quad \lambda \sim \mathrm{Gam}(\alpha)/\rho$$

Mixture of Gaussians

$$f(x) = \sum_{j=1}^{M} \alpha_j \mathcal{N}(\mu_j, \sigma_j^2), \quad \sum_j \alpha_j = 1, \quad \alpha_j \geqslant 0$$

1) Random J, $\mathbb{P}(J=j) = \alpha_j$

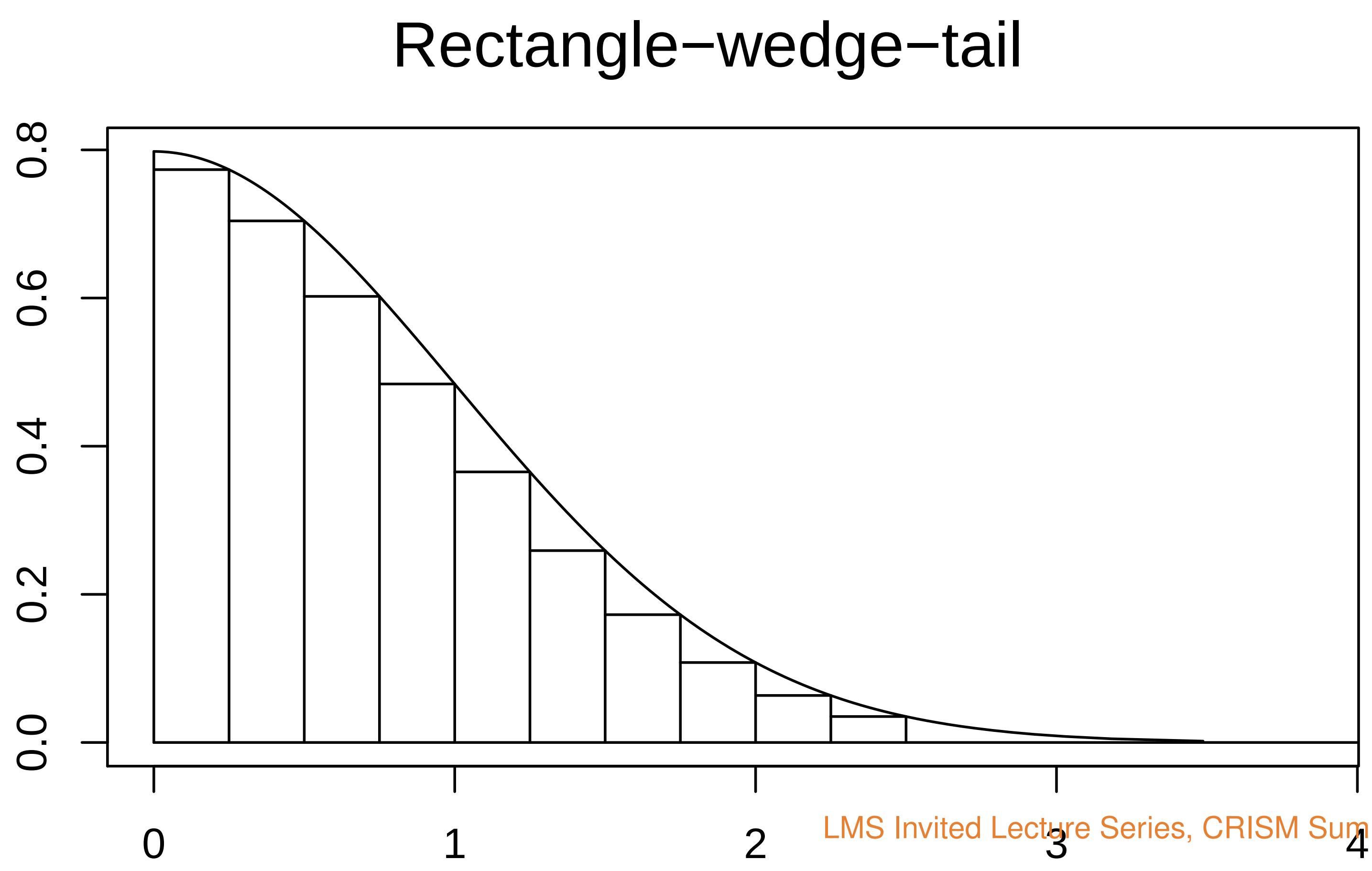
2) $X \sim \mathcal{N}(\mu_j, \sigma_j^2)$

Carpentry

Split region R under f into subregions R_j

Choose J with $\mathbb{P}(J = j) = \mathrm{Area}(R_j)$

Sample as if $\mathbf{U}(R_j)$.



Example of a ‘trick’

For $E \sim \mathrm{Exp}(1)$ let $X = \lfloor E \rfloor = \max\{z \in \mathbb{Z} \mid E \geqslant z\}$

Now

$$\mathbb{P}(X = x) = \int_x^{x+1} e^{-z}\,\mathrm{d}z = -e^{-z}\Big|_x^{x+1} = e^{-x} - e^{-x-1} = (1-\theta)^x\theta$$

for $\theta = 1 - e^{-1}$. This is a geometric distribution, number of trials to first success.

It would be nice to have $\mathbb{P}(X = x) = (1-\theta)^x\theta, \quad x = 0, 1, \ldots$
for **any** success probability $\theta \in (0, 1]$ that we like.

Discrete random variables

Arbitrary ones can require some cumbersome bookkeeping.

Summary

- MC is used on problems that we cannot do otherwise.
- We have several tricks for non-uniform distributions.
- We can usually find one that works.
- Things are different for random vectors, objects, processes.

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli