An estimate for the probability of heads on a coin whose relative error is independent of the true value

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

29 Feb, 2016

Gregorian Calendar

365.242189 days (reality)

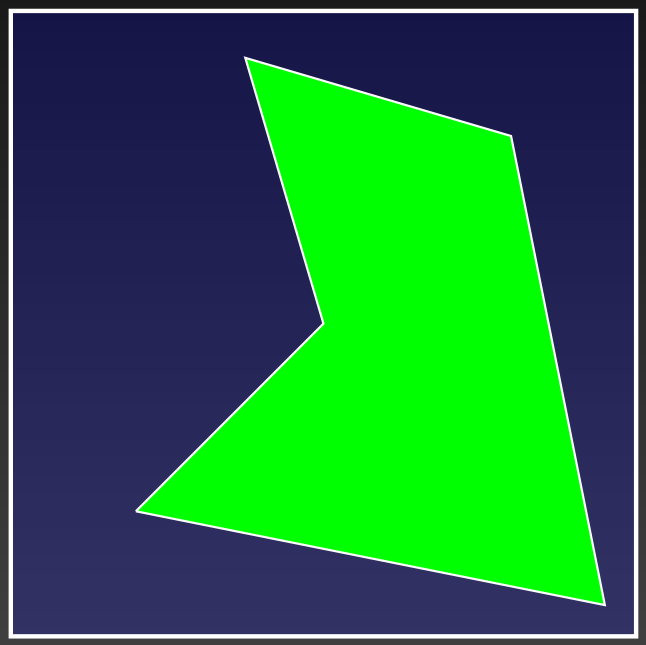
365
365 + 1/4 = 365.25 (for leap day)
365 + 1/4 - 1/100 = 364.24 (don't leap every 100 years)
365 + 1/4 - 1/100 + 1/400 (don't don't leap on multiples of 400)

365.2425 days (Gregorian)

What is it I do?

I flip coins for a living!

To do high dimensional integration



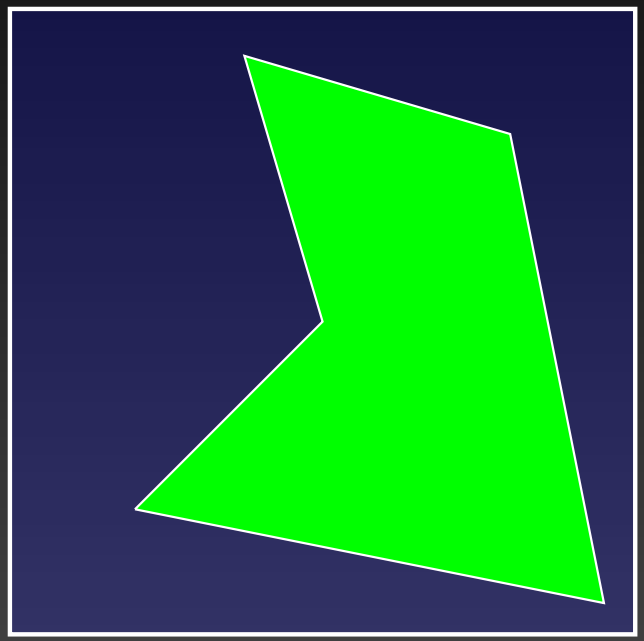
The chance that a uniform draw from the blue area lands inside the green area is

$$\frac{\text{green area}}{\text{blue area}}$$

Just like flipping a coin!

- Heads if you land in the green area
- Tails if you don't

Monte Carlo high dimensional integration



1. Using repeated uniform draws from the blue area…

2. estimate p, the chance a draw falls in the green area by $\hat{p}$…

3. and return blue area $\cdot\ \hat{p}$ as estimate for green area.

Today

An estimator $\hat{p}$ for the probability of heads p of a coin, where $\hat{p}/p$ does not depend on p.

Applications

- $\#P$ complete problems
  - Example: how many independent sets of a graph are there?
- Entropy for statistical physics models
- Exact p-values
- Normalizing constants of posterior distributions
- Bayes' Factors for hypothesis testing

Relative Error

$$\epsilon_{\mathsf{rel}} = \frac{\hat{\ell}}{\ell} - 1$$

Some reasons you might need relative error

1. Unitless quantity

2. Normalizing constants of high dimensional distributions grow exponentially in input size

3. Bayes' Factors are ratios of quantities, both of which should have relative error tolerances

4. Importance sampling effectiveness measured by coefficient of variation = relative standard deviation

5. In theoretical computer science, gives randomized approximation scheme for $\#P$ complete problems.

Randomized approximation scheme

Definition

An estimate $\hat{a}$ for a is an (ϵ, δ)-*randomized approximation scheme* if the probability that the relative error is at least ϵ is at most δ, so

$$\mathbb{P}\left(\left|\frac{\hat{a}}{a} - 1\right| > \epsilon\right) < \delta.$$

Coins as information

A bit is a number that is 0 or 1.

The smallest unit of information in a digital world.

A coin flip

Gives one bit of information

The bit is random

Calling the coin

- England: Heads or Tails
- Ancient Rome: Navia aut Caput (Ship or Head)
- Argentina: *Cara o Cruz* (Face or Cross)
- Russia: *орел*, *решка* (Eagle or Reshka)

h. The reverse side of a coin; esp. in phr. *head(s) or tail(s)*: see HEAD *n.*[1] 4b.

1684 T. OTWAY *Atheist* II. 17 As the Boys do by their Farthings..go to Heads or Tails for 'em.

1767 T. BRIDGES *Homer Travestie* (ed. 2) I. III. 101 'Tis heads for Greece, and tails for Troy... Two farthings out of three were tails.

1801 J. STRUTT *Sports & Pastimes* IV. ii. 251 The reverse to the head being called the tail without respect to the figure upon it.

1884 *Punch* 16 Feb. 73/1 A sovereign, a half sovereign,..or farthing, so long as it has a 'head' one side, and..a 'tail' the other.

1893 F. W. L. ADAMS *New Egypt* 267 The goddess who sits on the 'tails' side of our bronze currency.

i. The lower, inner, or subordinate end of a long-shaped block or brick; the bottom or visible part of a roofing slate or tile.

1793 J. SMEATON *Narr. Edystone Lighthouse* (ed. 2) §82 The tail of the header was made to..bond with the interior parts.

1856 S. C. BREES *Terms & Rules Archit.* *Tail*,..the lower end of the slate or tile.

The problem

Let p be the probability of heads

Can flip coin as often as I want

Estimate p

Basic estimate

Basic estimate $\hat{p}_n$:

1. Flip coin n times (Draw $X_1, \ldots, X_n \leftarrow \text{Bern}(p)$ iid.)
2. Let $\hat{p}_n$ be fraction of time coin came up heads.

$$\hat{p}_n \leftarrow \frac{X_1 + \cdots + X_n}{n}.$$

Example: Flip coin 5 times

4 out of 5 heads makes $\hat{p}_5 = 4/5 = 0.8000$.

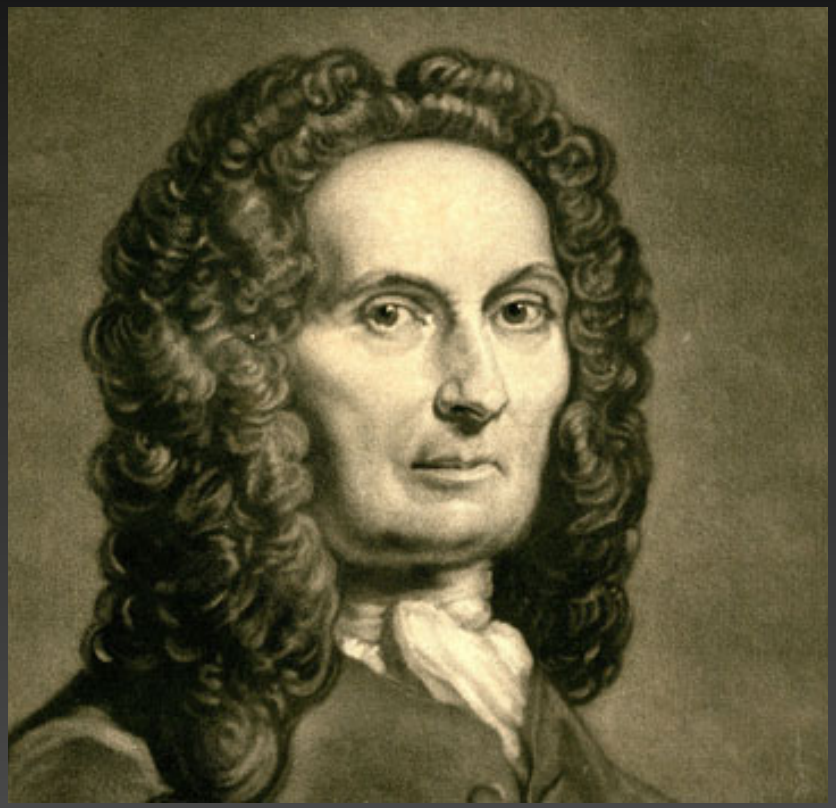
The basic estimate has lead to some great statistics

Jacob Bernoulli took 20 years to prove that Law of Large Numbers holds for $\{0, 1\}$ random variables. (Published posthumously in 1713)

Strong Law of Large Numbers

$\lim_{n \to \infty} \hat{p}_n = p$ with probability 1

How accurate is the basic estimate?



Abraham de Moivre proved in 1733 an early version of the Central Limit Theorem in order to study how the simple estimate behaves

Relative Error

$$\epsilon_{\mathsf{rel}} = \frac{\hat{\ell}}{\ell} - 1$$

Example

Suppose $p = 20\%$ and $\hat{p} = 22\%$. Relative error is:

$$\frac{22\%}{20\%} - 1 = 1.1 - 1 = 10\%.$$

Relative error using CLT

Use Central Limit Theorem to get rel error at most ϵ with probability at least $1 - \delta$, need

$$2\epsilon^{-2}p^{-1}(1-p)\ln(\delta^{-1})$$

samples.

Problem

- Do not know p
- CLT inaccurate when ϵ, δ small

Relative error for Basic estimate

Relative error depends both on p and n

Example: $n = 5$:

$$\frac{\hat{p}_5}{p} - 1 \in \left\{ \frac{0}{5p} - 1, \frac{1}{5p} - 1, \frac{2}{5p} - 1, \frac{3}{5p} - 1, \frac{4}{5p} - 1, \frac{5}{5p} - 1 \right\}$$

No way is relative error for basic estimate independent of p

Properties of a known relative error estimate

Theorem (H. 2016)

Let $\hat{p}$ be an nonnegative estimator for p such that the distribution of $\hat{p}/p$ does not depend on $p \in (0, 1]$. Then

1. *$\hat{p}$ can have positive probability of being o, but $\mathbb{P}(\hat{p} = a) = 0$ for all $a > 0$.*
2. *$\hat{p}$ is unbounded: for all $a > 0$, $\mathbb{P}(\hat{p} > a) > 0$.*

Note: in particular, $\mathbb{P}(\hat{p} > 1) > 0$

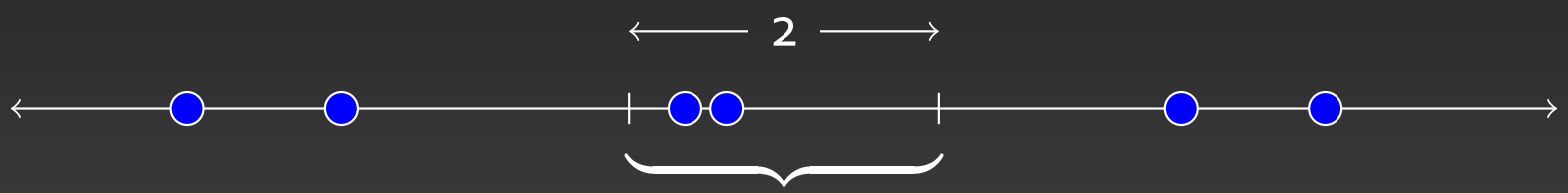
There are applications where allowing $\hat{p} > 1$ is important!

The New Algorithm

Point process

Definition

A *point process* is a collection of a random number of points N drawn from a region A, so $\{X_1, \ldots, X_N\} \subseteq A$.

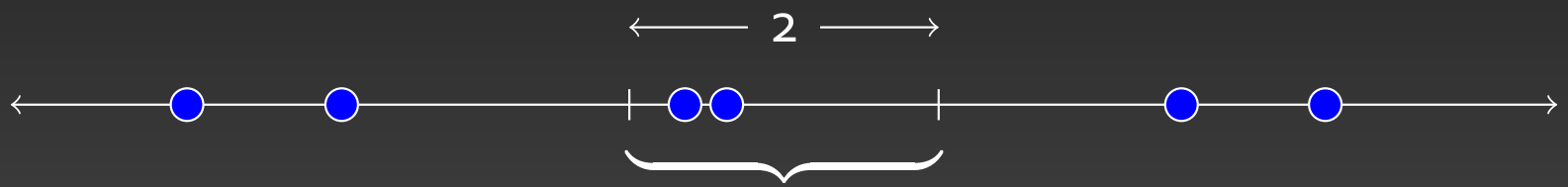


Expected number in interval of length 2 is 2λ

Poisson point process

Definition

A point process is *Poisson* if there is a parameter λ such that for any interval of length a, the average number of points of the process that fall into the interval is λa.



Expected number in interval of length 2 is 2λ

Example: Collins Dining Hall

Suppose students arrive at Collins as a Poisson point process

$$\lambda = 90/\text{hour}.$$

Average number of customers that arrive in the first half-hour is

$$\lambda(1/2)\text{hour} = \frac{90}{\text{hour}} \cdot \frac{1}{2}\,\text{hour} = 45.$$

Two ways to change the rate

The rate λ can be changed by using

1. Thinning
2. Scaling

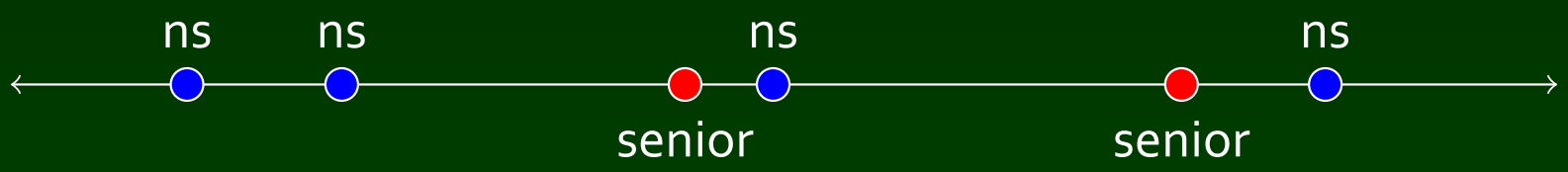
By combining these effects, can make p disappear from $\hat{p}/p$!

Changing the rate through thinning

Suppose each student arriving has a 26% chance of being a senior.

Then the rate at which seniors arrive is

$$\frac{90}{\text{hour}}(0.26) = \frac{23.4}{\text{hour}}.$$

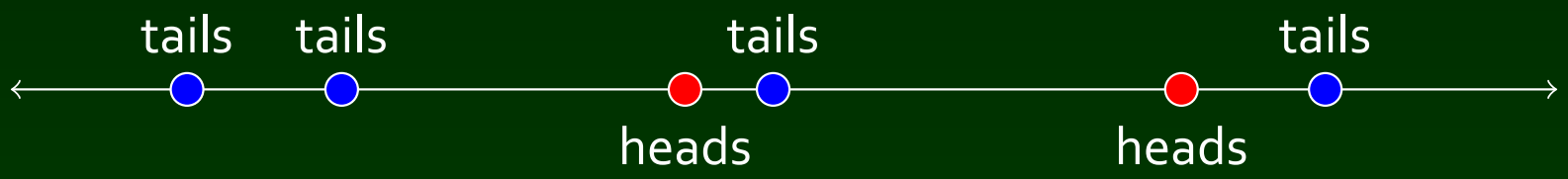


Process called *thinning*

Changing the rate through thinning

Suppose for each point flip Bern(p)

Only keep points that get heads



Old expected number in interval $[a, b]$ is $\lambda(b - a)$

Expected number in interval $[a, b]$ is $\lambda p(b - a)$

New effective rate: λp

Changing the rate by scaling

Suppose customers arrive McDonald's at rate 90/hour

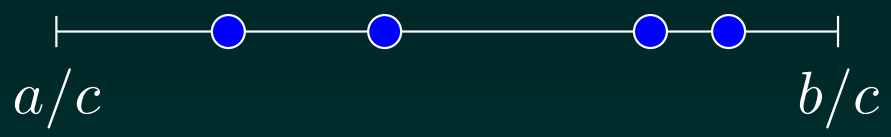
First customer arrives at time 0.4 hour

Change to minutes:

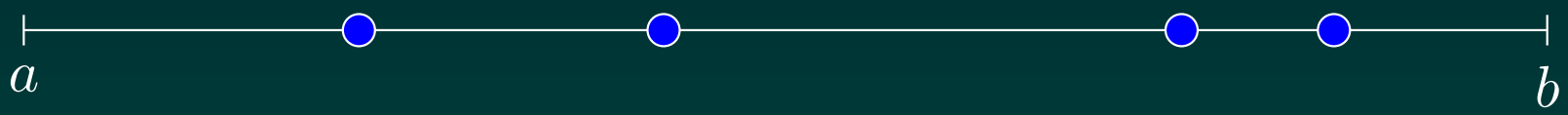
$$90/\text{hour} \mapsto (90/60) = 1.5/\text{minute}$$
$$0.4\,\text{hour} \mapsto (0.4)(60) = 24\,\text{minutes}$$

Changing the rate by scaling

Start with Poisson point process rate λ over $[a/c, b/c]$



Now scale by multiplying by c



Average # of pts. in $[a, b]$ in new is average # in $[a/c, b/c]$ in old

$$\lambda(b/c - a/c) = (\lambda/c)(b - a).$$

Changing the rate by scaling

For a Poisson point process of rate λ:

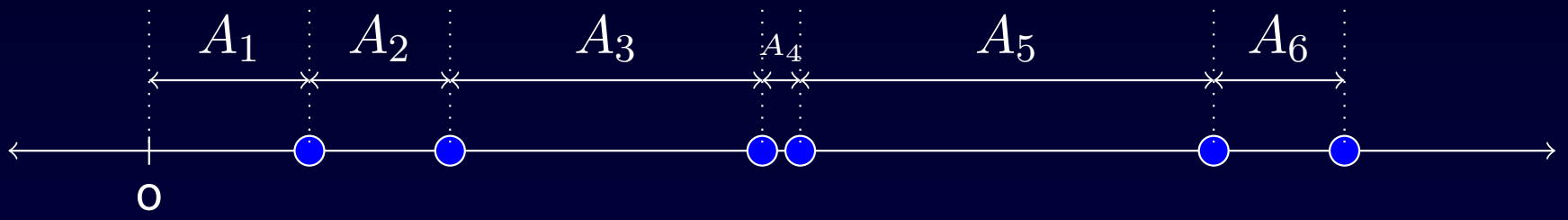
$$P_1, P_2, P_3, \ldots$$

and any constant c:

$$P_i \mapsto cP_i$$
$$\lambda \mapsto \lambda/c$$

Time between points are exponentially distributed

Distances between points are iid exponential r.v.'s of rate λ



$$A_1, A_2, A_3, \ldots \sim \text{Exp}(\lambda) \text{ iid}$$

Drawing $P_1 \sim$ *Exp*(p)

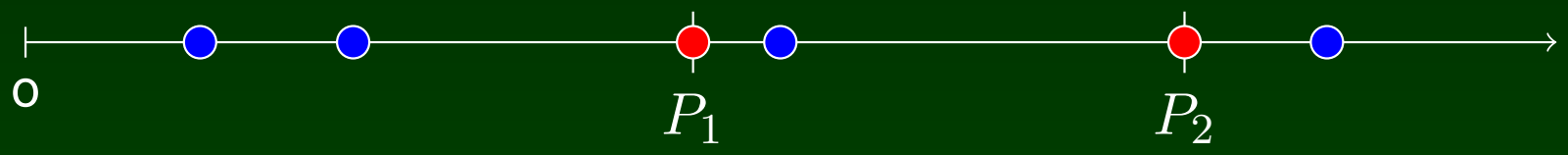
- Generate Poisson process of rate 1 on $[0, 1]$
- Thin it using the p-coin

Time until first head is Exp(p)!

Gamma Bernoulli Approximation Scheme

New estimate for p:

- Run Poisson point process of rate 1 forward in time from 0
- Thin the process as it is run forward using p-coin
- Continue until reach k heads
- Let P_k be time of the kth head



[P_k has a gamma distribution with parameters k and p]

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

An Example

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

An Example

1. Decide that $k = 4$ is sufficient

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

An Example

1. Decide that $k = 4$ is sufficient
2. Suppose it takes 22 flips to get 4 heads

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

An Example

1. Decide that $k = 4$ is sufficient
2. Suppose it takes 22 flips to get 4 heads
3. Generate $A_1, \ldots, A_{22}$ iid Exp(1)

The Algorithm

1. Decide what value of k you want to use
2. Flip p-coin until get k heads. Say it takes N flips
3. Generate $A_1, \ldots, A_N$ iid Exp(1) random variables
4. Estimate is $\hat{p} = (k-1)/(A_1 + \cdots + A_N)$.

An Example

1. Decide that $k = 4$ is sufficient
2. Suppose it takes 22 flips to get 4 heads
3. Generate $A_1, \ldots, A_{22}$ iid Exp(1)
4. Final estimate $(4-1)/(A_1 + \cdots + A_{22}) = 0.1823\ldots$

Easy to implement

Six lines of pseudocode

GBAS *Input:* $k \geq 2$

1) $R \leftarrow 0$, $S \leftarrow 0$
2) Repeat
3) $\quad X \leftarrow \text{Bern}(p)$, $A \leftarrow \text{Exp}(1)$
4) $\quad S \leftarrow S + X$, $R \leftarrow R + A$
5) Until $S = k$
6) $\hat{p} \leftarrow (k-1)/R$

The cool part

Consider the relative error

$$\frac{\hat{p}}{p} - 1 = \frac{k-1}{P_k p} - 1$$

But $P_k p$ is the equivalent of scaling time by a factor of p

- Started with rate 1 process
- Thinned to get rate p process
- Scaling time by p gives rate $p/p = 1$ process again!

Relative error does not depend on p!

Distribution of relative error known exactly

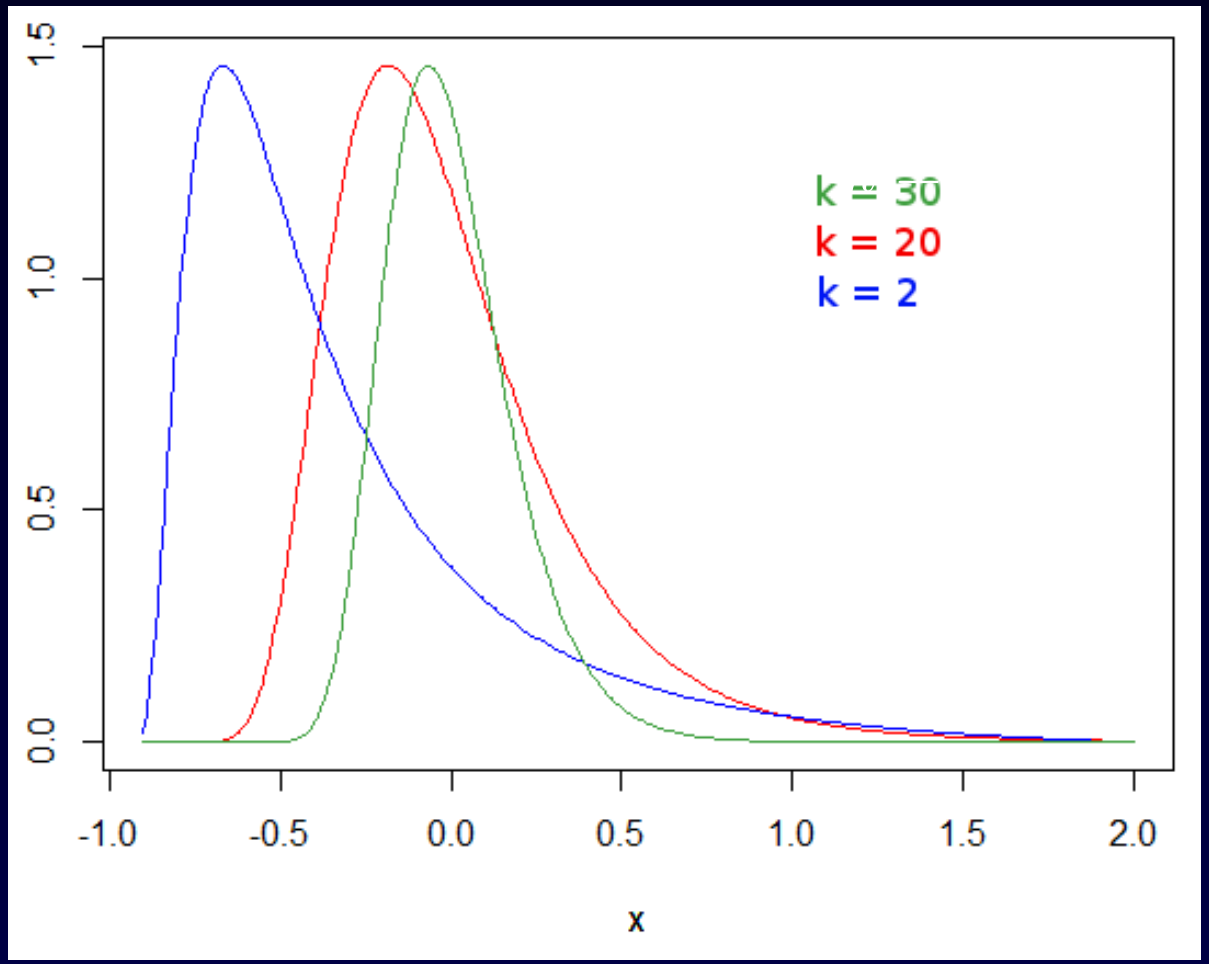
Adding exponentials

- When $T_1, \ldots, T_k \sim \text{Exp}(\lambda)$...
- ...$T_1 + \cdots + T_k \sim \text{Gamma}(k, \lambda)$
- So $pP_k \sim \text{Gamma}(k, 1)$
- If $X \sim$ Gamma, $1/X \sim$ InvGamma

$$\frac{\hat{p}}{p} \sim \text{InvGamma}(k, 1/(k-1))$$

- $\mathbb{E}(\hat{p}/p) = 1/[(k-1)/(k-1)] = 1)$

As k increases, relative error concentrates about zero



Benefits

Since we know distribution of $\hat{p}/p$ exactly

- Get exact confidence intervals for p easily
- Yields faster randomized approximation schemes
- Theory gives first order same as CLT

Does it work well in practice? YES!

If we knew p exactly

- Exactly find probabilities of tails of binomial distribution
- Use this to find the exact n needed for the basic estimate to be an (ϵ, δ) approximation

The results

- Suppose I want an estimate with absolute relative error at most 10% with probability at least 95%

$\epsilon = 0.1, \delta = 0.05$

p	Exact n	$\mathbb{E}[T_p]$	$\mathbb{E}[T_p]/n$
1/20	7 219	7 700	1.067
1/100	37 545	38 500	1.025

Why we might want to allow $\hat{p} > 1$

M. L. Huber and R. L. Wolpert, Likelihood-based inference for Matérn type-III repulsive point processes, Advances in Applied Probability, *41(4), pages 958-977, 2009*

Created an MLE by finding likelihood as product of several p_i:

$$\ell = p_1 p_2 \cdots p_n$$

Say $p_i \in [0.4, 0.6]$. Then estimators might be something like:

$$\hat{\ell} = (0.55)(1.3)(0.45) \cdots (1.2).$$

Rounding 1.3 to 1 throws off estimate!

Poisson

Can we do the same thing for Poisson?

Definition

Say that X is a *Poisson* random variable with mean μ, write $X \sim \text{Pois}(\mu)$ if

$$\mathbb{P}(X = i) = \exp(-\mu)\frac{\mu^i}{i!} \text{ for } i \in \{1, 2, 3, \ldots\}.$$

Why is the Poisson point process called Poisson?

For a Poisson point process $\{P_i\}$ of rate λ, the number of points that fall into $[a, b]$ satisfies:

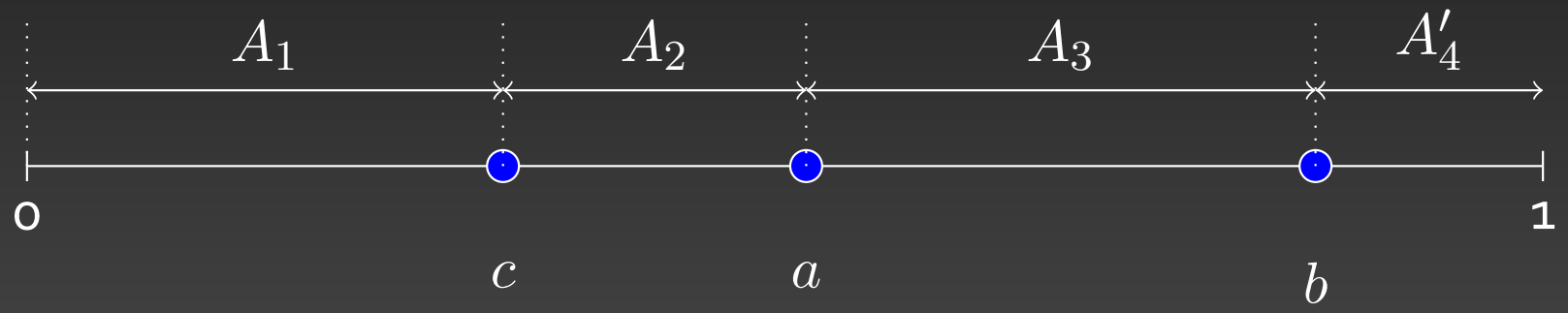
$$N_{[a,b]} = \#(\{P_i\} \cap [a, b])$$
$$\mathbb{E}[N_{[a,b]}] = \lambda(b - a)$$
$$N_{[a,b]} \sim \text{Pois}(\lambda(b - a))$$

Conditioning on number of points

Suppose we know $N_{[0,1]} = 3$.

Say $\{a, b, c\} = \{P_i\} \cap [0, 1]$

Then $a, b,$ and c are iid uniform over $[0, 1]$



How to get A_4

Generate Poisson point process on $[1, 2]$:

$$A_4 = A_4' + A_4''$$

Convert Poisson to Exp

Start with

$$N_{[0,1]}, N_{[1,2]}, N_{[2,3]}, \ldots \overset{\text{iid}}{\sim} \text{Pois}(\mu),$$

end with

$$A_1, A_2, A_3, \ldots \overset{\text{iid}}{\sim} \text{Exp}(\mu)$$

Conversion Rates

On average

$$\frac{1}{p} \text{ Bern}(p) = 1 \text{ Exp}(p)$$

$$1 \text{ Pois}(\mu) = \mu \text{ Exp}(\mu)$$

To get k Exp random variables, need on average:

$$k/p \text{ Bern}(p)$$

$$\lceil k/\mu \rceil \text{ Pois}(\mu)$$

Application

The Tootsie Pop Algorithm (TPA)

> *M. L. Huber and S. Schott. Random construction of interpolating sets for high dimensional integration.* Journal of Applied Probability, *51(1), pages 92–105, 2012*

turns the problem of integrating a very general class of problems (including finding the partition function of a Gibbs distribution) into a problem of estimating the mean of a Poisson random variable.

References

Huber, M., "An unbiased estimate for the mean of a $\{0, 1\}$ random variable with relative error distribution independent of the mean", arXiv:1309.5413, 2013

Huber, M., "A Bernoulli mean estimate with known relative error distribution", Random Structures & Algorithms, to appear