An estimate for the chance p of heads on a coin where the relative error does not depend on p

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

The Problem

Flipping coins

Heads

Tails

The question

What is the probability the coin is heads?

Basic estimate

Step 1: Numerically encode coins:

Heads = 1, Tails = 0

Step 2: Assign probability distribution:

$$C \sim \text{Bernoulli}(p) \Rightarrow \mathbb{P}(X = 1) = p, \; \mathbb{P}(X = 0) = 1 - p$$

Step 3: Basic estimate:

$$\hat{p}_n = \frac{C_1 + \cdots + C_n}{n}, \quad C_i \overset{\text{iid}}{\sim} \text{Bern}(p)$$

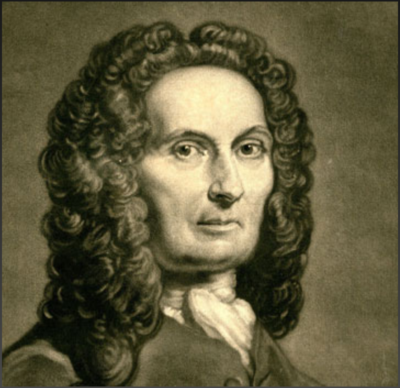
This has lead to some great mathematics

Jacob Bernoulli proved in 1713 an early version of the Strong Law of Large Numbers.

Strong Law of Large Numbers:

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

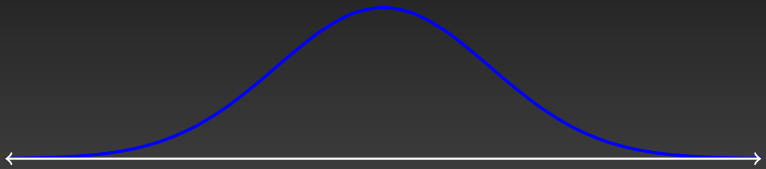
But how fast does it converge?



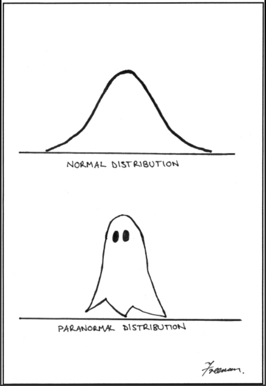
Abraham de Moive
proved in 1733 an early version of the Central Limit Theorem in order to study how the simple estimate behaves

The Central Limit Theorem

The CLT says that in the limit as you add independent, identically distributed random variables, the resulting density approaches a normal distribution:



Our coin flips are iid, so $\hat{p}$ approximately normal...



M. Freeman, A visual comparison of normal and paranormal distributions, *J. of Epidemiology and Community Health*, 60(1), p. 6, 2006

Central Limit Theorem

The CLT has some drawbacks for this problem

- Convergence to normal polynomial in n
- Tails exponentially small in n
- Not accurate out in the tails
- For confidence close to 1
- Bad for small p
- Gives additive error, not relative error

Relative Error

Definition

The *relative error* of an estimate $\hat{p}$ for p is

$$\frac{\hat{p}}{p} - 1 = \frac{\hat{p} - p}{p}.$$

Example: Actual answer: 20%
Estimate: 23%
Relative error: $3\%/20\% = 15\%$

Today

I will present an unbiased estimate $\hat{p}$ for p where the relative error does not depend in any way on p.

Estimate properties

The new estimate

- Requires a random number of flips
- Unbiased
- Number of flips very close to optimal
- Relative error distribution known exactly
- Allows easy construction of exact confidence intervals

Relative error and basic estimate: An example

For the basic estimate, relative error heavily depends on p

Suppose $n = 5$ and $p = 0.25$.

Then

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

and

$$\frac{\hat{p}_n}{p} \in \left\{ \frac{0}{5}, \frac{4}{5}, \frac{8}{5}, \frac{12}{5}, \frac{16}{5}, \frac{20}{5} \right\}.$$

The values that $[\hat{p}_n/p] - 1$ takes on depends on p!

More generally

More generally:

$$\frac{\hat{p}_n}{p} \in \left\{ \frac{0}{np}, \frac{1}{np}, \frac{2}{np}, \ldots, \frac{n}{np} \right\}$$

New estimate

- Distribution of $\hat{p}/p - 1$ does not depend on p
- Allows us to easily find exact confidence intervals

But is it fast?

Goal: for $\epsilon > 0$ and $\delta > 0$,

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

Suppose we knew p ahead of time, what should n be exactly for

$$\epsilon = 10\% \text{ and } \delta = 5\%?$$

Can directly calculate tails of a binomial to get:

p	Exact n
1/20	7219
1/100	37546

New algorithm

Let T_p be number of flips required by new estimate

$\epsilon = 0.1$, $\delta = 0.05$			
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

$\epsilon = 0.01$, $\delta = 10^{-6}$			
p	Exact n	$\mathbb{E}[T_p]$	$\mathbb{E}[T_p]/n$
1/20	4 545 010	4 789 800	1.053
1/100	236 850 500	239 490 000	1.011

Estimate properties

The new estimate

- Requires a random number of flips
- Unbiased
- Number of flips very close to optimal
- Relative error distribution known exactly
- Allows easy construction of exact confidence intervals

How did I get started on this problem?

My work is in perfect simulation

- Drawing samples exactly from high dimensional models...
- ...usually using a random number of samples.

Examples

- Ising (and autonormal) model
- Strauss model
- Widom-Rowlinson
- Allele frequency tables
- Colorings of graphs
- Matérn type III process
- Weighted assignments

Numerous applications

Applications

Application: Finding the permanent of a matrix

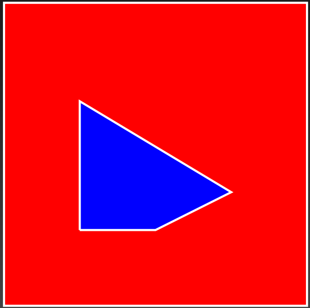
M. Huber and J. Law. Fast approximation of the permanent for very dense problem. In *Proc. of 19th ACM-SIAM Symp. on Discrete Alg.*, pp. 681–689, 2008

Definition

Suppose n workers are to be assigned to jobs $\{1, 2, \ldots, n\}$, but each worker is only qualified for a specified set of jobs. The number of such assignments is called the *permanent*.

Acceptance/Rejection

HL 2008 was an example of an acceptance/rejection method...



Goal: Estimate size of blue

1. Draw $X_1, \ldots, X_n$ from red region
2. Let k be the number of X_i that fell into blue
3. Estimate is (k/n) times size of red region

This is just the coin problem!

Probability of heads p is size of blue over size of red

- Want to minimize number of draws from red region...
- ...is the same as number of flips of a coin

In the paper, used a Chernoff bound to bound Binomial tails:

$$14p^{-1}\epsilon^{-2}\ln(2/\delta)$$

flips of the coin sufficed

Then I was asked to referee a paper...

The authors had referenced the 2008 permanent paper...

- They used the 14 constant
- This constant is way too large
- So I started work to reduce this constant

How should number of samples vary with p?

To get a rough idea of how many samples needed, consider $\hat{p}_n$

$$\mathbb{E}[\hat{p}_n] = p, \quad \mathrm{SD}(\hat{p}_n) = \sqrt{\frac{(p)(1-p)}{n}}$$

So to get $\mathrm{SD}(\hat{p}_n) \approx \epsilon p$...

$$n \approx \epsilon^{-2}(1-p)/p$$

Number of samples should be $\Theta(1/p)$, but we don't know p

DKLR

P. Dagum, R. Karp, M. Luby, and S. Ross. An optimal algorithm for Monte Carlo estimation. *Siam. J. Comput.*, 29(5):1484–1496, 2000.

They solved the $1/p$ problem in the following clever way:

1. Keep flipping coins until get k heads
2. Estimate is k divided by number of flips

Run time On average, draw k/p samples

[Biased estimate, however]

Their theorem: to get $\mathbb{P}(|(\hat{p}_{\text{DKLR}}/p) - 1| > \epsilon) < \delta$,

$$k \geq 1 + 4(e - 2)(1 + \epsilon)\ln(2/\delta)\epsilon^{-2}$$

Note $4(e - 2) \approx 2.873\ldots$

Running time for DKLR

$\epsilon = 0.1$, $\delta = 0.05$			
p	Exact n	$\mathbb{E}[T_p]$	$\mathbb{E}[T_{\text{DKLR}}]$
1/20	7 219	7 700	23 340
1/100	37 545	38 500	116 700

Application: Exact p values

M. Huber, Y. Chen, I. Dinwoodie, A. Dobra, and M. Nicholas, Monte Carlo algorithms for Hardy-Weinberg proportions, *Biometrics*, 62(1), pp. 49–53, 2006.

Definition

A *p value* is the probability that a statistic applied to a draw from the null hypothesis model is more unusual than the statistic applied to the data.

Low p-value = evidence that null hypothesis is untrue

Estimating p*-values with perfect samples*

Want p-value for a statistic $S(\cdot)$

A p-value is just

$$\mathbb{P}(S(X) \text{ is weirder than} S(\text{data}))$$

where X is a draw from statistical model

So if have algorithm for drawing X exactly from model...

This is again exactly the coin flipping problem!

The Estimate

Uniform and exponential random variables

Say $U \sim \text{Unif}([0,1])$ if for all $0 < a < b < 1$,

$$\mathbb{P}(a < U < b) = b = a.$$

To get an exponential random variable (with rate 1):

$$U \sim \text{Unif}([0,1]) \Rightarrow -\ln(U) \sim \text{Exp}(1)$$

The algorithm in words:

Before you begin:

- Fix k a positive integer

The estimate:

1. Flip a coin
2. Draw an exponential random variable of rate 1
3. Add the exponential to a total of time
4. Keep doing 1 through 3 until you have k heads
5. The final estimate is $k - 1$ divided by the sum of the exponentials

The algorithm in pseudocode

Gamma Bernoulli Approximation Scheme

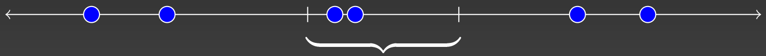
GBAS *Input:* $k \geq 2$

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

Poisson point process

Definition

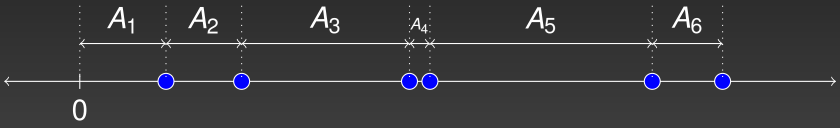
P is a *Poisson point process* on a region A of rate λ if for any $B \subseteq A$ of finite size, the mean number of points of P that fall in B is λ times the size of B. Also, the # of points in an in interval is independent of the # of points in a disjoint interval.



Expected number in interval of length 2 is 2λ

Equivalent Formulation

Distances between points are iid
exponential random variables of rate λ
(take exp rate 1, divide by λ)

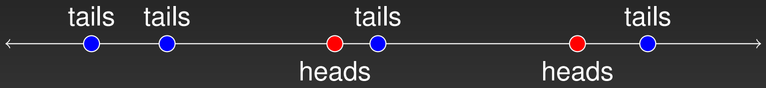


$$A_1, A_2, A_3, \ldots \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$$

Back to mean formulation...

Suppose for each point flip Bern(p)

Only keep points that get heads



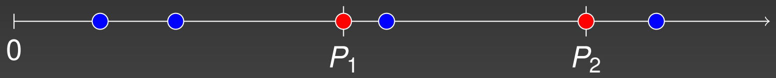
Expected number in interval $[a, b]$ is $\lambda p(b - a)$

New effective rate: λp

Process called *thinning*

Estimate: Poisson formulation

- Run Poisson process forward in time from 0
- Each point flip a coin–only keep heads
- Continue until have k heads
- Let P_k by time of the kth head
- Estimate is $(k-1)/P_k$



Gamma Bernoulli Approximation Scheme

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)	$\hat{p} \leftarrow (k-1)/R$

Gamma distributions

Because $P_i - P_{i-1} \sim \text{Exp}(p)$

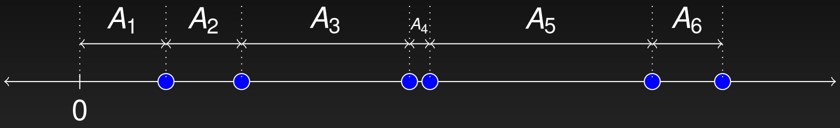
$P_k \sim \text{Gamma}(k, p)$ [sum of k exponentials]

So $1/P_k \sim \text{InverseGamma}(k, p)$

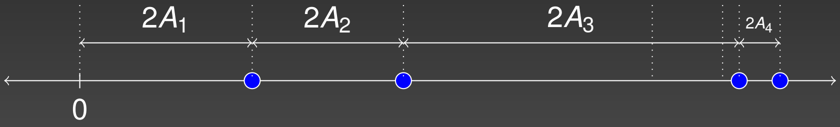
$$\mathbb{E}\left[\frac{k-1}{P_k}\right] = (k-1)\mathbb{E}[P_k^{-1}] = (k-1)\frac{p}{k-1} = p$$

Estimate is unbiased!

Back to exponential formulation...



Multiply all the A_i by 2:



New expected number in $[0, t]$ = old expected in $[0, t/2]$

So $\lambda(t - 0)/2 \Rightarrow$ new rate is $\lambda/2$

Scaling exponentials

Fact

If $X \sim Exp(\lambda)$, then $cX \sim Exp(\lambda/c)$.

Fact

If $X \sim Gamma(k, \lambda)$, then $X \sim Gamma(k, \lambda/c)$.

That means

$$\frac{\hat{p}}{p} = \frac{k-1}{pP_k} = (k-1)A,$$

where $A \sim \text{InverseGamma}(k, p/p) = \text{InverseGamma}(k, 1)$

Relative error independent of p

Theorem

For $\hat{p}$ given earlier,

$$\mathbb{E}[\hat{p}] = p, \quad \frac{\hat{p}}{p} - 1 \sim (k-1)A - 1,$$

where $A \sim \mathit{InverseGamma}(k, 1)$, making the relative error independent of p. The expected number of flips used by the estimate is k/p.

Filling in the table

Recall the table we had earlier...

$\epsilon = 0.1$, $\delta = 0.05$			
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

How I filled in those entries:

$$\min_n \mathbb{P}(|(\text{Bin}(n, 1/20)/n)/(1/20) - 1| > 0.1) < 0.05 = 7219$$

$$\min_k \mathbb{P}(|(k-1)\text{InverseGamma}(k, 1) - 1| > 0.1) < 0.05 = 385,$$

and $385/p = 385/(1/20) = 7700$.

Comparison to

How many samples should be taken if CLT exact?

What should the constant be?

For basic estimate $\hat{p}_n$:

$$\mathbb{E}[C_i] = p, \quad \mathrm{SD}(C_i) = \sqrt{p(1-p)},$$

by CLT

$$\hat{p}_n = \frac{C_1 + \cdots + C_n}{n} \approx \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$$

which means

$$\frac{\hat{p}_n}{p} \approx \mathrm{N}\left(1, \frac{1-p}{np}\right)$$

So for relative error...

Subtracting 1

$$\frac{\hat{p}_n}{p} - 1 \approx N\left(0, \frac{1-p}{np}\right)$$

Hence

$$\sqrt{\frac{np}{1-p}}\left(\frac{\hat{p}}{p} - 1\right) \approx \mathrm{N}(0,1)$$

Bounding the normal tail

$$Z \sim \mathrm{N}(0, 1) \text{ with density } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$$

$$\left(\frac{1}{a} - \frac{1}{a^3}\right) \phi(a) \leq \mathbb{P}(Z \geq a) \leq \left(\frac{1}{a}\right) \phi(a)$$

Combining these results

$$\mathbb{P}\left(\left|\frac{\hat{p}_n}{p} - 1\right| > \epsilon\right) = \mathbb{P}\left(\sqrt{\frac{np}{1-p}}\left|\frac{\hat{p}_n}{p} - 1\right| > \epsilon\sqrt{\frac{np}{1-p}}\right)$$

$$\approx \mathbb{P}\left(|Z| > \epsilon\sqrt{\frac{np}{1-p}}\right)$$

Note

$$\phi\left(\sqrt{\frac{np}{1-p}}\epsilon\right) = \frac{1}{\sqrt{2\pi}}\exp(-np\epsilon^2/(1-p))$$

The result

When CLT holds exactly

Let $C_i \sim \mathrm{N}(p, p(1-p))$, then

$$n = \left[\frac{2(1-p)}{p}\epsilon^2\right]\ln(2/\delta) + \text{ lower order terms}$$

New estimate

Let $C_i \sim \mathrm{Bern}(p)$, then

$$\mathbb{E}[T] = \left[\frac{2}{p}\epsilon^2\right]\ln(2/\delta) + \text{ lower order terms}$$

Final thoughts

Some current projects

Bernoulli Factory

Given a p coin, can you flip a $2p$ coin?

Concentration

If you only bound standard deviation can you get concentration as if you had a normal random variable?

Current results: Assuming CLT need ϵ^{-2}, new method $64 + \epsilon^{-2}$

Partition functions

How many samples are necessary to estimate the normalizing constant of a Gibbs distribution?

Simulation with fixed correlation

Copulas are not the only method (with Nevena Marić)

Summary

Applications

- Numerical integration
- Finding exact p values

The new estimate

- Unbiased
- Easy to build
- Nearly optimal number of samples (lose factor of $1 - p$)
- Relative error $(\hat{p}/p) - 1$ independent of p
- Easy to get exact confidence intervals