

Computational Statistics

Section 3.2: Small probability estimation

Why estimate something that almost never happens?

- **Engineering:** a load exceeds a component's strength.
- **Finance:** a modelled loss exceeds a severe threshold.
- **Computing:** a queue exceeds its available capacity.

A simulator may produce millions of ordinary outcomes before one failure.

Our question

Can we make the event occur more often in the simulation, while still estimating its probability under the original model?

An event probability is an expectation

Let $X \sim p$ and let A be the event of interest. Write $P_A = \mathbb{P}_p(X \in A)$.

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Then, as in equation (3.2) of the notes,

$$P_A = \int_A p(x) dx = \int \mathbf{1}_A(x) p(x) dx = \mathbb{E}_p[\mathbf{1}_A(X)].$$

Interpretation: average the answers to the yes/no question “Did the event occur?”

Direct Monte Carlo counts hits

Draw $X_1, \dots, X_N \stackrel{\text{iid}}{\sim} p$ and set

$$K = \sum_{i=1}^N \mathbf{1}_A(X_i), \quad \hat{P}_{\text{MC}} = \frac{K}{N}.$$

- K is the number of simulations that fall inside A .
- $K \sim \text{Binomial}(N, P_A)$.
- $\mathbb{E}[\hat{P}_{\text{MC}}] = P_A$: the estimator is unbiased.

Example

If 12 out of 1,000 simulations hit the event, the estimate is 0.012.

Variance of the direct estimator

Let $Y = \mathbf{1}_A(X)$. Since $Y^2 = Y$,

$$\mathbb{E}[Y] = P_A, \quad \text{Var}(Y) = P_A - P_A^2 = P_A(1 - P_A).$$

Independence gives

$$\boxed{\text{Var}(\hat{P}_{\text{MC}}) = \frac{P_A(1 - P_A)}{N}}.$$

$$\text{SD}(\hat{P}_{\text{MC}}) = \sqrt{\frac{P_A(1 - P_A)}{N}}.$$

The absolute standard deviation gets smaller as P_A approaches zero. Why can the problem still become harder?

Rare events require relative accuracy

For $P_A > 0$, define the relative standard deviation

$$\frac{\text{SD}(\hat{P}_{\text{MC}})}{P_A} = \sqrt{\frac{1 - P_A}{NP_A}} \approx \frac{1}{\sqrt{NP_A}}.$$

To achieve relative standard deviation at most r ,

$$N \geq \frac{1 - P_A}{r^2 P_A}.$$

A probability of one in a million

For $P_A = 10^{-6}$ and $r = 0.1$, we need approximately 10^8 draws.

The key quantity is NP_A , the **expected number of hits**.

No hits does not mean no risk

$$\mathbb{P}(K = 0) = (1 - P_A)^N \approx e^{-NP_A}.$$

If $NP_A = 0.1$, the probability of seeing no hits is about 90.5%.

- A zero estimate can occur even though $P_A > 0$.
- The empirical variance of all-zero indicators is also zero.
- Reporting “ 0 ± 0 ” would give misleading reassurance.

For *direct iid Bernoulli trials*, a one-sided exact 95% upper bound after zero hits is

$$P_A \leq 1 - 0.05^{1/N} \approx \frac{3}{N}.$$

Practical importance sampling for a probability

Draw $Z_i \stackrel{\text{iid}}{\sim} q$ and use

$$H_i = \mathbf{1}_A(Z_i) \frac{p(Z_i)}{q(Z_i)}, \quad \hat{P}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N H_i.$$

If $q > 0$ wherever $\mathbf{1}_A p$ contributes, then

$$\mathbb{E}_q[H_i] = \int_{A \cap \{q > 0\}} \frac{p(x)}{q(x)} q(x) dx = P_A.$$

Two changes: hit the event more often; count each hit with its correct weight.

Example 1: exactly nine heads in ten flips

Under the target model, ten flips are independent and fair:

$$X = (X_1, \dots, X_{10}), \quad X_j \sim \text{Bernoulli}(1/2), \quad A = \left\{ \sum_{j=1}^{10} X_j = 9 \right\}.$$

$$P_A = \binom{10}{9} (1/2)^{10} = \frac{10}{1024} = 0.009765625.$$

- This means **exactly nine**, not at least nine.
- The analytic answer lets us check a simulation method.

One sequence versus the number of heads

For a specific ordered sequence x with k heads,

$$p(x) = (1/2)^{10}.$$

For the total number of heads K ,

$$\mathbb{P}_p(K = k) = \binom{10}{k} (1/2)^{10}.$$

Avoid a double count

Use either sequence probabilities or count probabilities consistently. Do not attach an extra binomial coefficient to a simulated sequence's weight.

Make nine heads easier to observe

Under the proposal, flip an independent biased coin with head probability θ :

$$Z_j \sim \text{Bernoulli}(\theta), \quad 0 < \theta < 1.$$

The proposal hit probability is

$$r_\theta = \mathbb{P}_q(K = 9) = 10\theta^9(1 - \theta).$$

For $\theta = 0.9$,

$$r_{0.9} \approx 0.38742 \quad \text{versus} \quad P_A \approx 0.00977.$$

The raw fraction of hits now estimates $r_{0.9}$, not P_A .

Multiply the likelihood ratios across the flips

For a sequence containing k heads and $10 - k$ tails,

$$q_{\theta}(z) = \theta^k (1 - \theta)^{10-k},$$

$$w_{\theta}(z) = \frac{p(z)}{q_{\theta}(z)} = \left(\frac{0.5}{\theta}\right)^k \left(\frac{0.5}{1-\theta}\right)^{10-k}.$$

- Each head contributes a factor $0.5/\theta$.
- Each tail contributes a factor $0.5/(1 - \theta)$.
- Independence makes the joint probabilities products.

What does one simulated experiment contribute?

For $\theta = 0.9$ and exactly nine heads,

$$c = \left(\frac{0.5}{0.9}\right)^9 \frac{0.5}{0.1} \approx 0.0252068.$$

Thus

$$H = \begin{cases} c, & K = 9, \\ 0, & K \neq 9. \end{cases} \quad \hat{P}_{\text{IS}} = c \frac{M}{N},$$

where M is the number of proposal experiments with nine heads.

Illustration

If $M = 390$ out of $N = 1000$, then $\hat{P}_{\text{IS}} \approx 0.009831$.

Check the expectation directly

Since $c_\theta = P_A/r_\theta$,

$$\mathbb{E}_q[H] = r_\theta c_\theta + (1 - r_\theta)0 = P_A.$$

Equivalently,

$$\underbrace{10\theta^9(1-\theta)}_{\text{chance of a hit}} \underbrace{\frac{(1/2)^{10}}{\theta^9(1-\theta)}}_{\text{weight of a hit}} = 10(1/2)^{10}.$$

The event occurs more often, but its contribution is reduced by exactly the required amount.

How much variance do we save?

For the coin event,

$$\text{Var}_q(H) = c_\theta^2 r_\theta (1 - r_\theta) = P_A^2 \left(\frac{1}{r_\theta} - 1 \right).$$

Method	Single-output variance	SD of mean, $N = 1000$
Fair coin	0.00967026	0.003110
Biased coin, $\theta = 0.9$	0.000150793	0.0003883

Variance reduction: about 64.1 times. Standard deviation reduction: about 8.0 times.

Two-minute exercise: spot the incorrect estimators

A biased-coin simulation generates $N = 1000$ experiments and $M = 400$ hits. Every hit has weight $c = 0.0252068$.

1. Which estimates the fair-coin probability: M/N or cM/N ?
2. Calculate the correct estimate.
3. Would $\theta = 1$ be a better proposal because it gives more heads?

Explain each answer in words before calculating.

Exercise solution

1. The correct estimator is cM/N .
2. Its value is $0.0252068 \times 0.4 \approx 0.010083$.
3. No. If $\theta = 1$, every experiment produces ten heads, so the proposal never visits the event.

Why the M/N fails

M/N estimates the probability under the biased coin.

Example 2: a Gaussian point inside a disk

Let $X \sim \mathcal{N}(0, I_2)$, with density

$$p(x) = \frac{1}{2\pi} \exp\left(-\frac{\|x\|^2}{2}\right).$$

First consider the disk in the notes:

$$A_1 = \{x : (x_1 - 1)^2 + (x_2 - 1)^2 < 1\}.$$

- Generate two independent standard normal coordinates.
- Count the points inside the disk.
- The reference probability is approximately 0.18069.

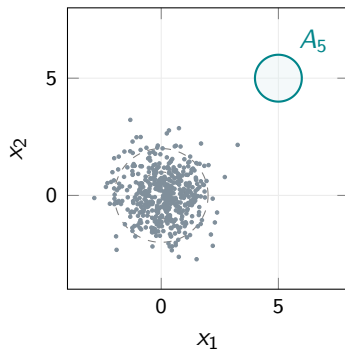
Direct simulation is quite effective here: the event is not particularly rare.

Move the disk far from the origin

$$A_5 = \{x : (x_1 - 5)^2 + (x_2 - 5)^2 < 1\}.$$

Reference: $P_A \approx 2.23465 \times 10^{-10}$.

- About one hit per 4.5 billion direct draws on average.
- A 10% relative SD needs roughly 4.5×10^{11} draws.



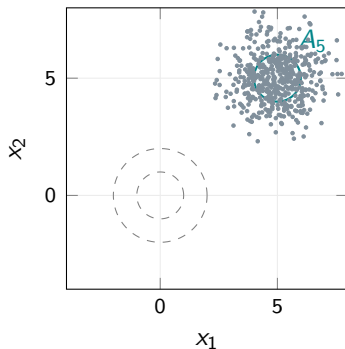
Shift the sampling distribution towards the event

Use $q_\mu = \mathcal{N}(\mu, I_2)$. For $\mu = (5, 5)$, the proposal is centred on the disk:

$$\mathbb{P}_{q_\mu}(A_5) = 1 - e^{-1/2} \approx 0.39347.$$

The event is now common.

Its original probability is still about 2.23×10^{-10} .



Derive the Gaussian likelihood ratio

With the same covariance I_2 ,

$$q_{\mu}(z) = \frac{1}{2\pi} \exp\left(-\frac{\|z - \mu\|^2}{2}\right).$$

Therefore

$$\begin{aligned} \frac{p(z)}{q_{\mu}(z)} &= \exp\left(\frac{\|z - \mu\|^2 - \|z\|^2}{2}\right) \\ &= \boxed{\exp\left(-\mu^{\top} z + \frac{\|\mu\|^2}{2}\right)}. \end{aligned}$$

Check: for $\mu = 0$, the weight is one.

The shifted Gaussian estimator

Draw $Z_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, I_2)$ and compute

$$H_i = \mathbf{1}\{\|Z_i - (5, 5)\|^2 < 1\} \exp\left(-\mu^\top Z_i + \frac{\|\mu\|^2}{2}\right).$$

$$\hat{P}_{\text{IS}} = \frac{1}{N} \sum_i H_i.$$

- Both Gaussian densities are positive everywhere: coverage holds.
- Hits in different parts of the disk have different weights.
- Unlike the coin example, a single constant does not weight all hits.

Executed experiment: more hits need not mean less error

$N = 100,000$ for each proposal $\mathcal{N}((m, m), I_2)$; NumPy seed 2026.

Reference $P_A = 2.23465 \times 10^{-10}$. Estimates, MCSEs and SDs below are in units of 10^{-10} .

m	Hits	Estimate	MCSE	Theoretical SD
0	0	0	0*	472.72
3	1,619	2.23655	0.07512	0.07542
4	17,848	2.17918	0.02991	0.03051
5	39,322	2.26741	0.03511	0.03483
6	18,062	2.11005	0.10115	0.11026

*All-zero outputs give zero empirical MCSE, which is uninformative here.

Compare weighted variability

$m = 5$ produces more hits than $m = 4$, but has a larger theoretical estimator SD. This does not establish that $m = 4$ is globally optimal.

Quantify uncertainty using the weighted outputs

For independent draws and a finite weighted second moment,

$$s_H^2 = \frac{1}{N-1} \sum_{i=1}^N (H_i - \hat{P}_{\text{IS}})^2, \quad \widehat{\text{MCSE}} = \frac{s_H}{\sqrt{N}}.$$

With sufficient sampling for a normal approximation,

$$\hat{P}_{\text{IS}} \pm 1.96 \widehat{\text{MCSE}}$$

is an approximate 95% confidence interval.

- Use the variability of H_i , not just the hit indicators.
- A few dominant weights or very few hits undermine confidence in the approximation.
- Assess efficiency at a comparable computational cost.

Checkpoint: what are we estimating?

Discuss briefly:

1. Why can direct MC have small absolute error but huge relative error?
2. Does the proposal hit rate estimate the original probability?
3. Why not always maximise the number of hits?
4. Should the importance sampling sum be divided by N or by the number of hits?

Answers: the target is tiny; no, it estimates $\mathbb{P}_q(A)$; weights also vary; divide by all N draws.

Lecture summary

- Probabilities are expectations of indicators: $P_A = \mathbb{E}_p[\mathbf{1}_A(X)]$.
- Direct relative SD is approximately $1/\sqrt{NP_A}$ for rare events.
- Importance sampling makes the event easier to observe and corrects with p/q .
- Multiply the per-flip ratios for a sequence; use the Gaussian ratio for a mean shift.
- Judge proposals by weighted error and cost, not by hit counts alone.

Next: Section 3.3

How can we choose proposal parameters systematically?