

A motivating integral

Consider

$$\mathcal{I} = \int_0^1 f(x) \, dx.$$

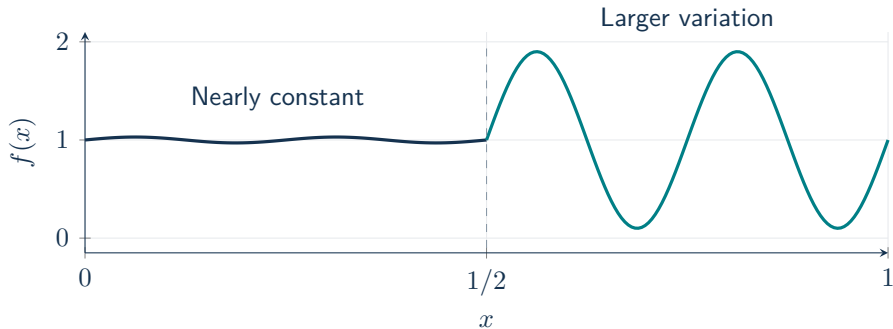
For $X \sim \text{Uniform}(0, 1)$, the density is one on the interval, hence

$$\mathcal{I} = \mathbb{E}[f(X)].$$

Ordinary Monte Carlo draws uniformly over the whole interval:

$$\hat{\mathcal{I}}_{\text{MC}} = \frac{1}{N} \sum_{i=1}^N f(X_i), \quad X_i \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1).$$

Different variability in different regions



Illustrative curve for the qualitative sketch in the script. The recording gives no exact function.

The variability of function values matters for Monte Carlo error.

Random counts and a planned allocation

With ten ordinary uniform samples,

$$\mathbb{E}[\text{number in the first half}] = 5, \quad \mathbb{E}[\text{number in the second half}] = 5.$$

- The realised counts are random, for example four and six.
- A nearly constant region may need relatively few evaluations.
- We might instead choose two samples in the first half and eight in the second.

Stratification lets us choose the counts in advance.

We will derive the conditions under which the allocation $2 + 8$ is optimal.

Splitting the integral

$$\mathcal{I} = \int_0^{1/2} f(x) \, dx + \int_{1/2}^1 f(x) \, dx.$$

Use separate uniform distributions:

$$X^{(1)} \sim p_1 = \text{Uniform}(0, 1/2), \quad X^{(2)} \sim p_2 = \text{Uniform}(1/2, 1).$$

Both intervals have length $1/2$, so their uniform densities equal two.

$$\mathbb{E}_{p_1}[f(X^{(1)})] = 2 \int_0^{1/2} f(x) \, dx,$$

$$\mathbb{E}_{p_2}[f(X^{(2)})] = 2 \int_{1/2}^1 f(x) \, dx.$$

The weights of the conditional means

Each integral is one half of its conditional expectation:

$$\int_0^{1/2} f(x) \, dx = \frac{1}{2} \mathbb{E}_{p_1}[f(X^{(1)})],$$

$$\int_{1/2}^1 f(x) \, dx = \frac{1}{2} \mathbb{E}_{p_2}[f(X^{(2)})].$$

Therefore,

$$\mathcal{I} = \frac{1}{2} \mathbb{E}_{p_1}[f(X^{(1)})] + \frac{1}{2} \mathbb{E}_{p_2}[f(X^{(2)})].$$

The two weights are the original probabilities of the two halves.

Two independent Monte Carlo calculations

Draw independent samples within each stratum, and independently across strata:

$$X_i^{(1)} \stackrel{\text{iid}}{\sim} p_1 \quad (i = 1, \dots, n_1), \quad X_i^{(2)} \stackrel{\text{iid}}{\sim} p_2 \quad (i = 1, \dots, n_2).$$

For fixed positive n_1, n_2 ,

$$\widehat{\mathcal{I}}_{\text{SS}} = \frac{1}{2n_1} \sum_{i=1}^{n_1} f(X_i^{(1)}) + \frac{1}{2n_2} \sum_{i=1}^{n_2} f(X_i^{(2)}).$$

$$\mathbb{E}[\widehat{\mathcal{I}}_{\text{SS}}] = \frac{1}{2} \mathbb{E}_{p_1}[f(X^{(1)})] + \frac{1}{2} \mathbb{E}_{p_2}[f(X^{(2)})] = \mathcal{I}.$$

The sample sizes can differ while the estimator remains unbiased.

A constant or nearly constant stratum

If $f(x) = a$ throughout the first half, then

$$f(X^{(1)}) = a \quad \text{almost surely,} \quad \text{Var}_{p_1}(f(X^{(1)})) = 0.$$

One evaluation recovers the conditional mean exactly. Its integral contribution is

$$\frac{1}{2} \mathbb{E}_{p_1}[f(X^{(1)})] = \frac{a}{2}.$$

- Nearly constant outputs have small within-stratum variance.
- A more variable stratum may require more samples.

Allocation and probability weights

For a budget of ten evaluations, one possible choice is

$$n_1 = 2, \quad n_2 = 8.$$

Let $\bar{f}_1$ and $\bar{f}_2$ be the two separate sample means. Then

$$\hat{\mathcal{I}}_{\text{SS}} = \frac{1}{2}\bar{f}_1 + \frac{1}{2}\bar{f}_2.$$

A pooled average of all ten values would instead be

$$\frac{2\bar{f}_1 + 8\bar{f}_2}{10} = 0.2\bar{f}_1 + 0.8\bar{f}_2.$$

Changing the sample allocation leaves the target weights unchanged.

The pooled average generally estimates a different quantity.

Several strata

- Divide the state space into strata.
- Estimate a conditional expectation within each stratum.
- Use separate sample sizes, for example ten in one stratum and a hundred in another.
- Combine the contributions using the original stratum probabilities.

The conditional distributions need not be uniform. The original distribution and the chosen partition determine them.

The partition and sample allocation are the design choices.

A partition of the state space

Let $X \sim p$ have state space Ω . A finite measurable partition satisfies

$$\Omega = \bigcup_{j=1}^J \Omega_j, \quad \Omega_j \cap \Omega_k = \emptyset \quad (j \neq k).$$

Ω_1	Ω_2
Ω_3	Ω_4

The strata cover the state space and do not overlap.

Boundaries with probability zero can be assigned to either neighbouring stratum.

Known stratum probabilities

For each stratum define

$$\pi_j = \mathbb{P}(X \in \Omega_j), \quad j = 1, \dots, J.$$

- We need these probabilities to be known or readily computable.
- Retain strata with $\pi_j > 0$. Probability-zero strata contribute nothing.
- The partition implies $\sum_{j=1}^J \pi_j = 1$.

Knowing p defines the probabilities, but does not always make them easy to compute.

Conditional distributions within strata

Within stratum j , sample from

$$\mathcal{L}(X \mid X \in \Omega_j).$$

If p is a density and $\pi_j > 0$, the conditional density is

$$p_j(x) = \frac{p(x)\mathbf{1}_{\Omega_j}(x)}{\pi_j}.$$

- The indicator restricts the density to Ω_j .
- The restricted mass is π_j .
- Dividing by π_j normalises the density to one.

Sampling from the conditional distribution

The basic method needs a practical way to generate samples from every retained stratum.

- A formula for p_j alone does not guarantee an efficient sampler.
- An elementary closed-form expression is not essential if a suitable sampling procedure is available.
- The chosen partition should make both stratum probabilities and conditional sampling tractable.

Conditional sampling can be the main implementation challenge.

The two conditional uniform distributions

For $X \sim \text{Uniform}(0, 1)$ and strata $[0, 1/2)$ and $[1/2, 1]$,

$$X \mid \{0 \leq X < 1/2\} \sim \text{Uniform}(0, 1/2),$$

$$X \mid \{1/2 \leq X \leq 1\} \sim \text{Uniform}(1/2, 1).$$

Both stratum probabilities are

$$\pi_1 = \pi_2 = \frac{1}{2}.$$

Both conditional densities equal two on their own intervals and zero elsewhere.

These are exactly the distributions used in the two separate averages.

The general stratified estimator

Write the conditional mean and variance of the *output* as

$$\mathcal{I}_j = \mathbb{E}_{p_j}[f(X)], \quad \sigma_j^2 = \text{Var}_{p_j}(f(X)).$$

The law of total expectation gives $\mathcal{I} = \sum_{j=1}^J \pi_j \mathcal{I}_j$. For fixed $n_j \geq 1$, draw independent samples within and across strata:

$$\bar{f}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} f(X_i^{(j)}), \quad X_i^{(j)} \stackrel{\text{iid}}{\sim} p_j.$$

Then

$$\boxed{\hat{\mathcal{I}}_{\text{SS}} = \sum_{j=1}^J \pi_j \bar{f}_j}, \quad \mathbb{E}[\hat{\mathcal{I}}_{\text{SS}}] = \sum_{j=1}^J \pi_j \mathcal{I}_j = \mathcal{I}.$$

Allocation changes precision. The probability weights stay fixed.

Variance of the stratified estimator

Within stratum j , independent sampling gives

$$\text{Var}(\bar{f}_j) = \frac{\sigma_j^2}{n_j}.$$

Independence across strata removes the covariance terms:

$$\begin{aligned}\text{Var}(\hat{\mathcal{I}}_{\text{SS}}) &= \text{Var}\left(\sum_{j=1}^J \pi_j \bar{f}_j\right) \\ &= \sum_{j=1}^J \pi_j^2 \text{Var}(\bar{f}_j) = \boxed{\sum_{j=1}^J \frac{\pi_j^2 \sigma_j^2}{n_j}}.\end{aligned}$$

Increasing n_j reduces stratum j 's contribution to the variance.

Proportional allocation: a natural baseline

Let $N = \sum_{j=1}^J n_j$ be the total number of function evaluations. Allocate samples in proportion to the stratum probabilities:

$$n_j = N\pi_j.$$

Substitution into the variance formula gives

$$V_{\text{prop}} = \sum_{j=1}^J \frac{\pi_j^2 \sigma_j^2}{N\pi_j} = \boxed{\frac{1}{N} \sum_{j=1}^J \pi_j \sigma_j^2}.$$

For two equally probable strata, this assigns $N/2$ samples to each.

Proportional allocation uses probability, but ignores variability.

For now, treat sample counts as continuous. Integer allocation comes later.

What proportional stratification removes

The law of total variance decomposes the ordinary output variance:

$$\text{Var}_p(f(X)) = \underbrace{\sum_j \pi_j \sigma_j^2}_{\text{within-stratum variation}} + \underbrace{\sum_j \pi_j (\mathcal{I}_j - \mathcal{I})^2}_{\text{between-stratum variation}}.$$

Ordinary Monte Carlo with N independent samples has variance

$$V_{\text{MC}} = \frac{1}{N} \text{Var}_p(f(X)).$$

Therefore,

$$V_{\text{MC}} - V_{\text{prop}} = \frac{1}{N} \sum_j \pi_j (\mathcal{I}_j - \mathcal{I})^2 \geq 0.$$

Fixed proportional counts remove the between-stratum contribution.

Optimal allocation: the optimisation problem

For a fixed partition, set $a_j = \pi_j^2 \sigma_j^2$. We want to solve

$$\min_{n_1, \dots, n_J > 0} F(n_1, \dots, n_J) = \sum_{j=1}^J \frac{a_j}{n_j} \quad \text{subject to} \quad \sum_{j=1}^J n_j = N.$$

- The probabilities π_j and variances σ_j^2 are fixed.
- The sample counts n_j are our decision variables.
- Assume $\pi_j > 0$, $\sigma_j > 0$, and equal evaluation costs.

We first solve the continuous problem, then choose integer counts.

Lagrange multipliers: the first-order equations

Use the same sign convention as the lecture notes:

$$L(n_1, \dots, n_J, \lambda) = \sum_{j=1}^J \frac{a_j}{n_j} + \lambda \left(N - \sum_{j=1}^J n_j \right).$$

For each j , differentiate and set the derivative to zero:

$$\frac{\partial L}{\partial n_j} = -\frac{a_j}{n_j^2} - \lambda = 0.$$

Thus $\lambda < 0$, and the positive solution is

$$n_j^2 = \frac{a_j}{-\lambda}, \quad \boxed{n_j = \frac{\sqrt{a_j}}{\sqrt{-\lambda}}}.$$

The budget determines the multiplier

Sum the first-order solution over all strata:

$$N = \sum_{j=1}^J n_j = \frac{\sum_{j=1}^J \sqrt{a_j}}{\sqrt{-\lambda}}.$$

Hence

$$\sqrt{-\lambda} = \frac{1}{N} \sum_{k=1}^J \sqrt{a_k}.$$

Substitute back and use $\sqrt{a_j} = \pi_j \sigma_j$:

$$n_j^* = N \frac{\sqrt{a_j}}{\sum_k \sqrt{a_k}} = N \frac{\pi_j \sigma_j}{\sum_k \pi_k \sigma_k}.$$

Optimal allocation is proportional to probability times standard deviation.

The resulting minimum variance

Write $S = \sum_{k=1}^J \pi_k \sigma_k$, so that $n_j^* = N \pi_j \sigma_j / S$. Substituting into the variance formula,

$$\begin{aligned} V_{\text{opt}} &= \sum_{j=1}^J \frac{\pi_j^2 \sigma_j^2}{N \pi_j \sigma_j / S} \\ &= \frac{S}{N} \sum_{j=1}^J \pi_j \sigma_j \\ &= \boxed{\frac{1}{N} \left(\sum_{j=1}^J \pi_j \sigma_j \right)^2}. \end{aligned}$$

Larger probability and larger output variability justify more samples.

Why this is a global minimum

For any positive allocation, Cauchy–Schwarz gives

$$\begin{aligned}\left(\sum_j \pi_j \sigma_j\right)^2 &= \left(\sum_j \frac{\pi_j \sigma_j}{\sqrt{n_j}} \sqrt{n_j}\right)^2 \\ &\leq \left(\sum_j \frac{\pi_j^2 \sigma_j^2}{n_j}\right) \left(\sum_j n_j\right) = N \operatorname{Var}(\hat{\mathcal{I}}_{\text{SS}}).\end{aligned}$$

Therefore every allocation satisfies

$$\operatorname{Var}(\hat{\mathcal{I}}_{\text{SS}}) \geq \frac{S^2}{N}.$$

Equality requires $\pi_j \sigma_j / \sqrt{n_j} = c \sqrt{n_j}$, so $n_j \propto \pi_j \sigma_j$.

When does optimal allocation improve on proportional?

Let $S = \sum_j \pi_j \sigma_j$. The difference is

$$\begin{aligned} V_{\text{prop}} - V_{\text{opt}} &= \frac{1}{N} \left[\sum_j \pi_j \sigma_j^2 - S^2 \right] \\ &= \frac{1}{N} \sum_j \pi_j (\sigma_j - S)^2 \geq 0. \end{aligned}$$

- Equal σ_j : proportional allocation is already optimal.
- Different σ_j : adjusting the allocation strictly improves variance.

$$\boxed{V_{\text{opt}} \leq V_{\text{prop}} \leq V_{\text{MC}}.}$$

A worked example: when two plus eight is optimal

Consider two equally probable strata with the following output variability:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \sigma_1 = 1, \quad \sigma_2 = 4, \quad N = 10.$$

Then $S = \frac{1}{2}(1) + \frac{1}{2}(4) = 2.5$, so

$$n_1^* = 10 \frac{0.5 \times 1}{2.5} = 2, \quad n_2^* = 10 \frac{0.5 \times 4}{2.5} = 8.$$

The estimator still uses the probability weights:

$$\widehat{\mathcal{I}}_{\text{SS}} = \frac{1}{2}\bar{f}_1 + \frac{1}{2}\bar{f}_2.$$

A standard-deviation ratio of 1 : 4 gives an allocation ratio of 1 : 4.

The variance reduction in the example

For $\sigma_1^2 = 1$ and $\sigma_2^2 = 16$,

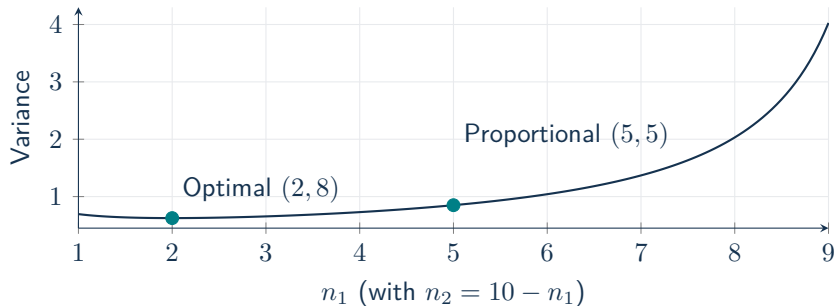
$$V(n_1, n_2) = \frac{0.25}{n_1} + \frac{4}{n_2}.$$

Allocation	(n_1, n_2)	Variance	Standard error
Proportional	(5, 5)	$0.05 + 0.80 = 0.85$	0.922
Optimal	(2, 8)	$0.125 + 0.50 = 0.625$	0.791

$$\text{Variance reduction} = 1 - \frac{0.625}{0.85} \approx 26.5\%.$$

Reallocating the same ten evaluations reduces variance by about a quarter.

The best allocation balances marginal gains



At the continuous optimum, $a_j/(n_j^*)^2$ is the same in every stratum.

The marginal variance reduction is equal across strata.

When the within-stratum variances are unknown

Use $m_j \geq 2$ pilot samples in each stratum to estimate

$$s_j^2 = \frac{1}{m_j - 1} \sum_{i=1}^{m_j} (f(X_{i,\text{pilot}}^{(j)}) - \bar{f}_{j,\text{pilot}})^2.$$

For a new production budget N , choose counts approximately as

$$n_j \approx N \frac{\pi_j s_j}{\sum_k \pi_k s_k}.$$

- Freeze the counts, then draw independent production samples.
- Retain a minimum count in each stratum, even when a pilot s_j is zero.

Pilot estimates guide effort. They do not guarantee the exact optimum.

If all $s_j = 0$, use proportional counts and collect more pilot data.