Unit 1: Monte Carlo Foundations

Lecture 2

Estimators, bias and variance

Mean squared error and the Monte Carlo idea

Learning outcomes

By the end of this lecture, you should be able to:

- distinguish a parameter, estimator and realised estimate;
- define mean squared error, bias and variance;
- derive the bias–variance decomposition of MSE;
- explain why unbiasedness alone does not establish accuracy;
- describe Monte Carlo estimation using simulated observations.

The central identity

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \mathrm{Bias}(\widehat{\theta})^2.$$

Recap: a sample and a target

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F, \qquad \mu = \mathbb{E}_F[X].$$

Random sample The collection $X_1, \ldots, X_n$.

Parameter A fixed feature of F, such as μ.

Statistic A specified function $T(X_1, \ldots, X_n)$ of the sample.

Pause and discuss

What is a natural statistic for estimating μ?

An estimator is a statistic with a purpose

A statistic used to approximate a parameter θ is an estimator of θ:

$$\widehat{\theta} = T(X_1, \ldots, X_n).$$

For the population mean, a natural choice is

$$\widehat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Estimator and estimate

Before sampling, $\bar{X}_3$ is random. For the data $(2, 4, 9)$, its realised value is $\bar{x}_3 = 5$.

The parameter being estimated must be specified.

Several candidates for estimating the mean

With $n \geq 2$, consider

$$T_A = \bar{X}_n, \qquad T_B = X_1, \qquad T_C = X_1 + X_2, \qquad T_D = 0.$$

All four are statistics. We can propose each as an estimator of μ.

Pause and discuss

Which would you expect to work well? What mathematical criterion would justify your preference?

Key point

A plausible choice needs a precise assessment of its error.

Mean squared error

For a fixed target θ, define

$$\boxed{\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2].}$$

1. Compute the error $\widehat{\theta} - \theta$.
2. Square it so opposite signs do not cancel.
3. Average over the randomness of the sample.

We assume $\mathbb{E}[\widehat{\theta}^2] < \infty$ for the finite decomposition below.

Our comparison criterion

Smaller MSE means smaller expected squared error at the parameter value under consideration.

The expectation is over repeated samples

Imagine independently repeating the same experiment:

$$\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}, \ldots, \widehat{\theta}^{(R)}.$$

If the benchmark θ is known, we can estimate the MSE by

$$\frac{1}{R} \sum_{r=1}^{R} \left(\widehat{\theta}^{(r)} - \theta\right)^2.$$

- The target stays fixed across repetitions.
- The sample and the estimate change.
- One unusually accurate run does not establish a low MSE.

Student question: what is fixed?

Object	Meaning	Status in this calculation
θ	True parameter	Fixed, possibly unknown
$\widehat{\theta}$	Estimator	Random
$m = \mathbb{E}[\widehat{\theta}]$	Mean of the estimator	Fixed at the chosen model

Key point

An unknown value can be fixed. An expectation of a random variable is a number once the model is specified.

Decomposition: add and subtract the mean

Write $m = \mathbb{E}[\widehat{\theta}]$. Then

$$\begin{aligned}\widehat{\theta} - \theta &= \widehat{\theta} - m + m - \theta \\ &= \underbrace{(\widehat{\theta} - m)}_{\text{centred random part}} + \underbrace{(m - \theta)}_{\text{fixed offset}}.\end{aligned}$$

- We have only added and subtracted the same number.
- The first part has expectation zero.
- The second part measures the displacement of the mean from the target.

Decomposition: expand the square

Using $(a+b)^2 = a^2 + 2ab + b^2$,

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}\left[((\widehat{\theta} - m) + (m - \theta))^2\right] \\ &= \mathbb{E}[(\widehat{\theta} - m)^2] \\ &\quad + 2\mathbb{E}[(\widehat{\theta} - m)(m - \theta)] \\ &\quad + \mathbb{E}[(m - \theta)^2].\end{aligned}$$

The middle expression is the expected cross term.

Pause and discuss

Which factors in that term are random, and which are constants?

In-class exercise: the cross term

Exercise

Let $m = \mathbb{E}[\widehat{\theta}]$. Show that

$$\mathbb{E}[(\widehat{\theta} - m)(m - \theta)] = 0.$$

Explain:

1. which factor can be taken outside the expectation;
2. which property of expectation you use;
3. whether the product itself must be zero for every sample.

Take 5 minutes. No independence assumption is needed.

Exercise solution: the cross term

Because $m - \theta$ is constant,

$$\begin{aligned}\mathbb{E}[(\widehat{\theta} - m)(m - \theta)] &= (m - \theta)\mathbb{E}[\widehat{\theta} - m] \\ &= (m - \theta)(\mathbb{E}[\widehat{\theta}] - m) \\ &= (m - \theta)(m - m) = 0.\end{aligned}$$

Key point

The cross term vanishes after taking expectation. It need not vanish for each realised sample.

The first term is the estimator's variance

By definition,

$$\boxed{\mathrm{Var}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2].}$$

It measures spread around the estimator's own mean.

A constant estimator

If $\widehat{\theta} \equiv c$, then $\mathrm{Var}(\widehat{\theta}) = 0$. Its squared error is still $(c - \theta)^2$.

Pause and discuss

Does zero variance guarantee a correct answer?

The second term is squared bias

Define the bias by

$$\boxed{\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta = m - \theta.}$$

Since $m - \theta$ is constant,

$$\mathbb{E}[(m-\theta)^2] = (m-\theta)^2 = \text{Bias}(\widehat{\theta})^2.$$

- Positive bias: overestimation on average.
- Negative bias: underestimation on average.
- Squared bias is always non-negative.

The bias–variance decomposition

Combining the terms gives

$$\boxed{\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \mathrm{Bias}(\widehat{\theta})^2.}$$

Contribution	What it measures
Variance	Spread around the estimator's mean
Squared bias	Squared displacement of that mean from the target

Key point

To understand total squared error, examine both contributions.

Systematic displacement and random variation

The error itself decomposes as

$$\widehat{\theta} - \theta = \underbrace{\widehat{\theta} - \mathbb{E}[\widehat{\theta}]}_{\text{random variation}} + \underbrace{\text{Bias}(\widehat{\theta})}_{\text{systematic displacement}} .$$

- Repeating a fixed procedure changes its realised random error.
- Repetition alone does not change that procedure's expected offset.
- Changing the procedure or sample size can change its bias.

Interpretation

A biased estimator can sometimes be very close to the target. An unbiased estimator can sometimes be far away.

Unbiased estimators

An estimator is unbiased for θ if

$$\mathbb{E}_\theta[\widehat{\theta}] = \theta$$

for every parameter value in the model. Equivalently, its bias is zero throughout the model.

Consequence for finite variance

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}).$$

Unbiasedness concerns the average over repeated samples.

A single observation can be unbiased

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with mean μ and variance $\sigma^2 < \infty$. For $T_B = X_1$,

$$\mathbb{E}[T_B] = \mu, \qquad \text{Bias}(T_B) = 0.$$

But

$$\text{Var}(T_B) = \sigma^2, \qquad \text{MSE}(T_B) = \sigma^2.$$

Key point

The estimator is unbiased, but it does not use the additional observations as n increases.

The sample mean uses all the observations

Under the same iid assumptions,

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu.$$

Independence removes the covariance terms, giving

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

Therefore

$$\mathrm{MSE}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

For $n > 1$ and $\sigma^2 > 0$, this is smaller than the MSE of X_1.

Comparing the candidate estimators

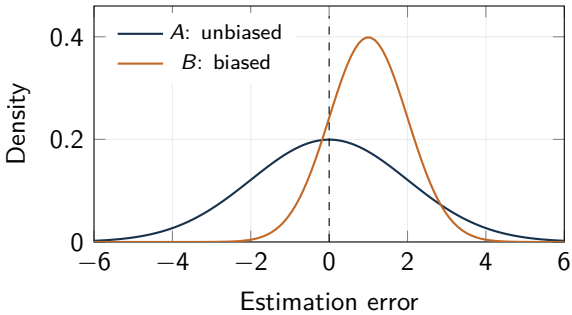
For iid observations with mean μ and variance σ^2, and $n \geq 2$:

Estimator of μ	Bias	Variance	MSE
$\bar{X}_n$	0	σ^2/n	σ^2/n
X_1	0	σ^2	σ^2
$X_1 + X_2$	μ	$2\sigma^2$	$2\sigma^2 + \mu^2$
0	$-\mu$	0	μ^2

Pause and discuss

Why is the bias of $X_1 + X_2$ equal to μ? Can the constant estimator ever have the smallest MSE?

An unbiased estimator need not have the lowest MSE



Illustrative sampling laws:

$$A - \theta \sim N(0, 4),$$
$$B - \theta \sim N(1, 1).$$

$$\mathrm{MSE}(A) = 4 + 0^2 = 4,$$
$$\mathrm{MSE}(B) = 1 + 1^2 = 2.$$

B has smaller MSE despite its bias.

In-class exercise: a numerical MSE comparison

Exercise

For this theoretical comparison, suppose

$$n = 4, \qquad \mu = 2, \qquad \sigma^2 = 9.$$

Compare the estimators $\bar{X}_4$, X_1 and 0.

1. Calculate the bias, variance and MSE of each.
2. Rank them by MSE.
3. Which conclusions would change if $\mu = 0$?

Work for 4 minutes. The specified μ is for assessing performance.

Exercise solution: a numerical MSE comparison

For $n = 4$, $\mu = 2$ and $\sigma^2 = 9$:

Estimator	Bias	Variance	MSE
$\bar{X}_4$	0	$9/4$	2.25
X_1	0	9	9
0	-2	0	4

The ranking is $\bar{X}_4$, then 0, then X_1.

If $\mu = 0$

The first two MSEs are unchanged, but the zero estimator has MSE zero. The preferred estimator can depend on the true parameter.

Returning to the expectation problem

Let $X \in \mathbb{R}^d$ follow a specified distribution, and let $f : \mathbb{R}^d \to \mathbb{R}$. Our target is

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

Define the scalar output

$$Y = f(X), \qquad I = \mathbb{E}[Y].$$

Key point

Computing the expectation is now a problem of estimating the mean of Y.

The missing ingredient is a sample

If we had iid observations $Y_1, \ldots, Y_N$, a natural estimator would be

$$\widehat{I}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i.$$

But the computational problem supplies a model and a function, rather than observed data.

The Monte Carlo idea

Use a computer to generate the random sample, then average the simulated outputs.

The realised average approximates the target expectation.

The basic Monte Carlo estimator

Assume we can generate $X_1, \ldots, X_N \overset{\text{iid}}{\sim} p$.

1. Draw an input X_i from the required distribution.
2. Calculate the scalar output $Y_i = f(X_i)$.
3. Average the outputs:

$$\boxed{\widehat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(X_i).}$$

Sampling assumption

Knowing the formula for a density does not automatically provide an efficient sampler. Here we assume independent draws are available.

Worked Monte Carlo setup: a square integral

Let $U \sim U(0,1)$ and $f(u) = u^2$.

$$I = \int_0^1 u^2 \, \mathrm{d}u = \mathbb{E}[U^2] = \frac{1}{3}.$$

Generate $U_1, \ldots, U_N \overset{\text{iid}}{\sim} U(0,1)$ and use

$$\widehat{I}_N = \frac{1}{N} \sum_i U_i^2.$$

An illustrative realised sample

For $(u_1, u_2, u_3, u_4) = (0.1, 0.4, 0.7, 0.9)$,

$$\widehat{i}_4 = \frac{0.01 + 0.16 + 0.49 + 0.81}{4} = 0.3675.$$

Lecture 2 summary

- An estimator is a statistic used to approximate a specified parameter.
- MSE averages squared estimation error over possible samples.
- $\text{MSE} = \text{Var} + \text{Bias}^2$ under finite second moments.
- The cross term vanishes in expectation.
- Zero bias alone does not ensure a small MSE.
- Monte Carlo generates data and averages $f(X_i)$ to estimate $\mathbb{E}[f(X)]$.

Pause and discuss

Which part of the error changes between runs? What does Monte Carlo supply when the original problem has no data?

Extension: unbiasedness of Monte Carlo

Assume $\mathbb{E}[|f(X)|] < \infty$ and every X_i has distribution p.

$$\begin{aligned}\mathbb{E}[\widehat{I}_N] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[f(X_i)] \\ &= \frac{1}{N}\sum_{i=1}^{N} I = I.\end{aligned}$$

Key point

Linearity of expectation proves unbiasedness. Independence is not required for this step.

Extension: covariance terms in a sample mean

For variables $Y_1, \ldots, Y_N$ with finite second moments,

$$\operatorname{Var}\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\left[\sum_i \operatorname{Var}(Y_i) + 2\sum_{i<j} \operatorname{Cov}(Y_i, Y_j)\right].$$

- The factor $1/N$ becomes $1/N^2$ when taking variance.
- Dependence enters through the covariance terms.
- Independence implies zero covariance when second moments exist.

Extension: variance and RMSE of Monte Carlo

Let $\sigma_f^2 = \mathrm{Var}_p(f(X)) < \infty$. For iid draws, the cross-covariances vanish:

$$\begin{aligned}\mathrm{Var}(\widehat{I}_N) &= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(f(X_i)) \\ &= \frac{N\sigma_f^2}{N^2} = \frac{\sigma_f^2}{N}.\end{aligned}$$

Unbiasedness now gives

$$\boxed{\mathrm{MSE}(\widehat{I}_N) = \frac{\sigma_f^2}{N}, \qquad \mathrm{RMSE}(\widehat{I}_N) = \frac{\sigma_f}{\sqrt{N}}.}$$

Extension: exact error for the square integral

For $X \sim U(0,1)$ and $f(X) = X^2$,

$$I = \frac{1}{3}, \qquad \mathbb{E}[f(X)^2] = \mathbb{E}[X^4] = \frac{1}{5}.$$

Thus

$$\sigma_f^2 = \frac{4}{45}, \qquad \mathrm{SD}(\widehat{I}_N) = \sqrt{\frac{4}{45N}}.$$

N	Variance of the estimator	SD of the estimator
100	8.89×10^{-4}	0.02981
1000	8.89×10^{-5}	0.00943
10000	8.89×10^{-6}	0.00298

Extension: estimating a probability

Let $q = \mathbb{P}(X \in A)$ and $Y_i = \mathbf{1}_A(X_i)$. Since $Y_i^2 = Y_i$,

$$\mathbb{E}[Y_i] = q, \qquad \mathrm{Var}(Y_i) = q - q^2 = q(1-q).$$

Therefore

$$\widehat{q}_N = \frac{1}{N}\sum_i Y_i, \qquad \mathrm{Var}(\widehat{q}_N) = \frac{q(1-q)}{N}.$$

Because $q(1-q) \le 1/4$, the standard deviation is at most $1/(2\sqrt{N})$.

Extension: relative error for rare events

For $0 < q < 1$, the relative standard deviation is

$$\frac{\mathrm{SD}(\widehat{q}_N)}{q} = \sqrt{\frac{1-q}{Nq}}.$$

For small q, this is approximately $1/\sqrt{Nq}$.

A probability of one in a million

If $q = 10^{-6}$ and $N = 10^6$, the expected number of events is $Nq = 1$ and the relative SD is approximately 100%. About 10^8 draws are required for a relative SD of 10%.

Key point

A small absolute error can coexist with a very large relative error.

Extension: repeated copies of one draw

Suppose $Y_i = Z$ for all i, with $\mathbb{E}[Z] = I$ and $\mathrm{Var}(Z) = \sigma_f^2$.

$$\widehat{I}_N = \frac{1}{N}\sum_i Z = Z, \qquad \mathrm{Var}(\widehat{I}_N) = \sigma_f^2.$$

The estimator is unbiased, but averaging creates no variance reduction.

Practical implication

The formula σ_f^2/N relies on the dependence assumptions. Correlated simulation outputs require a different error calculation.

Extension: shrinkage towards a fixed value

Let $\bar{X}_N$ estimate μ, with variance $v = \sigma^2/N$. Consider shrinking towards a fixed value c:

$$\widetilde{\mu} = a\bar{X}_N + (1-a)c, \qquad 0 \le a \le 1.$$

Then

$$\begin{aligned}
\text{Bias}(\widetilde{\mu}) &= (1-a)(c-\mu), \\
\text{Var}(\widetilde{\mu}) &= a^2 v, \\
\text{MSE}(\widetilde{\mu}) &= a^2 v + (1-a)^2(c-\mu)^2.
\end{aligned}$$

The variance decreases, but a bias term appears unless $c = \mu$ or $a = 1$.

Extension: shrinkage depends on the target

For illustration, suppose $v = 1$, $\mu = 0.2$, $c = 0$ and $a = 1/2$.

$$\text{Bias}(\widetilde{\mu}) = -0.1, \qquad \text{Var}(\widetilde{\mu}) = 0.25, \qquad \text{MSE}(\widetilde{\mu}) = 0.26.$$

The unbiased sample mean has MSE 1.

The comparison depends on the target

With the same v, c, a but $\mu = 3$, the shrinkage MSE is $0.25 + 2.25 = 2.5$.

There is no universal improvement in this example.

Extension exercise: the cube integral

Exercise

Let $U_i \overset{\text{iid}}{\sim} U(0,1)$ and $\widehat{I}_N = N^{-1} \sum_i U_i^3$.

1. Find the target I.
2. Show that the estimator is unbiased.
3. Compute $\text{Var}(U^3)$ and $\text{MSE}(\widehat{I}_N)$.
4. What happens to its RMSE when N increases from 100 to 400?

Work for 5 minutes. State the assumptions used at each step.

Extension exercise: solution

$$I = \mathbb{E}[U^3] = \frac{1}{4}, \qquad \mathbb{E}[U^6] = \frac{1}{7}.$$

Linearity of expectation gives $\mathbb{E}[\widehat{I}_N] = 1/4$.

$$\mathrm{Var}(U^3) = \frac{1}{7} - \frac{1}{16} = \frac{9}{112}.$$

Independence and unbiasedness give

$$\mathrm{MSE}(\widehat{I}_N) = \frac{9}{112N}, \qquad \mathrm{RMSE}(\widehat{I}_N) = \frac{3}{\sqrt{112N}}.$$

Multiplying N by four halves the RMSE.

Extension: assessing an estimator by simulation

If a benchmark target I is known and runs are independent,

$$\widehat{\mathrm{MSE}} = \frac{1}{R}\sum_{r=1}^{R}(\hat{I}_N^{(r)} - I)^2.$$

Its square root estimates RMSE.

- Increasing R improves the accuracy of this performance assessment.
- Increasing N improves the accuracy of each Monte Carlo estimate.
- If I is unknown, this direct benchmark calculation is unavailable.

Key point

The number of repetitions and the sample size within a run play different roles.