Unit 1: Monte Carlo Foundations

Lecture 1

Expectations and statistical estimation

The integration problem and the language of sampling

Learning outcomes

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

- formulate an expectation of a function of a random vector;
- explain why a tensor-product integration grid becomes expensive;
- distinguish a random sample from its observed values;
- distinguish a parameter, statistic, estimator and estimate;
- explain why estimating a mean is a statistical problem.

The question to keep in mind

What makes a sample-based approximation a good approximation?

A common computational problem

Many applications require an average under a probability distribution:

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

- In engineering, $f(X)$ may be the output of a model with uncertain inputs.
- In finance, $f(X)$ may be a payoff under a specified model.
- In Bayesian inference, $f(X)$ may be a quantity averaged under a posterior.

Key point

The target is simple to write down. Computing it can be difficult.

A random vector

Let X be a d-dimensional random vector:

$$X = (X^{(1)}, \ldots, X^{(d)}) \in \mathbb{R}^d.$$

- d is the number of coordinates in one outcome.
- Each coordinate $X^{(j)}$ is a scalar random variable.
- The coordinates need not be independent.

Two uncertain inputs

$X = (X^{(1)}, X^{(2)})$ could represent the temperature and pressure supplied to a physical model.

A real-valued function of the input

A function $f : \mathbb{R}^d \to \mathbb{R}$ maps a vector to a scalar. For $d = 2$, examples include

$$f(x_1, x_2) = x_1 x_2, \qquad f(x_1, x_2) = x_1 + x_2, \qquad f(x_1, x_2) = 3x_1.$$

- The input can have several coordinates.
- The output is a single real number.
- The function need not depend on every input coordinate.

Pause and discuss

For input $(2, 5)$, what are the outputs of these three functions?

The output is another random variable

Define

$$Y = f(X).$$

Before observing X, the value of Y is also random.

Squaring a uniform input

If $U \sim U(0,1)$ and $Y = U^2$, then an input $u = 0.4$ gives an output $y = 0.16$. The input and output have different distributions.

A function can change the type of distribution

For the same continuous input, $\mathbf{1}_{\{U>1/2\}}$ takes only the values zero and one. A transformed variable need not have a density.

The expectation we want to calculate

Our target is the fixed number

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

We assume $\mathbb{E}[|f(X)|] < \infty$, so the expectation is finite.

A benchmark with a known answer

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

$$I = \mathbb{E}[U^2] = \frac{1}{3}.$$

Pause and discuss

Is $\mathbb{E}[U^2]$ equal to $(\mathbb{E}[U])^2$?

Expectation as an integral

If X has a known density p, then

$$\boxed{I = \mathbb{E}[f(X)] = \int_{\mathbb{R}^d} f(x)p(x)\,\mathrm{d}x.}$$

- $f(x)$ is the quantity we want to average.
- $p(x)$ weights the possible inputs.
- Computing the integral computes the expectation.

Notation

The density $p(x)$ differs from a cumulative distribution function. For a scalar variable, the latter is $F(t) = \mathbb{P}(X \leq t)$.

Analytical integration

Sometimes we can calculate the integral exactly. For example,

$$\mathbb{E}[U^2] = \int_0^1 u^2 \, \mathrm{d}u = \left[\frac{u^3}{3} \right]_0^1 = \frac{1}{3}.$$

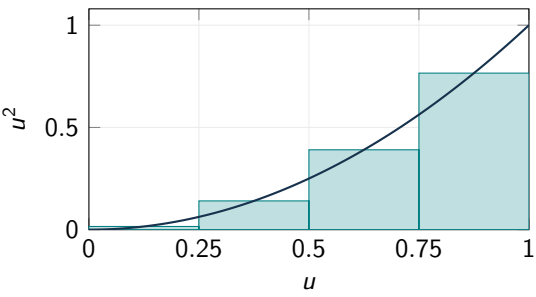
In more complicated applications:

- the function f may have no convenient antiderivative;
- the density p may be complicated;
- the integral may involve many coordinates.

Pause and discuss

If an exact calculation is unavailable, which numerical methods could we try?

Numerical integration with rectangles



Midpoint rule with m rectangles:

$$Q_m = h \sum_{j=1}^{m} g((j - \tfrac{1}{2})h), \qquad h = \frac{1}{m}.$$

For $g(u) = u^2$ and $m = 4$,

$$Q_4 = 0.328125, \qquad I = \frac{1}{3}.$$

The integral becomes a weighted sum of function values.

A tensor-product grid

With m grid points per coordinate in d dimensions,

$$\boxed{\text{number of grid points} = m^d.}$$

Dimension	Points per coordinate	Total points
1	10	10
2	10	100
3	10	1000
6	10	1000000
10	10	10000000000

Pause and discuss

What happens to the point count when we add one coordinate?

The curse of dimensionality

For a tensor-product grid, the number of evaluations grows exponentially with d.

An illustrative cost

If each evaluation takes one millisecond, 10^{10} evaluations take

$$10^7 \text{ seconds} \approx 116 \text{ days}$$

when performed sequentially, before other costs.

What the argument establishes

Straightforward tensor grids can become impractical. Holding coordinate resolution fixed does not guarantee fixed total error, and this is not a statement about every numerical integration method.

A statistical approach to the integration problem

Monte Carlo methods use random samples to approximate quantities such as

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

The central questions are:

1. How do we generate samples from the required distribution?
2. How do we use those samples to estimate I?
3. How do we assess the quality of the estimate?

Key point

Before studying the algorithm, we need the language of statistical estimation.

A random sample

Let F denote a probability distribution. If

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F,$$

then the collection is an **iid random sample of size** n from F.

- Each X_i has distribution F.
- The variables are mutually independent.
- The number n is the sample size.

For the statistics examples

We will initially take each observation X_i to be scalar.

The two parts of iid

Identically distributed Every draw follows the same probability law.

Independent The joint law factors into the individual laws.

For scalar observations, independence implies

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Same distribution without independence

Draw $Z \sim U(0,1)$ once and set $X_1 = \cdots = X_n = Z$. Each entry is uniform, but all entries repeat the same draw.

Dimension and sample size

Quantity	Meaning	Example
Dimension d	Coordinates in each observation	Temperature and pressure: $d = 2$
Sample size n	Number of observations	100 simulated input vectors: $n = 100$

A sample can consist of vectors:

$$X_1, \ldots, X_n \in \mathbb{R}^d.$$

Independence between sampled vectors does not require independence between coordinates within each vector.

Random observations and realised data

Before observing the sample	After observation	Role
X_1	$x_1 = 2$	First observation
X_2	$x_2 = 4$	Second observation
X_3	$x_3 = 9$	Third observation

- Upper-case symbols represent random variables in the model.
- Lower-case symbols represent the realised values.
- Another repetition could produce a different data set.

Key point

The sampling model describes possible data sets. A realised sample is one particular data set.

Parameters describe the distribution

A **parameter** is a fixed property of the underlying model or distribution.

$$\mu = \mathbb{E}_F[X], \qquad \sigma^2 = \mathrm{Var}_F(X).$$

Other features can include a median, a mode, or higher moments.

- The parameter belongs to the distribution, not to one realised sample.
- It may be unknown to us.
- We must check that the quantity exists and is well defined.

A specified distribution

For $X \sim U(0,1)$, the mean is the fixed number $\mu = 1/2$.

A statistic is a function of the sample

For a random sample $X_1, \ldots, X_n$, a statistic has the form

$$T = T(X_1, \ldots, X_n).$$

- Its rule is specified and evaluable from the data.
- Its calculation does not require an unknown model parameter.
- It may be scalar-valued or vector-valued.
- It need not depend on every observation.

Key point

A statistic is itself a random variable because it is a function of random observations.

Examples of statistics

Statistic	Formula	Observations used
Sample mean	$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$	All
Sum of squares	$Q = \sum_{i=1}^n X_i^2$	All
First observation	$T = X_1$	First only
Sum of first two	$T = X_1 + X_2$	First two, if $n \geq 2$

A distinction

The sum of squares is a statistic. It is not generally the sample variance.

Student question: is the sample mean random?

Question

If the observed data are fixed numbers, why do we say that a statistic is random?

Before observing the sample,

$$\bar{X}_3 = \frac{X_1 + X_2 + X_3}{3}$$

is a random variable. After observing $(x_1, x_2, x_3) = (2, 4, 9)$,

$$\bar{x}_3 = \frac{2 + 4 + 9}{3} = 5$$

is its realised value.

In-class exercise: identifying the objects

Suppose $X_1, \ldots, X_n$ are iid observations with an unknown finite mean μ.

Exercise

Classify each expression as a parameter, a statistic, or a realised value.

1. μ.
2. $\bar{X}_n = n^{-1} \sum_i X_i$.
3. $\bar{x}_3 = 5$, calculated from the data $(2, 4, 9)$.
4. X_1.

Is $\bar{X}_n - \mu$ a statistic when μ is unknown?

Discuss in pairs for 3 minutes.

Exercise solutions

1. μ is a **parameter**: the fixed population mean.
2. $\bar{X}_n$ is a **statistic**: a specified function of the sample.
3. $\bar{x}_3 = 5$ is a **realised value** of the statistic.
4. X_1 is a **statistic**, even though it ignores the other observations.

The error $\bar{X}_n - \mu$

It is a random quantity, but it is not a statistic under the usual definition when its calculation requires the unknown μ.

Statistical inference: estimating a parameter

Suppose observations come from an underlying distribution F:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F.$$

We want to estimate a fixed feature θ of F using the sample. Choose a statistic

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

for this purpose.

Estimator The statistic used to approximate θ.

Estimate Its numerical value for the observed data.

A natural estimator of the mean

For the target $\mu = \mathbb{E}_F[X]$, a natural choice is

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

One realised data set

For $(x_1, x_2, x_3) = (2, 4, 9)$, the estimate is $\widehat{\mu} = 5$. This calculation does not show that the true population mean equals five.

Other targets generally require other statistics. For example, estimating a population median suggests using a sample median.

Inference and Monte Carlo computation

	Statistical inference	Monte Carlo computation
Starting point	Observed data from an unknown model	A specified model and an expectation
Main unknown	A property of the distribution	The numerical value of an expectation
Sample	Observed data	Simulated draws

Both settings use statistics to approximate a fixed target and study variation across possible samples.

What makes an estimator good?

To estimate μ, we could consider

$$\widehat{\mu}_1 = \frac{1}{n}\sum_i X_i, \qquad \widehat{\mu}_2 = X_1, \qquad \widehat{\mu}_3 = 0.$$

All are statistics. How should we compare their use as estimators?

- Where are their sampling distributions centred?
- How much do their values vary across samples?
- How do their errors change with sample size?
- How much computation do they require?

Key point

A plausible formula still needs a mathematical assessment of its accuracy.

Lecture 1 summary

- Our computational target is an expectation $I = \mathbb{E}[f(X)]$.
- A density representation turns it into an integral.
- Tensor-product grids have m^d points and can be expensive.
- A parameter describes a distribution; a statistic is a function of data.
- An estimator is a statistic used to approximate a parameter.
- An estimate is the resulting number for a realised sample.

Next lecture

Bias, variance and mean squared error give us criteria for comparing estimators.

Extension: the basic Monte Carlo estimator

Assume we can generate independent draws $X_1, \ldots, X_N \overset{\text{iid}}{\sim} p$. For $I = \mathbb{E}_p[f(X)]$, define

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

1. Simulate each input X_i.
2. Evaluate the scalar output $Y_i = f(X_i)$.
3. Average the outputs.

Key point

Monte Carlo integration uses a sample mean to estimate an expectation.

Extension: uniform integration and scale factors

For $U \sim U(0,1)$,

$$\int_0^1 g(u)\,\mathrm{d}u = \mathbb{E}[g(U)].$$

For $X \sim U(a,b)$, the density is $1/(b-a)$ on the interval, so

$$\boxed{\int_a^b g(x)\,\mathrm{d}x = (b-a)\mathbb{E}[g(X)].}$$

A non-unit interval

To estimate $\int_2^5 x^2\,\mathrm{d}x$, generate $X_i \overset{\text{iid}}{\sim} U(2,5)$ and return

$$\frac{3}{N}\sum_i X_i^2.$$

Extension: probabilities are expectations

For an event A, define

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

Then

$$\mathbb{P}(X \in A) = \mathbb{E}[\mathbf{1}_A(X)].$$

A normal tail probability

For $Z_i \overset{\text{iid}}{\sim} N(0,1)$, estimate $q = \mathbb{P}(Z > 1.5)$ by

$$\widehat{q}_N = \frac{1}{N} \sum_i \mathbf{1}_{\{Z_i > 1.5\}}.$$

The estimator is the fraction of simulated values exceeding the threshold.

Extension exercise: setting up an estimator

Exercise

For each target, specify the sampling distribution, the output function and the estimator.

1. $A = \int_0^1 u^3 \, \mathrm{d}u$.
2. $B = \int_2^5 x^2 \, \mathrm{d}x$.
3. $C = \mathbb{P}(U^2 > 1/2)$ for $U \sim U(0, 1)$.

Which target is a probability, and why can we use the same averaging principle?

Allow 4 minutes in an extended class or lab.

Extension exercise: solutions

1. $U_i \overset{\text{iid}}{\sim} U(0,1)$:
$$\widehat{A}_N = \frac{1}{N}\sum_i U_i^3, \qquad A = \frac{1}{4}.$$
2. $X_i \overset{\text{iid}}{\sim} U(2,5)$:
$$\widehat{B}_N = \frac{3}{N}\sum_i X_i^2, \qquad B = 39.$$
3. $U_i \overset{\text{iid}}{\sim} U(0,1)$:
$$\widehat{C}_N = \frac{1}{N}\sum_i \mathbf{1}_{\{U_i^2 > 1/2\}}, \qquad C = 1 - \frac{1}{\sqrt{2}}.$$

A probability is an expectation of an indicator, so sample averaging applies.

Extension: a first Python implementation

```
import numpy as np

rng = np.random.default_rng(2026)
N = 1000
u = rng.uniform(0.0, 1.0, size=N)
y = u**2
estimate = y.mean()
print(estimate)
print("Benchmark:", 1/3)
```

- Average the squared inputs, rather than square their average.
- Change the seed to see another realised estimate.
- A reproducible result still needs an assessment of simulation error.