Statistical error of Monte Carlo estimates

Now that you've had some exposure to Monte Carlo simulation, we turn to the issue of the size of the variability in estimates from one run to another and how this varies with the number of repetitions of the random experiment in each run. This variability is called the statistical error in the estimate.

It is different from other kinds of error we shall meet shortly, which are those relevant for deterministic problems. In those cases, the answer is the same every time we run the calculation — but it's not the correct answer! Such errors are also present in stochastic estimates but they are dominated by the statistical errors (otherwise we wouldn't regard them as stochastic).

We call each simulated random experiment an instance or realization. The whole simulation consists of repetitions of such instances. Finally, we can run the simulation many times.

Variability of random samples

To explain how Monte Carlo estimates vary, we model the estimates from each repetition as observations of some quantity derived from the same random experiment, with each observation being independent of the rest.

We now treat the observations from each repetition in the simulation just as one would treat observations of an actual experiment using the tools of statistics. Since typically it is easy to generate many repetitions in a simulation, it will be sufficient to use statistical methods suitable for large sample sizes. This will allow us to construct confidence intervals i.e. error bars, by using the power of the Central Limit Theorem.

We refer to the document 'StatsNotes.pdf' for details and proofs, but describe the key ideas here.

Random variables

In any simulation, we could examine many possible properties of interest. A random variable is a way of summarizing the outcome of a random experiment so we can focus on some particular aspect of it.

◁ **Example:** Assume we throw two dice. There are 36 outcomes in the sample space but we may only be interested in the total of the two faces showing, or the maximum value showing or just in whether we threw a 'double' or not. So we can define 3 different random variables acting on the same sample space:

- $X_1(\omega) =$ total of two faces showing
- $X_2(\omega) =$ maximum face showing
- $X_3(\omega) = 1$ if ω is a double, 0 otherwise.

X_3 is called an indicator random variable because it just indicates whether a 'double' has occurred.

Definition of a random variable

A **random variable** is a function from the sample space Ω of a random experiment to the real numbers $\mathbb{R}$. Random variables are usually denoted by upper case letters $X, Y, Z, \ldots$. So if $X : \Omega \to \mathbb{R}$ is a random variable, X assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

◁ **Example:** Consider the random experiment of tossing a coin 3 times and observing the sequence of results. If $Y =$ the number of heads obtained, then Y is a random variable.

ω	$Y(\omega)$
hhh	3
thh	2
hth	2
hht	2
htt	1
tht	1
tth	1
ttt	0

Discrete versus continuous

In this example, the range of Y is $Y(\Omega) = \{0, 1, 2, 3\}$. This is an example of a discrete random variable, which means that its values are only a finite or countably infinite set. Usually this will happen when the sample space itself is a finite or countably infinite set.

If the sample space is uncountably large and the set of values of X is an interval $(a, b) \subseteq \mathbb{R}$, then X is a continuous random variable.

$\triangleleft$ **Example:** Some common discrete random variables: Bernoulli, binomial, geometric, Poisson

$\triangleleft$ **Example:** The most common continuous random variable: normal.

Probability mass function

Since a random variable X assigns to each outcome in Ω a number, and each outcome has some probability, it follows that each value in the range of X has a probability. The **probability distribution** of a discrete random variable X is defined by its **probability mass function** (pmf)

$$p_X(x) = Pr(X = x).$$

It has the following properties:

- $p_X(x)$ is non-zero at only a finite or countably infinite set of x values, say either $x_1, x_2, \ldots, x_n$ or $x_1, x_2, \ldots$;
- $p_X(x) \geq 0$;
- $\sum_x p_X(x) = 1$.

where the sum is over all possible values of x.
Any function satisfying these conditions is said to be a probability mass function.

Continuous random variables have a similar function, called the probability density function, but where we integrate over a range of possible values.

Independent random variables

We can extend the idea of independence from events to random variables.

Random variables X and Y are *independent* if any event defined using X is independent of any event defined using Y
i.e. for any sets A and B, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.
$\triangleleft$ **Example:** for independent experiments such a tossing a coin and rolling a die at the same time, any random variable describing the coin toss will be independent of any random variable describing the die roll.

Now consider repeated random experiments of the same kind that are independent of each other
$\triangleleft$ **Example:** sampling from a large population with replacement (so the population is identical at each sampling)

Random sample from a distribution

We say a random sample from a distribution is a sequence of mutually independent random variables $X_1, X_2 \ldots X_n$, with the same distribution. Also called an independent identically distributed (iid) sequence.

Given a random sample, we can estimate the pmf by

$$\hat{p}(x) = \frac{|X_i = x|}{n}$$

We hope that $\hat{p}(x) \to p_X(x)$ as the sample size $n \to \infty$ in accordance with our notion of probability as a long-run frequency.

◁ **Example:** When you simulate your random experiment n times (n repetitions), you are generating a random sample of size n from a population of possible simulation runs the size of the period of the random number generator. So, unless you run your simulation an awful lot or your random number generator is poor, you can safely regard your repetitions as independent observations.

Expectation

Faced with the varying results from many instances of a random experiment, we often want to know the average behaviour. This is given by the expectation or expected value of a random variable.

Let X be a discrete random variable with set of possible values D and pmf $p_X(x)$. Then the **expected value** or **mean value** of X denoted by

$$\mathrm{E}(X) = \mu_X = \sum_{x \in D} x \cdot p_X(x)$$

Thus $\mathrm{E}(X)$ is the weighted sum of the values of X where the weight of x is its probability $p_X(x)$. It represents the centre of mass of the probability distribution.

$\triangleleft$ **Example:** If Y is the number of heads in three coin tosses,

$$\mathrm{E}(Y) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Expectation is linear

Generally if X is a discrete random variable with pmf $p_X(x)$ and $h(x)$ is any function then the **expected value of** the discrete random variable $h(X)$ is

$$\mathrm{E}[h(X)] = \sum_{x \in D} h(x) \cdot p_X(x).$$

If the function h is *linear* i.e. $h(x) = ax + b$, then we have the special property that

$$E(h(X)) = h(E(X))$$

but this is usually *not true*!

If $a, b \in \mathbb{R}$ are constants

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

Variance is not linear

We also want to know how much spread there is about the mean μ.

The **variance** of X denoted $\mathrm{var}(X)$ or σ_X^2 or just σ^2 provides a measure of *spread* or variability or dispersion. It is defined by

$$\mathrm{var}(X) = \sum_{x\in\Omega}(x-\mu)^2 \cdot p_X(x) = E((X-\mu)^2)$$

and measures how close is the distribution to its mean.

If all possible values of X are near μ then $\mathrm{var}(X)$ is small, while if the spread is large so is $\mathrm{var}(X)$.

$$\boxed{\mathrm{var}(aX+b) = a^2 \cdot \sigma_X^2 = a^2\mathrm{var}(X)}$$

Notice that $\mathrm{var}(X+b) = \mathrm{var}(X)$ reflecting the fact that the variance is unchanged by a simple translation.
So $\mathrm{var}(aX+b)$ is quadratic in a but not a function of b.

Standard deviation

The **standard deviation** of X is

$$\sigma = \sigma_X = \sqrt{\mathrm{var}(X)} = \sqrt{\sigma_X^2}.$$

We often use the standard deviation as a scale for the variation in X since it has the same units as X.

From the properties of the variance, it follows that

$$\sigma_{aX+b} = |a| \cdot \sigma_X.$$

We will use a suitable standard deviation to measure the statistical error of our Monte Carlo estimates.

Statistics

Consider taking a random sample of size n from a population. Before the sample we are uncertain about what the value of each of the n observations will be. The first observation must be considered as a random variable X_1, the second observation another random variable X_2 and so on.

After taking a sample, each observation is a number

$$x_1, x_2, \ldots, x_n$$

These numbers are the **observed values** of the X_i. We assume the population distribution is not known and we want to make inferences about this distribution.

We next have to decide what *information* about the population we want and how to find it from some *function* of the random variables $X_1 \ldots X_n$ — called a **statistic**. A statistic that is used to estimate a parameter or characteristic of a population is called an **estimator**. We use the notation $\hat{\theta}$ to represent a statistic which estimates a parameter θ.

Estimators

For a statistic $\hat{\theta}$ to estimate θ to be useful, it should be:

1. accurate
   the expected value of the statistic $E(\hat{\theta})$ should be close to the parameter θ
2. precise
   $\mathrm{var}(\hat{\theta})$ not too big

In particular,

$$\boxed{\text{an estimator } \hat{\theta} \text{ is } \textbf{unbiased} \text{ if } E(\hat{\theta}) = \theta}$$

Usually we prefer unbiased estimators unless this causes a big price to be paid in terms of $\mathrm{var}(\hat{\theta})$.

In the context of simulations, each random experiment is an observation and the relevant property of each instance is an observed value.

Sample mean

An important statistic in simulations is the sample mean. The **sample mean** is the random variable defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is a natural **estimator** $\hat{\mu}$ of the population mean μ.

After the sample is taken the corresponding **point estimate** for the population mean is the single number

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

which is the **observed value** of $\overline{X}$ in the particular sample.

The sample mean is generally how we compute the average behaviour of a random experiment, by combining the observed values from each instance.

The sample mean is unbiased

The important result:

$$\boxed{E(X+Y) = E(X) + E(Y)}$$

is true for *any* random variables, discrete or continuous, provided the expected values exist. Combining this with the linearity result shows that

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n).$$

If each $E(X_i) = \mu$ (as is the case for a random sample), then

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

So the expected value of the sample mean (an estimator of the parameter μ) is just μ itself.

The sample mean is an unbiased estimator of the population mean.

But how precise is it?

Variance of the sample mean

For independent random variables X, Y

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$$

Now we can find $\text{var}(\overline{X})$.

$$\text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_X^2 = \frac{1}{n}\sigma_X^2$$

If $X_1, \ldots, X_n$ is a random sample (independent and identically distributed) from a distribution with variance σ_X^2, then

$$\text{var}(\overline{X}) = \frac{1}{n}\sigma_X^2$$

This means that the variability of the sample mean *falls* as the sample size increases. A large sample size gives a more precise estimate (as you would expect).

Sample proportion

An important special case is where the random variables X_i are indicator variables of some property of the population. Then $E(X_i) = p$, the probability of observing that property in the population. The sample mean of these indicator variables is just the **sample proportion** $\hat{P}$ (the proportion of trials with the property in question).

As above, we get

$$E(\hat{P}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}(np) = p$$

The sample proportion is an unbiased estimator of the population proportion (probability).

Variance of sample proportion

The sample proportion is special because we know everything about its variance, just by knowing its mean!

Since $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where the X_i are Bernoulli random variables:

$$\mathrm{var}(\hat{P}) = \mathrm{var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n}$$

If $X_1, \ldots, X_n$ is a random sample of some property with probability p, then

$$\mathrm{var}(\hat{P}) = \frac{p(1-p)}{n}$$

Again, the variability of the sample proportion *falls* as the sample size increases. This explains qualitatively the observations from Lecture 5.

Standard errors

The standard deviation of an estimator is also called the **standard error**. It is the standard error that gives a measure of the statistical error.

◁ **Example:** The sample mean has standard error $\frac{\sigma_X}{\sqrt{n}}$
Note that we don't know σ_X without further work.

◁ **Example:** The sample proportion has standard error $\sqrt{\frac{p(1-p)}{n}}$
We know this once we have an estimate for p.

To get a more quantitative description, we must know the distribution of the estimator.

Central Limit Theorem

This is where we use large sample statistics — if the random sample is large enough, then any sum of random variables is well-approximated by a normal random variable. The sample mean (and sample proportion) are special cases.

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. If n is sufficiently large, then $\overline{X}$ has approximately a **normal distribution** with mean μ and variance σ^2/n.

Thus approximately $\overline{X} \stackrel{d}{=} N(\mu, \sigma^2/n)$ for large n, *no matter what the distribution of* X *!!*.

This remarkable result explains the widespread occurrence of the normal distribution in many different circumstances — any random variable that arises as the sum of many independent random variables is approximately normally distributed.

Confidence intervals

Suppose we have an estimator $\hat{\theta}$ of a parameter θ with a *known* sampling distribution. Then we can find a *random interval* (L, U) which has a fixed probability $1 - \alpha$ of including the fixed but unknown parameter θ.

$$Pr(\theta \in [L, U]) = 1 - \alpha$$

We call (L, U) the $100(1 - \alpha)$% confidence interval (CI).

We will take $\alpha = 0.05$ i.e. a 95% (two-sided) confidence interval.
◁ **Example:** For a 95% CI, we want $Pr(\hat{\theta} < L) = 0.025$ and $Pr(\hat{\theta} < U) = 0.975$. L and U are each random variables

Coverage

◁ **Example:** For a normal RV with variance σ^2, $L = \mu - 1.96\sigma$ and $U = \mu + 1.96\sigma$

Now we take a sample $\{x_1 \ldots x_n\}$ to find an observed value of θ and hence observed values of L and U — call them l and u.
We also call the interval (l,u) a $100(1-\alpha)\%$ confidence interval (CI) for θ BUT IT IS NOT TRUE that

$$Pr(l < \theta < u) = 1 - \alpha$$

since θ is fixed — either θ is in (l,u) or it isn't!
But if you repeated this process many times, the set of intervals (l_i, u_i) would contain θ about $100(1-\alpha)\%$ of the time.

Confidence interval for the sample proportion

For n large enough we have

$$\hat{P} \overset{d}{\sim} N(p, \frac{p(1-p)}{n})$$

Then the **approximate 95% confidence interval for** p is

$$\left(\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; , \; \hat{p} + 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) .$$

where $\hat{p}$ is the observed value of $\hat{P}$.

$\triangleleft$ **Example:** Confidence intervals for de Méré's bet.

Confidence interval for the mean

The CLT tells us that approximately $\overline{X} \overset{d}{=} N(\mu, \sigma^2/n)$ for large n.
If don't know the variance σ^2, we will have to estimate the variance of X using an estimator S^2, the **sample variance**. Then for large n,

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{d}{\sim} N(0, 1).$$

In this circumstance we can use the observed values $\overline{x}$ and s of $\overline{X}$ and S to get confidence intervals. For example,

for a two-sided 95% confidence interval we would have

$$\left(\overline{x} - 1.96 \cdot \frac{s}{\sqrt{n}} \;,\; \overline{x} + 1.96 \cdot \frac{s}{\sqrt{n}}\right).$$

So to construct a CI for the mean, we need an estimate of the sample variance S^2.

Sample variance

The **sample variance** is defined as

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{n=1}^{n}(X_i-\overline{X})^2 \\
&= \frac{1}{n-1}\left[\sum_{n=1}^{n}X_i^2 - n\overline{X}^2\right].
\end{aligned}
$$

The sample variance S^2 is an estimator $\hat{\sigma}^2$ for the population variance σ^2. Again after the sample is taken the observed value of S^2 is the single number given by

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{n=1}^{n}(x_i-\overline{x})^2 \\
&= \frac{1}{n-1}\left[\sum_{n=1}^{n}x_i^2 - n\overline{x}^2\right].
\end{aligned}
$$

which is the point estimate of σ^2.

Sample variance

$$E(S^2) = \frac{1}{n-1}\left[n\sigma_X^2 - n\text{var}(\overline{X})\right] = \frac{1}{n-1}\left[n\sigma_X^2 - n \cdot \frac{1}{n}\sigma_X^2\right] = \sigma_X^2$$

Hence

$\boxed{S^2 \text{ is an unbiased estimator of } \sigma^2.}$

Note: the divisor $n-1$ is used in calculating the sample variance since it gives an unbiased estimate of σ^2. We have "used up" one of the members of the sample in calculating the sample mean which replaced the population mean in the formula for variance.

The **sample standard deviation** $S = \sqrt{S^2}$ is an estimator for the population standard deviation σ. Note: S is *not* an unbiased estimator of σ, but often the bias is small.

End of Lecture 6