Error Analysis in Direct Monte Carlo

Error Propagation and CLT

Sachin Shanbhag

Department of Scientific Computing
Florida State University,
Tallahassee, FL 32306.

Contents

We will take a closer look at error analysis in direct Monte Carlo

$$\text{error} \sim \frac{1}{\sqrt{n}},$$

and its link to the CLT.

We will also introduce the idea of importance sampling which can be thought of as a way decrease the rejection rate in accept/reject sampling.

- Propagation of Error
- Central Limit Theorem
- Error in Direct Monte Carlo Integration
- Importance Sampling: Variance Reduction

Propagation of Error

Consider independent variables x and y with errors σ_x and σ_y.

- if $z = x + y$, or $z = x - y$, then the error in z is given by:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2. \tag{1}$$

- if $z = xy$, or $z = x/y$ then

$$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2. \tag{2}$$

This has the appearance of Pythogorean triples.

- add/subtract: absolute error
- multiply/divide: relative error

Propagation of Error

If $z = f(x)$, then

$$\sigma_z = |f'(x)|\sigma_x. \tag{3}$$

Q1: Volume $V = s^3$, with $s = 2.00 \pm 0.02$ cm. What is the uncertainty in volume?
A1: $\sigma_V = |3s^2|\sigma_s$. Therefore, $V = 8 \pm 0.2$ cm^3.

Q2: $x = 100 \pm 6$, and $z = \sqrt{x}$. What is σ_z?
A2: $\sigma_z = \sigma_x / 2\sqrt{x} =$ 0.3

Q3: Write $x = z \times z$ to find the error in x using the "product rule".

Q4: If $x = 1 + 0.1$, $y = 3 + 0.2$. What is the error in $x - y$ and x/y?

General Formula

If x_1, x_2, ..., x_n are independent random variables, and

$$q = q(x_1, x_2, ..., x_n),$$

then the error in q,[1]

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x_1}\sigma_{x_1}\right)^2 + \left(\frac{\partial q}{\partial x_2}\sigma_{x_2}\right)^2 + ... + \left(\frac{\partial q}{\partial x_n}\sigma_{x_n}\right)^2. \quad (4)$$

Exercise: Verify that the eqns 1-3 above are special cases of eqn 4.

This formula is a linear approximation, and is guaranteed to work well when the errors are "small".

[1] derivation at the end of the lecture notes

Average of Independent Observations

Consider the average of a number of independent and identically distributed (iid) samples/observations,

$$z = \frac{x_1 + x_2 + ... + x_n}{n}.$$

That is, $\sigma_{x_i}^2 = \sigma_x^2$ for all x_i.

Using the general formula above, we get:

$$\sigma_z = \sigma_x \sqrt{\frac{1}{n^2} + \frac{1}{n^2} + ... \frac{1}{n^2}} = \frac{\sigma_x}{\sqrt{n}}. \tag{5}$$

The error of the average is smaller than the error in an individual observation, due to cancellation of errors.

Central Limit Theorem

Suppose n iid random variables $X_i \sim \pi(x)$, $1 \leq i \leq n$.

Let $E_\pi[x] = \langle x \rangle$ and σ_x^2 be the mean and variance of $\pi(x)$.

Further, let

$$Z = \frac{X_1 + X_2 + ... + X_n}{n},$$

be the sample average or mean.

We know from the general formula for error propagation of iid random variables that, $\sigma_z^2 = \sigma_x^2/n$.

- $\sigma_z^2 \to 0$, as $n \to \infty$, and hence,
- the sample mean Z approaches the true mean $\langle x \rangle$.

Central Limit Theorem

The CLT tells us something more about Z, in addition to the first and second moments of $\pi(z)$ (the mean and variance, respectively).

It tells us something about the entire probability distribution Z, provided n is sufficiently large, for its effects to kick in.

Informal Statement

As n increases, the distribution of Z, say $g_Z(z)$ approaches a normal distribution

In the limit, $n \to \infty$, the distribution $g_Z(z) \to \mathcal{N}\left(\langle x \rangle, \sigma_z^2\right)$.

Note that the distribution $\pi(x)$ can be anything: discrete, continuous, whatever!

CLT

What is this "distribution of Z" business?

Note that the average of n random variables is itself a random variable (with tighter variation).

That is, if we repeat the "experiment" of calculating the sample average two different times, we don't expect the answers to be identical.

Let's pick an example to illustrate the CLT.

Dice Example

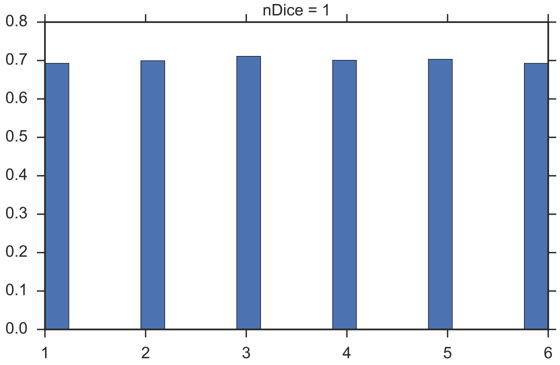
Suppose we roll n dice ($x_i \sim U[1,6]$) and define Z as the sample average. What is the distribution of Z?

We can try to do an analytical calculation to figure out $g(z)$, but let's just do a MC simulation.

Dice Example

```
def avgDiceRoll(nDice):
    """nDice = n in the description above"""

    nSamples = 50000 # num of indep experiments
    Z = np.zeros((nSamples))

    for i in range(nSamples):
        Z[i] = np.mean(np.random.randint(1,7,nDice))

    # histogram
    n, bins, patches = plt.hist(Z, 21, density=True)
    plt.xlim(1,6)
    plt.title('nDice = {0:d}'.format(nDice))

# one die
avgDiceRoll(1)
```

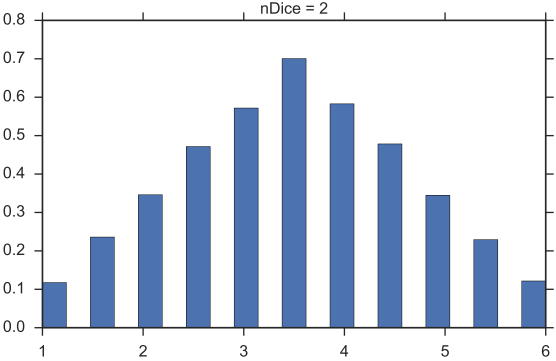
Dice Example



As n (=nDice) increases, observe how an initially flat distribution, increasingly starts to peak at its mean.

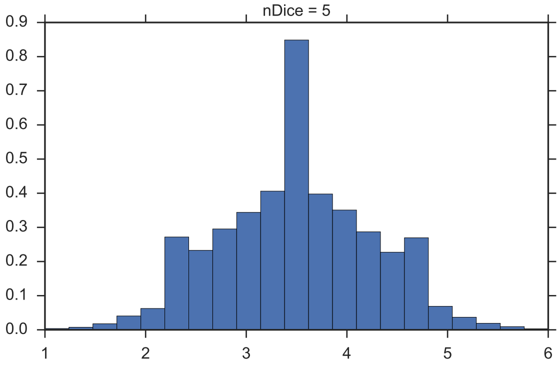
Dice Example

```
z = avgDiceRoll(2)
```



Dice Example

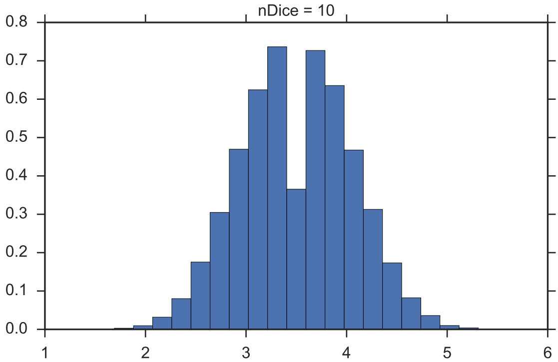
```
z = avgDiceRoll(5)
```



Also note that the distribution gets increasingly narrow.

Dice Example

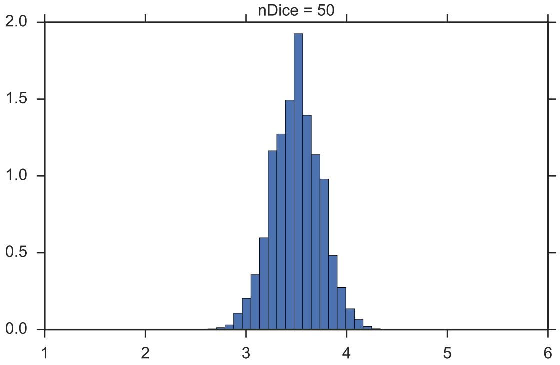
```
z = avgDiceRoll(10)
```



For "large" n the distribution looks unmistakably Gaussian.

Dice Example

```
z = avgDiceRoll(50)
```



The distribution starts looking like a continuous distribution.

Insight

MC "integration by darts"

We compute the integral by estimating the mean or average value of the integrand over the domain.

We draw a large number of samples. Hence, we can trust the law of large numbers to apply.

Equipped with the CLT, we will now turn our attention to a more careful analysis of error in direct Monte Carlo.

Error in Direct Monte Carlo

Consider a 1D integral,

$$\begin{aligned} I &= \int_a^b f(x)dx \\ &\approx \sum_{i=1}^{n} f(x_i)\Delta x \\ &\approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) \\ &\approx (b-a)\bar{f} \end{aligned}$$

If we use integration by n darts, and note that the f_i are iid random variables with variance σ_f^2, what is the variance σ_I^2?

Error in Direct Monte Carlo

$$\begin{aligned}\sigma_I^2 &= (b-a)^2 \frac{\sigma_f^2}{n} \\ &= (b-a)^2 \frac{\langle f^2 \rangle - \langle f \rangle^2}{n}\end{aligned}$$

Notice that the variance in the integral is proportional to:

(i) the variance of the integrand $f(x)$

(ii) the square of the domain size:

(iii) inversely proportional to n

Error in Direct Monte Carlo

In general, for integrals of any dimension:

$$\sigma_I^2 = \frac{V^2}{n}\sigma_f^2, \text{ where } V = \int d\mathbf{x}$$

Practical Implication

- In MC, we evaluate the average of f_i to get $\bar{f}$
- Similarly, we can estimate the variance of f simply by keeping track of $\bar{f^2}$.
- This will let us put error-bounds on any single direct MC estimate

Let us see this in action!

Example

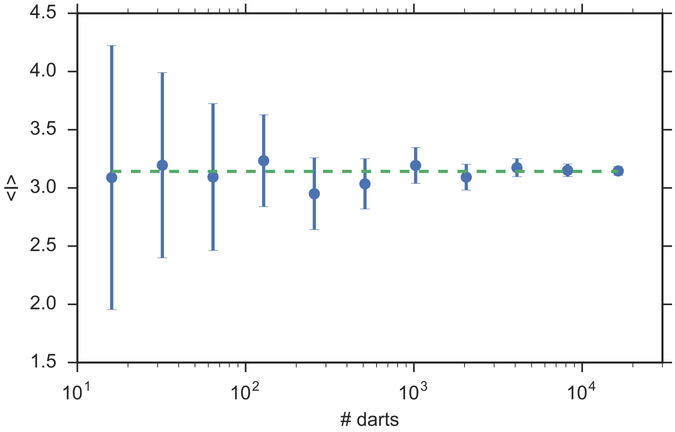
Let's consider the example, we've seen before:

$$I = \int_0^{\pi} x \sin x \; dx = \pi.$$

```
def simpleIntegral(npts):

    xi     = np.random.uniform(0, np.pi, npts)
    fi     = xi * np.sin(xi)

    intg   = np.pi * np.mean(fi)
    stdInt = np.pi/np.sqrt(npts) * np.std(fi)

    return intg, stdInt
```

Example

```
I, errI = simpleIntegral(1000)

print(I, errI)
3.16046202212 0.0198906574291
```



Postface

For

$$I = \int_a^b f(x)\,dx$$

The key results for MC integration,

$$\begin{aligned} I &= (b-a)\langle f \rangle \\ \sigma_I &= \frac{(b-a)\sigma_f}{\sqrt{n}} \end{aligned}$$

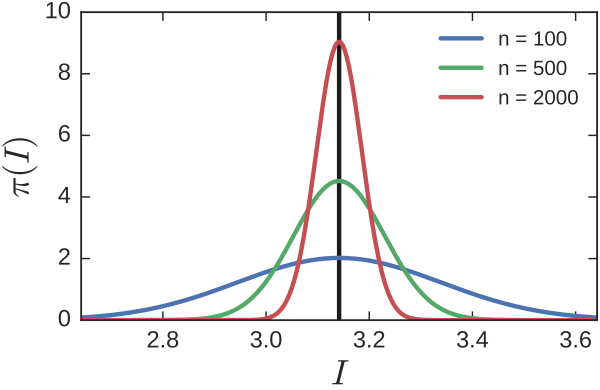
Consider the particular example $f(x) = x \sin x$.

$$\begin{aligned} \langle f \rangle &= \frac{1}{\pi}\int_0^\pi x \sin x\,dx = 1 \\ \langle f^2 \rangle &= \frac{1}{\pi}\int_0^\pi x^2 \sin^2 x\,dx = \frac{\pi^2}{6} - \frac{1}{4} \end{aligned}$$

Postface: Example

Thus, theoretically,

$$\sigma_f^2 = \langle f^2 \rangle - \langle f \rangle^2 = \frac{\pi^2}{6} - \frac{5}{4} \approx 0.3949$$



$$\pi(I) = \mathcal{N}(\pi, \pi^2 \sigma_f^2 / n)$$

Postface: Example

The MC estimate I is itself a random variable with its own distribution

CLT tells us that for large enough n,

$$\pi(I, n) = \mathcal{N}(\pi, \pi^2 \sigma_f^2 / n)$$

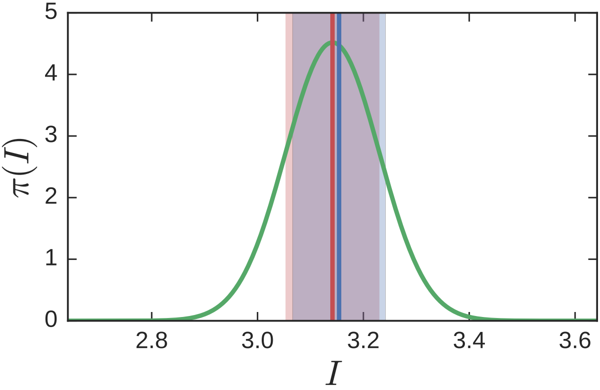
When we run an MC simulation with a particular n, the resulting estimate of the integral is sampled from $\pi(I, n)$.

As n increases, $\pi(I, n)$ becomes narrower, and our estimates are closer to the expected value $I = (b - a)\langle f \rangle$ more often

Our on-the-fly estimate of σ_I depends on our on-the-fly estimate of σ_f

Postface

For $n = 500$, the true $\sigma_f \approx \sqrt{0.3949} = 0.6284$.



estimated $\sigma_f = 0.6168$

red and blue areas show range of the true and estimated σ_I

Supplementary Information

1. Derivation of Error Propagation Formula

1. Derivation of Error Propagation Formula

Without loss of generality, assume $q(x_1, x_2)$. Let X_1 and X_2 be particular random variables and $Q = q(X_1, X_2)$ be the resultant.

Further suppose that (μ_1, μ_2), and $(\sigma^2_{x_1}, \sigma^2_{x_2})$ are the expected values, and variances, respectively, of X_1 and X_2.

Step 1: Taylor series expansion

$$\begin{aligned} Q &= q(X_1, X_2) \\ &= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(X_1 - \mu_1) + \frac{\partial q}{\partial x_2}(X_2 - \mu_2) + \text{h.o.t} \end{aligned} \tag{6}$$

We can ignore higher order terms if $(X_i - \mu_i)$ are small.

In experiments, this often requires measurement error to be "reasonably" small.

Derivation

Step 2: Take Expected Value

Note for independent variables:

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2).$$

Using this property, we can write eqn 7,

$$\begin{aligned}
E(Q) &= E(q(\mu_1, \mu_2)) + E\left(\frac{\partial q}{\partial x_1}(X_1 - \mu_1)\right) + E\left(\frac{\partial q}{\partial x_2}(X_2 - \mu_2)\right) \\
&= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(E(X_1) - \mu_1) + \frac{\partial q}{\partial x_2}(E(X_2) - \mu_2)) \\
&= q(\mu_1, \mu_2) + \frac{\partial q}{\partial x_1}(\mu_1 - \mu_1) + \frac{\partial q}{\partial x_2}(\mu_2 - \mu_2)) \\
&= q(\mu_1, \mu_2)
\end{aligned}$$

Derivation

Step 3: Take Variance

Note for independent variables:

$$V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2).$$

Using this property, we can write eqn 7,

$$\begin{aligned}
V(Q) &= V(q(\mu_1, \mu_2)) + V\left(\frac{\partial q}{\partial x_1}(X_1 - \mu_1)\right) + V\left(\frac{\partial q}{\partial x_2}(X_2 - \mu_2)\right) \\
&= 0 + \left(\frac{\partial q}{\partial x_1}\right)^2 V((X_1) - \mu_1)) + \left(\frac{\partial q}{\partial x_2}\right)^2 V((X_2) - \mu_2)) \\
&= \left(\frac{\partial q}{\partial x_1}\right)^2 (V(X_1) - 0) + \left(\frac{\partial q}{\partial x_2}\right)^2 (V(X_2) - 0) \\
\sigma_q^2 &= \left(\frac{\partial q}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial q}{\partial x_2}\right)^2 \sigma_{x_2}^2
\end{aligned}$$

2. Signal to Noise Ratio

If x_1, x_2, $\ldots$, x_n are independent random variables, and

$$q = q(x_1, x_2, ..., x_n),$$

then,

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x_1}\sigma_{x_1}\right)^2 + \left(\frac{\partial q}{\partial x_2}\sigma_{x_2}\right)^2 + ... + \left(\frac{\partial q}{\partial x_n}\sigma_{x_n}\right)^2.$$

For $q = x_1 \pm x_2$, this implies

$$\sigma_q^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2. \tag{7}$$

Similarly, for $q = x_1 x_2$, or $q = x_1/x_2$,

$$\left(\frac{\sigma_q}{q}\right)^2 = \left(\frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\frac{\sigma_{x_2}}{x_2}\right)^2. \tag{8}$$