Vinayak Ravichandran

Github: @vravich01

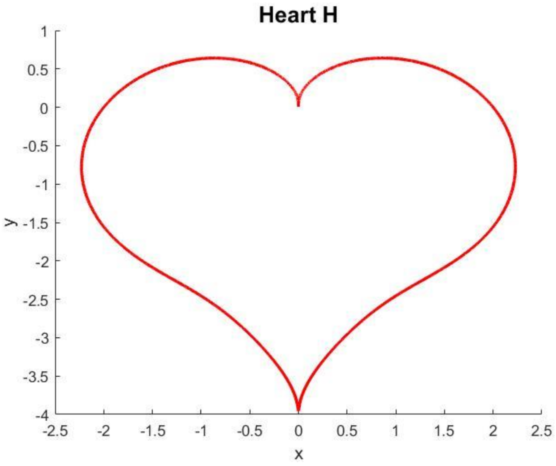
Email: vinayak1ravichandran@gmail.com

Monte Carlo Methods

Section 1: Introduction

In this analysis, the area of a heart, $\boldsymbol{H}$, was estimated using two variations of Monte Carlo methods. Monte Carlo methods rely on repetitions of simulations of random sampling to ensure the most accurate results for the desired value. However, there are other factors to consider when trying to maximize the accuracy of estimations produced by Monte Carlo methods such as control variates and sampling methods. The known approximate area of $\boldsymbol{H}$ is 12.5230317, and the polar equation which defines $\boldsymbol{H}$ is:

$$r_{\boldsymbol{H}}(\theta) = 2 - 2\sin(\theta) + \sin(\theta)\frac{\sqrt{|\cos(\theta)|}}{\sin(\theta) + 1.4}$$



Section 2: Using Random Points & Indicator Functions

2.1) Methods

2.1.a) N random cartesian points, (x, y), were generated such that $x \in [-2.5, 2.5]$ and $y \in [-1, 4]$. Then, the random variable X was defined as:

$$X = 25 \times I_{\boldsymbol{H}}$$

Where $I_{\boldsymbol{S}}$ is defined as an indicator function for any closed curve $\boldsymbol{S} \in \mathbb{R}^2$:

$$I_{\boldsymbol{S}} = \begin{cases} 1 & if\ (x, y) \in \boldsymbol{S} \\ 0 & otherwise \end{cases}$$

Since we can expect the proportion of number of points which fall within $\boldsymbol{H}$ to the total number of points to be equal to the proportion of $area(\boldsymbol{H})$ to $area([-2.5, 2.5] \times [-1, 4])$, we can expect $E[X] = area(\boldsymbol{H})$**.** This Monte Carlo method relies on the following assumption, which becomes more accurate as $N \to \infty$:

$$\frac{area(\boldsymbol{H})}{area([-2.5, 2.5] \times [-1, 4])} \approx \frac{\#\ of\ (x, y) \in \boldsymbol{S}}{N}$$

Which can be rearranged as:

$$area(\boldsymbol{H}) \approx area([-2.5, 2.5] \times [-1, 4]) \frac{\#\ of\ (x, y) \in \boldsymbol{S}}{N} = 25 \frac{\#\ of\ (x, y) \in \boldsymbol{S}}{N} \approx E[X]$$

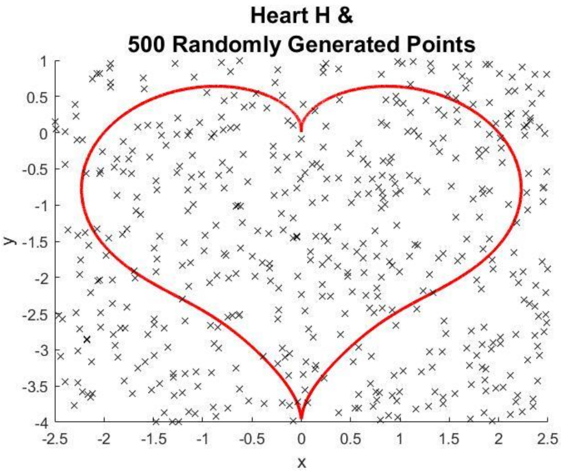
Then, another random variable $\bar{X}_N$ estimates $E[X]$ for N randomly generated points. This estimation for identically independent X_i is defined as:

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$\bar{X}_N$ was used to compute estimates for $E[X]$ to be used as the center of the population distribution, μ. μ, in the form of $\bar{X}_N \approx E[X]$, can now be the center of a confidence interval of X for which the standard error was computed by using the standard deviation of X for N points (the standard deviation of the sample) as an estimate of the population standard deviation. Dividing the estimate of the population standard deviation by $\sqrt{N}$ produces the standard error:

$$stE(X) = \sigma(X)/\sqrt{N}$$

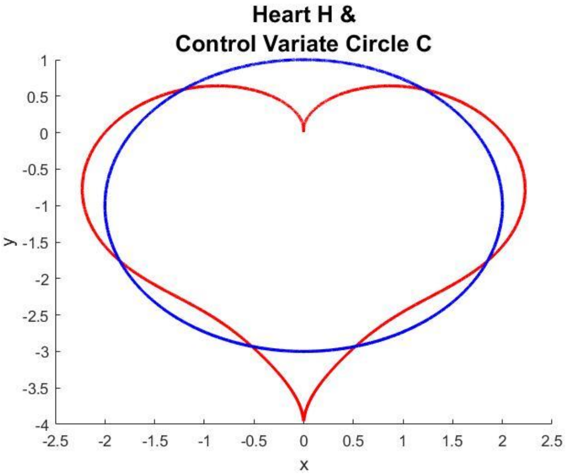
Note that the Central Limit Theorem allows for the assumption that the population is normally distributed and allows the use of the sample center $\mu = \bar{X}_N$ as an approximation of the population center $E[X]$. Then a 99.7% confidence interval for $E[X]$ was computed as $[\mu - 3stE, \mu + 3stE]$. The plot is shown below.



Heart H &
500 Randomly Generated Points

2.1.b) To increase the accuracy of the estimated area of $\boldsymbol{H}$, a control variate ($\boldsymbol{C}$) was introduced; the polar equation that defines $\boldsymbol{C}$ is:

$$r_{\boldsymbol{C}}(\theta) = \sqrt{4 - \cos(\theta)^2} - \sin(\theta)$$



Heart H &
Control Variate Circle C

C is a circle which has a known radius; therefore, it has a known area. Now, define a new random variable Z as:

$$Z = X + \alpha(Y - E[Y])$$

Where the random variable Y is defined as:

$$Y = 25 \times I_{\boldsymbol{C}}$$

Note that $E[X] = E[Z]$ due since the expected value operator is a linear operator. The following manipulations will show that $E[X] = E[Z]$:

$$E[Z] = E[X + \alpha(Y - E[Y])$$

$$E[Z] = E[X] + E[\alpha(Y - E[Y])]$$

$$E[Z] = E[X] + \alpha E\big[Y - E[Y]\big]$$

$$E[Z] = E[X] + \alpha E[Y] - \alpha E[E[Y]]$$

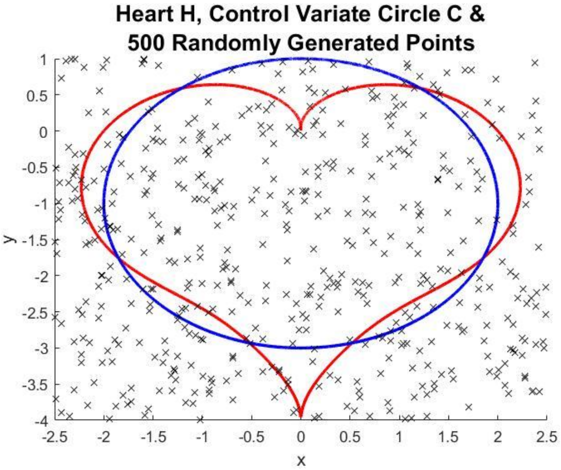
Here, it is important to note that $E[Y]$ is the area of $\boldsymbol{C}$, which is constant. So then $E\big[E[Y]\big] = E[Y]$:

$$E[Z] = E[X] + \alpha E[Y] - \alpha E[Y]$$

$$E[Z] = E[X] + 0$$

$$\therefore E[Z] = E[X]$$

Using Z, the area of $\boldsymbol{H}$ can be estimated more accurately than by using X. The improvement reveals itself in the form of a lower standard error for Z, so the 99.7% confidence interval for Z should be over a smaller range than for X. Both intervals are computed using the same set of random cartesian points. A plot is shown below.



2.1.c) In addition to a control variate, an optimized value of the constant α can be found such that the standard error is minimized. A linearly spaced array of possible best α values was created. Then, the possibilities were tested using the randomly generated points and the α value which resulted in the lowest $stE(Z)$ was recorded.

2.2) Results

The parameters of interest were $E[X]$, $E[Z]$, $\sigma(X)$, $\sigma(Z)$, $stE(X)$, $stE(Z)$, and confidence intervals for both X and Y (note that the first four are computed as estimates). The results are recorded in the table below. One run of the corresponding program (*Part1.m*) uses $N = 500$, so 500 random cartesian points were randomly generated for this sample.

Random Variable:	X	Z
Estimated Expected Value	12.2000	12.2665
Estimated Standard Dev	12.5089	8.9634
Standard Error (N = 500)	0.5594	0.4009
99.7 % Confidence Interval	$[10.5218, 13.8782]$	$[11.0640, 13.4691]$
Estimated Best α & Associated Standard Error	N/A	-0.7824, 0.1635

Note that each value for Z is "better" than for X: the estimated expected value is closer to $area(\boldsymbol{H})$, and the standard error is lower which leads to a tighter confidence interval. The expected values seem to be lower than $area(\boldsymbol{H})$, however, $E[Z]$ is still more accurate.

2.3) Conclusion

Although Monte Carlo algorithms are useful in simulating problems pseudo-deterministically, they can be improved upon as shown by the introduction of control variate $\boldsymbol{C}$. Since $\mathrm{area}(\boldsymbol{C})$ is known, the addition of $\alpha(Y - E[Y])$ to X "stabilizes" the set of random variables Z_i so that its variance is less than X despite $E[X] = E[Z] = area(\boldsymbol{H})$. Therefore, with the same level of confidence, it can be asserted that $E[X] \in [10.5218, 13.8782]$, which is a wider interval than $E[Z] \in [11.0640, 13.4691]$. Stabilizing the data by adding a control variate allows us to make more precise predictions about datasets.

Section 3: Estimating Integrals

3.1) Methods

3.1.a) The approach in this section will be to use Monte Carlo methods on the polar area integral for $\boldsymbol{H}$:

$$area(\boldsymbol{H}) = \int_0^{2\pi} \frac{r_{\boldsymbol{H}}(\theta)^2}{2}$$

Instead of random cartesian point, random values of θ will be generated from a $unif(0, 2\pi)$ distribution. Then, the random variable X can be defined as:

$$X = \pi r_{\boldsymbol{H}}(\theta)^2$$

Note that circular areas are being sampled randomly and the radii at which the area is calculated corresponds to $r_{\boldsymbol{H}}(\theta)$, which is a radius measurement from $\boldsymbol{H}$ itself. Also note that the expected value of X is still approximately:

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

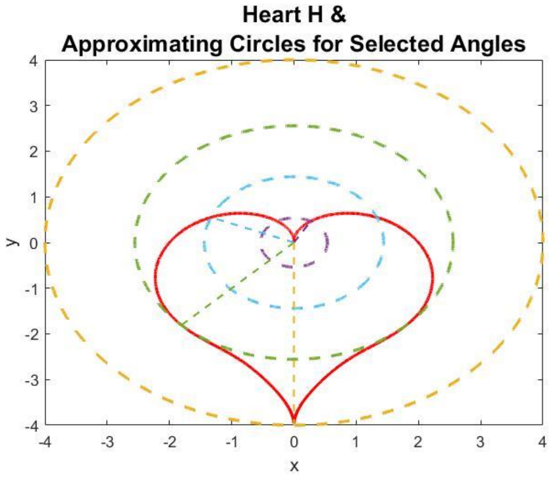
Note that since θ comes from $unif(0, 2\pi)$, so the probability of $\theta = c$ such that c is any value from $[0, 2\pi]$ is $\frac{1}{2\pi - 0}$. So we can establish that:

$$PDF_\theta = \frac{1}{2\pi}$$

It is important to recall that:

$$E[X] \approx \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{i=1}^{N} \pi r_{\boldsymbol{H}}(\theta)_i^{\,2} = \frac{1}{N}\sum_{i=1}^{N} 2\pi \times \frac{r_{\boldsymbol{H}}(\theta)_i^{\,2}}{2} = \frac{1}{N}\sum_{i=1}^{N} \frac{\frac{r_{\boldsymbol{H}}(\theta)_i^{\,2}}{2}}{PDF_\theta}$$

Note that the numerator of within the summation is equal to the integrand of the area integral. Therefore, we can assert that $E[X] = area(\boldsymbol{H})$ since summing up radial areas over $[0, 2\pi]$ (integral) is the same as the average value of the areas of randomly selected circles, where both the radial areas and circles are defined by $r_{\boldsymbol{H}}(\theta)$. Four of these random, approximating circles are shown below for specific values of θ.



A 99.7% confidence interval for $E[X]$ was computed as $[\mu - 3stE, \mu + 3stE]$ where:

$$\mu = \bar{X}_N$$

$$stE(X) = \sigma(X)/\sqrt{N}$$

3.1.b) A circle control variate ($\boldsymbol{C}$) was introduced here as well; the polar equation that defines $\boldsymbol{C}$ is still:

$$r_{\boldsymbol{C}}(\theta) = \sqrt{4 - \cos(\theta)^2} - \sin(\theta)$$

Then, define random variable Z just as *Section* 2:

$$Z = X + \alpha(Y - E[Y])$$

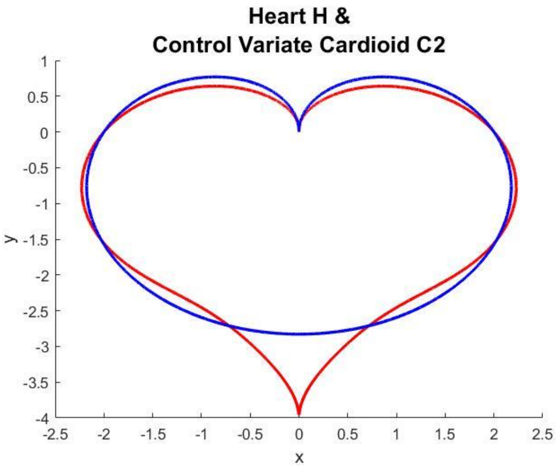
Except now, random variable Y is defined as:

$$Y = \pi r_{\boldsymbol{C}}(\theta)^2$$

Just as with *Section 2*, it is expected that Z will vary less than X. A 99.7% confidence interval was constructed for both X and Z. The optimized α value for Z was also found using the same algorithm as *Section 2*.

3.1.c) Now, since two Monte Carlo methods have been used and the use of a control variate has been established as helpful, a more fitting control variate can be analyzed. Let $\boldsymbol{C_2}$ be a cardioid such that $area(\boldsymbol{C_2}) = area(\boldsymbol{C})$; it is defined by the polar equation:

$$r_{\boldsymbol{C_2}}(\theta) = 2\sqrt{1 - \sin(\theta)}$$


Heart H &
Control Variate Cardioid C2

Then, define a new random variable Z_2 as:

$$Z_2 = X + \alpha(Y_2 - E[Y_2])$$

Where the random variable Y_2 is defined as:

$$Y_2 = \pi r_{\boldsymbol{C_2}}(\theta)^2$$

A 99.7% confidence interval was constructed for Z_2.

3.2) Results

The parameters of interest in this section were $E[X]$, $E[Z]$, $E[Z_2]$, $\sigma(X)$, $\sigma(Z)$, $\sigma(Z_2)$, $stE(X)$, $stE(Z)$, $stE(Z_2)$, and confidence intervals for X, Y, Z_2. Again, note that the expected values and standard deviations are estimates. Results for $N = 500$ are recorded in the table below.

Random Variable:	X	Z	Z_2
Estimated Expected Value	13.0191	12.7827	12.7520
Estimated Standard Dev	9.6068	6.3534	6.2473
Standard Error (N = 500)	0.4296	0.2841	0.2794
99.7 % Confidence Interval	[11.7302, 14.3080]	[11.9303, 13.6351]	[11.9138, 13.5901]
Estimated Best α & Associated Standard Error	N/A	-1.0767, 0.1444	-1.0767, 0.1243

Here, the values obtained using control variates (Z and Z_2), are more accurate than for X. Note that Z_2 is more accurate than Z. Although the expected value of all three random variables are the same, the estimations produced by the simulations reveal that some computational forms reduce variance and lead to a more accurate result.

3.3) Conclusion

Monte Carlo methods can be useful to evaluate integrals a demonstrated by estimating the area of $\boldsymbol{H}$, which happens to have an exact polar integral representation. The methods described in *Section 3* allow the situation to be simulated with random sampling, but it can also be thought of as approximating an integral. These methods can be expanded to any symbolically unsolvable integral to obtain a numerical approximation by determining the proper PDF (see *Section 3.1*.a). Although the results of this analysis could not determine a general formula for α in terms of operations on sets of random variables, it can be concluded that α can be optimized when a control variate is used.