# Lab # 3 Monte Carlo Simulation & Confidence Intervals

For the questions of lab n°3 under Linux, even if it is simpler to use **rand()** for a first test (or more `precisely (double) rand() / (double) RAND_MAX` using `stdlib)`, note that this function is not suitable for scientific work. You should now know that for scientific applications the default system random number generators are often statistically weak**. Such generators have to be banned for modern scientific programming**. For simulations, it is far better to use a fast generator such as Mersenne Twister proposed by Matsumoto, discovered in lab 2 and proposed by many language APIs and library.

For a quick start work in a single file, later, if you can, uyou can separate your code into:

- one file with the Mersenne Twister source code, and
- another with your main simulation code (with only one main() function).

Be careful to keep the original MT initialization that you checked in Lab 2 and use the function proposed to draw pseudo-random numbers between 0 and 1 (included). All code should be written in C.

1) **Propose a function to estimate π by sampling like seen in the lecture** (or sample the volume of a sphere with a radius of 1 if you are already familiar with the Monte Carlo method). In the case of a sphere, the volume is known analytically (see below), where is the radius.

$$\mathrm{Vol} = \frac{4}{3}\pi r^3$$

Your function will accept in input the number of points used for the sampling.

Test your code with :

- 1 000 points,
- 1 000 000 points,
- and lastly 1 000 000 000 points.

Each point in $\mathbb{R}^2$ or $\mathbb{R}^3$ needs two or three random numbers: $x_r$, $y_r$ (and $z_r$, for a volume sampling). For this case study let see how many drawings you may need to get a precision below $10^{-2}$ (two decimals), then below $10^{-3}$ and then below $10^{-4}$.

Remember that the convergence rate of this method is very slow: `sqrt(of the total numbers of points drawn)`. But like a peasant tractor this method "goes everywhere". In this question, you do not have to make replicates (independent experiments) and a side effect is that you also have a rough estimate of π – this comes in the next question.

**2) Computing independent experiments and obtaining the mean**. Launching – ie drawing the "dice" many times, means calling many times the previous function which estimates π or he volume of a sphere (and each function call needs many drawings). When simulating a dice rolling, we needed only one random number for an experiment. When we loop on the "dice rolling", we have many simulations (independent experiments). When we estimate this volume, we launch a `simPi` or `simSphere` function (with a specified number of points: nbPoints, and each point needs three random numbers between [0..1]). In order to compute '*n'* independent experiments (replicates) of this simulation, first initialize a good random number generator like MT, then propose a loop, which will call '*n*' times the estimation of the volume with "nbPoints". The independence is simply achieved if we do not reinitialize the pseudo-random number generator between two experiments (replicates). This can be done by keeping the proper and original initialization found in Makoto's web site. When we run the loop calling the Volume simulation function, we can store all the results of the '*n'* experiments (replicates) in an array (of numbers in double precision). We can then propose the mean of all the 'estimated Volumes' (arithmetic mean). Run 10 to 40 experiments with:

1 000, 1000 000 and 1 000 000 000 points, compute the difference between your `meanVolume` and the analytically expected volume. If you divide by this value you will also obtain what is called the relative error.

2) **Computing of confidence intervals around the simulated mean**. The user inputs the number of replicates (experiments) he wants to perform before computing a confidence interval at 99% ($\alpha$=0.01) – **to do so, inspire yourself from the technique and table given in the appendix** of this lab for ($\alpha$=0.05). See whether the number of replicates improves your results and decreases the confidence radius. The number of random drawings (sampling) for each individual replicate (each experiment) is a sensitive parameter. Be careful when you do this comparison. We have very few significant numbers in a float variable (only 7 significant decimal digits for `floats`), for scientific computing prefer `double` instead. It is preferred for scientific computing. When you have an estimation of the standard deviation, you can also see how to compute the 'standard error' of the sample mean.

The Monte Carlo method needs a lot of random sampling to increase its precision (remember the very slow convergence rate). However, the Monte Carlo (MC) method is the only method we have to compute hyper-volumes in hyperspaces, to achieve evenly distributed space filling in large dimensions for experimental designs and where accurate mathematical methods are often intractable or non-existing. Uniformity is guaranteed up to 623 dimensions for the Mersenne Twister (eg. hyperspace of parameters for sensitivity analysis : 100 parameters => 100 dimensions to explore). AI training also requires massive MC sampling far above the 623 dimensions.

## Appendix: Computing confidence Intervals for a sample of results: we don't have the full population

**Introduction:** If *X* is a simulation result, ($X_1$ , ... , $X_n$) is the set obtained with *n* independent replications (experiments) of this stochastic simulation. This means that each independent stochastic simulation experiment is run (under the same conditions) but with a different (and independent) random stream. Usually the computing of confidence intervals is achieved on the observed arithmetical means.

$$\overline{X}(n) = \sum_{i=1}^{n} X_i \,/\, n$$

The computing of a confidence interval is simple when we consider that the $X_i$ variables (simulation results) have identical independent Gaussian distributions.

**Principle:** The confidence interval is centered on the arithmetical mean, so we just compute what is called the confidence radius (error margin). Here is the theoretical statistical hypothesis that is used to obtain this radius. If $X_i$ have identical independent Gaussian distributions with a theoretical mean $\mu$ and a variance of $\sigma^2$, then the following random variable :

$T(n) = \dfrac{\overline{X}(n) - \mu}{\sqrt{S^2(n)/n}}$ is distributed according to a Student law with *n-1* degrees of liberty.

$$S^2(n) = \frac{\sum_{i=1}^{n} \left[X_i - \overline{X}(n)\right]^2}{n-1}$$

$S^2(n)$ is an estimate without bias of the $\sigma^2$ variance. Since we don't know the theoretical standard deviation $\sigma$, we estimate the variance with the results we have and thus use this approximation. We compute $S^2(n)$ and use it in the following formula below to obtain the confidence radius (error margin) at the 1-$\alpha$ level :

$$R = t_{n-1,1-\alpha/2} \times \sqrt{\frac{S^2(n)}{n}}$$

William Sealy Gosset introduced the Student to correct the fact that we only use an estimate of $\sigma$ and not its true value. The table below gives the values $t_{n-1,1-\alpha/2}$ for a student law with $\alpha = 0.05$.

If $\alpha = 0.05$, the confidence interval is said at 95%. The computing of R gives the following interval $\left[\overline{X} - R, \overline{X} + R\right]$, with a 1-$\alpha$ confidence (95%).

**Table 1:** Values of $t_{n-1,1-\alpha/2}$ of a Student law starting with $\alpha = 0.05$ depending on *n* experiments.

| 1≤n≤10 | $t_{n-1,1-\alpha/2}$ | 11≤n≤20 | $t_{n-1,1-\alpha/2}$ | 21≤n≤30 | $t_{n-1,1-\alpha/2}$ | n>30 | $t_{n-1,1-\alpha/2}$ |
|---|---|---|---|---|---|---|---|
| **1** | **12.706** | **11** | **2.201** | **21** | **2.080** | **40** | **2.021** |
| **2** | **4.303** | **12** | **2.179** | **22** | **2.074** | **80** | **2.000** |
| **3** | **3.182** | **13** | **2.160** | **23** | **2.069** | **120** | **1.980** |
| **4** | **2.776** | **14** | **2.145** | **24** | **2.064** | **+∞** | **1.960** |
| **5** | **2.571** | **15** | **2.131** | **25** | **2.060** | | |
| **6** | **2.447** | **16** | **2.120** | **26** | **2.056** | | |
| **7** | **2.365** | **17** | **2.110** | **27** | **2.052** | | |
| **8** | **2.308** | **18** | **2.101** | **28** | **2.048** | | |
| **9** | **2.262** | **19** | **2.093** | **29** | **2.045** | | |
| **10** | **2.228** | **20** | **2.086** | **30** | **2.042** | | |

The Central Limit Theorem (CLT) says that for non-normal data, the distribution of the sample means has an approximate normal distribution, no matter what the distribution of the original data looks like, as long as the sample size is large enough (usually at least 30) and all samples have the same size. This is true not only for the sample mean but also for other sample statistics (find the table for $\alpha$ = 0.01).