实验室#2--随机变量的产生
(用C语言实现--查看讲座幻灯片并仔细阅读主题)

1) 找到 "**Matsumoto主页** "和原始Mersenne Twister（MT）的 "**最后C实现**"。

在第一个实验室中，你已经实现了生成随机数的不同技术。你已经看到，掌握高质量的随机数不是那么容易。st在这个实验室中，我们将使用21世纪为科学界提出的顶级生成器之一（在623个维度上等分布，周期为2^{19937} 号）。尽管它不是加密安全的，但它将是你在本实验和下一个实验中使用的发生器。

你可以找到目前用C语言实现的Mersenne Twister (MT) google到 "Matsumoto Home Page"，然后找到Mersenne Twister/2002版本--解释和C代码。下载带有源代码的.tar文件+预期输出和readme（采取32位版本）。

编译并测试你是否获得了预期的输出（为了可移植性和**可重复性**）。在实验题中使用genrand_int32或genrand_real(1/2)函数。解压缩档案（Unix命令："`tar zxvf yourfile.tgz`"）并使用该例子。将你在你的计算机会话中在本地获得的结果与预期的输出进行比较（可重复性--见松本提出的README文件和预期输出）。从现在开始，总是使用精细的生成器，如MT或其他非常好的生成器。

一旦你测试了这段代码的位数重现性，你将通过在Makoto代码的主函数前加入你的代码来测试本实验室的下一个函数，你将修改Makoto的测试函数来测试你的实验室函数。

2) 产生A和B之间的均匀随机数

实施：使用MT函数提供[0...1]之间的数字，提出一个名为 "uniform "的C函数，有2个参数'a'和'b'（实数），在'a'和'b'之间产生伪随机数。对-89,2°C和56,7°C之间的温度测试这个函数。

3) 离散经验分布的再现

假设我们有3类的实地数据。A类有350个观察值，B类有450个观察值，C类有200个观察值，得出3个物种（A、B和C）的分布概率如下。A为35%，B为45%，C为20%。

用MT复制（模拟）一个具有相同分布的个体群体。

a) **实施并测试**一个模拟这种离散分布的程序，其中有A、B和C三个等级，用1 000、10000、100000和1 000 000张图进行测试。在3个变量中累积每个物种的个体数量，并显示获得的百分比。

b) **实现**一个更通用的函数，其输入参数如下：一个类数组的大小，然后是数组本身与每个类中

观察到的个体数量（见讲座幻灯片中的例子--HDL "好 "胆固醇，用这些值来检查这个问题）。

a. 首先计算对应的数组中每个类别的概率（分布函数），并对此进行测试。
b. 然后计算另一个给出累积概率的数组。这个函数输出的是后一个数组。测试一下吧。
c. 用幻灯片中给出的数据（和/或用你自己的数据）测试整个函数，并检查有1000和1000000张图的模拟分布。

4) 连续分布的再现

有可能通过反转分布函数来重现连续分布。当在0和1之间抽取一个伪随机数时，有可能得到一个根据给定的连续分布函数（F）分布的数字，假设后者是可逆的。

$$x = F^{-1} \text{(抽出的随机数)}$$

这种名为 "变形 "的技术并不完全通用，但它可以应用于许多分布规律（二项式、Weibull、Uniform...）。例如，指数分布的分布函数（负指数法）在方程（8）中给出，导致方程（9）的反法必须实施。我们可以把它看作是泊松分布的一个类似物。实际上，泊松过程中两个事件之间的时间（直观地说：两个罕见事件之间的时间）遵循一个指数分布。例如，两个放射性物质解体之间的时间。

$$F(x) = \int_0^x \frac{1}{M} e^{-\frac{1}{M}z} \, dz = 1 - e^{-\frac{1}{M}x} \qquad (8)$$

$$\text{随机抽取的号码} = 1 - e^{-\frac{1}{M}x}$$

$$1 - \text{随机抽取的数字} = e^{-\frac{1}{M}x}$$

$$\ln(1 - RandomNumberDrawn) = -\frac{1}{M}x$$

$$x = -M \ln(1 - RandomNumberDrawn) \qquad (9)$$

A和B之间的统一法律	: x = F^{-1} (抽取的随机数) = A+(B-A) * 抽取的随机数	
	平均值=（B+A）/2	差异 = 1/12 * (B - A)^2
负指数法（平均）：x = F^{-1} （抽出的随机数）= - 平均数x对数（1-抽出的随机数）。		
	平均值=M	差异=M

图2.均匀指数和负指数法的反函数

以下是你要编码的内容。

a. 实现negExp函数，接受平均值作为参数。
b. 检查抽出1000个（然后是1000 000个）后得到的平均值，它应该接近11。这假定使用0和1之间的精细随机数

在方程（9）中，以获得正确的分布。例如，这些数字可以对应于提交给一个计算集群的两个作业之间的到达时间。

c. 检查这个离散分布（有偏差的骰子）。使用一个有23个bin的数组，测试0到1之间，1到2之间，……的数字的频率。保留最后一个仓，以累积22以上的数值的数量。对于每一个抽出的数字，计算它出现在哪个仓中，并对所有抽出的数字进行累计（1 000，1 000 000）。

```
Test22bins[ (int) negExp(11) ] ++;
```

如果平均数设置为11，你会产生许多0和1之间的数字，1和2之间的数字会少一点，等等。如果你显示直方图，会产生类似图3的东西（斜率不同）。你还应该测试观察到的平均值是否与理论平均值相对应。

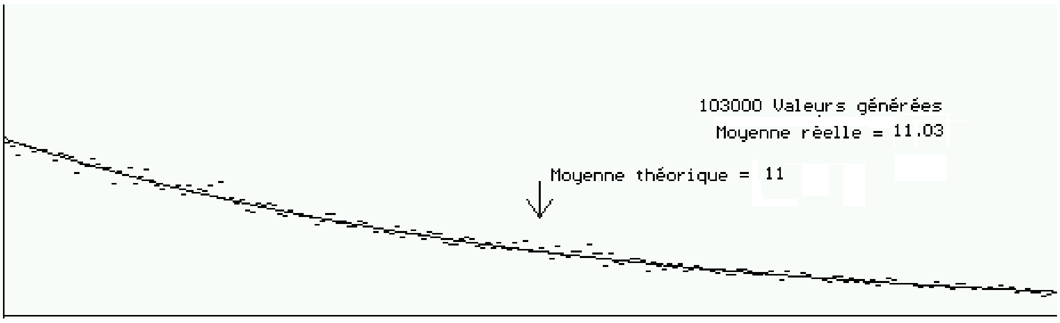

图2.统一和负指数法的反函数的模拟。

5) 模拟非可逆分布规律

在不可逆分布规律的情况下，我们可以使用拒绝技术，这是一种受蒙特卡洛启发的技术。下面是一个标准的拒绝算法，用于根据概率分布f(x)在2个值MinX和MaxX（+Min Y和MaxY，它们是提供概率分布（密度）函数（PDF）周围方框的值）之间生成一个数字。

图4.对任何分布的通用拒绝算法

```
(1)   产生2个随机数Na1和Na2
(2)   计算X = MinX + Na1 * (MaxX - MinX)
(3)   计算Y = MaxY * Na2
(4)   如果Y<f(X)
      那么X被认为是按照密度函数为f(x)的规律分布的。
      否则拒绝(X)并转到(1)即：再次抽取2个0和1之间的伪随机数，等等。
      结束语
```

高斯分布的特殊情况。

正态法的密度（缩小和居中：平均数=0，标准差=1）被记为*N*(0,1)，并在下文中用公式（10）给出。

$$p(x) \quad \frac{1}{\sqrt{(2)}} e^{\frac{x^2}{2}} \qquad (10)$$

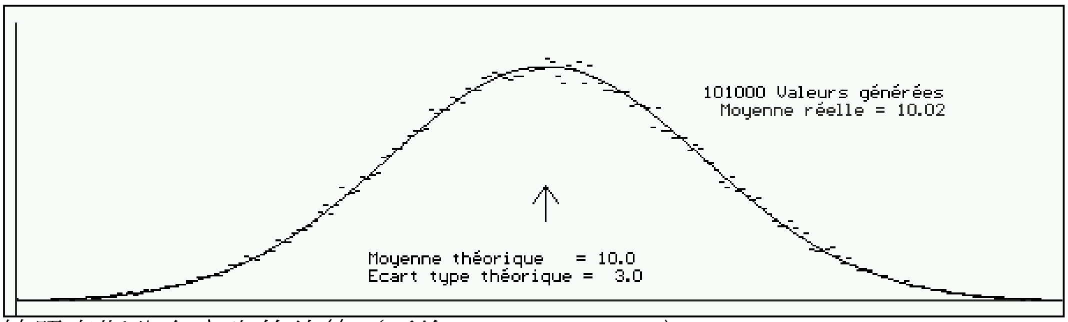

图5.按照高斯分布产生的估值（平均=10，std.dev.=3）

5.1 第一次实施。

考虑一个实验，抽出30次普通骰子。总结得到的结果。预期的结果是在30（最小：30×面1和潜在的最大180（30×面6）之间，概率非常低（$1 / 6^{30}$））。

模拟这个实验'许多'次以获得平均数（和标准差）的近似值。然后，你可以在平均值周围定义统计分档，以查看（预期的）钟形曲线（使用150个分档--为每个可能的总和建立一个数组，例如用EXCEL显示结果）。

5.2 对高斯分布的分析模型的检验

1958年，Box和Muller提出了一种不使用中心极限定理并使用两个伪随机数的精确方法。方程（14）使用两个随机数Rn1和Rn2，并产生两个分布在中心和缩小的高斯定律两边的数字--*N*（0,1）。存在许多变种来近似高斯分布，有些更快，有些更精确...

$$x1\ \cos(2Pv2)(2\ln(Rn1))^{\frac{1}{2}}$$
$$x2\ \sin(2Pv2)(2\ln(Rn1))^{\frac{1}{2}} \qquad (14)$$

实施。测试Box和Muller函数，在*N*(0,1)之后产生0左右的数字。两个伪随机数给出2个数字。检查1000张和1000000张图，有多少数字分布在-5和5周围的20个仓中（在[-3...-2.5[，[-2.5，-2[，...[2...2.5[，[2.5...3[。打印你的结果，看看它是否符合高斯分布的已知统计数据？

6) 在C/C++和Java中找到能生成随机变量的库，就像你在前面的问题中做的那样。