# 通过计算机数值计算结果学习中心极限定理

胡锋 重庆

最近浏览生物群体行为的文献时，注意到了所谓的“ Many Wrong Principle ”。其基本思想是，运动时群体中的个体向着目标方向行进时，会有误差$\sigma$（高斯分布中的标准差）。但是作为群体的整体，运动方向是个体运动方向的平均，其运动的误差会变小。群体越大，群体运动的方向越精确，误差变为$\frac{\sigma}{\sqrt{n}}$（$n$是群体中个体数量）。

这个原理是数学中的中心极限定理（Central limit theorem CLT）的应用，而这个应用很让我感兴趣。所以花了一些时间，通过计算机进行数值计算，检验、学习了中心极限定理。

中心极限定理可以表述为（粗糙表述）：对独立的，同分布的随机变量来说，假设其均值为$\mu$，有限方差为$\sigma^2$，则来自于该随机变量总体的，含有$n$个这样的随机变量的样本的平均值符合高斯分布，其均值为$\mu$，方差为$\sigma^2/n$。

解释：(1）这个表述中实际涉及到两类随机变量。第一类是“ 个体 ”的随机变量，记为$x_1, x_2, x_3 .....x_n$，这$n$个变量是某一个样本的随机变量。它们是独立的，概率分布相同的，可以是高斯分布或者均匀分布。第二类随机变量是“ 整体 ”随机变量，记为$X_{n1}, X_{n2}, X_{n3} .....$，每一个$X_{ni}$都与一个样本$i$对应。整体随机变量与个体随机变量的关系是$X_{ni}=(\frac{x_1+x_2+.....x_n}{n})_{sample\,i}$，是这个样本“ 个体 ”随机变量的平均值。

（2）如果样本很大（即 n 很大，每个样本中的“ 个体 ”随机变量很多），则“ 整体 ”随机变量的高斯分布会很窄，方差为$\sigma^2/n$，标准差为$\sigma/\sqrt{n}$。如果统计“ **整体** ”随机变量的概率，画在分布图上，可以看到这个高斯分布的宽度（对应标准差）随着 n 增大，以$1/\sqrt{n}$变小。

下文中用计算机数值计算来检验中心极限定理，所用到的“ 个体 ”随机变量分别为[0-1）间均匀分布的随机数和高斯分布（用醉汉散步的理论产生）的随机

数。原始数据点由 c 程序产生，经 Origin 数量统计（Frequency account）后，被 Mathematica 画在了图上。理论曲线是 CLT 预测的高斯分布曲线。

为了画出“ **整体** ”随机变量的概率分布图，必须产生大量的“ 整体 ”随机变量，然后统计其出现的概率，画出统计图。在我的程序中，每处理一个样本是一次 trial，即每一个 trial 会产生一个“ 整体 ”随机变量。

（1） 用[0-1)间均匀分布的随机数来检验 CLT。

（说明：[0-1)间均匀分布的随机数的产生器来自于 *Numerical Recipes in C : the art of scientific computing* 这本书中 *Chapter 7: Random Numbers* 中推荐的 ran1 程序。）

```
trial = 10 001; n = 100; σ = Sqrt[1/12];
Show[Plot[1/500 1/Sqrt[2 π * σ^2 / n] * Exp[-x^2/(2 * σ^2 / n)], {x, -0.5, 0.5}, PlotRange → All, AxesOrigin → {0, 0}],
 ListPlot[Import["E:\Central limit theorem\countrand2.dat"], PlotStyle → RGBColor[1, 0, 0]]]
```

（作图的 Mathematica 代码）

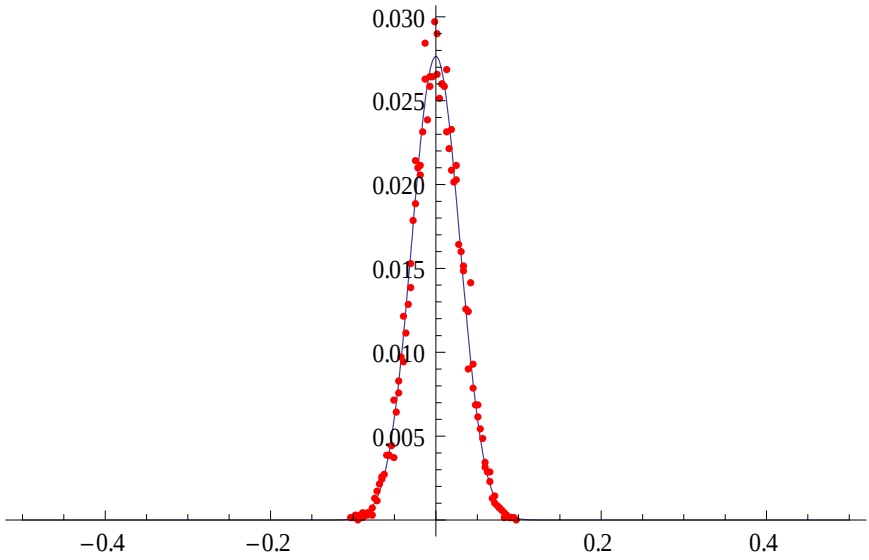



图 1：“ 整体 ”随机变量的分布图：横轴为其大小，纵轴为其出现的几率。均值为 0 是因为每个整体随机变量都减去了 0.5。

说明：程序运行了 10 001 个 trials，处理了 10 001 个样本，即产生了 10 001 个“ 整体 ”随机变量。每一个样本有 n=100 个[0-1) 间均匀分布的“ 个体 ”随机变量。注意[0-1)间均匀分布的随机数的方差 $\sigma^2 = \int_0^1 (x-\frac{1}{2})^2 dx = \frac{1}{12}$。图 1 的“ 整体 ”随机变量其分布的宽度 $\Delta x = 0.1-(-0.1) = 0.2$。

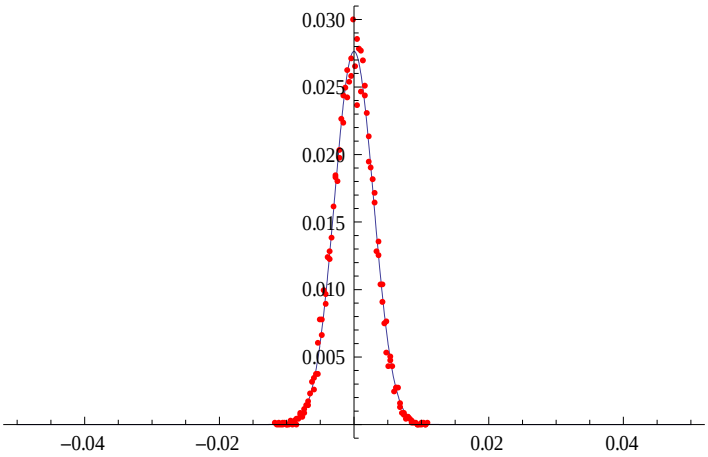
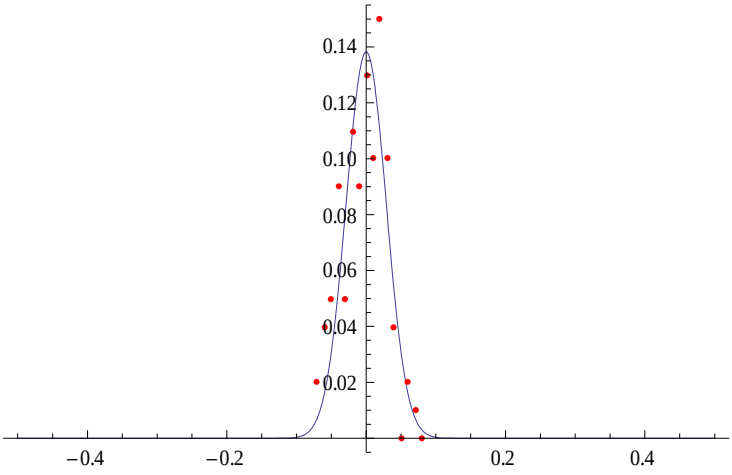



图 2：“ 整体 ” 随机变量的分布图：横轴为其大小，纵轴为其出现的几率。均值为 0 是因为每个整体随机变量都减去了 0.5。

说明：程序运行了 10 001 个 trials，处理了 10 001 个样本，即产生了 10001 个 “ 整体 ” 随机变量。每一个样本有了 n=10 000 个[0-1) 间均匀分布的 “ 个体 ” 随机变量。其分布的宽度 $\Delta x = 0.01 - (-0.01) = 0.02$ 。

图 2 的分布宽度比图 1 小了 10 倍，其原因是样本容量由 n=100 变到 n=10 000，增大了 100 倍，导致 “ 整体 ” 随机变量的方差减小了 100 倍，标准差减小了 $\sqrt{}100=10$ 倍，即图中的宽度减小了 10 倍。**中心极限定理得到了验证。**

另外，trial 的多少，即产生的 “ 整体 ” 随机变量的多少，**仅仅**影响数据点对理论曲线的符合程度。理论上，如果产生无穷多个 “ 整体 ” 随机变量（即运行无穷个 trial），数据点会完全吻合理论曲线。




程序运行了 101 个 trials，处理了 100 个样本，产生了 101 个 “ 整体 ” 随机变量。每一个样本有了 n=100 个[0-1) 间均匀分布的 “ 个体 ” 随机变量。其分布的宽度与上面的结论吻合 $\Delta x = 0.2$。但是因为 “ 整体 ” 随机变量太少，数据点与理论曲线符合不好。

（2） 用高斯分布的随机数来检验 CLT。

高斯分布用“ 醉汉散步 ”的理论产生，醉汉从坐标为 100 处出发，单位时间内以相同的概率向左或向右迈出一步（单位距离），T 时间后，n 个醉汉的分布为高斯分布。

为了避免出现的“ T 的奇偶性 ”问题（比如 T 为奇数时，醉汉最终的位置必在奇数坐标位置），在程序中，让 T 等概率的取 T 和 T＋1。

```
trial = 50 000; drunkard = 100; T = 100.5;
Show[Plot[1/20 * 1/Sqrt[2 π * T / drunkard] * Exp[-(x - 101)^2/(2 * T / drunkard)], {x, 98, 103}, PlotRange → All, AxesOrigin → {100, 0}],
 ListPlot[Import["E:\Central limit theorem\Count3.dat"], PlotStyle → RGBColor[1, 0, 0]]]
```

（作图用的 Mathematica 代码）

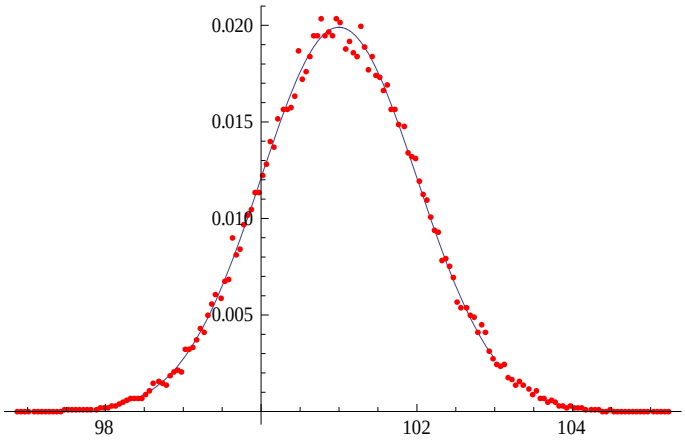



图 3：“ 整体 ”随机变量的分布图：横轴为其大小，纵轴为其出现的几率。

说明：程序运行了 50 000 个 trials，处理了 50 000 个样本，产生了 50 000 个“ 整体 ”随机变量。每一个样本有了 n=100 个[0-1) 高斯分布的“ 个体 ”随机变量（醉汉的最终位置），其平均值为 100，醉汉行走时间是 100.5 。其分布的宽度 $\Delta x = 104 - 98 = 6$。

不知什么原因，图中的平均值是 101，而非 100？

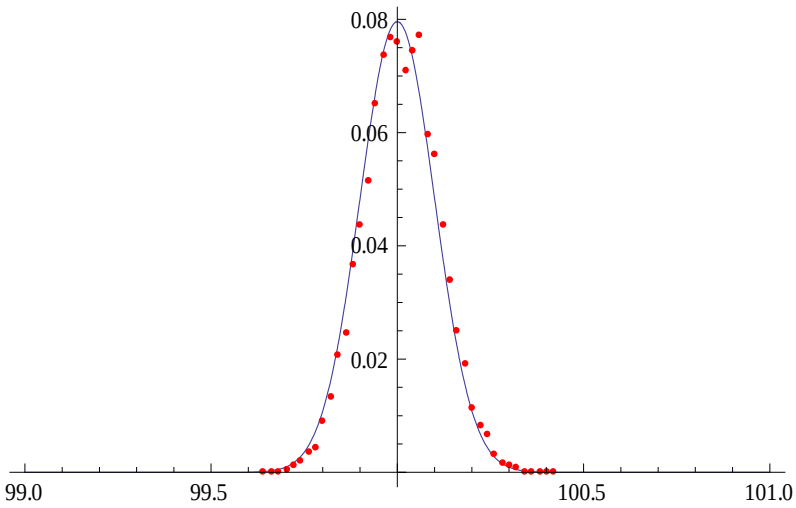



图 4：“ 整体 ” 随机变量的分布图：横轴为其大小，纵轴为其出现的几率。

说明：程序运行了 10 000 个 trials，处理了 10 000 个样本，即产生了 10 000 个 “ 整体 ” 随机变量。每一个样本有了 n=10 000 个[0-1) 高斯分布的 “ 个体 ” 随机变量（醉汉的最终位置），其平均值为 100，醉汉行走时间是 100.5。其分布的宽度 $\Delta x = 100.3 - 99.7 = 0.6$。

图4的分布宽度比图3小了10倍，其原因是样本容量由n=100变到n=10 000，增加了 100 倍，导致 “ 整体 ” 随机变量的方差减小了 100 倍，标准差减小了 $\sqrt{}100=10$ 倍，即图中的宽度减小了 10 倍。**中心极限定理得到了验证。**

**附：**均值为 $\mu$，方差为 $\sigma^2$ 的高斯分布（正态分布）表达式为：

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

均值为 $\mu$，方差为 $\sigma^2/n$ 的高斯分布（正态分布）表达式为：

$$\rho(x) = \frac{1}{\sqrt{2\pi\,\sigma^2/n}} \exp[-\frac{(x-\mu)^2}{2\,\sigma^2/n}]$$

其标准差变为了 $1/\sqrt{n}$，概率分布图上的宽度变为了 $1/\sqrt{n}$。

2009-7-28