

# 随机分布的生成与估计

---

李一鸣

2018 年 9 月 28 日

## 随机数和随机序列的产生

---

生成随机序列  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ ，其中每个  $X_i$  服从  $[-\frac{a}{2}, \frac{a}{2}]$  的均匀分布。

生成随机序列  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ ，其中每个  $Y_i$  服从  $[-\frac{a}{2}, \frac{a}{2}]$  的均匀分布。

蒙特卡罗投点法：

在边长为  $a$  的正方形内随机投点，设该点落入此正方形的内切圆中的概率为  $P_{circle}$ ，则：

$$\begin{aligned} P_{circle} &= \frac{S_{circle}}{S_{square}} \\ &= \frac{\pi(\frac{a}{2})^2}{a^2} \\ &= \frac{\pi}{4} \end{aligned} \tag{1}$$

假定生成的数据中有  $m$  个在圆内， $n$  个在圆外。则：

$$f_{circle} = \frac{m}{n} \tag{2}$$

对于任一点  $(X_i, Y_i)$ ，如果满足：

$$X_i^2 + Y_i^2 \leq (\frac{a}{2})^2 = \frac{a^2}{4} \tag{3}$$

则其在圆内，计入  $m$  中，否则计入  $n$  中。

以频率估计概率，我们有：

$$\pi = \frac{4m}{n} \tag{4}$$

实验结果：

在实验中取  $a=1$ （其实多少都没有关系，精确度只与样本数相关），取  $n=1\,000, 10\,000, 100\,000, 1\,000\,000$  分别进行实验。

```
n = 1000
m = 792

pi = 3.168
```

```
n = 10000
m = 7902

pi = 3.1608
```

```
n = 100000
m = 78567

pi = 3.14268
```

```
n = 1000000
m = 785107

pi = 3.140428
```

## 随机分布的计算机模拟

---

### 高斯分布的模拟

生成均值为  $\mu = 10$ 、方差为  $\sigma = 5$  的正态分布，并画出均值和方差随样本数增加而变化的图。

设总样本数为  $N$ ，记前  $n$  个样本数据的均值、方差分别为  $E_n$  和  $D_n$ ，则得到均值、方差矩阵：

$$\begin{aligned}\mathbf{E} &= (E_1, E_2, \dots, E_N) \\ \mathbf{D} &= (D_1, D_2, \dots, D_N)\end{aligned}\tag{5}$$

样本个数矩阵：

$$\mathbf{N} = (1, 2, \dots, N)\tag{6}$$

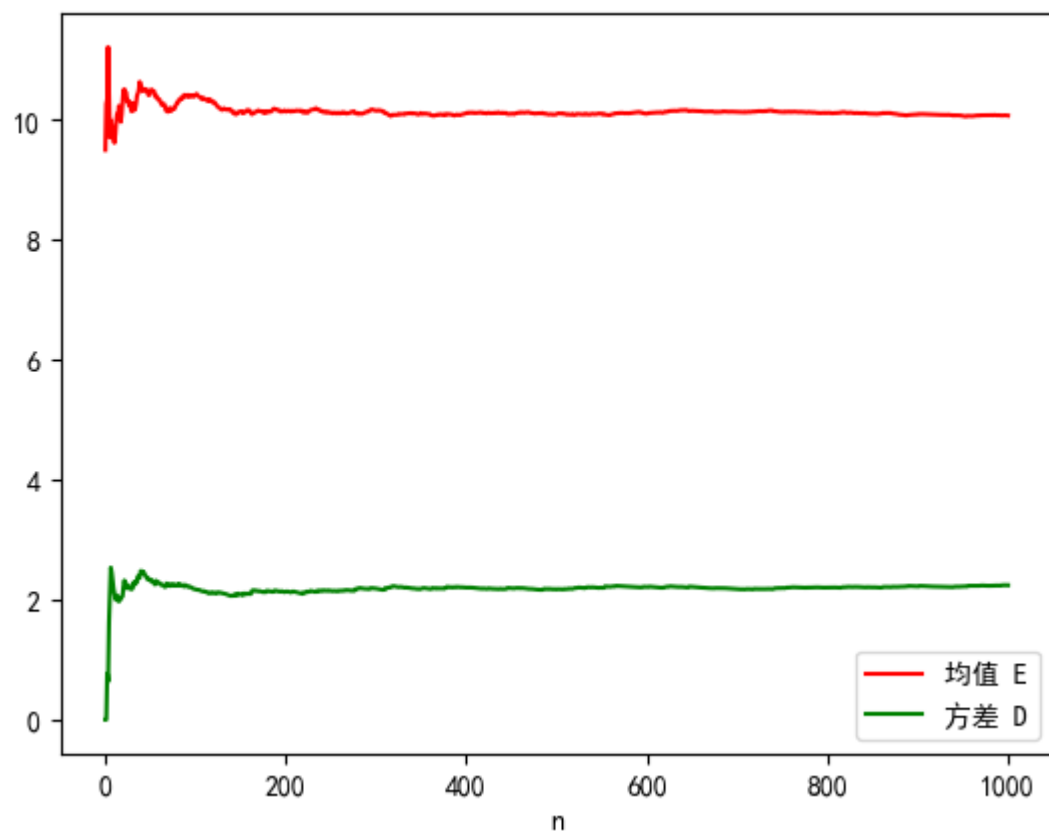
我们只需要作出  $(\mathbf{N}, \mathbf{E})$  和  $(\mathbf{N}, \mathbf{D})$  的图像即可。

实验结果：

在实验中取  $a = 1$ ，取  $N = 1\,000, 10\,000$  分别进行实验。

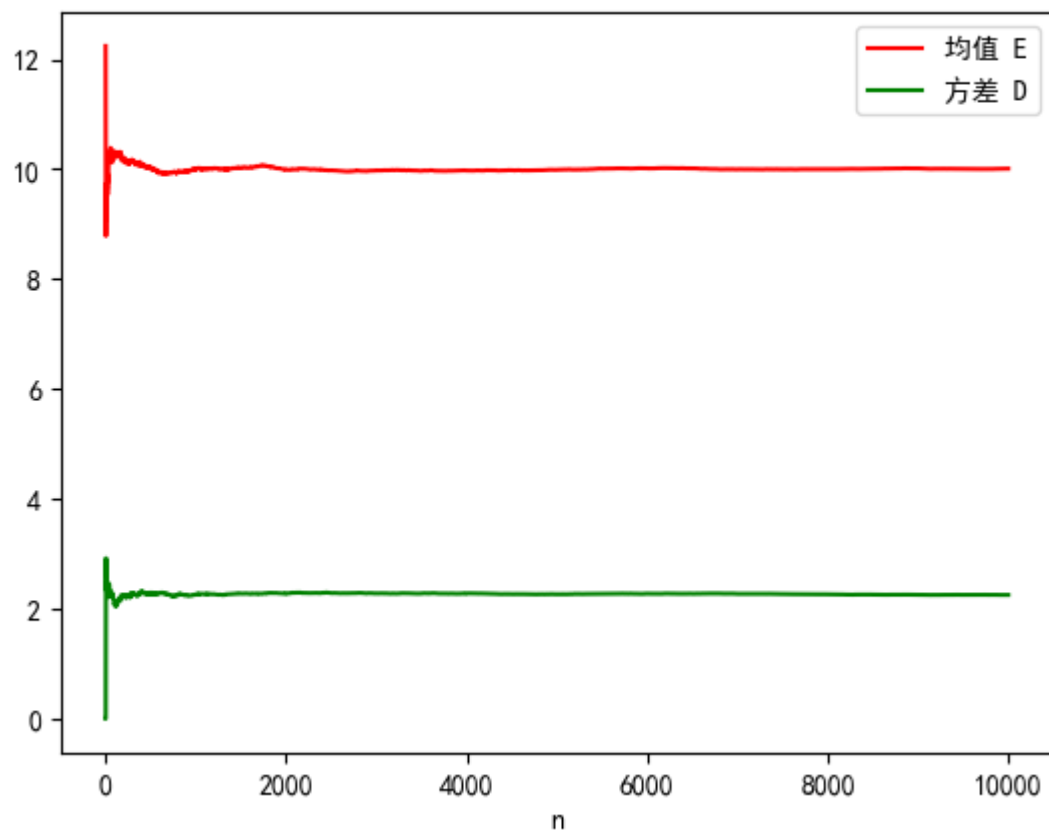
正态分布样本的均值、方差随样本数增加而变化的图像

$$\mu = 10.0, \sigma^2 = 5.0, N = 1000$$



正态分布样本的均值、方差随样本数增加而变化的图像

$$\mu = 10.0, \sigma^2 = 5.0, N = 10000$$



## 敌军坦克到达情况的模拟

敌军坦克分队到达我方阵地规律服从泊松分布，平均每分钟到达  $\lambda$  辆。

泊松分布的期望值是  $\lambda$ ，也就是说在一分钟之内，到达的坦克数量  $T$  的分布列为：

$$P(T = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (7)$$

我们可以生成  $N$  组数据  $\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_N$ ，分别用它们的均值  $E(\mathbf{T}_1), E(\mathbf{T}_2), \dots, E(\mathbf{T}_N)$  表示第  $1, 2, \dots, N$  分钟内到达的坦克数量存入  $\mathbf{A} = (A_1, A_2, \dots, A_N)$  中，则在  $N$  分钟内坦克到达总数量  $A_{total}$  满足：

$$A_{total} = \sum_{n=1}^N A_n \quad (8)$$

实验结果：

取  $\lambda = 4, N = 3$ ，对样本数进行改变，得到如下实验结果：

```
size = 1000
lambda = 4.0
N = 3

A = [4.078 4.006 3.855]
A_total = 11.939
```

```
size = 10000
lambda = 4.0
N = 3

A = [4.0126 4.0009 3.9689]
A_total = 11.9824
```

```
size = 100000
lambda = 4.0
N = 3

A = [3.98306 4.0062 3.99864]
A_total = 11.9879
```

```
size = 1000000
lambda = 4.0
N = 3

A = [4.000975 4.000943 4.001314]
A_total = 12.003232
```

每辆敌军坦克到达的时刻服从期望为  $\frac{1}{\lambda}$  的指数分布，也就是说坦克到达的时间  $S$  的分布函数为：

$$F_S(x) = e^{-\lambda x} \quad (9)$$

我们可以生成  $M$  组数据  $\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_M$ ，分别用它们的均值  $E(\mathbf{S}_1), E(\mathbf{S}_2), \dots, E(\mathbf{S}_M)$  表示第  $1, 2, \dots, M$  辆坦克到达所需的时间，将其存入  $\mathbf{B} = (B_1, B_2, \dots, B_M)$  中，则在  $N$  分钟内每辆敌军坦克到达时间为：

$$\mathbf{B}' = (B'_i = \sum_1^j B_j \mid j \in [1, M] \text{ where } B'_i < N) \quad (10)$$

实验结果：

取  $\lambda = 4, N = 3$ ，对样本组数  $M$  进行改变，得到如下实验结果：

```
size = 1000000
lambda = 4.0
M = 1000
N = 3

B' = [0.          0.03970983 0.66690408 0.75143449 1.02948559 1.19015077
1.5940807  1.66309342 1.75191405 1.90816193 1.93829474 2.03826935
2.04132296 2.04792707 2.17952641 2.50712507 2.59624773 2.61581964
2.81367139 2.9882194 ]
size of B' = 20
```

```
size = 1000000
lambda = 4.0
M = 10000
N = 3

B' = [0.          0.12969758 0.54244581 1.2514086  1.65576013 1.80392056
2.08619065 2.09500261 2.10120925 2.73365359 2.83698835]
size of B' = 11
```

```
size = 1000000
lambda = 4.0
M = 100000
N = 3

B' = [0.          0.1147557  0.4439659  0.5103366  0.81882969 1.08811278
1.15349092 1.34232144 2.03658054 2.14223747 2.24801589 2.45649107
2.46995655 2.6791674  2.8093229 ]
size of B' = 15
```

```

size = 1000000
lambda = 4.0
M = 1000000
N = 3

B' = [0.          1.23851918  1.41039667  1.61506133  1.6686324   2.26447849
       2.37059676  2.43934594  2.87615955  2.93745805  2.96028765]
size of B' = 11

```

## 基于高斯分布混合模型的模式分类方法

考虑水果聚类问题，水果的属性  $\mathbf{X}$  满足高斯分布，其均值向量、协方差矩阵分别为  $\mu, \Sigma$ ，将其概率密度记为  $p(\mathbf{X}|\mu, \Sigma)$ 。

定义高斯混合分布：

$$\begin{aligned}
 p_M(\mathbf{X}) &= \sum_{i=1}^k \alpha_i p(\mathbf{X}|\mu_i, \Sigma_i) \\
 p(\mathbf{X}|\mu_i, \Sigma_i) &= \frac{\exp\{-\frac{1}{2}(\mathbf{X} - \mu_i)^T \Sigma_i^{-1}(\mathbf{X} - \mu_i)\}}{(2\pi)^{D/2} |\Sigma_i|^{\frac{1}{2}}}
 \end{aligned} \tag{11}$$

该分布由  $k$  个混合分布组成，每个混合成分对应一个高斯分布，其中  $\mu_i, \Sigma_i$  是第  $i$  个高斯混合成分的参数， $\alpha_i$  为选择第  $i$  个混合成分的概率，满足：

$$\alpha_i > 0, \sum_{i=1}^k \alpha_i = 1 \tag{12}$$

记样本  $X_j$  属于第  $i$  个高斯成分的后验概率为  $y_{ji}$ ，有：

$$\begin{aligned}
 y_{ji} &= \frac{\alpha_i p(x_j|\mu_i, \Sigma_i)}{p_M(x_j)} \\
 &= \frac{\alpha_i p(x_j|\mu_i, \Sigma_i)}{\sum_{l=1}^k \alpha_l p(x_j|\mu_l, \Sigma_l)}
 \end{aligned} \tag{13}$$

为了得到混合分布的各个组成部分的分布参数，我们需要利用 EM 算法 (Expectation-maximization algorithm) 不断迭代来获取  $k$  个类的均值和方差参数。

**E 步：**

根据当前参数计算样本后验概率  $\mathbf{Y} = (y_{ji})_{ji}$

**M 步：**

根据后验概率更新模型参数  $\{\alpha_i, \mu_i, \Sigma_i | 1 \leq i \leq k\}$ ，新的参数与后验概率应该满足下面的关系：

$$\begin{aligned}\alpha'_i &= \frac{\sum_{j=1}^N y_{ji}}{N} \\ \mu'_i &= \frac{\sum_{j=1}^N y_{ji} x_j}{\sum_{j=1}^N y_{ji}} \\ \Sigma'_i &= \frac{\sum_{j=1}^N y_{ji} (x_j - \mu'_i)(x_j - \mu'_i)^T}{\sum_{j=1}^N y_{ji}}\end{aligned}\tag{14}$$

不断重复 E、M 两步直到收敛。

实验结果：

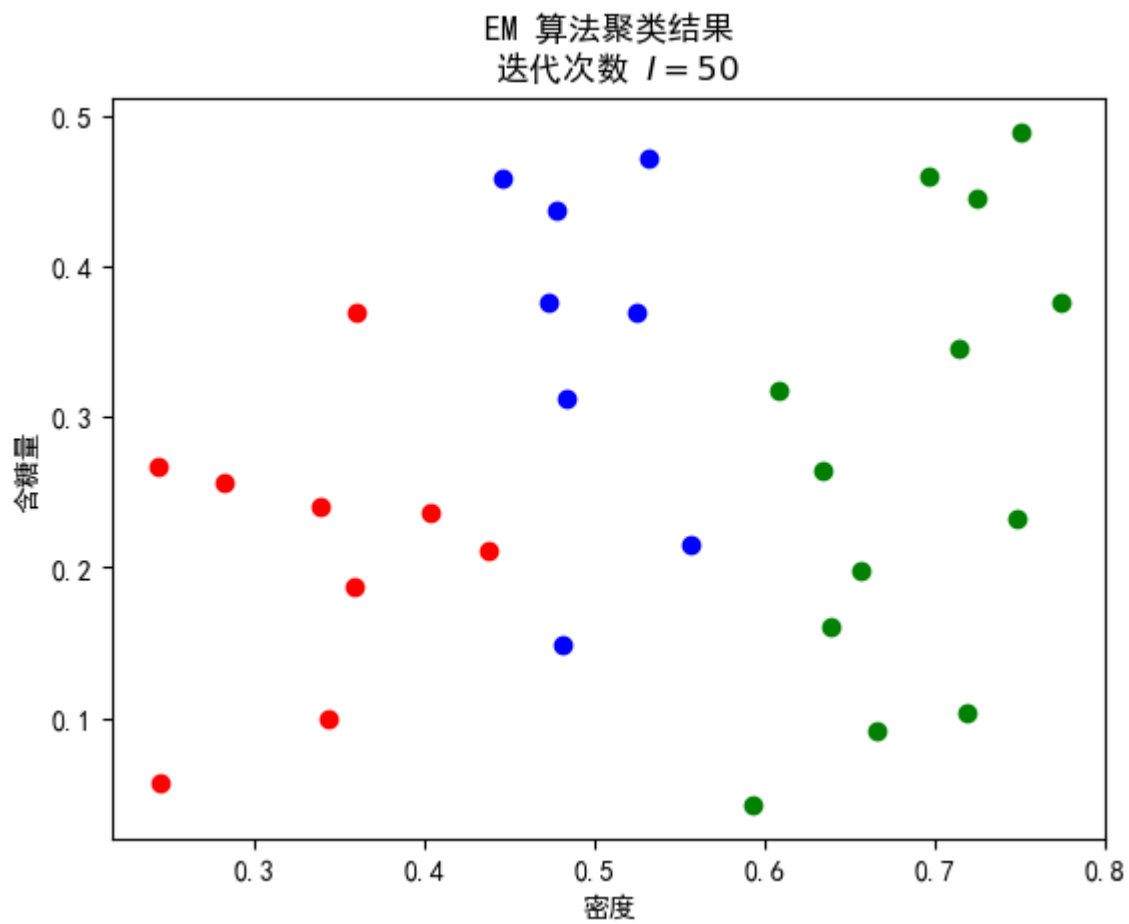
现有水果数据  $\mathbf{S}$ ：

$$\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_N)\tag{15}$$

其中  $N = 30$ ， $\mathbf{S}_i$  为二维列向量，包含密度、含糖率两个属性，我们随机初始化一组参数：

$$\begin{aligned}\alpha_1 &= \alpha_2 = \alpha_3 = \frac{1}{3} \\ \mu_1 &= \mathbf{S}_6, \mu_2 = \mathbf{S}_{22}, \mu_3 = \mathbf{S}_{27} \\ \Sigma_1 &= \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.1 & 0.0 \\ 0.0 & 0.1 \end{pmatrix}\end{aligned}\tag{16}$$

令迭代次数  $I = 50$ ，将得到的结果以散点图绘制如下：



详细计算过程参见 `em-50.txt`。

## 参考文献

1. [\[Monte Carlo method | Wikipedia\]](#)
2. [\[Expectation-maximization algorithm\]](#)