数据科学和大数据处理第一次作业

刘群*

地球系统科学研究中心

Center for Earth System Science

2015 年 4 月

1. 假设 p 是分布在 0 到 1 之间的均匀分布 (UniformDistribution), 即关于 x 的概率密度函数 p 满足: x 在 0 和 1 之间时 $p(x) = 1$, 其余 x 取值时 $p(x) = 0$。推导关于 p 的两个无关采样 x_1 和 x_2, 其平均值的分布与方差特征。并使用 MATLAB 验证多次采样下的均值分布趋近于正态分布。

一般来说，多次采样的均值和方差的无偏估计分别为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

从而有，当 $n = 2$ 时，均值为

$$\bar{x} = \frac{x_1 + x_2}{2}$$

方差为

$$s^2 = \frac{1}{2-1}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2\right) = \frac{(x_1 - x_2)^2}{2}$$

下面用 MATLAB 验证多次采样下的均值分布趋于正态分布, 如图??所示，当采样点个数为 1 时，样本是均匀分布，随着采样点的数目增加，比如说增加到 30 时，所有样本点的均值已经趋于正态分布。

2. 设定基于随机变量 x 的 n 个采样的偏斜度 (Skewness) 样本估计为:

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{3}{2}}}$$

其中 x_i 为 x 的第 i 个采样, 而 x 为其样本均值。对以下问题进行说明:

1) b_1 是否为对真实偏斜度的无偏估计;

2) b_1 是否能正确估计概率密度分布的偏斜度其符号 (即左偏或右偏).

*电子邮件: liu-q14@mails.tsinghua.edu.cn，学号: 2014211591

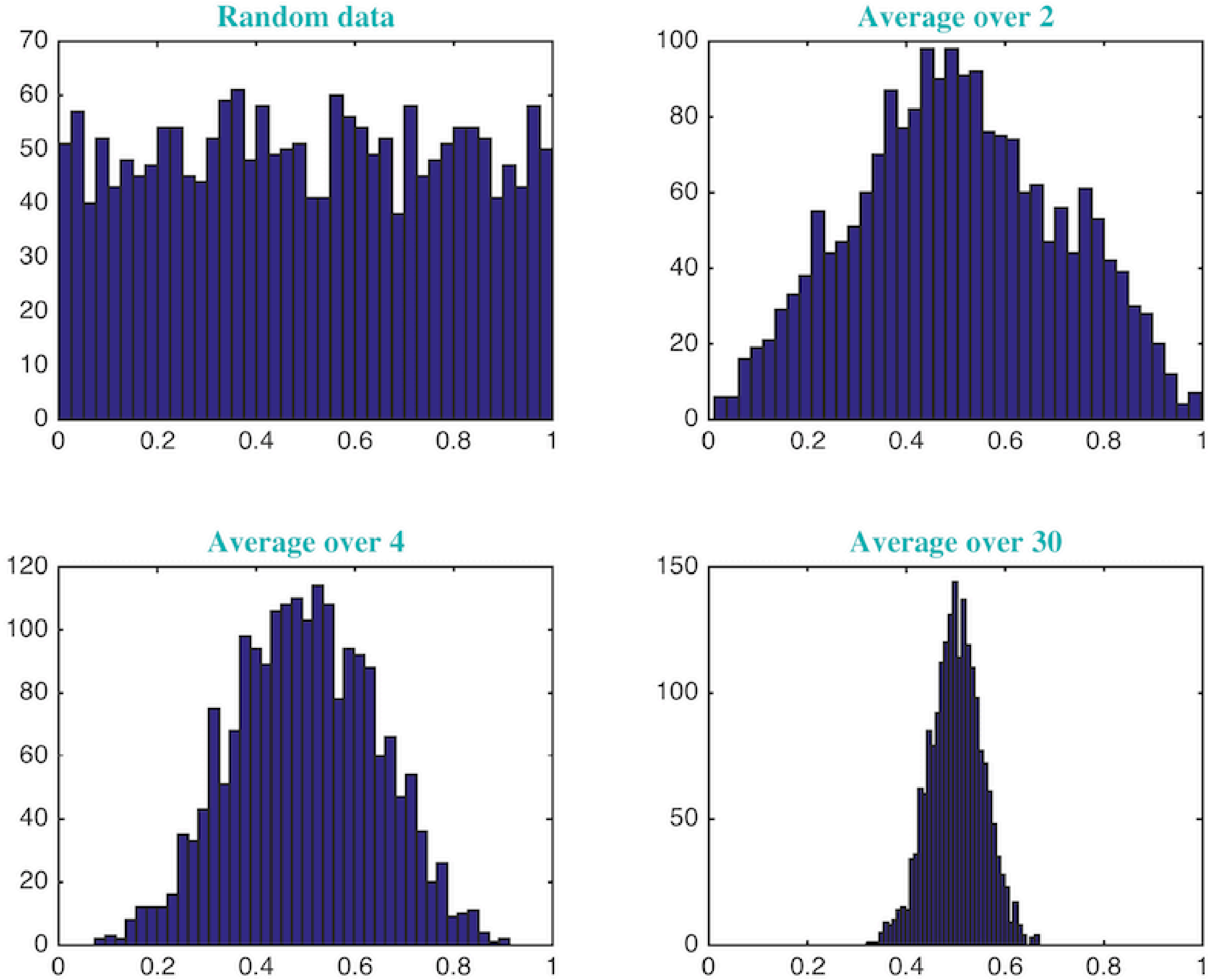

图 1: 中心极限定理的验证

(1) b_1 不是为对真实偏斜度的无偏估计, 原因如下：随机变量 X 的偏斜度 γ_1 的定义为:

$$
\begin{aligned}
\gamma_1 = & E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \\
= & \frac{E\left[X^3 - 3X^2\mu + 3X\mu^2 + \mu^3\right]}{\sigma^3} \\
= & \frac{E\left[X^3\right] - E[3X^2\mu] + E[3X\mu^2] - E\left[\mu^3\right]}{\sigma^3} \\
= & \frac{E\left[X^2\right] E[X] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{\sigma^3} \\
= & \frac{(\mu^2+\sigma^2)\mu - 3\mu(\mu^2+\sigma^2) + 3\mu^3 + \mu^3}{\sigma^3} \\
= & \frac{-2\mu\sigma^2}{\sigma^3} = \frac{-2\mu}{\sigma}
\end{aligned}
$$

$$
\begin{aligned}
E(b_1) &= E\left(\frac{\frac{1}{n}\sum i=1^n(x_i-\bar{x})^3}{\left[\frac{1}{n-1}\sum i=1^n(x_i-\bar{x})^2\right]^{\frac{3}{2}}}\right)=\frac{\frac{1}{n}\sum\limits_{i=1}^{n}E(x_i-\bar{x})^3}{E(s^3)}=\frac{E(x_i-\bar{x})^3}{E(s^2)E(s)}\\
&=\frac{E(x_i^3-3x_i^2\bar{x}+3x\bar{x}^2-\bar{x}^3)}{\sigma^2E(s)}=\frac{E(x_i^3)-3E(x_i^2\bar{x})+3E(x\bar{x}^2)-E(\bar{x}^3)}{\sigma^2E(s)}\\
&=\frac{(\mu^2+\sigma^2)\mu-3\left[\frac{-2\mu\sigma^2}{n}+\frac{n+2}{n}\sigma^2\mu+\mu^3\right]+2E(\bar{x}^3)}{\sigma^2E(s)}(\text{Because } E(x_i\bar{x}^2)=E(\bar{x}^3))\\
&=\frac{(\mu^2+\sigma^2)\mu-3\left[\frac{-2\mu\sigma^2}{n}+\frac{n+2}{n}\sigma^2\mu+\mu^3\right]+2\left(-2\mu\sigma^2/n^2+3\sigma^2\mu/n+\mu^3\right)}{\sigma^2E(s)}\\
&=\frac{-2(n-2)(n-1)\mu\sigma^2}{n^2\sigma^2E(s)}=\frac{-2\mu}{E(s)}\frac{(n-2)(n-1)}{n^2}\\
&\neq\frac{-2\mu}{\sigma}
\end{aligned}
$$

故 b_1 是真实偏斜度的有偏估计。

(2) 当 n 很大时，b_1 可以看成是真实偏斜度的一个近似估计，因此可以正确估计偏斜度的符号，但是当 n 很小、即样本比较少时，可能会造成一些误判。

3. 证明最优插值 (Optimal Interpolation, OI) 所导出的增益矩阵:

$$\left(B^{-1}+H^TR^{-1}H\right)^{-1}H^TR^{-1}$$

与基于三维变分法 (Three-Dimensional Variational Method, 3D-Var) 导出的增益矩阵:

$$\left(BH^T\right)\left(R+HBH^T\right)^{-1}$$

相等。

证明：分别在两个矩阵的左边乘以 $\left(B^{-1}+H^TR^{-1}H\right)$, 右边乘以 $\left(R+HBH^T\right)$, 有

最优插值的增益矩阵变为:

$$
\begin{aligned}
&\left(B^{-1}+H^TR^{-1}H\right)\left(B^{-1}+H^TR^{-1}H\right)^{-1}H^TR^{-1}\left(R+HBH^T\right)\\
&=H^TR^{-1}\left(R+HBH^T\right)\\
&=H^T+H^TR^{-1}HBH^T
\end{aligned}
$$

基于三维变分法的增益矩阵变为:

$$
\begin{aligned}
&\left(B^{-1}+H^TR^{-1}H\right)\left(BH^T\right)\left(R+HBH^T\right)^{-1}\left(R+HBH^T\right)\\
&=\left(B^{-1}+H^TR^{-1}H\right)\left(BH^T\right)\\
&=H^T+H^TR^{-1}HBH^T
\end{aligned}
$$

对比上面最终的结果可以看到，两种方法导出的增益矩阵是相等的。