核密度估计的 R 模拟报告

数 41 李博扬 2014012118

【模拟目的】

在 R 中，利用 density 计算混合正态分布核密度估计(kernel density estimation)。并与准确值比较评价估计好坏。

【模拟步骤】

1、 生成样本量为 n（分别取 n=30 和 100），满足混合正态分布$pN(0,1)+(1-p)N(2.5,1)$的随机样本（其中 p=1/3）。

通过书面计算可知其 pdf 为

$$f(x)=\frac{1}{3\sqrt{2\pi}}e^{-\frac{x^2}{2}}+\frac{2}{3\sqrt{2\pi}}e^{-\frac{(x-2.5)^2}{2}}$$

2、 利用 density 计算出核密度估计函数。

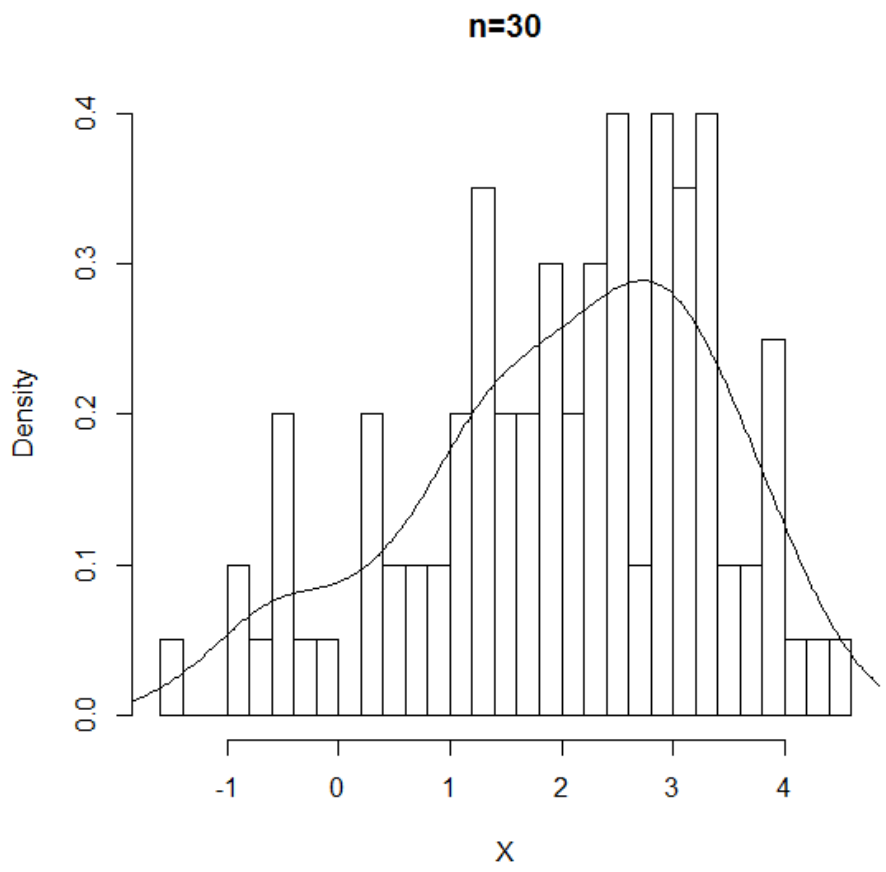
3、 将估计得到的函数图像与混合正态分布密度直方图画在同一图中。

4、 计算估计的 MISE(平均积分平方误差)，评价核密度估计的准确性

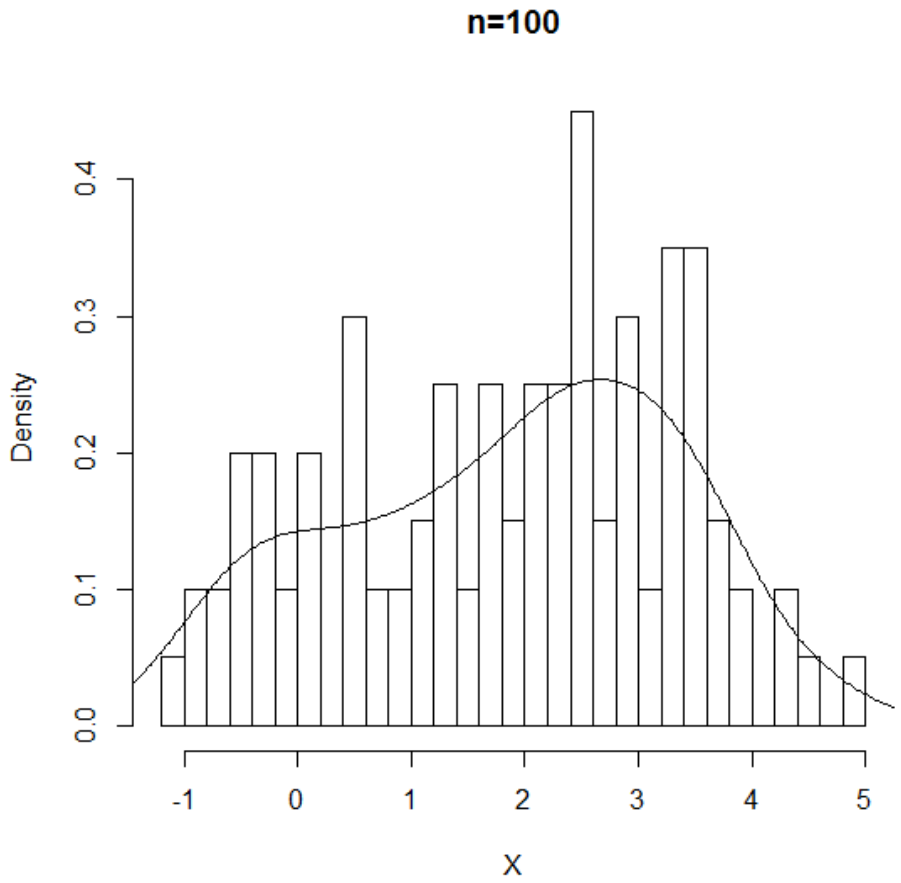
【模拟结果】

一、估计函数与总体的概率密度分布直方图

1、 n=30



2、 n=100



二、计算 MISE

重复模拟 1000 次取 MISE 的均值，结果如下：

	MISE
n=30	0.004684
n=100	0.003595

【模拟结论】

1、 从图中直观上来说，核密度估计与总体概率分布直方图有较好契合性。且 n 越大，契合度越高，估计越精确。

2、 从计算得出的 MISE 结果来看，无论 n 取值如何，估计得到平均积分平方误差数量级大约在10^{-3}左右，较为精确。

对核密度估计的看法：

核密度估计的原理其实是很简单的。在我们对某一事物的概率分布的情况下。如果某一个数在观察中出现了，我们可以认为这个数的概率密度很大，和这个数比较近的数的概率密度也会比较大，而那些离这个数远的数的概率密度会比较小。基于这种想法，针对观察中的第一个数，我们都可以 f(x-xi)去拟合我们想象中的那个远小近大概率密度。当然其实也可以用其他对称的函数。

针对每一个观察中出现的数拟合出多个概率密度分布函数之后，取平均。如果某些数是比较重要，某些数反之，则可以取加权平均。但是核密度的估计并不是，也不能够找到真正的分布函数。在某些极端情况下是不适用的。