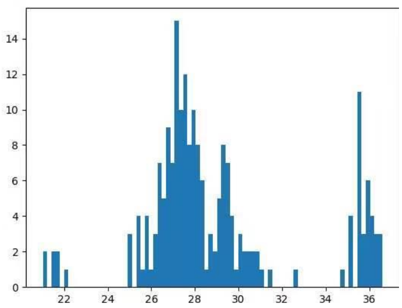
Assignment1 作业报告

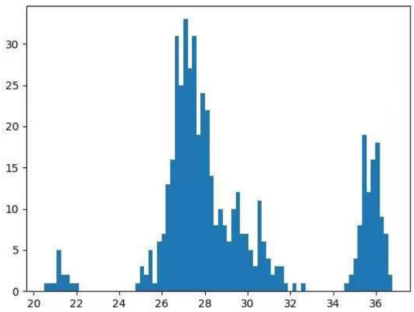
10%：观察数据量的改变对估计结果的影响

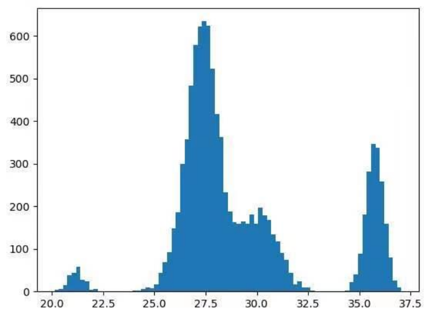
三种估计方法结果相似。随着数据样本量的增大，估计算法得出的数据分布曲线更加平滑，更能体现一般化的分布。在样本很少的时候，统计出的结果十分杂乱，几乎不能体现概率分布的峰值等统计性质。随着数据增多，分布估计的曲线变得平滑，更直观，更能让我们了解真实的观察数据分布。

以直方图的情况为例（从左至右分别是数据量 N=200, 500, 10000）

可以发现直方图估计出的结果和我们上述结论十分吻合。事实上，核密度估计和 k 近邻估计的结果也符合我们的结论。这说明，数据样本的量越大，我们对数据分布的估计越能趋近于真实情况。







20%：直方图估计中，块数对估计结果的影响

从数学的角度来分析，直方图以落在块中的点数个数来近似统计概率密度：

$$p(xi) = Ni * bins / N$$

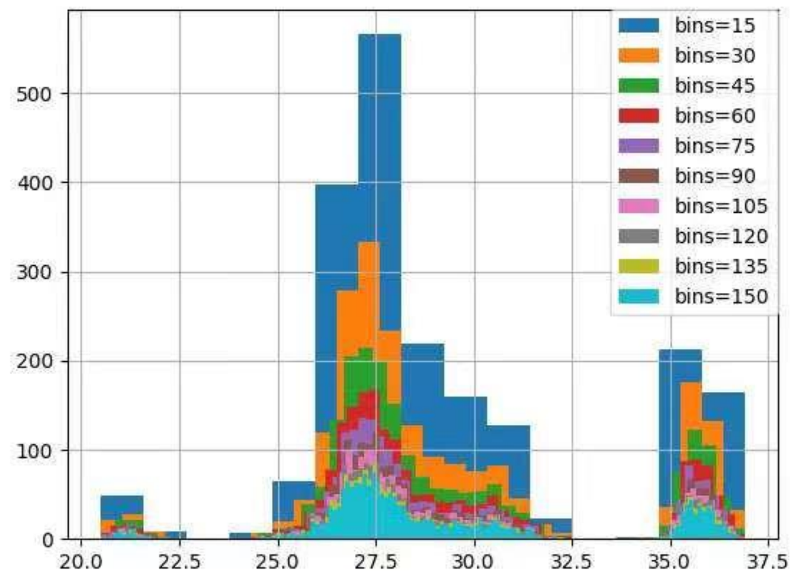
bins / N 越大，Ni 越小，函数 p(x)的频率越高，图像也越接近曲线。这意味着当块数不多时，图像失真程度会比较严重，即所谓的过拟合。在块数增大时图像曲线耿平滑，估计结果更加趋于真实分布。

另一方面，在块数非常大的时候，阻碍我们估计真实程度的因素落到了样本量上。当频数增大时，样本量的不够导致不能保证每个块内都落有点（甚至，大部分块中都没有点），

这样会导致图像的缺失增加，结果欠拟合，反而影响我们直观的观察分布。

因此，选取适当块数 bins 的值有助于我们观察分布。在样本量 N=2000 的时候，我们对样本量做直方图估计，得出结果如下。我们发现当 bins=15 的时候，一个块过大，不能很好的体现数据分布；当 bins=150 的时候，分布结果的整体高度却非常小，与真实结果有很大偏差。

这说明我们对这个问题的分析大致符合实际情况。在 N=2000 的这个样本中，bins=45 的结果相对更好。



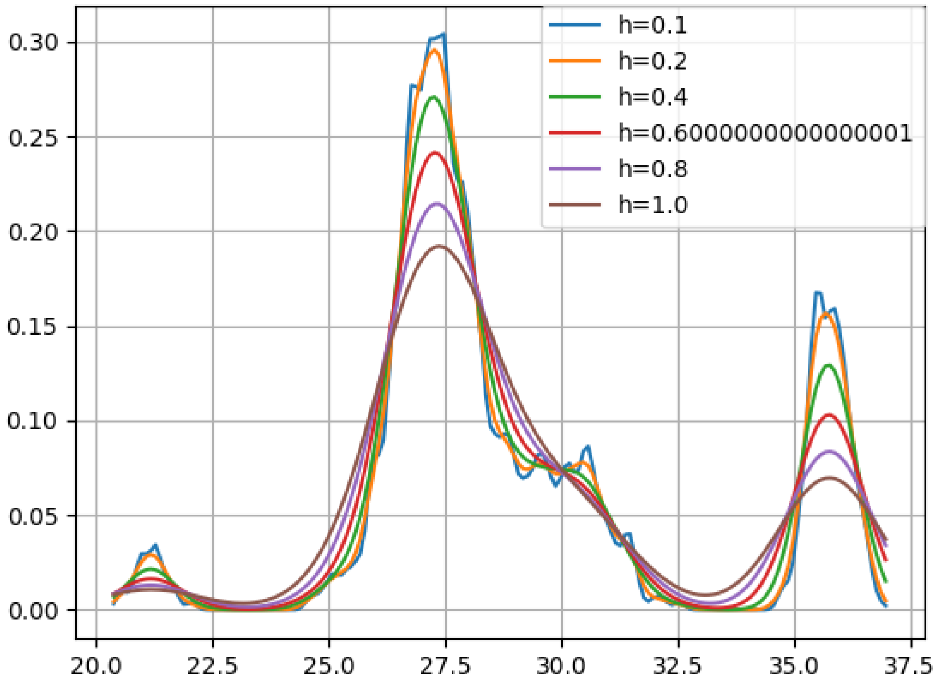
30%：核密度估计中，邻域大小 h 对估计结果的影响

本次作业选用高斯核的核函数，计算概率密度 p(x)为：

$$\mathrm{p(x)} = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\frac{1}{\sqrt{2\pi h^2}}e^{-\frac{|x-x_n|^2}{2h^2}}$$

当 h 非常大的时候，邻域过大，这会导致邻域中点概率密度的计算结果被 “平均”，即图像出现 “欠拟合” 现象。另一方面，在 h 很小的时候，邻域过小，图像分布的细节被强调的太重，导致过拟合，图像无法很好的诠释一般情况。

如图，在 N=2000 的样本上做核密度估计。以 0.1 为精度撒点近似实现上述连续函数。



显然地，h=0.1，上图中的蓝色曲线相比其他更加不光滑；同时棕色的曲线（即 h=1）的高度明显低于其他曲线，可见真实数据的一些分布被忽略了。这个结果类似于上图的直方图估计，最佳带宽应是适中的。这也能体现出核密度估计的原理是直方图估计的延伸。

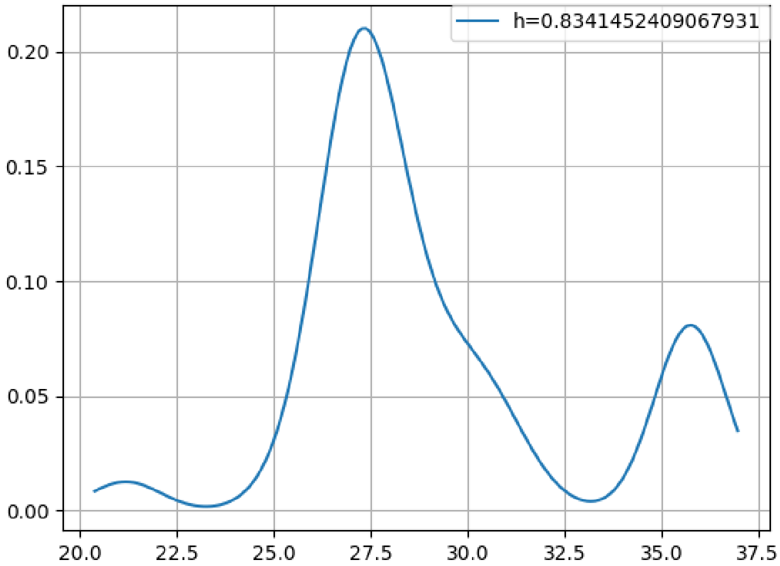
下面来探讨最佳的带宽 h′最佳取值选择。既然要寻找最值，我们首先需要定义误差函数。不妨采用比较常见的均平方积分误差函数，即：

$$\mathrm{MISE(h)} = \int \left(\mathrm{F(x)} - \mathrm{f(x)}\right)^2 dx$$

这次样本采用的核函数是高斯核，根据已有结论

$$\mathrm{h} = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}}$$

经过计算，取 n=2000，标准差σ=3.59867。于是可以计算处 h=0.834。作密度曲线如



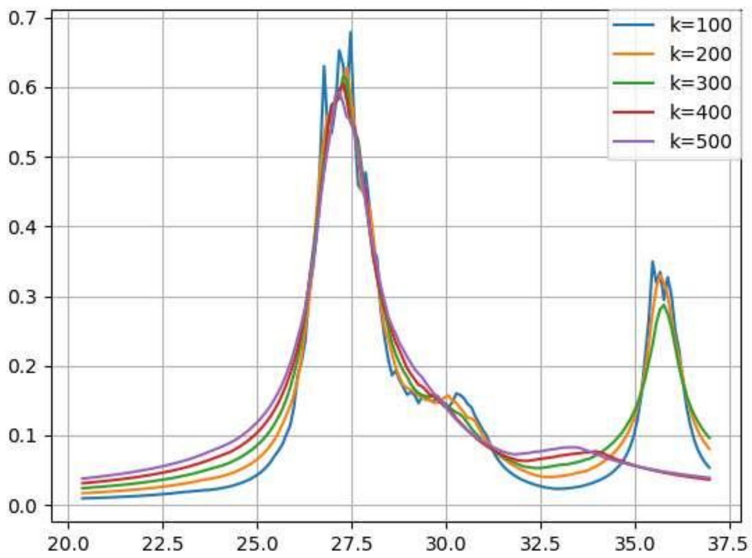
下图：

40%：k 近邻估计中，近邻数 k 对估计结果的影响

若说核密度估计是固定了邻域 h，统计落入邻域中点的数量，那么 k 近邻估计正好相反。算法中固定了点的数量 K，转而计算区间的长度：

$$p(x) = K / (N * V)$$

k 的变化对分布结果影响类似于前两种方法。如下图 k=100, 200, 300, 400, 500 时的分布。当 k 校时，图像仍有非常突兀的尖峰，这显示出图像此时欠拟合；当 k 大的时候，曲线光滑，但是在峰值处高度略低，且整体过于平滑，真实数据的一些分布特征被忽视（比如右边的峰）。



下面分析 k 近邻的分布不总是收敛到 1。根据我们对 p(x)的定义：

$$\int_{-\infty}^{\infty} p(x)dx = \sum_{\Delta V} \frac{K}{N * V} \Delta V = \frac{K}{N} \sum_{\Delta V} \frac{1}{V} \Delta V = \frac{K}{N} \int \frac{1}{x} dx$$

而 1/x 的积分不收敛，这是显然的。故而我们可以得出与题设相同的结论。

PS. 关于代码

本题的图通过 python3 作得，在核密度和 k 近邻估计中，以 0.1 为精度撒点而作图。因为本题的作图需求非常低，所以如 h, k, interval_size(撒点精度)等参数均为手动调整。

源代码实现了三种作图的方法，同时保留了一部分主程序的作图逻辑。