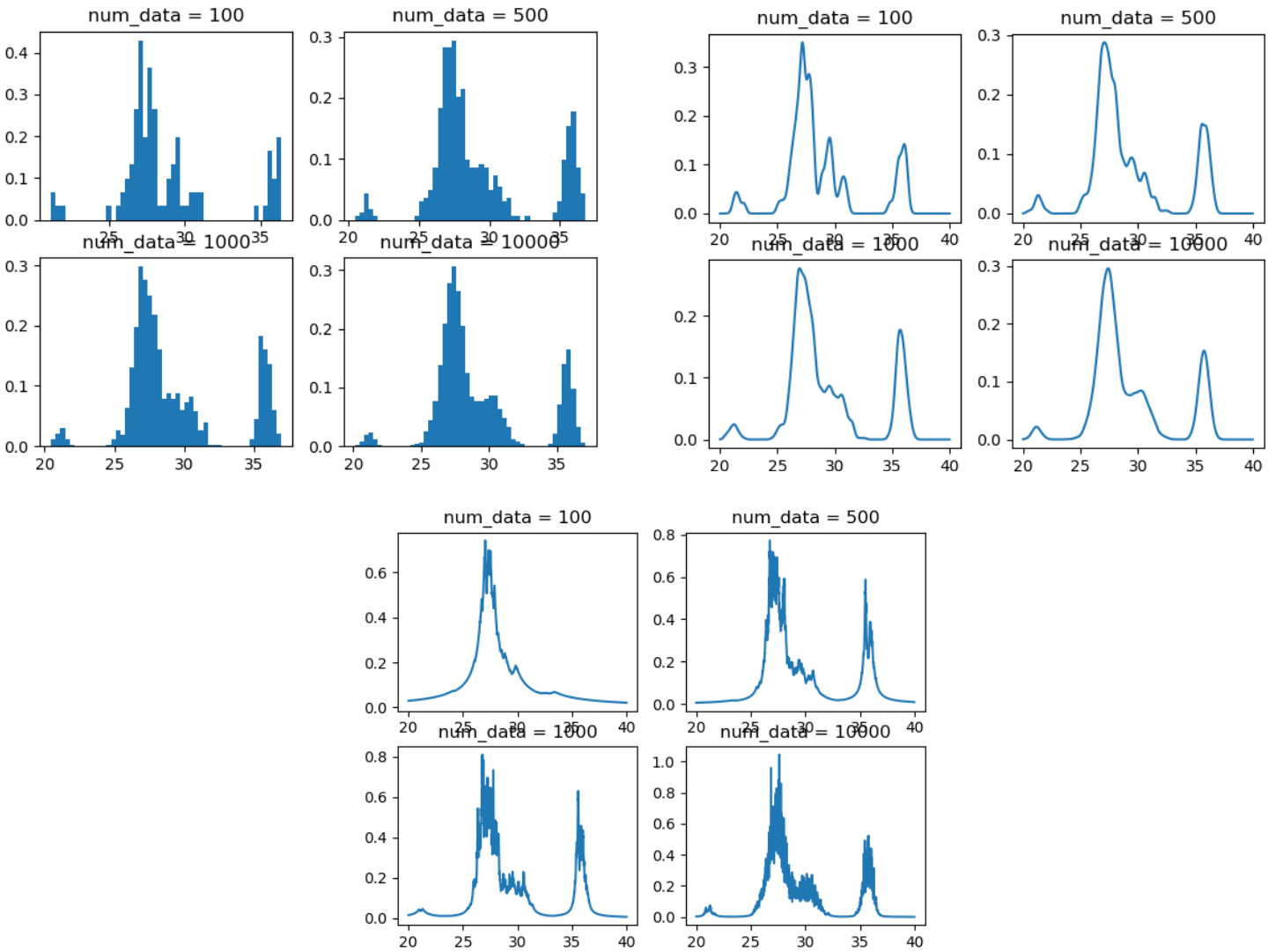
# Assignment 1 Brief Report

## 1. Size of training data

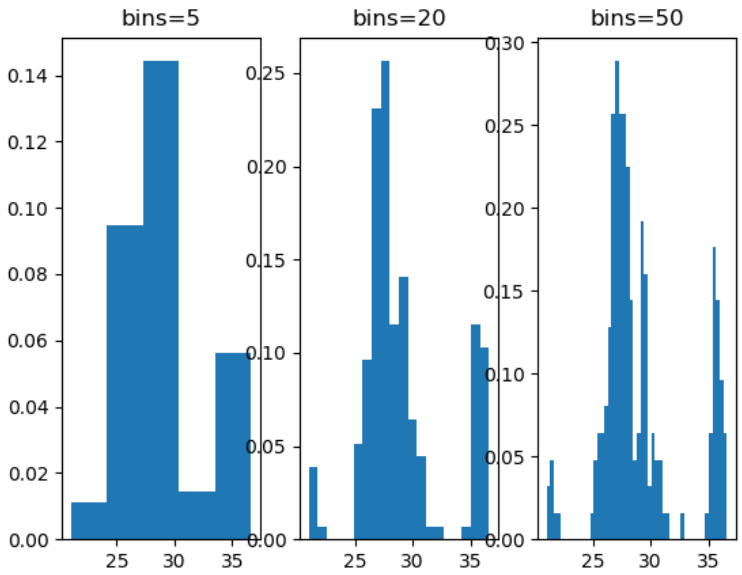
总的来说，数据量越大，三种密度估计方法的结果都会变得更好，同时对于参数的敏感性都会相对下降。对于直方图估计法，数据量的增加抑制了**随机性误差**对于密度估计的影响；对于核密度法，从下图中明显可以看出数据量增大之后的概率密度**图像更加平滑**（并且接近真实分布）；对于 KNN 来说，虽然数据量的增大带来了**巨大的波动和方差**，但是我们可以明显地从 10000 的图中直观的看到第二个峰值，这是前两幅图都没有的效果。




## 2. Size of bins for histogram

直方图法简单易懂，而且没有参数，完全的使用蒙塔卡罗估计，通过采样对概率密度的分布进行估计。但是同时也带来了一个问题就是人为的引入了不连续性，i.e. 在 bin 的边缘处不可避免的有断崖情况出现。当 bin 的数量比较小的时候（e.g. 5），这种现象尤其明显；但是当 bin 特别多的时候，又会出现某些 bin 中没有数据落入的情况（毕竟采样过程是离散的），因此我们需要选择适当的 bin 大小，不能太大也不能太小。

实际中进行 bin 大小的选择的时候，可以直观的根据概率密度的连续程度只管判断 bin 的大小，或者可以根据一些经验性的取值公式进行选择，主要有以下三种：

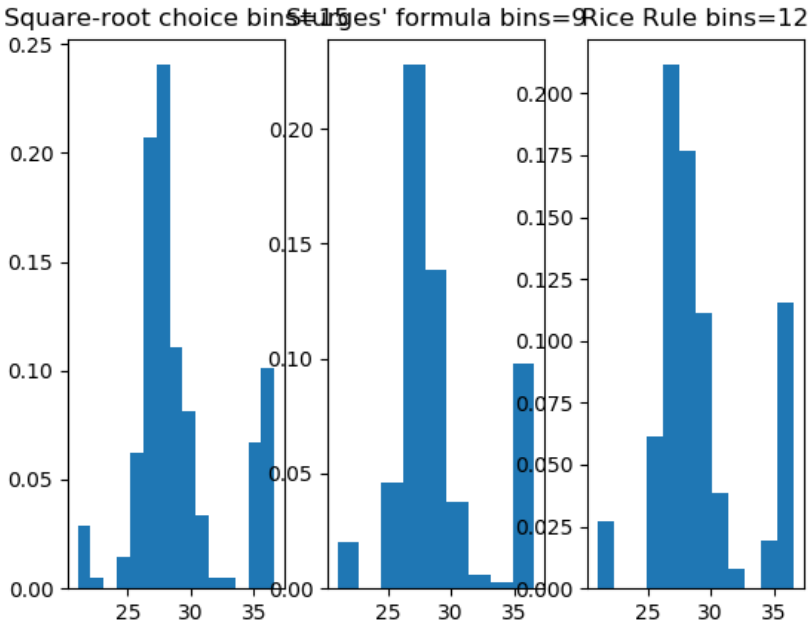



① Square-root choice：bins=⌈$\sqrt{n}$⌉

② Sturges' formula：bins=⌈$log2n$⌉+1

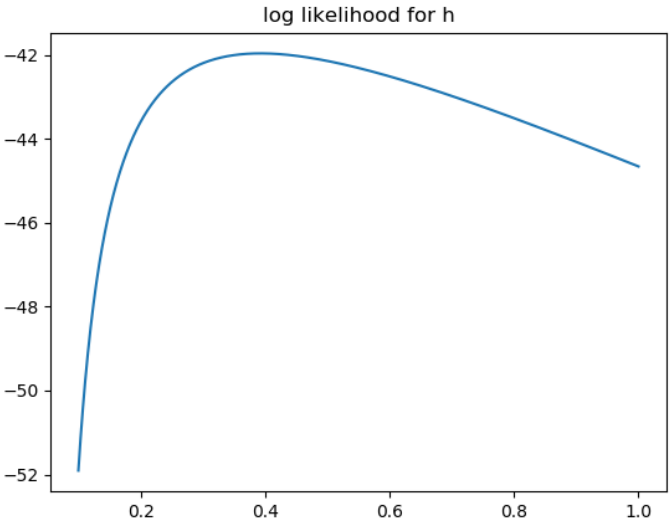
③ Rice Rule：bins=⌈2$n$^1/3⌉

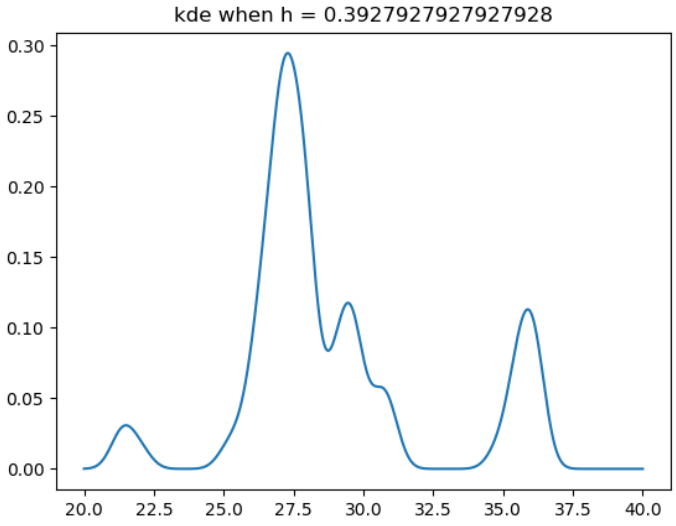
当 num_data = 200 时，使用以下三种方式进行实验，发现 n=15 时效果最好



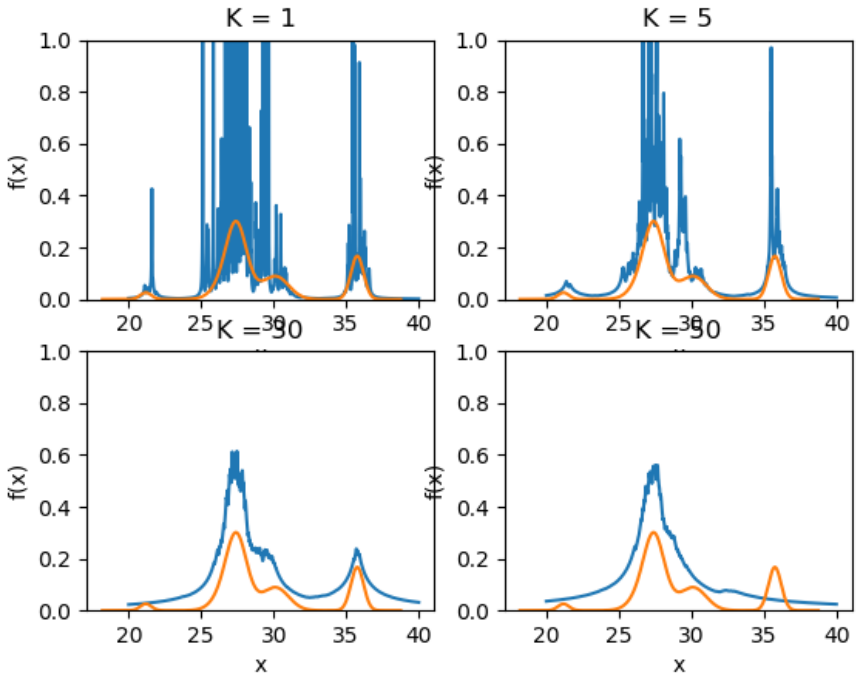

### 3. Find best h

使用最大似然估计 P(x | h)，寻找最适合的参数 h，使用 log likelihood 简化连乘运算。同时为了计算似然，在训练数据上使用 5 折交叉验证，绘制对数似然相对于参数 h 的曲线，发现当 h 约等于 0.39 的时候似然值最大。此时，高斯核密度法估计效果良好




## 4. K nearest neighbors




类似书中例图我们取了四个不同的 K 值进行实验，可以发现和 bin 的大小效应相同，当 K 很小的时候波动和方差特别大，密度估计结果很容易受到很多局部 cluster 的影响而变得很大，当 K 很大的时候，模型出现 over-fit 情况，即模型过于复杂，数据量太小无法做到很准确的拟合。

另外我们可以证明，当 x 的取值范围是有限集合的时候，即 x 存在最大值和最小值时，KNN 无法得到一个概率分布。其中 l 和 s 分别为 x 的最大值和最小值。

$$\int_{-\infty}^{+\infty} p(x)dx \geq \int_{-\infty}^{+\infty} \frac{K}{N*(l-s)}dx \geq \frac{K}{N*(l-s)}\int_{-\infty}^{+\infty} dx \to \infty$$