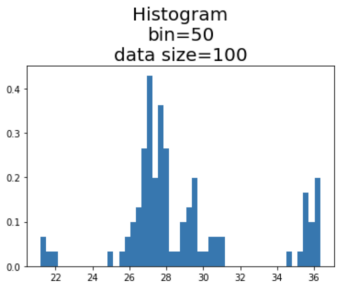
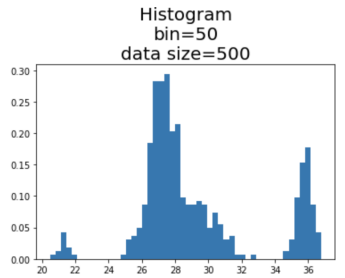
Assignment-1 Report

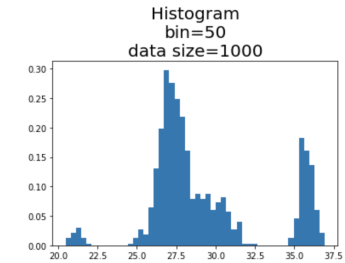
Influence by number of data

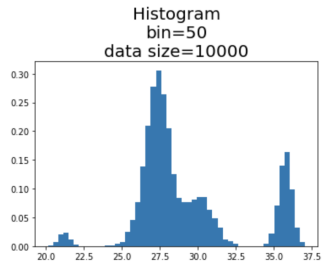
Empirical accertion: it is obvious that more data can lead to more accurate estimation. Experiments testify this idea.

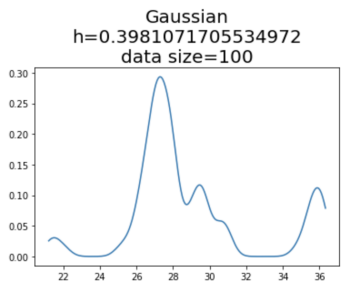
Experiments



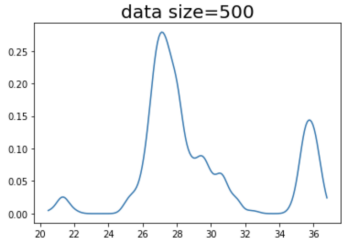




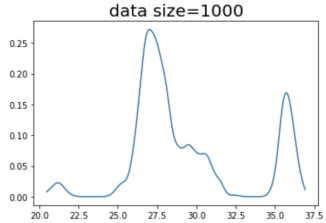




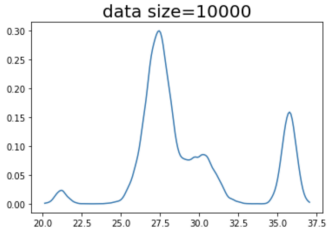
Gaussian
h=0.2885399811814427
data size=500



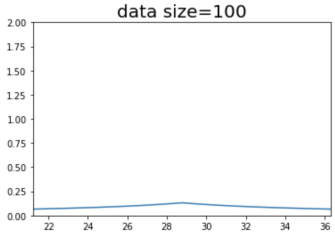
Gaussian
h=0.251188643150958
data size=1000



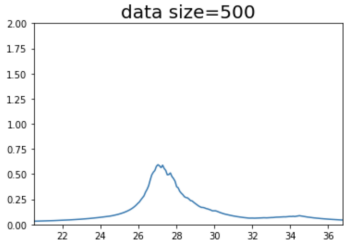
Gaussian
h=0.15848931924611134
data size=10000



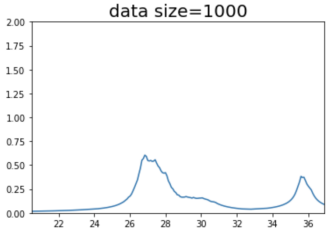
Nearest Neighbor
K=100
data size=100



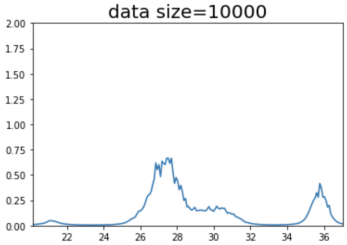
Nearest Neighbor
K=100
data size=500



Nearest Neighbor
K=100
data size=1000



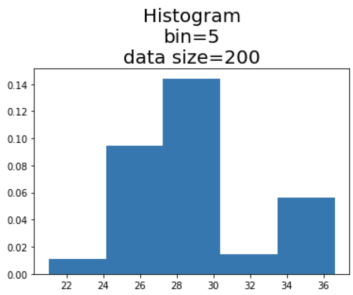
Nearest Neighbor
K=100
data size=10000

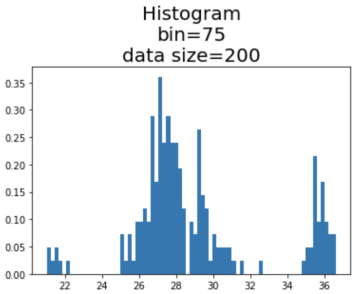


- More data can make estimation more accurate. In digrams which use nearest neighbour method, as amout of data increase, estimation result changes -- first unimodal, then bimodal, finally trimodal -- and more and more approach truth.
- More data can define boundaries of density more precisely. It is especially clear histograms. When data gets more and more, the histograms become smoother.

Tips: Sample data must represent the whole picture of rather than be part of original data.

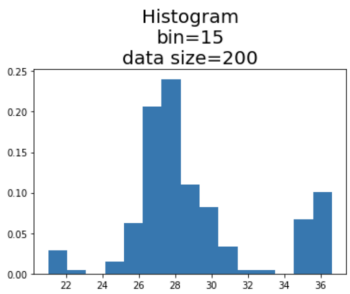
Histogram Estimation





Above pictures tell us Too few bins shown as left picture, the histogram doesn't really portray the data very well(The data is not a unimodal). Too many bins shown as right, we will get a broken comb look, which also doesn't give a sense of the distribution.

Method to find a suitable bins We can first find a small bins *sb* and a big one *bb*. Then we choose a series of data all evenly spaced between the *sb* and *bb* to draw the picture and pick the ideal one. As for me, I set *sb*=5, *bb*=75 and *interval*=5 end up choosing *bins*=15.(Shown as follow)



More exploration

from wikipedia, I found some methods to select a resonable *bins*

Square-root choice

$$bins = \lceil \sqrt{N} \rceil$$

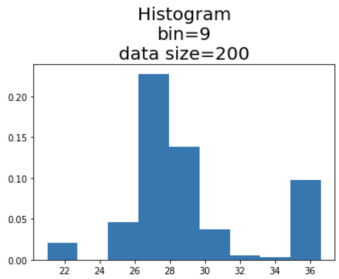
N is number of sampled data. This case is exactly when bins=15.

Sturges' formula

$$bins = \lceil \log_2 N \rceil + 1$$

N is number of sampled data. This case is exactly when bins=15.

Sturges' formula is derived from a binomial distribution and implicitly assumes an approximately normal distribution. It implicitly bases the bin sizes on the range of the data and can perform poorly if n < 30, because the number of bins will be small—less than seven—and unlikely to show trends in the data well. It may also perform poorly if the data are not normally distributed.(from wikipedia) Using Sturges' formula, bins=9. Shown as following.

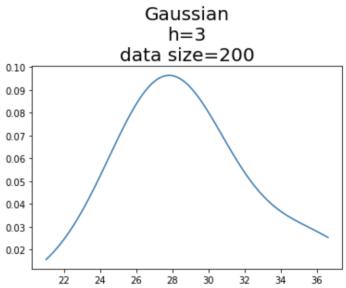


Conclusion bins=9 is nearly as good as bins=15. Square-root choice may be more suitable here.

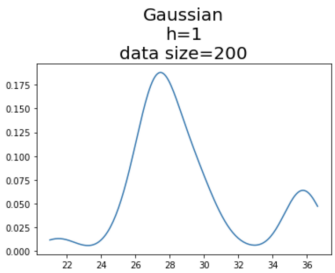
Kernel Density Estimation

Tune h

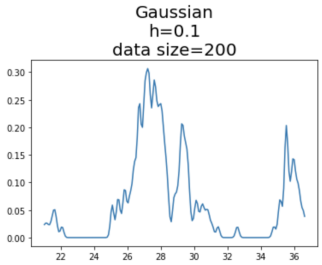
When h>2, gaussian kde gives unimodal which is not the truth.(Shown as following picture)



When 0.1< h <2, gaussian kde gives trimodal which is the truth.



When h<=0.1, gaussian ked gives the spiky curve. As the h decreases, curve can no more show trimodal.



Method to tune h

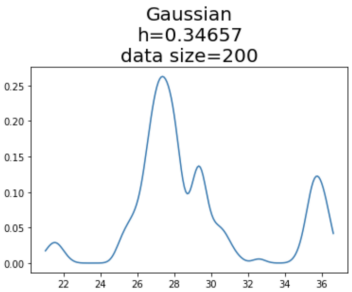
It is is the same as 'the method to find a suitable bins'. We start with finding a big h *bh* and small *sh* by baseline experiments. Then we choose a series of h evenly spaced between *bh* and *sh* and select a good one. Next, we reset *bh* and *sh* and do former precedure iteratively until find a ideal one.

Scott's Rule

In scipy documentay, I found a heuristic method to find a suitable *h*.

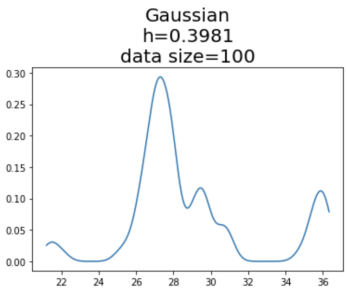
$$h = N^{-\frac{1.0}{d+4}}$$

N is the number of sampled data, d is the dimension of single data point.
In our case, d=1,N=200, we get h=0.34657



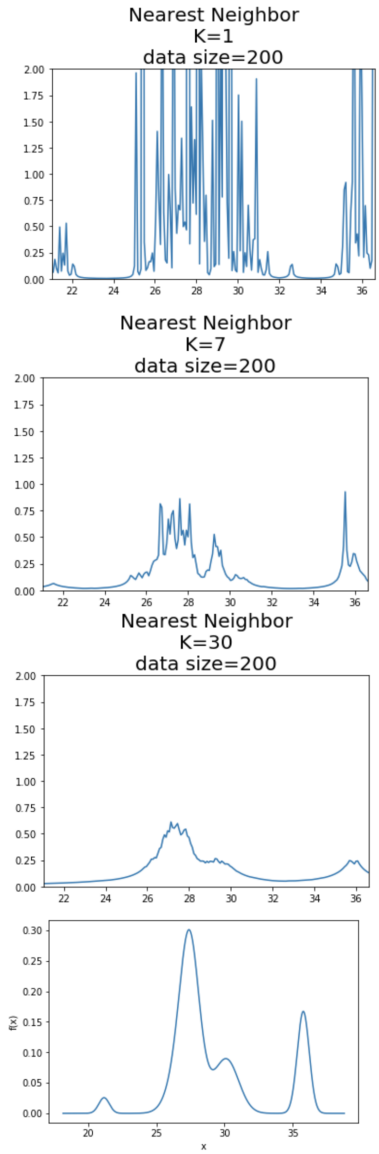
This result can be a good start.

Tune h with number of sampled data is 100



Nearest Neighbour Method

Vary K



K governs the degree of smoothing, so that a small value of K leads to a very noisy density model (up left), whereas a large value (bottom left) smoothes out the trimodal nature of the true distribution (shown by bottom right) from which the data set was generated. K=7 is a decent choice(shown as up right).

Yield a invalid distribution

Empirically: if K=1, probabiliy mass at sampled data is infinite. And

$$\hat{P}(x_i) = \frac{K * N}{V_i} \rightarrow P(x_i)$$

requires V_i small enough. Otherwise, small sphere centered x_i will overlap, leading to divergence.

References

Tune hitograms' bins
Understanding nearest neighbour

materials for understanding nearest neighbour more

Tune h of gaussian kde