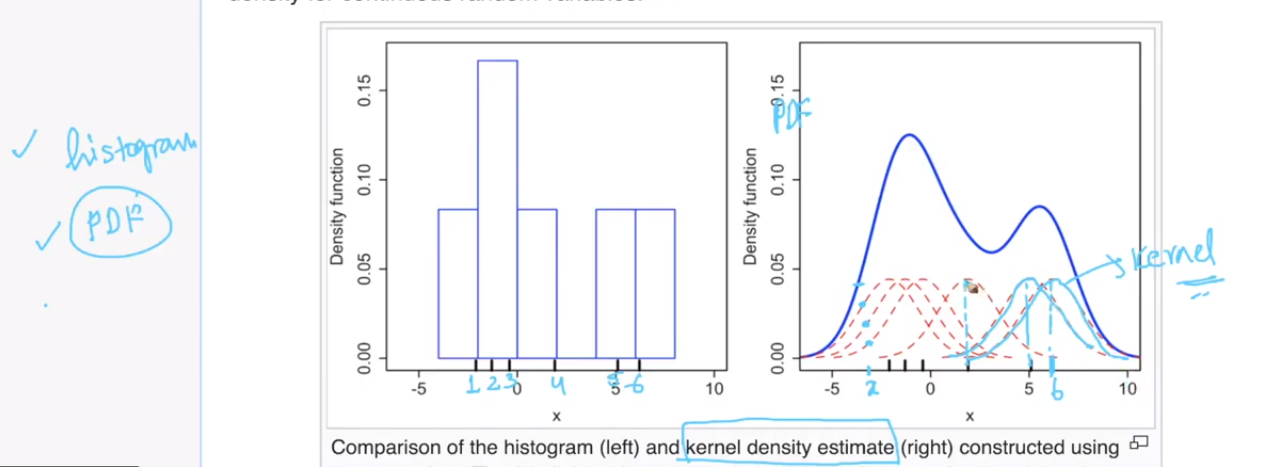
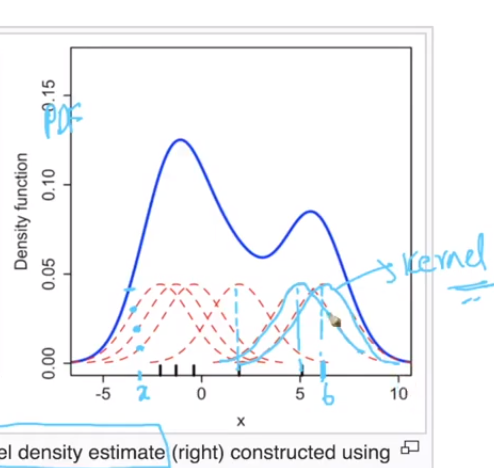
**Kernel density estimation(KDE):**

Kernel density estimation or KDE is a non-parametric way to estimate the probability density function of a random variable. In other words the aim of KDE is to find probability density function (PDF) for a given dataset. How it differs from normalized histogram approach? Well, it smooths the around values of PDF

As we already know that pdf are generated using histograms but how they are generated, they are generated using Kernel density estimation, at each observation, considering that observation as mean we draw a symmetric curve, known as **guassian kernel,** as seen in below fig we have created guassian kerne for 5, 6 and some values also(shown in light blue color).





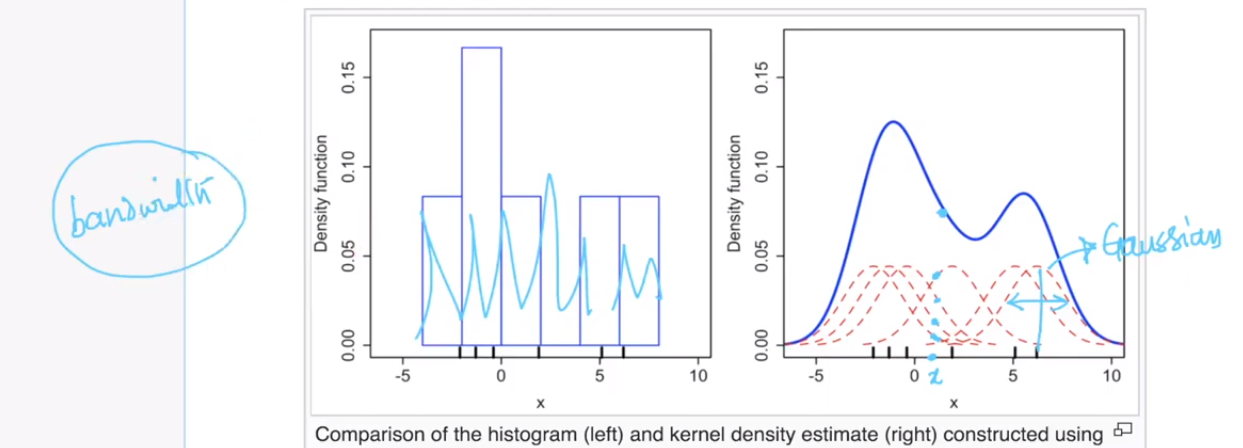
**How we choose no. of observations:**

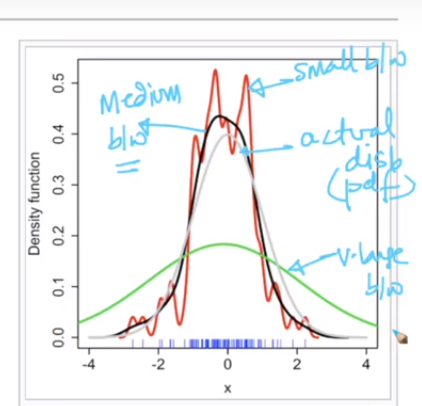
if I have iris dataset and 50 observations for PL, then I need to find KDE for each and every datapoint and we accumulate the sum of the gaussian kernels which are overlapping to find the height of PDF.

Now in order to find the pdf or height of pdf at that point, we simply adds the height at each guassian kernel, as given in below figure, for finding height at x, we adds the height of all 4 gaussian kernel that are available at x (shown in blue dots).

Now the question is that we know the mean here but how we know the spread, which in gaussian kernel is known as Bandwidth.

Bandwith is same as std-dev (=sqrt(variance)) for a Gaussian kernel. By adjusting the bandwidth, we are changing the shape of the kernel thereby changing the shape of final PDF. If the variance of the kernel is too small, we get a jagged and non-smooth PDF and if the variance is reasonable, we get a smooth and reasonable PDF.





**Some Comments:**

* In Layman's term, what does pdf and cdf tell about feature/variable and also what does is depict upon visualizing these in EDA for any dataset.

If you want to present your finding about say, "How long users stays on a a certain website", it may be good to show it in CDF as it shows the accumulated time he spent on that website, through the pages etc. On the other hand, if you want to simply show the probability of users clicking on an advert link (e.g. Google adwords link) then you may want to present it in PDF form as it will probably be a normal distribution bell curve and you can show the probability of that happening

* What do you mean by smoothed form of histogram or smoothing one histogram in layman terms?

A histogram is an accurate representation of the distribution of numerical data in the form of counts and shows the underlying frequency distribution. Kernel density estimation is a really useful statistical which let’s you create a smooth curve given a set of data. For the kernel density estimate, we place a normal kernel with a certain variance on each of the observations. The kernels are summed to make the kernel density estimate(in most cases the kernel is of type gaussian with a parameter called bandwidth)

* Some references:

<https://mathisonian.github.io/kde/>

<https://www.homeworkhelponline.net/blog/math/tutorial-kde>