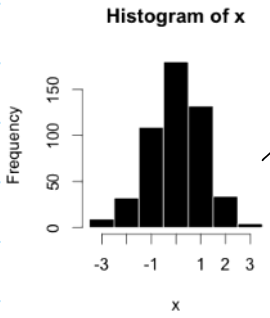
Kernel Density Estimation (KDE)

Tuesday, October 5, 2021 12:59

For the glory of God

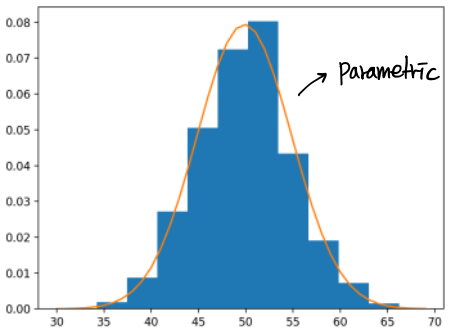
What is Kernel Density Estimation (KDE)?

- In statistics, we are typically able to generate a histogram using a given dataset.
  - ↳ A histogram is a plot that involves first grouping the observations into bins and counting the number of events that fall into each bin.
    - ↳ Divide the entire range of values into a series of intervals and count how many values fall into each interval



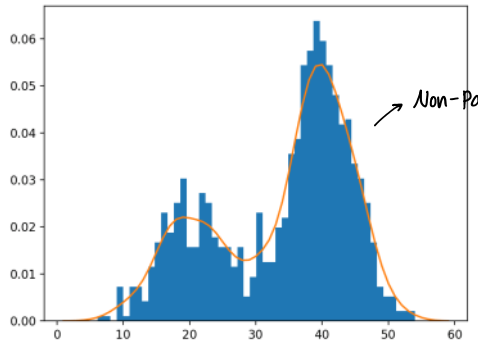
↝ Reviewing a histogram of data will help to identify whether the density looks like a common probability distribution (e.g., Gaussian distribution) or not.

- This histogram may be good enough to analyze some datasets; however, sometimes we are interested in calculating a smoother estimate.
  - ↳ Here, if the shape of a histogram matches a well-known probability distribution, we can use Parametric Density Estimation to calculate a smoother estimate.
    - If not, we can consider Non-parametric Density Estimation to calculate a smoother estimate.
- In terms of Parametric Density Estimation,
  - ↳ This can be achieved by estimating the parameters of the distribution from the dataset.
    - e.g. Using mean and standard deviation, we can estimate a normal (or Gaussian) distribution.



↝ Parametric Density Estimation

- In some cases, a data sample may not resemble a common probability distribution or can't be easily made to fit the distribution.
  - ↳ In this case, Parametric Density Estimation is not feasible.
  - ↳ Instead, an algorithm is used to approximate the probability distribution of the data without a pre-defined distribution.
    - ↳ This is referred to as Non-parametric Density Estimation.



↗ Non-Parametric Density Estimation

- The most common non-parametric approach for estimating the PDF of a continuous random variable is called Kernel Density Estimation (KDE).
- In theory, KDE is defined as a non-parametric way to estimate the probability density function of a random variable.

(In practice, it is a technique that enables us to create a smooth curve for a given dataset that is not feasible with well known parameters)

How do we get a KDE line?

- The high-level idea is to calculate a probability for each given value of a random variable by weighting the contribution of observations from data samples.

(or density)

- Mathematically, KDE means that we want to estimate the shape of $f(x)$.

$$f(x) = \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right)$$; where
- $f(x)$ = kernel density estimator
- K = kernel function ⇝ The following functions are commonly used: Uniform, Triangle, Gaussian, Epanechnikov, ⋯
- x = given point
- x_i = observations
- b = bandwidth

- Intuitively, KDE will be obtained by the following procedure:

Step 1. Let's say that we have a few sample points from an unknown distribution as follows:

Sample	1	2	3	4	5	6
Value	-2.1	-1.3	-0.4	1.9	5.1	6.2

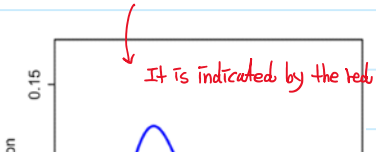
Step 2. Define a kernel function with a bandwidth.

- Note that
  - A bandwidth is a parameter that controls the number of samples or window of samples used to estimate the probability for a given point.
  - As the bandwidth is a free-parameter, it will make influence on the final estimate, leading to validation works such as comparing with histogram.

e.g.  Gaussian with b_1 vs.  Gaussian with b_2 ; Here, $b_2 > b_1$

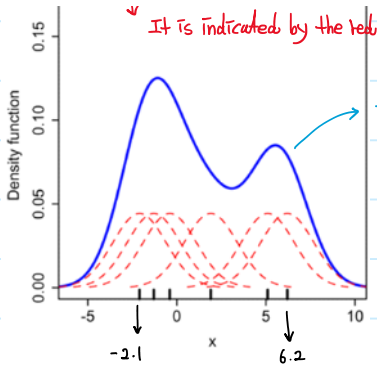
Step 3. Let's say that we decide to use Gaussian distribution with the specified bandwidth parameter.

↳ Then the kernel is placed on each of the samples.



It is indicated by the red dashed lines below:

It is indicated by the red dashed lines below :

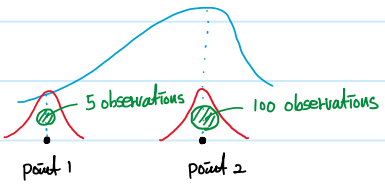


This is what KDE produces from step 4 and 5.

Step 4. Count how many observations from the kernel range.

↪ The logic is : If we see more points (or observations) nearby given point (x), then put more weight ; thus, the estimate is higher.

e.g.



Here, note that

: Kernel is used to control the contribution of samples in the dataset toward estimating the probability of the given point.

Step 5. The kernels are summed to make the kernel density estimate.