Random Variables

* What are Algebic Variables?

In Algebra a variable, like x, is an unknown value

$$x + 5 = 10$$

$$\boxed{x = 5}$$

* What are Random Variables in stats and Probability?

A Random variable is a set of possible values from a random experiment.

coin → H
coin → T

$x = \{1, 0\}$ → sample space

H = 1 , T = 0
denote

* Random Variable denote with capital ~~small~~ letter and Algebic variable denote small letter

Types of Random Variable

Discrete Random Variable

$\{H, T\}$

$\{1, 2, 3, 4, 5, 6\}$

Continous Random Variable

$x = (0, 10)$

Range of value or any value like 8.5, 2.4...

Probability Distribution

* What are Probability Distributions?

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability value.

coin Toss	1 (H)	0 (T)
Probab	1/2	1/2

Dice	1	2	3	4	5	6
Prob	1/6	1/6	1/6	1/6	1/6	1/6

2 Dice Rolled

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Prob distribution

2 → 1/36

3 → 2/36

4 → 3/36

5 → 4/36

6 → 5/36

7 → 6/36

8 → 5/36

9 → 4/36

10 → 3/36

11 → 2/36

12 → 1/36

Problem with Distribution

In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

Example:- Height of people, Rolling 10 dice together

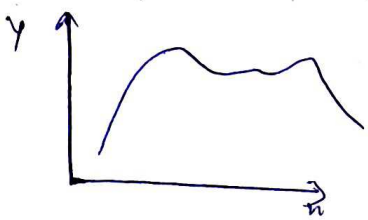
Solution — function?

What if we use a mathematical function to model the relationship between outcome and probability?

x → outcome	1	2	3	4	5	6
y → Prob	1/6	1/6	1/6	1/6	1/6	1/6

$$y = f(x)$$

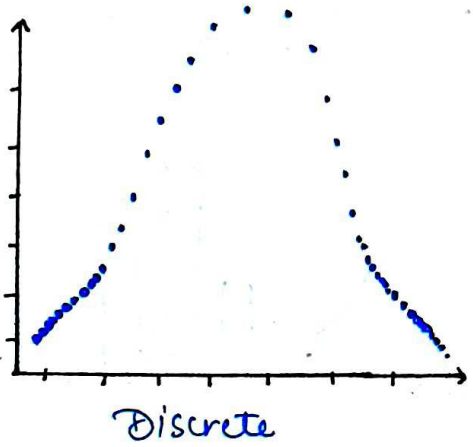
Prob dist func



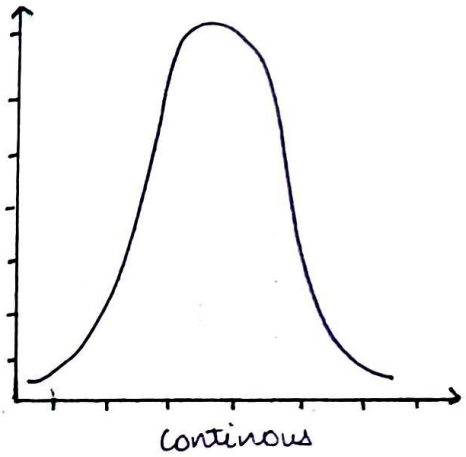
Note - A lot of time Probability Distribution (Table) and Probability distribution functions are used interchangeably. (↳ function)

Types of Probability Distribution (PDF)

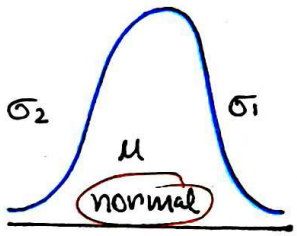
Descrete Random Variable

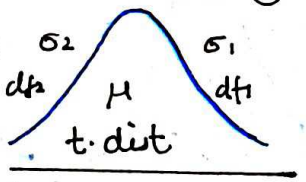


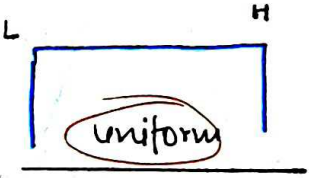
Continous Random Variable

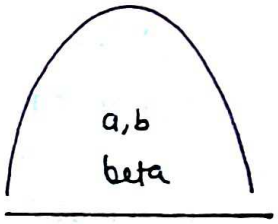


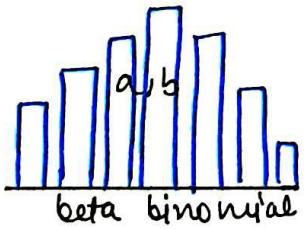
Probability Distribution

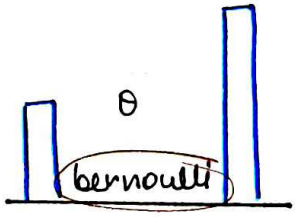


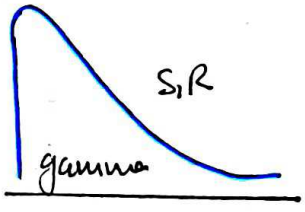


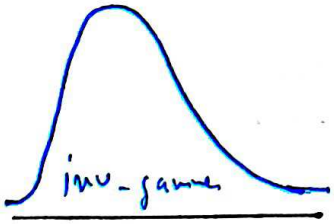


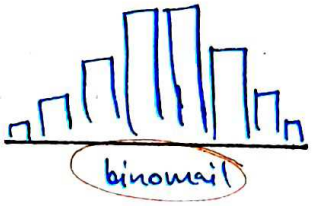


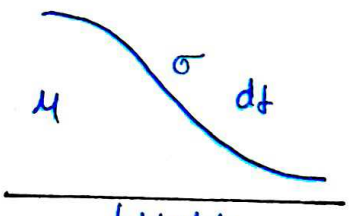


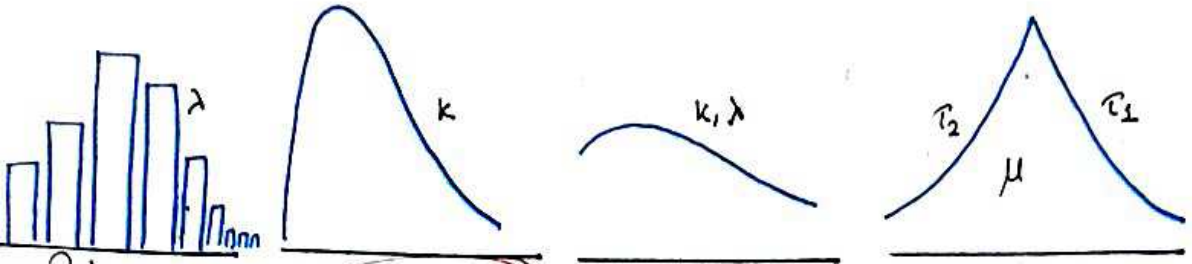




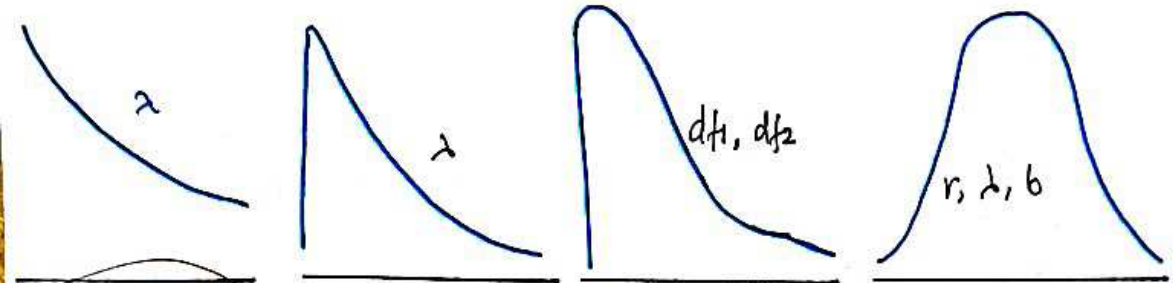




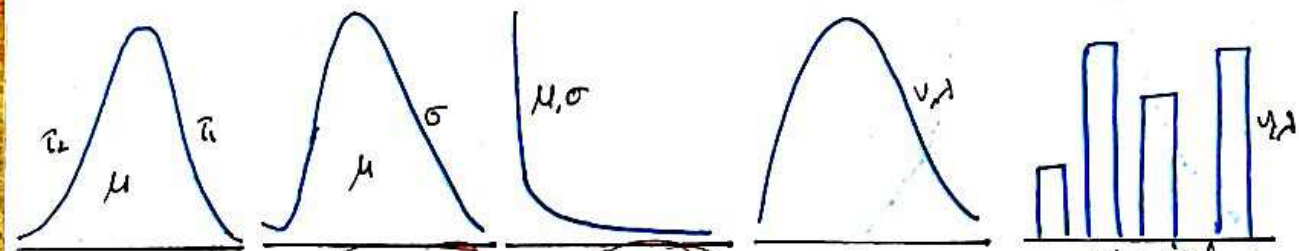




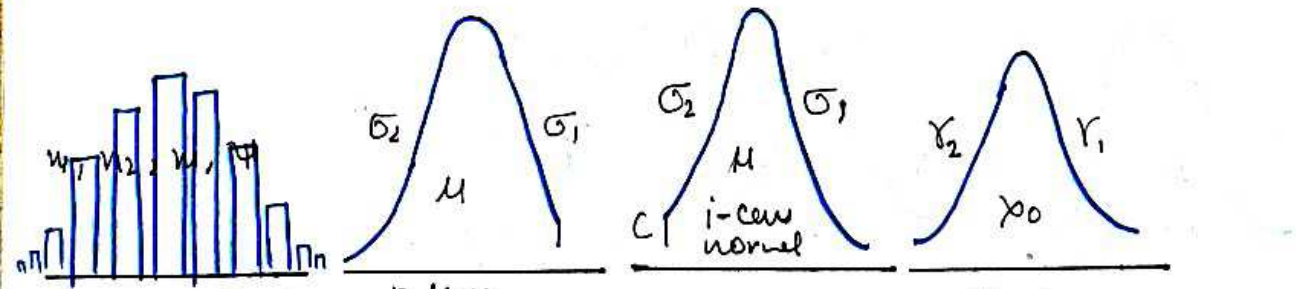
Poisson | Chi-square | non central chi-square | double exp.



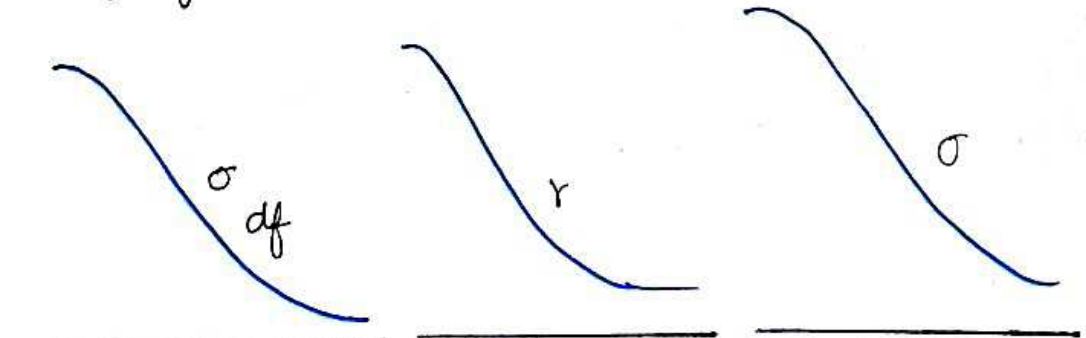
exponential | Shifted exp | F dist | gen gamma



logistic | lg-normal | Pareto | Weibull | categorical



non central Hyper geom | r cens normal | i-cens normal | Cauchy



half-t | half Cauchy | half-normal

* Why are Probability Distribution important?

→ Gives an idea about the shape / distribution of the data.

→ And if our data follows a famous distribution then we automatically know a lot about the data.

A note on Parameter (PDFs)

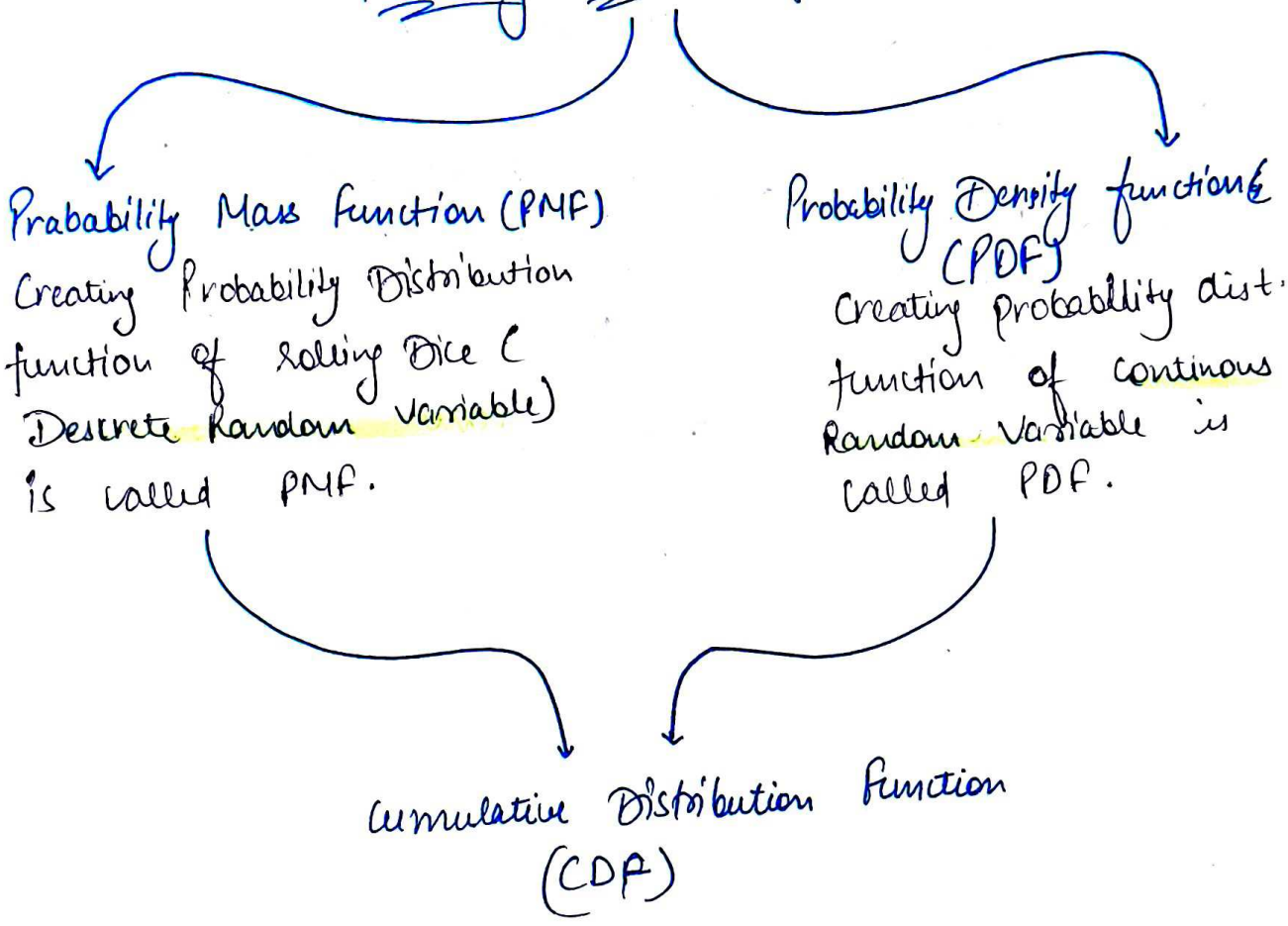
Parameter in probability distribution are numerical values that determine the shape, location, and scale of the distribution.

Different probability distribution have different sets of parameter that determine their shape and character -stics and understanding these parameter essential in statistical analysis and inference.

Probability Distribution Functions

A probability distribution function is a mathematical function that describes the probability of obtaining different values of a random variable in a particular probability distribution.

$y = f(x)$



PDF → CDF

PMF → CDF

Probability Mass function (PMF)

It is a mathematical function that describes the probability distribution of a discrete random variable.

The PMF of a discrete random variable assigns a probability to each possible value of the random variable. The probabilities assigned by the PMF must satisfy two condition:

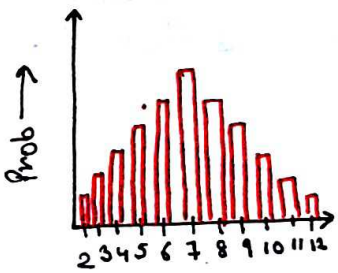
a. The probability assigned to each value must be non-negative (i.e., greater than or equal to zero)

One dice rolled

$$y = \begin{cases} 1/6 & \text{if } x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases}$$

Two dice rolled

$$y = \begin{cases} 1/36 & x \in \{2,12\} \\ 2/36 & x \in \{3,11\} \\ \vdots & \\ 0 & \text{otherwise} \end{cases}$$



* We can find the PMF and draw the the graphs.

Cumulative Distribution function (CDF) of PMF

The cumulative distribution function (CDF) $F(x)$ describe the prob that a random variable X with a given prob dist. will be found at a value less than or equal to x

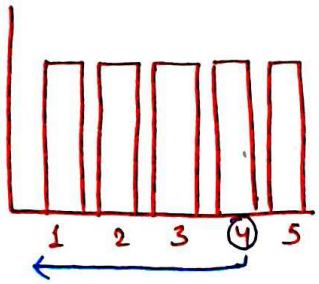
$$F(x) = P(X <= x)$$

Let a dice is rolled

$\{1, 2, 3, 4, 5\}$

$$f(x \leq 4) = f(x=4) + f(x=3) + f(x=2) + f(x+1)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6}$$

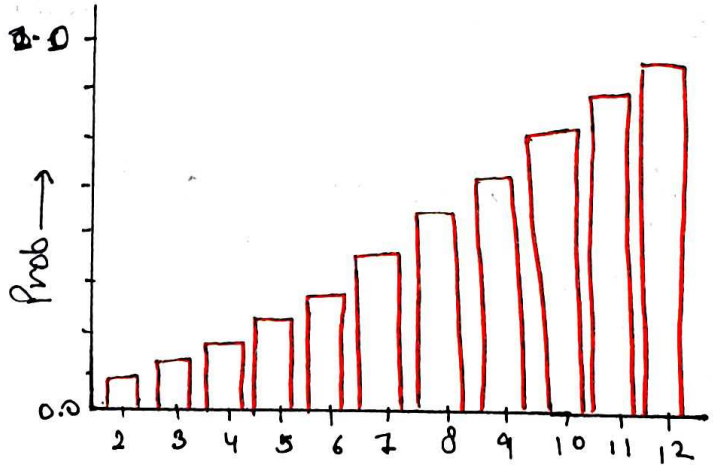


In PMF if $x = 4$ then Prob of 4.

but In CDF if $x=4$ then find prob is equal to and less than 4.

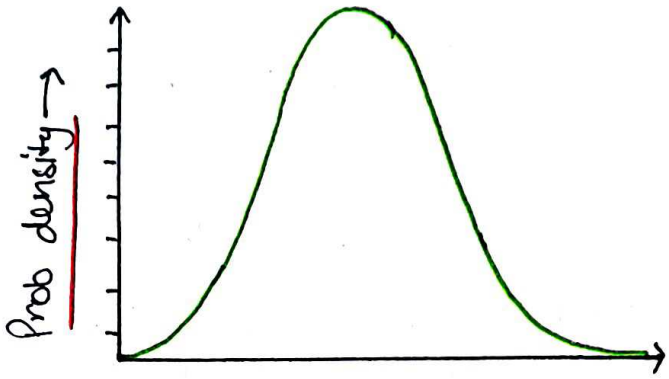
Two dice are rolled

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



Probability Density Function

It is a mathematical function describe the prob dist. of a continous random variable.



why Prob. density and why no probability?

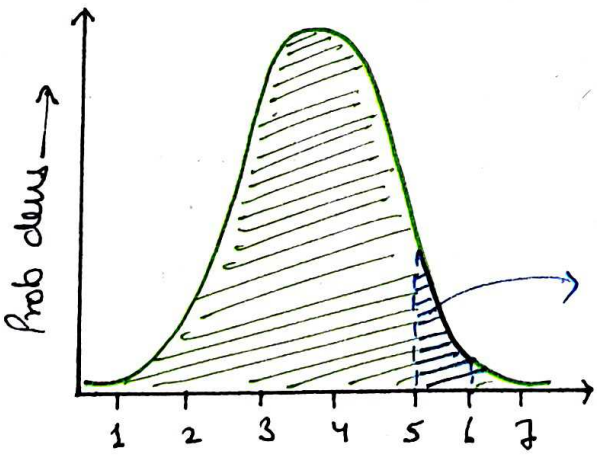
let assume gpa = 7.912...

We cannot able to find prob at a ~~xxxxxx~~ low level like 7.912... something. Prob of this may be near to zero or zero. We can dealing with infinite no. of values. That's why we can not use prob at y-axis.

What does the area of this Graph represents?

PDF helps to find the prob betn the range like (5 to 6).

$P(0 \leq x \leq 10) = 1 \Rightarrow$ Total area



Find the area and this area is the prob.

$$\int_5^6 f(x)\,dx$$

How is Graph calculated ?

Density Estimation

Density estimation is a statistical technique used to estimate the prob density func (PDF) of a random variable based on a set of observation or data. In simpler terms, it involves estimating the underlying distribution of a set of data points.

Density estimation can be used for a variety of purposes, such as hypothesis testing, data analysis, and data visualization. It is particular useful in areas such as machine learning, where it is often used to estimate the probability distribution of input data or to model the likelihood of certain events or outcomes.

There are various methods for density estimation, including parametric and non-parametric approaches. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumption about the distribution and instead estimate it directly from the data.

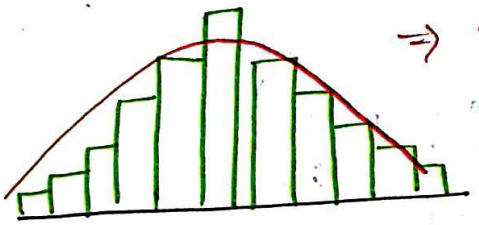
Commonly used techniques for density estimation include kernel density estimation (KDE), histogram estimation and Gaussian mixture models (GMMs). The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density function

Parametric density estimation is a method of estimating the prob density function (PDF) of random variable by assuming that the underlying distribution belong to specific parametric family of probability distribution such as the normal, exponential or poisson distribution.

→ I took 1000 data 4.3, 9.1, 6.5, ... 1000

→ and plot a histogram (mean, std. dev.)



⇒ It seems like Normal distribution

So, we use pdf of normal dist.

$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\sigma-\mu}{\sigma}\right)^2}$ we have to find mean and standard dev.

After plotting the graph and graph doesnot match with any other graph then we use Non parametric Density Estimation (KDE)

Non-Parametric Density Estimation

But sometimes the distribution is not clear or its not one of the famous distributions.

Non parametric density estimation is a statistical technique used to estimate the probability density function of a random variable without making any assumption about the underlying distribution. It is also referred to as non-parametric density estimation because it does not require the use of a predifined

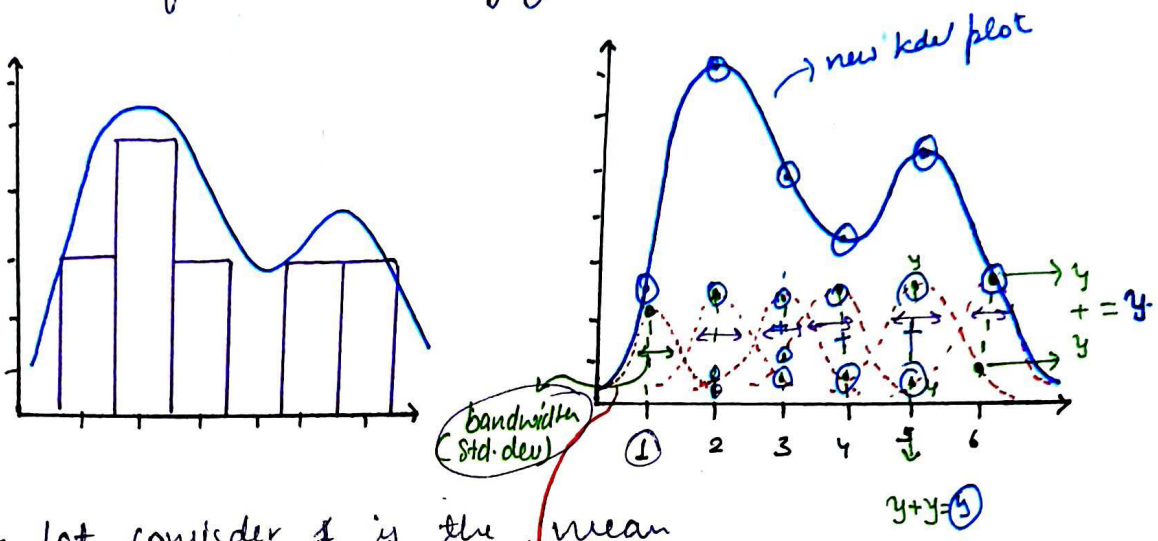
prob distribution function, as opposed to parametric method such as the Gaussian distribution.

The non-parametric density estimation technique involves constructing an estimate of the prob density func. using the available data. This is typically done by creating a kernel density estimate.

Non-parametric density estimation has several advantages over parametric density estimation. One of the main advantages is that it does not require the assumption of a specific distribution, which allows for more flexible and accurate estimation in situations where the underlying distribution is unknown or complex, However, non-parametric density estimation can be computationally intensive and may require more data to achieve accurate estimate compared to parametric methods.

Kernel Density Estimate (KDE)

The KDE technique involves using a kernel function to smooth out the data and create a continous estimate of the underlying density function.



Step1:- Let consider 1 is the mean and create normal dist on 1 and same as for 2, 3, 4, 5, 6

Step2:- Add the value of y which is intercept to normal dist of 1.

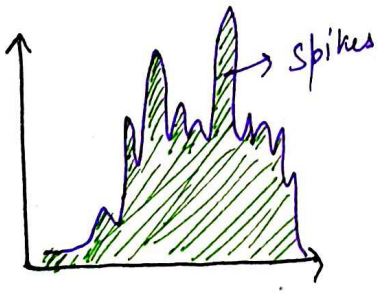
Step 3:- Make new kde plot with some of y and original x.

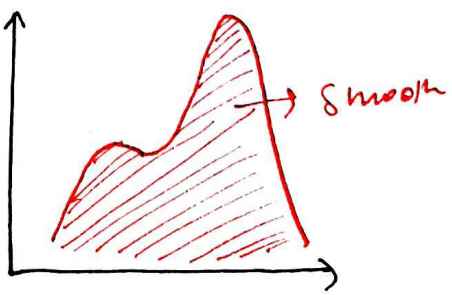
Step4:- Bandwidth is the hyperparameter.

Kernel = Gaussian mostly used.

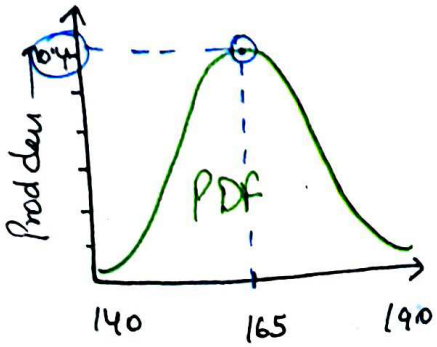
bandwidth ↑↑

bandwidth ≃ 0

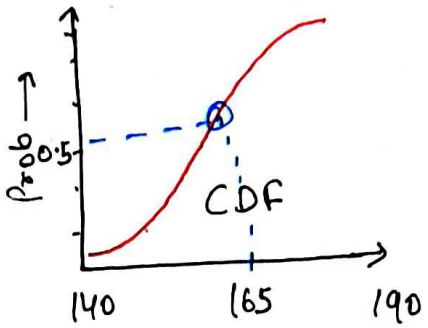




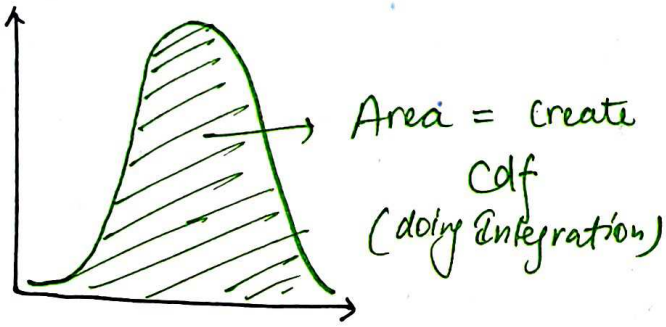
Cumulative Density function → PDF

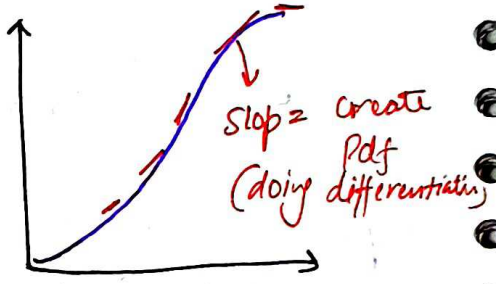


$P(x=165)$



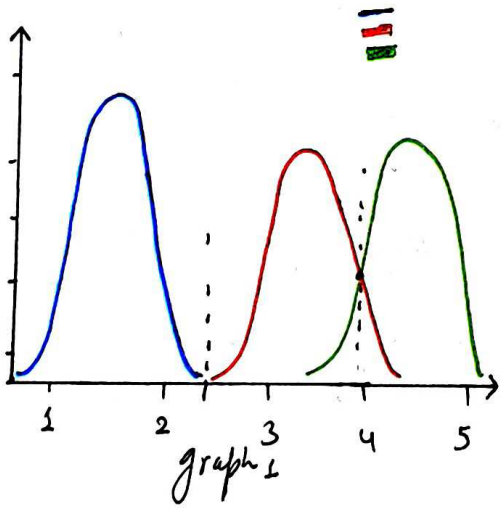
$P(x \leq 165)$

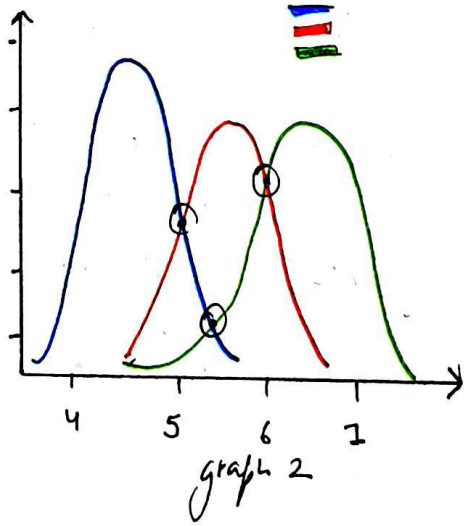




How to use PDF in Data Science ?

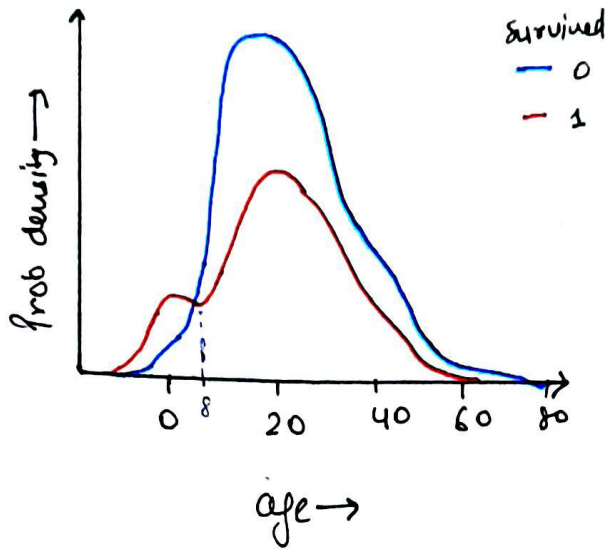
1)





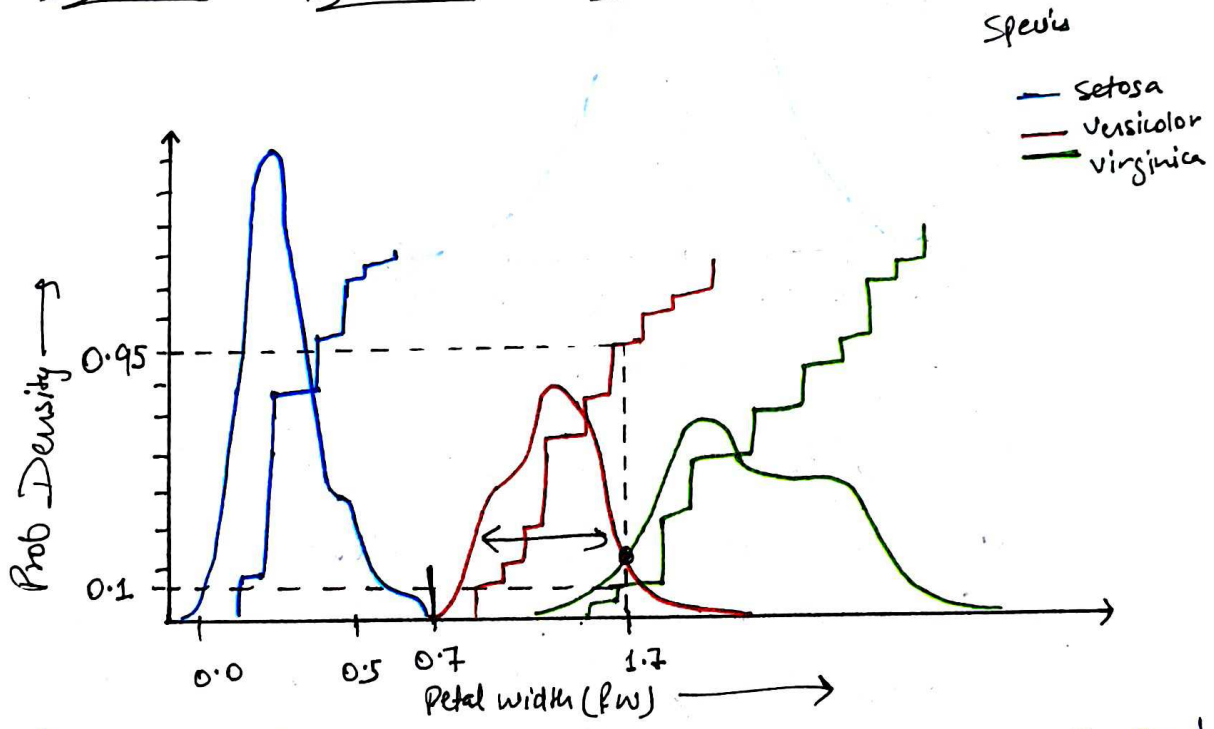
PDF helps in the features selection. In graph 1 we can easily differentiate categories bet^n the catgories. but In graph 1 we can see the kde plot near to each other and difficult to differenciate.

2



In this graph, 0-8 age have higher chance to survive but after 8-80 have higher chance for prob density to not survive.

How to use CDF in data science?



if $0.7 < PW < 1.7$ is versicolor? ⇒ 95% of versicolor under 1.7 and between 0.7 and 1.7.

5% can be mistake

0.1 → 10% of virginca data in versicolor data ⇒ So PW > 1.7 Virginica

↳ 90% of data virginica

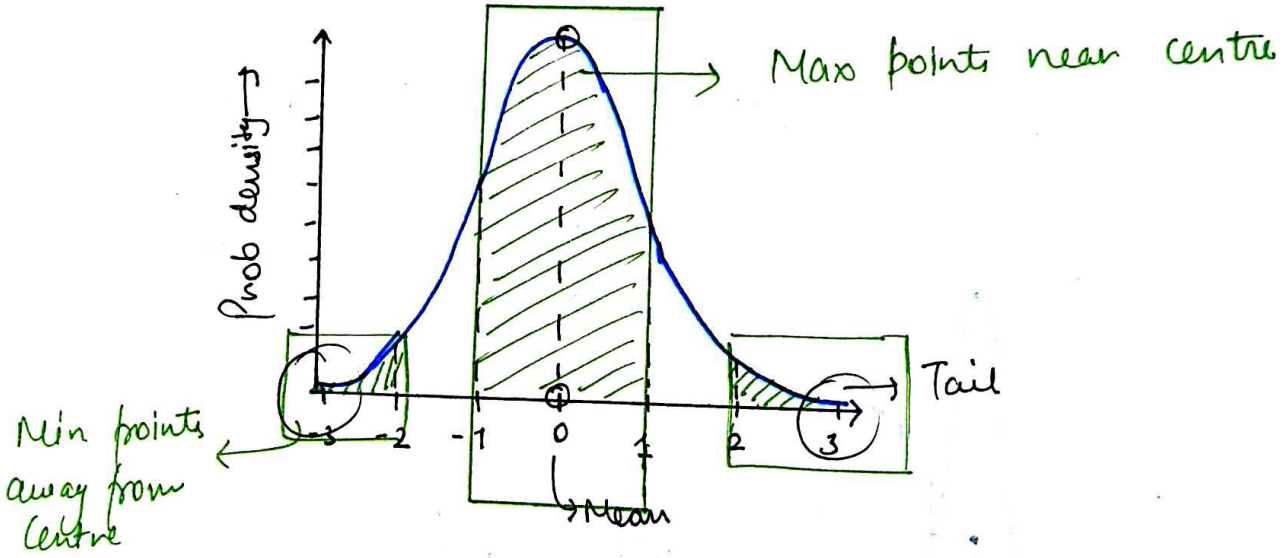
10% mistake can be

* JO decision making PDF ke through kiya use ek quantative label de skte hn (95% right 5% wrong)

Normal Distribution

1. What is Normal Distribution?

Normal Distribution, also known as Gaussian distribution is a prob distribution that is commonly used in statistical analysis. It is a continous prob distribution that is symmetrical around the mean, with a bell-shaped curve.



The normal distribution is characterized by two parameters: the mean (μ) and the standard deviation (σ). The mean represent the centre of the distribution, while the standard deviation represent the spread of the distribution.

Why is it so important?

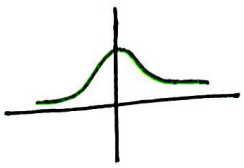
Commonality in Nature Many natural phenomena follow a normal distribution, such as the height of people, the weights of objects, the IQ scores of a population and many more. Thus, the normal distribution provides a

convenient way to model and analysis such data.

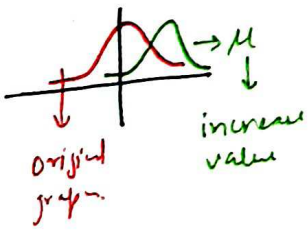
PDF eqn of Normal distbn

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

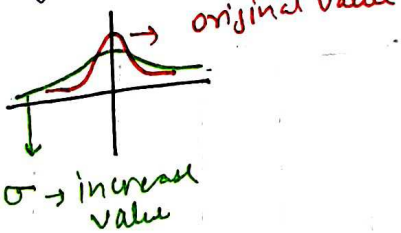
Eqn in detail:

$y = e^{-x^2} \Rightarrow$ 

$y = e^{-(x-\mu)^2} \Rightarrow$ Normal distribution change position left to right

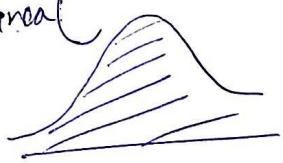


$y = e^{\frac{-(x-\mu)^2}{2\sigma^2}} =$ help to spread of the graph



$y = e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ → integeration = $\sigma\sqrt{2\pi}$ for cancelling this divide by this area

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Standard Normal Variate (z)

or

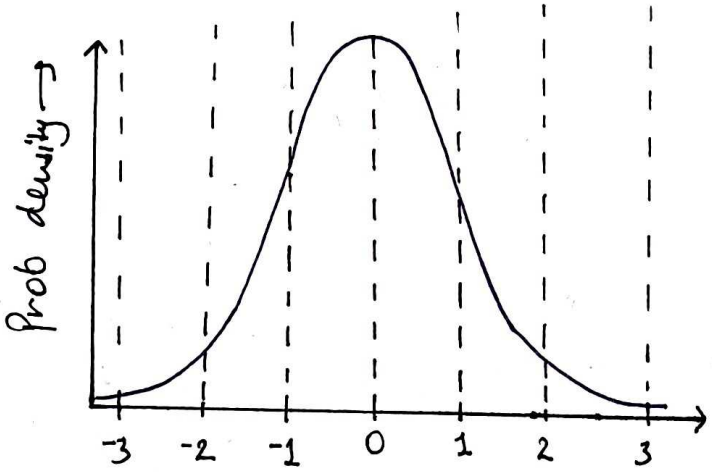
Standard Normal Distribution

* What is standard Normal variate

A Standard Normal Variate (z) is a standardized form of the normal distribution with mean = 0 and standard deviation = 1.

$X \sim N(\mu, \sigma)$ → Normal distbⁿ

$Z \sim N(0, 1)$ → Standard Normal Variate

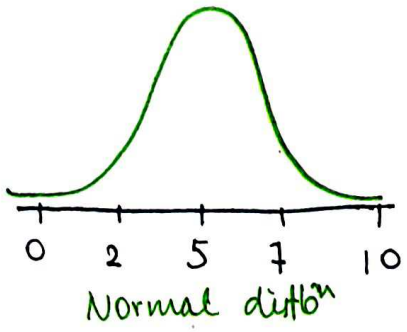


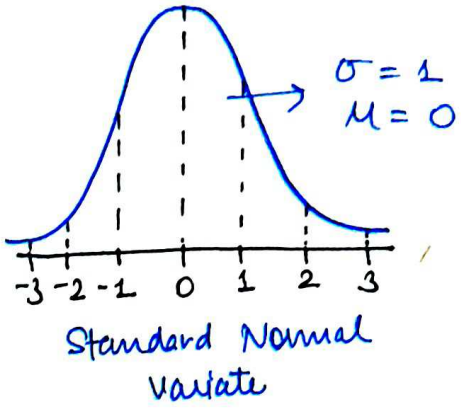
Standardizing a normal distribution allows us to compare different distribution with each other, and to calculate probabilities using standardized tables or software.

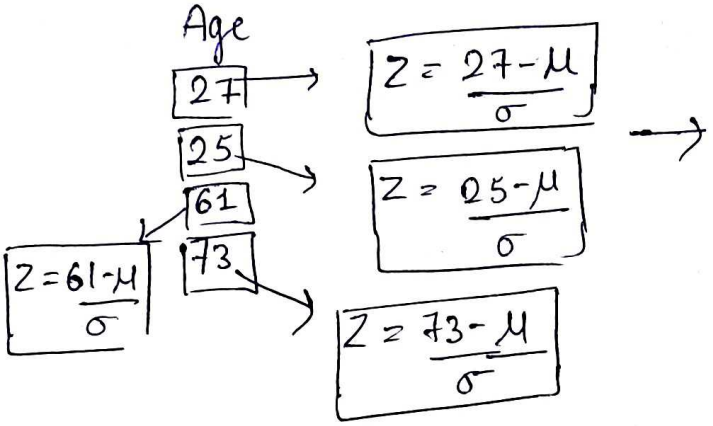
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \Rightarrow \boxed{\frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(x)^2}}$$

Standard Normal Variate

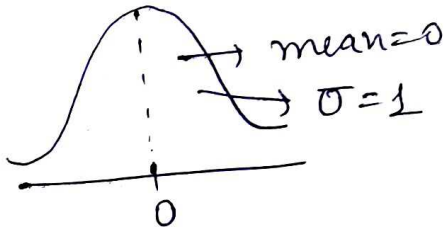
* How to transform a normal distribution to Standard Normal variate.







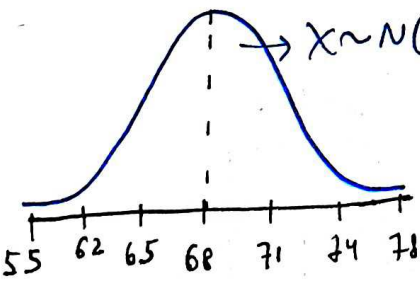
Now Plot the graph



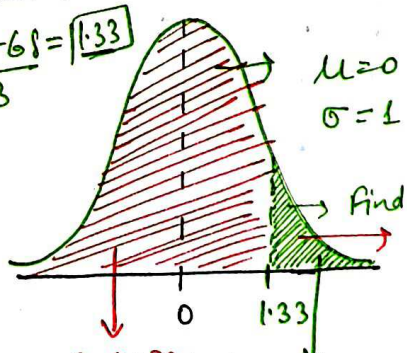
Suppose the heights of adult males in a certain population follow a normal distbn with a mean of 68 inches and a Standard deviation of 3 inches. What is the prob that a randomly selected adult male from this population is taller than 72 inches?

* 72 inches or usse kam hone ki prob = 90.8

Given → population ⇒ Normal distbn



$Z = \frac{72-68}{3} = 1.33$



using Z table → the table and the norm is 0.90824

1.33 → 1.3 find on the table and find 0.03 →

0.90824 (90.8%) Prob

72 inches ⇒ 1.33

Total area - Find area

100 - 90.8

= 9.2%

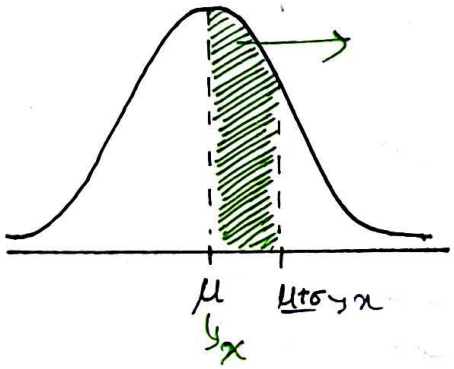
* What are z-tables?

A z-table tells you the area underneath a normal distbⁿ curve, to the left of the z-score.

Check table :- https://www.ztable.net/

* For a Normal Distribution $x \sim (\mu, std)$ what percent of population lie betⁿ mean and 1 std, 2 std, 3 std?

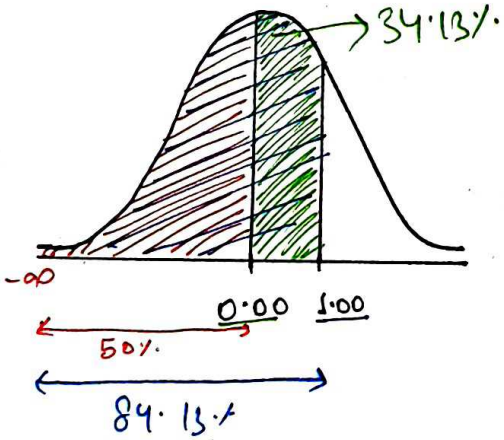
$X \sim N(\mu, \sigma)$



SNV (z) $z = \frac{x-\mu}{\sigma}$

$z = \frac{\mu-\mu}{\sigma} = 0.00$

SNV (z) $z = \frac{\mu+\sigma-\mu}{\sigma} = 1.00$



z table of 0.00 = 50%.

z table of 1.00 = 84.13%.

84.13 - 50 = 34.13

$z = \frac{\mu+2\sigma-\mu}{\sigma} = 2$

z table of 2 → 0.9772

97%.

97.72 - 84.13 = 13.59