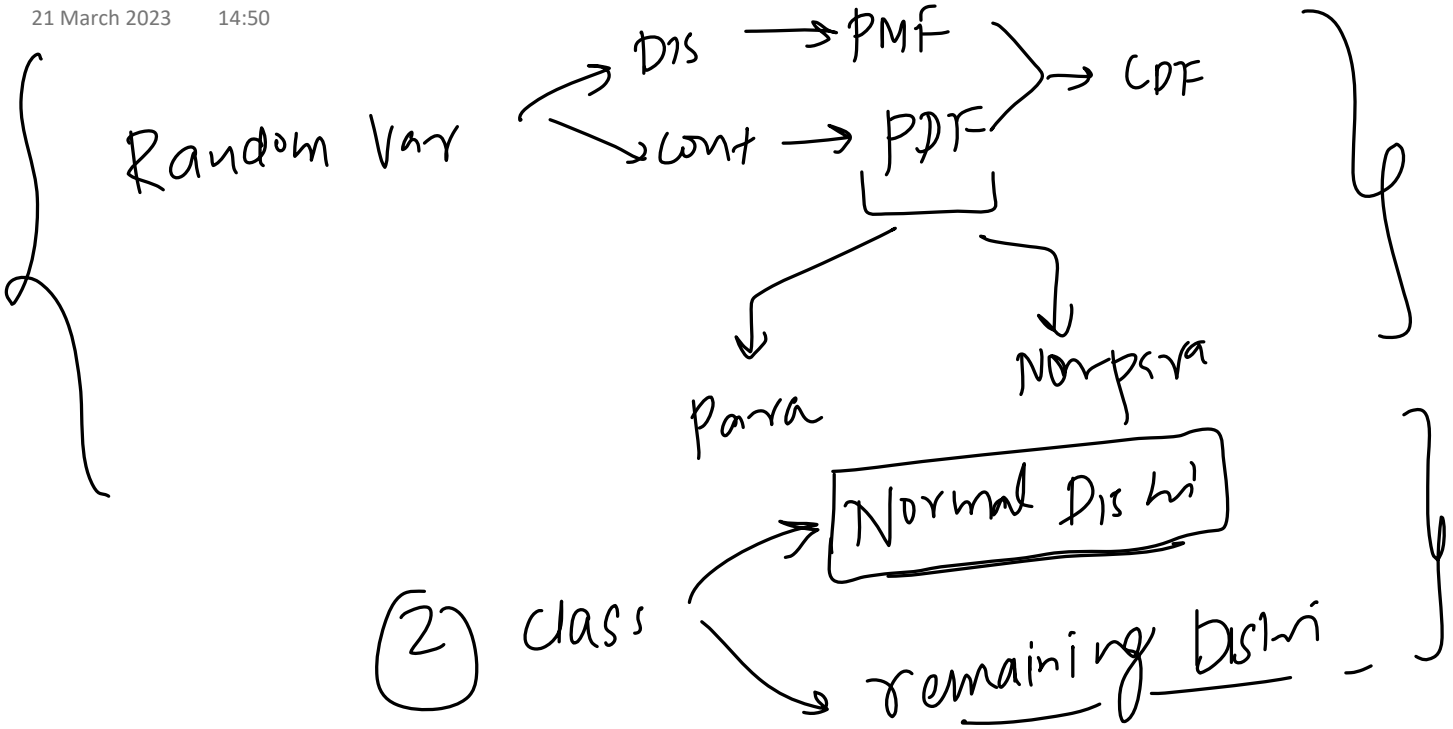
Recap

21 March 2023 14:50



আমরা feature selection করি যে, কোনটা আমাদের flower identification কাজে লাগবে। Petal. length কাজে লাগবে কিনা তা PDF দিয়ে দেখতে পারি।

How to use PDF in Data Science

20 March 2023 18:11

PDF → CDF

PDF $P(X = x)$

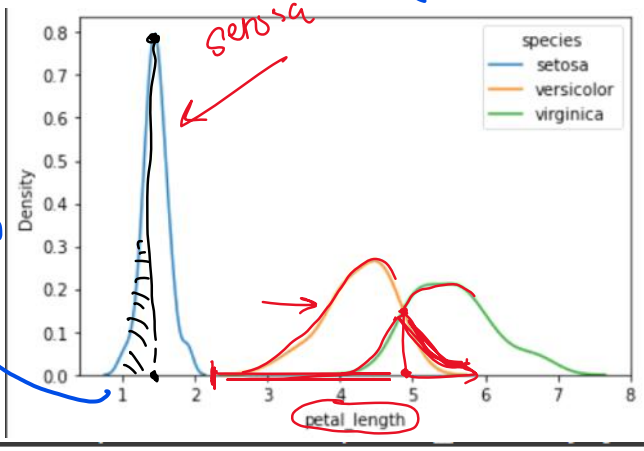
(Petal length) ← $2.3 < pl < 5$ → versicolor

$pl > 5$ virginica

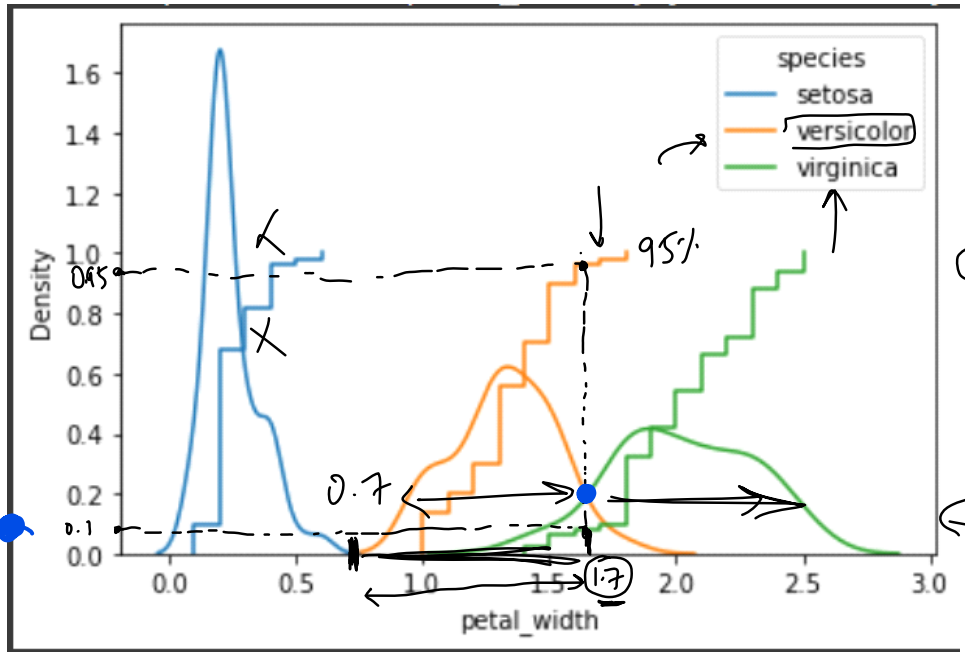
$P(X \leq x)$

$pl < 2.3$ → setosa

> 2.3



Petal length (1~2) হলে setosa



95% of times

If $0.7 < pw < 1.7$ → versicolor

$pw > 1.7$ → virgin → 90%

10% virgin

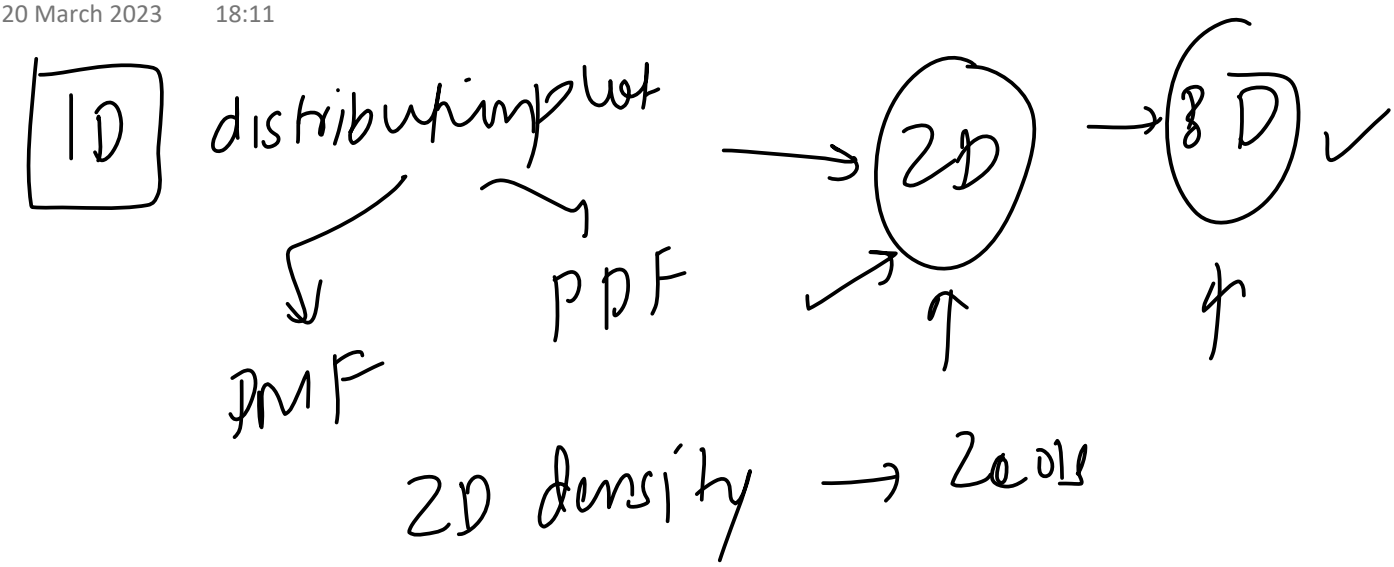
Petal-width দিয়ে কত % কোন ফুলকে আলাদা করে identify করতে পারবো তা CDF দিয়ে ৯টা ধারণা পেতে পারি।

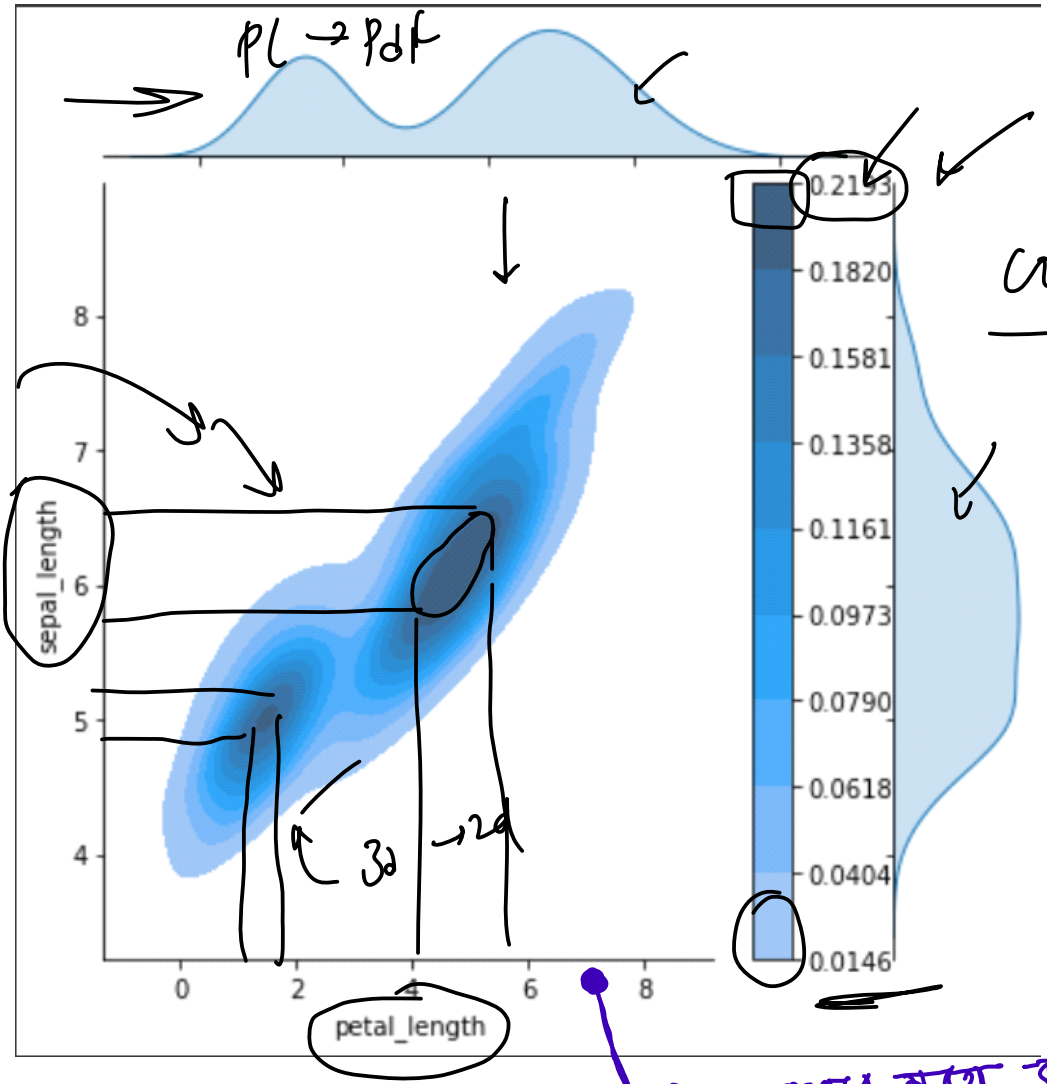
উপরে, Petal-width দিয়ে (0.7 ~ 1.7) পর্যন্ত versicolor কে identify করলে প্রায় 95% versicolor ফুলকে আলাদা করতে পারবো।

(1.7 পর্যন্ত কারণ এখানে Kdeplot dominate করতেছে virginica ফুলের চেয়ে।) ⊛

2D Density Plots

20 March 2023 18:11



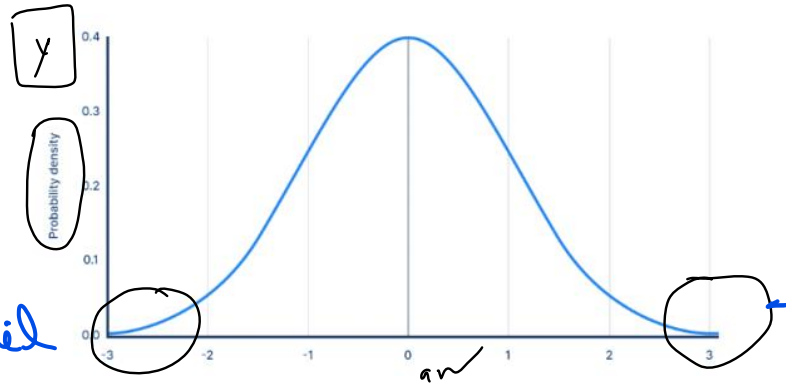


Normal Distribution

20 March 2023 18:06

1. What is normal distribution?

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



-> Tail
-> Asymptotic in nature
-> Lots of points near the mean and very few far away

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

Denoted as: $X \sim N(\mu, \sigma)$ where $\mu \rightarrow$ mean, $\sigma \rightarrow$ std

Why is it so important?

Commonality in Nature: Many natural phenomena follow a normal distribution, such as the heights 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 analyse such data.

PDF Equation of Normal Distribution

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

Parameters in Normal Distribution

https://samp-suman-normal-dist-visualize-app-lkntug.streamlit.app/

Equation in detail:

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

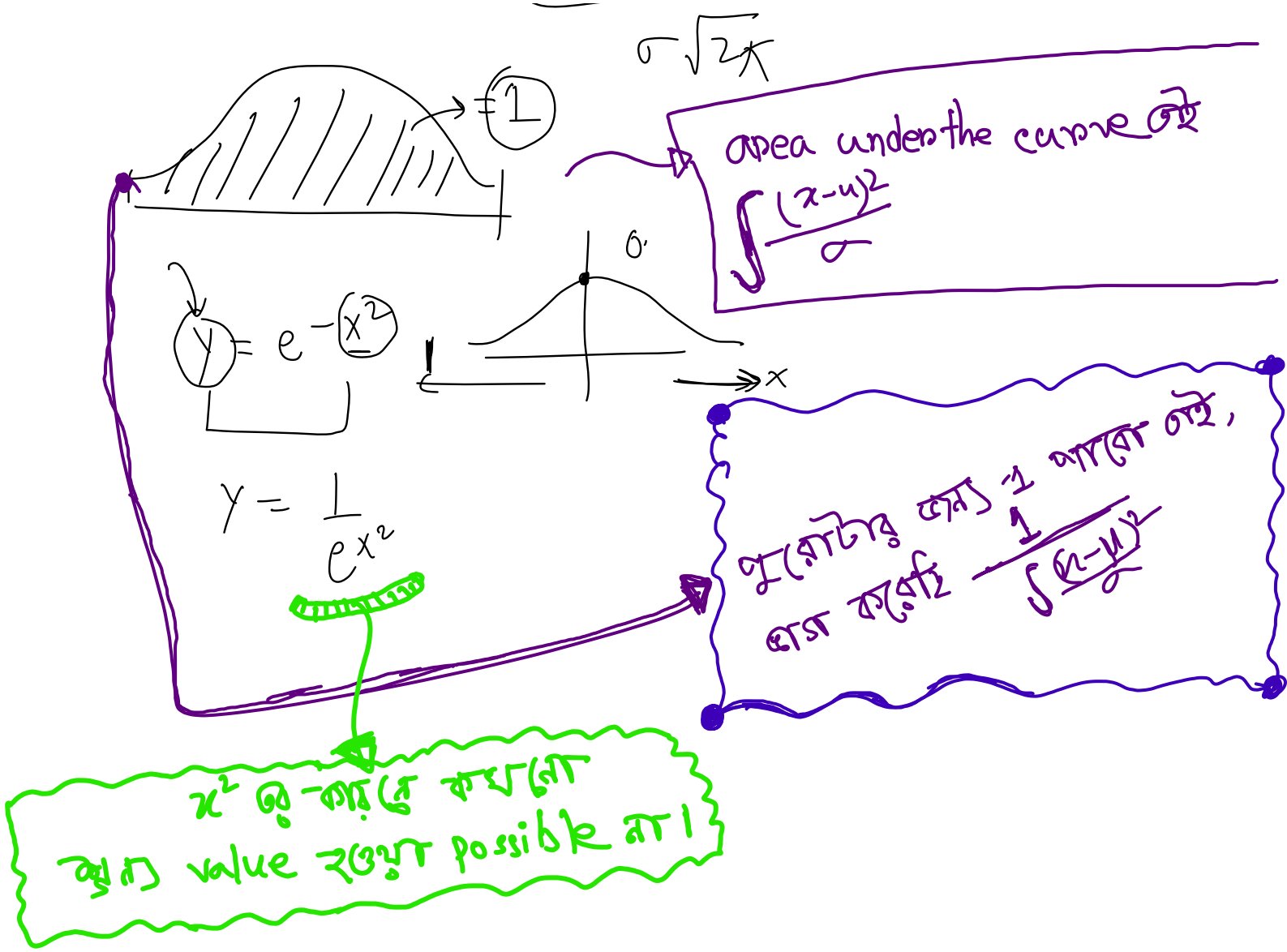
$$e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \div \sigma\sqrt{2\pi}$$

In desmos we can visualize the equation :

i) e^{+x}

ii) $e^{-x+\mu}$

iii) $e^{-\frac{x+\mu}{\sigma}}$

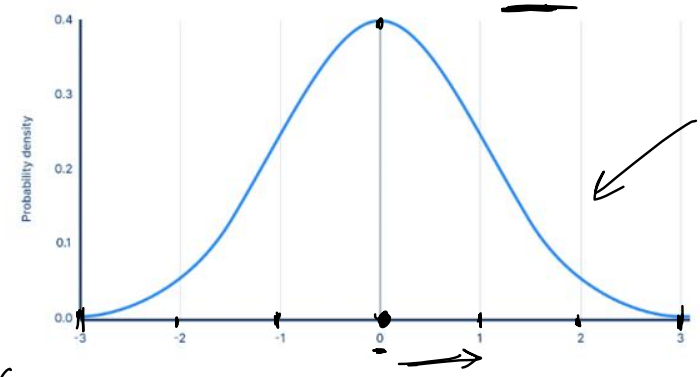

$\sigma\sqrt{2\pi}$
=1
Area under the curve এই
$\int \frac{(x-u)^2}{\sigma}$
0
$y = e^{-x^2}$
x
$y = \frac{1}{e^{x^2}}$
সুতোটার জন্য -1 থাকবে এই,
এরর করেছি $\frac{1}{\int (x-u)^2}$
x^2 ঋণাত্মক কখনো
ঋণাত্মক value হওয়া possible না।

Standard Normal Variate (Z)

20 March 2023 18:08

- 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.



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

Equation:

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

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- How to transform a normal distribution to Standard Normal Variate

Refer Python code

Kya Fayda Standardize karne ka?

Question

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

- What are Z-tables

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

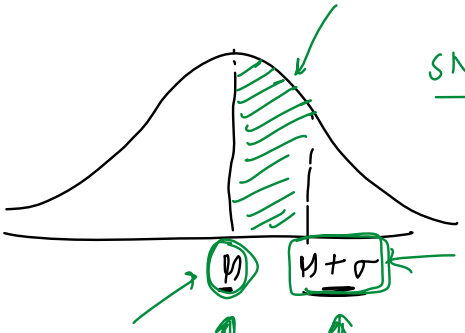
https://www.ztable.net/

Question

For a Normal Distribution X~(u,std) what percent of population lie between mean and 1 standard deviation, 2 std and 3 std?

standard deviation, 2 std and 3 std.

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



SNV(z) →

$\mu \rightarrow$ ↙

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

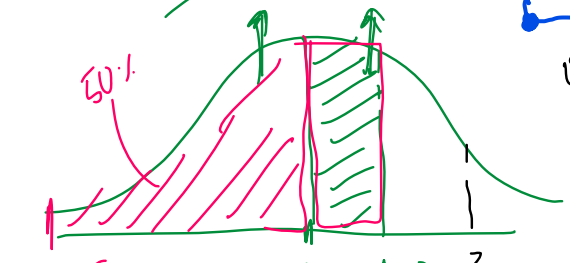
$$z = \frac{\mu-\mu}{\sigma} = 0 \quad \text{(i)}$$

$$z = \frac{x-\mu}{\sigma} = \frac{\mu+\sigma-\mu}{\sigma} \quad \text{(ii)}$$

0.9772 $= \frac{\sigma}{\sigma} = 1$

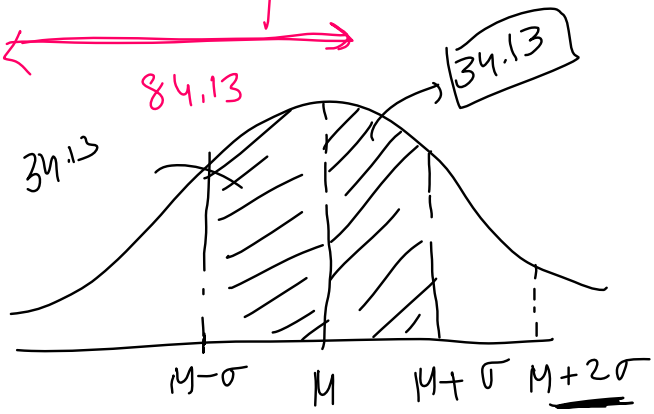
97%

84.13 − 50 = 34.13 from z table



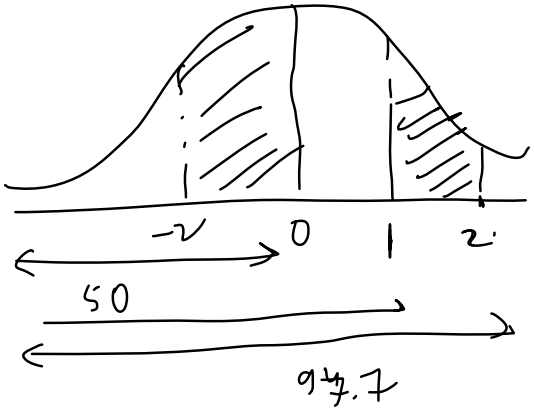
−1σ +1σ

64.28 %



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

49.

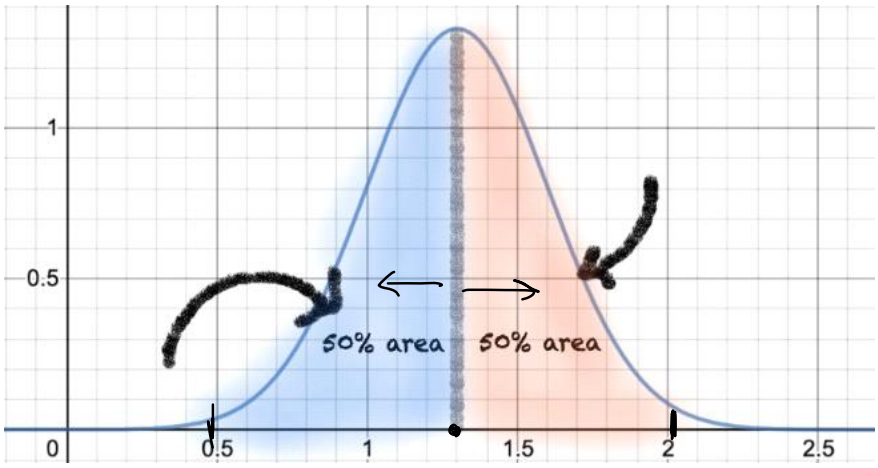


Properties of Normal Distribution

20 March 2023 18:06

1. Symmetricity

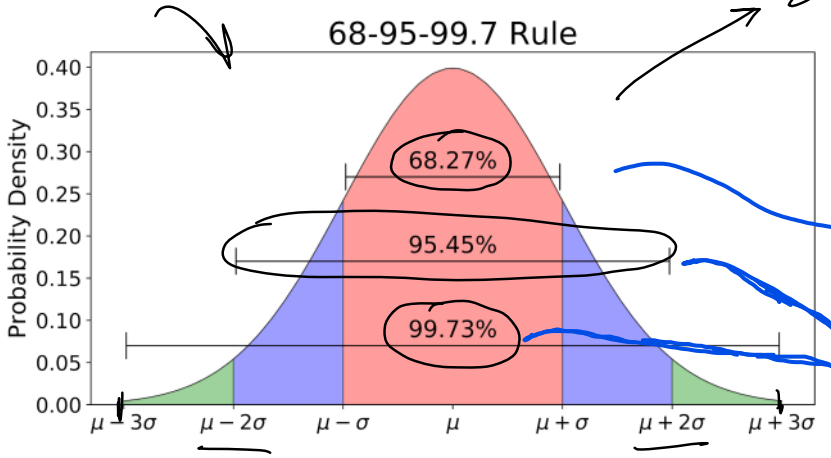
The normal distribution is symmetric about its mean, which means that the probability of observing a value above the mean is the same as the probability of observing a value below the mean. The bell-shaped curve of the normal distribution reflects this symmetry.

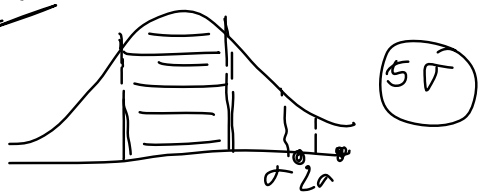


2. Measures of Central Tendencies are equal

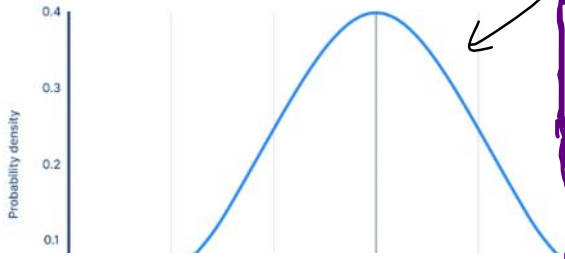
3. Empirical Rule

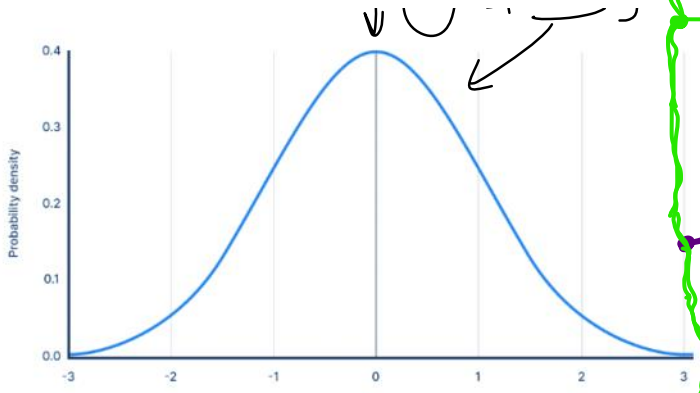
The normal distribution has a well-known empirical rule, also called the 68-95-99.7 rule, which states that approximately 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean, and about 99.7% of the data falls within three standard deviations of the mean.





4. The area under the curve





z-score হচ্ছে, area under curve বের করে দেয়।

যে কোন PDF এর জন্য area under curve 1 হবে। Always.

মজার ব্যাপার হচ্ছে, Broadman নামের যে scientist আছে, যেখানে normal distribution এ (3rd std about mean) 99% data fall করে বেশিরভাগ, 5th std deviation এর মধ্যে পড়তো। মানে তিনি বলতে চাচ্ছে, এর থেকে outliers এর বাপ বলতে পারি।

Skewness :

20 March 2023 18:07

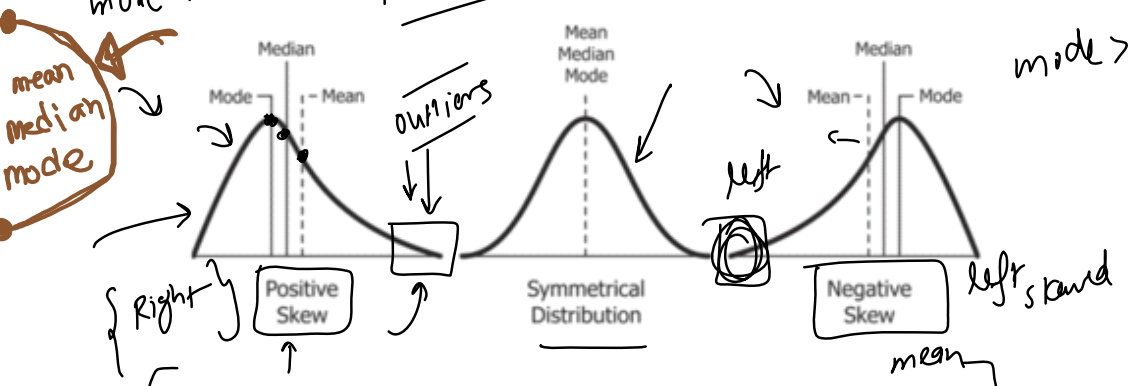
What is skewness?

A normal distribution is a bell-shaped, symmetrical distribution with a specific mathematical formula that describes how the data is spread out. Skewness indicates that the data is not symmetrical, which means it is not normally distributed.

Skewness is a measure of the asymmetry of a probability distribution. It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.

In a symmetrical distribution, the mean, median, and mode are all equal. In contrast, in a skewed distribution, the mean, median, and mode are not equal, and the distribution tends to have a longer tail on one side than the other.

Skewness can be positive, negative, or zero. A positive skewness means that the tail of the distribution is longer on the right side, while a negative skewness means that the tail is longer on the left side. A zero skewness indicates a perfectly symmetrical distribution.

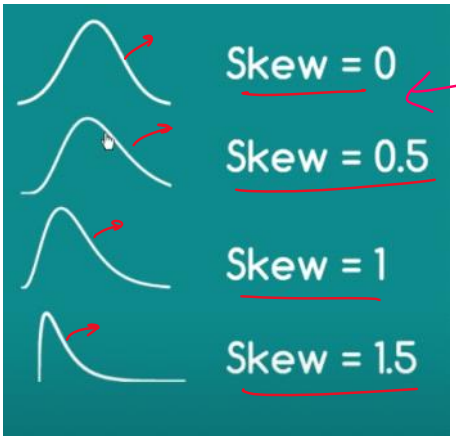


The greater the skew the greater the distance between mode, median and mean

How skewness is calculated?

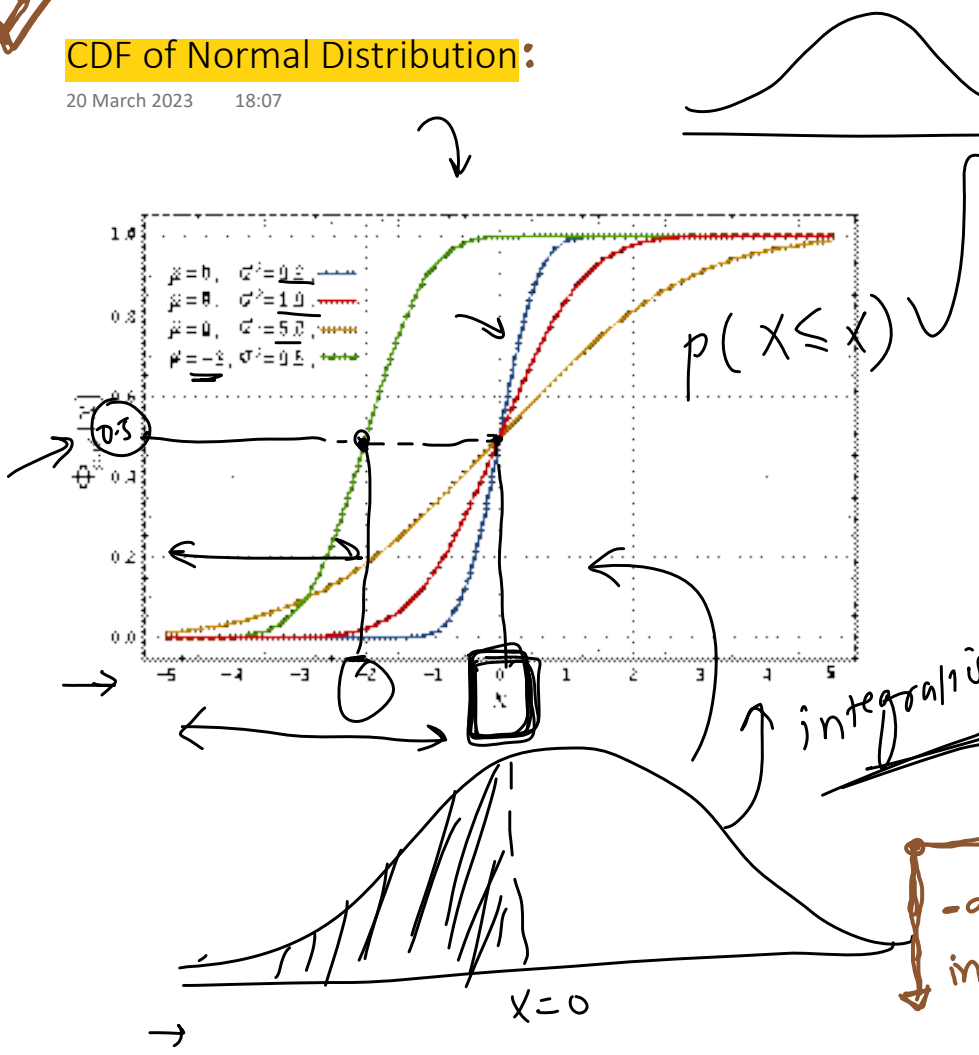
$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{(x-\bar{x})}{s}\right)^3$$

- Python Example
- Interpretation



CDF of Normal Distribution:

20 March 2023 18:07



$$F(x) = p(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

CDF ে আমরা যে কোন point দেওয়া থাকলে তার আগ পর্যন্ত করবো। অর্থাৎ, সেই point টুকুর পোবেবল বের করার জন্য -∞ থেকে সেই point পর্যন্ত integration করবো।

Use in Data Science

Normal distribution

20 March 2023 18:08

- Outlier detection
- Assumptions on data for ML algorithms -> Linear Regression and GMM
- Hypothesis Testing
- Central Limit Theorem