Recap

21 March 2023 14:50

Random Var → Dis → PMF

Random Var → Cont → PDF

PMF, PDF → CDF

PDF → Para

PDF → Nonpara

2 class → $\boxed{\text{Normal Distri}}$

2 class → remaining Distri

→ 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

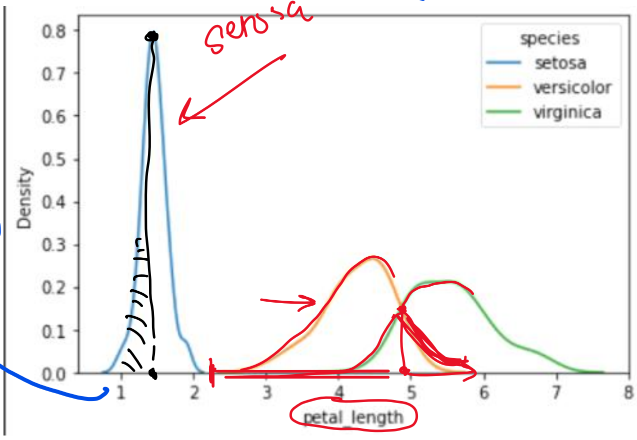
$pl > 5$ vir

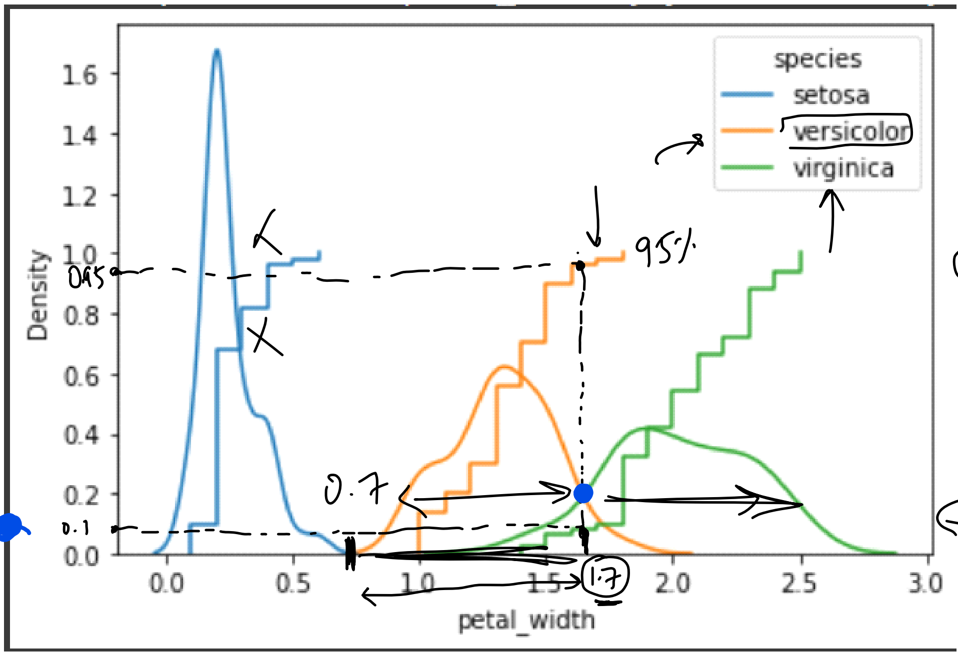
$P(X \le x)$

$pl < 2.3$

>2.3

petal length (১~২) 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 ১টা 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 ওখানে Kdeplot dominate করতেছে virginica ফুলের চেয়ে।) ⊛

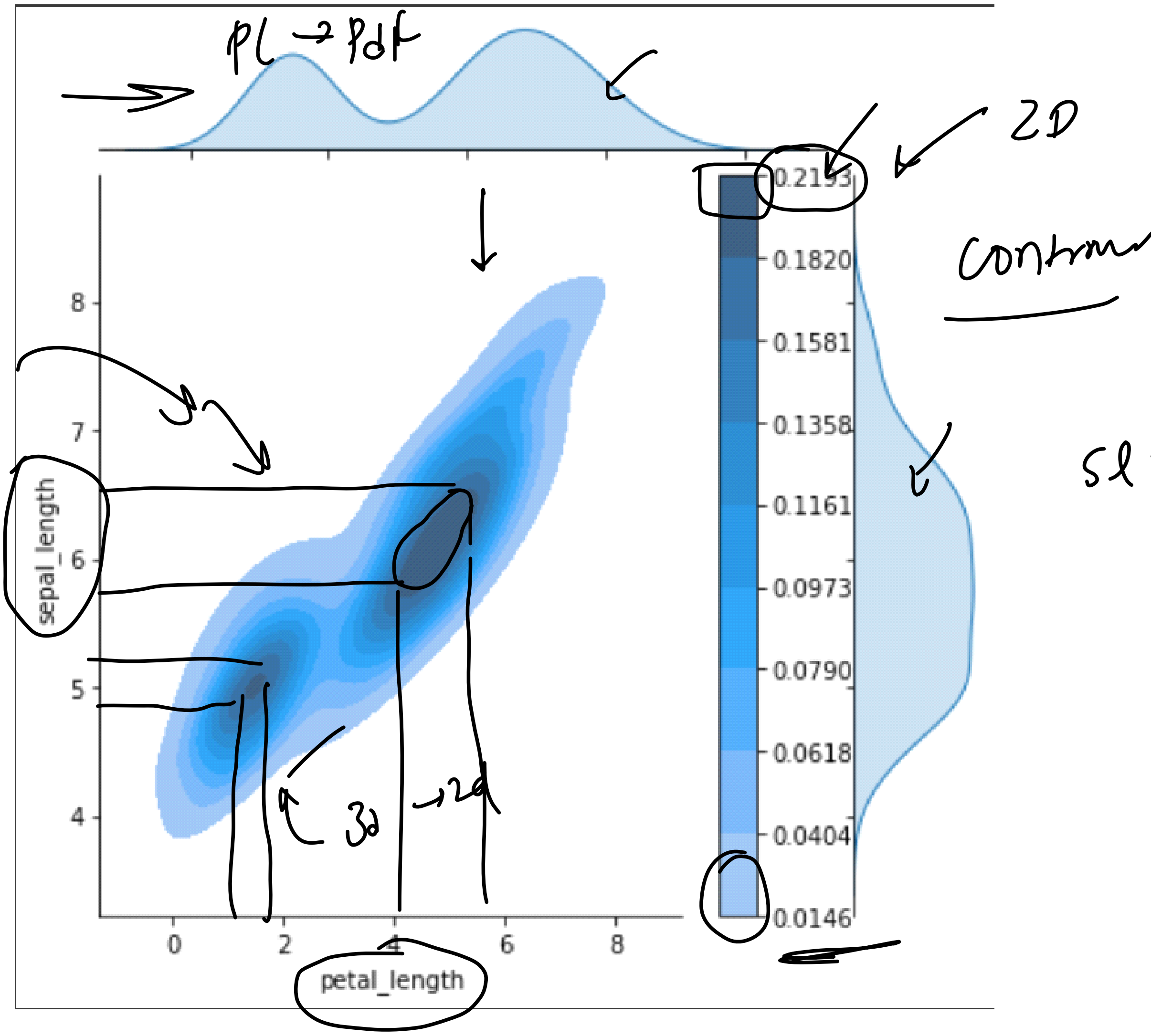
2D Density Plots

20 March 2023 18:11

1D distributionplot → 2D → 3D

PMF PDF 4

2D density → 2 cols



color

Normal Distribution

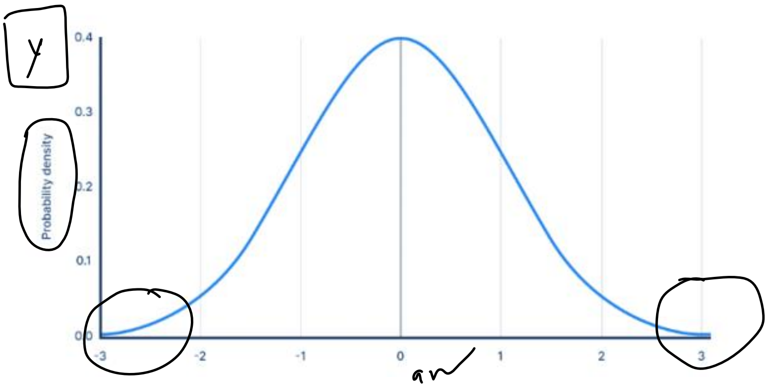
20 March 2023 18:06

bell curve

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.

→ pdf



-> 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.

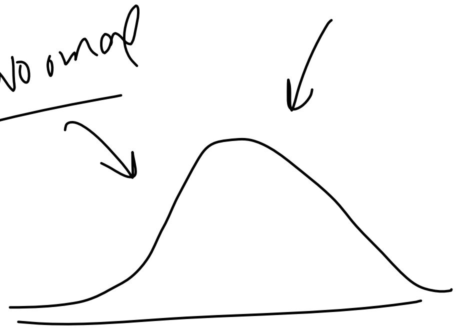
Denoted as:

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

$\mu \to$ mean

$\sigma \to$ std

No normal



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

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

$\boxed{\mu, \sigma}$

data → Normal → $\boxed{\mu, \sigma}$

$\sigma\sqrt{2\pi}$

Parameters in Normal Distribution

intpr

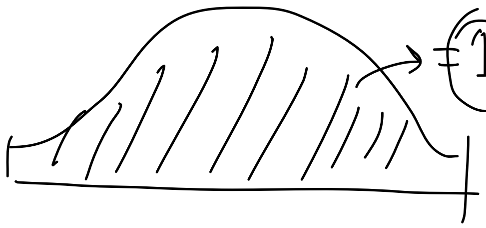
https://samp-suman-normal-dist-visualize-app-lkntug.streamlit.app/

Equation in detail:

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

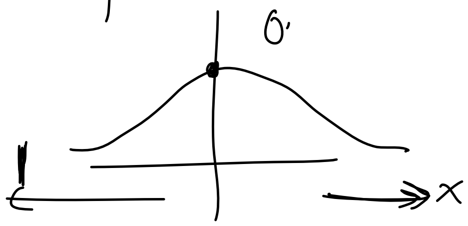
$\sigma\sqrt{2\pi}$

$\sigma\sqrt{2\pi}$



$= 1$

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



$$y = \frac{1}{e^{x^2}}$$

Standard Normal Variate(Z) → Standard Normal distribution

20 March 2023 18:08

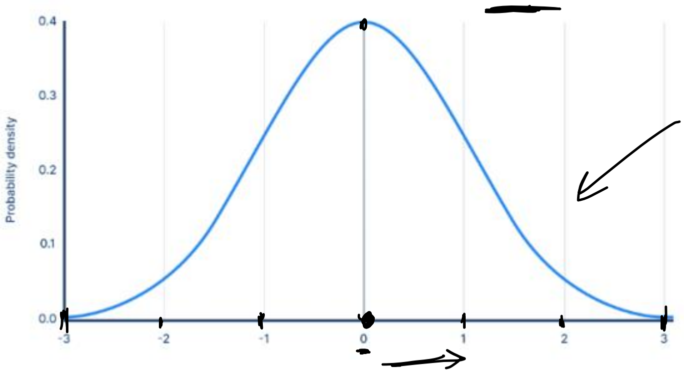
$X \sim N(\mu, \sigma)$ $\mu = 0$, $\sigma = 1$

$Z \sim N(0, 1)$

- 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.

(Z)



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

Equation:

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

$X \sim N(5, 2.5)$

$\sigma = 1$, $\mu = 0$

- How to transform a normal distribution to Standard Normal Variate

$X \sim N(29, 14) \rightarrow N(0,1)$

Refer Python code

Kya Fayda Standardize karne ka?

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?

Age: 27, 25, 61, 73 → μ, σ

$Z = \frac{27-\mu}{\sigma}$

$X \sim N(68, 3)$ → (Z) → $Z = \frac{72-68}{3} = \frac{4}{3}$

1.3, 1,33, 0.03, 1.33 → 0.90824 → 90.8

100 − 90.8 = 9,02

- 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/

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

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

$\mu \rightarrow$

$Z = X - \mu$

standard deviation, 2 std and 3 std?

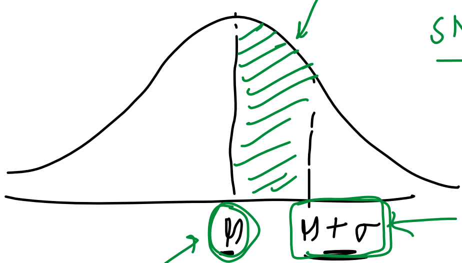
$X \sim N(\mu, \sigma)$ %

$\mu \rightarrow$

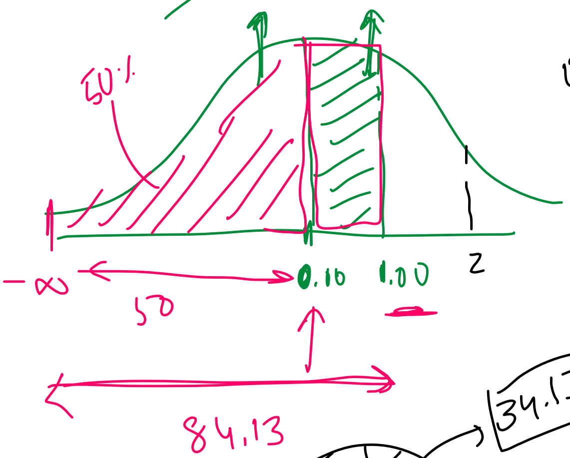
$$z = \frac{x - \mu}{\sigma}$$

SNV(z) $\rightarrow$

$$z = \frac{\mu - \mu}{\sigma} = 0$$



$$z = \frac{x-\mu}{\sigma} = \frac{\cancel{\mu} + \sigma - \cancel{\mu}}{\sigma} = \frac{\sigma}{\sigma} = 1$$



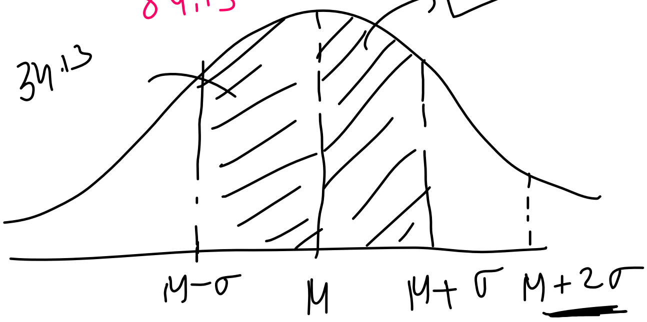
0.9772

97%

$84.13 - 50 = \boxed{84.13}$

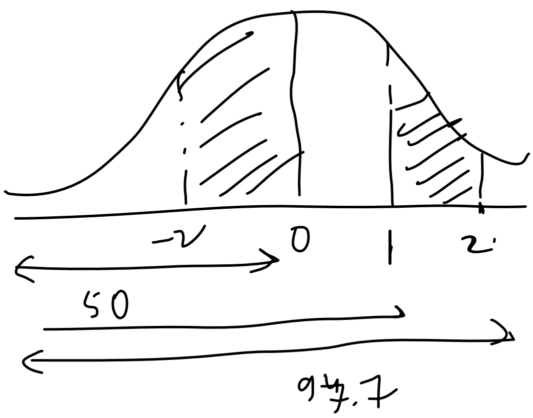
$-1\sigma \qquad +1\sigma$

$\boxed{64.28}$ %



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

69.

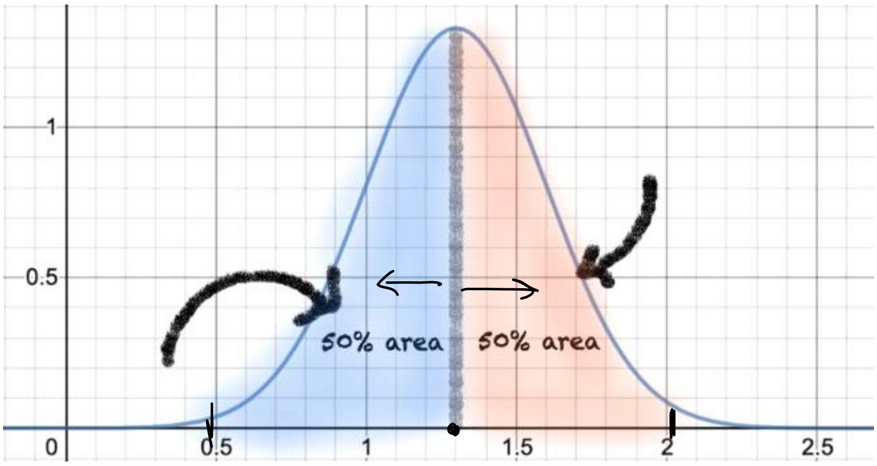


Properties of Normal Distribution

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.



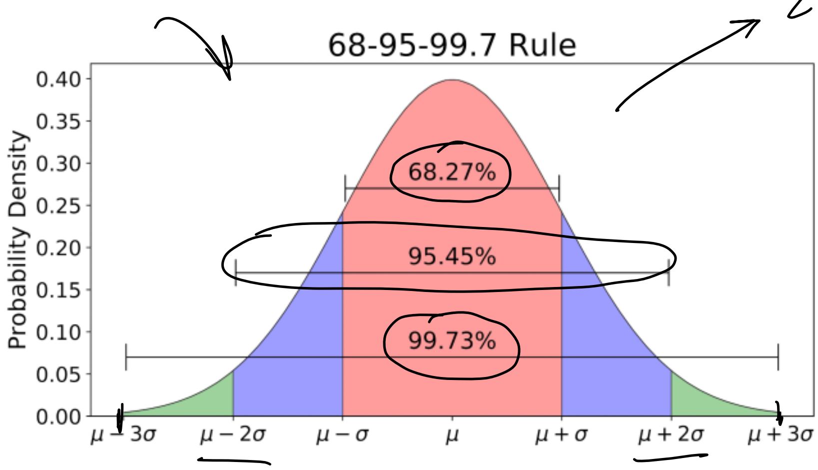
2. Measures of Central Tendencies are equal → mean → median → mode

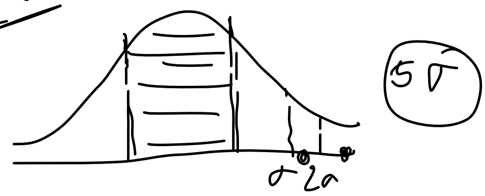
3. Empirical Rule

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.

Standard normal variate

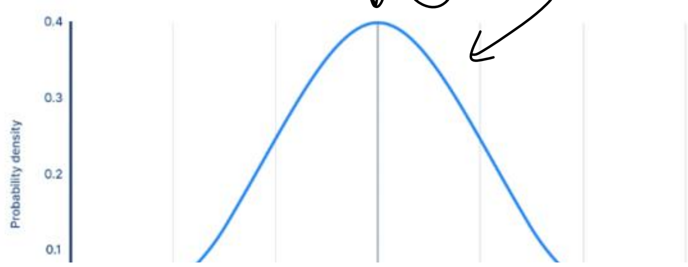
Z-table

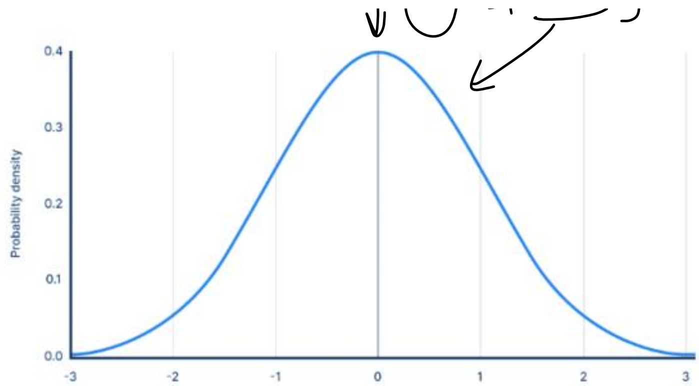




4. The area under the curve

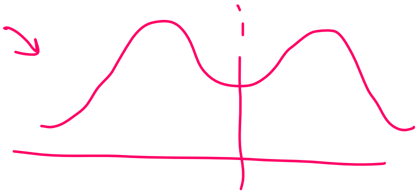
1 → PDF




Probability density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

Skewness

20 March 2023 18:07



→ symmetric

- 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.

moment

1 mom mean

2 mome — varia

3 momt → skew

4th → kurtosis

mode < median < mean

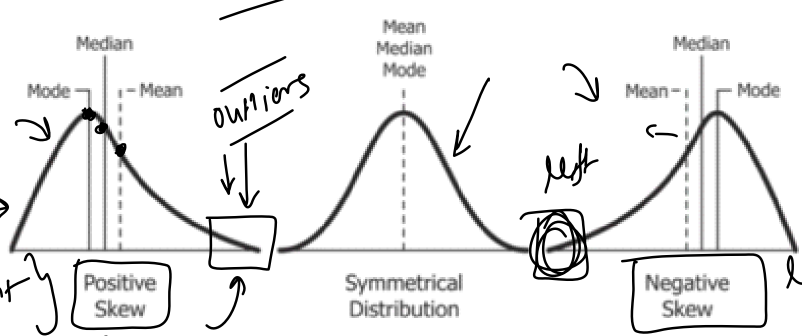
tail event

mode > median > mean

outliers

left

Right



left skewed

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

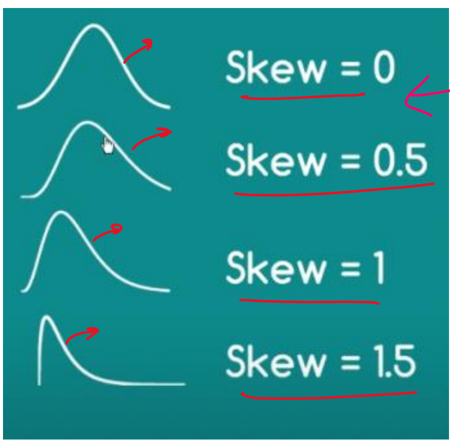
- How skewness is calculated?

moment

sample skew

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

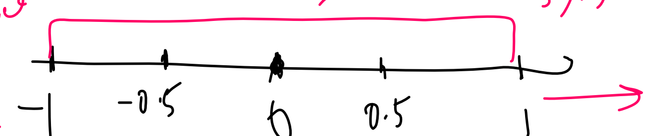
- Python Example
- Interpretation



highly skewed

0.38

moderately sym

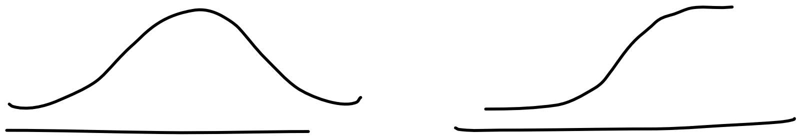


-1 -0.5 0 0.5 1

almost sym

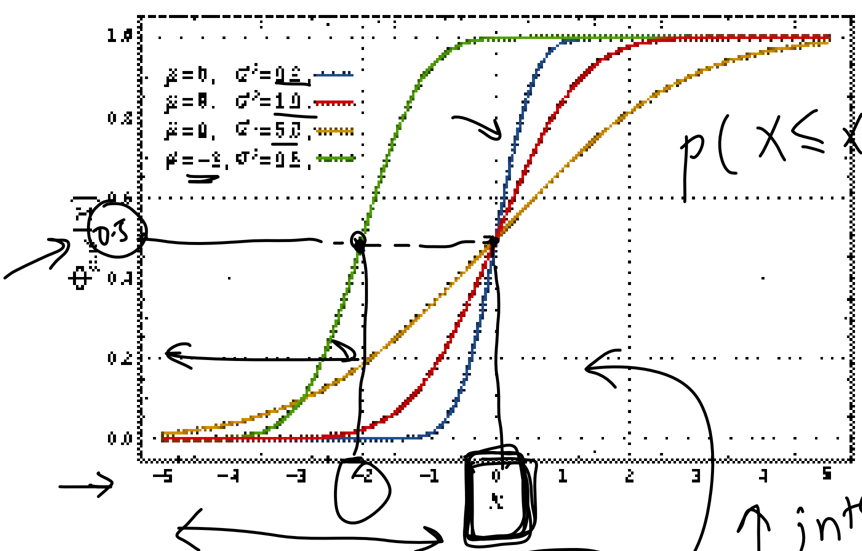
CDF of Normal Distribution

20 March 2023 18:07



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

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



$p(X \le x)$

integration

Use in Data Science

20 March 2023 18:08

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