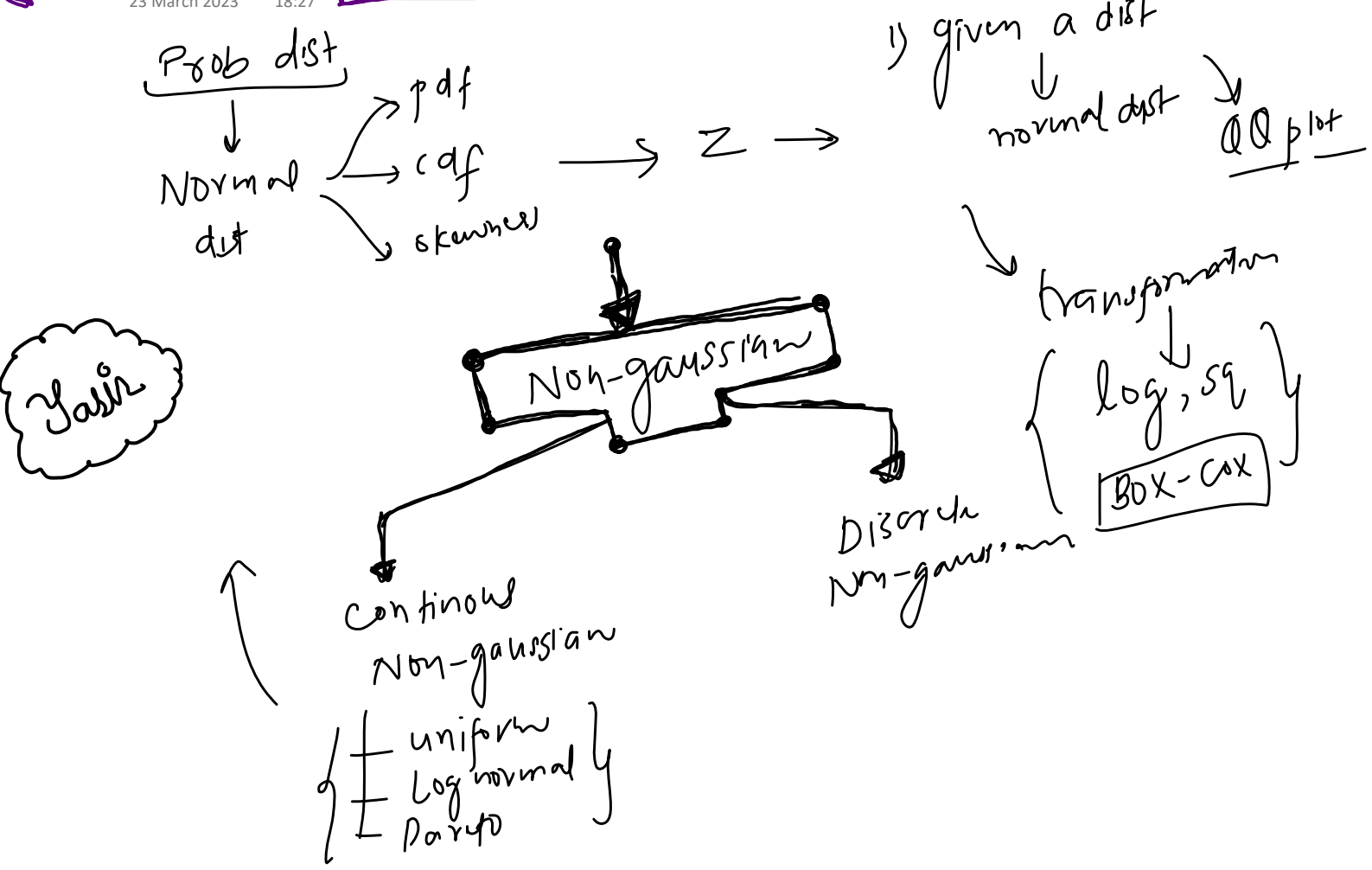
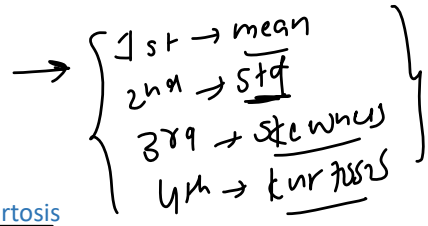
আমরা, আগের Lecture এ, gaussian (Normal) distribution নিয়ে পড়েছি।
Recap, এখন, আমরা Non-gaussian distribution নিয়ে পড়বো।

23 March 2023 18:27



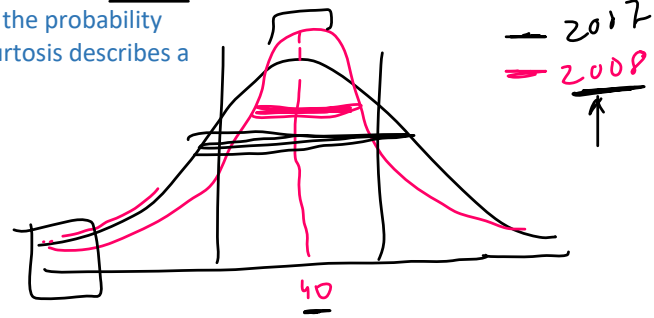
Kurtosis

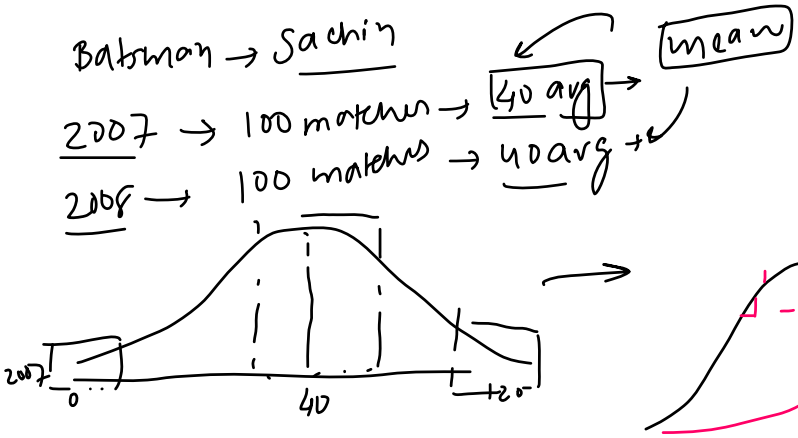
23 March 2023 13:18

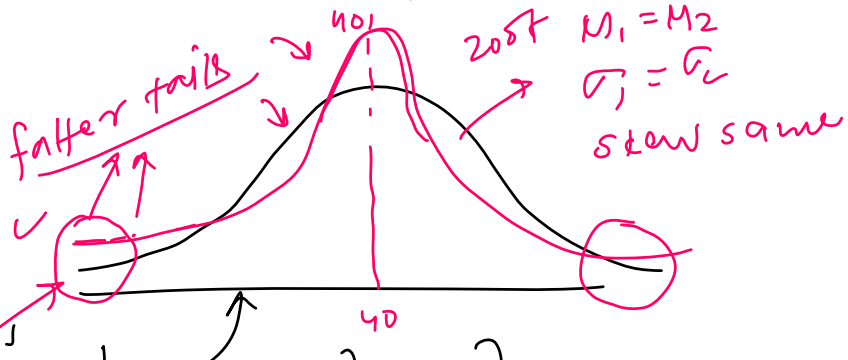


- What is Kurtosis?

Kurtosis is the 4th statistical moment. In probability theory and statistics, kurtosis (meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describes a particular aspect of a probability distribution.







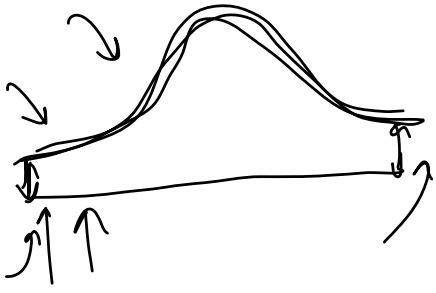
- False notation about Kurtosis

https://en.wikipedia.org/wiki/Kurtosis

- Formula

sample kurtosis

$$\left\{\frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_i^n \left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$



- Practical Use-case

In finance, kurtosis risk refers to the risk associated with the possibility of extreme outcomes or "fat tails" in the distribution of returns of a particular asset or portfolio.

If a distribution has high kurtosis, it means that there is a higher likelihood of extreme events occurring, either positive or negative, compared to a normal distribution.

In finance, kurtosis risk is important to consider because it indicates that there is a greater probability of large losses or gains occurring, which can have significant implications for investors. As a result, investors may want to adjust their investment strategies to account for kurtosis risk.

- Excess Kurtosis & Types

Excess kurtosis is a measure of how much more peaked or flat a distribution is

Excess Kurtosis & Types

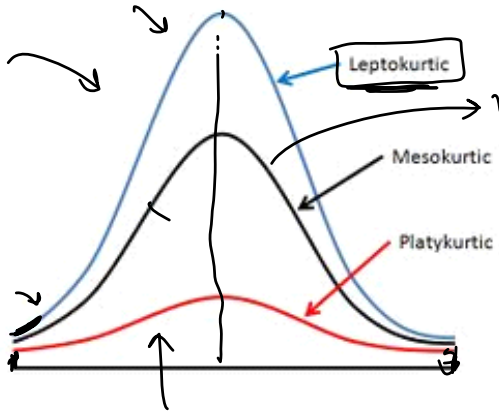
Excess kurtosis is a measure of how much more peaked or flat a distribution is compared to a normal distribution, which is considered to have a kurtosis of 0. It is calculated by subtracting 3 from the sample kurtosis coefficient.

Types of Kurtosis

Leptokurtic

A distribution with positive excess kurtosis is called leptokurtic. "Lepto-" means "slender". In terms of shape, a leptokurtic distribution has fatter tails. This indicates that there are more extreme values or outliers in the distribution.

Example - Assets with positive excess kurtosis are riskier and more volatile than those with a normal distribution, and they may experience sudden price movements that can result in significant gains or losses.



Platykurtic

A distribution with negative excess kurtosis is called platykurtic. "Platy-" means "broad". In terms of shape, a platykurtic distribution has thinner tails. This indicates that there are fewer extreme values or outliers in the distribution.

Assets with negative excess kurtosis are less risky and less volatile than those with a normal distribution, and they may experience more gradual price movements that are less likely to result in large gains or losses.

Mesokurtic

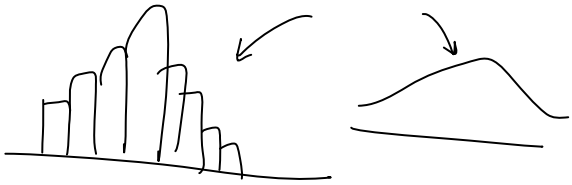
Distributions with zero excess kurtosis are called mesokurtic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters.

Mesokurtic is a term used to describe a distribution with a excess kurtosis of 0, indicating that it has the same degree of "peakedness" or "flatness" as a normal distribution.

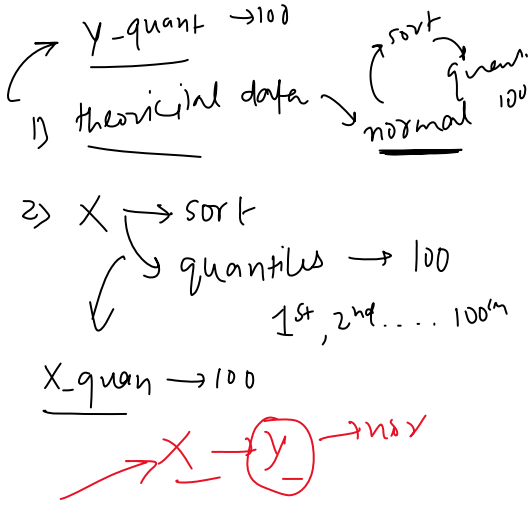
Example -
In finance, a mesokurtic distribution is considered to be the ideal distribution for assets or portfolios, as it represents a balance between risk and return.

QQ Plot

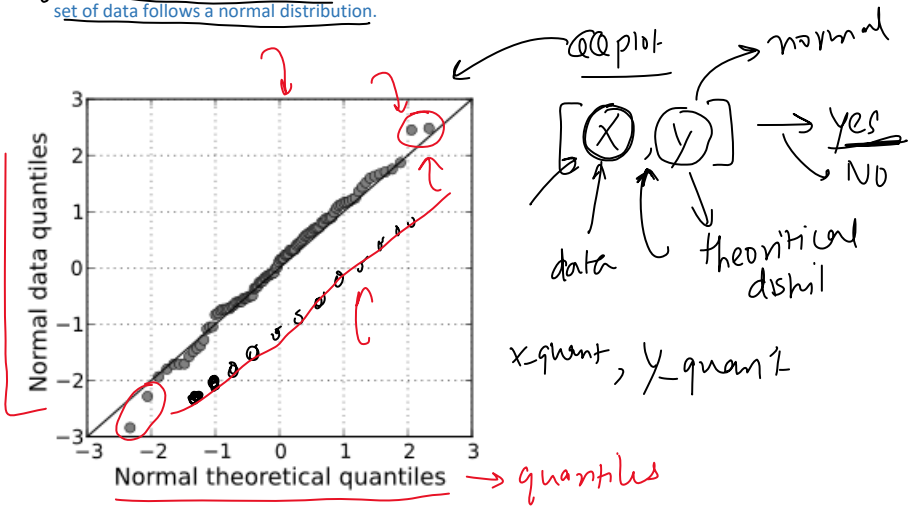
23 March 2023 13:19



- How to find if a given distribution is normal or not?
  - *Visual inspection* One of the easiest ways to check for normality is to visually inspect a histogram or a density plot of the data. A normal distribution has a bell-shaped curve, which means that the majority of the data falls in the middle, and the tails taper off symmetrically. If the distribution looks approximately bell-shaped, it is likely to be normal.
  - *QQ Plot* Another way to check for normality is to create a normal probability plot (also known as a Q-Q plot) of the data. A normal probability plot plots the observed data against the expected values of a normal distribution. If the data points fall along a straight line, the distribution is likely to be normal.
  - *Statistical tests*: There are several statistical tests that can be used to test for normality, such as the Shapiro-Wilk test, the Anderson-Darling test, and the Kolmogorov-Smirnov test. These tests compare the observed data to the expected values of a normal distribution and provide a p-value that indicates whether the data is likely to be normal or not. A p-value less than the significance level (usually 0.05) suggests that the data is not normal.



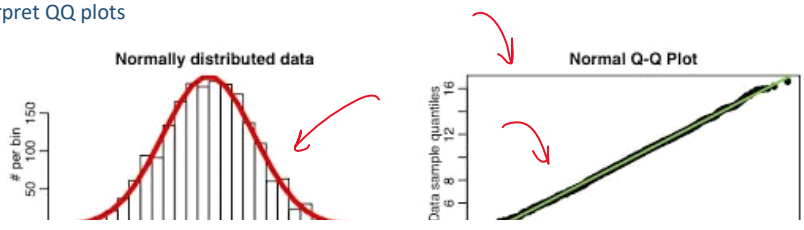
- What is a QQ Plot and how is it plotted?
  - A QQ plot (quantile-quantile plot) is a graphical tool used to assess the similarity of the distribution of two sets of data. It is particularly useful for determining whether a set of data follows a normal distribution.



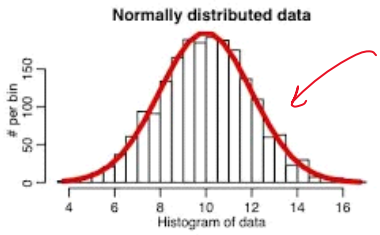
  - In a QQ plot, the quantiles of the two sets of data are plotted against each other. The quantiles of one set of data are plotted on the x-axis, while the quantiles of the other set of data are plotted on the y-axis. If the two sets of data have the same distribution, the points on the QQ plot will fall on a straight line. If the two sets of data do not have the same distribution, the points will deviate from the straight line.

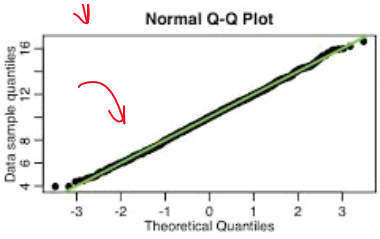
- Python example
- https://www.statsmodels.org/dev/generated/statsmodels.graphics.gofplots.qqplot.html

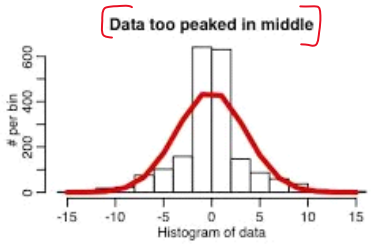
- How to interpret QQ plots

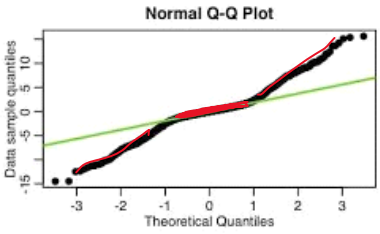


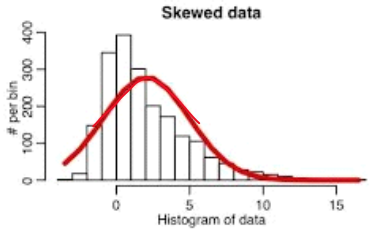
1.

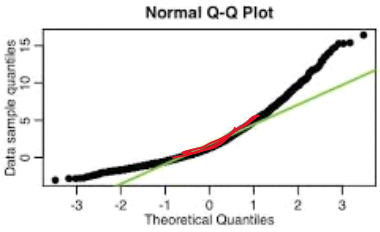

Normally distributed data
per bin
Histogram of data

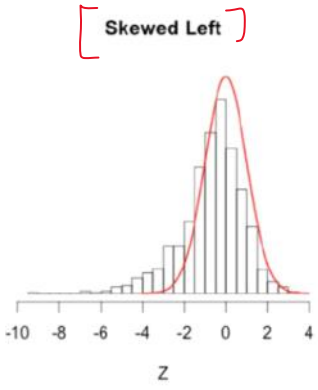

Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles

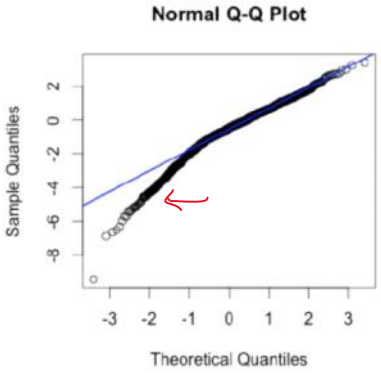

Data too peaked in middle
per bin
Histogram of data


Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles

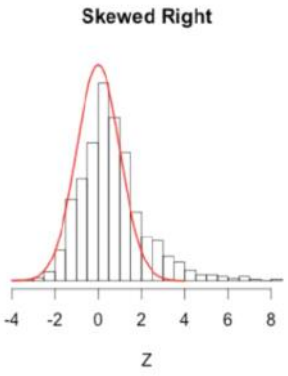

Skewed data
per bin
Histogram of data

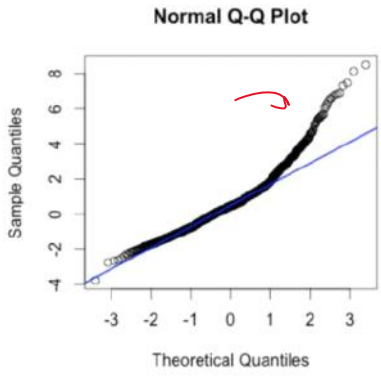

Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles


Skewed Left
Z

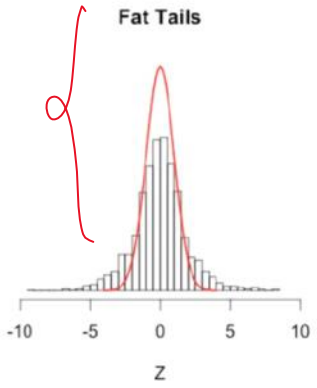

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

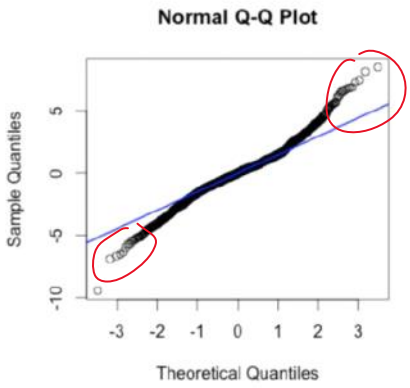
1.

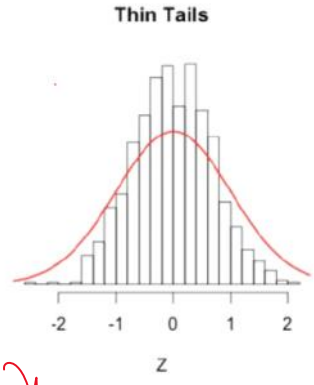

Skewed Right
Z

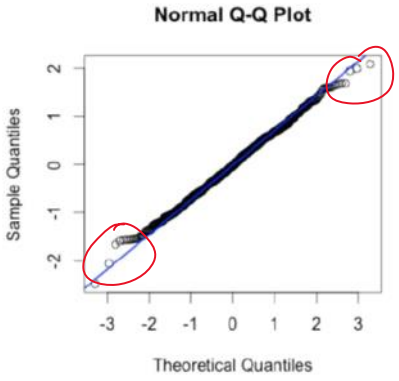

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

1.

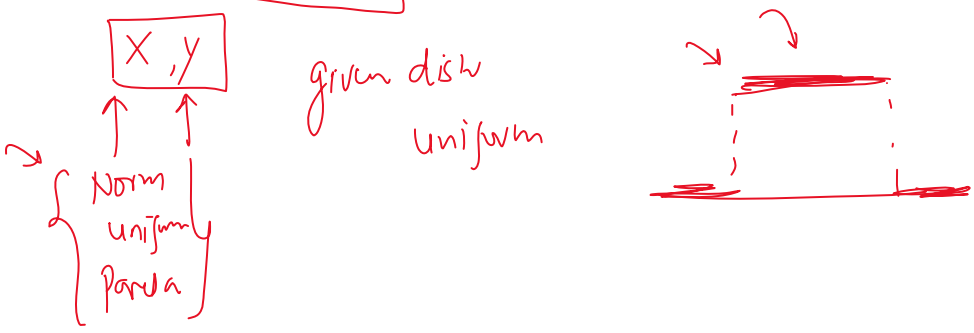








- Does QQ plot only detect normal distribution?



Uniform Distribution

23 March 2023 13:19

- What is Uniform Distribution and it's types

In probability theory and statistics, a uniform distribution is a probability distribution where all outcomes are equally likely within a given range. This means that if you were to select a random value from this range, any value would be as likely as any other value.

Types

Uniform (X)

Discrete — Continuous

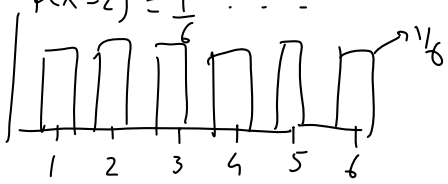
Denoted as $X \sim U(a,b)$ → parameters (a ↔ b; a: lower, b: high)

a=1 b=6

dice roll → {1, 2, 3, 4, 5, 6}

$P(X=1) = \frac{1}{6}$

$P(X=2) = \frac{1}{6}$ . . .

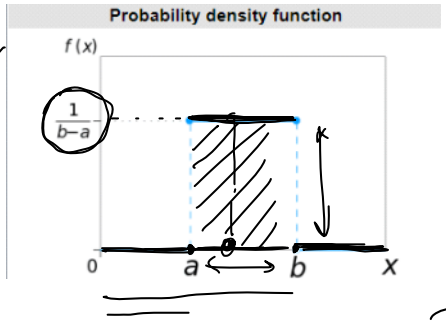


- Examples

a. The height of a person randomly selected from a group of individuals whose heights range from 5'6" to 6'0" would follow a continuous uniform distribution.
b. The time it takes for a machine to produce a product, where the production time ranges from 5 to 10 minutes, would follow a continuous uniform distribution.
c. The distance that a randomly selected car travels on a tank of gas, where the distance ranges from 300 to 400 miles, would follow a continuous uniform distribution.
d. The weight of a randomly selected apple from a basket of apples that weighs between 100 and 200 grams, would follow a continuous uniform distribution.

- PDF CDF and Graphs

PDF

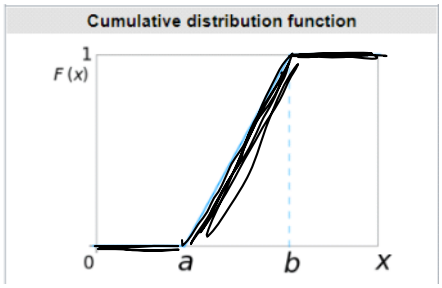


$\frac{1}{b-a}$ ← $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for otherwise} \end{cases}$

$x < a$ or $x > b$

$b-a$

CDF

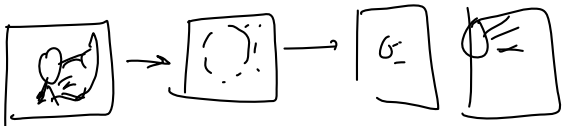


https://en.wikipedia.org/wiki/Continuous_uniform_distribution

- Skewness → 0 → Symmetric → Normal

- Application in Machine learning and Data Science

a. **Random initialization** In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected

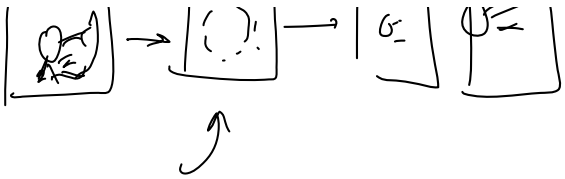


a. **Random initialization**: In many machine learning algorithms, such as neural networks and k-means clustering, the initial values of the parameters can have a significant impact on the final result. Uniform distribution is often used to randomly initialize the parameters, as it ensures that all values in the range have an equal probability of being selected.

b. **Sampling**: Uniform distribution can also be used for sampling. For example, if you have a dataset with an equal number of samples from each class, you can use uniform distribution to randomly select a subset of the data that is representative of all the classes.

c. **Data augmentation**: In some cases, you may want to artificially increase the size of your dataset by generating new examples that are similar to the original data. Uniform distribution can be used to generate new data points that are within a specified range of the original data.

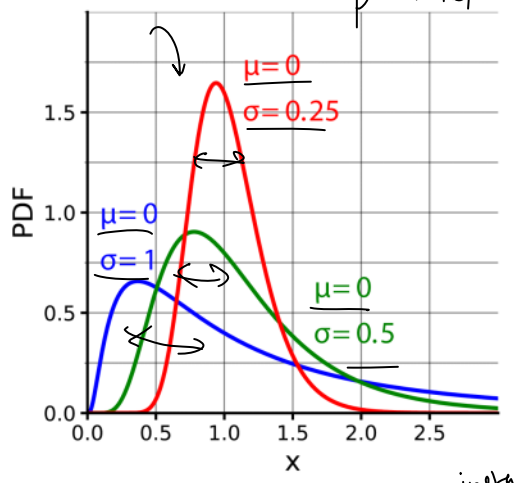
d. **Hyperparameter tuning**: Uniform distribution can also be used in hyperparameter tuning, where you need to search for the best combination of hyperparameters for a machine learning model. By defining a uniform prior distribution for each hyperparameter, you can sample from the distribution to explore the hyperparameter space.



Log Normal Distribution

23 March 2023 13:19

In probability theory and statistics, a lognormal distribution is a heavy tailed continuous probability distribution of a random variable whose logarithm is normally distributed.



Examples

- The length of comments posted in Internet discussion forums follows a log-normal distribution.
- Users' dwell time on online articles (jokes, news etc.) follows a log-normal distribution.
- The length of chess games tends to follow a log-normal distribution.
- In economics, there is evidence that the income of 97%–99% of the population is distributed log-normally.

Denoted as

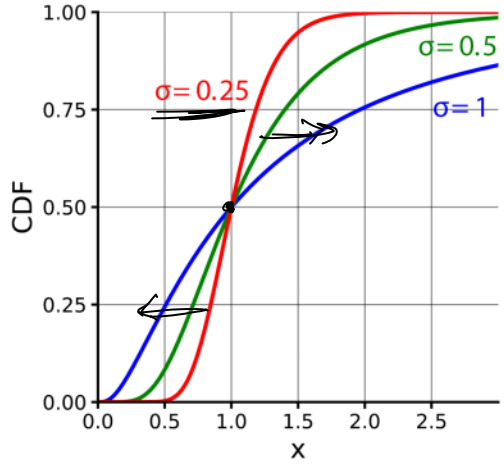
$X \sim \text{lognormal}(\mu, \sigma) \rightarrow \ln(x) \sim N(\mu', \sigma')$

PDF Equation

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

← similar Normal dist.

CDF



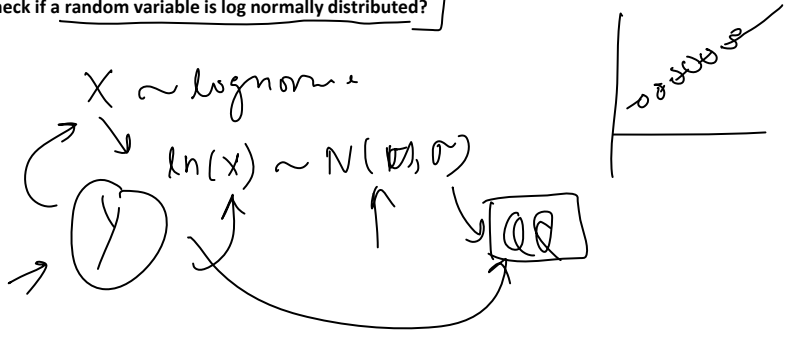
Skewness

Skewed

How to check if a random variable is log normally distributed?

Skewness

How to check if a random variable is log normally distributed?



Pareto Distribution

23 March 2023 13:19

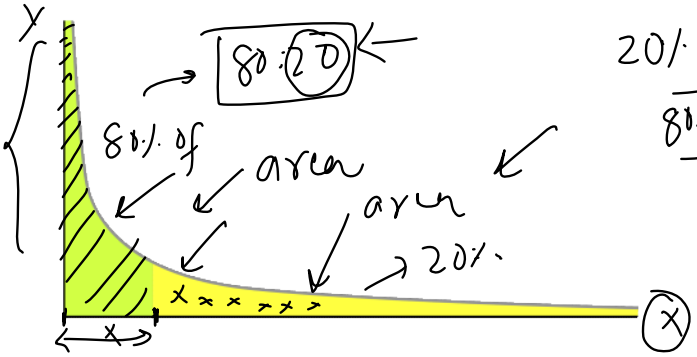
Pareto Distribution

The Pareto distribution is a type of probability distribution that is commonly used to model the distribution of wealth, income, and other quantities that exhibit a similar power-law behaviour

What is Power Law

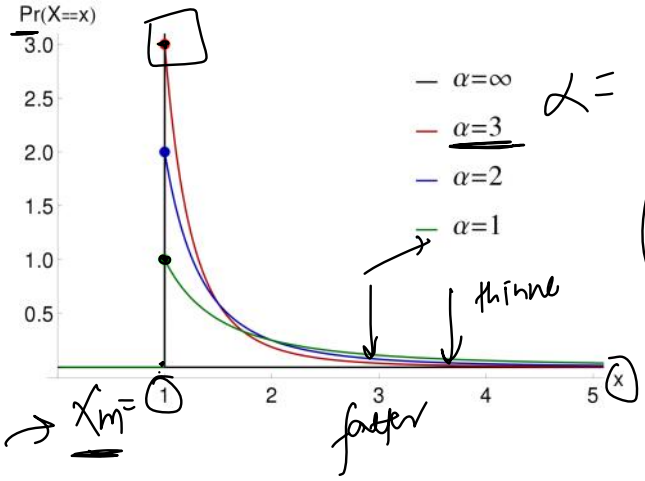
In mathematics, a power law is a functional relationship between two variables, where one variable is proportional to a power of the other. Specifically, if y and x are two variables related by a power law, then the relationship can be written as:

y = k * x^a



Vilfredo Pareto originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. He also used it to describe distribution of income. This idea is sometimes expressed more simply as the Pareto principle or the "80-20 rule" which says that 20% of the population controls 80% of the wealth

Graph & Parameters

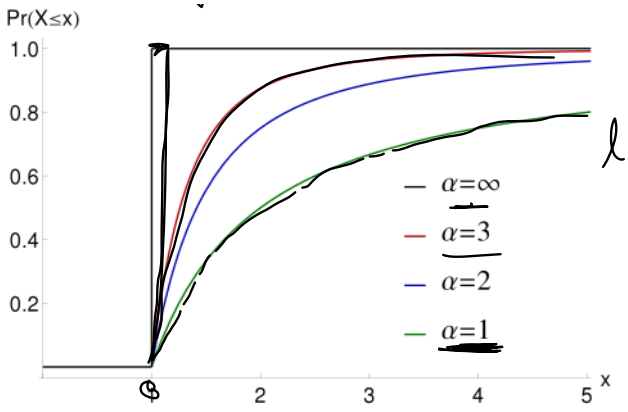


Examples

- The sizes of human settlements (few cities, many hamlets/villages)
- File size distribution of Internet traffic which uses the TCP protocol (many smaller files, few larger ones)

CDF

Pr(X<x)



log (y)

pareto

x log

Skewness → skewed

log-log plot → log

log x

[How to detect if a distribution is Pareto Distribution?]

$$y = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$$

QQ plot

x, y

pareto

Transformations

23 March 2023 13:24