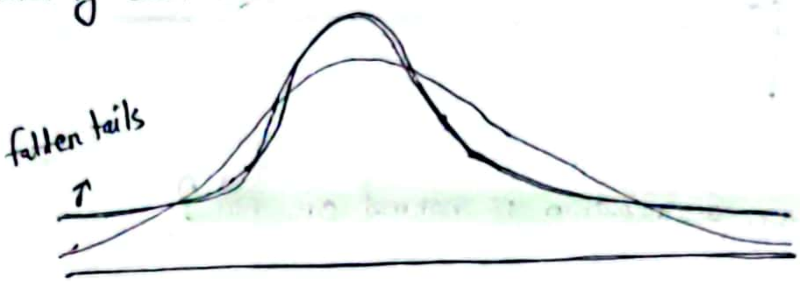
Non Gaussian Distribution:

Kurtosis: It is the 4th statistical moment. It is a measure of the "tailedness" of the probability distribution of a real-valued random variable.



There are three types of Kurtosis:

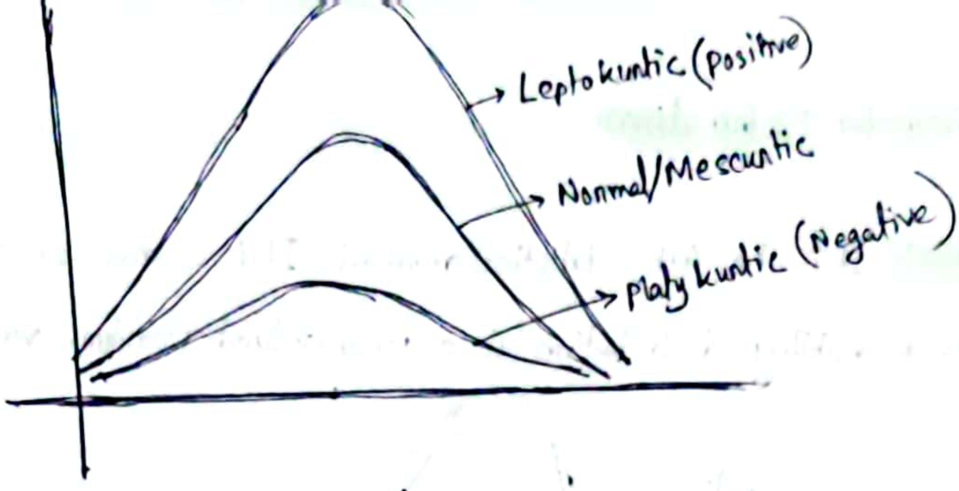
1) **Mesokurtic (Normal Kurtosis)**: A distribution with kurtosis = 3, which is a kurtosis of a normal distribution. In mesokurtic distribution, the tails have the same weight as a normal distribution.

2) **Leptokurtic (Positive Kurtosis)**: A distribution with kurtosis greater than 3, The tails are heavier than those of a normal distribution, indicating that there are more extreem values or outliers in the data.

3) **Platykurtic (Negative Kurtosis)**: A distribution with kurtosis less than 3. The tails are lighter than those of a normal distribution, indicating that there are few extreme values or outliers in the data.

Formula for **sample kurtosis:**

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



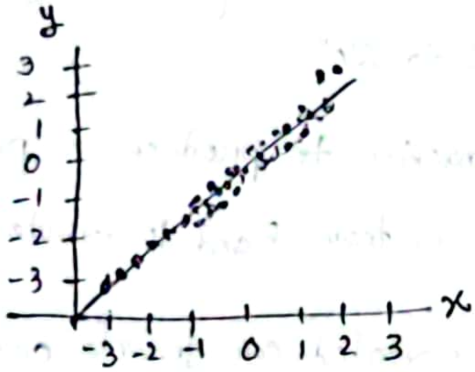
How to find if a given distribution is normal or not?

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 of symmetrically. If the distribution looks approximately bell shaped, it is likely to be normal distribution.

QQ Plot: Another way to check for normality is to create a normal probability plot (QQ plot). It 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 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 pvalue that indicates whether the data to be normal or not. A p value less than the significance level ~~(0.5)~~ (0.05) suggests that the data is not normal.

QQ Plot: A graphical tool to asses the similarity of distribution of two sets of data. It is particularly useful for determining whether a set of data follows a normal distribution



[x, y]

→ ~~Theoretical~~ Your data distribution

→ Theoretical ~~Your data~~ distribution

In QQ plot, there is a theorietical distribution (Generally Normal distribution) ~~who~~ with ~~y~~ whom you compare your data distribution similarity.

If all the comes under that line, that means distribution x is similar to y.

If ~~y~~ x was Normal distribution (we took) then we can say ~~x~~ y is also normally distributed.

Does QQ plot only detect Normal Distribution?

No, we also can compare Uniform, pareto etc.

Uniform Distribution: It is noted on PW-Skills part.

Example of Continous Uniform Distribution:

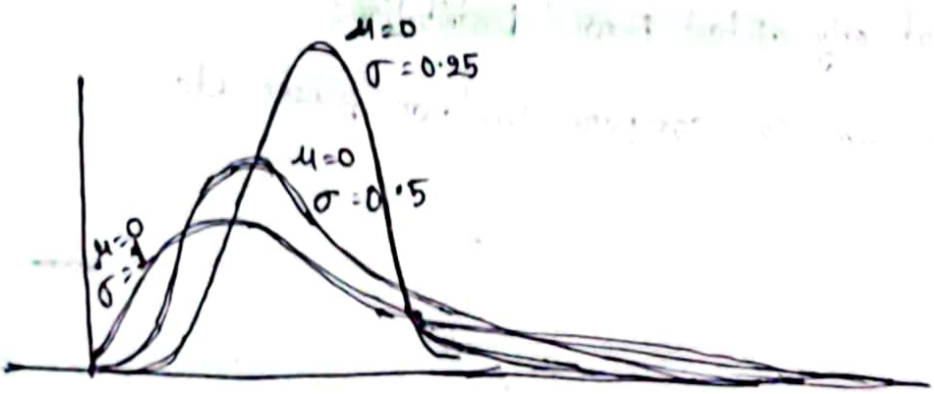
1) The height of a person selected randomely from a group of individuals whose heights range from 5'6" to 6'0"

2) The time it takes for a machine to produce a product, where the production time ranges from 5 and 10 minutes.

3) The distance that a randomly selected car travels on a tank of gass, where the distance ranges from 300 to 400 miles.

4) The weight of a randomly selected apple from a basket of apples, that weighs between 100 and 200 grams,

Log Normal Distribution: It is a heavy tailed (Right skewed) continuous probability distribution of a random variable whose logarithm is normally distributed.



Examples: 1) The length of comments posted on internet disscussion forums follows a log normal distribution

2) The length of a chess game tends to follow a log normal distribution

3) In economics, there is evidence that the income of 97-99% of the population, is distributed log-normally.

Formula: PDF → $\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$

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

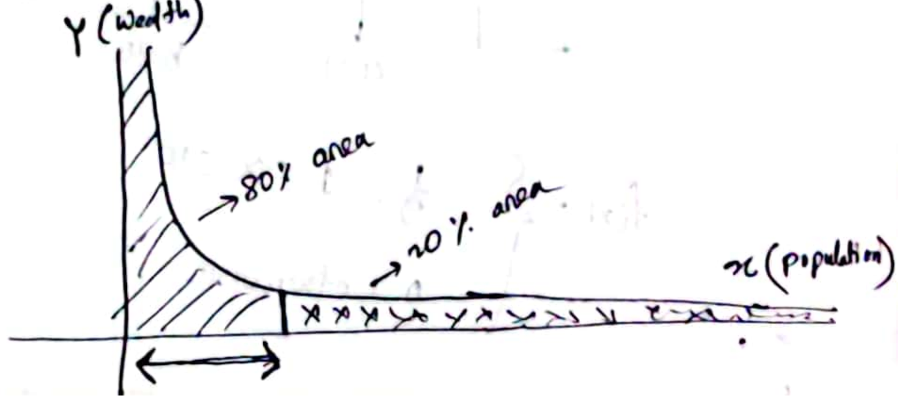
We have to take the log $(x \rightarrow \text{random var})$, then if it is converted to normal distribution, then that random variable is log ~~nn~~ normally distributed.

We will first take the log of random variable, then will plot it with qq plot. Then if y from the graph is normal, then ~~so~~ the random variable is log normal.

Pareto Distribution: It 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.

Power Law: A power law is a functional relationship between two variable, where one variable is proportional to a power of the other. Specially if y and x are two variables related ~~to~~ by a power law, then the relationship can be written as →

$y = K * x^a$



If this graph shows population (x) any ~~(y)~~ wealth (y), then it can be said, 20% of the population controls 80% of wealth. and 80% of population control 20% of wealth.

Some questions on Uniform Distribution:

Question 1: Suppose the wait time for a bus at a particular bus stop follows a uniform distribution between 5 minutes and 15 minutes. What is the probability that a person waiting at the bus stop will have to wait more than 10 mins for the next bus?

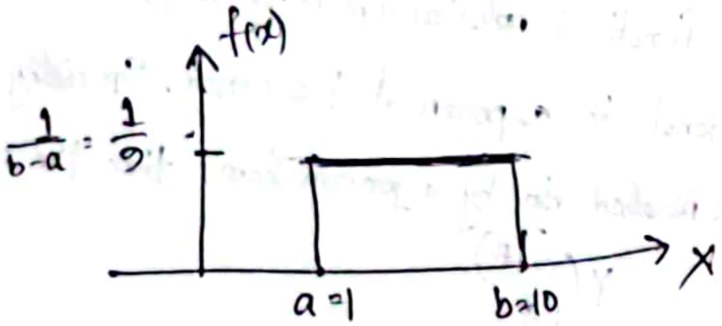
Ans: Probability (waiting >10 minutes) = (15 minutes − 10 minutes)/(15 minutes − 5 minutes)

$$= \frac{5}{10} = 0.5 \text{ or } 50\%$$

Question 2: If X is uniformly distributed in the interval $[1,10]$ then find →

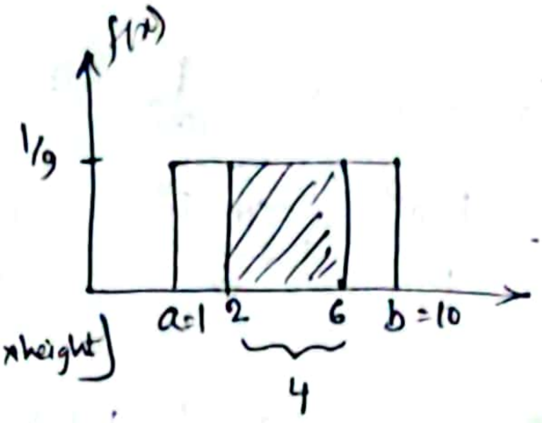
1. $P(2<X<6)$ 2. $P(X>3)$ 3. $P(X<6)$ 4. $P(2<X<9)$

5. $f(x)$, $E(X)$, $Var(X)$ and std. Deviation



$$f(x) = \begin{cases} \frac{1}{9}; & 1 \le x \le 10 \\ 0; & \text{otherwise} \end{cases}$$

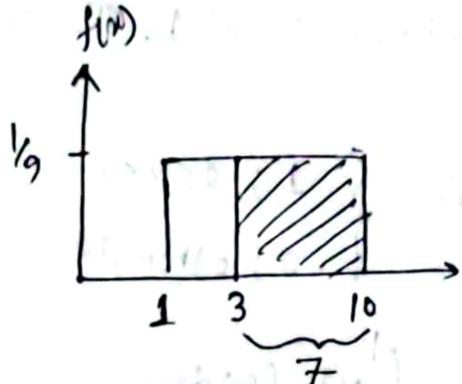
i) $P(2<X<6) = \int_{2}^{6} f(x)\,dx$



$= 4 \times \frac{1}{9}$ [width × height]

$= \frac{4}{9}$

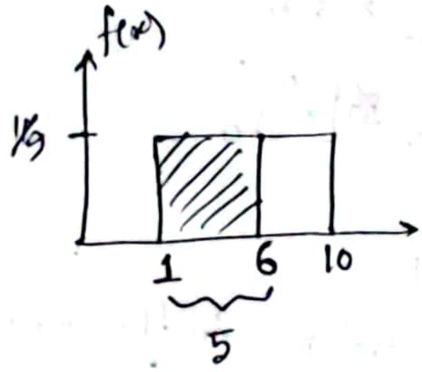
ii) $P(X>3) = \int_{3}^{\infty} f(x)\,dx$



$= 7 \times \frac{1}{9}$

$= \frac{7}{9}$ [Because max width can be 10]

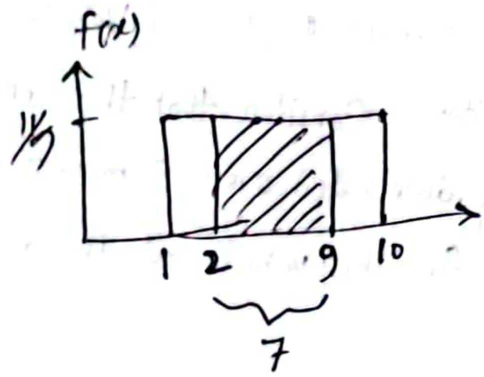
iii) $P(X<6) = \int_{-\infty}^{6} f(x)\,dx$



$= 5 \times \frac{1}{9}$

$= \frac{5}{9}$

iv) $P(2<X<9) = \int_{2}^{9} f(x)\,dx$



$= 7 \times \frac{1}{9}$

$= \frac{7}{9}$

(v) $E(X)$ (Expectation of X) $= \frac{a+b}{2} = \frac{1+10}{2} = 5.5$

$$Var(X) = \frac{(b-a)^2}{12} = \frac{(10-1)^2}{12} = \frac{81}{12}; \quad \sigma = \sqrt{\frac{81}{12}}$$

Question 3: If X is uniformly distributed random variable that takes values between 0 and 1. The value of $E(x^3)$ will be →

$$f(x) = \begin{cases} 1; & 0 \le x \le 1 \\ 0; & \text{otherwise} \end{cases}$$

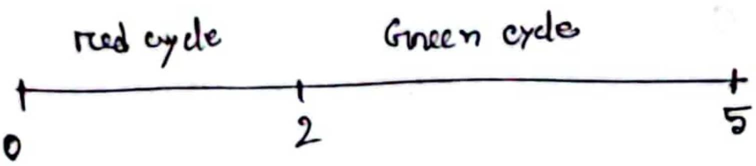
$$E(x^3) = \int_0^1 x^3 \cdot f(x)\, dx$$

$$= \int_0^1 x^3\, 1 \cdot dx$$

$$= \frac{1}{4} \left[x^4\right]_0^1$$

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

Question 4: Assume that in a traffic junction, the cycle of the traffic signal lights is 2 minutes of green (Vehicle does not stop) and 3 minutes of red (Vehicle stops). Consider that the arrival time of vehicles at the junction is uniformly distributed over 5 minute cycle. The expected waiting time (in minutes) for the vehicle at the junction is?



As the arrival of the vehicle is uniformly distributed over 5 min cycle,

$$\therefore f(x) = \begin{cases} \frac{1}{5} ; 0 \le x \le 5 \\ 0 ; \text{otherwise} \end{cases}$$

for the vehicle, waiting time. y,

$$y = \begin{cases} 0, & 0 \le x \le 2 \text{ (For red signal)} \\ 5-x, & 2 \le x \le 5 \text{ (For green signal)} \end{cases}$$

$$\text{Expected waiting time } E(Y) = \int_{-\infty}^{\infty} y . f(x)\, dx$$

$$= \int_{2}^{5} y . \frac{1}{5} dx$$

$$= \frac{1}{5} \int_{2}^{5} (5-x)\, dx$$

$$= \frac{1}{5} \left[5x - \frac{x^2}{2} \right]_2^5$$

$$= \frac{1}{5} \left[(25 - 12.5) - (10 - 2) \right]$$

$$= 0.9$$

So expected waiting minute $= 0.9$ minute.

Transformation: How to convert a similar like Normal distribution data to actual normal distribution?

There are two types of transformation for this.

They are →
1) Log transform
2) Box-cox transform.

Transformation is needed because many ML algorithms needs to have their data normally distributed.

Scikit-Learn
- Function Transformer
- Power Transformer
- Quantile Transformer

There are three ways to find if data is normally distributed or not?

1. sns.distplot → you can plot the data and see the distribution shape
2. pd.skew() → if it is 0, then data is normally distributed else skewed
3. QQ plot → It is the more reliable way to check if the data is normally distributed

Log transformer:

→ It doesn't work on negative valued data

→ It generally Transform right-skewed data to Normally distributed data.

Squared Transform:

→ It transform left-skewed data to normally distributed data.

There are other transformation techniques like reciprocal, sqrt transformation.

Power Transformer:

Box cox Transform: This transform converts any given distribution to normal distribution.

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$$

The exponent here is a variable called lambda (λ) that varies over the range of -5 to 5 and in the process of searching, we examine all values of λ, Finally, we choose the optimal value (resulting in the best approximation to a normal distribution) for your variable.

This method is only applicable to dataset where values strictly ≥ 0.

CamScanner