Agenda

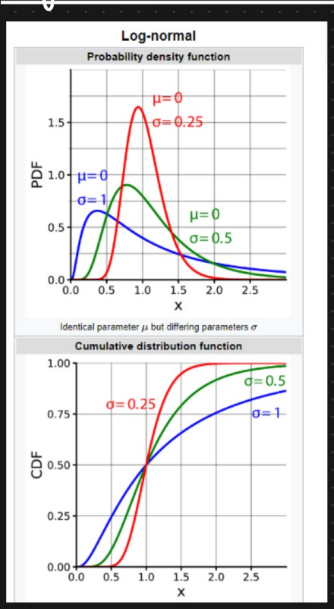
Different types of Distribution

1. Normal / Gaussian Distribution ✓
2. Standard Normal Distribution ✓
3. Log Normal Distribution ✓
4. Power law Distribution ✓
5. Bernoulli Distribution ✓
6. Binomial Distribution ✓
7. Poisson Distribution ✓
8. Uniform Distribution → Discrete, Continuous ✓
9. Exponential Distribution ✓
10. F distribution ✓
11. Chi square distribution ✓ } ANNOVA
12. Hypothesis Testing } ✓

All the distribution that we will probably study:
--> first ask whether that distribution deals with categorizing or distribution of discrete random variable (probability mass fun^) or continuous random variable (probability density function)

--> notation
-->parameters
-->pdf/pmf and cdf graph understanding(not by heart)

1. Log Normal Distribution {Continuous Random Variable}



In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, X = exp(Y), has a log-normal distribution.

Understanding above def below

$X \sim \text{lognormal}(\mu, \sigma^2)$ If X is continuous random var which is log normally distributed with some mean and variance.

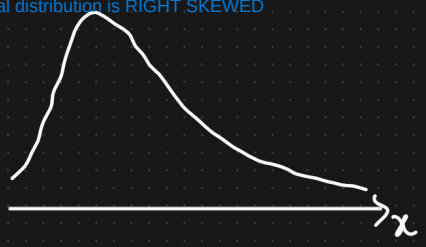
$Y \sim \ln(x) \Rightarrow \text{Normal Distribution}(\mu, \sigma^2)$.

If we take logarithmic of X then the distribution thus formed will be normally distributed represented with y random var

ln = natural log ($\log_e$)

ln is natural log represented as log base e

This is log normally distributed.
Note that log normal distribution is RIGHT SKEWED



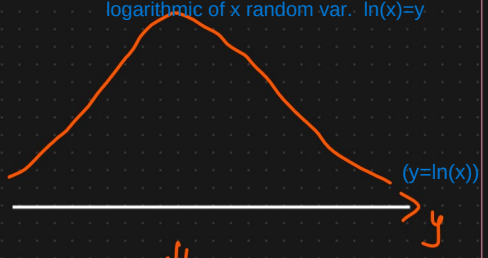
Logarithm Transformation

$\Rightarrow$ $\boxed{\ln(x)}$

$\Leftarrow$ $\boxed{X = \exp(y)}$

Reverse condition:
If you want to convert normally distributed random var y in log normal distribution then simply take exponential of y. ie; exp(y)=x

This is normally distributed formed by taking the logarithmic of x random var. ln(x)=y

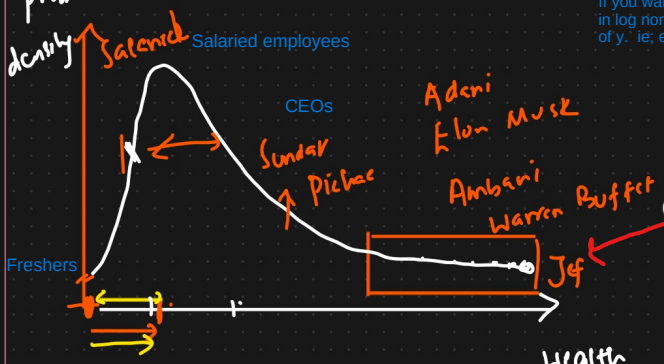


$\Downarrow$

Model will get Trained Efficient

In ML we try to convert data distribution into normal distribution because learning of ML models improves (ie; model can be trained efficiently) when data is normally distributed.

Age



DATA

Below are the examples of data following log normal distribution

① Wealth distribution of the world

② Salaried distribution in a Company

③ People writing length of comments

④ User spending time in read Articles

Wealth

$\Downarrow$

Notation: $\text{logNormal}(\mu, \sigma^2)$

Parameters: $\mu \in (-\infty, +\infty)$

$\sigma > 0$

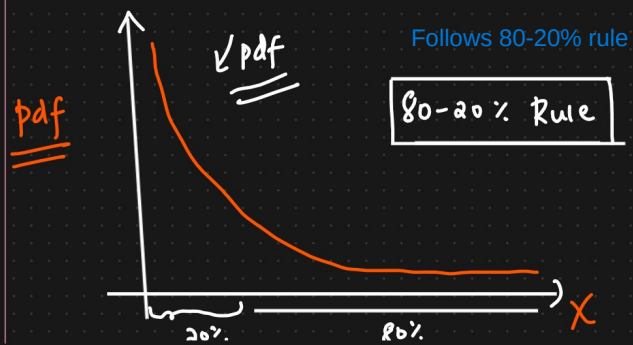
Pdf : $\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$

basically, this is the equation for the smoothen curve formed after distributing data.

(This is of y=f(x) form)

Note: To identify distribution is normal or use qq plot (will be discussed later)

④ Power Law Distribution [Continuous Random Variable]



Follows 80-20% rule

80-20% Rule

x and y(probability density)

In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another. For instance, considering the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four

Eg: IPL Games — RCB

Project

1A 1PM 1BA

5-6 s/o

① 20% of the team is responsible for winning 80% of the match

② 80% of the sales in Amazon is derived from 20% of the product. Read (Products which are hero/best selling products)

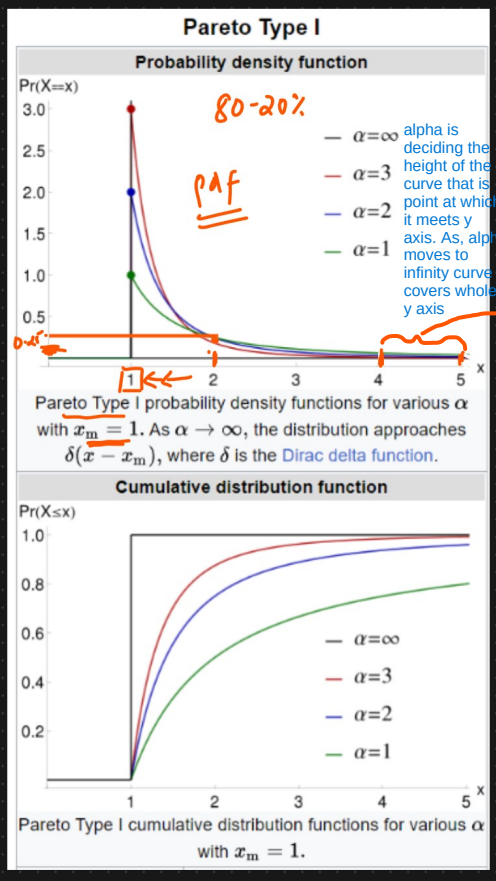
③ 80% of the wealth is distributed among 20% of the people.

④ 80% of the projection complete by 20% of team.

Types of Power Law Distribution

① Pareto Distribution ② Exponential Distribution (Only pareto is discussed for exponential you may do the wiki search and see likewise we have seen other distributions)

① Pareto Distribution [Continuous Random Variable]



If X is a random variable with a Pareto (Type I) distribution, then the probability that X is greater than some number x, i.e. the survival function (also called tail function), is given by For type 1; x suffix m is 1 but it can be different in other cases

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^\alpha & x \geq x_m, \\ 1 & x < x_m, \end{cases}$$

X = random var
x= value of random var at a pt.
x suffix m = 1

pdf =

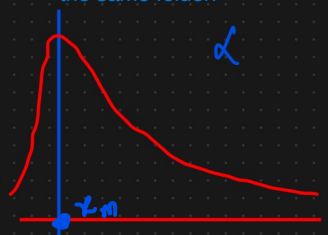
wrt pdf area within this section

$\Pr(x > 4) = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$ (here, x=4 and x suffix m=1 therefore, x>x suffix m)

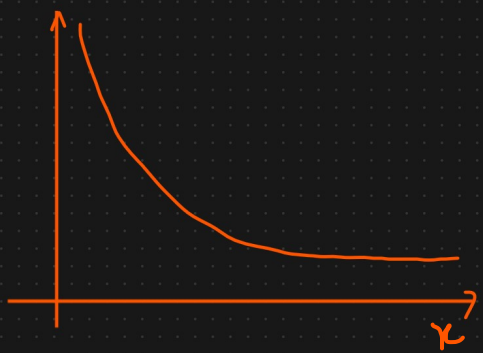
$\Pr(X = 4)$ $= \left(\frac{1}{4}\right)^2 = \frac{1}{16} = 0.25$

$x_m = 2$ $\left(\frac{1}{4}\right)^2 = 1$

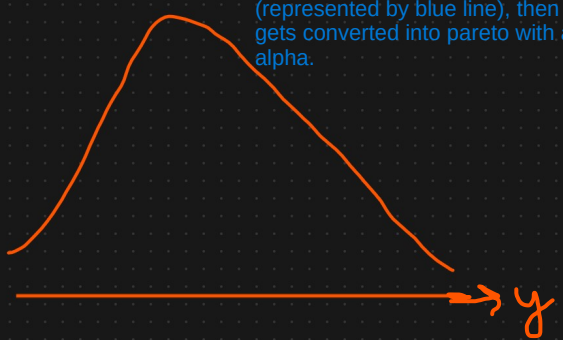
For pdf and cdf formula just see Pareto_pdf_cdf.png in the same folder.



Red curve represents log normal distribution. In this if we draw a line parallel to the y axis at some point ie; x suffix m (represented by blue line), then log normal gets converted into pareto with a parameter alpha.



Box Cox Transformation ⇒



Using Box Cox transformation we can convert Pareto distribution for random var x into normal distribution with random var y(See Box_Cox_Tranfoamtion.png for formula).

Sholay movie -->tossing a coin

⑧ Bernoulli Distribution : Outcome of the process is binary {1,0}

{Success, Failure}.

Deal with distribution of discreate random variables

Tossing a Coin (Fair)

Probability of head $Pr(H) = 0.5 \Rightarrow p$ //

Probability of tail $Pr(T) = 0.5 \Rightarrow 1-p = q$ //

$\Rightarrow p+q = 1$

$$\text{Pdf/Pmf} = p^k (1-p)^{1-k}$$

$k \{1,0\}$

$$= p(1-p)^0$$

$$= p \text{ //}$$

$$\text{pmf} = \begin{cases} p & \text{if } k=1 \\ 1-p=q. & \text{if } k=0 \end{cases}$$

For pmf and cdf see Bernouli.png in the same folder

⑨ Binomial Distribution → Combination of multiple Bernoulli Distribution

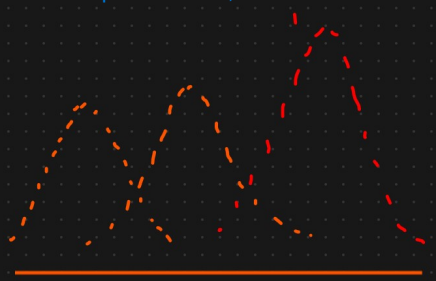
(deals with the distribution of discrete random variables)

In Bernoulli we take about single experiment. And for Binomial we combine multiple Bernoulli experiments. ie; Binomial is n times Bernoulli experiment.

n, p $\quad k$

$$\text{pmf} = {}^{n}C_{k}\, p^k (1-p)^{n-k}$$

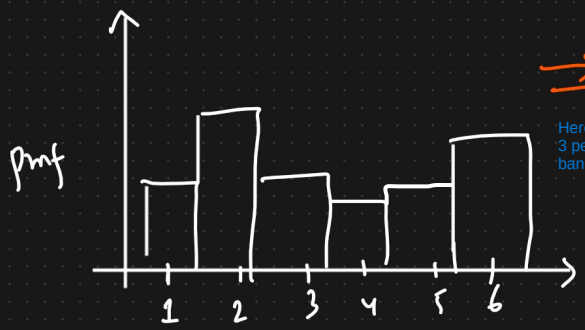
n is number of trials



In pmf curve each dot represents single trial. For more info can see in wiki.

⑩ Poisson Distribution (Deals with the distribution of discreate random variable)

Understanding with the help of an example: No. of people visiting bank every hour



$\Rightarrow \boxed{\lambda = 3} \Rightarrow$ Expected no. of people to come at that specific time interval

Here lambda = 3 means that at least 3 people are expected to visit the bank every hour

pmf $\quad Pr(X=5) = \dfrac{e^{-\lambda} \lambda^{x}}{x!}$

pdf

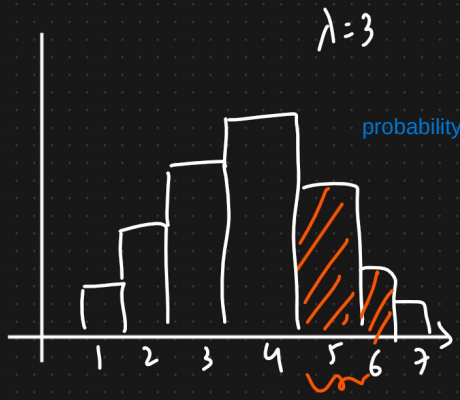
Probability density

$f(x)$

pmf

Probabilities

wrt pdf we find probability density and wrt pmf we find probability.



$$Pr(x=5) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-3}3^{5}}{5!}$$

probability of a person coming in the 5th hour

$$= 0.101 = 10.1\%$$

$$Pr(X=5 \text{ or } 6) = Pr(x=5) + Pr(x=6)$$

$$= \frac{e^{-\lambda}\lambda^{5}}{5!} + \frac{e^{-\lambda}\lambda^{6}}{6!}$$

Probability of a person coming in the 5th or 6th hour

How to identify Poisson distribution?
Ans: A specific event will occur at each instant.
For eg; No. of people visiting hospitals every hour, no. of people visiting railway stations and airports every hour, no. of people visiting every 2 hours etc. all these will follow the Poisson distribution.

④ Uniform Distribution

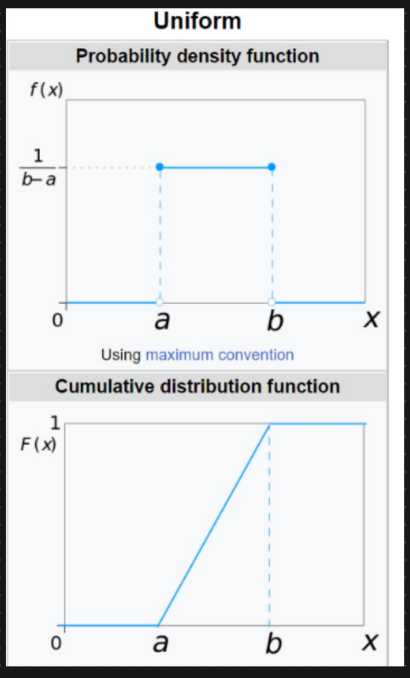
① Continuous Uniform Distribution (pdf)

② Discrete Uniform Distribution (pmf)

① Continuous Uniform Distribution { Continuous Random Variable }

Eg: The number of candies sold daily at a shop is uniformly distributed

Minimum 15 and max 30 candies sold daily [15-30] [Max, Min] ⇒ Interval



Notation: $U(a,b)$ $\boxed{b>a}$

Parameters: $-\infty < a < b < \infty$

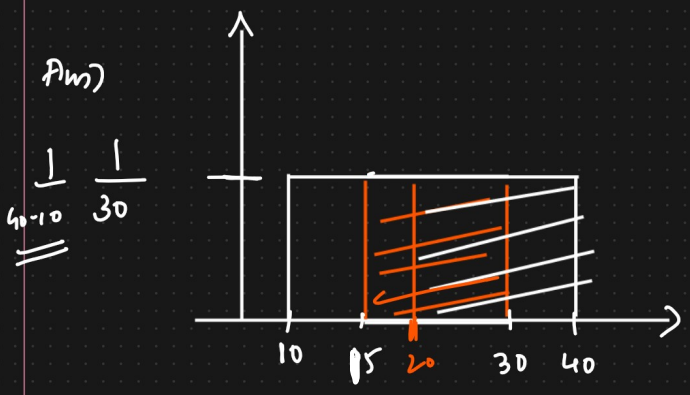
a=min interval
b=max interval

$$pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{Otherwise} \end{cases}$$

Eg: The number of Landrus sold daily at a shop is uniformly dishributed with a maximum of 40 and a minimum of 10.

(i) Probability of daily sales to fall between 15 and 30.

Ans)



$x_1 = 15$

$x_2 = 30$

$Pr(15 \leq x \leq 30)$

Area under graph will give Probability

$= (30-15) * \frac{1}{b-a}$

$= 15 \times \frac{1}{30} = \frac{1}{2} = 0.5$ //

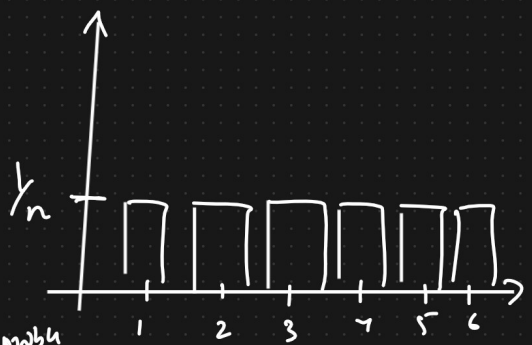
(ii) $Pr(x \geq 20) = (40-20) * \frac{1}{30}$

$= \frac{20}{30} = 66.66\%$

② Discrete Uniform Distribution {Discrete Random Variables}.

Rolling a dice = {1, 2, 3, 4, 5, 6}

$Pr(1) = 1/6$ $Pr(2) = 1/6$



$n = b - a + 1$

$n = 6 - 1 + 1 = 6$

a=min outcome
b=max outcome

Cml prob

Notation $U(a,b)$

Parameters a, b with $b > a$

PMF $1/n$

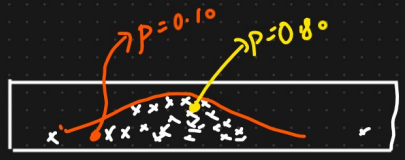
$Pr(X = 1 \text{ or } 2)$

In Bernoulli possible outcome is always 2 (binary) whereas, in Discrete Uniform distribution possible outcome can be more than 2.

⑧ Hypothesis Testing [Inferential Stats]

① p value

Out of 100 touches in their Space bar 10 times touching in that area.



Hypothesis Testing

Person → Crime

What is hypothesis?
Ans: Hypothesis is a means through which we map input with output. h(theta)=y=f(x)=theta0 +theta1*x --> hypothesis for simple linear regression.

As a first step in hypotheses testing we define the null and alternate hypothesis.

① Null Hypothesis H_0 – Person has not committed crime.

Default hypothesis is the default information available in question or problem through which one is testing.

Alternate Hypothesis H_1 – Person has committed crime

Complement or opposite of null hypothesis is Alternate hypothesis.

As a second step we do experiments based on which judgement/prediction can be made.

② Experiments : Proofs, DNA, fingerprints, evidences ⇒ Judge ⇒ Person has committed crime

As a final step we make the conclusion whether to accept or reject the null hypothesis.

③ Reject the Null Hypothesis

OR

We fail to Reject H_0.

Example:

① Coin is fair or No through 100 experiments

H_0 = Coin is fair

H_1 : Coin is not fair

Experiment :

C.I = 95%

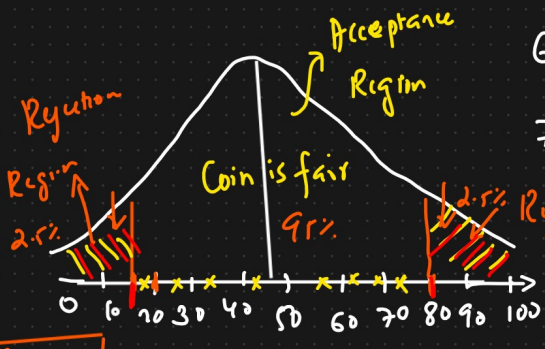
$\alpha = 1 - C.I = 0.05$

Significance level

50 Head 50 Tails

60 Head 40 Tails

70 Head 30 Tail



This graph is plotted based on number of heads

PValue < 0.05

probability value for the Null hypothesis To be True

Confidence Interval

In Experiment phase lets say that based on number of heads we are deciding whether coin is fair or not. Say we plotted our 100 tossing and plotted the number of heads obtained(represented in above graph). There will be a domain expert who will tell us that in 100 experiments what should be the range of number of heads within which coin will act fair and below/beyond that range it will become an unfair coin. Through this we are basically asking domain expert for confidence interval(C.I).

- C.I represents the region/range where hypothesis is true. It is acceptance area.
-Significance level = 1-C.I represents the rejection region where hypothesis is not true.
-Significance level is represented by alpha.

Assignment : Exploring Distribution

F distribution:
-Used to compare variance for 2 groups.
-Check pdf graph in wiki. d1 and d2 are degree of freedom. With initial value of d1 and d2 curve seems to be pareto/exponential. As we increase these parameters curve start starts appearing like log normal distribution. On further increasing these parameters curve becomes normally distributed.
(This was interview question just pfd of f distribution was shown and insights one get by seeing was asked. Tip start from lower parameter value and start observing the trend in the curve as we increase parameter value to maximum.)

Chi Square distribution:
-Similar to log normal. When k increases (degree of freedom) it becomes log nomal (right skewed) distribution.
-Exponential, pareto and log normal distribution can be chi square distribution(check pdf curve in wiki tha by changing parameter value k all mentioned distributions can be obtained in chi square distribution).
- If asked for example then simply give examples of pareto,log normal as chi square is similar to it based on parameter (degree of freedom)