Distributions in probability

We are doing some experiment and collecting data. Now we want to see what is the distribution of porbability of getting given observation. Distribution gives you theoratical model to define chance of observation value or range of value for given experiment(collection of data).

1. Bernauli distribution : Bernauli(p)

Simplest distribution of experiment whose outcomes are binary. Either 0 or 1.
Gender : Male(1) or Female(0)
coin toss : Head(1) or Tail(0)

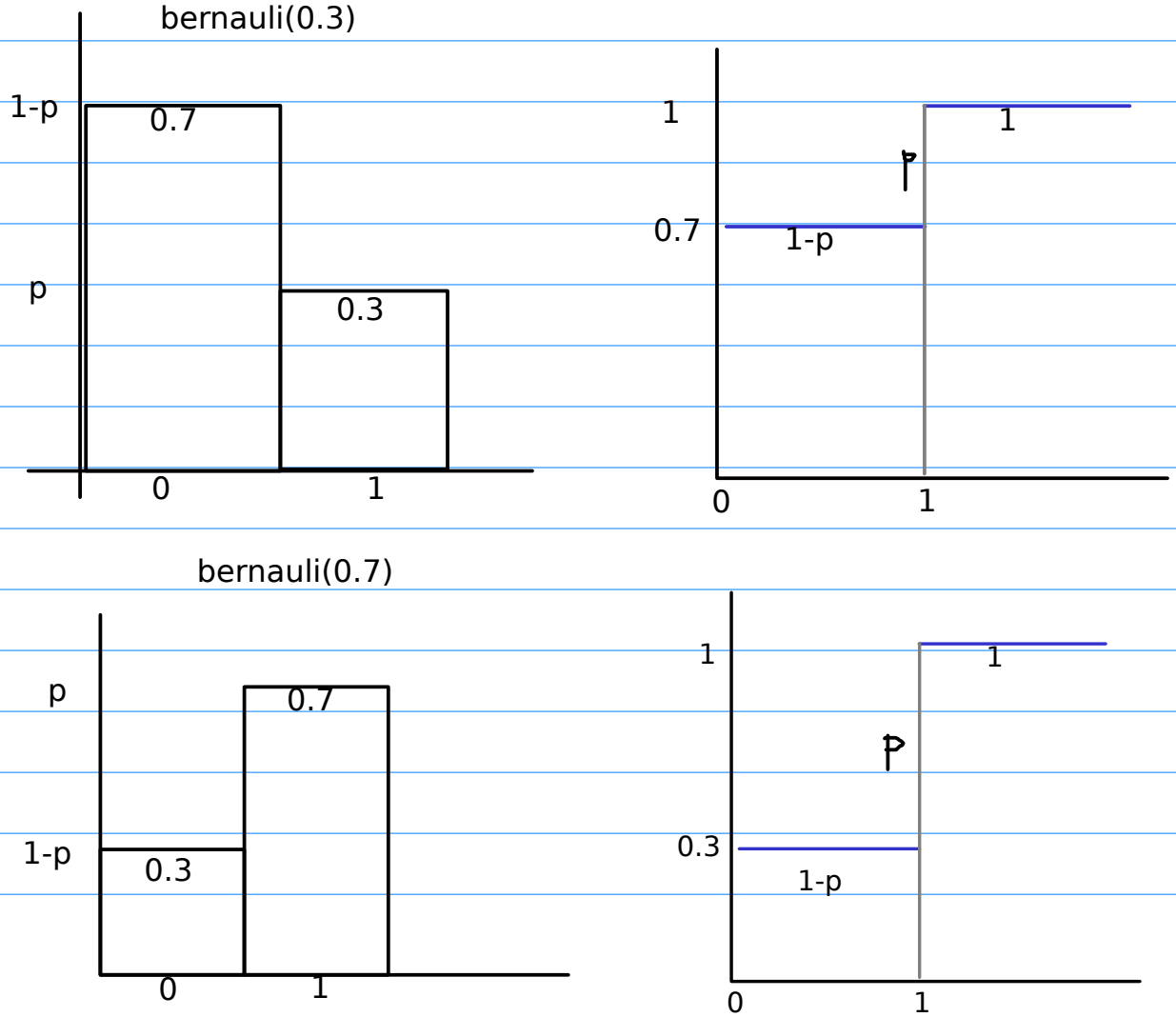
Parameter p is probability of getting value 1 or true probability.
False probability (probability of getting value 0) q = 1 - p

PMF = $$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

E(x) = mean = p

Var(x) = pq = p(1-p)

PMF **CDF**



2. Binomial Distribution: Binomial(n, p)

Probability of getting value 1 for k times out of n times is distributed as binomial.

p = 0.5 probability of getting head
What is the probability of getting 7 heads(1) if you toss a coin 10 times.

p = 0.1 probability of getting faulty bulb
What is the probability of getting 2 faulty bulbs(1) in box of 20.

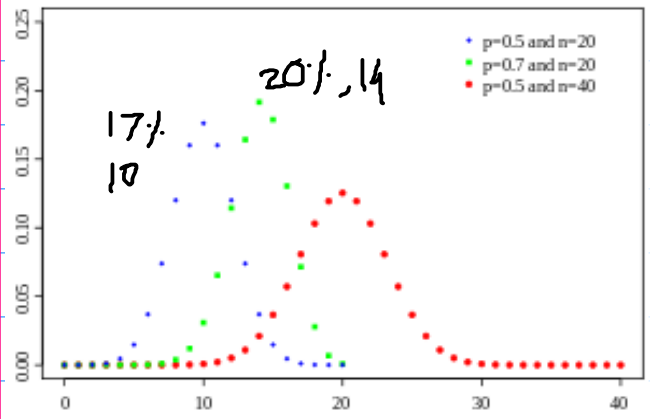
PMF $$P(X = k) = P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

CDF $$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

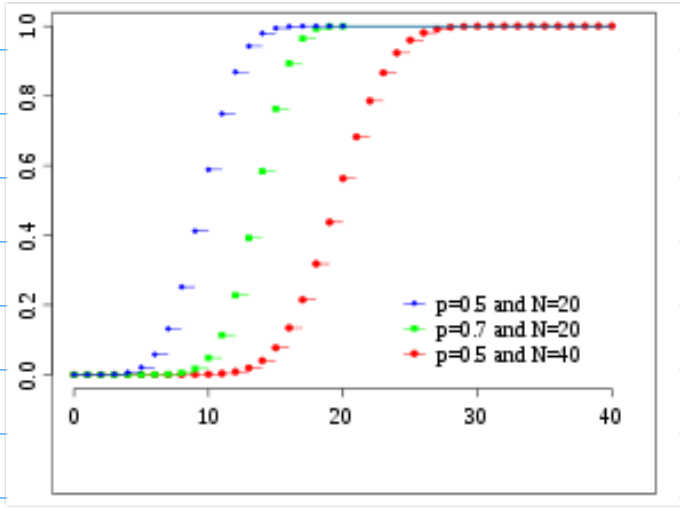
E(x) = mean = np

Var(x) = npq = np(1-p)

PMF



CDF



for n = 20 we have two values of p 0.5 and 0.7

blue — When p = 0.5, it is high probability to have 10 points to be equal to 1

Green — When p = 0.7, it is high probability to have 14 points to be equal to 1

3. Uniform (Continuous) Distribution: uniform(a, b)

Probability of getting any value is similar across range a to b.

uniform colour, weight of cricket ball, random value generator

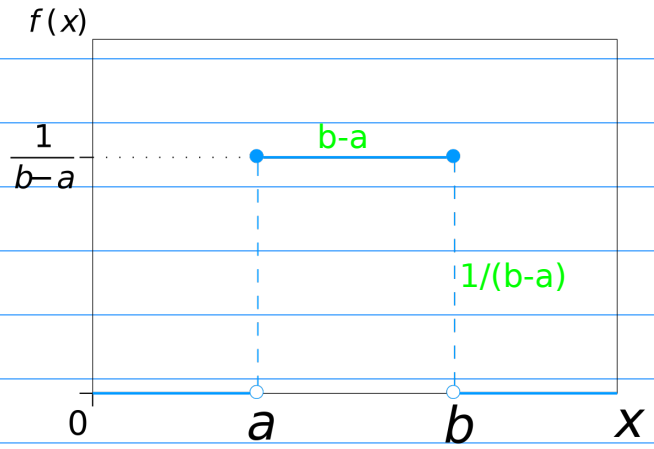
PDF =
$$\begin{cases} = \frac{1}{b-a} & \text{if } x \in [a,b] \\ = 0 & otherwise \end{cases}$$

CDF =
$$\begin{cases} = 0 & \text{if } x < a \\ = \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ = 1 & \text{if } x > b \end{cases}$$

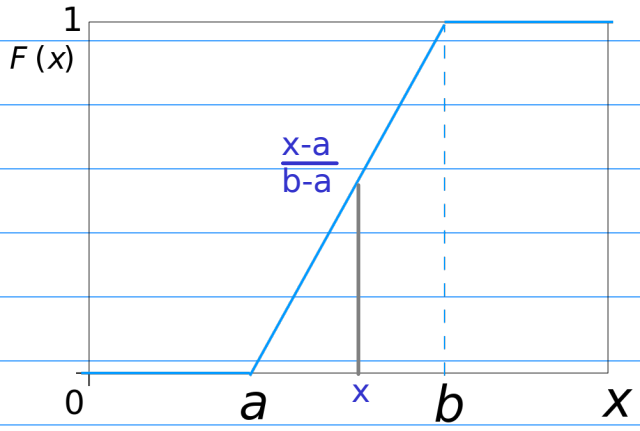
E(x) = mean = $(a+b)/2$

Var(x) = $(b-a)^2/12$

PDF



CDF



4. Poisson Distribution: poisson(λ)

Some event is occurring **λ** times per given time interval (sec, min, hr, day,...).
What is the probability that it will occur x times in given time.

Let's say **λ**=10, means we are getting average 10 calls per hour. Now we want to calculate what is the probability that you will get 10, 9, 11, 15, 6 calls in hour?

This probability distribution for given **λ** is called poisson distribution.

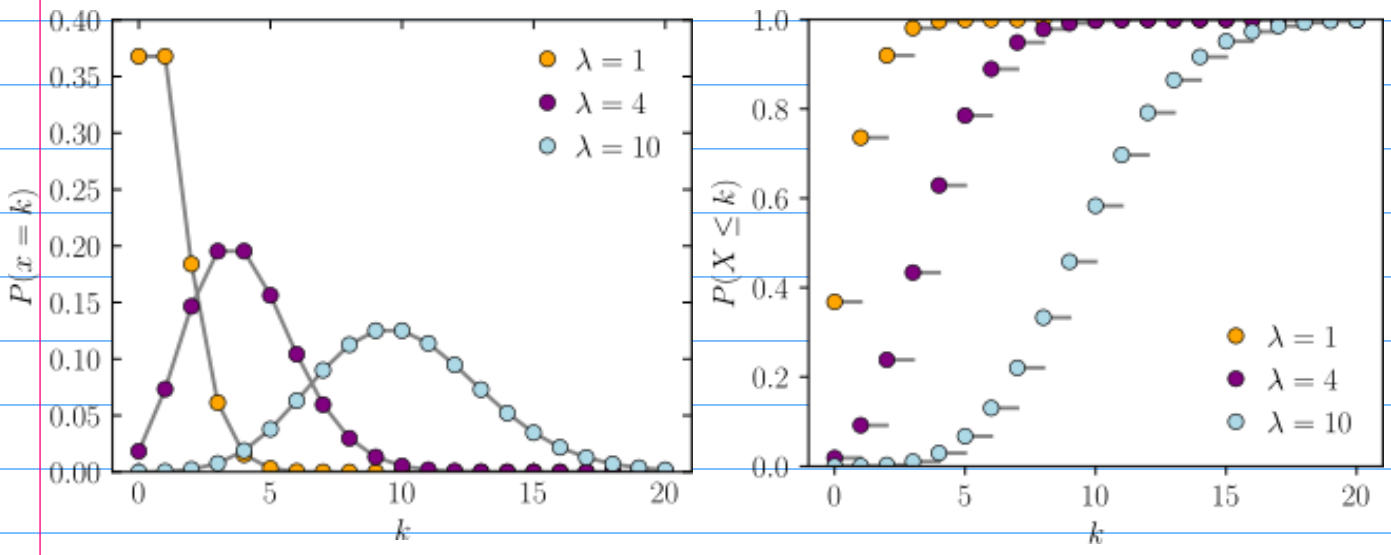
PMF = P(X=k) = f(k) = $$\frac{\lambda^k e^{-\lambda}}{k!}$$

CDF = P(X≤k) = F(k) = $$e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$$

E(x) = mean = **λ**

Var(x) = **λ**

PMF **CDF**



If X1 and X2 are two independent random variable with poisson distribution with rates λ1 and λ2 then we can say, (X1+X2) is poisson distribution of rate (λ1+λ2)

5. Exponential Distribution: exponential(λ), β=1/λ

Some event is occurring λ times per given time interval (sec, min, hr, day,...).
What is the probability that event will occur after t time.

Let's say λ=10, means we are getting average 10 calls per hour. Now we want to calculate what is the probability that you will get call in 10 min, 30 min, 1hr, 1.5hr?

This probability distribution for given β=1/λ is called exponential distribution which is inverse of poisson distribution. Or it is probability of time distribution between poisson events (with rate λ).

PMF = $$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

CDF = $$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

E(x) = mean = $\beta = 1/\lambda$

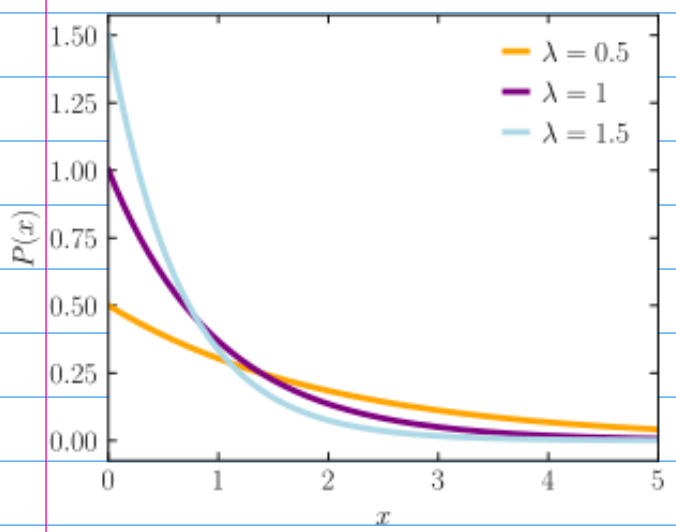
Var(x) = $\beta^2 = 1/\lambda^2$

Exponential distribution is memory less distribution. Probability that event will occur in at least t time is same as it occur after waiting for s time.

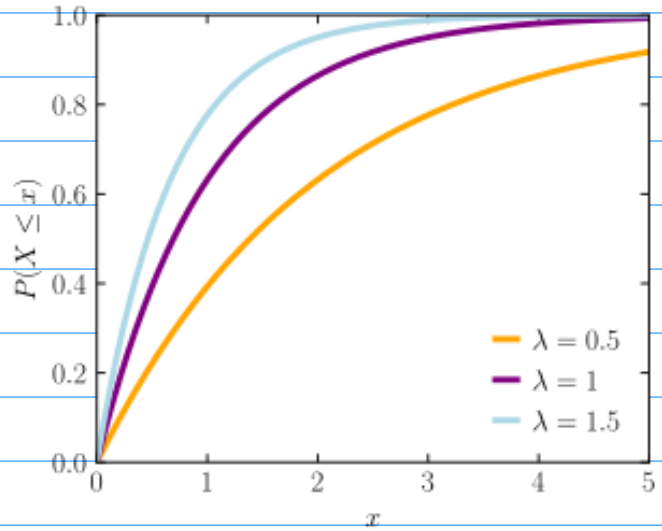
Pr(X > s+t | X > s) = Pr(X > t)

Pr(event occur in 4-0 min| waited for 30 min) = Pr(event occur in 10 min)

PDF



CDF



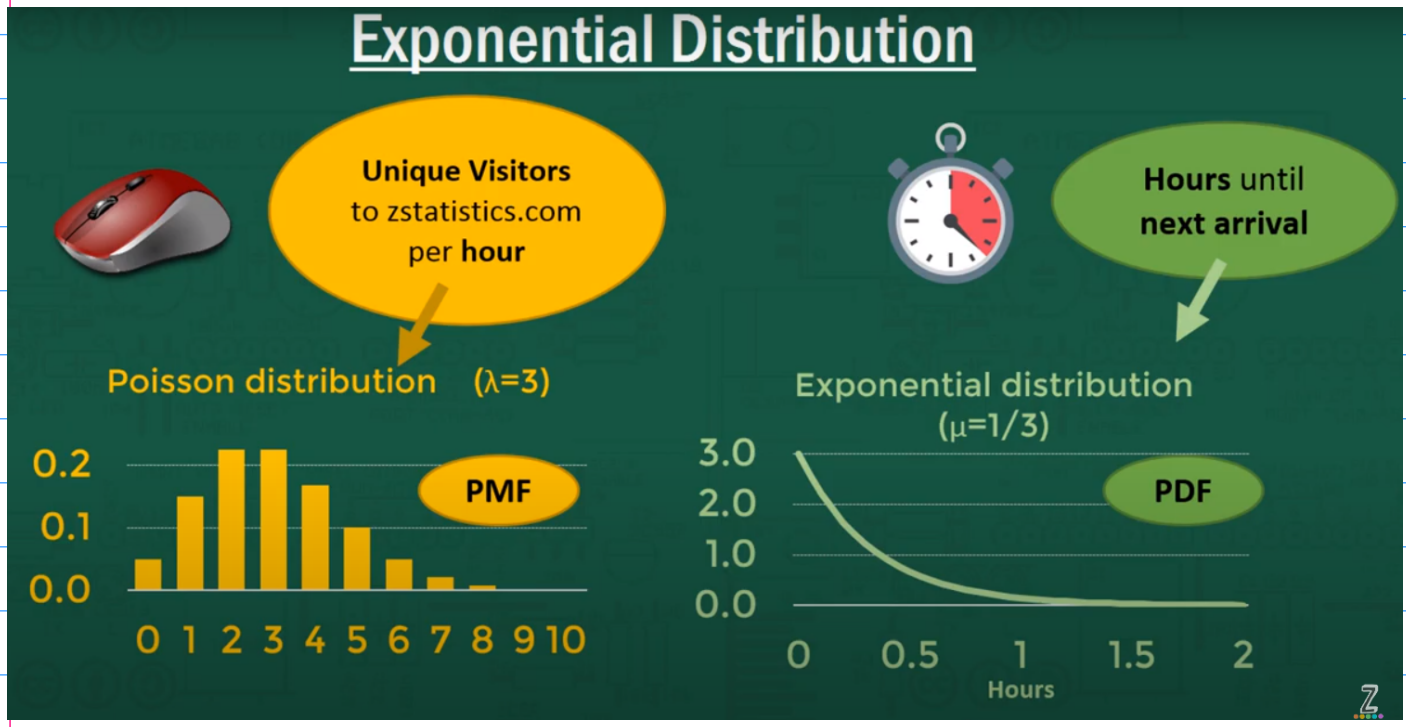
Let's take an example of unique visitors on website. Let's say number of unique visitors on webpage is posisson distributed with λ = 3 /hour. So you get average 3 unique visitors per hour or β = 1/3 hours per visitor. That means on an average you get visitor at 20 minutes interval.

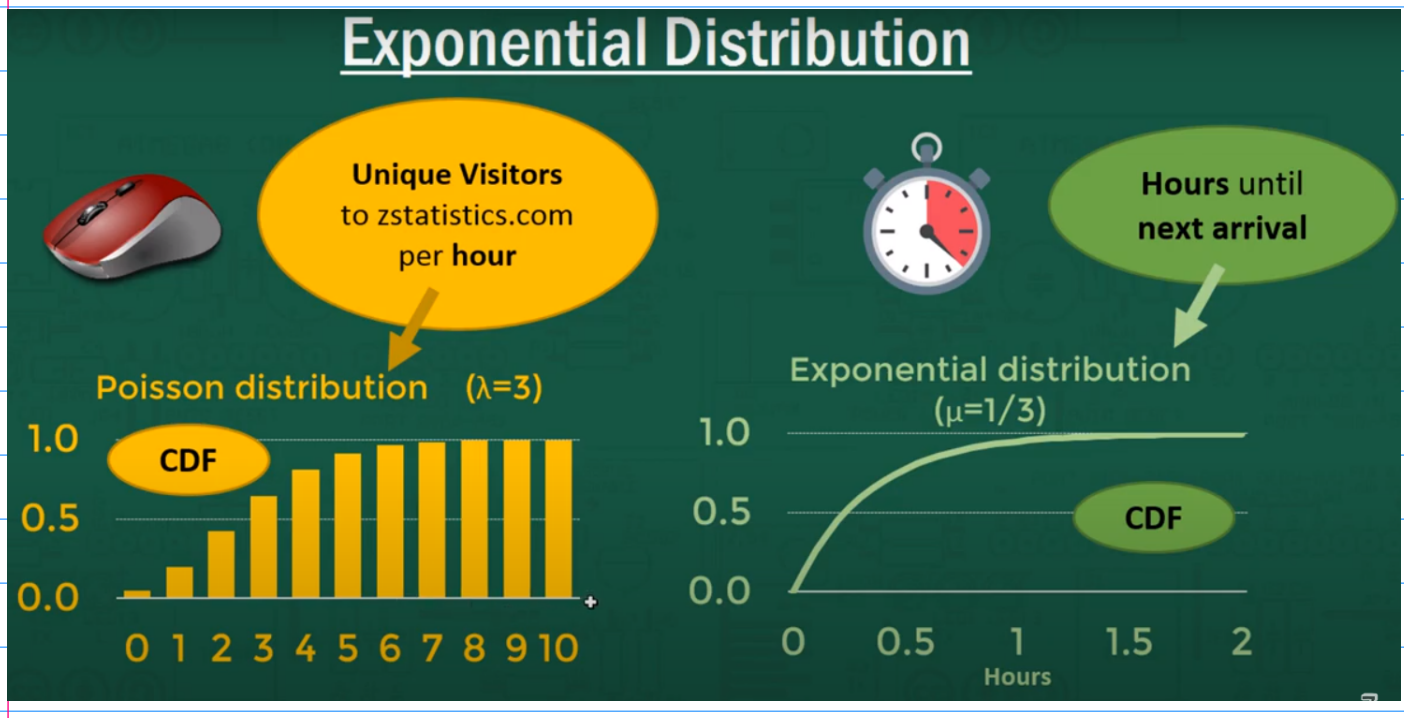
Note= In below pictures we have taken β as $\mu = 1/\lambda$

PDF: We can observe that it is really unlikey to get visitor at 2 hours interval.

CDF: It is really likey (~100%) that we get visitor in 2 hours.

It is more likely to be interval of 0 to 30 min with $\beta = \mu$ = 20min





From - https://www.youtube.com/watch?v=2kg1O0j1J9c&list=PLTNMv857s9WVzutwxaMb0YZKW7hoveGLS&index=5
http://www.zstatistics.com/

6. Normal/Gaussian Distribution: N(μ, σ^2)

There are so many natural phenomenon that follow normal distribution, which is also known as Gaussian distribution. Normal distribution shows that large part of population tends to have value near to mean value **μ** and as you move away from the mean value probability start decreasing according to variance or σ^2 standard deviation **σ**. It forms bell curve of probability.

Height weight and length of living things follows Normal Distribution.

When we draw large number of samples from population of unknown distribution if follows Normal distribution.

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

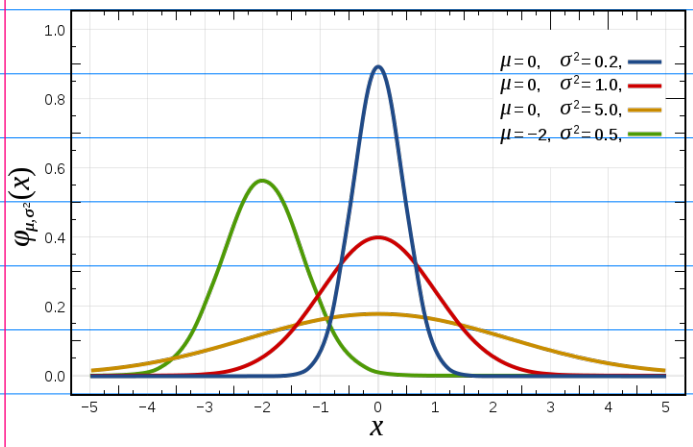
CDF=F(x) = $$P(X \leq x) = \int_0^x f(x)dx$$

E(x) = mean = **μ**

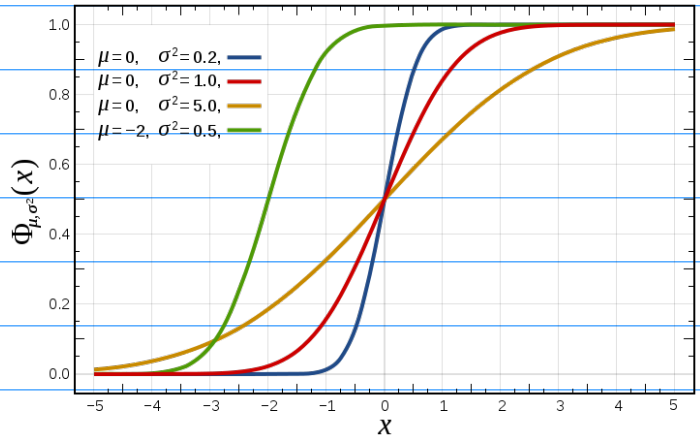
Var(x) = σ^2

standard deviation = **σ**

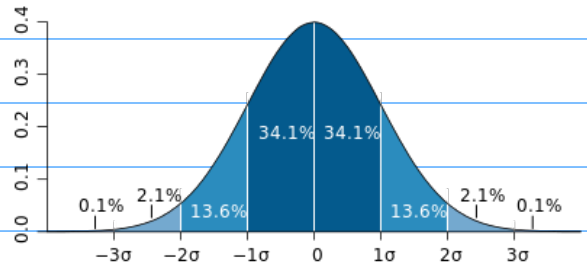
PDF



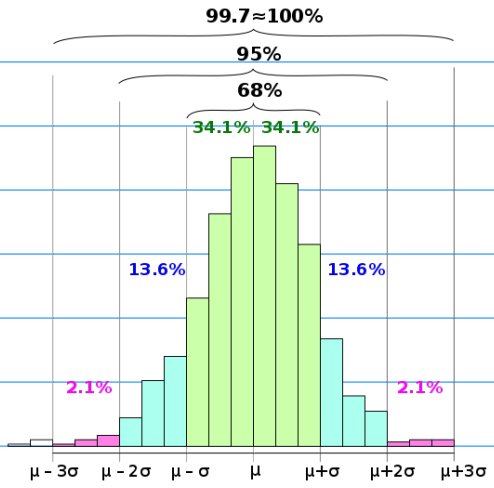
CDF



Standard Deviation and Coverage: 68-95-99.7 rule



$$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$$
$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$$
$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$$



7. Log Normal Distribution: lognormal(μ, σ^2)

There are so many phenomenon whose logarithm follow normal distribution, which is known as Log Normal Distribution. That means if random variable X is log normally distributed than ln(X) follows normal/gaussian distribution.

If something is more likely to happen in short interval and then probability of that happening decrease slowly over time is generally Log normally distributed.

Length of Comments posted in internet discussion forums

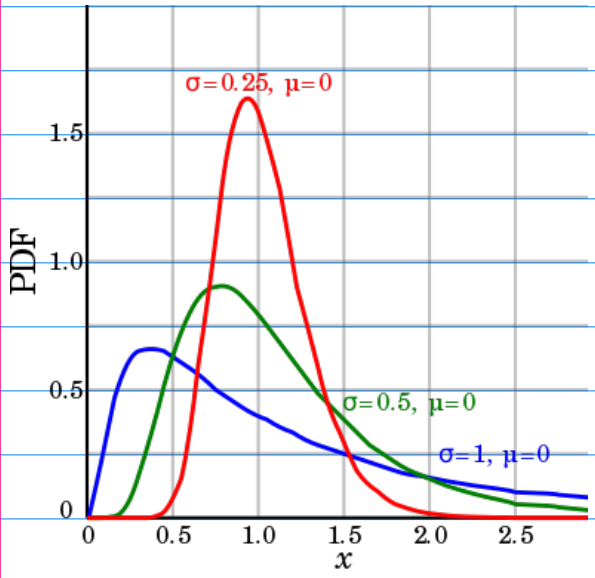
User's dwell time on online articles. Length of chess games.

Income of 97-99% of the population. (top 2-3% follows pareto distibution)

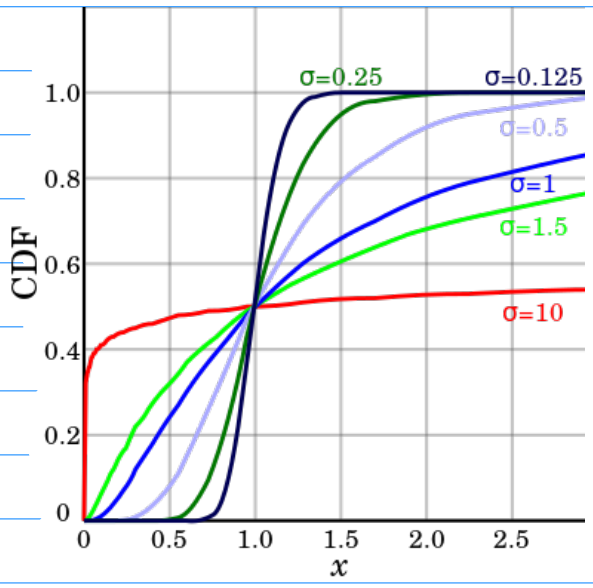
PDF=f(x) = $$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2}$$

CDF=F(x) = $$P(X \le x) = \int_0^x f(x)dx$$

PDF



CDF



Practically if we find any distribution which is following log normal distribution

we take log of all data points so that new RV follows normal distribution.

If we find RV X is following

log-normal distribution

Apply natural log, Y = ln(x)
So Y follows normal/gaussian distribution.

$$X = \{x_1, x_2, x_3, \ldots, x_n\}$$
$$Y = \{y_1, y_2, y_3, \ldots, y_n\}$$

pareto distribution

Apply box cox transformation.
λ *= box-cox(X). Now use this* ***λ****,*

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

now Y follows normal distribution

8. Pareto distribution : $pareto(x_m, \alpha)$ *infinite distribution*

There are so many phenomenon whose logarithm follow normal distribution, which is known as Log Normal Distribution. That means if random variable X is log normally distributed than ln(X) follows normal/gaussian distribution.

If most of the part of RV (80%) is distributed in short interval and then rest 20% part's probability of happening decrease slowly over range is pareto distributed.

The sizes of human settlements (few cities, many villages)

Oil reserves in oil fields (a few large fields, many small fields)

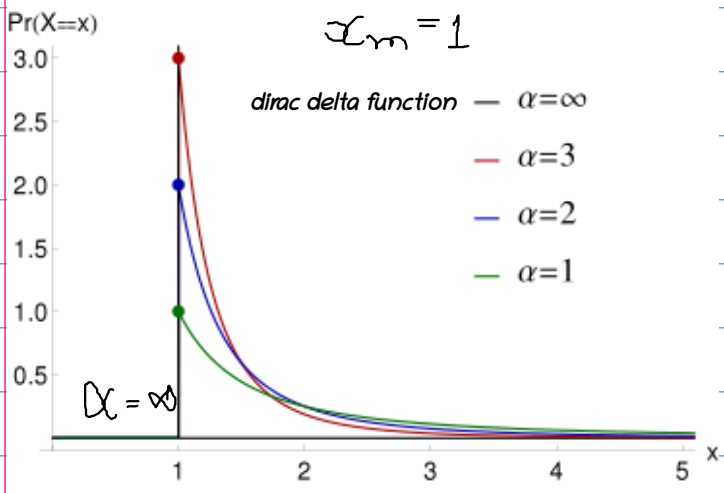
Size of meteorites.

parameters

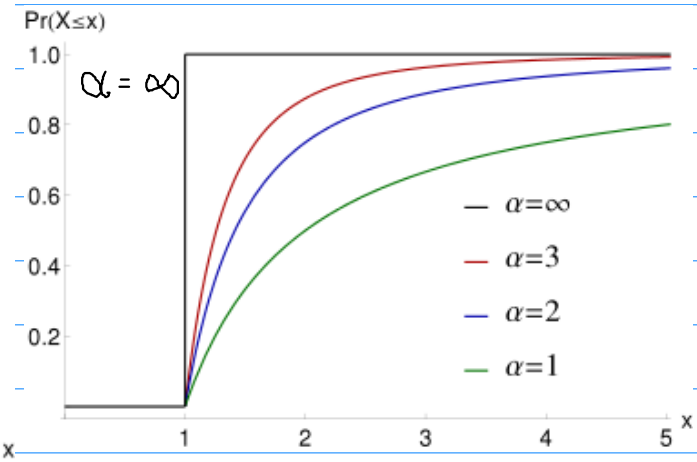
x_m >0, scale parameter - which decides the scale of distribution

α >0, shape parameter - which decided the tail of distribution (tail index)

PDF



CDF



Q: *How to check if RV is following pareto distribution? What to do if it is Pareto?*

A: *Plot log of probability of number p(x) versus log of that number, y axis= log(prob(x)) and x axis= log(x). If it happens to be straight line than we can say that it is Pareto distribution(power law distribution).*

We can also plot pareto-QQ plot also to check if RV is following pareto dist.

limitation: Pareto QQ will not work good for very small values of **α** is very small.

Log-log plot is necessary but insufficient evidence for a power law relationship, as many non power-law distributions will appear as straight lines.

If you find the distribution is pareto you can apply box-cox transformation to convert it into normal/gaussian distribution.