Probability distributions

What and why :

PD is probability vs the value, the y axis represents PDF or PMF, PDF for continuous RV and PMF for discrete RV. Note that PDF and probability are different. An experiment might follow one of these distributions might not. It is not necessary to follow these distributions. The variable may have its own distribution. Consider the following example,

You take survey of 500 people in your town. Lets say 20% of them are teachers, 15% are lawyers and so on. So this is the distribution of your population. It may follow given distributions may not. These distribution are standard distributions which are observed in normal life. There are mainly four distributions,

1. Gaussian
2. Bernoulli
3. Poison
4. Binomial

PD is very useful in analysing the data, by PD we can get insights of data and decide which algorithm / technique to use.

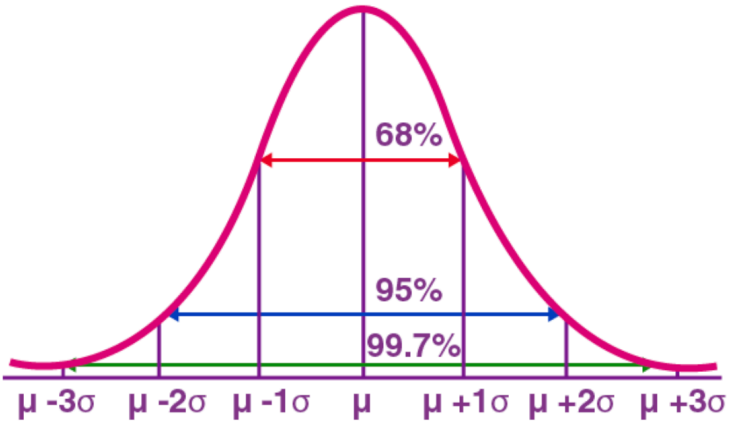
Gaussian Distribution :

Gaussian Distribution also known as normal distribution or bell shaped curve, is one of the widely used distribution. The PDF is given by,

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

Where the x is random variable.

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. The standard deviations are used to subdivide the area under the normal curve. Each subdivided section defines the percentage of data, which falls into the specific region of a graph.



From the graph we can say that, 68% of data falls within range of mean - std to mean + std, 95% of data falls within range of mean - 2*std to mean + 2*std. that's why it is called 68-95-97.5 rule.

Properties

1. Mean = Median = Mod

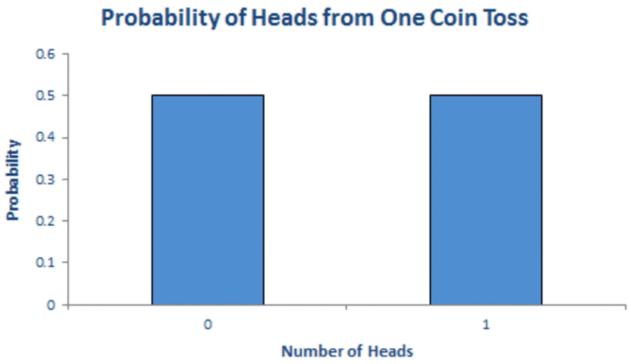
By Karan Hadiyal

2. Symmetric about centre
3. 50% data on left & 50% data on right
4. Area under the curve is 1
5. Never touches X axis
6. Has only one peak

Application :
1. Min-Max feature scaling on attributes following GD
2. Add Gaussian noise to image in image processing
3. Naive bayes assumes GD on numerical feature

Bernoulli Distribution :

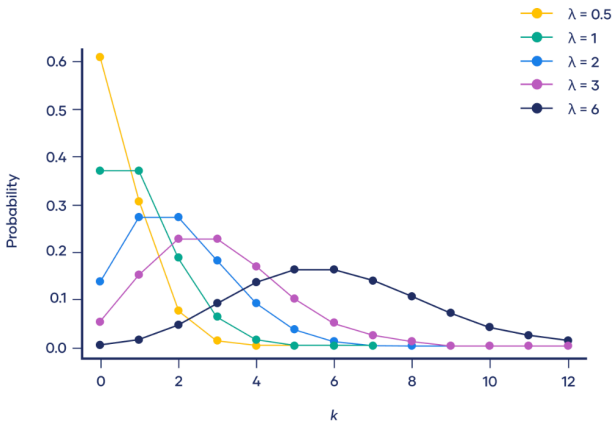
BD is used when we have discrete random variable, and we have two categories, success and failure. Lets take our coin example where head represents success
and tail represents failure. So the probability distribution of success and failure is 0.5 and 0.5. Where as in spam detection the probability distribution of success and failure can be 0.2 and 0.8. (0.2 means spam). Bernoulli trial is the event that with one trial it gives two outcomes, BD works on Bernoulli trails only. Consider the graph. Note that probability of success and failure need not to be same.



There are three conditions for BD
1. Each outcome has two results “success” or “failure”
2. At each trial the probability of outcome is same, like tossing a coin first time the probability of getting head is 0.5 and tossing a coin 100th time the probability of getting head is 0.5
3. Each trial must be independent of each other

Poison Distribution :

PD is used when we have discrete random variable, when we want to find the probability of events occurring in particular time interval. It gives the probability of an event happening a certain number of times (*k*) within a given interval of time or space. The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events or rate of event occurrence in time interval. Consider the graph of Poison distribution.



PD can be applied if,
1. We know the rate of event occurrences in a time interval
2. Events happen randomly and independently, i.e., occurrence of one event doesn’t affect another

The most probable number of events is represented by the peak of the distribution - the mode.

- When λ is a non-integer, the mode is the closest integer smaller than λ.
- When λ is an integer, there are two modes: λ and λ−1.

When λ is low, the distribution is much longer on the right side of its peak than its left (i.e., it is strongly right-skewed). As λ increases, the distribution looks more and more similar to a normal distribution. In fact, when λ is 10 or greater, a normal distribution is a good approximation of the Poisson distribution. The mean and variance for poison distribution is λ.

The PDF is given by,

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

P(X = k) represents the probability of the random variable X taking the value k.
e is the base of the natural logarithm (approximately 2.71828).
λ is the average rate parameter (λ) of the Poisson distribution.

Consider the example, we have customers, data when from 10 AM to 5 PM 100 customers visit. To total events is 100, now we want to find customer between 10 AM and 12 PM. So the lambda, is 100 * (2/7). Lambda is the rate of customers arrive in time interval length, so in 7 hours 100 customers arrive so in 2 hours approximately, 28.56 customers will arrive. Now the pdf for k in this time interval will give the probability of arriving k customers, where k varies from 1 to 100 or any. Like probability of 1 customer arriving in given time interval, 2 customers arriving and so on.

The applications of poison distribution are,
1. Text messages per hour
2. Machine malfunctions per hour
3. Website visitors per month

Binomial Distribution :

BD is a discrete probability distribution, it gives the probability of success in number of trials. BD can be applied in Bernoulli trials, in BT each experiment has two outcomes, success or failure, and each experiment is independent of each other.

Binomial distribution gives probability of k number of success in n number of trials, by the probability of success in one trial. i.e., in a coin toss lets say head is success, so probability of occurring head is 0.5 in one trial. So p = 0.5, now if we want to calculate probability of 5 success in 10 trials (probability of occurring head 5 times in 10 tosses) then binomial distribution is there. The PMF is,

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

P(x=k) means probability of success occurring k times, n is the number of experiments, and p is the probability of one event.

The Bernoulli distribution is the special case of binomial distribution when n = 1. The mean of BD is np, the variance is np(1-p). When p = 0.5 then the distribution is symmetric, when p < 0.5 it is right skewed and when p > 0.5 it is left skewed. P < 0.5 means in n trials the probability of success is very lees, hence most of the data points are concentrated at **right side**. That means the distribution is stretched or skewed towards left.

Left skewed - winning in casino - in 100 trials we win for 4-5 times
Right skewed - goal miss by Ronaldo - in 100 trials he misses for 4-5 times