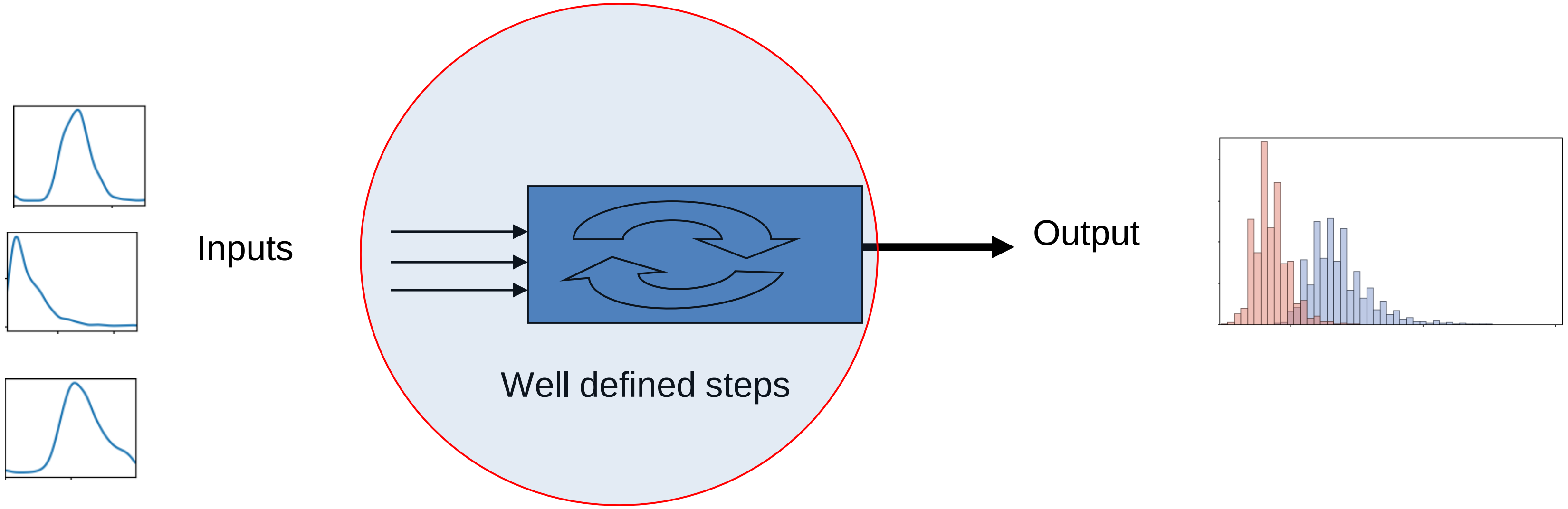
Probability, Probability Distribution And Modeling

Well defined stable process



Distribution of the values of inputs and output of all well defined processes are characterized by a distribution shape with central values (a.k.a expected values) and standard deviation (spread of the values around the central values)

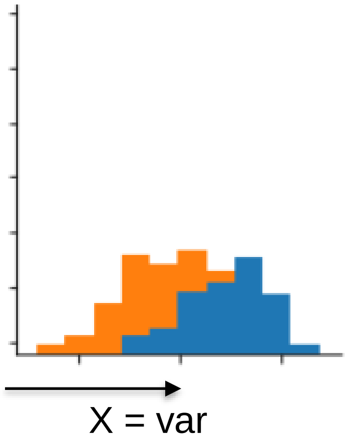
The ML algorithm will take into account the distribution on inputs and output and represents their relationship as a model

Probability and distributions

1. Objective of modeling is to understand current behavior of a process and predict future behavior
2. Future behavior predictions can never be 100% accurate. The only way to make 100% accurate prediction is "Back to the future" ☺
3. Since we cannot be 100% sure, we associate predictions with probabilities. Probability and probability distributions play important role in modeling
4. What is probability and why do we need it? Probability help us quantify uncertainty and provides us a rational way to decide in uncertain situations
5. When we use probability, we have an event of interest in mind. The event may be "currency note being fake" or "mpg of car being 15 or above" etc.
6. Probability in it's basic form is a ratio based on past experience. How many times the event of interest under similar conditions occurred in the past. For e.g.
   a. given the characteristics of the car, how many cars in the past gave us at least 15 miles per gallon
   b. given the characteristics of a currency note, how many notes in the past with similar characteristics were fake
7. Same probability concept is used both while building the model and also to express our confidence in model's ability to predict accurately

Probability and distributions

1. Consider the attribute X = “var” in the “banknote_authentication” data set. The orange distribution is for fake note and blue one for real
2. As the value of X increases from 0, the frequency of fake notes increases and peaks at certain value before falling down
3. Similarly, with increase in X the frequency of real note increases but from a different minimum value and peaks and falls at different values
4. If we separate the two distributions, what we get is a frequency distribution on X attribute for fake and real notes from past data
5. The different position of the peaks, the difference in spread i.e. range of values for the two classes help us build a model that distinguishes fake from real
6. The farther these distributions are i.e. lesser the overlap, more able will the model be in detecting fake from real notes
7. The ML algorithms form a probability function for X. For e.g. Given a fake currency note, what different values can X obtain with what probability
8. X is called the random variable which could be continuous or discrete
9. To get the probabilities, the algorithms need distributions in X for the given class and their similarity with known statistical distributions such as Normal, Binomial etc.
10. Knowledge of statistical distributions help selecting appropriate parametric ML algorithm. For e.g. Poisson regression if the values of target column show similarity with Poisson distribution



Probability and distributions

Binomial distribution–

1. There are many types of probability distributions in statistics. They are statistical models that describes the likelihood with which a random variable (X) can acquire different possible values.
2. A physical process can be mapped to one of these statistical models provided it’s output has distribution similar to the distribution
3. Three most commonly used distribution statistical models are discussed here -

4. Binomial distribution (Discrete Probability Distribution)–
   1. Prerequisites -
      - i. Number of observations n is fixed
      - ii. Each observation / trial is independent of other
      - iii. Each observation leads to one of two possible outcomes
      - iv. Probability of ‘success’ p is the same each trial

   2. The model allows us to compute $P(X\,\text{"successes"}) = \frac{n!}{x!\,(n-x)!}\,p^{x}(1-p)^{(n-x)}$ of "successes" in the n trials

   3. Given probability of sever going down on any day = .25 (long term average), what is the probability of 6 servers going down in a batch of 11 on the same day… Assume servers are independent of each other
      - i. F(6) = C(11,6) * (.25)^6 * (1 – .25)^(11-6)   = 0.0267 (truncated)

   4. Mean of the distribution is = μ_x = np and variance = σ^2 = n(p)(1-p). Mean is larger than the variance

   5. The distribution tends to normal as n tends to infinity

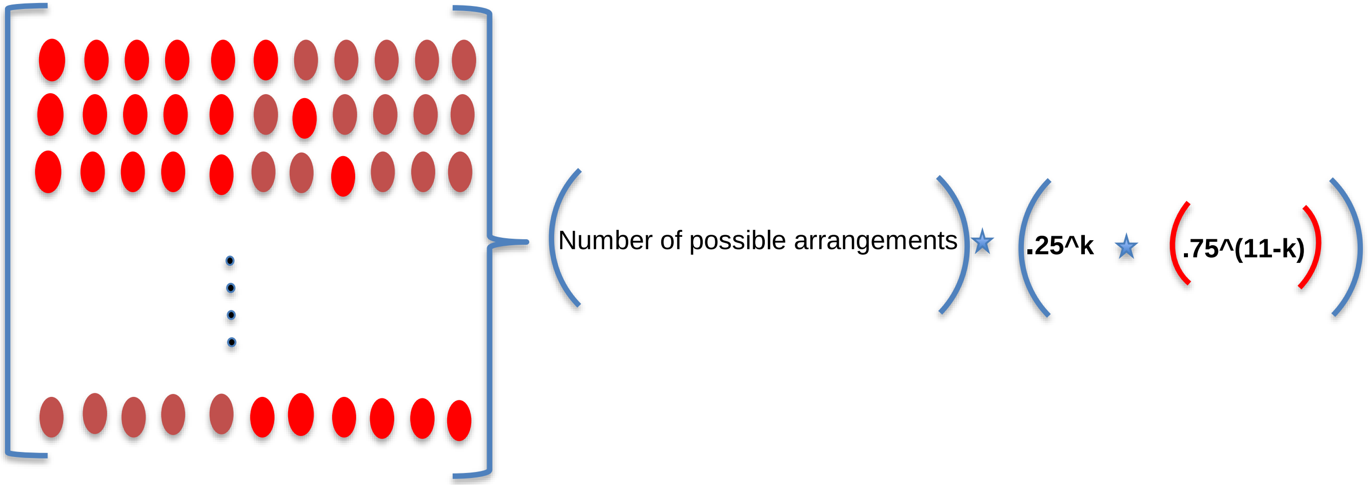
Probability and distributions

Binomial distribution (Contd) –

What is the probability of six servers going down out of a rack of 11 given long term probability of .25 that a server will go down on a given day. Servers down events are independent.

1. The probability of a server down = .25 on any day
2. Probability of k(red) number of server down on same day = no of arrangements X (.25^ no of reds) X ((1-.25) ^ no of greens)

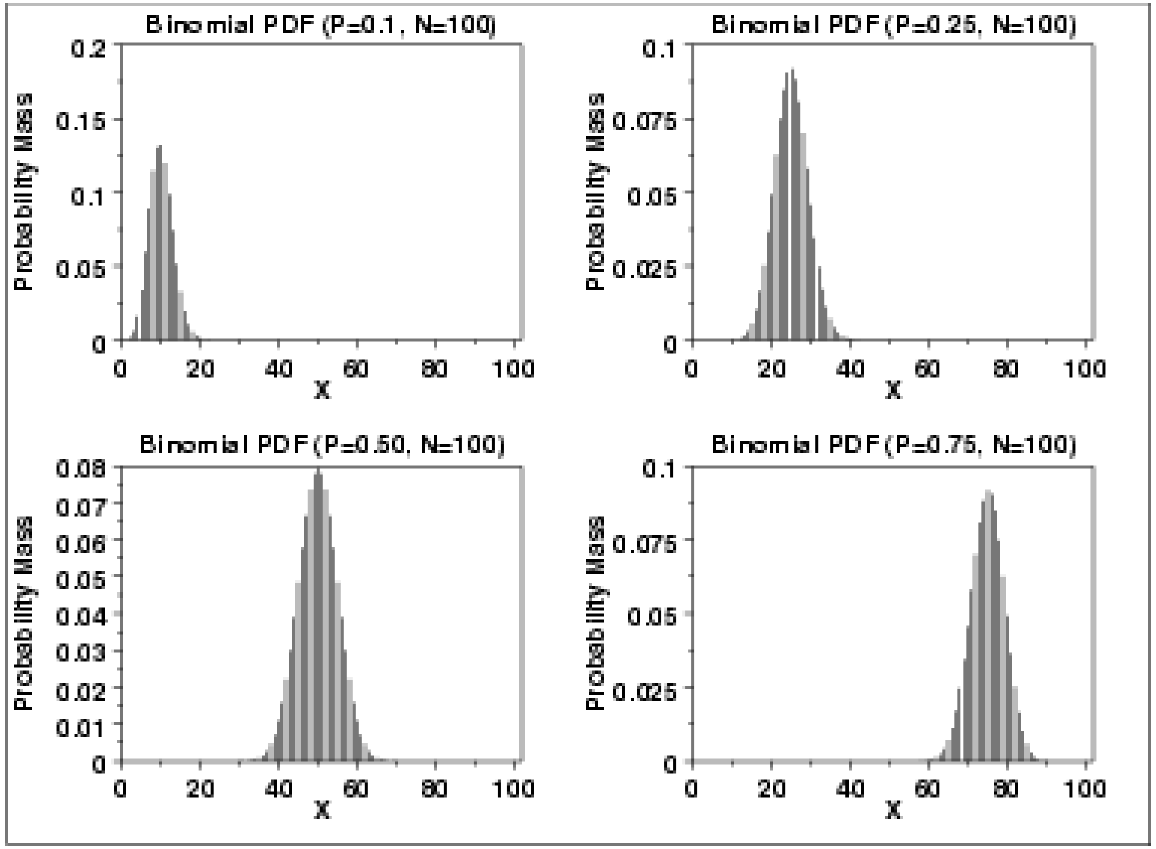
No of ways six of eleven severs can go down
k = 6 , N = 11



(Number of possible arrangements) ★ (.25^k ★ (.75^(11-k)))

Probability and distributions

Binomial Distribution



Source: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366i.htm

Poisson Distribution-

1. discrete distribution characterized by a single parameter λ. Both mean and variance are same and expressed as λ.

2. Random variable X takes only non-negative integer values i.e. 0 and above

3. The events occur randomly but average rate of event ‘r’ i.e. number of events per time period is constant.

4. Average number of events i.e. λ, in a time period = (r * time period) **

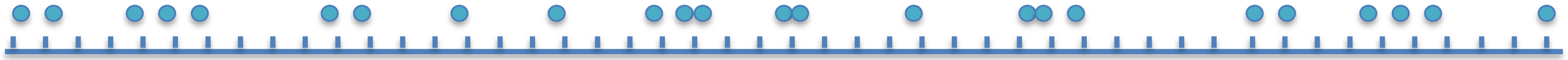
5. The Poisson distribution is the limiting case of a binomial distribution where N approaches infinity and p approaches zero while Np = λ.

6. Prerequisites -
   1. Number of observations n is large
   2. Each observation / trial is independent of other
   3. Probability of ‘success’ p is the same each trial and P is very small

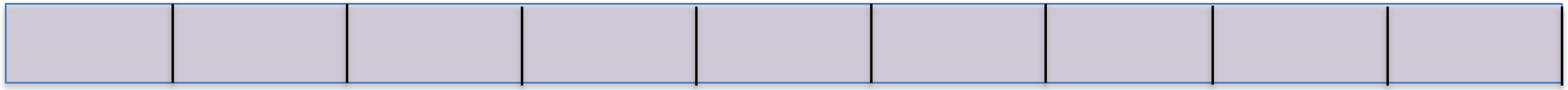
** Note: Poisson distribution can also be based on space, volume area etc.

Probability and distributions

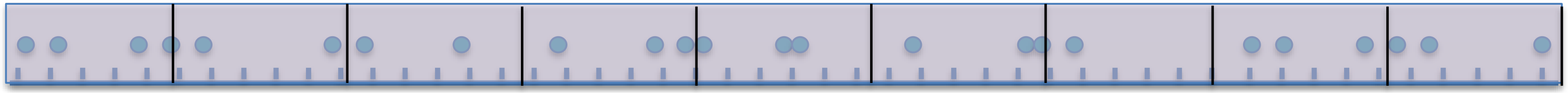
Poisson Distribution (Contd…) -



Rate of events (r) = number of observed events / time duration of observation = 25 / 50 = .5 per unit time



Time interval (t) = as per requirement = 6 units



Avg per time interval (λ) = r * t - This will be constant for all time intervals = 0.5 * 6 = 3
Variance across time intervals = λ = 3 in this example

Probability of K events in a time interval.

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Poisson Distribution (Contd…) -

Some examples of Poisson processes include –

1. Number of footprint in a mall in a day
2. Number of tech support calls in a day
3. Number of visitors to a website
4. Gene mutations in a culture
5. Radioactive decay in atoms
6. Photons arriving at a space telescope
7. Number of defects per motherboard
8. Number of bugs in a software
9. Bus arrival time at a bus stop

greatlearning

Probability and distributions

Poisson Distribution (Contd…) -

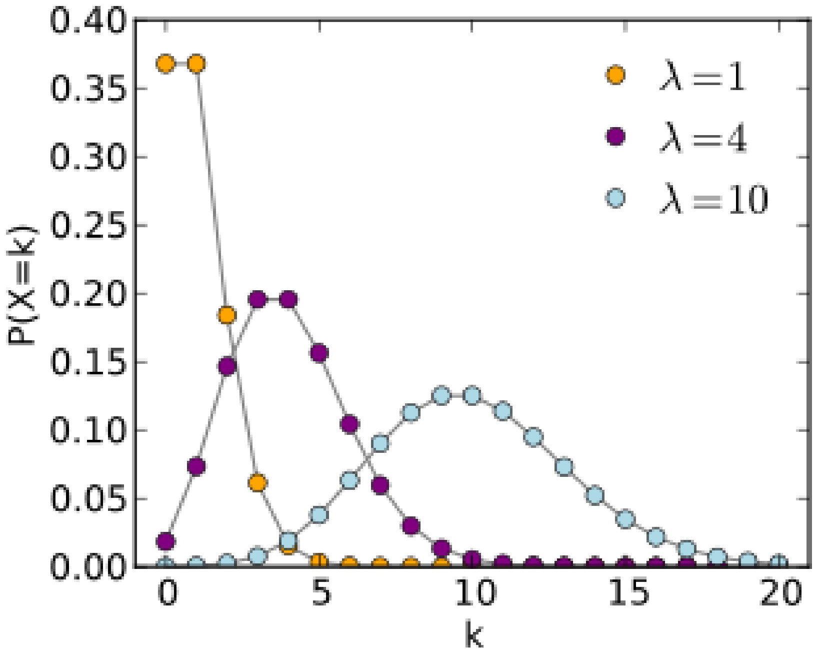
Assume a research conducted among an American Indian tribe found 10 per 1000 diabetic . What is the probability that a sample of 300 people with similar conditions will contain exactly 5 diabetic patients? (PIMA dataset)

a. • The average number of diabetics in 300 people is $\mu = 0.01 \times 300 = 3$

b. • The probability of finding 5 diabetics in the sample of 300 is:

c. • P(X)=(e^-3 * 3^5) / 5! = 0.10082

Poisson Distribution



Source: http://sherrytowers.com/2013/08/29/aml-610-fall-2013-module-ii-review-of-probability-distributions/#binom

Note: This URL has excellent introduction to hypothesis testing under various distributions

Probability and distributions

Normal Distribution – also called the Gaussian distribution

1. Is an important family of continuous probability distributions.
2. Each member of the family may be defined by two parameters, the mean and the variance
3. Standard normal distribution is the normal distribution with a mean of zero and variance of 1
4. Carl Friedrich Gauss used it extensively for his astronomical studies and discovered the equation (statistical model). The function is also called bell curve.
5. Apparently this type of distribution is common in nature. Hence the word "Normal" is also used for this distribution. It has following characteristics –
    1. The mean, median and mode of the distribution coincide.
    2. The distribution curve is symmetric bell-shaped about the line x = µ.
    3. The to

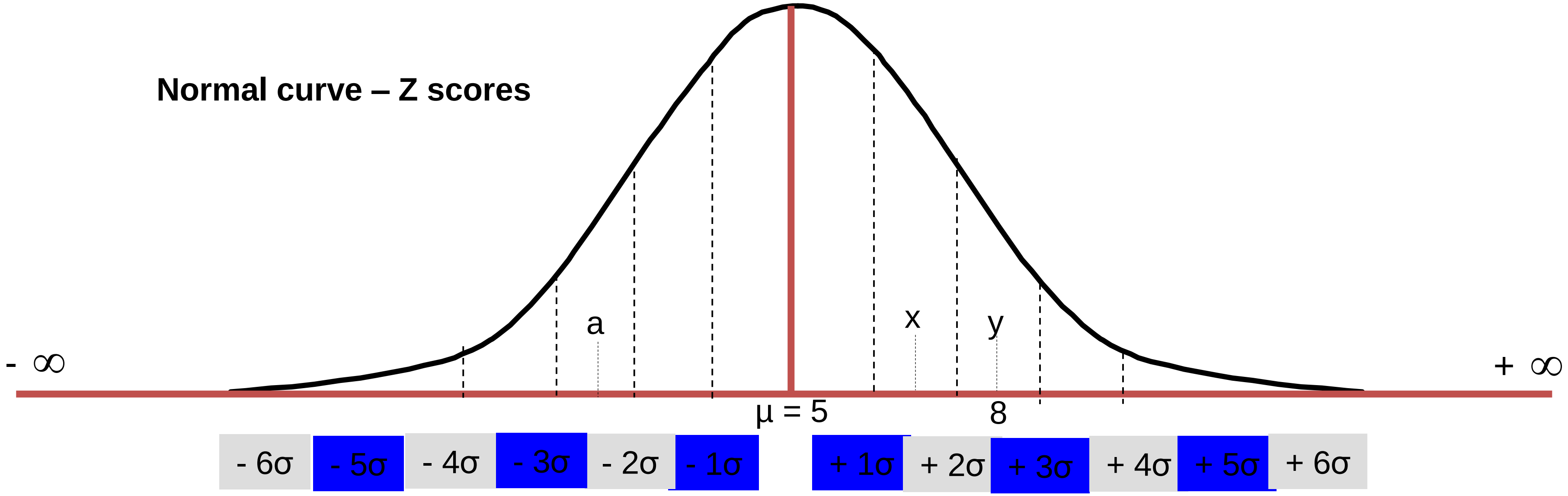
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\sigma^2$$

    4. Expected value of x = µ , variance =
6. Most of the Scikit_Learn parametric machine learning algorithms assume the data on the individual attributes with continuous values are normally distributed

Probability and distributions

Normal curve – Z scores



1. Z score is the distance of a data point in terms of standard deviation

   For e.g. “x” is +1.5σ from the mean, z = +1.5

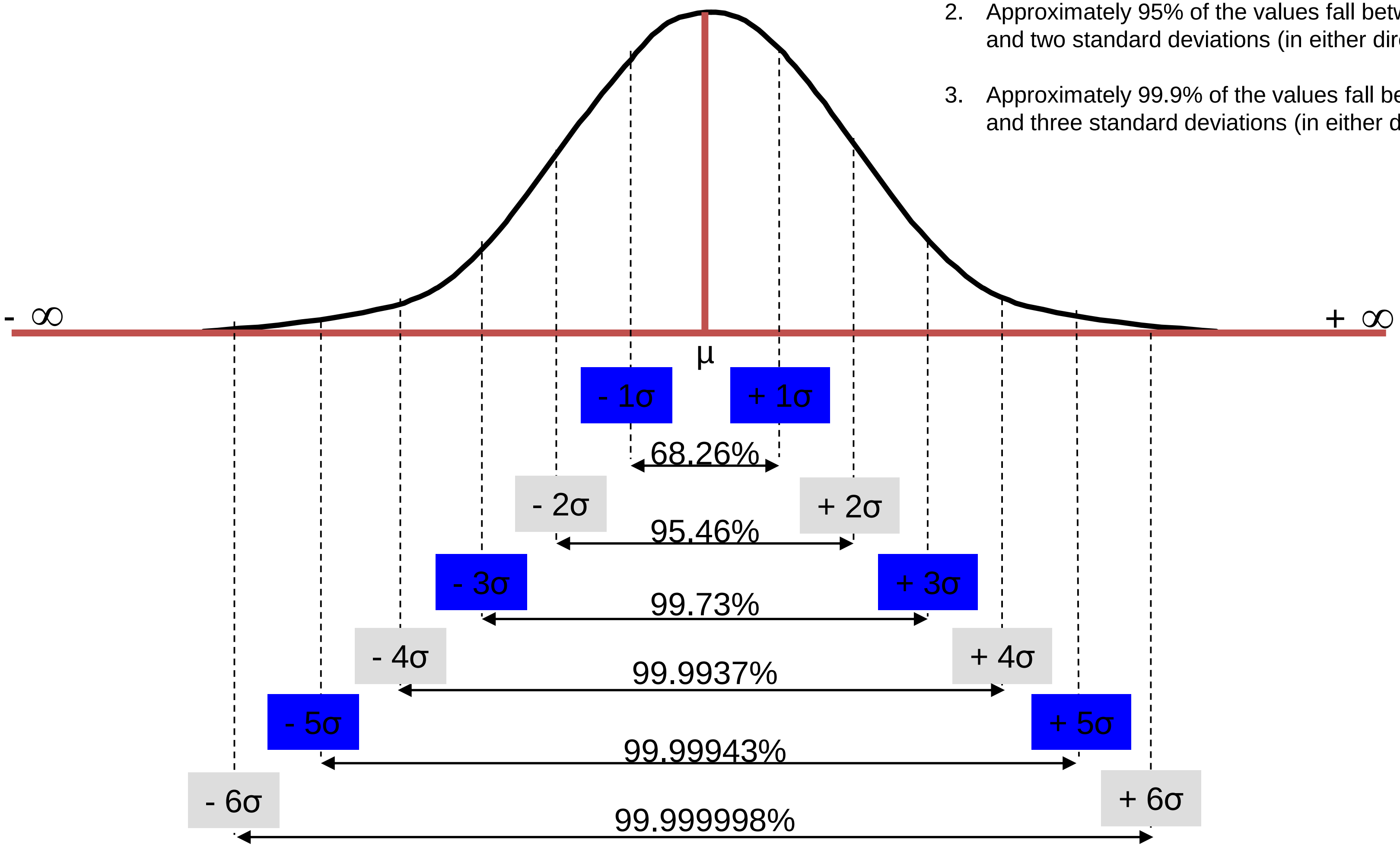
   “y” is +2.5σ from the mean, z = +2.5

   “a” is -2.5σ from the mean, z = -2.5

2. Values with larger magnitude Z score occur less and less frequently

Probability and distributions

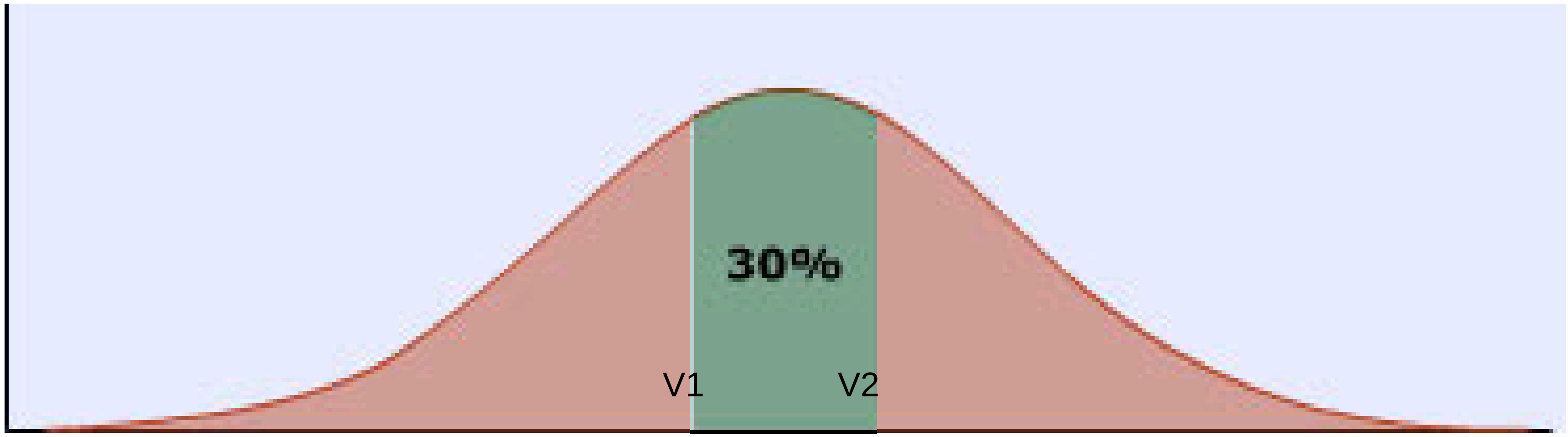
Normal curve – Z score and area of the curve



Normal distributions have the following characteristics:

1. Approximately 68% of the values fall between the mean and one standard deviation (in either direction)
2. Approximately 95% of the values fall between the mean and two standard deviations (in either direction)
3. Approximately 99.9% of the values fall between the mean and three standard deviations (in either direction)

Normal distribution with area under the curve as probability



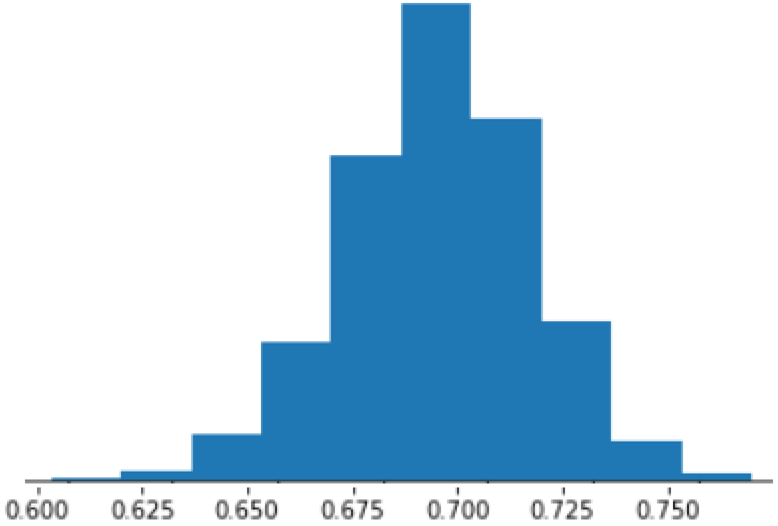
Possible values of a given parameter e.g. diameter of a bolt up to infinity

1. Probability of producing output with values between V1 and V2 is area of the curve under between the two values / total area of the bell curve
2. What is the probability of producing output with
   - values less than V1
   - Values >v1
   - Values < v2
   - Value = v1 or v2

Note: normal distribution is one of the many possible distribution that may be used to represent a process. This is only for explanation

Probability and distributions

1. Let the accuracy scores of a model executed using bootstrapping constitute a normal distribution with mean of 72% and standard deviation of 6 %. What is the probability that the model will give a score greater than 75% in production environment?
   a. Let random variable X corresponds to the accuracy of the model in production.
   b. We want to calculate P(X > 75%).
   c. Calculate the Z-score of 75 % accuracy, = $\frac{x-\mu}{\sigma}$ 5 – 72) / 6 = 0.5
   d. The mean µ of the distribution is 72 % and the standard deviation σ is 6%
   e. P(X > 75%) = 1 – P(X<=75) i.e. P(Z <= 0.5). From the distribution table = .69
   f. Hence P(X>75%) = 1 - .69 = .31



Joint & Conditional Probabilities

Joint & Conditional Probabilities

Joint Probability – is the probability of multiple events occurring together (we are not talking of causality here i.e. one event leads to another). For e.g.

1. probability of drawing a king from a deck of cards is 4/52
2. Probability of drawing a red colour card from a deck of cards is 26/52
3. Probability of drawing a red colour king = 2 / 52

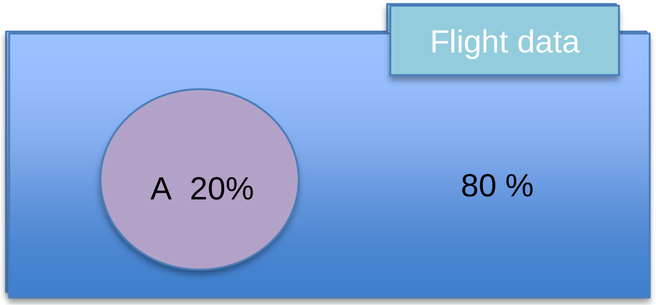
Conditional Probability – it is the probability that an event has occurred (not yet observed) given another event has occurred. For e.g.

1. given the card drawn is red (an event has occurred)
2. what is the probability it is a king (event not yet observed)?
3. Since the card is red, there are 26 likely values for red
4. Of these 26 possible values we are interested in king which is 2 (king of diamonds and heart)
5. Thus the conditional probability that the card is a king given red card is 2 /26
6. Compare this with joint probability of red king (2/52).
7. Given an event has occurred, it increases the probability of the other event

Joint & Conditional Probabilities

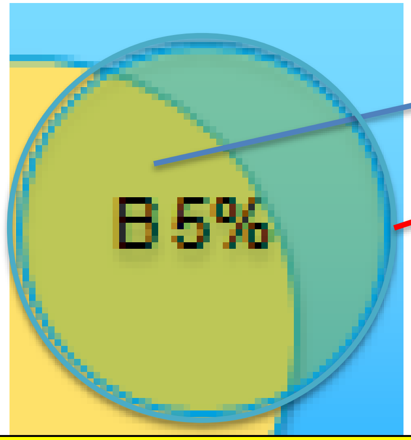
a. Imagine you represent all the flight experience you had till date as the blue area in a mathematical space. The dimensions of the boxes and circles are immaterial

b. Of these experiences, 20% of the time you experienced flight delay



Joint & Conditional Probabilities

P(flight delay given fog) = P (A n B) / P(B)



More the overlap, more the occurrences of flight delay and fog

P(A| B) = P(A n B) / P(B) → eq 1

Naïve Bayes Classifier -

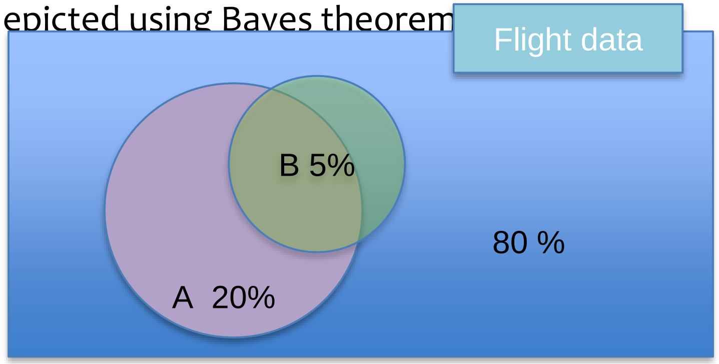
a. The relationship between dependent events is depicted using Bayes theorem

Likelihood Prior prob

Posterior

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Evidence



b. Probability of event A given that event B has occurred (fog has formed) depends on

I. Apriori probability of fog occurring whenever there was flight delay – P (B/A)

II. Apriori probability of flight delay P(A) which is 20% in the example

III. Apriori probability of flight facing fog P(B) which is 5% in the example

c. When it is a matter of deciding the class of an output such as whether flight will get delayed or not, we calculate P(A/B) and P(!A/B), compare which is higher. Since in both the denominator is P(B), it is ignored as it has no influence on which class will it be

d. However, to calculate the updated probability of a class, denominator P(B) is required

Hypothesis & Hypothesis testing

Hypothesis

- A hypothesis is an educated guess or proposition that attempts to explain a set of facts or natural phenomenon.

- Hypotheses could be formulated based on initial analysis of available data, domain knowledge, prior experience etc.

- The goal of a hypothesis is to explain an observation and set the direction for further research

Hypothesis

1. Null Hypothesis (H0)–
    a. Claims no significant change, no difference, no effect. For e.g.
        a. in process improvement H0 claims the productivity before and after the change is same, i.e. process improvement had no effect
        b. Modelling to predict potential default/non-default status based on attributes such as income, H0 states income has no effect on default/non-default status

    b. It is the default hypothesis which is either accepted or rejected based on the evidence in the data

    c. When H0 is accepted, it is said that the data does not have sufficient evidence to reject H0 because it is not rejected

Hypothesis

	Alternate Hypothesis	NULL Hypothesis
1	Student’s performance in college is impacted by quality of social life	Student’s performance in college is not related to quality of social life
2	Drop in service usage is linked to customer churn	Drop in service usage is independent of customer decision to churn
3	Polar caps melting is result of increase in air pollution	Polar ice caps melting is not linked to increase in air pollution.
4	There is impact of colour on resale value of different sedan models	Colour of a car has no impact on resale value of a car model
5	Demonetization will improve GDP	Demonetization will have no impact on GDP
6	The machine learning model is an indication that it has learnt the task	The accuracy is a statistical chance