Andhra Pradesh State Skill Development Corporation

Machine Learning

Thinking Statistically in Machine Learning

CHAPTER 3

Thinking Statistically in Machine Learning

Importance of Statistics:

Statistics is a collection of tools that you can use to get answers to important questions about data. Statistics is a required prerequisite for most books and courses on applied machine learning. But **what exactly is statistics?**
Statistics is a subfield of mathematics. It refers to a collection of methods for working with data and using data to answer questions.

It is because the field consists of a grab bag of methods for working with data that can seem large and amorphous to beginners. It can be hard to see the line between methods that belong to statistics and methods that belong to other fields of study.

When it comes to the statistical tools that we use in practice, it can be helpful to divide the field of statistics into two large groups of methods: descriptive statistics for summarizing data, and inferential statistics for drawing conclusions from samples of data.

- **Descriptive Statistics**: Descriptive statistics refer to methods for summarizing raw observations into information that we can understand and share.
- **Inferential Statistics**: Inferential statistics is a fancy name for methods that aid in quantifying properties of the domain or population from a smaller set of obtained observations called a sample..

Statistical Concepts:

1. Statistics in Data Preparation

Statistical methods are required in the preparation of train and test data for your machine learning model.
This includes techniques for:

- Outlier detection.
- Missing value imputation.
- Data sampling.
- Data scaling.
- Variable encoding. And much more.

A basic understanding of data distributions, descriptive statistics, and data visualization is required to help you identify the methods to choose when performing these tasks.

2. Statistics in Model Evaluation

Statistical methods are required when evaluating the skill of a machine learning model on data not seen during training.
This includes techniques for:

- Data sampling.
- Data Resampling.
- Experimental design.

Resampling techniques such as k-fold cross-validation are often well understood by machine learning practitioners, but the rationale for why this method is required is not.

3. Statistics in Model Selection

Statistical methods are required when selecting a final model or model configuration to use for a predictive modelling problem.
These include techniques for:

- Checking for a significant difference between results.
- Quantifying the size of the difference between results.

This might include the use of statistical hypothesis tests.

4. Statistics in Model Presentation

Statistical methods are required when presenting the skill of a final model to stakeholders.
This includes techniques for:

- Summarizing the expected skill of the model on average.
- Quantifying the expected variability of the skill of the model in practice.

This might include estimation statistics such as confidence intervals.

5. Statistics in Prediction

Statistical methods are required when making a prediction with a finalized model on new data.
This includes techniques for:

- Quantifying the expected variability for the prediction.

This might include estimation statistics such as prediction intervals.

Let's take a look as some examples of real analyses or applications you might need to implement as a data scientist:

1. Experimental design: Your company is rolling out a new product line, but it sells through offline retail stores. You need to design an A/B test that controls for differences across geographies. You also need to estimate how many stores to pilot in for statistically significant results.
2. Regression modelling: Your company needs to better predict the demand of individual product lines in its stores. Under-stocking and overstocking are both expensive. You consider building a series of regularized regression models.
3. Data transformation: You have multiple machine learning model candidates you're testing. Several of them assume specific probability distributions of input data, and you need to be able to identify them and either transform the input data appropriately or know when underlying assumptions can be relaxed.

Probability mass function

The **probability mass function** is the function which describes the probability associated with the random variable x. This function is named P(x) or P(x = x) to avoid confusion. P(x = x) corresponds to the probability that the random variable x takes the value x (note the different typefaces).

Let’s roll a die an infinite number of times and look at the proportion of 1, the proportion of 2 and so on. We call x the random variable that corresponds to the outcome of the dice roll. Thus the random variable x can only take the following discrete values: 1, 2, 3, 4, 5 or 6. It is thus a **Discrete random variable**.

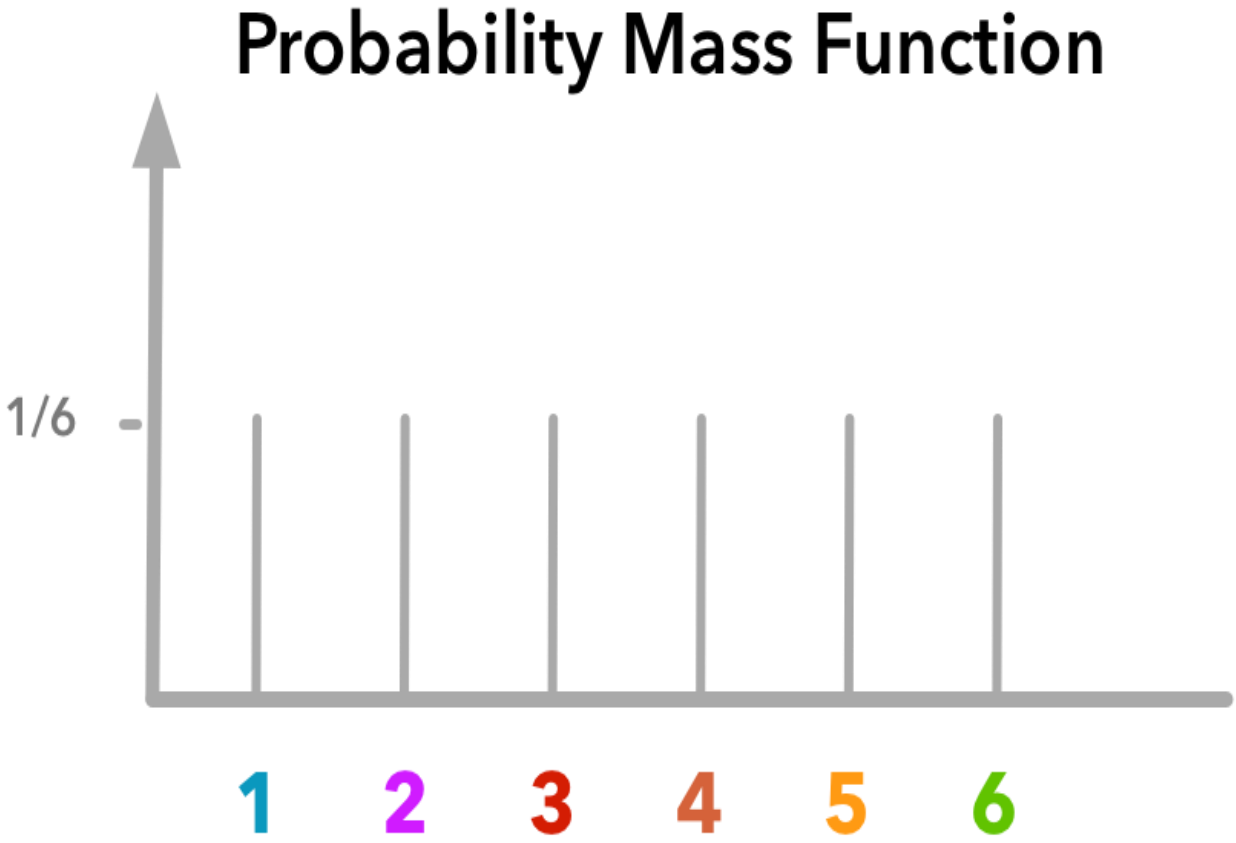
The aim of the probability mass function is to describe the probability of each possible value. In our example, it describes the probability to get a 1, the probability to get a 2 and so on. In the case of a dice rolling experiment, we have the same probability to get each value (if we assume that the die is perfect). This means that we can write:

$P(x = 1) = P(x = 2) = P(x = 3) = P(x = 4) = P(x = 5) = P(x = 6)$

Now, how can we calculate the probabilities $P(x = 1)$, $P(x = 2)$ etc.? Since we have 6 possible outcomes and that they are equiprobable we have:

$P(x = 1) = P(x = 2) = P(x = 3) = P(x = 4) = P(x = 5) = P(x = 6) = 1/6$

By the way, this distribution shows the same probability for each value: it is called the **Uniform Distribution.**



Probability mass function of the dice experiment
The y-axis gives the probability and x-axis the outcome.

Probability Density Function

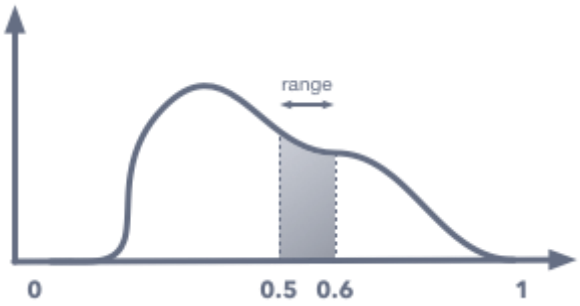
Some variables are not discrete. They can take an infinite number of values in a certain range. But we still need to describe the probability associated with outcomes. The equivalent of the probability mass function for a continuous variable is called the **probability density function**.In the case of the probability mass function, we saw that the y-axis gives a probability. For instance,

in the plot we created with Python, the probability to get a 1 was equal to 1/6 = 0.16 (check on the plot above). It is 1/6 because it is one possibility over 6 total possibilities.
However, we can't do this for continuous variables because the total number of possibilities is infinite. For instance, if we draw a number between 0 and 1, we have an infinite number of possible outcomes (for instance 0.320502304…). In the example above, we had 6 possible outcomes, leading to probabilities around 1/6. Now, we have each probability equal to $1/+\infty = 0$. Such a function would not be very useful.

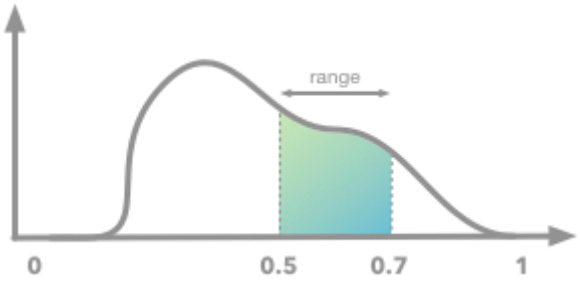
For that reason, the y-axis of the probability density function doesn't represent probability values. To get the probability, we need to calculate the **area under the curve** (we will see below some details about the area under the curve). The advantage is that it leads to the probabilities according to a certain range (on the x-axis): the area under the curve increases if the range increases. Let's see some examples to clarify all of this.

Example

Let's say that we have a random variable x that can take values between 0 and 1. Here is its probability density function:
Probability density function
We can see that 0 seems to be not possible (probability around 0) and neither 1. The pic around 0.3 means that will get a lot of outcomes around this value.
Finding probabilities from probability density function between a certain range of values can be done by calculating the area under the curve for this range. For example, the probability of drawing a value between 0.5 and 0.6 corresponds to the following area:



Probability density function and area under the curve between 0.5 and 0.6.We can easily see that if we increase the range, the probability (the area under the curve) will increase as well. For instance, for the range of 0.5–0.7:



Probability density function and area under the curve between 0.5 and 0.7.

We will see in a moment how to calculate the area under the curve and get the probability associated with a specific range.

Properties of the probability density function

These differences between the probability mass functions and the probability density function lead to different properties for the probability density function:

In this case, p(x) is not necessarily less than 1 because **it doesn't correspond to the probability** (the probability itself will still need to be between 0 and 1).

Common Continuous Distributions - Uniform Distribution

Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$

The graph of a uniform distribution curve looks like

You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10.

Let's try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution is:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = $(b-a)^2/12$

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

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

Exponential Distribution:

Let's consider the call center example one more time. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.
Other examples are:
1. Length of time between metro arrivals,
2. Length of time between arrivals at a gas station
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.
A random variable X is said to have an exponential distribution with PDF:

$$f(x) = \{ \lambda e^{-\lambda x}, \quad x \ge 0$$

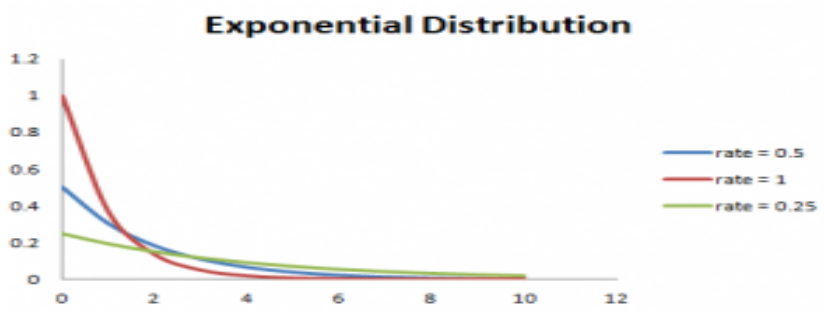
and parameter $\lambda > 0$ which is also called the rate.

For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.
Mean and Variance of a random variable X following an exponential distribution:

Mean -> $E(X) = 1/\lambda$

Variance -> $Var(X) = (1/\lambda)^2$

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



To ease the computation, there are some formulas given below.
$P\{X \le x\} = 1 - e^{-\lambda x}$, corresponds to the area under the density curve to the left of x.
$P\{X > x\} = e^{-\lambda x}$, corresponds to the area under the density curve to the right of x.

$P\{x1<X\leq x2\} = e^{-\lambda x1} - e^{-\lambda x2}$, corresponds to the area under the density curve between x1 and x2.

Common Discrete Distributions - Binomial and Multinomial Distributions

Binomial Distribution

Let's get back to cricket. Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let's assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker? So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

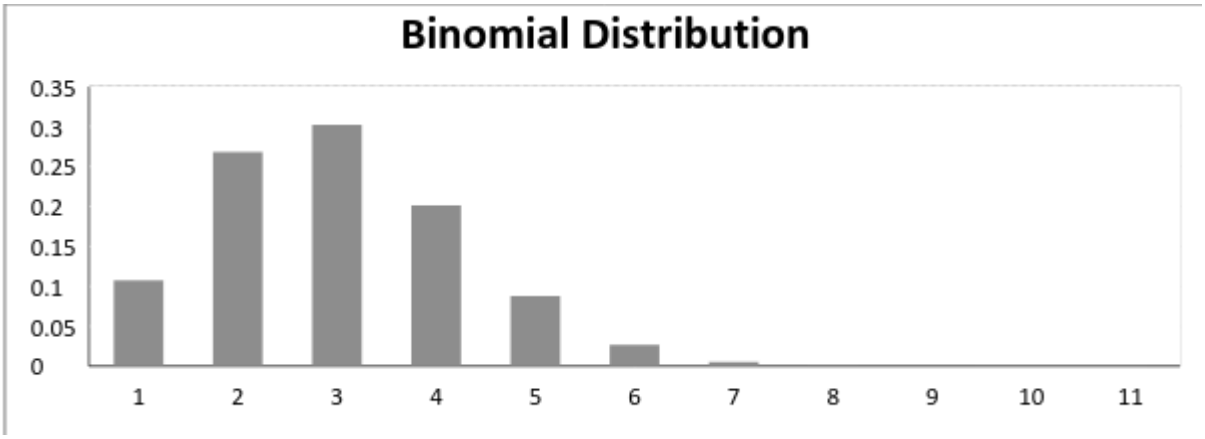
On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

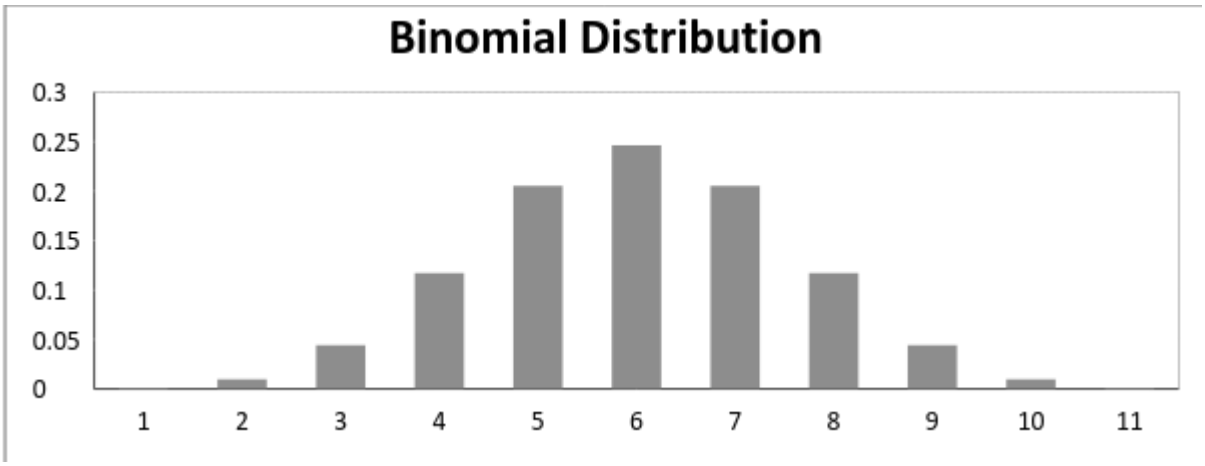
The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean -> $\mu = n*p$

Variance -> $Var(X) = n*p*q$

Multinoulli Distribution

The Multinoulli distribution, also called the categorical distribution, covers the case where an event will have one of K possible outcomes.

- x in {1, 2, 3, …, K}

It is a generalization of the Bernoulli distribution from a binary variable to a categorical variable, where the number of cases K for the Bernoulli distribution is set to 2, K=2.
A common example that follows a Multinoulli distribution is:

- A single roll of a die that will have an outcome in {1, 2, 3, 4, 5, 6}, e.g. K=6.

A common example of a Multinoulli distribution in machine learning might be a multi-class classification of a single example into one of K classes, e.g. one of three different species of the iris flower.
The distribution can be summarized with K variables from p1 to pK, each defining the probability of a given categorical outcome from 1 to K, and where all probabilities sum to 1.0.

- P(x=1) = p1
- P(x=2) = p1
- P(x=3) = p3
- …
- P(x=K) = pK

In the case of a single roll of a die, the probabilities for each value would be 1/6, or about 0.166 or about 16.6%

Poisson Distribution

Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The number of emergency calls recorded at a hospital in a day.
2. The number of thefts reported in an area on a day.
3. The number of customers arriving at a salon in an hour.
4. The number of suicides reported in a particular city.
5. The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called **Poisson distribution** when the following assumptions are valid:
1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.
Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

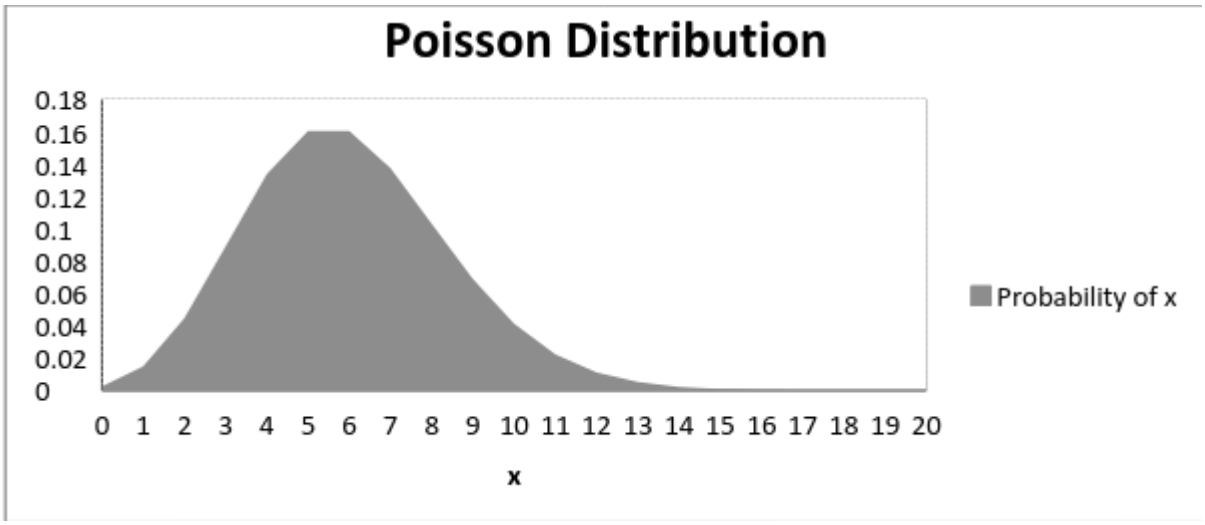
- λ is the rate at which an event occurs,
- t is the length of a time interval,
- And X is the number of events in that time interval.

Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.
Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

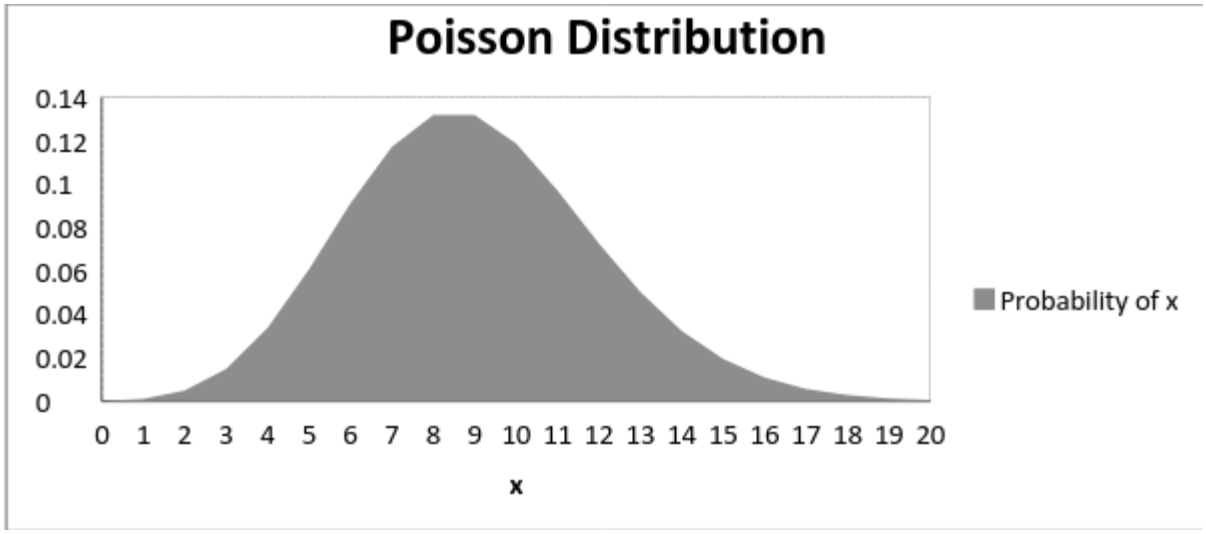
The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots..$$

The mean μ is the parameter of this distribution. μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to increase in mean.



It is perceptible that as the mean increases, the curve shifts to the right.
The mean and variance of X following a Poisson distribution:

Mean -> E(X) = μ

Variance -> Var(X) = μ