Click to prove you're human

Welcome to the world of Probability in Data Science! Let me start things off with an intuitive example. Imagine you are a Data Analyst or someone making Machine Learning models or working on algorithms or python scripts, and you need to analyze trends. Still, you dont have enough data set with you to analyze the trend in your dataset. Through this article, lets find a way to solve this problem using probability distribution. In this article we will exploring different types of probability distribution and their use cases and need of probability distribution in various data types. Table of contents A probability distribution is a mathematical function that defines the likelihood of different outcomes or values of a variable. This function is commonly represented by a graph or probability table, and it provides the probabilities of various possible results of an experiment or random phenomenon based on the sample space and the probabilities of events. Probability distributions are fundamental in probability theory and statistics for analyzing data and making predictions. Checkout this article about the Probability Distributions for Data Science for Beginners Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students. You gave these graded papers to a data entry guy in the university and told him to create a spreadsheet containing the grades of all the students. But the guy only stores the grades and not the corresponding students. He made another blunder; he missed a few entries in a hurry, and we have no idea whose grades are missing. One way to find this out is by visualizing the grades and seeing if you can find a trend in the data. The graph you plotted is called the frequency distribution of the data. You see that there is a smooth curve-like structure that defines our data, but do you notice an anomaly? We have an abnormally low frequency at a particular score range. So the best guess would be to have missing values that remove the dent in the distribution. This is how you try to solve a real-life problem using data analysis. Distribution is a must-know concept for any Data Scientist, student, or practitioner. It provides the basis for analytics and inferential statistics. Probability distributions are versatile tools used in various fields and applications. They primarily model and quantify uncertainty and variability in data, making them fundamental in data science, statistics, and decision-making processes. Probability distributions enable us to analyze data and draw meaningful conclusions by describing the likelihood of different outcomes or events. In statistical analysis, these distributions play a pivotal role in parameter estimation, hypothesis testing, and data inference. They also find extensive use in risk assessment, particularly in finance and insurance, where they help assess and manage financial risks by quantifying the likelihood of various outcomes. Machine learning algorithms leverage probability distributions to model uncertainty in predictions, enhancing their ability to make accurate forecasts. Additionally, probability distributions support quality control efforts, allowing for the monitoring and controlling processes by identifying deviations from expected values. Probability distributions are not confined to data analysis alone; they also play crucial roles in fields like engineering, environmental science, epidemiology, and physics. In these diverse domains, probability distributions enable reliable modeling, simulation, and prediction, ultimately contributing to informed decision-making and problem-solving. Before we jump on to the explanation of distributions, lets see what kind of data we can encounter. The data can be discrete or continuous. Discrete Data, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5, or 6, not 1.5 or 2.45. (Discrete Probability Distribution) Continuous Data can take any value within a given range. The range may be finite or infinite. For example, a girls weight or height, the length of the road. The weight of a girl can be any value 54 kgs, 54.5 kgs, or 54.5436kgs. (Continuous Probability Distribution) Now let us start with the types of distributions. Here is a list of distribution types: Bernoulli Distribution Uniform Distribution Binomial Distribution Normal or Gaussian Distribution Exponential Distribution Poisson Distribution Lets start with the easiest distribution, which is Bernoulli Distribution. It is actually easier to understand than it sounds! All you cricket junkies out there! At the beginning of any cricket match, how do you decide who will bat or ball? A toss! It all depends on whether you win or lose the toss, right? Lets say if the toss results in a head, you win. Else, you lose. Theres no midway. A Bernoulli distribution has only two bernoulli trials or possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X with a Bernoulli distribution can take the value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p. Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure.Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes. The probability mass function is given by: px(1-p)1-x where x (0, 1) It can also be written as: The probabilities of success and failure need not be equally likely, like the result of a fight between Undertaker and me. He is pretty much certain to win. So, in this case probability of my success is 0.15, while my failure is 0.85 Here, the probability of success(p) is not the same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight. Here, the probability of success = 0.15, and the probability of failure = 0.85. The expected value is exactly what it sounds like. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Bernoulli distribution is found as follows: The variance of a random variable from a bernoulli distribution is: V(X) = E(X) [E(X)] = p p = p(1-p) There are many examples of Bernoulli distribution, such as whether it will rain tomorrow or not, where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game. When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely, which 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: 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. Lets 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 are: Mean -> E(X) = (a+b)/2 Variance -> V(X) = (b-a)/12 The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by: Lets get back to cricket. Suppose you won the toss today, indicating a successful event. You toss again, but you lose this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Lets 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, the 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 the same for all the trials is called a Binomial Distribution. The outcomes need not be equally likely. Remember the example of a fight between Undertaker and me? 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 doesnt 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. Based on the above explanation, the properties of a Binomial Distribution are: Each trial is independent. There are only two possible outcomes in a trial success or failure. A total number of n identical trials are conducted. The probability of success and failure is the same for all trials. (Trials are identical.) The mathematical representation of binomial distribution is given by: A binomial distribution graph where the probability of success does not equal the probability of failure looks like this. Now, when the 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 -> = n*p Variance -> Var(X) = n*p*q The normal distribution represents the behavior of most of the situations in the universe (That is why its called a normal distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics: The mean, median, and mode of the distribution coincide. The curve of the distribution is bell-shaped and symmetrical about the line x=. The total area under the curve is 1. Exactly half of the values are to the left of the center, and the other half to the right. A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity, then the shapes will be quite similar. The PDF of a random variable X, following a normal distribution, is given by: The mean and variance of a random variable X, which is said to be normally distributed, is given by: Mean -> E(X) = Variance -> Var(X) = ^2 Here, (mean) and (standard deviation) are the parameters.The graph of a random variable X ~ N (, ) is shown below. A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1. For such a case, the PDF becomes: 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: The number of emergency calls recorded at a hospital in a day. The number of thefts reported in an area in a day. The number of customers arriving at a salon in an hour. The number of suicides reported in a particular city. The number of printing errors on 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. Check about Different Probability Distributions in this Article A distribution is called a Poisson distribution when the following assumptions are valid: Any successful event should not influence the outcome of another successful event. The probability of success over a short interval must equal its probability over a longer interval. 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: 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 denote the mean number of events in an interval of length t. Then, = *t. The PMF of X following a Poisson distribution is given by: The mean is the parameter of this distribution. is also defined as the times the 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 the increase in the 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) = Variance -> Var(X) = Lets consider the call center example one more time. What about the interval of time between the calls? Here, the exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls. Other examples are: Length of time between metro arrivals Length of time between arrivals at a gas station The life of an air conditioner The 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: And parameter >0, which is also called the rate. For survival analysis, 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/ Variance -> Var(X) = (1/) Also, the greater the rate, the faster the curve drops, and the lower the rate, the flatter the curve. This is explained better with the graph shown below. To ease the computation, there are some formulas given below: P{Xx} = 1 e-x corresponds to the area under the density curve to the left of x P{X>x} = e-x corresponds to the area under the density curve to the right of x P{x1