Statistics Essentials for Data Science

Probability Distributions

Learning Objectives

By the end of this lesson, you will be able to:

- Explain the fundamental concept of random variables
- Build statistical analysis based on the principles of probability distribution
- Evaluate discrete distributions against continuous distributions
- Comprehend the commonly used continuous and discrete probability distributions

Business Scenario

ABC is a government agency. The agency stores a lot of data related to the citizens of the country, such as population statistics, forex rates, and other relevant topics.

The agency wants to analyze and determine the population’s data in the next few years. However, different theorems are supposed to be used to calculate the probability of this data being accurate.

In order to do this, the agency will determine discrete distributions and explore the commonly used continuous and discrete probability distributions.

Discussion: Random Variable

Duration: 15 minutes

- What is a random variable, and what are the types of random variables?
- What is a probability distribution?

Random Variable

Random Variable

It is a type of variable whose value depends on the numerical outcomes of certain random events.

These variables can only take values, as they are used to determine the results of a random event.

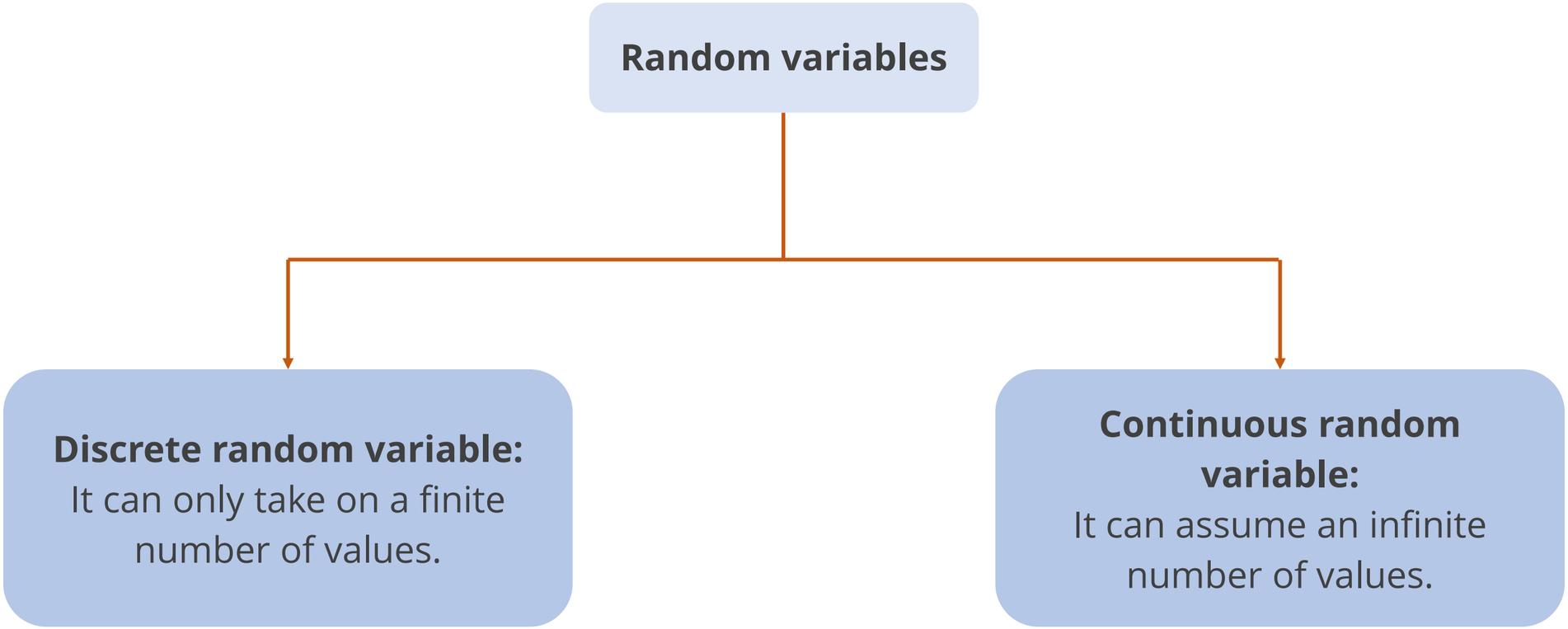
These variables must be quantifiable and often take the form of real numbers.

Random Variable: Example

Assume X as the outcome of rolling three dice

- Total number of outcomes = 6 × 6 × 6 = 216
- Therefore, when rolling three dice, there are 216 possible outcomes.
- Each outcome represents a unique combination of three numbers, ranging from 1 to 6.

Types of Random Variables



Discrete Random Variable

Examples of discrete random variables include:

Number of misprints on a page

Number of customers at a theatre

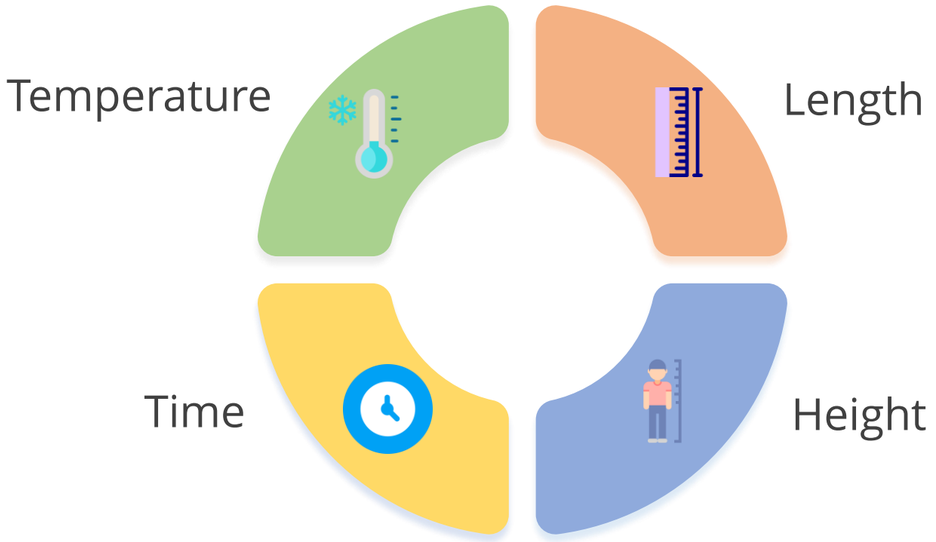
Outcome of throwing a dice

Number of cars at a gas station

These random variables take only selected values in a range, like whole numbers.

Continuous Random Variable

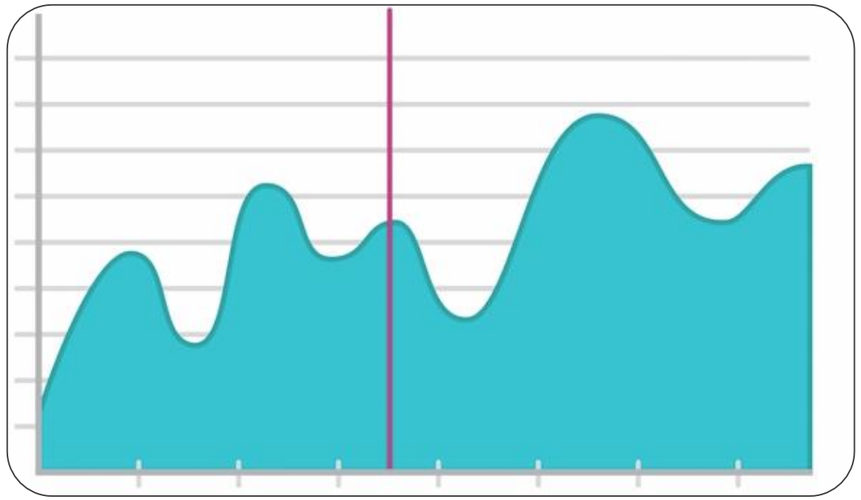
Examples of continuous random variables include:



These random variables can take any value in a certain range.

Probability Distribution

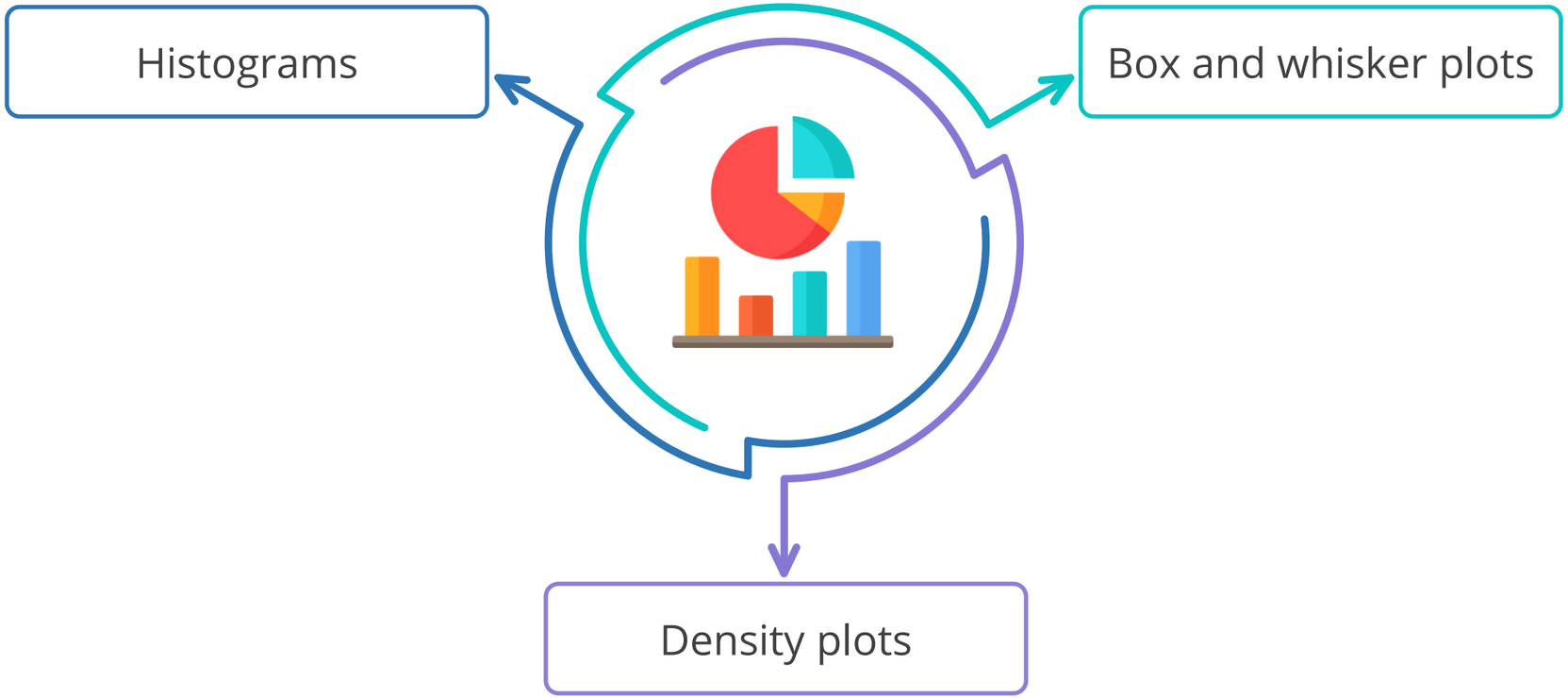
A probability distribution is a statistical function that describes all the possible values and probabilities for a random variable within a given range.



- This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution will be determined by several factors.
- These include the mean (average), standard deviation, skewness, and kurtosis of the distribution.

Probability Distribution

Probability distributions can be visualized using:



Probability Distribution: Example

Consider a production process that produces both non-defective and potentially defective pieces

After inspecting a large number of pieces, the following probabilities were determined:

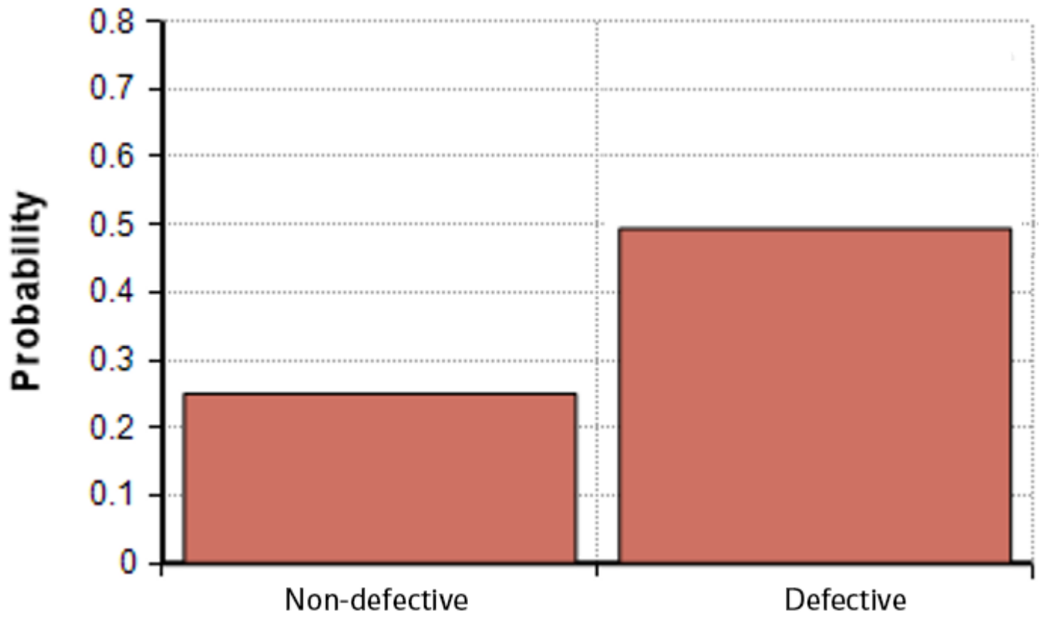
Outcomes	Probability
Non-defective	0.85
Defective	0.15

- The probability distribution indicates the production outcomes' likelihood.
- About 85% of pieces are likely non-defective (probability 0.85), while around 15% may be defective (probability 0.15).

Analyzing this probability distribution helps assess the distribution of defective and non-defective pieces in production, aiding in quality control evaluation, product acceptance decisions, and estimation of rework or reject rates.

Probability Distribution: Example

The probability distribution indicates how the total probability is distributed across all values. A random variable constitutes an event for each value it takes.



Probability Distribution: Example

The outcomes of throwing a dice are as follows:

Value	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

Probability Distribution: Example

The numerical value obtained by multiplying each value taken by the variable with its probability when added on is called the expected value of the random variable.

Value	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

Expected value
E[x] or μ

Probability Distribution: Expected Value

Example: Calculating the expected value for a random variable X that represents the outcome of throwing a dice

$$E[x] = \mu = 1/6 * (1+2+3+4+5+6) = 3.5$$

The average value expected to be seen over a significant number of dice rolls is 3.5.

Probability Distribution: Expected Value

An expected value represents the expected outcome for a random variable over a large number of trials.

It represents the theoretical average outcome that can be anticipated based on the probabilities assigned to all possible outcomes.

The expected value is calculated as the sum of all possible outcomes of a random variable, weighted by their probabilities.

Discussion: Random Variable

Duration: 15 minutes

- What is a random variable, and what are the types of random variables?

Answer: It is a variable that quantifies the outcome of a random experiment. The two types of random variables are discrete random variables and continuous random variables.

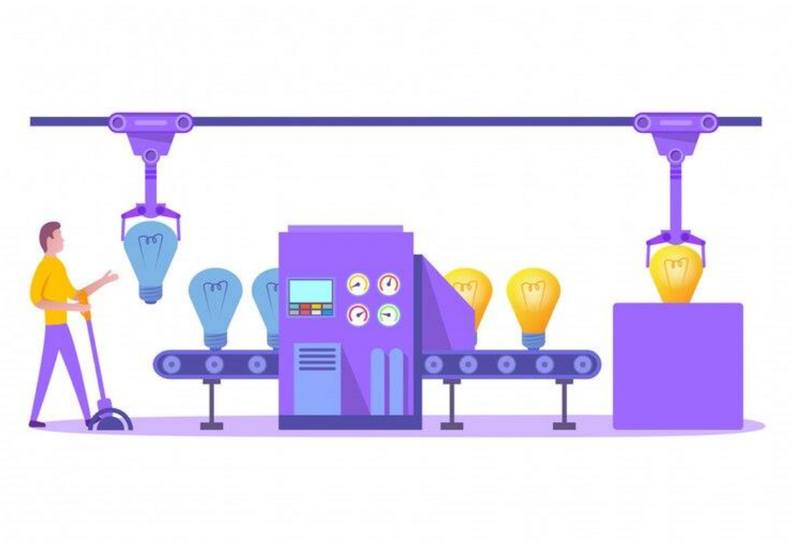
- What is a probability distribution?

Answer: It is a statistical function that describes all the possible values and probabilities for a random variable within a given range. This range will be bound by the minimum and maximum possible values.

Random Variable and Probability Distribution

Random Variable and Probability Distribution: Example

Consider the production process of a factory that produces light bulbs



Each light bulb can either be defective or non-defective.

Random Variable and Probability Distribution

Consider X as a random variable representing the number of defective light bulbs in a batch of n light bulbs

The probability distribution of X can be calculated using the following formula:

$$P(X = k) = (n \text{ choose } k) * p^k * (1 - p)^{(n - k)}$$

(n choose k) is the number of ways that k defective bulbs can be produced out of n total bulbs

- It is a binomial probability formula.
- We are dealing with a discrete random variable that represents the number of defective light bulbs in a batch of 1000.

Random Variable and Probability Distribution

Assume the production of a batch of 1000 light bulbs with a 0.05 probability of defective light bulbs

- The probability distribution of X is calculated as:
  $P(X = 0) = (1000 \text{ choose } 0) * (0.05)^0 * (0.95)^{1000} = 0.006$

  $P(X = 1) = (1000 \text{ choose } 1) * (0.05)^1 * (0.95)^{999} = 0.051$

  $P(X = 2) = (1000 \text{ choose } 2) * (0.05)^2 * (0.95)^{998} = 0.214$

  $P(X = 3) = (1000 \text{ choose } 3) * (0.05)^3 * (0.95)^{997} = 0.392$

  ...

  $P(X = 1000) = (1000 \text{ choose } 1000) * (0.05)^{1000} * (0.95)^0 = 0.00$
- The binomial coefficient (1000 choose 1) is calculated as 1000!/(1! * (1000 - 1)!), which simplifies to 1000.
- This probability distribution determines the probability of producing defective light bulbs in a batch.
- It is used to make informed decisions about quality control and other aspects of the production process.

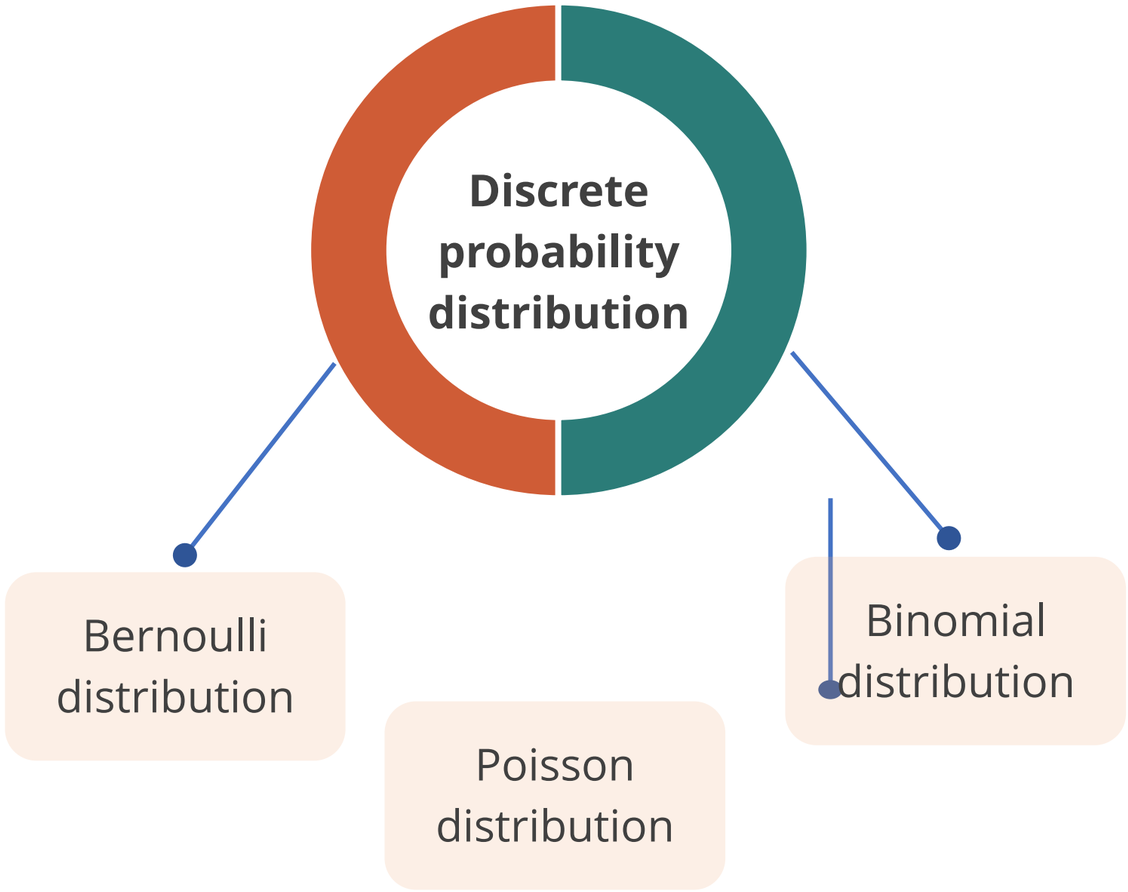
Commonly Used Discrete Probability Distributions

Discussion

Duration: 5 minutes

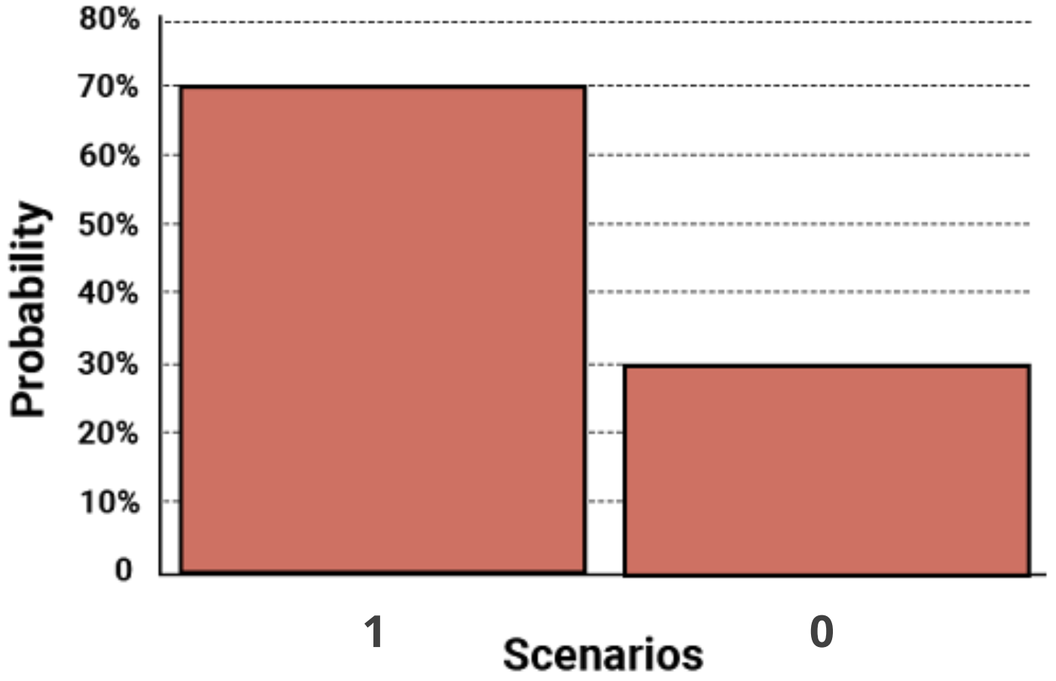
- What are the commonly used discrete probability distributions?
- What is a binomial distribution?

Discrete Probability Distribution



Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution that models a single trial with two possible outcomes, typically labeled as success (denoted as 1) and failure (denoted as 0).



Bernoulli Distribution

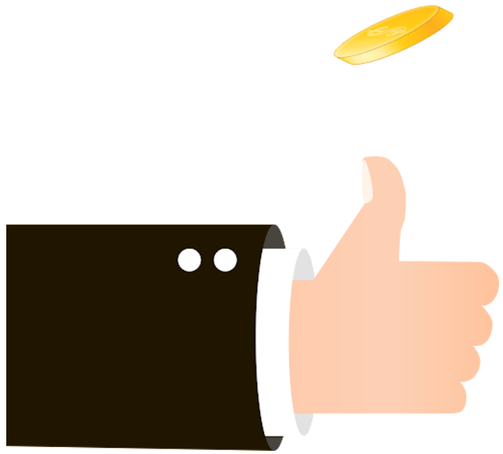
The Bernoulli probability is denoted by P and is computed using the following formula:

$$P(X = x) = p^x (1 - p)^{1-x}; \quad P(x) = \begin{cases} 1 - p \text{ for } x = 0 \\ p, \text{ for } x = 1 \end{cases}$$

- x represents the outcome, which can be a success (x=1) or failure (x=0).
- p denotes the probability of success.
- q equals 1 – p and represents the probability of failure.
- The value of "p" lies between 0 and 1 ($0 < p < 1$).

Bernoulli Distribution

It is commonly used to model situations with binary outcomes, such as flipping a coin, the success or failure of an event, or the presence or absence of a characteristic.



It serves as the building block for more complex distributions, such as the binomial distribution, which models the number of successes in a fixed number of Bernoulli trials.

Bernoulli Distribution: Example

Consider a scenario where there is a bag of marbles, and the objective is to determine the probability of drawing a red marble from the bag

Assume that the probability of drawing a red marble is 0.4

With the help of the Bernoulli distribution equation:

$$P(X = x) = p^x (1-p)^{1-x}; \quad P(x) = \begin{cases} 1 - p \text{ for } x = 0 \\ p, \text{ for } x = 1 \end{cases}$$

Bernoulli Distribution: Example

Find out the probabilities of drawing a red or non-red ball

The probability of drawing a red marble (success, X = 1) is:

$$P(X = 1) = p^1 * (1 - p)^{(1 - 1)} = p = 0.4$$

The probability of drawing a red marble is 0.4, or 40%.

The probability of drawing a non-red marble (failure, X = 0) is:

$$P(X = 0) = p^0 * (1 - p)^{(1 - 0)} = 1 - p = 0.6$$

The probability of drawing a non-red marble is 0.6, or 60%.

In this example, the Bernoulli distribution is used to model the probability of a single event (drawing a red marble) with two possible outcomes (success or failure).

Binomial Distribution

It is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, where each trial has two possible outcomes: success or failure.

The key characteristics include:

Fixed number of trials: The binomial distribution models a fixed number of trials, denoted by **n**.

Two possible outcomes: Each trial can result in one of the two outcomes.

Independence of trials: The outcome of one trial does not affect the outcome of any other trial. Each trial is independent.

Constant probability of success: The probability of success (often denoted as *p*) remains the same for each trial. The probability of failure is given by (1 - p).

Binomial Distribution

Let X denote a random variable that indicates the number of times an event occurs in *n* Bernoulli trials with a probability p of X occurring in the trial.

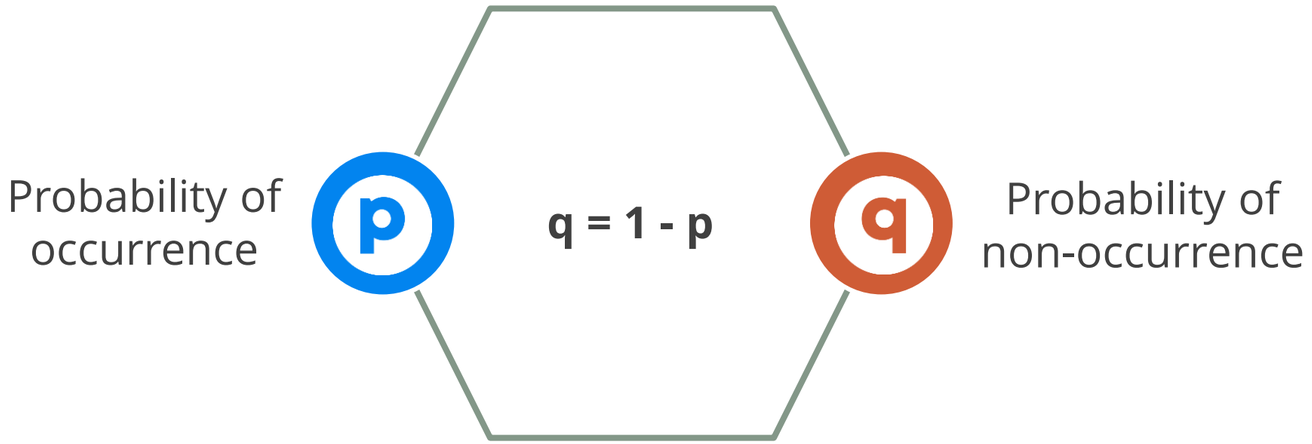
X follows the binomial distribution, and it indicates the number of event occurrences.

n indicates the number of trials.

p indicates the probability of success of the Bernoulli trials.

Binomial Distribution

If p is the probability of occurrence in a trial, then the probability of non-occurrence is q = 1 - p, which is also a constant.



Binomial Distribution

When X follows the binomial distribution with parameters n (the number of trials) and p (the probability of success in a single trial), the equation for calculating the probability of a specific outcome is as follows:

$$P(X = k) = C(n, k) * p^k * (1 - p)^{(n - k)}$$

Where,

- $P(X = k)$ represents the probability of achieving exactly k successes in n trials.
- $C(n, k)$ is the binomial coefficient, which signifies the number of ways to select k successes from n trials ($C(n, k) = n! / (k! * (n - k)!)$).
- p^k denotes the probability of k successes in the trials, obtained by raising the probability of success in one trial, p, to the k power.
- $(1 - p)^{(n - k)}$ indicates the probability of the remaining n - k trials resulting in failure, computed by raising the failure probability in a single trial, 1 - p, to the power of n - k.

Binomial Distribution: Example

Consider a supplier company that delivers electronic components to a manufacturing plant

- Historical data indicate a 10% defect rate in the supplied components.
- The manufacturing plant randomly selects 8 components from the supplier's shipment for inspection.

Calculating the probability of encountering exactly two defective components using the binomial distribution formula:

$$P(X = k) = C(n, k) * p^k * (1 - p)^{(n - k)}$$

- $n = 8$
- $k = 2$
- $p = 0.10$

Binomial Distribution: Example

Substituting the previous values of n, k, and p into the formula:

$$P(X = 2) = C(8, 2) * (0.10)^2 * (1 - 0.10)^{(8 - 2)}$$

Binomial coefficient: $C(8, 2) = 8! / (2! * (8 - 2)!) = 28$

$$\begin{aligned} P(X = 2) &= 28 * (0.10)^2 * (0.90)^6 \\ &= 28 * 0.01 * 0.531441 \\ &= 0.1488 \end{aligned}$$

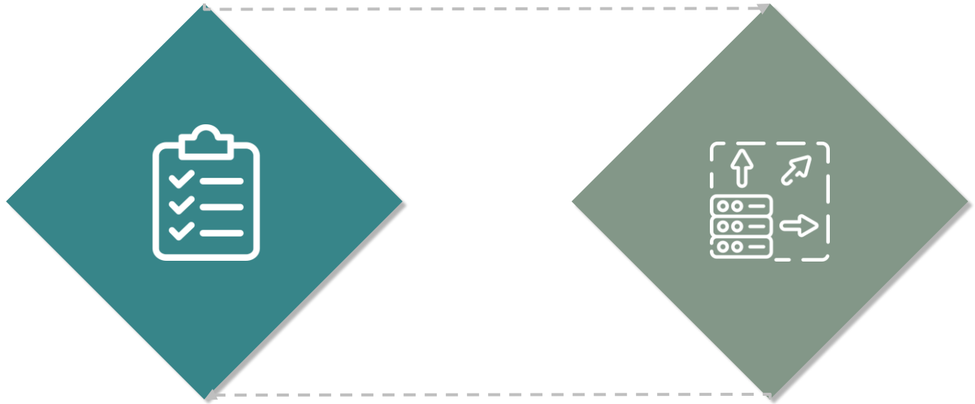
The probability of encountering exactly two defective components when inspecting eight components from the supplier's shipment is 0.1488, which is equivalent to 14.8%.

Poisson Distribution

The Poisson distribution can be used to analyze situations, where there are a limited number of outcomes from an experiment.

But there are some prerequisites to using Poisson distribution.

Numerous outcomes
even when an upper limit exists



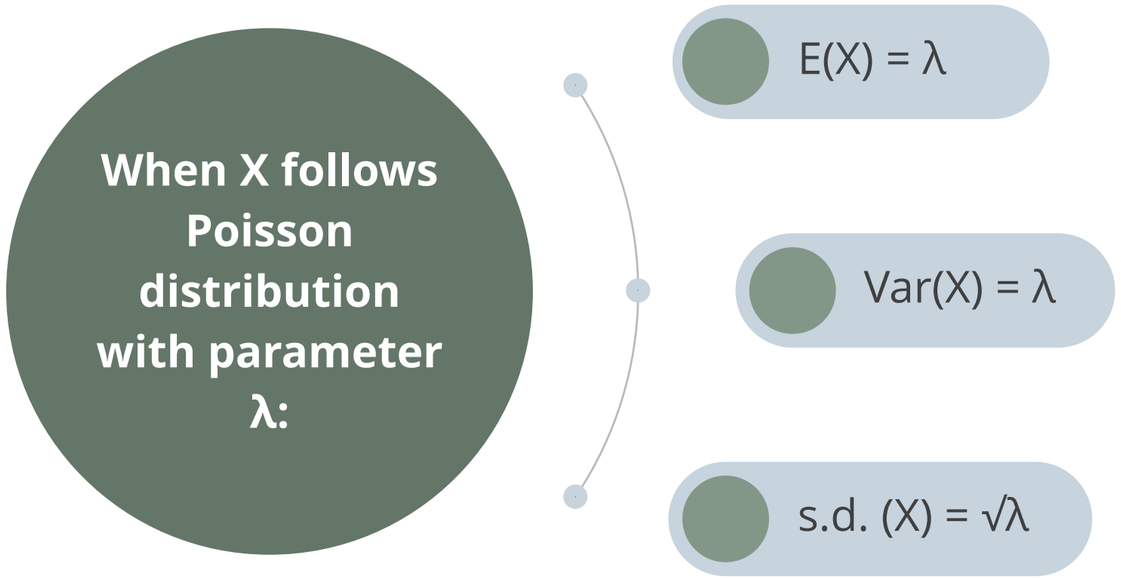
Very large value
with a very low probability of the outcome’s occurrence

Parameters of the Poisson Distribution

If X denotes the random variable indicating the number of occurrences that follows a Poisson distribution, the probabilities of X taking a given value [0, 1, 2, and so on] depend on its expected value λ.

Parameters of the Poisson Distribution

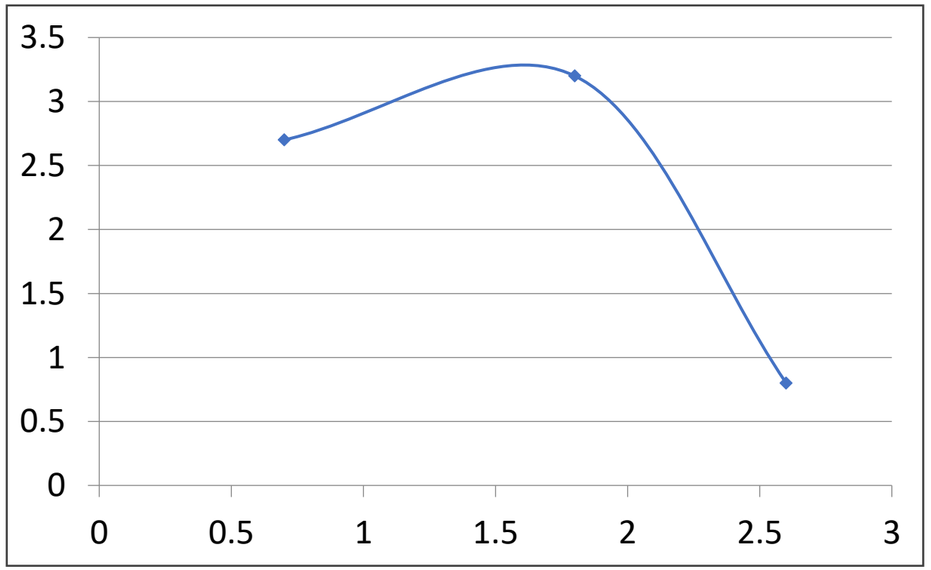
λ is the parameter of the distribution.



The probabilities can be obtained from tables or software.

Parameters of Poisson Distribution

Example: The number of spares required for any machine component during the equipment's life follows a Poisson distribution with $\lambda = 2$.



The firm must decide the number of spares required, so the probability of a stockout is at most 0.06.

Example

Solution: Let the random variable X denote the number of spares required. X follows a Poisson distribution with parameter $\lambda = 2$.

The table shows the stockout probabilities for different values of stocks. Note that the stockout probability for a given value = 1 – cumulative value.

K	0	1	2	3	4	5	6
P(X=k)	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.012
Cumulative	0.1353	0.406	0.6767	0.8571	0.9473	0.9834	0.9954
Stockout probability	0.8647	0.594	0.3233	0.1429	0.0527	0.0166	0.0046

Example

When three spares are stocked, the probability of a stockout is 0.1429.

K	0	1	2	3	4	5	6
P(X=k)	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.012
Cumulative	0.1353	0.406	0.6767	0.8571	0.9473	0.9834	0.9954
Stockout probability	0.8647	0.594	0.3233	0.1429	0.0527	0.0166	0.0046

When fewer units are stocked, stockout probabilities are higher.

Example

The stockout probability cannot exceed 0.06.

When three or fewer units are stocked	When four units are stocked
Stockout probability exceeds 0.06	Stockout probability = 0.0527 < 0.06
Inadequate	Adequate

Since the stockout probability should not exceed 0.06, the firm should stock 4 units.

Discussion

Duration: 5 minutes

- What are the commonly used discrete probability distributions?

Answer: The commonly used discrete probability distributions are the Bernoulli distribution, the Poisson distribution, and the Binomial distribution.

- What is a binomial distribution?

Answer: Binomial distribution is the probability distribution of the number of times an outcome occurs in a fixed number of trials.

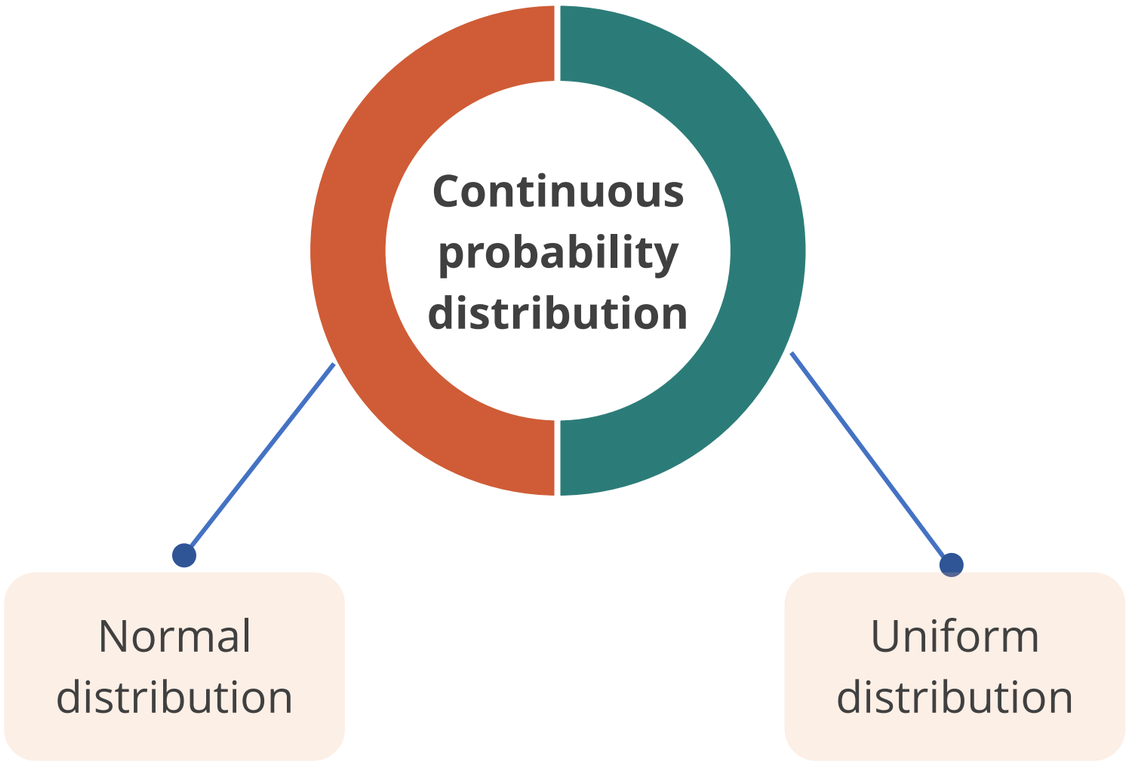
Commonly Used Continuous Probability Distributions

Discussion: Continuous Probability Distribution

Duration: 15 minutes

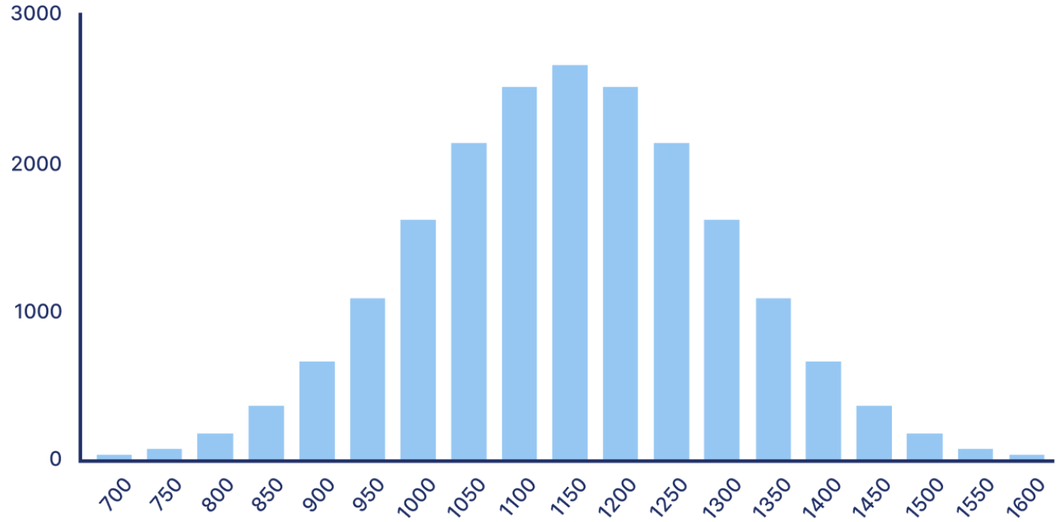
- What are the commonly used continuous probability distributions?
- What is a uniform probability distribution?

Continuous Probability Distribution



Normal Probability Distribution

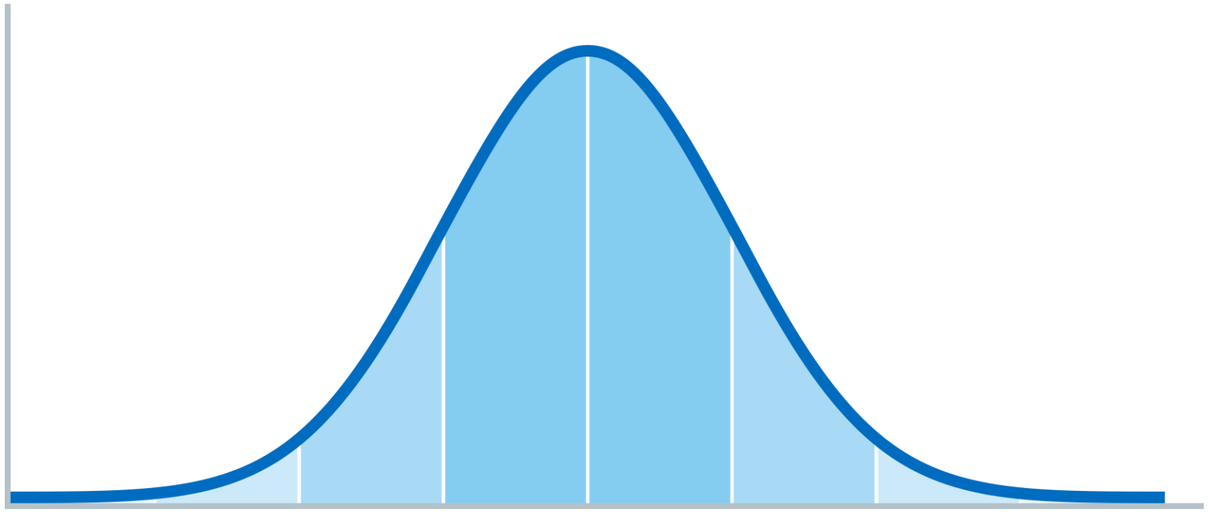
It is a continuous probability distribution that is widely used to describe many natural phenomena, such as height, weight, IQ scores, and many other measurements.



It is characterized by a bell-shaped curve that is symmetric around its mean.

Normal Probability Distribution

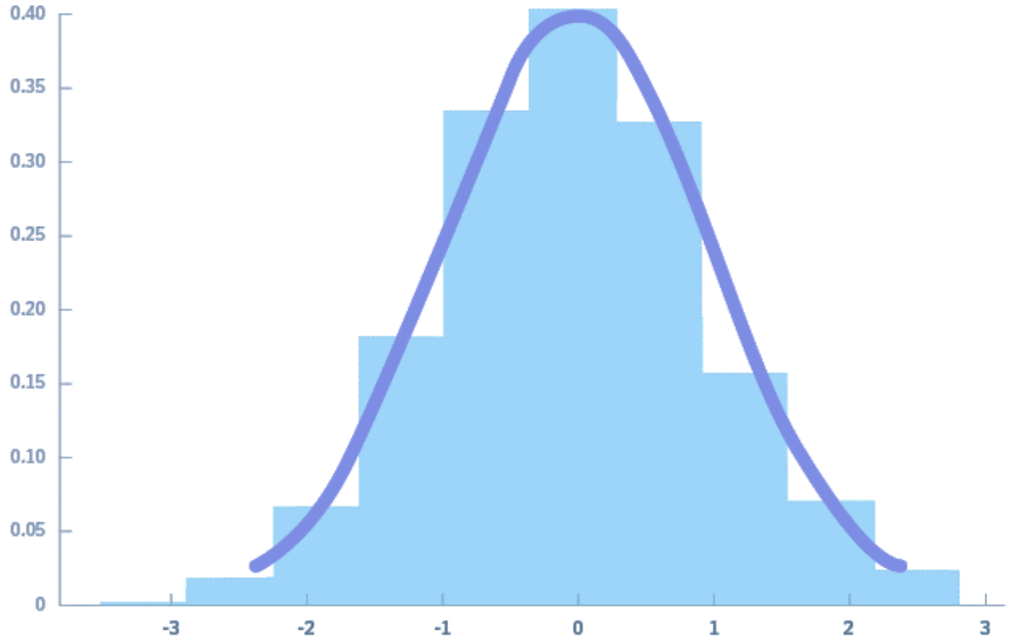
In a normal distribution, data is symmetrically distributed with no skew.



When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

Normal Probability Distribution

Normal distributions are also called Gaussian distributions or bell curves because of their shape.



Parameters of Normal Distribution

The normal distribution is defined by two parameters:

Mean (μ)

This represents the center of the distribution and determines where the bell curve is centered.

Standard deviation (σ)

This measures the spread or dispersion of the distribution. It determines the shape and width of the bell curve.

Properties of Normal Distribution

The properties of a normal distribution are:

The mean, median, and mode are exactly the same.

The distribution is symmetric. Half the values fall below the mean and half are above the mean.

The mean and the standard deviation are the two values that describe the distribution.

Normal Probability Distribution

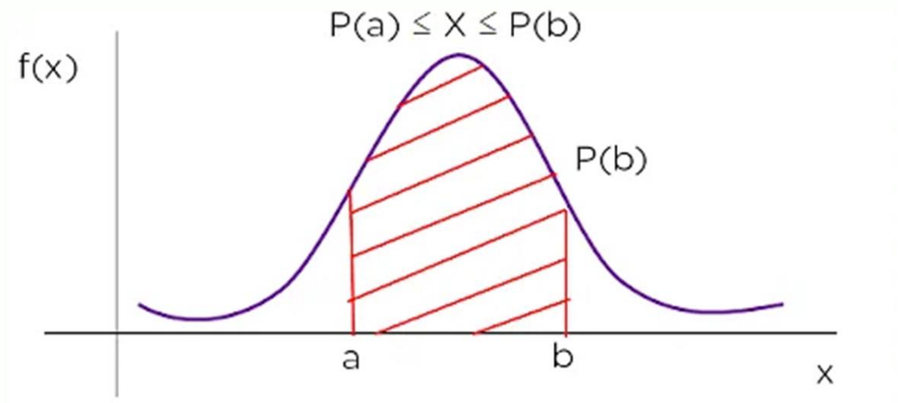
The equation for a normal distribution is given by:

$$f(x) = (1 / \sigma\sqrt{(2\pi)}) * e^{(-(x-\mu)^2 / (2\sigma^2))}$$

- x is the variable that represents a particular value within the distribution.
- μ (mu) is the mean of the distribution, representing the center of the curve.
- σ (sigma) is the standard deviation, which measures the dispersion of the data points around the mean.
- π is a mathematical constant representing pi (approximately 3.14159).
- e is Euler's number, a mathematical constant approximately equal to 2.71828.

Probability Density Function (PDF)

The probability density function (PDF) is a mathematical function that describes the probability distribution of a continuous random variable.



The PDF satisfies two conditions:

- It is non-negative for all possible x-values, and the integral of the PDF over the entire range of possible values equals 1.
- This ensures that the total probability of the distribution is normalized to 1.

The PDF is widely used in statistics and probability theory to analyze and make inferences about continuous random variables that follow a normal distribution.

Normal Distribution: Example

Consider a factory outlet that uses a filling machine to produce low-pressure oxygen shells

Assume that the filling machine has a mean output of $\mu = 500$ oxygen shells per hour and a standard deviation of $\sigma = 10$ shells per hour

- The goal is to determine the probability that the filling machine will produce between 490 and 510 oxygen shells in an hour.
- In order to find the probability, integrate the probability distribution function (PDF) of the normal distribution between the two values of interest.
- This integration represents the area under the curve, reflecting the probability of the machine output falling within that specific range.

Normal Distribution: Example

Step 1: Standardize the values

Standardize the values of 490 and 510 using the formula: $z = (x - \mu) / \sigma$

For 490: $z_1 = (490 - 500) / 10 = -1$

For 510: $z_2 = (510 - 500) / 10 = 1$

Normal Distribution: Example

Step 2: Calculate the cumulative probabilities

Calculate the cumulative probabilities corresponding to the standardized values using a standard normal distribution table

For $z = -1$, the cumulative probability is $P(Z \leq -1) \approx 0.1587$

For $z = 1$, the cumulative probability is $P(Z \leq 1) \approx 0.8413$

Normal Distribution: Example

Step 3: Calculate the desired probability

Subtract the cumulative probability of 490 from the cumulative probability of 510 to find the probability between these two numbers:

$$P(490 \leq x \leq 510) = P(Z \leq 1) - P(Z \leq -1)$$

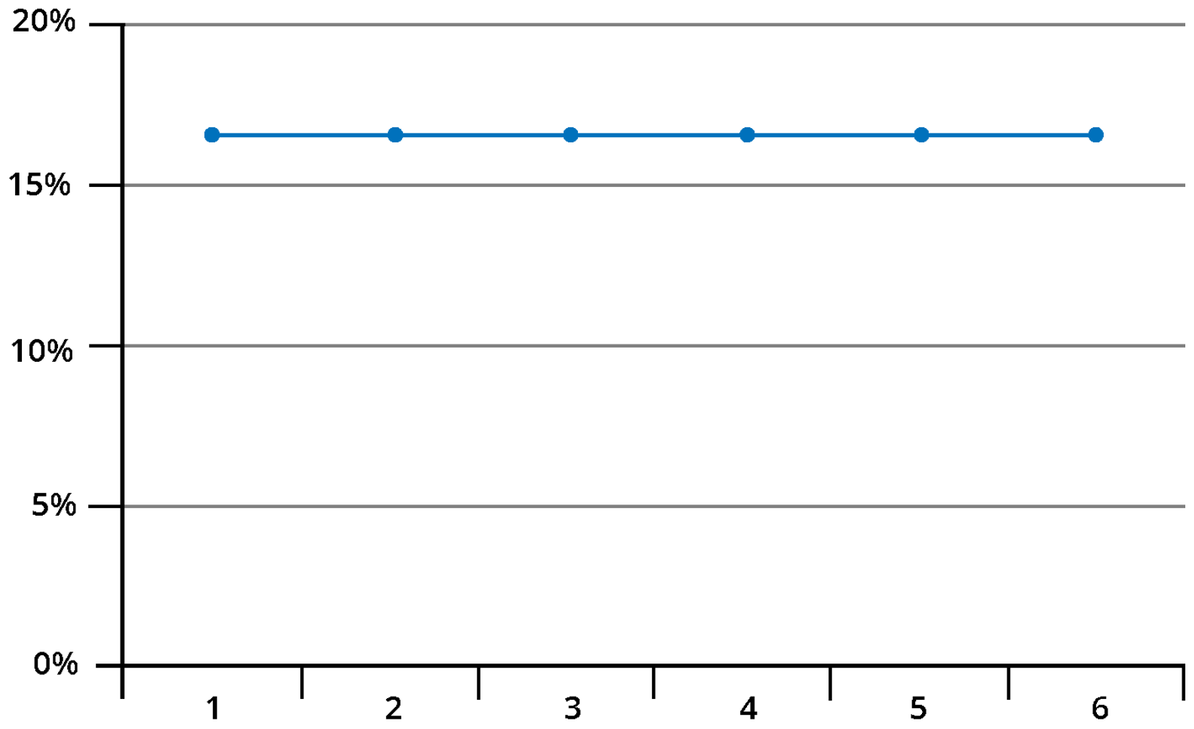
$$\approx 0.8413 - 0.1587$$

$$\approx 0.6826$$

There is approximately a 68.26% probability of the filling machine producing between 490 and 510 oxygen shells per hour.

Uniform Probability Distribution

A uniform probability distribution is a type of probability distribution where all values within a given range have an equal probability of occurring.



Uniform Probability Distribution

In statistics, a uniform probability distribution is a continuous probability distribution in which all values in a given range have an equal probability of occurring.

A deck of cards contains, uniform distributions because it is equally likely to draw a heart, club, diamond, or spade.

A coin has a uniform distribution because the probability of getting either heads or tails in a coin toss is the same.

Parameters of Uniform Distribution

The uniform distribution is defined by two parameters:

Lower bound (a)

This parameter represents the lower end, or the minimum value, of the range over which the uniform distribution is defined. All values within the range must be greater than or equal to this lower bound.

Upper bound (b)

This parameter represents the upper end, or the maximum value, of the range over which the uniform distribution is defined. All values within the range must be less than or equal to this upper bound.

Properties of Uniform Distribution

The properties of a uniform distribution are:

The probability density function (PDF) remains constant over the entire range. This means that all values within the range have an equal probability of occurring.

Every value within the specified range has the same probability of occurring. There are no peaks or valleys in the distribution, and the probabilities are evenly distributed across the interval.

A uniform distribution is defined over a specific range, typically denoted as [a, b]. All values within this range have a non-zero probability, while values outside the range have a probability of zero.

Uniform Probability Distribution

The equation for the continuous uniform probability distribution is as follows:

The probability density function (PDF) of the continuous uniform distribution for a random variable X defined over a continuous range [a, b] is as follows:

$$f(x) = 1 / (b - a), \text{ for } a \leq x \leq b$$

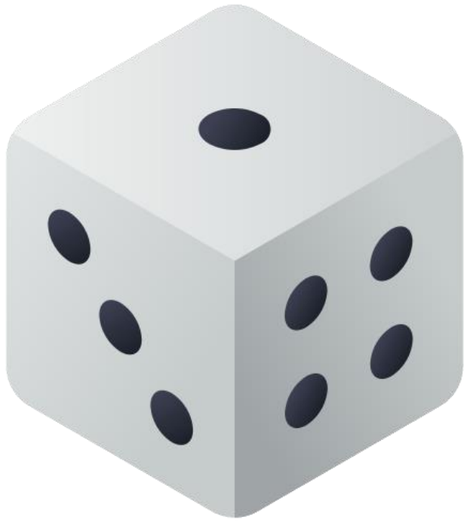
$$0 \text{ otherwise}$$

Where,

- a represents the lower bound of the range.
- b represents the upper bound of the range.
- (b - a) represents the width or length of the range.

Uniform Probability Distribution: Example

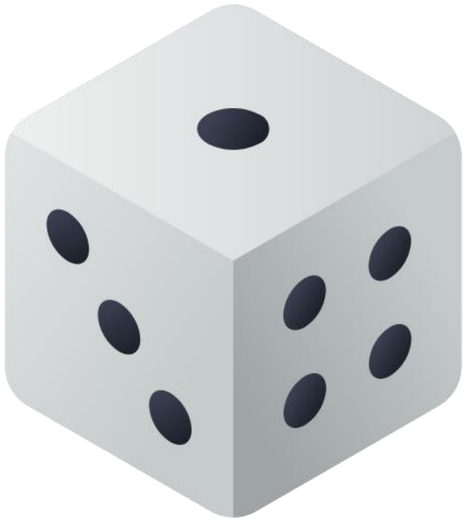
Consider the scenario of rolling a fair six-sided dice



- Each side of the dice has an equal chance of appearing.
- In this case, the range of possible outcomes is from 1 to 6.
- Therefore, a (lower bound) = 1 and b (upper bound) = 6.

Uniform Probability Distribution: Example

The probability density function (PDF) of the uniform distribution becomes:



$$f(x) = 1 / (6 - 1) = 1/5 \text{ for } 1 \leq x \leq 6$$
$$0 \text{ otherwise}$$

In this case, all numbers from 1 to 6 have an equal probability of 1/5 or 0.2, while values outside this range have a probability of 0.

Discussion

Duration: 15 minutes

- What are the commonly used continuous probability distributions?

Answer: The commonly used continuous probability distributions are the normal distribution and the uniform distribution.

- What is a uniform probability distribution?

Answer: A uniform probability distribution is a continuous probability distribution in which all values in a given range have an equal probability of occurring.

Key Takeaways

- Random variables and probability distributions are used when the plausible values and their probabilities are studied.
- The values of random variables and corresponding probabilities constitute probability distribution.
- The three types of discrete probability distributions are the binomial distribution, the Bernoulli distribution, and the Poisson distribution.
- Continuous probability distributions include uniform probability distributions and normal distributions.
- The two parameters of a uniform probability distribution are the lower bound (a) and the upper bound (b).

Knowledge Check

Knowledge Check

1

Which of the following random variables can take any value in a certain range?

A. Discrete random variables

B. Continuous random variables

C. Probability distribution

D. Continuous distribution

Knowledge Check

Which of the following random variables can take any value in a certain range?

A. Discrete random variables

B. Continuous random variables

C. Probability distribution

D. Continuous distribution

The correct answer is **B**

The continuous random variables can take any value in a certain range.

Knowledge Check 2

Which of the following are the parameters of a uniform probability distribution?

A. PDF and mean μ

B. Lower bound (a) and upper bound (b)

C. Mean μ and standard deviation σ

D. PDF and standard deviation σ

Knowledge Check 2

Which of the following are the parameters of a uniform probability distribution?

A. PDF and mean μ

B. Lower bound (a) and upper bound (b)

C. Mean μ and standard deviation σ

D. PDF and standard deviation σ

The correct answer is **B**

The lower bound (a) and upper bound (b) are the two parameters of a uniform probability distribution.

Knowledge Check 3

Which of the following is a distribution completely specified by its mean μ and standard deviation σ?

A. Normal distribution

B. Probability distribution

C. Continuous distribution

D. Uniform probability distribution

Knowledge Check 3

Which of the following is a distribution completely specified by its mean μ and standard deviation σ?

A. Normal distribution

B. Probability distribution

C. Continuous distribution

D. Uniform probability distribution

The correct answer is **A**

A normal distribution is completely specified by its mean and standard deviation.

Thank You