PROBABILITY DISTRIBUTION

Imagine you have a bag of differently colored marbles, and you want to understand how likely it is to pick each color. A probability distribution is like a chart or a set of rules that tells you the chances of picking each color.

If all the marbles are equally likely to be picked, you have a **uniform distribution**. Each color has the same chance.

If some colors are more common, you might have a distribution where certain colors have higher probabilities of being picked. For instance, you're more likely to pick red marbles than blue marbles.

On the other hand, some distributions might make it very unlikely to pick certain colors. Maybe there's a rare, special marble that's hard to get.

Now, think about this in terms of everyday situations:

- In weather forecasting, a probability distribution could describe the chances of different weather conditions (sunny, rainy, cloudy) on a given day.
- In finance, it could tell you the probability of making different amounts of money when investing in stocks.
- In a classroom, it might show the likelihood of getting different scores on a test.

In essence, a probability distribution is a **tool to understand and quantify uncertainty or randomness**. It helps us make informed decisions by knowing the odds of different outcomes in various situations.

DEFINITION

A probability distribution is a mathematical function that specifies the likelihood of different outcomes in a random experiment.

Probability distributions can be categorized into two primary groups based on the type of data they are intended to model:

1. **Discrete Distributions**: These distributions are used to model discrete data, where the random variable can only take on specific, distinct values with gaps in between. These values are typically finite and countable. Discrete distributions are often associated with situations involving counts, integers, or events that are distinct and separate from each other. Common examples include the Poisson distribution, binomial distribution, and geometric distribution. **A Probability Mass Function is used for discrete random variables.** Discrete random variables take on a countable number of distinct values, and the PMF assigns a probability to each of these values.

2. **Continuous Distributions**: Continuous distributions are designed for continuous data, where the random variable can take on an infinite number of possible values within a given range. These distributions are used when data can vary smoothly and can include any real number within an interval. Continuous distributions are characterized by probability density functions and are often used to model measurements or observations that are not restricted to specific values. The normal

(Gaussian) distribution, exponential distribution, and uniform distribution are examples of continuous distributions. **A Probability Density Function is used for continuous random variables**. Continuous random variables can take on an uncountable number of values within a range. Unlike the PMF, which assigns probabilities to specific values, the PDF assigns probabilities to intervals of values.

DESCRETE PROBABILITY DISTRIBUTION

1. BINOMIAL DISTRIBUTION

It models the probability of getting a certain number of successes (usually denoted as **r**) in a series of experiments or Bernoulli trials, where each trial has only two possible outcomes: success or failure. The term **binomial** comes from the fact that there are two possible outcomes.

The formula for finding **binomial probability** is given by:

$$P(X=r) = {}^{n}C_{r}(p)^{r}(1-p)^{n-r}$$

Where **n** is **the number of trials**, **p** is the **probability of success**, and **r** is the **number of successes after n trials**.

However, there are some **conditions** that need to be met in order for us to be able to apply the formula.

1. The **total number** of trials is **fixed** at **n**.
2. Each trial is **binary**, i.e., it has **only two possible outcomes:** success or failure.
3. **Probability of success** is the **same** in all trials, denoted by **p**.

   Bernoulli trail – random experiment with 2 possible outcomes

2. POISSON DISTRIBUTION

The Poisson distribution is a probability distribution that models the number of rare events occurring within a fixed time or space interval. It is used when you have a known average rate of occurrence, but the exact timing or count of events is uncertain.

The probability mass function (PMF) of the Poisson distribution is given by:

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

Key characteristics of a Poisson distribution include:

1. **Count of Rare Events:** It is used to model the count of rare events within a fixed time or space interval. These events are often rare, random, and independent of each other.
2. **Fixed Interval:** The distribution is defined for a fixed interval of time or space. For example, the number of phone calls received at a customer service center in an hour or the number of accidents at a specific intersection in a day.

3. **Constant Rate:** The average rate of events occurring within the interval is denoted as "λ" (lambda) and remains constant. Lambda represents the expected number of events in the given interval.

Example:

Suppose you manage a small coffee shop, and on average, you get 5 customers arriving every 15 minutes. You want to know the probability of getting exactly 7 customers in the next 15 minutes.

In this case, we have the following information:

- λ (average rate of customer arrivals) = 5 customers per 15 minutes.
- K = 7 (7 is rare or less likely since average is 5)

We want to find P(X=7), which is the probability of having 7 customers arrive in the next 15 minutes using the Poisson distribution formula:

$$\boxed{P(X = 7) = \frac{e^{-5} \cdot 5^7}{7!}}$$

3. GEOMETRIC DISTRIBUTION

The geometric distribution is a probability distribution that models the number of trials needed to achieve the first success in a series of independent trials. In other words, it helps us answer questions like "How many attempts does it take to succeed for the first time?"

Key Characteristics:

1. **Success on the nth Trial:** The geometric distribution focuses on the number of trials (usually denoted as "n") required to achieve the first success.
2. **Independent Trials:** Each trial is independent of the others, and the probability of success remains constant from trial to trial.
3. **Two Outcomes:** Each trial has only two possible outcomes: success or failure.

The Probability Mass Function (PMF) of the geometric distribution is given by:

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

Example:

Let's say you work in sales, and the probability of making a successful sales call to a potential client is 0.2 (20% chance of success). You want to know how many calls you'll need to make before achieving your first successful sale.

In this scenario:

- *p* (probability of a successful call) = 0.2.

We want to determine *P*(*X*=*n*), the probability of achieving your first successful sale on the nth sales call using the geometric distribution formula:

$$P(X=n)=(1-p)^{\wedge}(n-1) \text{ x } p$$

Now, let's calculate it step by step:

1. Calculate 1−*p*: 1−0.2=0.81
2. Calculate (0.8) ^(*n*−1) for the first *n*−1 failed calls.
3. Multiply the result by *p* to account for the successful call.

CONTINOUS PROBABILITY DISTRIBUTION

1. UNIFORM DISTRIBUTION

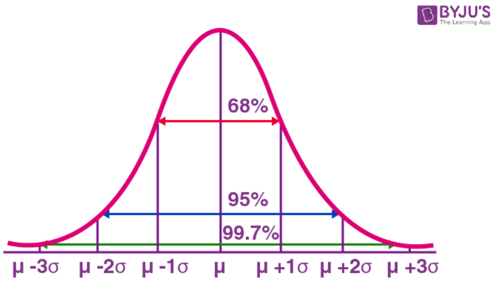
The uniform distribution, also known as the rectangular distribution, is a simple and fundamental probability distribution that assigns equal probability to all values within a specified range. It's characterized by its constant probability density within that range.

2. NORMAL (GAUSSIAN) DISTRIBUTION

The normal distribution, also known as the Gaussian distribution, is one of the most important and widely used probability distributions in statistics. It is characterized by its bell-shaped curve and is often used to model a wide range of natural phenomena, including measurements, errors, and various real-world data.

Key Characteristics:

1. **Symmetry:** The normal distribution is symmetric, with the mean (average) at the center of the curve. The mean, median, and mode are all equal and located at the peak of the curve.
2. **Bell-Shaped Curve:** The curve is bell-shaped and symmetrically tapers off as you move away from the mean. This shape is a result of the probability being higher for values close to the mean and decreasing as you move further away in either direction.
3. **68-95-99.7 Rule:** This rule states that for a dataset that follows a normal distribution,
   - Approximately 68% of the data falls within one standard deviation of the mean.
   - About 95% falls within two standard deviations.
   - Nearly 99.7% falls within three standard deviations.



The value of σ is an indicator of how wide the graph is. This will be true of any graph, not just a normal distribution. A **low** value of σ means that the graph is **narrow**, while a **high** value implies that the graph is **wider**. This is because a wider graph has more values away from the mean, resulting in a high standard deviation.

Many natural phenomena, such as the height of individuals in a population, errors in measurements, and the distribution of IQ scores, closely follow a normal distribution. This makes it a valuable tool for modeling and understanding a wide range of real-world data.

Formula for Normal Distribution

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

Where:

- µ (mu) is the mean.
- σ (sigma) is the standard deviation.
- π (pi) is a mathematical constant, approximately equal to 3.14159.
- e is the base of the natural logarithm, approximately equal to 2.71828.
- x is the value for which you want to find the probability.

3. STANDARD NORMAL DISTRIBUTION

To find the probability, you do not need to know the value of the mean or the standard deviation; just knowing the number of standard deviations away from the mean your random variable is suffices. That is given by: Standard Normal distribution.

The **standard normal distribution** is one of the forms of the normal distribution. It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one. In other words, a normal distribution with a mean 0 and standard deviation of 1 is called the standard normal distribution. Also, the standard normal distribution is centered at zero, and the standard deviation gives the degree to which a given measurement deviates from the mean.

The random variable of a standard normal distribution is known as the **standard score or a z-score or a standard normal variable**. It is possible to transform every normal random variable X into a z score using the following formula:

$$z = (X - \mu) / \sigma$$

Probability Density Function is given by the formula,

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}$$

Example: What is the probability of a normally distributed random variable lying within 1.65 standard deviations of the mean?

Answer: You have to find the probability of the variable lying between μ-1.65σ and μ+1.65σ. i.e. P(μ-1.65σ < X < μ+1.65σ). In terms of Z, this becomes P(-1.65 < Z < +1.65). This would be equal to P(1.65)- P(-1.65) = 0.95- 0.05 = 0.90.

PRACTICE QUESTIONS:

Comprehension 1:

Suppose a new **cancer treatment** has been discovered, claiming to increase the **one-year survival rate** for pancreatic cancer patients to **40%.** In other words, the probability that a patient suffering from pancreatic cancer would survive for at least one year after receiving this treatment is 40%. Suppose a hospital is planning to use this treatment for its pancreatic cancer patients.

Question1: The hospital has a total of 10 patients suffering from pancreatic cancer. What is the probability that exactly 4 of these patients would survive the first year after receiving this treatment?

Answer: There are a total of 10 patients and the probability of surviving the first year is 40%, or 0.4 for each of them. Hence, the probability of 4 patients surviving is given by P(X=4)= 10C4(0.4)^4(0.6)^6 = 0.251 or 25.1%.

Question 2: What is the probability that the number of patients that survive the first year after receiving the treatment would not be more than 2?

Answer: Let's define X as the number of patients that will survive the first year after treatment. Now, according to the question, you have to find the probability of that number being less than or equal to 2, i.e. P(X<=2). You know that P(X<=2) = P(X=0) + P(X=1) + P(X=2)
= 10C0(0.4)0(0.6)10+ 10C1(0.4)1(0.6)9+ 10C2(0.4)2(0.6)8= 0.167 or 16.7%

Comprehension 2:

There are two lottery contests going on. You have to participate in both these contests. The schemes of the two lotteries are as follows:

Tiger Lottery:

You have purchased a non-refundable ticket worth ₹5,000. If you lose, you'll get nothing. If you win, you'll get a cash prize of ₹60,000. A total of 20 participants (including you) have participated in the contest. There will be only one winner, where all participants have an equal chance of winning.

Society lottery:

An initial investment of ₹4,000 was required. There's a 30% chance of winning a cash prize of ₹25,000. If you win, you get the cash prize as well as your initial investment. If you lose, you'll get nothing.

Question 1: What is the expected profit/loss on average from the Tiger Lottery?

Answer: If you win, you get ₹60,000. But you also spent ₹5,000. So net profit on winning is ₹55,000. Similarly, in case you lose, it'll cost you just ₹5,000 which was your ticket price. Therefore,

1. X = {+55000,-5000}

2. P(X) = {1/20, 19/20}

Using above terms, E[X] =-2000

Question 2: What is the expected profit/loss on average from society's lottery?

The net profit on winning is:

-4000+25000+4000 = ₹25000. If you lose, it'll cost you =-4000+0 = -₹4000

Let's define the X:

1. X = {+25000,-4000}

2. P(X) = {0.3, 0.7}

3. E[X] = 25000(0.3) + (-4000) (0.7) = +4700

Comprehension 3:

The **normal distribution**, aka the **Gaussian distribution**, was discovered by **Carl Friedrich Gauss** in 1809. Gauss was trying to create a probability distribution for **astronomical errors**. Astronomical errors are the errors that were made by astronomers while observing phenomena such as distances in space.

For example, Gauss found that an astronomer trying to estimate the distance between Earth and Uranus always makes an error. This **error** is **normally distributed**, with **μ = 0 km** and **σ = 1,000 km**.

Question 1: Based on the information above, what is the probability of the astronomer overestimating the distance by 2,330 km or more?

Answer: Let’s define X as the astronomical error, which is normally distributed with mean 0 km and standard deviation 1,000 km. Now, you have to find the probability that X > 2330, i.e. P(X>2330). Converting this to Z, it becomes P(Z>2.33). Since P(Z<=2.33) + P(Z>2.33) = 1, P(Z>2.33) = 1- P(Z<2.33) = 1- 0.9901 = 0.0099 or 0.99%, which is approximately 1%.

Question 2: Hence, what is the probability that the astronomer under- or over-estimates the distance by less than 500 km?

Answer: Let’s define X as the astronomical error, which is normally distributed with mean 0 km and standard deviation 1,000 km. Now, you have to find the probability that -500 < X < 500, i.e. P(-500 < X < 500). Converting this to Z, it becomes P(-0.5 < Z < 0.5) = P(Z < 0.5)- P(Z <-0.5) = 0.6915- 0.3085 = 0.3830, or 38.30%.

Comprehension 4:

In the light of the heavy rains and the consequent floods in Kerala, the state government decided to find out how probable these types of rainfalls are during the month of August so that they can redesign the current state infrastructure and prepare themselves for such future events.

From historical data, it was found that the average rainfall received by Kerala in the month of August was 1600 mm with a standard deviation of 400 mm. Assuming that the rainfall data follows a normal distribution, answer the following questions.

Question 1: It was calculated that the floods were caused by a rainfall of 2200 mm. From the given information, what would be the probability that the state receives more than 2200 mm of rainfall in this period?

Answer: Let X be defined as the rainfall (measured in mm) with $\mu = 1600$ and $\sigma = 400$. Now for the given question, we need to calculate $P(X>2200)$. Converting it to Z we have to find $P(Z> (2200-1600)/400)$ or $P(Z>1.5)$. Now we have $P(Z>1.5)= 1- P(Z<=1.5)$.From the Z-table we have $P(X<=1.5) = 0.9332$. Thus $P(X>1.5) = 1- 0.9332= 0.0668$= 6.7%

Question 2: At what cutoff rainfall value should the current infrastructure be redesigned so that there is only a 3% chance that either similar or heavier rains are observed in the future?

Answer: Let X be defined as the rainfall and the cutoff rainfall be denoted by x. Given that $\mu= 1600$ and $\sigma= 400$. Thus, the Z-value corresponding to x would be $(x-1600)/400$. Let it be denoted by z. Now as per the question $P(Z>=z)= 0.03$. This can be written as $1-P(Z<z) =0.03$ or $P(Z<z)= 0.97$. Then from the tables, you get the value of z that satisfies the previous equation as 1.88. Thus, you have $(x-1600)/400 =1.88$ or $x =1600+1.88*400 = 2352$