

Statistics

Topics	Definition	Formulae
What is statistics	Is a branch of mathematics. it is a collecting, analyzing and Interpreting the large amount of data.	--
Why stats is important	It is used for prediction, decision and classifications .etc	--
Where stats is used	In Medical research, Stock market, sales projection and Weather forecasting.etc	Ex:1 drugs &vaccine 2 Investing. 3 weather conditions
Types of Statistics	Two types → 1.Descriptive 2. Inferential Statistics	
1.Descriptive Stats	It is Describes, Organize and Summarize information about Entire data population .	Ex: Satisfaction of all customers
2.Inferential Stats	It is generalize about a population based on sample of data	Ex:Satisfaction of sample(few) customer = all customers Satisfaction
Sampling techniques Or methods	Is a procedure for selecting sample members from a population .they are two types... 1.Probability sample 2.Non Probability sample	--

Sampling Method/Sampling Techniques

Probability or Random

Non-Probability or Non-random

Simple Random Sampling

Systematic Sampling

Stratified Sampling

Cluster Sampling

Area Sampling

Multi-Stage Sampling

Judgement Sampling

Convenience Sampling

Quota Sampling

Panel Sampling

Snowball Sampling

Types of Sampling Method Simple Techniques and Examples

Probability Sampling Method

Non-probability Method

Simple Random Sampling

Every member of the population has an equal chance of being selected.



Convenience Sample

It includes the individuals who are most accessible to the researcher.

Systematic Sampling

Individuals of the population are chosen at regular intervals. It is easier to conduct than simple random method.



Voluntary Response

Here people volunteer themselves, instead of researchers choosing individuals.

Stratified Sampling

When the population shows mixed character then this method is used. The population divides into subgroups.



Purposive Sampling

Researchers use judgements to select a sample that is most useful for research.

Cluster Sampling

Instead of sampling individuals from subgroups, the subgroups are randomly selected.



Snowball Sampling

In this sampling, the number of people who have access to "snowballs" as you come in contact with more people.

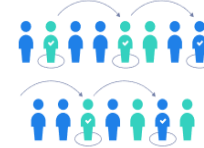
www.homeworkjoy.com

homeworkjoy
Study with satisfaction

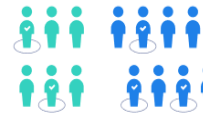
Simple random sample



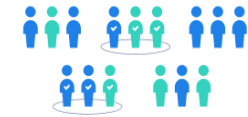
Systematic sample



Stratified sample



Cluster sample

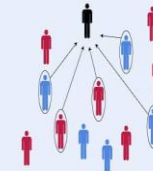


Scribbr

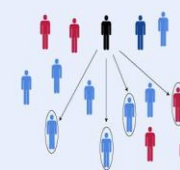
Convenience sample



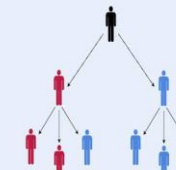
Voluntary response sample



Purposive sample

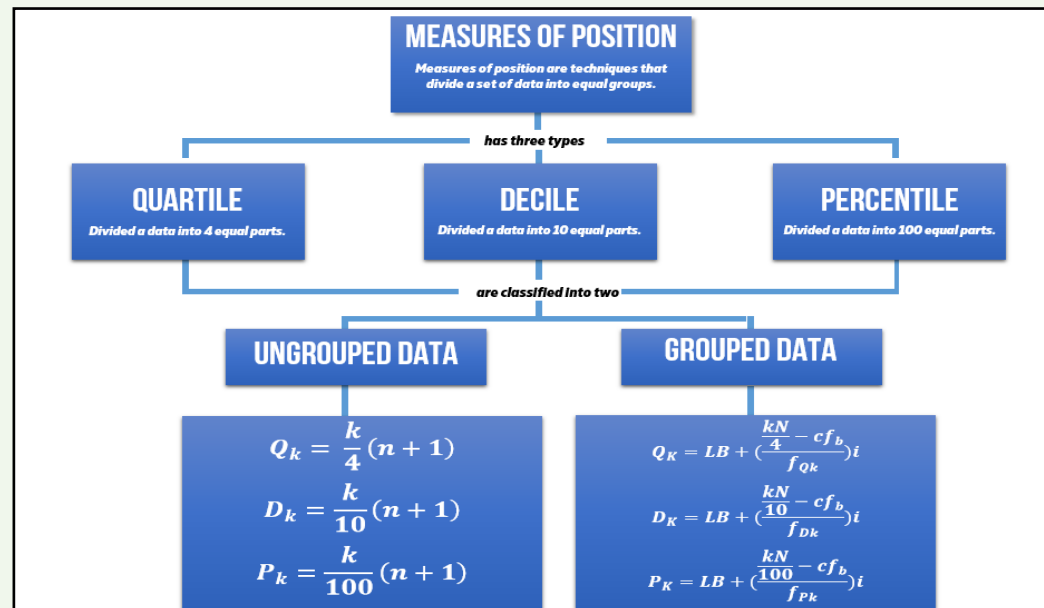


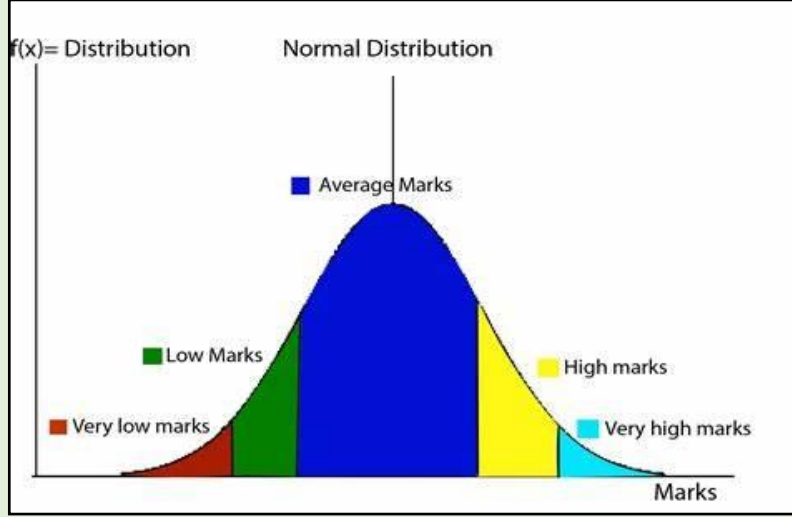
Snowball sample



Median	Middle value of the data set Note : first sort the element If the data set I even find avg of middle two number and divide it 2	Even=(n/2) observation Odd=[(n+1)/2]observation
Mode	Most occurred number is called mode	--
Range	Difference between highest number and lowest number	Max-min
Standard deviation	Is a measure of how dispersed the data is in relation to the mean	<div style="text-align: center;"> <h3>Standard Deviation</h3> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <u>Sample</u> $S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ </div> <div style="text-align: center;"> <u>Population</u> $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ </div> </div> </div>
Variance	The value of variance is equal to the square of standard deviation	$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$

Percentiles
Quartiles
Decile



Outliers in Data	An outlier is a data point that differs significantly from other observations	
Skewness for the data	Refers to a distortion or asymmetry that deviates from the Symmetrical bell curve or normal distribution in set of data	
The normal curves		
Z –text	Z Test is the statistical hypothesis which is used in order to determine that whether the two samples means calculated are different in case the standard deviation is available	
T –tests	sample is large whereas the T test is used in order to determine a how averages of different data sets differs from each other in case standard deviation or the variance is not known.	

Comparison of z- and t-Tests

- One-sample z-test
- Sample is drawn from a population whose mean is known or hypothesized
- σ^2 is known, compute σ_M

$$\sigma_M = \sqrt{\frac{\sigma^2}{n}}$$

- $z = (M - \mu) / \sigma_M$

- One sample t-test
- Sample is drawn from a population whose mean is known or hypothesized
- σ^2 not known
 - Use s^2 to compute s_M

$$s^2 = \frac{SS}{df} \quad \text{so} \quad s_M = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{SS/df}{n}}$$

- $df = n - 1$
- $t = (M - \mu) / s_M$

Advanced Statistics

1. Binomial Distribution

Binomial Distribution

- The number of trials must be fixed. ✓
- Each trial must have the same two possible outcomes. ✓
- The trials must be independent. ✗
- The probability of success must be the same in each trial.

Binomial Distribution Formula

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

where

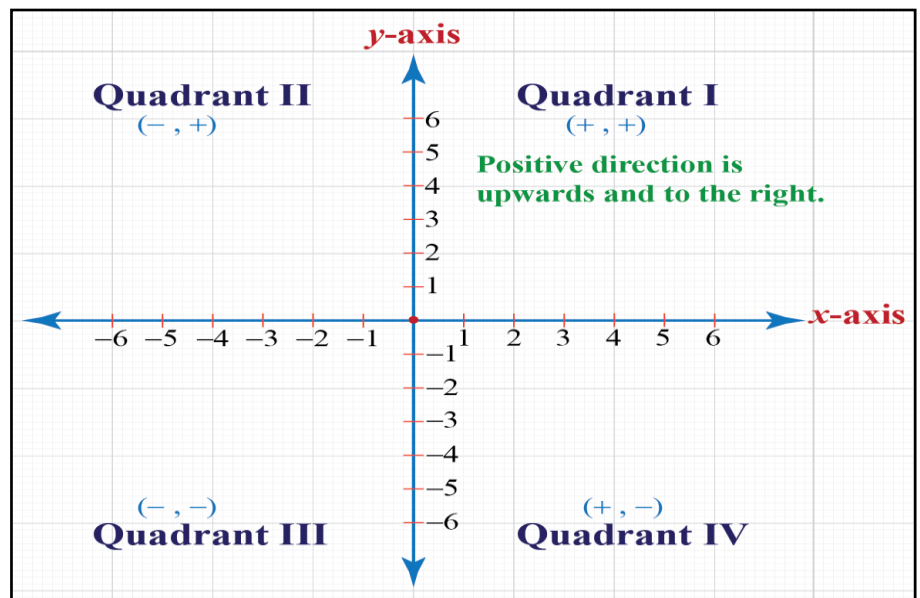
n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

2. Quadrants



3. Pearsons correlation

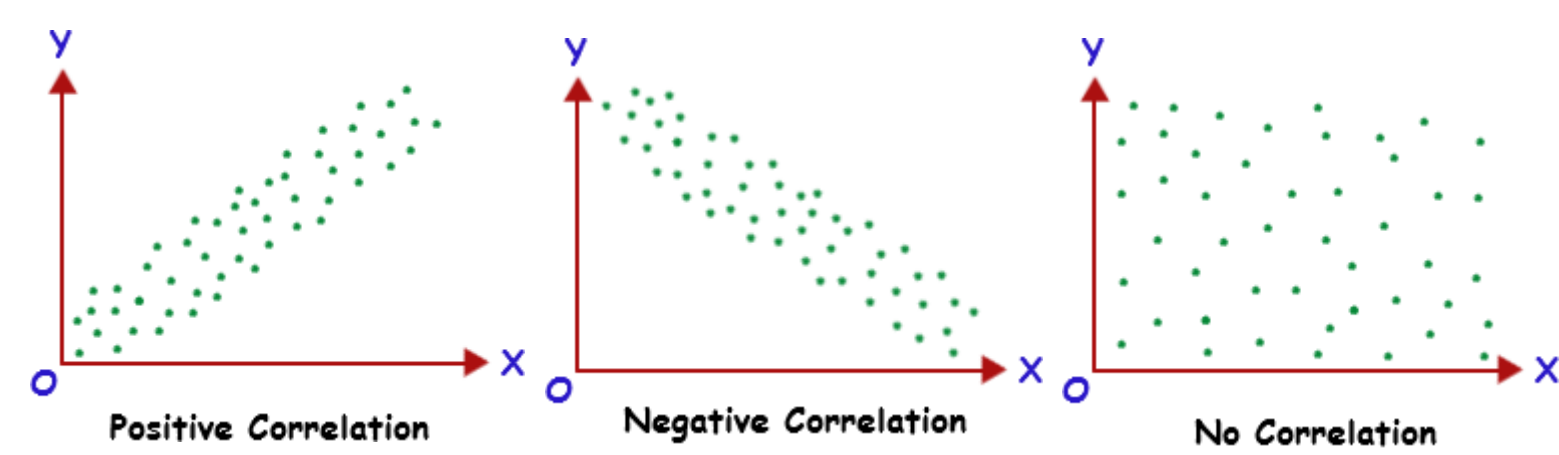
The name correlation suggests the relationship between two variables as their Co-relation. The correlation coefficient is the measurement of correlation. To see how the two sets of data are connected.

Positive Correlation

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Where,
 r = Pearson correlation coefficient
 x = Values in the first set of data
 y = Values in the second set of data
 n = Total number of values.

The Pearson correlation coefficient is denoted by the letter “ r ”. The formula for Pearson correlation coefficient r is given by:



4. Hypothesis testing with Pearson's r

Recall that the Pearson r statistic tells us how much and in what way two measured variables are related. We can also use this statistic to conduct hypothesis tests about population correlation values.

The population correlation value is indicated by ρ (Greek letter *rho* corresponding to the sample r). This would be the correlation if the entire population provided scores on the two measured variables you are interested in.

This means that we can state a null and alternative hypothesis for the population correlation ρ based on our predictions for a correlation. Let's look at how this works in an example.

Suppose that we wanted to know if students who live near campus have higher GPAs than students who live farther away and commute to campus. We could measure students' GPAs and also measure how far away they live by measuring the distance to their residence from the middle of the quad. These are the two measured variables we're interested in.

$$H_0; \rho = 0$$
$$H_1; \rho \neq 0$$

5.Spearman correlation

Spearman's Correlation is the **feature selection method**. Spearman's Correlation determines the strength and direction of the monotonic relationship between your two variables. What is a Monotonic Relationship? when the value of one variable increases the values of another variable is also increases or vice versa but not in a linear manner

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = Difference between the two ranks of each observation

n = Number of observations

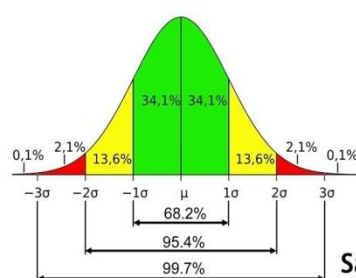
The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

6.Central limit theorem

The **Central Limit Theorem defines** that the mean of all the given samples of a population is the same as the mean of the population (approx) if the sample size is sufficiently large enough with a finite variation. It is one of the main topics of **statistics**

Central Limit Theorem



(Probability
&
Statistics)

Sample Standard
Deviation $(\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$

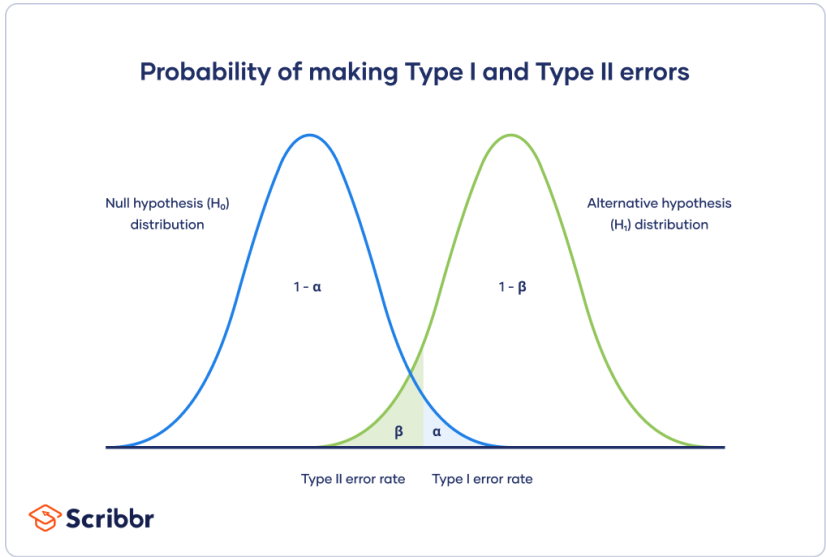
7. Type I and Type II Errors

In [statistics](#), a **Type I error** is a false positive conclusion, while a **Type II error** is a false negative conclusion.

Example: Type I vs Type II error

You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

- **Type I error (false positive):** the test result says you have coronavirus, but you actually don't.
- **Type II error (false negative):** the test result says you don't have coronavirus, but you actually do.

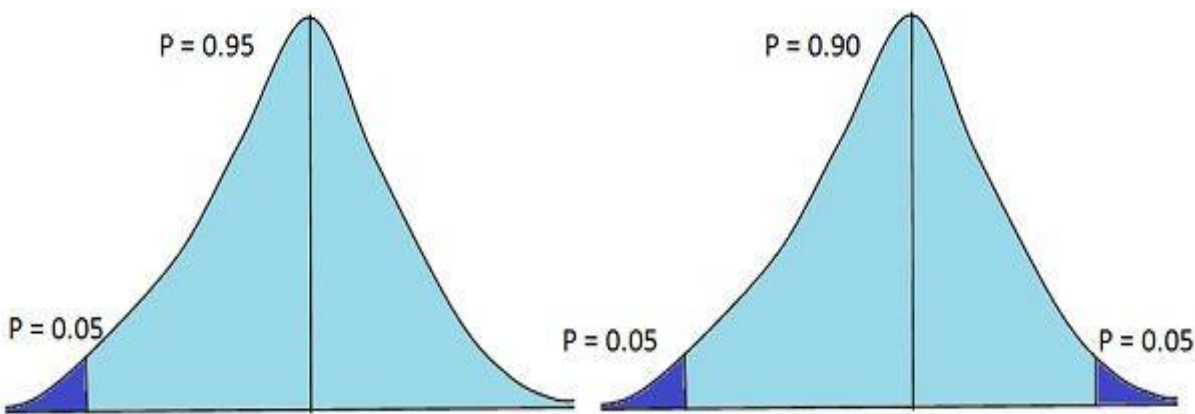


		Reality	
		True	False
Measured or Perceived	True	Correct 😊	Type 1 error False Positive
	False	Type 2 error False Negative	Correct 😊

8. One tailed & Two tailed

Comparison Chart

BASIS OF COMPARISON	ONE-TAILED TEST	TWO-TAILED TEST
Meaning	A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test.	A significance test in which alternative hypothesis has two ends, is called two-tailed test.
Hypothesis	Directional	Non-directional
Region of rejection	Either left or right	Both left and right
Determines	If there is a relationship between variables in single direction.	If there is a relationship between variables in either direction.
Result	Greater or less than certain value.	Greater or less than certain range of values.
Sign in alternative hypothesis		



One-tailed Test Vs Two-tailed Test

Probability

1. Some basic Probability Rules

Rule 1: The probability of any event E is a number between and including 0 and 1. $0 \leq P(E) \leq 1$

Rule 2: If an event E cannot occur, which means the event E is not in the sample space, then its probability is 0.

$$P(E) = 0$$

Rule 3: If an event E is certainly occur, then its probability is 1.

$$P(E) = 1$$

Rule 4: Let a sample space consist of n distinct events

$$s_1, s_2, s_3, \dots, s_n.$$

Then the sum of the probabilities of all those events is 1

$$P(s_1) + P(s_2) + P(s_3) + \dots + P(s_n) = 1$$

8

2. Addition rule

The addition rule states the probability of two events is the sum of the probability that either will happen minus the probability that both will happen.

Key Points

- The addition rule is: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- The last term has been accounted for twice, once in $P(A)P(A)$ and once in $P(B)P(B)$, so it must be subtracted once so that it is not double-counted.
- If AA and BB **are disjoint**, then $P(A \cap B) = 0$, so the formula becomes

$$P(A \cup B) = P(A) + P(B).$$

3. Multiplication Rule

The multiplication rule can be written as: $P(A \cap B) = P(B) \cdot P(A|B)$ and $P(A \cap B) = P(A) \cdot P(B|A)$.

Independent Events

Two events are independent if any of the following are true:

1. $P(A|B) = P(A)P(A|B) = P(A)$
2. $P(B|A) = P(B)P(B|A) = P(B)$
3. $P(A \text{ and } B) = P(A) \cdot P(B)$

4. Permutations

A permutation is an arrangement of objects in a definite order. The members or elements of sets are arranged here in a sequence or linear order. For example, the permutation of set $A=\{1,6\}$ is 2, such as $\{1,6\}$, $\{6,1\}$. As you can see, there are no other ways to arrange the elements of set A

Formula

$$P(n, r) = n(n-1)(n-2)(n-3)\dots\dots\dots\text{up to } r \text{ factors}$$

$$P(n, r) = n(n-1)(n-2)(n-3)\dots\dots\dots(n-r+1)$$

$${}_nP_r = \frac{n!}{(n-r)!}$$

Here, “ ${}_nP_r$ ” represents the “n” objects to be selected from “r” objects without repetition, in which the order matters.

5. Combination

In simple words, combination involves the **selection of events out of a larger group where order doesn't matter**. To calculate the probability of a combination, you will need to consider the number of favorable events over the number of total events. Combinations are used to calculate events where the order of events does not matter.

Combination Formula

$${}_nC_r = \frac{n!}{(n-r)!r!}$$

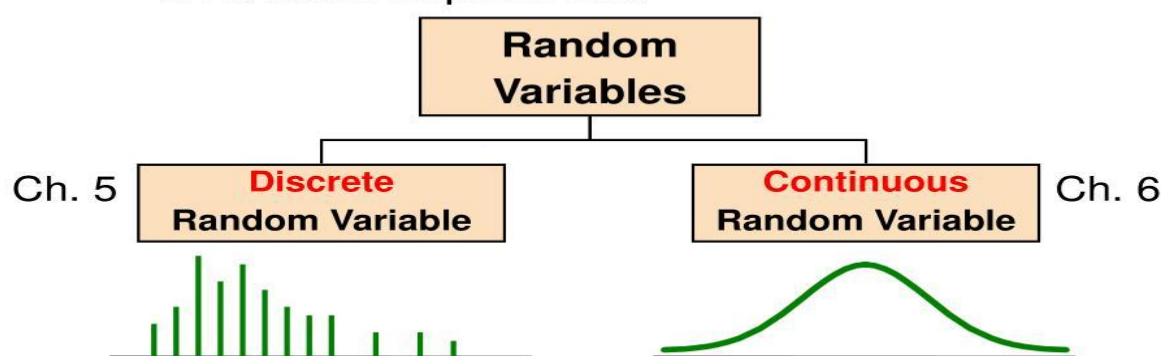
6. Random Variables

is a mathematical formalization of a quantity or object which depends on random events.

Introduction to Probability Distributions

▪ Random Variable

- Represents a possible numerical value from a random experiment



Chap 5-3

Discrete vs Continuous Variables

- **Discrete Variables:**
Can take on only certain values along an interval
 - the number of sales made in a week
 - the volume of milk bought at a store
 - the number of defective parts
- **Continuous Variables:**
Can take on any value at any point along an interval
 - the depth at which a drilling team strikes oil
 - the volume of milk produced by a cow
 - the proportion of defective parts

© 2002 The Wadsworth Group

7. discrete probability distribution

These distributions model the probabilities of random variables that can have discrete values as outcomes.

Examples: Bernoulli, Binomial, Negative Binomial, Hypergeometric, poisson, etc.,

1. Bernoulli Distribution

This distribution is generated when we perform an experiment once and it has only two possible outcomes – success and failure.

$$f(x) = \begin{cases} p^x * (1 - p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

2. Binomial Distribution

This is generated for random variables with only two possible outcomes.

$$\begin{aligned} P_n(x) &= C(n, x) p^x q^{n-x} \\ &= \frac{n!}{x! (n - x)!} p^x q^{n-x} \end{aligned}$$

3. Poisson Distribution

This distribution describes the events that occur in a fixed interval of time or space.

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$x = 0, 1, 2, 3, \dots$

λ = mean number of occurrences in the interval

e = Euler's constant ≈ 2.71828

4. Geometric Distribution

This is a special case of the negative binomial distribution where the desired number of successes is 1. It measures the number of failures we get before one success.

Geometric Probability Distribution
$$P(X = n) = p(1 - p)^{n-1}$$
$$P(X > n) = (1 - p)^n$$
$$\text{Mean } \mu = \frac{1}{p}$$
$$\text{Variance } \sigma^2 = \frac{1-p}{p^2}$$

p – probability of success
 n – number of first successful trial

8. Continuous probability distributions

These distributions model the probabilities of random variables that can have any possible outcome.

Examples: Normal, Student's T, Chi-square, Exponential, etc.,

9. Mean of discrete random variable

➤ Calculating the **Mean of a Discrete Random Variable**

Below is the probability distribution for a golfer on a par 3 hole.

x = Number of Strokes to Complete Course

x	$P(x)$	$x * P(x)$
1	0.10	$1 * 0.10 = 0.10$
2	0.30	$2 * 0.30 = 0.60$
3	0.45	$3 * 0.45 = 1.35$
4	0.15	$4 * 0.15 = 0.60$
		$\mu_x = 2.65$

Mean Formula:

$$\mu_x = \sum [x * P(x)]$$

10. variance of discrete random variable

The variance of a discrete random variable is given by:

$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 f(x_i)$ The formula means that we take each value of x , subtract the expected value, square that value and multiply that value by its probability. Then sum all of those values.

$$\text{Var}(X) = \sum_{i=1}^k (x_i - E(X))^2 p_i$$

11. Standard Deviation of discrete random variable

For a discrete random variable the standard deviation is calculated by **summing the product of the square of the difference between the value of the random variable and the expected value**, and the associated probability of the value of the random variable, taken over all of the values of the random variable, and finally taking the square root.

The **standard deviation** of a discrete random variable is given by

$$\sigma = \sqrt{\sigma^2}.$$

Example:

Find the standard deviation of the probability distribution for the sum of the two spins. The variance is 0.376.

x	$P(x)$	$x - \mu$	$(x - \mu)^2$	$P(x)(x - \mu)^2$
2	0.0625	-1.5	2.25	0.141
3	0.375	-0.5	0.25	0.094
4	0.5625	0.5	0.25	0.141

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{0.376} \approx 0.613\end{aligned}$$

Most of the sums differ from the mean by no more than 0.6 points.