PROBABILITY AND STATISTICS

SYLLABUS

- DESCRIPTIVE STATISTICS AND METHODS FOR DATA SCIENCE
- PROBABILITY
- PROBABILITY DISTRIBUTIONS
- ESTIMATION AND TESTING OF HYPOTHESIS (LARGE SAMPLE TEST)
- SMALL SAMPLE TESTS

Descriptive Statistic

- In Descriptive statistics, we are describing our data with the help of various representative methods like by using charts, graphs, tables, excel files etc. In descriptive statistics, we describe our data in some manner and present it in a meaningful way so that it can be easily understood.
- Most of the times it is performed on small data sets and this analysis helps us a lot to predict some future trends based on the current findings. Some measures that are used to describe a data set are measures of central tendency and measures of variability or dispersion.
- Types of Descriptive statistic:

➢ 1 Measure of central tendency

➢ 2 Measure of variability

❖ **Measure of central tendency**: It represents the whole set of data by single value .It gives us the location of central points

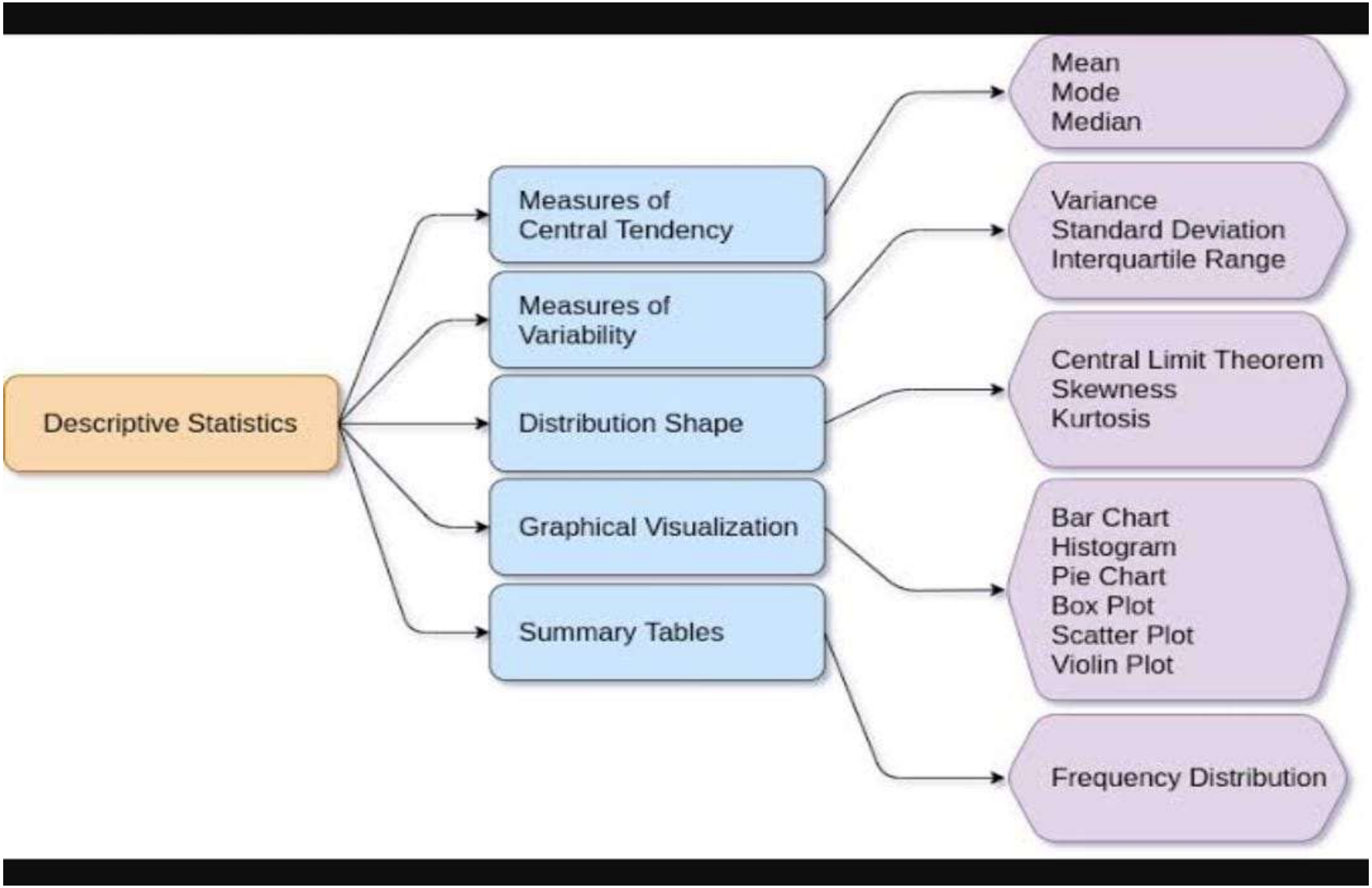
❖. There are three main measures of central tendency:

- Mean
- Mode
- Median

❖**Measure of variability**: Measure of variability is known as the spread of data or how well is our data is distributed.

❖The most common variability measures are:

- Range
- Variance
- Standard deviation



Descriptive Statistics

Measures of Central Tendency

Measures of Variability

Distribution Shape

Graphical Visualization

Summary Tables

Mean
Mode
Median

Variance
Standard Deviation
Interquartile Range

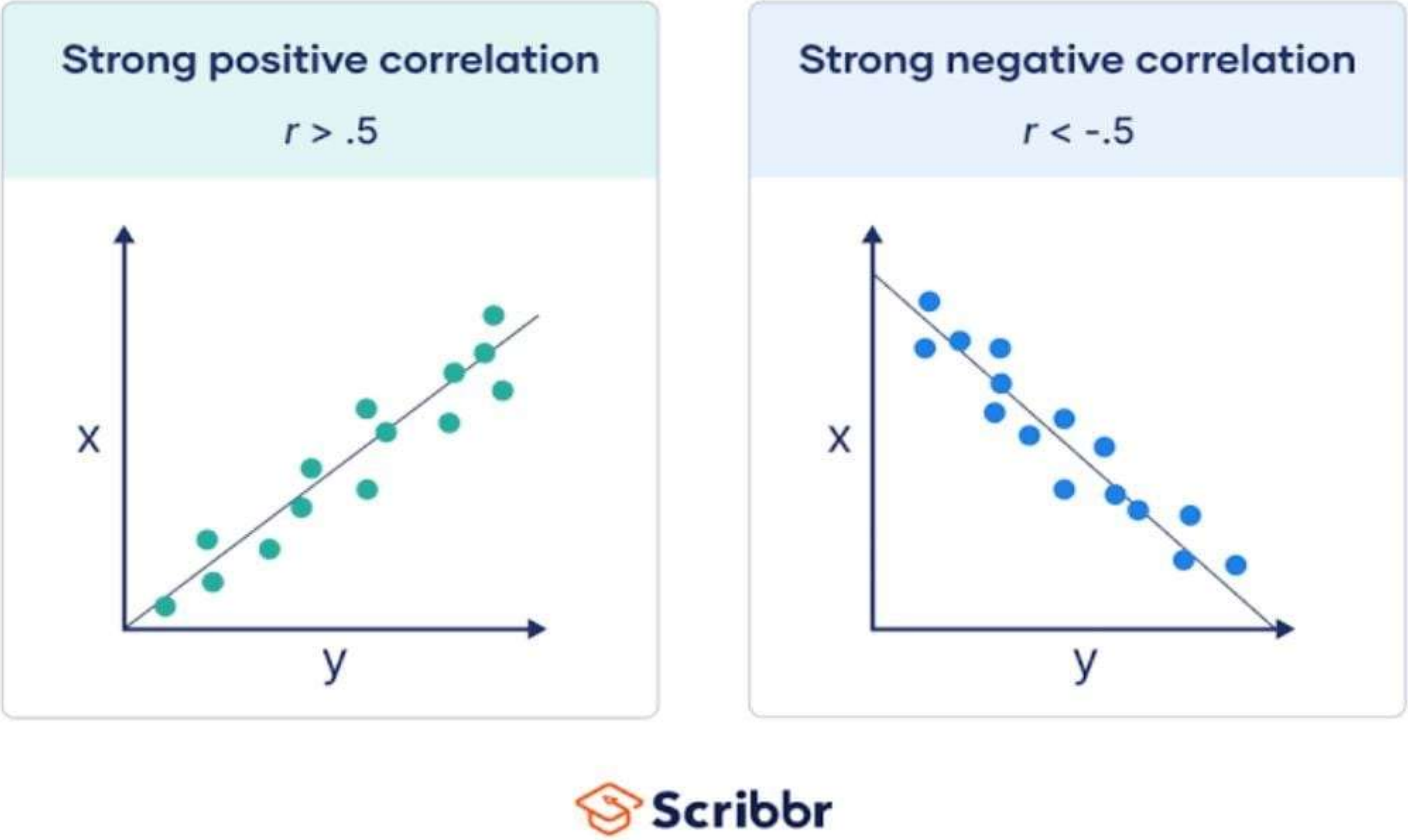
Central Limit Theorem
Skewness
Kurtosis

Bar Chart
Histogram
Pie Chart
Box Plot
Scatter Plot
Violin Plot

Frequency Distribution

CORRELATION

- In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Although in the broadest sense, "correlation" may indicate any type of association, in statistics it normally refers to the degree to which a pair of variables are linearly related. Familiar examples of dependent phenomena include the correlation between the height of parents and their offspring, and the correlation between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so-called demand curve.



Strong positive correlation

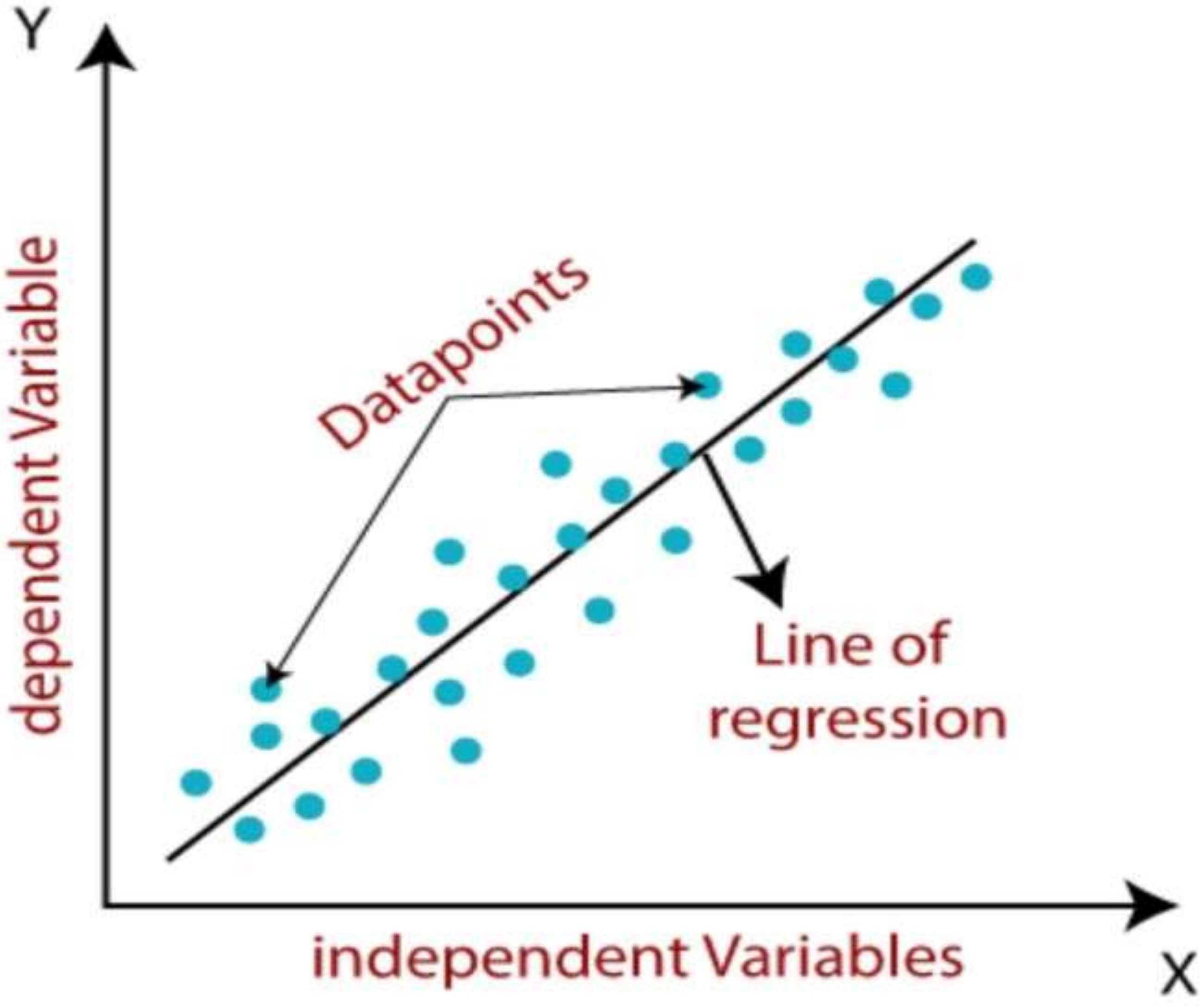
$r > .5$

Strong negative correlation

$r < -.5$

REGRESSION

- A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables.
- A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.
- It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line.
- Regression helps economists and financial analysts in things ranging from asset valuation to making predictions.
- In order for regression results to be properly interpreted, several assumptions about the data and the model itself must hold.


Y
dependent Variable
Datapoints
Line of regression
independent Variables
X

PROBABILITY ,RANDOM VARIABLES &DISTRIBUTION FUNCTIONS

- Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event.
- The value is expressed from zero to one. Probability has been introduced in Maths to predict how likely events are to happen.
- The meaning of probability is basically the extent to which something is likely to happen.
- This is the basic probability theory, which is also used in the probability distribution, where you will learn the possibility of outcomes for a random experiment
- .To find the probability of a single event to occur, first, we should know the total number of possible outcomes.

✓Probability = Number of Favorable Outcomes / Total Number of Outcomes

Or

$$P(A) = f / N$$

❖ Where:

▪P(A) = Probability of an event (event A) occurring

▪f = Number of ways an event can occur (frequency)

▪N = Total number of outcomes possible

RANDOM VARIABLES

- **A random variable** (also called random quantity, aleatory variable, or stochastic variable) is a mathematical formalization of a quantity or object which depends on random events. It is a mapping or a function from possible outcomes in a sample space to a measurable space, often the real numbers.
- **EX :**In an experiment a person may be chosen at random, and one random variable may be the person's height. Mathematically, the random variable is interpreted as a function which maps the person to the person's height. Associated with the random variable is a probability distribution that allows the computation of the probability that the height is in any subset of possible values, such as the probability that the height is between 180 and 190 cm, or the probability that the height is either less than 150 or more than 200 cm.

DISTRIBUTION FUNCTION OF A RANDOM VARIABLE

- In probability theory and statistics, the cumulative distribution function (CDF) of a real-valued random variable **X ,**or just **distribution function of X,** evaluated at x, is the probability that **X** will take a value less than or equal to x

❖ **F(x)=p(X<=x)**

F(x)=function of x

X=real value variable

p=probability that X will have a value <= x

PROBABILITY DISTRIBUTIONS

- In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

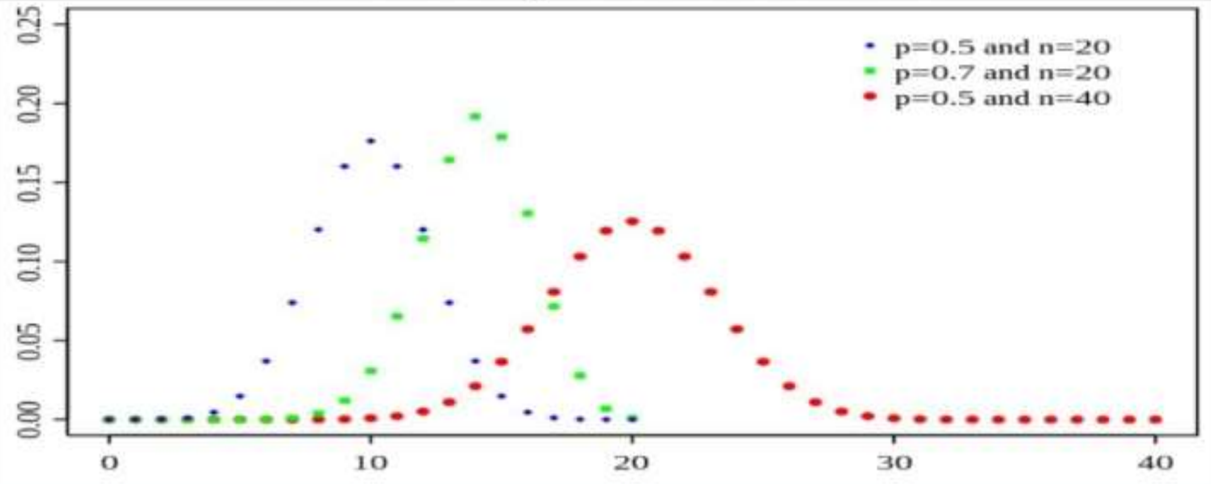
- For instance, if X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 (1 in 2 or 1/2) for X = heads, and 0.5 for X = tails (assuming that the coin is fair). Examples of random phenomena include the weather conditions at some future date, the height of a randomly selected person, the fraction of male students in a school, the results of a survey to be conducted, etc.

BINOMIAL DISTRIBUTION

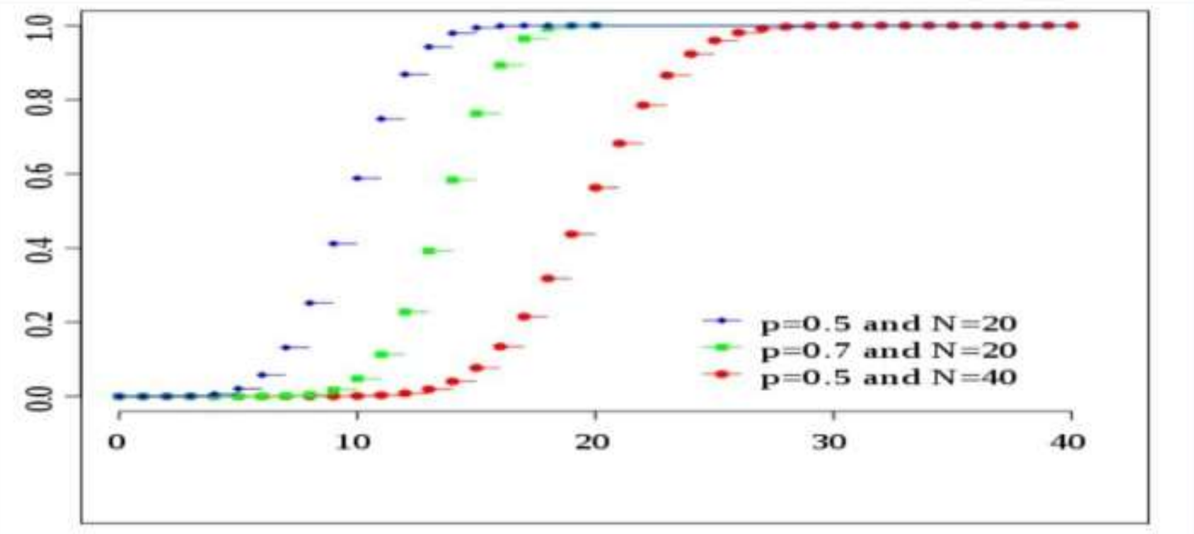
- In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure **q=1-p.** A single success/failure experiment is also called a **Bernoulli trial or Bernoulli experiment**, and a sequence of outcomes is called a **Bernoulli process**; for a single trial, i.e., n = 1, the binomial distribution is a **Bernoulli distribution**. The binomial distribution is the basis for the popular binomial test of statistical significance.

Binomial distribution

Probability mass function

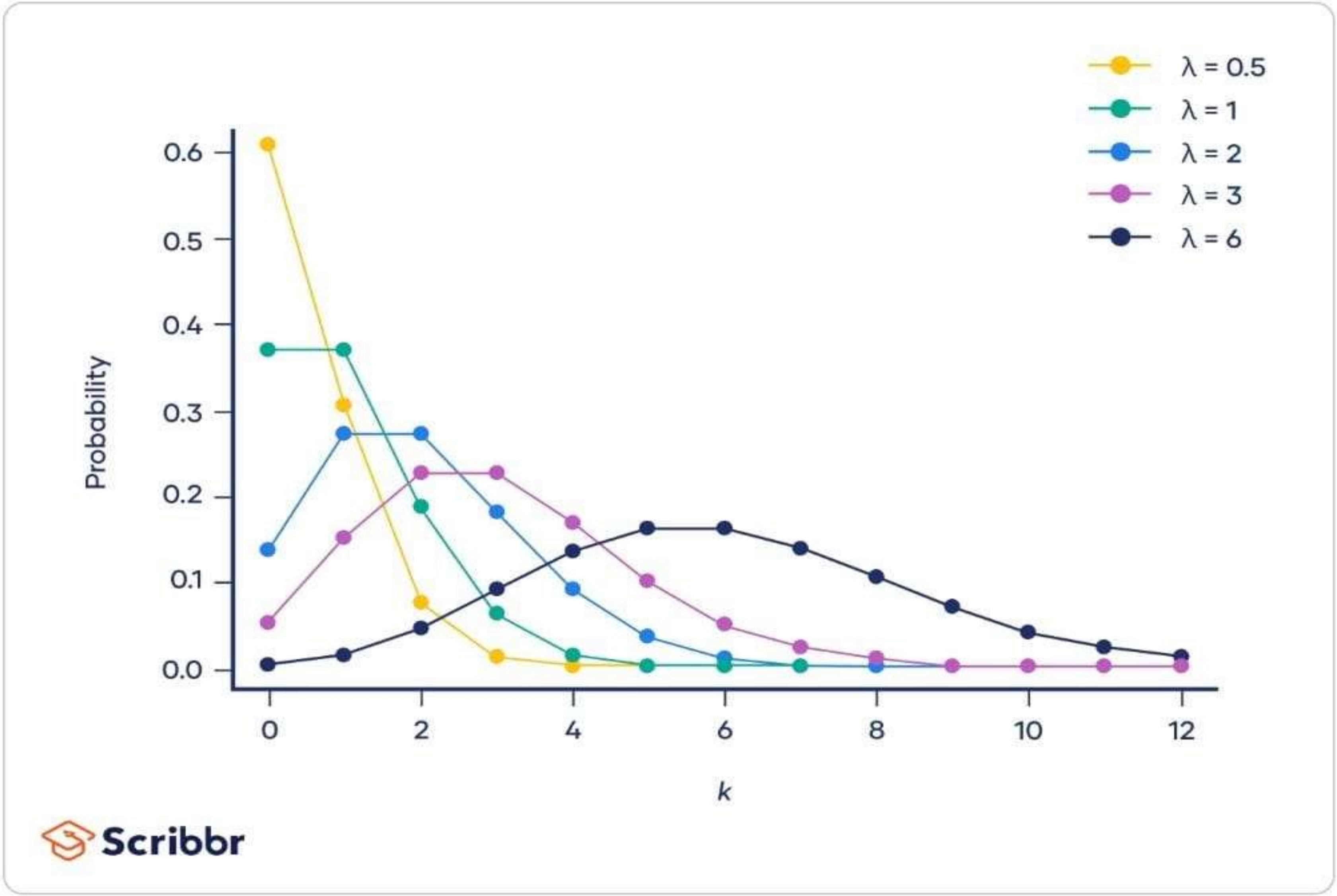


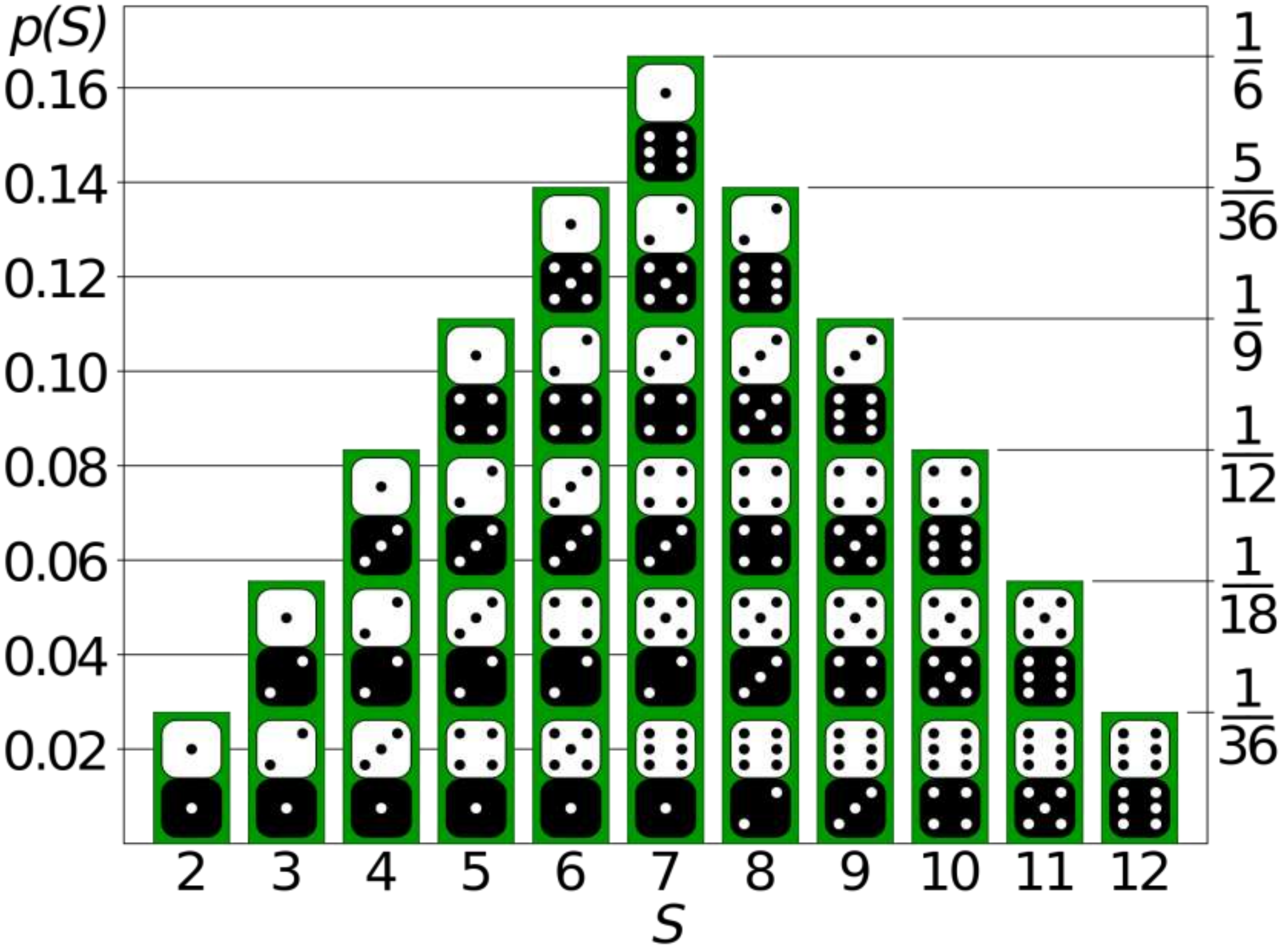
Cumulative distribution function



POSSION DISTRIBUTION

- A **Poisson distribution** is a discrete **probability distribution**. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space.
- The Poisson distribution has only one **parameter**, λ (lambda), which is the **mean** number of events. The graph below shows examples of Poisson distributions with different values of λ.
- A Poisson distribution is a discrete probability distribution, meaning that it gives the probability of a **discrete** (i.e., countable) outcome. For Poisson distributions, the discrete outcome is the number of times an event occurs, represented by k .
- You can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space. “**Events**” could be anything from disease cases to customer purchases to meteor strikes. The interval can be any specific amount of time or space, such as 10 days or 5 square inches.


λ = 0.5
λ = 1
λ = 2
λ = 3
λ = 6
Probability
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
k
Scribbr


p(S)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
2
3
4
5
6
7
8
9
10
11
12
S
$\frac{1}{6}$
$\frac{5}{36}$
$\frac{1}{9}$
$\frac{1}{12}$
$\frac{1}{18}$
$\frac{1}{36}$

ESTIMATION & TESTING OF HYPOTHESIS(LARGE SAMPLE TESTS)

❖A hypothesis in statistics, is a claim or statement about
a property of a population.
A statistical test of hypothesis consists of four parts:

1. **A null hypothesis** (the questioned hypothesis)
2. **An alternative hypothesis** (the hypothesis the researcher wishes to support)
3. **A test statistic**
4. **A rejection region**

- The null hypothesis is a statement about

➢**Null Hypothesis**, Ho
the value of a population parameter

- The null hypothesis contains a condition of
equality: =, <=,>=
- Test the Null Hypothesis directly

Results: Reject H0 or fail to reject H0

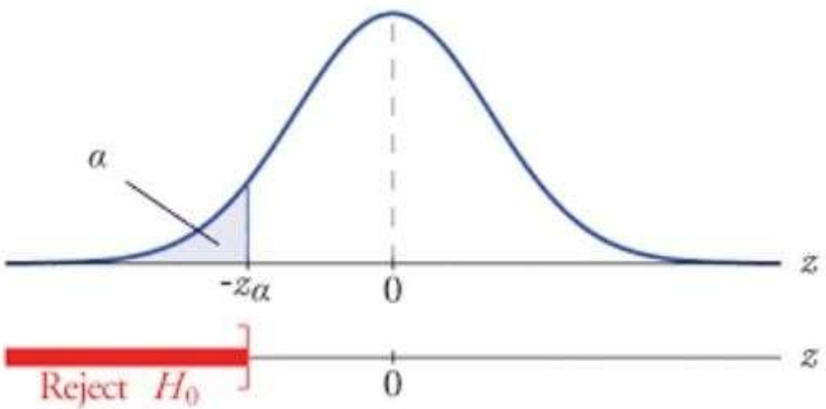
➢**Alternative Hypothesis**, Ha

- Hypothesis the researchers wishes to support
- Must be true if H0is false
- Contains !=, <, >
- Opposite of the Null

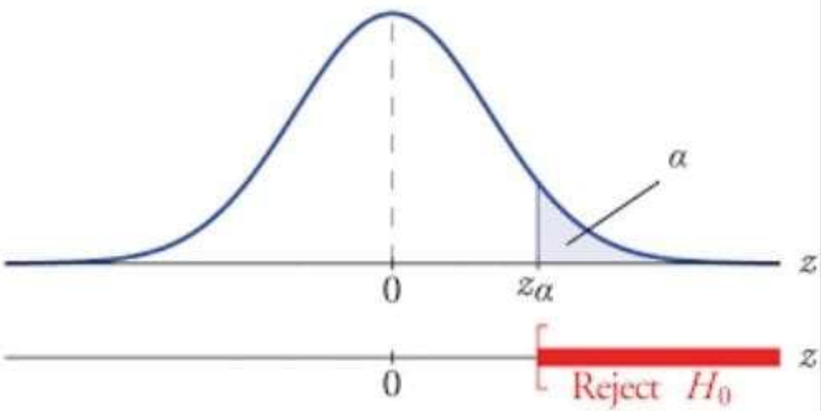
➢**Test Statistic**

- A value computed from the sample data that is used in making the decision about the rejection of the null hypothesis
- For large samples, testing claims about population means:

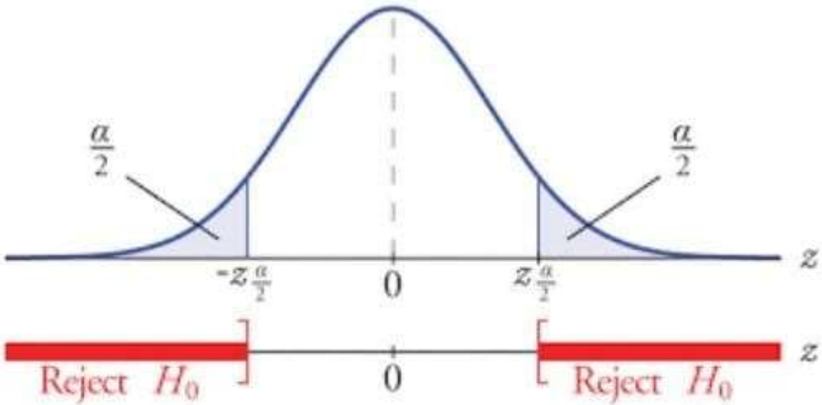
$H_a : \mu < \mu_0$



$H_a : \mu > \mu_0$



$H_a : \mu \neq \mu_0$



TEST OF SIGNIFICANCE (SMALL SAMPLES)

I. Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

II. Every test of significance begins with a null hypothesis H0. H0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write H0: there is no difference between the two drugs on average.

III. Hypotheses are always stated in terms of population parameter, such as the mean . An alternative hypothesis may be one-sided or two-sided. A one-sided hypothesis claims that a parameter is either larger or smaller than the value given by the null hypothesis. A two-sided hypothesis claims that a parameter is simply not equal to the value given by the null hypothesis — the direction does not matter.

Difference between Large and Small sample

Sr. No.	Large sample	Small sample
1.	The sample size is greater than 30.	The sample size is 30 or less than 30
2.	The value of a statistic obtain from the sample can be taken as an estimate of the population parameter.	The value of a statistic obtain from the sample can not be taken as an estimate of the population parameter.
3.	Normal distribution is used for testing.	Sampling distribution like t, F etc. are used for testing.