Binomial Distribution

[bī-ˈnō-mē-əl ˌdi-strə-ˈbyü-shən]

The likelihood of observing a certain outcome when performing a series of tests for which there are only two possible outcomes, such as getting heads or tails in a coin toss.

Investopedia

What Is Binomial Distribution?

- Binomial distribution is a statistical distribution that summarizes the probability that a value will take one of two independent values under a given set of parameters or assumptions.
- The underlying assumptions of binomial distribution are that there is only one outcome for each trial, each trial has the same probability of success, and each trial is mutually exclusive or independent of one another.

KEY :

- Binomial distribution is a statistical probability distribution that summarizes the likelihood that a value will take one of two independent values under a given set of parameters or assumptions.
- The underlying assumptions of binomial distribution are that there is only one outcome for each trial, that each trial has the same probability of success, and that each trial is mutually exclusive or independent of one another.
- Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution.

Understanding Binomial Distribution

- To start, the "binomial" in binomial distribution means two terms—the number of successes and the number of attempts. Each is useless without the other.
- Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution. This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure), given a number of trials in the data. Binomial distribution thus represents the probability for x successes in n trials, given a success probability p for each trial.
- Binomial distribution summarizes the number of trials, or observations, when each trial has the same probability of attaining one particular value.
- Binomial distribution determines the probability of observing a specific number of successful outcomes in a specified number of trials.

For example, the expected value of the number of heads in 100 trials of heads or tails is 50, or (100 × 0.5). Another common example of binomial distribution is estimating the chances of success for a free-throw shooter in basketball, where 1 = a basket made and 0 = a miss.

The binomial distribution function is calculated as:

$$P_{(x:n,p)} = {}^{n}C_{x}\, p^{x} (1 - p)^{n-x}$$

Where:

n is the number of trials (occurrences)

x is the number of successful trials

p is the probability of success in a single trial

${}^{n}C_{x}$ is the combination of n and x.

A combination is the number of ways to choose a sample of x elements from a set of n distinct objects where order does not matter, and replacements are not allowed. Note that ${}_{n}C_{x} = n! / r! (n - r) !)$, where ! is factorial (so, 4! = 4 × 3 × 2 × 1).

How Is Binomial Distribution Used?

- This distribution pattern is used in statistics but has implications in finance and other fields.

- Banks may use it to estimate the likelihood of a particular borrower defaulting, how much money to lend, and the amount to keep in reserve.

- It's also used in the insurance industry to determine policy pricing and assess.

Parameters

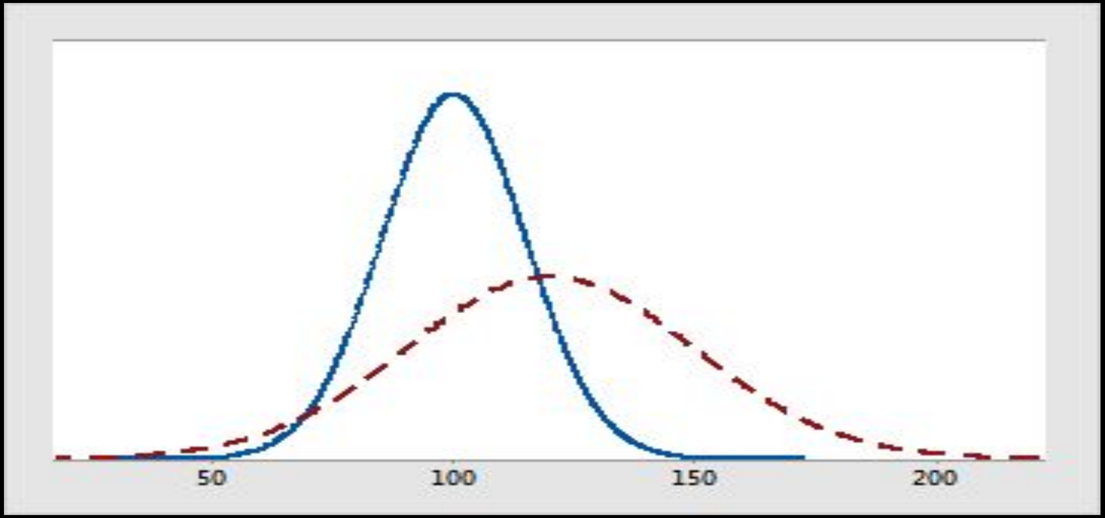
- parameters are descriptive measures of an entire population that may be used as the inputs for a probability distribution function (pdf) to generate distribution curves.
- parameters are usually signified by greek letters to distinguish them from sample statistics.
- for example, the population mean is represented by the greek letter mu (μ) and the population standard deviation by the greek letter sigma (σ).
- parameters are fixed constants, that is, they do not vary like variables.
- however, their values are usually unknown because it is infeasible to measure an entire population.
- each distribution is entirely defined by several specific parameters, usually between one and three.

The following table provides examples of the parameters required for three distributions.

The parameter values determine the location and shape of the curve on the plot of distribution, and each unique combination of parameter values produces a unique distribution curve.

Distribution	Parameter 1	Parameter 2	Parameter 3
Chi-square	Degrees of freedom		
Normal	Mean	Standard deviation	
3-Parameter Gamma	Shape	Scale	Threshold

For example, a normal distribution is defined by two parameters, the mean and standard deviation. If these are specified, the entire distribution is precisely known.



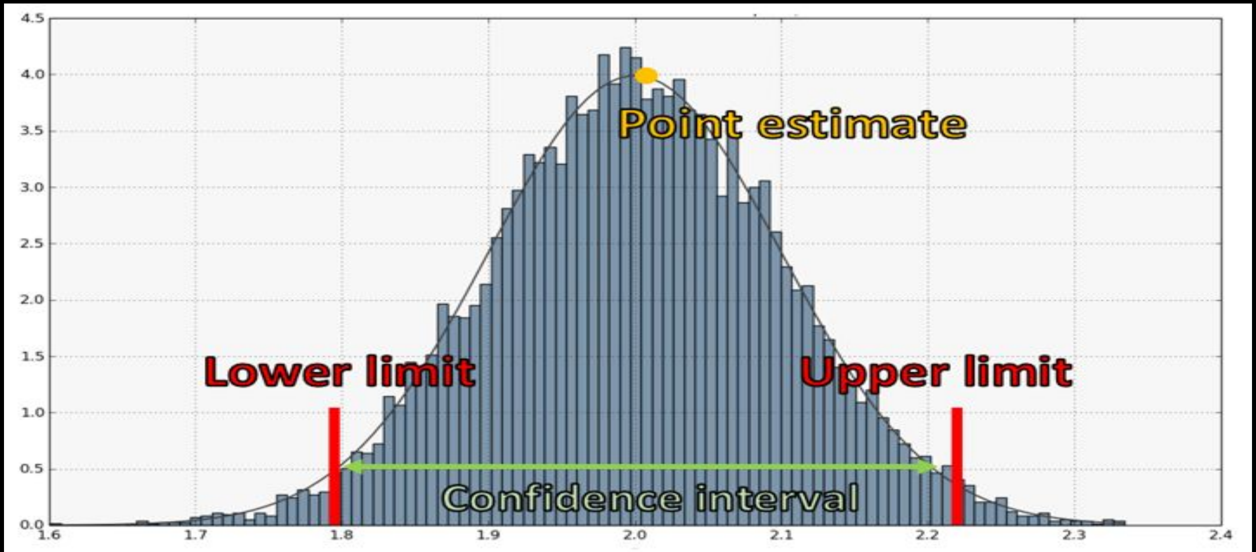
About parameter estimation (also called sample statistics)

- Parameters are descriptive measures of an entire population. However, their values are usually unknown because it is infeasible to measure an entire population.
- Because of this, you can take a random sample from the population to obtain parameter estimates.
- One goal of statistical analyses is to obtain estimates of the population parameters along with the amount of error associated with these estimates.
- These estimates are also known as sample statistics.
- Parameter Estimation is a branch of statistics that involves using sample data to estimate the parameters of a distribution.

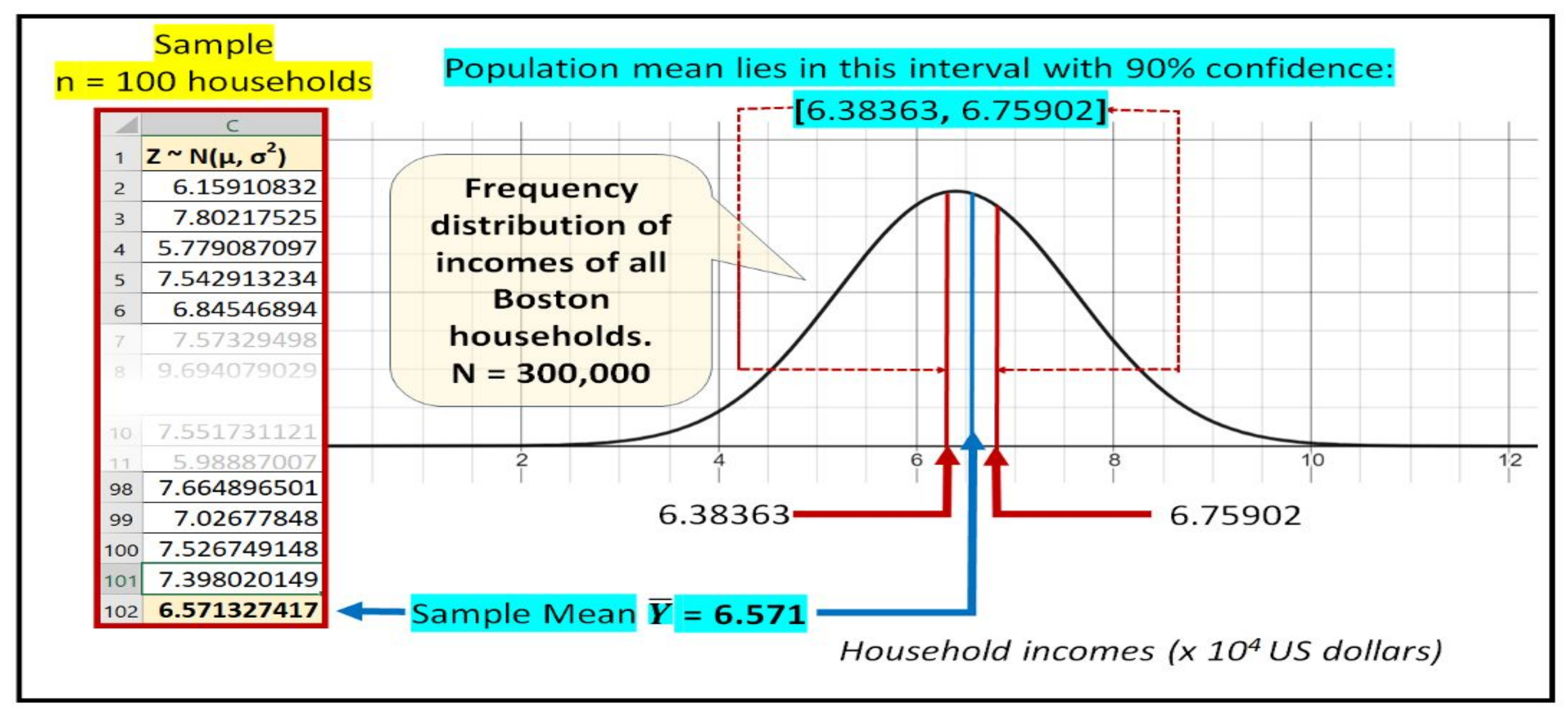
There are several types of parameter estimates

- Point estimates are the single, most likely value of a parameter. For example, the point estimate of population mean (the parameter) is the sample mean (the parameter estimate).
- Confidence intervals are a range of values likely to contain the population parameter.
- There are two important types of estimates you can make about the population parameter:
- point estimates and interval estimates.
- A point estimate is a single value estimate of a parameter based on a statistic. For instance, a sample mean is a point estimate of a population mean.
- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval is the most common type of interval estimate.

- point estimation: in statistics, the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population.
- The accuracy of any particular approximation is not known precisely, though probabilistic statements concerning the accuracy of such numbers as found over many experiments can be constructed.



❖ interval estimation, in statistics, the evaluation of a parameter—for example, the mean (average)—of a population by computing an interval, or range of values, within which the parameter is most likely to be located.



Example

parameter estimates, suppose you work for a spark plug manufacturer that is studying a problem in their spark plug gap.

It would be too costly to measure every single spark plug that is made. Instead, you randomly sample 100 spark plugs and measure the gap in millimeters. The mean of the sample is 9.2. This is the point estimate for the population mean (μ). You also create a 95% confidence interval for μ which is (8.8, 9.6). This means that you can be 95% confident that the true value of the average gap for all the spark plugs is between 8.8 and 9.6.

About sampling distributions

A sampling distribution is the probability distribution of a given statistic, such as the mean.

To illustrate a sampling distribution, let's examine a simple example where the complete population is known.

For example, the following table shows the weights of the entire population of 6 pumpkins.

The pumpkins can only be one of the weight values listed in the following table.

Pumpkin	1	2	3	4	5	6
Weight	19	14	15	12	16	17

.

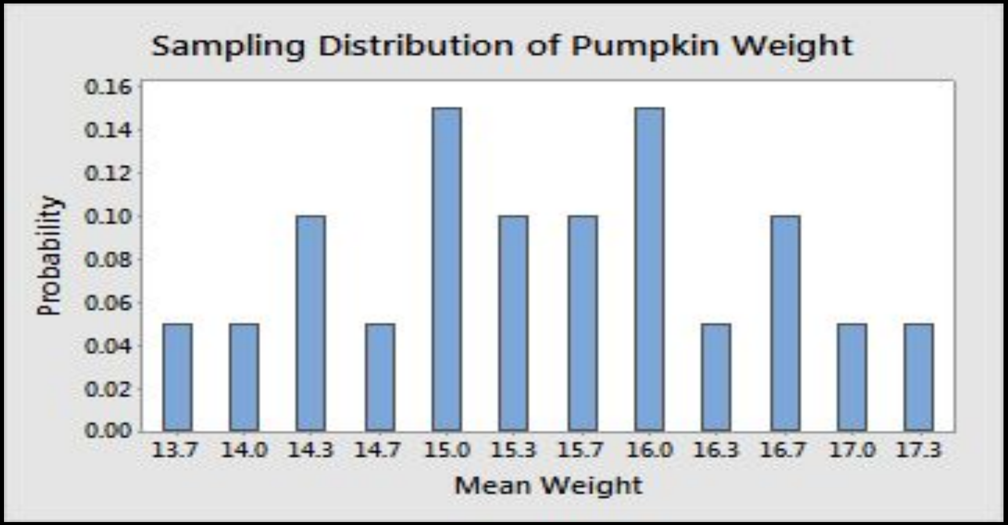
Even though the entire population is known, for illustrative purposes, we take all possible random samples of the population that contain 3 pumpkins (20 random samples).

Then, we calculate the mean of each sample.

The sampling distribution for the sample mean is described by all the sample means for every possible random sample of 3 pumpkins, which is shown in the following table

Sample	Weights	Mean Weight	Probability
2, 3, 4	14, 15, 12	13.7	1/20
2, 4, 5	14, 12, 16	14	1/20
2, 4, 6	14, 12, 17	14.3	2/20
3, 4, 5	15, 12, 16		
3, 4, 6	15, 12, 17	14.7	1/20
1, 2, 4	19, 14, 12	15	3/20
2, 3, 5	14, 15, 16		
4, 5, 6	12, 16, 17		
2, 3, 6	14, 15, 17	15.3	2/20
1, 3, 4	19, 15, 12		
1, 4, 5	19, 12, 16	15.7	2/20
2, 5, 6	14, 16, 17		
1, 2, 3	19, 14, 15	16	3/20
3, 5, 6	15, 16, 17		
1, 4, 6	19, 12, 17		
1, 2, 5	19, 14, 16	16.3	1/20
1, 2, 6	19, 14, 17	16.7	2/20
1, 3, 5	19, 15, 16		
1, 3, 6	19, 15, 17	17	1/20
1, 5, 6	19, 16, 17	17.3	1/20

The sampling distribution of the mean weights is displayed on this graph. The distribution is centered around 15.5, which is also the true value for the population mean. And the random samples with sample means closer to 15.5 have a greater probability of occurring than the samples with sample means farther away from 15.5.



What Is Distribution Fitting?

- Probability distribution fitting or simply distribution fitting is the fitting of a probability distribution to a series of data concerning the repeated measurement of a variable phenomenon.

- The aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval.

- Distribution fitting is the procedure of selecting a statistical distribution that best fits to a data set generated by some random process.

- In other words, if there are some random data available, and someone would like to know what particular distribution can be used to describe the data, then distribution fitting is what is being searched for.

Example

The “candidate” distributions that fit should be chosen depending on the nature of your probability data. For example, if someone needs to analyze the time between failures of technical devices, he/she should fit non-negative distributions such as Exponential or Weibull, since the failure time cannot be negative.

The selection of the appropriate distribution depends on the presence or absence of symmetry of the data set with respect to the mean value.

- Symmetrical distributions – When the data are symmetrically distributed around the mean while the frequency of occurrence of data farther away from the mean diminishes, one may for example select the normal distribution, the logistic distribution, or the Student's t-distribution.
- Skew distributions to the right – When the larger values tend to be farther away from the mean than the smaller values, one has a skew distribution to the right (i.e. there is positive skewness).
- Skew distributions to the left – When the smaller values tend to be farther away from the mean than the larger values, one has a skew distribution to the left (i.e. there is negative skewness).

Techniques of fitting

The following techniques of distribution fitting exist

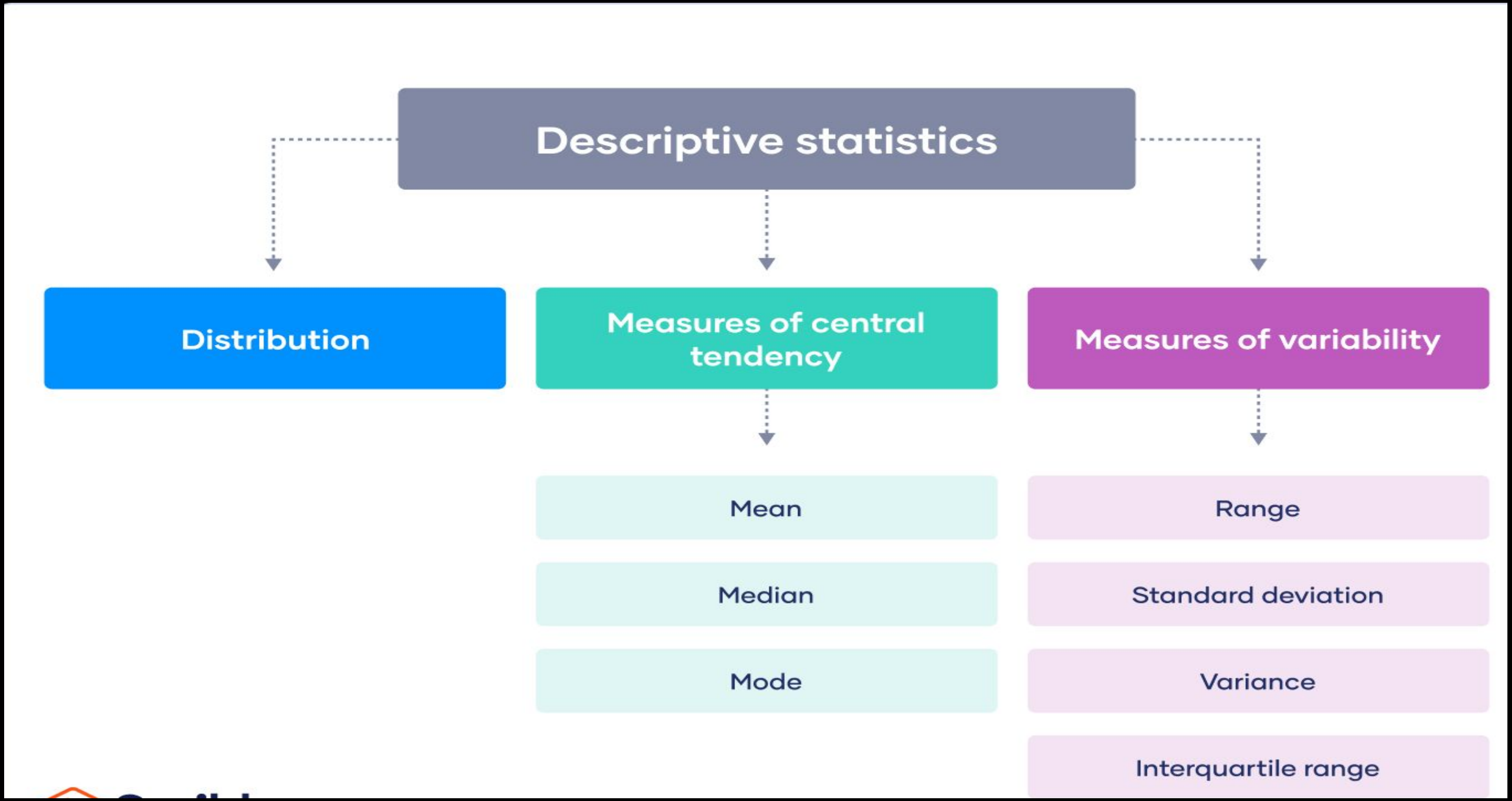
- Parametric methods, by which the parameters of the distribution are calculated from the data series. The parametric methods are – method of moments, method of L-moments and Maximum likelihood method
- Regression method, using a transformation of the cumulative distribution function so that a linear relation is found between the cumulative probability and the values of the data, which may also need to be transformed, depending on the selected probability distribution.

- **Descriptive statistics** summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).
- The next step is **inferential statistics**, which help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population.

Descriptive Statistics Definition

- Descriptive statistics can be defined as a field of statistics that is used to summarize the characteristics of a sample by utilizing certain quantitative techniques.
- It helps to provide simple and precise summaries of the sample and the observations using measures like mean, median, variance, graphs, and charts. Univariate descriptive statistics are used to describe data containing only one variable.
- On the other hand, bivariate and multivariate descriptive statistics are used to describe data with multiple variables.

Types


Descriptive statistics
Distribution
Measures of central tendency
Measures of variability
Mean
Median
Mode
Range
Standard deviation
Variance
Interquartile range

Descriptive Statistics	Inferential Statistics
It is used to describe the characteristics of either the sample or the population by using quantitative tools.	It is used to draw inferences about the population data from the sample data by making use of analytical tools.
Measures of central tendency and measures of dispersion are the most important types of descriptive statistics.	Hypothesis testing and regression analysis are the types of inferential statistics.
It is used to describe the characteristics of a known dataset.	It tries to make inferences about the population that goes beyond the known data.
Measures of descriptive statistics are mean, median, variance, range, etc.	Measures of inferential statistics are z test, f test, linear regression, ANOVA test, etc.

Graphical Statistics

- **Statistical graphics**, also known as **statistical graphical techniques**, are graphics used in the field of statistics for data visualization.
- Whereas statistics and data analysis procedures generally yield their output in numeric or tabular form, graphical techniques allow such results to be displayed in some sort of pictorial form.
- Exploratory data analysis (EDA) relies heavily on such techniques. They can also provide insight into a data set to help with testing assumptions, model selection and regression model validation, estimator selection, relationship identification, factor effect determination, and outlier detection.
- In addition, the choice of appropriate statistical graphics can provide a convincing means of communicating the underlying message that is present in the data to others.
- They include plots such as scatter plots, histograms, probability plots, spaghetti plots, residual plots, box plots, block plots and biplots.

Graphical statistical methods have four objectives:

- The exploration of the content of a data set
- The use to find structure in data
- Checking assumptions in statistical models
- Communicate the results of an analysis.
- If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data

Graphical Representation

- **Line Graphs** – Line graph or the linear graph is used to display the continuous data and it is useful for predicting future events over time.
- **Bar Graphs** – Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.
- **Histograms** – The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.
- **Line Plot –** It shows the frequency of data on a given number line. ‘ x ‘ is placed above a number line each time when that data occurs again.
- **Frequency Table** – The table shows the number of pieces of data that falls within the given interval.
- **Circle Graph** – Also known as the pie chart that shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc.
- **Stem and Leaf Plot –** In the stem and leaf plot, the data are organised from least value to the greatest value. The digits of the least place values from the leaves and the next place value digit forms the stems.
- **Box and Whisker Plot** – The plot diagram summarises the data by dividing into four parts. Box and whisker show the range (spread) and the middle (median) of the data.

What is Method of Moments?

- The method of moments is a way to estimate population parameters, like the population mean or the population standard deviation.
- The basic idea is that you take known facts about the population, and extend those ideas to a sample.

For example, it's a fact that within a population:
Expected value E(x) = μ

For a sample, the estimator

$$\hat{\mu}$$

is just the sample mean, $\bar{\mathbf{x}}$. The formula for the sample mean is:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i.$$

Continue……………………

- In statistics, the **method of moments** is a method of estimation of population parameters.
- The same principle is used to derive higher moments like skewness and kurtosis(Kurtosis is a statistical measure used to describe a characteristic of a dataset.).
- It starts by expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest.
- Those expressions are then set equal to the sample moments. The number of such equations is the same as the number of parameters to be estimated. T
- hose equations are then solved for the parameters of interest. The solutions are estimates of those parameters.

Advantages:--

Method of moments is simple (compared to other methods like the maximum likelihood method) and can be performed by hand.

Disadvantages:---

The parameter estimates may be inaccurate. This is more frequent with smaller samples and less common with large samples.

The method may not result in sufficient statistics. In other words, it may not take into account all of the relevant information in the sample.

Maximum Likelihood Estimation

- The maximum likelihood estimation is a method that determines values for parameters of the model. It is the statistical method of estimating the parameters of the probability distribution by maximizing the likelihood function.
- The point in which the parameter value that maximizes the likelihood function is called the maximum likelihood estimate.
- This principle was originally developed by Ronald Fisher, in the 1920s. He stated that the probability distribution is the one that makes the observed data “most likely”.
- Which means, the parameter vector is considered which maximizes the likelihood function.
- The goal of maximum likelihood estimation is to make inference about the population, which is most likely to have generated the sample i.e., the joint probability distribution of the random variables.

Simple Explanation – Maximum Likelihood Estimation using MS Excel.

Problem: What is the Probability of Heads when a single coin is tossed 40 times.

Observation: When the probability of a single coin toss is low in the range of 0% to 10%, the probability of getting 19 heads in 40 tosses is also very low. However, when we go for higher values in the range of 30% to 40%, I observed the likelihood of getting 19 heads in 40 tosses is also rising higher and higher in this scenario.

In some cases, after an initial increase, the likelihood percentage gradually decreases after some probability percentage which is the intermediate point (or) peak value. The peak value is called maximum likelihood.

Five Major Steps in MLE:

Perform a certain experiment to collect the data.

Choose a parametric model of the data, with certain modifiable parameters.

Formulate the likelihood as an objective function to be maximized.

Maximize the objective function and derive the parameters of the model.

Examples:

Toss a Coin – To find the probabilities of head and tail

Throw a Dart – To find your PDF of distance to the bull eye

Sample a group of animals – To find the quantity of animals