STATISTICS

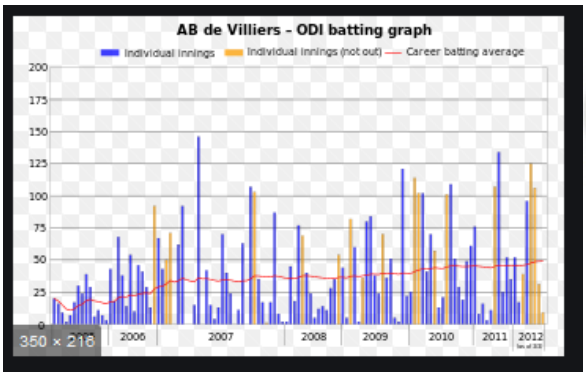
Statistics is the science that deals with methodologies to gather, review, analyze and draw conclusions from data. With specific Statistics tools in hand we can derive many key observations and make predictions from the data in hand.

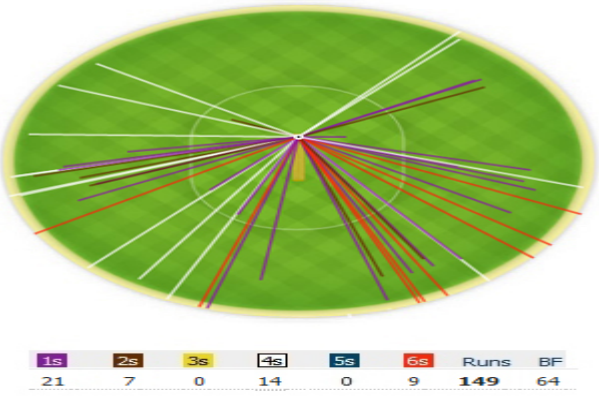
In Real world we deal with many cases where we use Statistics knowingly or unknowingly.

Let's talk about one such classic use of statistics in the most famous sports in India, yes you guessed it right, Cricket.

What makes Virat Kohli the best batsman in ODIs or Jaspreet Bumrah the best bowler in ODIs?

We all have heard about cricketing terms like batting average, bowler's economy, strike rate etc. We often see graphs like these





We see and talk about statistics all the time but very few of us know the science behind it.

Here by using different statistical methods ICC compare players, teams and rank them. So, if we learn the science behind it we can create our own rankings, compare players, teams or better if we debate with someone over who is the better player, we can debate now with facts and figures because we will understand the statistics behind it better. We can understand the above graphs better.

We will dive further in to the various methods and terminologies which will help to answer the question above as well as see the vast uses of Statistics in much complex scenarios such as medical science, drug research, stock markets, Economics, Marketing etc.

Types of Statistics

Descriptive Statistics:

The type of statistics dealing with numbers (numerical facts, figures, or information) to describe any phenomena. These numbers are descriptive statistics.

e.g. Reports of industry production, cricket batting averages, government deficits, Movie Ratings etc.

Inferential statistics

Inferential statistics is a decision, estimate, prediction, or generalization about a population, based on sample.

A **population** is a collection of all possible individual, objects, or measurements of interest.

A **sample** is a portion, or part, of the population of interest.

Inferential statistics is used to make inferences from data whereas descriptive statistics simply describes what's going on in our data.

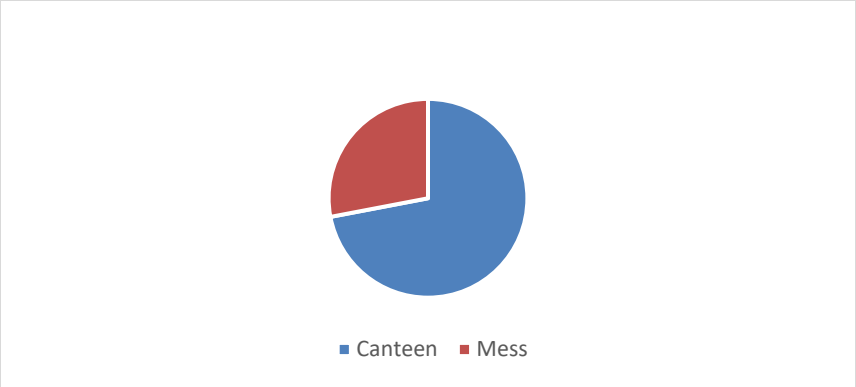
Let's clear our understanding about the above types with a basic scenario:

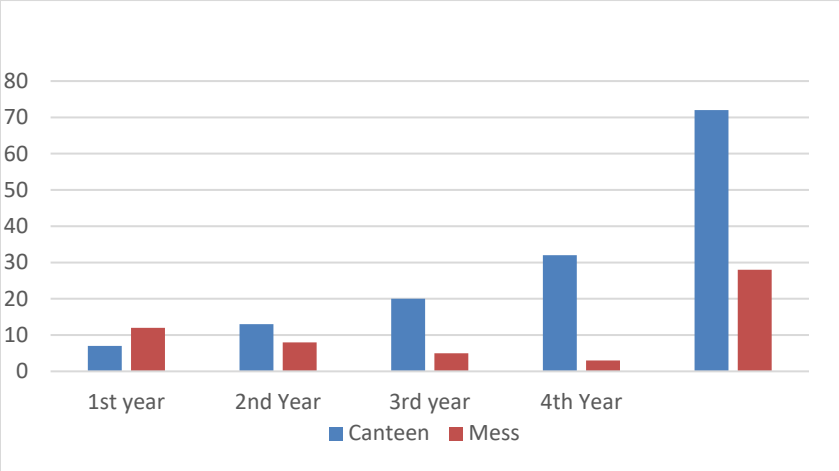
Suppose in your college there are 1000 students. You are interested in finding out the how many students prefer eating in the college canteen than college mess.

A random group of 100 students is selected. Here our population size is of 1000 students and the sample size is of 100 students. You surveyed the sample group and got the following results:

	1st year	2nd Year	3rd year	4th Year	Total
Canteen	7	13	20	32	72
Mess	12	8	5	3	28

Let's analyze the data:





1) 72 % of the students prefer eating in the canteen.
2) Of the total students who prefer canteen, 44.4 % are from the 4th year.
3) Of the total students who prefer canteen, 72% are from the 3rd year and 4th year.
4) 1st year students are more inclined towards eating in the mess.

The above statistics gives us a trend of variation among the students with their preference. We are using the numbers and figures to assess the data. This will be the part of Descriptive statistics.

Now, suppose you got a contract to open a canteen in the College. Now with the above data, you can make following assumptions:

1) 3rd and 4th year students are the main target for sales of restaurant.
2) You can give discounts to the 1st year students to increase the number count.
3) Since most students prefer eating in canteen, opening a canteen can be profitable business

You made the above inferences/estimations for the whole college based on the sample data. This is the part of Inferential statistics where you make decision based on the descriptive statistics of a sample data.

Though the above example is very basic and the real scenarios are much more complex, this would help in getting the underlying difference. We will see more complex examples ahead.

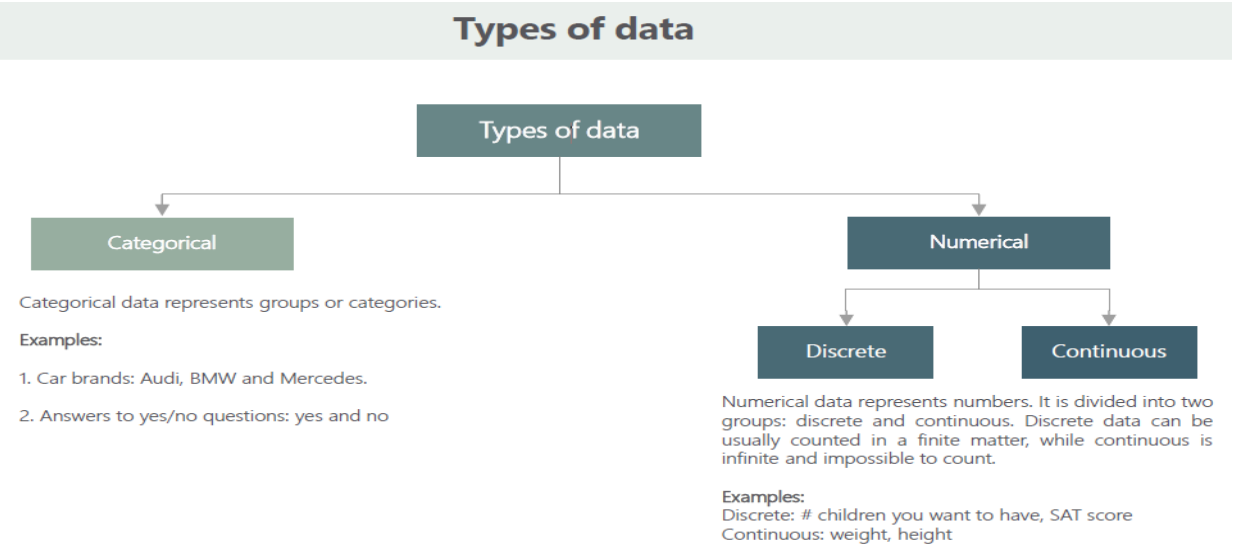
Q) The average salary of employees of company in 2017 is greater than the average salary of teachers of school in 2017.

Is the above statement an example of descriptive or inferential statistics?

Ans. Descriptive statistics because it summarizes the information in two samples.

Q) By 2030 , World will face shortage of waters. ---- Inferential Statistics

Types of Data



I) **Categorical data** represents characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning. You couldn't add them together

ii) **Numerical data** have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stocks shares a person owns etc.

* Numerical data can be further broken into two types: discrete and continuous.

Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite). For example, the number of heads in 100-coin flips takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the

fastest scenario) on up to infinity (if you never get to that 100th heads). Its possible values are listed as 100, 101, 102, 103, . . . (representing the countably infinite case).

Some more examples: -

1) Number of children in a school
2) Number of books in your library
3) Number of cases a Lawyer has won

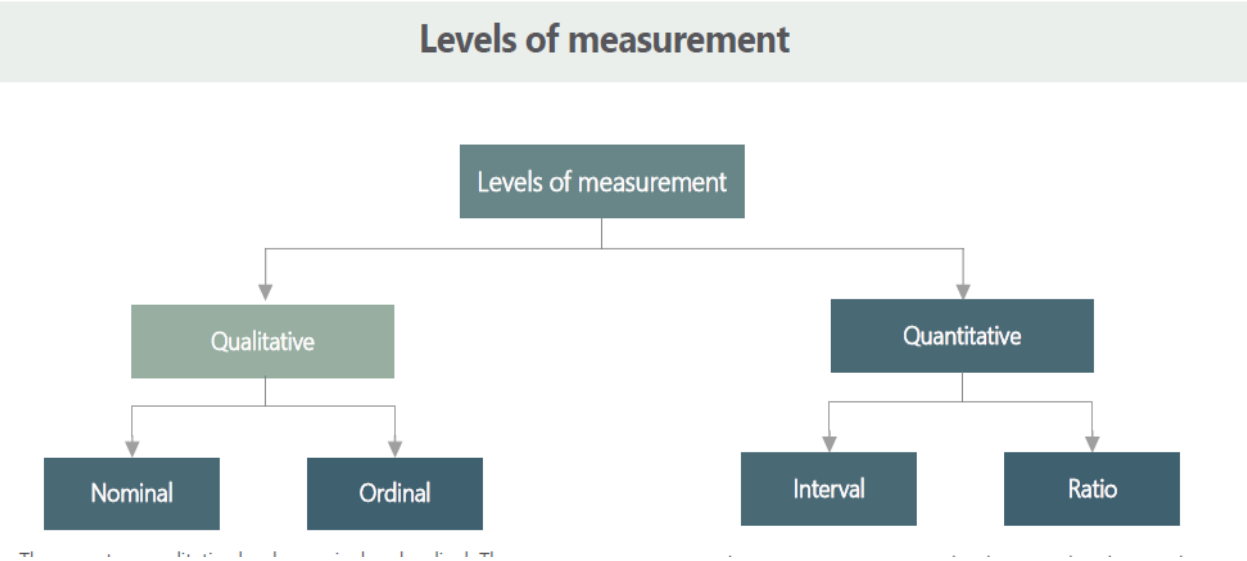
In all the above examples the values can be 1, 2,3 … so on but can never be 1.2,4.6, 8.7 etc. Thus, making the results countable.

Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line.

For example, the exact amount of petrol purchased at the petrol pump for bikes with 20-liters tanks would be continuous data from 0 liters to 20 liters, represented by the interval [0, 20], inclusive. You might pump 8.40 liters, or 8.41, or 8.414863 liters, or any possible number from 0 to 20. In this way, continuous data can be thought of as being uncountably infinite.

Another Example, you purchase a Light bulb and you are informed that the life of the light bulb is 2000 hours. Now, the life of bulb will be a continuous data as it can take any value such as 1998 hours,1998.56 hours, 1896.34 hours or 2000 hours as well.

Levels of measurement



1. Qualitative Data

Nominal data levels of measurement:

A nominal variable is one in which values serve only as labels, even if those values are numbers.

For example, if we want to categorize male and female respondents, we could use a number of 1 for male, and 2 for female. However, the values of 1 and 2 in this case do not represent any meaningful order or carry any mathematical meaning. They are simply used as labels. Nominal data cannot be used to perform many statistical computations, such as mean and standard deviation, because such statistics do not have any meaning when used with nominal variables.

However, nominal variables can be used to do cross tabulations. The chi-square test can be performed on a cross-tabulation of nominal data.

Examples:

Category	Code
Male	1
Female	2

Category	Code
Delhi	1
Mumbai	2
Bangalore	3

Here even though we code the different categories, we cannot say that since 2> 1 then Female>male.

Ordinal data levels of measurement:

Values of ordinal variables have a meaningful order to them. For example, education level (with possible values of high school, undergraduate degree, and graduate degree) would be an ordinal variable. There is a definitive order to the categories (i.e., graduate is higher than undergraduate, and undergraduate is higher than high school), but we cannot make any other arithmetic assumptions beyond that. For instance, we cannot assume that the difference in education level between undergraduate and high school is the same as the difference between graduate and undergraduate.

We can use frequencies, percentages, and certain non-parametric statistics with ordinal data. However, means, standard deviations, and parametric statistical tests are generally not appropriate to use with ordinal data.

Category	Code
Upper Class	1
Middle class	2
Lower class	3

Category	Code
Highly satisfied	5
Satisfied	4
Average	3
Below Average	2
Very bad	1

In the above example for ordinal data, the data gives a sense of comparability i.e. we can say that in the second table Highly satisfied is better than Average. Though, we can say that the difference between Highly satisfied and satisfied is same as Below Average and Very bad.

2. Quantitative Data

Interval scale data levels of measurement

For interval variables, we can make arithmetic assumptions about the degree of difference between values. An example of an interval variable would be temperature. We can correctly assume that the difference between 70 and 80 degrees is the same as the difference between 80 and 90 degrees. However, the mathematical operations of multiplication and division do not apply to interval variables. For instance, we cannot accurately say that 100 degrees is twice as hot as 50 degrees. Additionally, interval variables often do not have a meaningful zero-point. For example, a temperature of zero degrees (on Celsius and Fahrenheit scales) does not mean a complete absence of heat.

An interval variable can be used to compute commonly used statistical measures such as the average (mean), standard deviation etc. Many other advanced statistical tests and techniques also require interval or ratio data.

Example: Measurement of time of an historical event comes under interval scale, since year has no fix origin i.e. 0 year is different for different religions and countries.

Ratio scale data levels of measurement

All arithmetic operations are possible on a ratio variable. An example of a ratio variable would be weight (e.g., in pounds). We can accurately say that 20 pounds is twice as heavy as 10 pounds. Additionally, ratio variables have a meaningful zero-point (e.g., exactly 0 pounds means the object has no weight). Other examples of ratio variables include gross sales of a company, the expenditure of a company, the income of a company, etc.

Example: measurement of temperature in kelvin scale, since kelvin has an absolute 0, measurement of average height of students in a class.

A ratio variable can be used as a dependent variable for most parametric statistical tests such as t-tests, F-tests, correlation, and regression.

We can summarize different levels of measurements as below:

Offers:	Nominal	Ordinal	Interval	Ratio
Sequence of variables is established		yes	yes	yes
Mode	yes	yes	yes	yes
Median		yes	yes	yes
Mean			yes	yes
Difference between variables can be evaluated			yes	yes
Addition and subtraction of variables			yes	yes
Multiplication and Division of variables				yes
Absolute zero				yes

Measures of Central Tendency

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the mean, median, and mode. Each of these measures calculates the location of the central point using a different method.

Mean: The mean is the arithmetic average, for calculating the mean just add up all of the values and divide by the number of observations in your dataset.

Let you have dataset with n values as follows:

$$D= \{x_1, x_2, x_3, x_4, x_5 \ldots\ldots, x_n\}$$

Mean for the above data will be given as:

$$\text{mean}= \sum_{i=1}^{n} x_i/n$$

Median: The median is the middle value. It is the value that splits the dataset in half. To find the median, order your data from smallest to largest, and then find the data point that has an equal amount of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values.

Let you have dataset with n values as follows:

$D = \{x_1, x_2, x_3, x_4, x_5 \ldots\ldots, x_n\}$

Case I) When n is **odd**:

let mid value is at i position = x_i

Median = x_i

Case ii) when n is **even**

let mid value is at i position = x_i

Median = $(x_i + x_(i+1))/2$

Frequency: The number of times a variable occurs in the data set is called its frequency.

Mode: The mode is the value that occurs the most frequently in your data set i.e. has the highest frequency. On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

Q) Given a data set of heights of student in a class. Find out the mean, mode and median for the dataset. Heights (in cm) = {180, 167, 154, 142, 181, 145, 143, 145, 167, 145}

Solution: no. of observations (n)= 10

Mean= (180 + 167 +154 + 142 + 181 + 145 + 143 + 145 + 167 + 145) /10 = 156.9 cm

Rearranged Heights= {142, 143, 145, 145, 145, 154, 167, 167, 180, 181}

Since n is even,

i = n/2= 5

Value at 5^{th} position = 145

Value at 6^{th} position = 145

median = (145+154)/2 = 149.5

For calculating Mode, we will create a Frequency table for all the variables.

Variables	Frequency
180	1
167	2
154	1
142	1
181	1
145	3
143	1

the number with highest frequency of occurring is 145.

mode= 145

Let's take the above example and change values of some observations and check it's effect on our measurements:

Heights (in cm) = {180, 167, 154, 122, 181, 135, 123, 145, 166, 145}

Rearranged Heights= {122, 123, 135, 145, 145, 154, 166, 167, 180, 181}

Note: we have replaced the first three values on the left of the median

Mean= (122 + 123 + 135 + 145 + 145 + 154 + 166 + 167 + 180 + 181)/10 = 151.8

Median = (145+154)/2 = 149.5= 149.5

Note: We can see a significant change in the Mean whereas Median does not have any changes.

That's because the calculation of the mean incorporates all values in the data. If you change any value, the mean changes.

Unlike the mean, the median value doesn't depend on all the values in the dataset. Consequently, when some of the values are more extreme, the effect on the median is smaller. Of course, with other types of changes, the median can change.

Examples:

1) Given the below dataset, find out the mean, median and mode.

Variable	Frequency	Cumulative Frequency
20	7	7
40	5	12
60	4	16
80	3	19

Solution:

Variable(x)	Frequency(f)	Cumulative Frequency (c.f)	f * X
20	7	7	140
40	5	12	200
60	4	16	240
80	3	19	240
Total	19		820

Mean = sum(f*x)/ sum(f) = 820/19 = 43.15

Mode = 20 (highest frequency)

Median = value of variable corresponding to the (19/2)th cumulative frequency

= value of variable corresponding to the 9.5^{th} cumulative frequency

Since, there is no value of 9.5 in cumulative frequency column, we take the next cumulative frequency that is 12.

Median = value of variable corresponding to the 12^{th} cumulative frequency = 40

Mean, Median, Mode for Grouped Data

Mean for grouped is calculated the same way as we do in ungrouped data, just the variable(x) becomes the midpoint of the interval.

Median:

$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

Where:

l = lower class boundary of the median class

h = Size of the median class interval

f = Frequency corresponding to the median class

N = Total number of observations i.e. sum of the frequencies

c = Cumulative frequency preceding median class.

Mode:

$$Mode = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where,

l = Lower Boundary of modal class

$h = size\ of\ model\ class$

$f_m = Frequency\ corresponding\ to\ modal\ class$

$f_1 = Frequency\ preceding\ to\ modal\ class$

$f_2 = Frequency\ proceeding\ to\ modal\ class$

Let's solve an example to better understand the above formulas:

Q) Calculate the mean, median and mode for below given data:

Variable(x)	Frequency(f)	Cumulative Frequency(c.f)
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

Solution:

Group	Mid point(x)	Frequency(f)	Cumulative Frequency(c.f)	f * X
0-10	5	3	3	15
10-20	15	5	8	75
20-30	25	7	15	175
30-40	35	9	24	315
40-50	45	4	28	180
Total		28		760

Here for calculating mean, we chose the midpoints of the groups as variable(x).

Mean = 760/ 28 = 27.14

Median:

Median class = class with c.f value of (28/2)

Since 14 is not in c.f column, next closest value is 15.

Median class= [20-30]

Using the above formula for median:

l = 20

h= 10

f= 7

N=28

c= 8

Median = 20 +[(14-8) *10] / 7 = 28.57

Mode:

Modal class = [30-40] (It is the Group with the highest frequency 9)

l= 30

h= 10

f(m) = 9

f1= 7

f2= 4

Mode= 30 + 10* (9-7)/(2*9-7-4) = 32.85

SKEWNESS EFFECTS AND USES OF CENTRAL TENDANCIES

What is Skewness

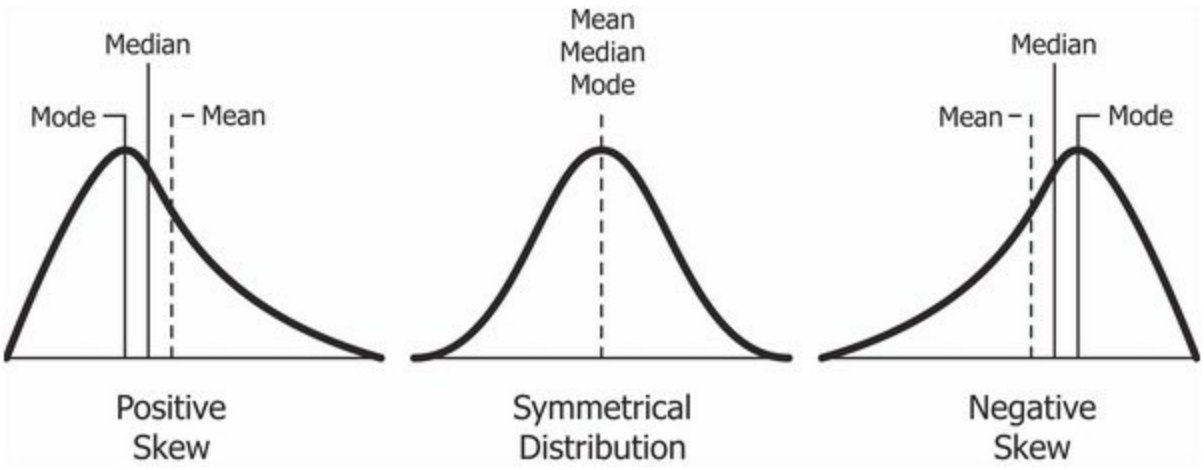
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal.

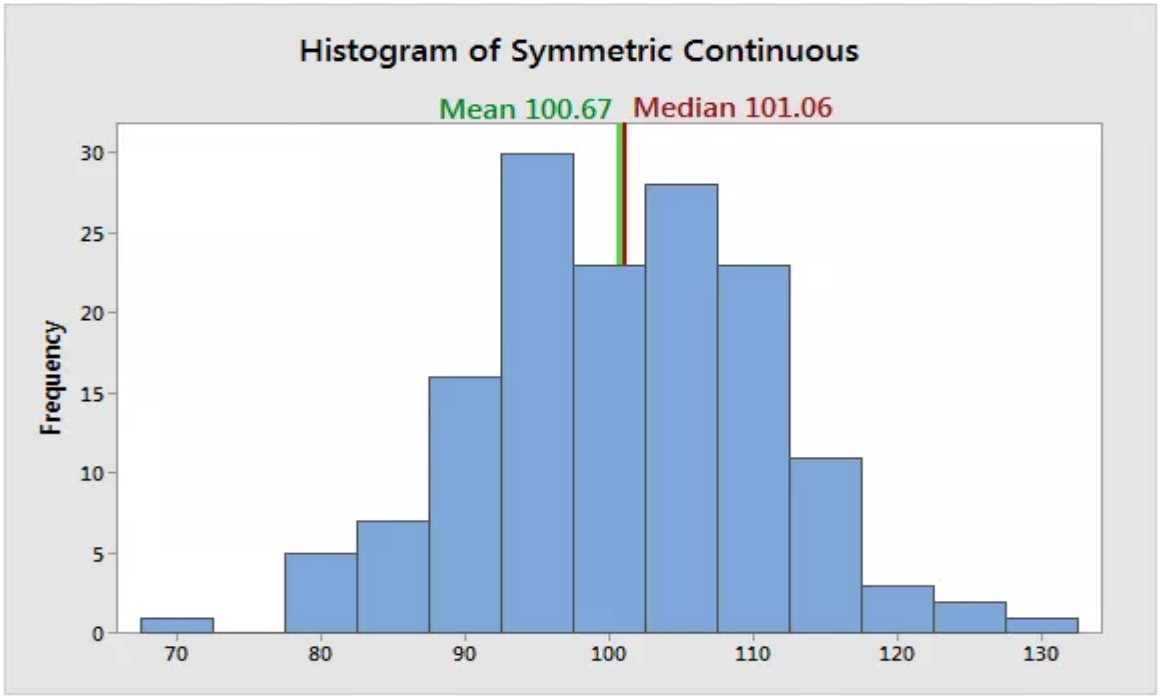
In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called **negative skewness**.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called **positive skewness**.

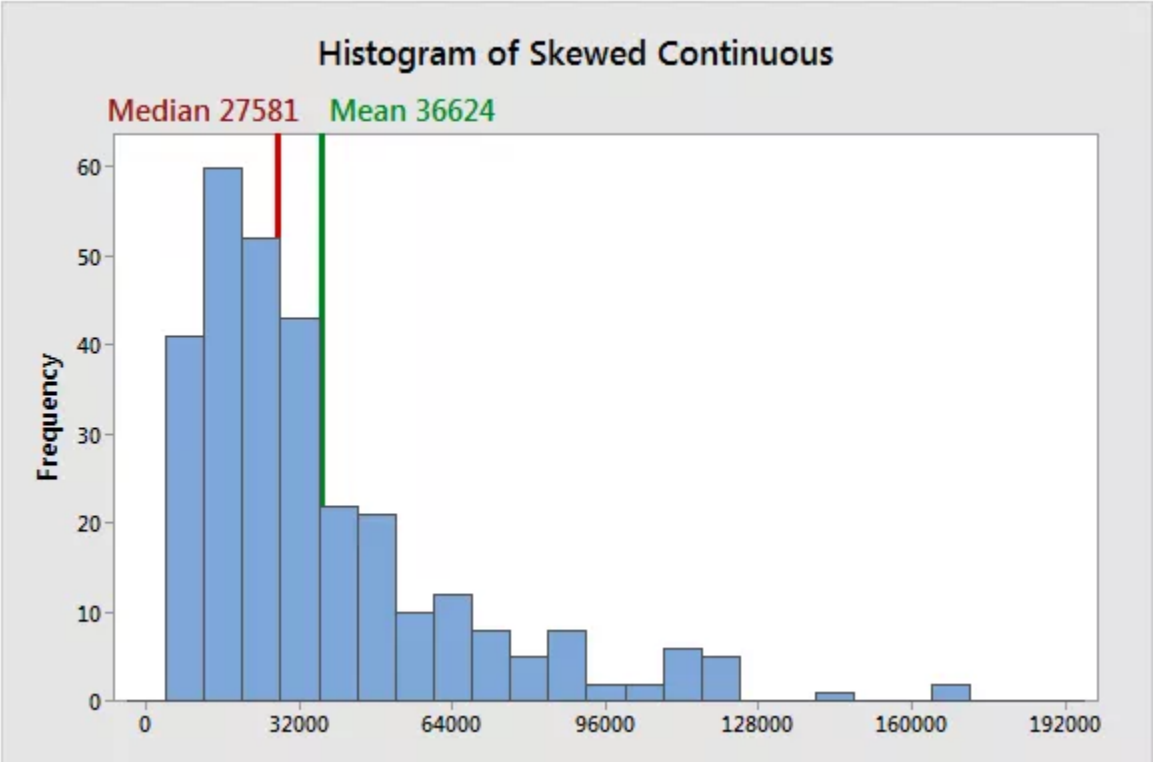


**In a symmetric distribution, the mean and median both find the center accurately. They are approximately equal.



**However, in a skewed distribution, the mean can miss the mark. In the histogram below, it is starting to fall outside the central area. This problem occurs because outliers have a substantial impact on the mean. Extreme values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center.

Here the median better represents the central tendency for the distribution.



Uses of Mean, Median and Mode

When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.

When you have ordinal data, the median or mode is usually the best choice. For categorical data, you have to use the mode.

Measures of Dispersion

The measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

How is it useful?

1) Measures of dispersion shows the variation in the data which provides information like how well the average of the sample represent the entire data. Less variation gives close representation while with larger variation average may not closely represent all the values in the sample.
2) Measures of dispersion enables us to compare two or more series with regard of their variations. It helps to determine consistency.
3) With the checking for variation in the data, we can try to control the causes behind the variations.

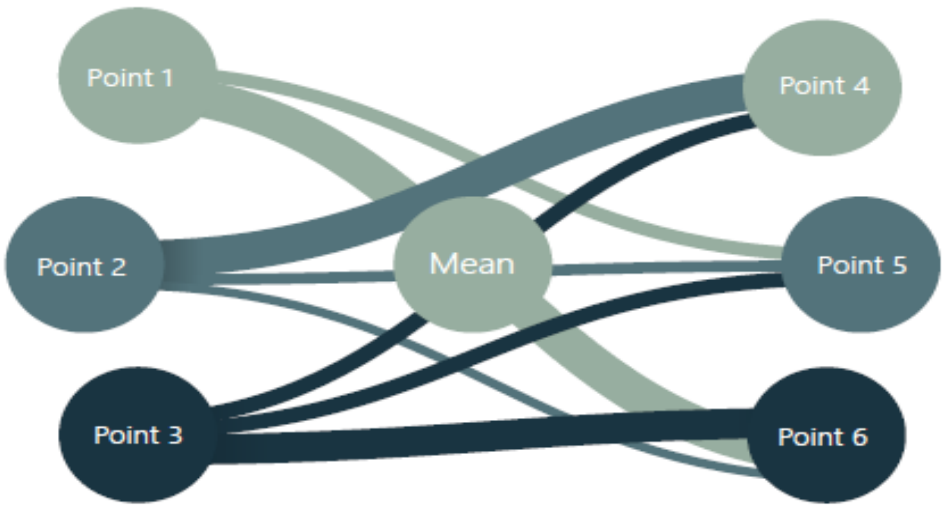
1) Range : A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X max and X min are the two extreme observations then

Range = X max – X min

Since it is based on two extreme observations so it gets affected by fluctuations.

Thus, range is not a reliable measure of dispersion

2) Standard Deviation



In statistics, the standard deviation is a very common measure of dispersion. Standard deviation measures how spread out the values in a data set are around the mean. More precisely, it is a measure of the average distance between the values of the data in the set and the mean. If the data values are all similar, then the standard deviation will be low (closer to zero). If the data values are highly variable, then the standard variation is high (further from zero).

Standard deviation

Let a population consist of n elements, $\{x_1; x_2; \ldots; x_n\}$, with a mean of $\overline{x}$. The standard deviation of the data is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}$$

The standard deviation is always a positive number and is always measured in the same units as the original data. Squaring the deviations overcomes the drawback of ignoring signs in mean deviations i.e. distance of points from mean must always be positive.

3) Variance

The Variance is defined as the average of the squared differences from the Mean.

Variance

Let a population consist of n elements, $\{x_1; x_2; \ldots; x_n\}$. Write the mean of the data as $\overline{x}$.

The variance of the data is the average squared distance between the mean and each data value.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

Example:

1. A class of students took a test in Language Arts. The teacher determines that the mean grade on the exam is a 65%. She is concerned that this is very low, so she determines the standard deviation to see if it seems that most students scored close to the mean, or not. The teacher finds that the standard deviation is high. After closely examining all of the tests, the teacher is able to determine that several students with very low scores were the outliers that pulled down the mean of the entire class's scores.

2. An employer wants to determine if the salaries in one department seem fair for all employees, or if there is a great disparity. He finds the average of the salaries in that department and then calculates the variance, and then the standard deviation. The employer finds that the standard deviation is slightly higher than he expected, so he examines the data further and finds that while most employees fall within a similar pay bracket, three loyal employees who have been in the department for 20 years or

more, far longer than the others, are making far more due to their longevity with the company. Doing the analysis helped the employer to understand the range of salaries of the people in the department.

Coefficient of Variation (CV)

The coefficient of variation (CV), also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation(σ) to the mean(μ). It gives the measure of variability

CV = Standard Deviation / Mean

Let's take one more example to try and understand how standard deviation and CV is helpful:

We are given batting score made by two batsmen in 10 matches:

Batsman	Match 1	Match 2	Match 3	Match 4	Match 5	Match 6	Match 7	Match 8	Match 9	Match 10	Sum	Mean
Batsman 1	54	35	68	12	13	120	6	0	18	184	510	51
Batsman 2	45	42	25	53	75	12	28	27	85	43	435	43.5

By seeing the above data, we can say that Batsman 1 is the better batsman than Batsman 2 and can be given preference since it's mean is greater. But is it really true?

Let's check the variance of the data:

Batsman	Batsman 1	Batsman 2	diff_1	diff_2	var sqaure1 = (diff_1)^2	var sq_2 = (diff_2)^2
Match 1	54	45	-3	-1.5	9	2.25
Match 2	35	42	16	1.5	256	2.25
Match 3	68	25	-17	18.5	289	342.25
Match 4	12	53	39	-9.5	1521	90.25
Match 5	13	75	38	-31.5	1444	992.25
Match 6	120	12	-69	31.5	4761	992.25
Match 7	6	28	45	15.5	2025	240.25
Match 8	0	27	51	16.5	2601	272.25
Match 9	18	85	33	-41.5	1089	1722.25
Match 10	184	43	-133	0.5	17689	0.25
sum	510	435			31684	4656.5

Variance (Batsman 1)= 31684/10 = 3168.4

Standard deviation(batsman1) = Variance (Batsman 1) ^ ½ = 56.288

Coeff. Of Variation (batsman1) = 56.288/51 = 1.10

Variance (Batsman 2)= 4656.5/10 = 465.65

Standard deviation(batsman1) = Variance (Batsman 2) ^ ½ = 21.57

Coeff. Of Variation (batsman2) = 21.57/43.5 = 0.50

We can clearly see that the standard deviation gives a different picture for both batsmen. Though batsman1 has a high average but his variance is very high. So, the batsman 1 is less reliable.

On the other hand, Batsman 2 has lower average but is much more consistent than Batsman 1.

Also, coeff. Of variation for batsman 2 is lower than batsman 1 which insures low variability and higher consistency.

If we only had taken mean into account, we wouldn't have gotten the true picture. This problem is solved by the dispersion measures.

Q) What is the variance and standard deviation of the possibilities associated with rolling a fair die?

Ans: - Possible outcomes = {1,2,3,4,5,6}

mean = (6+5+4+3+2+1)/6 = 3.5

variance = (6.25+2.25+0.25+0.25+2.25+6.25)/6=2.917

std. deviation = (2.917) ^0.5 = 1.71

Q) The following data set has a mean of 14.7 and a variance of 10.01.

18, 11, 12, a, 16, 11, 19, 14, b, 13

Compute the values of a and b.

Ans: -From the formula of the mean we have

14.7 = (114+a+b)/10

a+b = 147−114

a=33−b

From the formula of the variance we have

10.01 = (69.12+(a−14.7) ^2+(b−14.7) ^2)/10

Substitute a=33−b and solve: -

b= 13 0r b=20

Since a=33−b

we have a=20 or a=13.

So, the two unknown values in the data set are 13 and 20

We do not know which of these is a and which is b since the mean and variance tell us nothing about the order of the data.

Standard Deviation and Variance for Population and Sample Data

When you have "N" data values that are:

1) The Population: divide by N when calculating Variance

2) A Sample: divide by N-1 when calculating Variance

The formula for calculating Standard deviation and variance changes while dealing with Population and sample data:-

$$\text{The } \textbf{Population} \text{ Standard Deviation: } \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$\text{The } \textbf{Sample} \text{ Standard Deviation: } \quad s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Why do we divide by (n-1) instead of n?

How to calculate the standard deviation?

1. Compute the square of the difference between each value and the sample mean.

2. Add those values up.

3. Divide the sum by n-1. This is called the variance.

4. Take the square root to obtain the Standard Deviation.

Why n-1?

In step 1, you compute the difference between each value and the mean of those values. You don't know the true mean of the population; all you know is the mean of your sample. Except for the rare cases where the sample mean happens to equal the population mean, the data will be closer to the sample mean than it will be to the true population mean. So, the value you compute in step 2 will probably be a bit smaller (and can't be larger) than what it would be if you used the true population mean in step 1. To make up for this, divide by n-1 rather than n.

**This is called Bessel's correction. **

But why n-1? If you knew the sample mean, and all but one of the values, you could calculate what that last value must be. Statisticians say there are n-1 degrees of freedom.

Covariance and Correlation

Covariance

It is method to find the variance between two variables.

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.
5. It has dimensions.

For Population:

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

For Sample

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

Here,
x' and y' = mean of given sample set
n = total no of sample
xi and yi = individual sample of set

Covariance Relationship

Correlation

* It shows whether and how strongly pairs of variables are related to each other.

* Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive

* correlation and values close to -1 represents strong negative correlation.

* In this variable are indirectly related to each other.

* It gives the direction and strength of relationship between variables.

* It is the scaled version of Covariance.

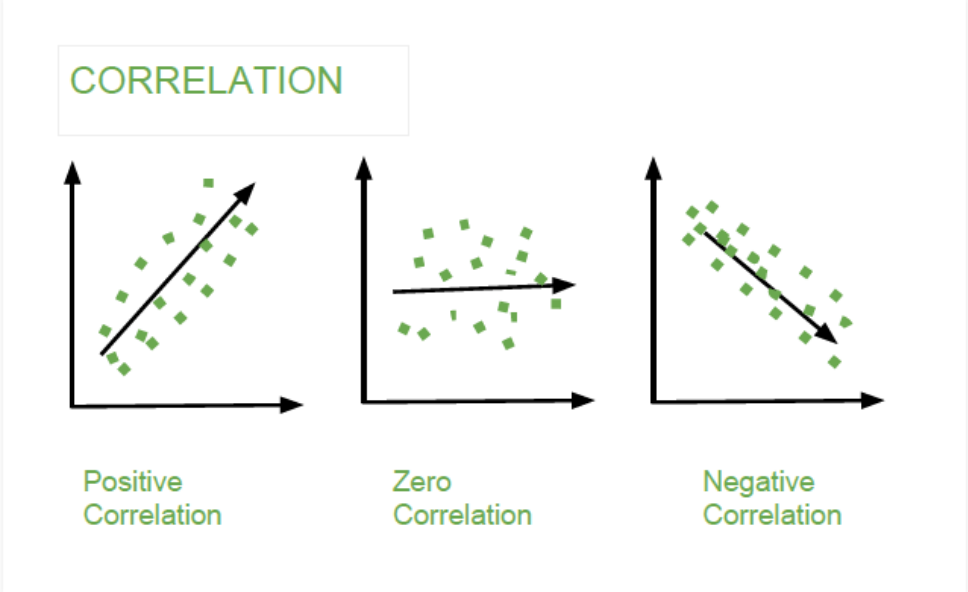
* It is dimensionless.

Formula –

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

Here,
x' and y' = mean of given sample set
n = total no of sample
xi and yi = individual sample of set

Correlation Relationship



Positive Correlation

When the values of variables deviate in the same direction i.e. when value of one variable increases(decreases) then value of other variable also increases(decreases).

Examples:

1) Height and weight of persons
2) Amount of rainfall and crops yield

3) Income and Expenditure of Households
4) speed of a wind turbine, the amount of electricity that is generated
5) The more years of education you complete, the higher your earning potential will be
6) As the temperature goes up, ice cream sales also go up
7) The more it rains, the more sales for umbrellas go up

Negative Correlation

When the values of variables deviate in the opposite direction i.e. when value of one variable increases(decreases) then value of other variable also decreases(increases).

Examples:

1) Price and demand of goods
2) Poverty and literacy
3) Sine function and cosine function
4) If a train increases speed, the length of time to get to the final point decreases
5) The more one works out at the gym, the less body fat one may have
6) As the temperature decreases, sale of heaters increases

Zero Correlation

When two variables are independent of each other, they will have a zero correlation.

Note: - When data is scaled covariance and correlation will give the same value. Also, correlation and Causality are not the same thing.

Example:

x = [1,2,3,4,5,6,7,8,9]

y = [9,8,7,6,5,4,3,2,1]

Find the correlation between x and y.

Ans: We can clearly see in the dataset that as x increase y decreases and vice versa.

Let's prove this with the formula we have studied above.

Solution:

x	y	x- x_mean	y- y_mean	(x- x_mean)^2	(y- y_mean)^2	(x- x_mean)*(y-y_mean)
1	9	-4	4	16	16	-16
2	8	-3	3	9	9	-9
3	7	-2	2	4	4	-4
4	6	-1	1	1	1	-1
5	5	0	0	0	0	0
6	4	1	-1	1	1	-1
7	3	2	-2	4	4	-4
8	2	3	-3	9	9	-9
9	1	4	-4	16	16	-16
5	5			60	60	-60

Corr(x,y) = -60/[(60*60)^1/2] = -1

As expected, we got perfect negative correlation between x and y.

As we will proceed further, we will use several statistical tools such as Python Statistics libraries to do the complex calculation and derive our descriptive statistics.

For e.g. if we solved the above problem using Python, it will pretty easy to do.

```
## Correlation Example 1
# Python code to demonstrate the
# use of numpy.corrcoef

from numpy import array
from numpy import corrcoef
x = array([1,2,3,4,5,6,7,8,9])
print(x)
y = array([9,8,7,6,5,4,3,2,1])
print(y)
Sigma = corrcoef(x,y)
print(Sigma)
```

```
[1 2 3 4 5 6 7 8 9]
[9 8 7 6 5 4 3 2 1]
[[ 1. -1.]
 [-1.  1.]]
```

Here is another example of Correlation coefficient calculation using python:

```
## Correlation Example 2
# Python code to demonstrate the
# use of numpy.corrcoef


import numpy as np

x = np.array([[0, 3, 4], [1, 2, 4], [3, 4, 5]])

print("Shape of array:\n", np.shape(x))

print("Correlation matrix of x:\n", np.corrcoef(x))
```

```
Shape of array:
 (3, 3)
Correlation matrix of x:
 [[ 1.          0.89104211  0.96076892]
 [ 0.89104211  1.          0.98198051]
 [ 0.96076892  0.98198051  1.        ]]
```

In the above example we are trying to find correlation between three set of variables.

The matrix represent the correlation between them, such as the 2nd value in the 1st row gives relation between 1st and 2nd set of variables, 3rd value in the 1st row gives correlation between 1st and 3rd set of variables and so on.