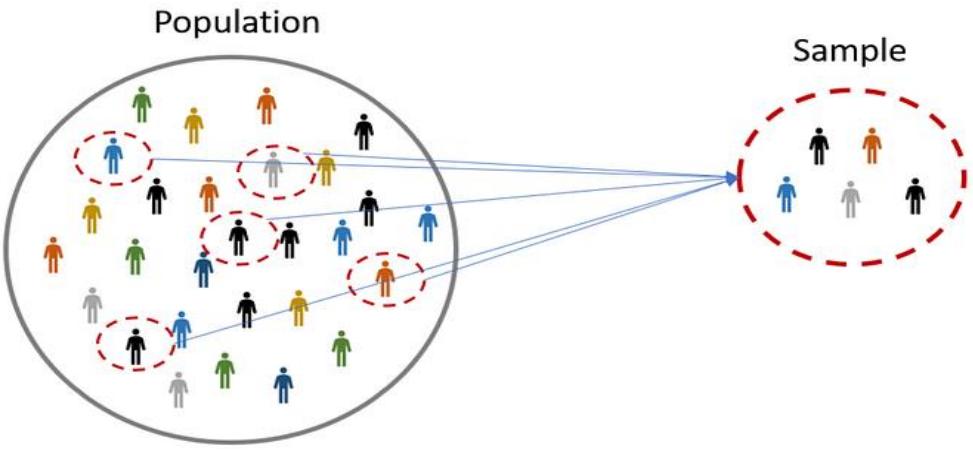
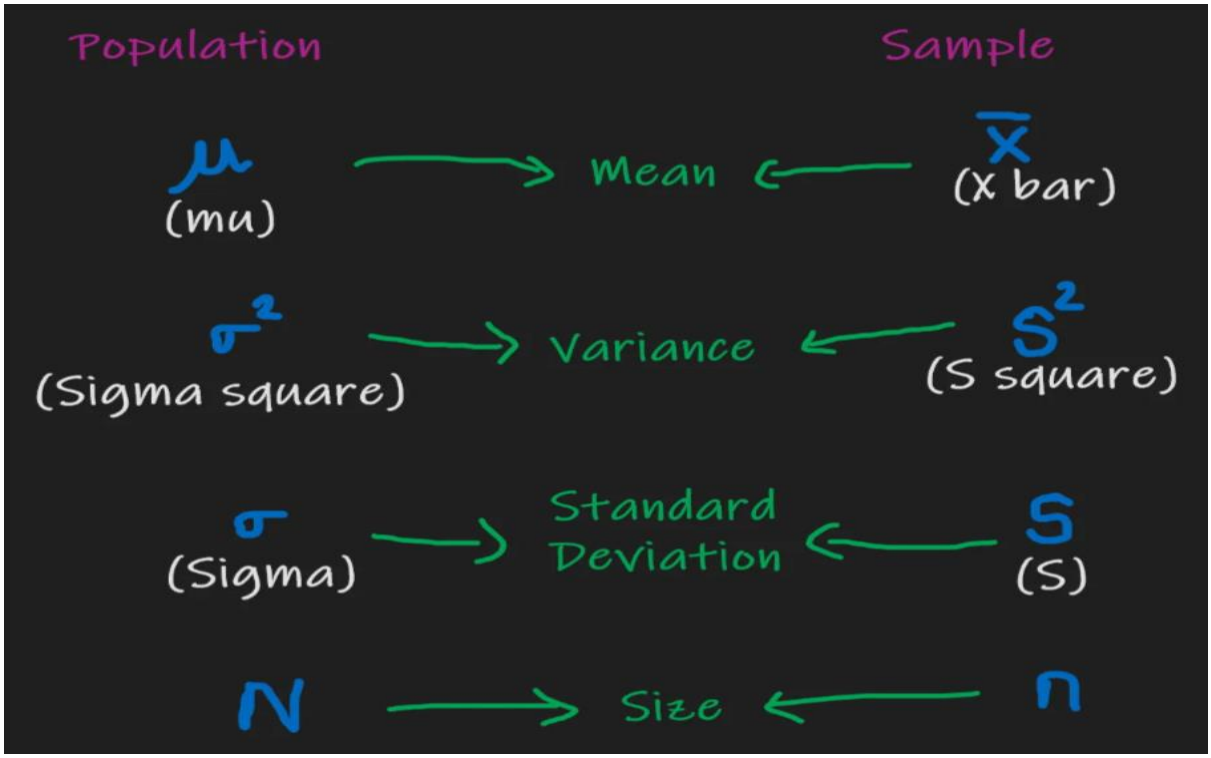
Population & Sample



Population : The Population is the Entire group that you are taking for analysis or prediction.
Sample : Sample is the Subset of the Population(i.e. Taking random samples from the population). The size of the sample is always less than the total size of the population.

Parameter & Statistic

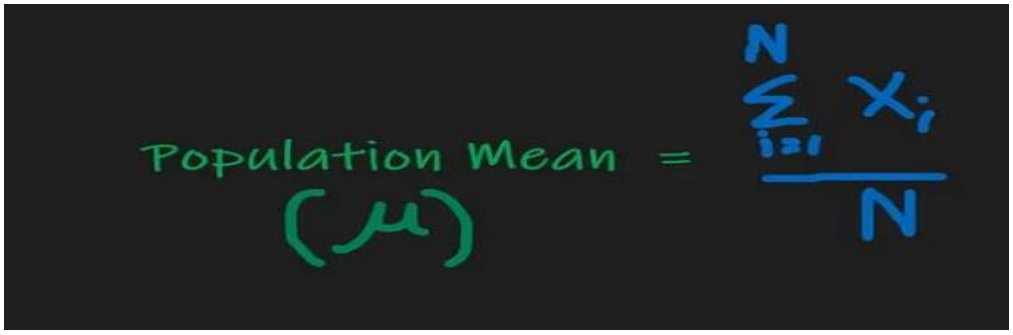
Parameters
Calculating **Mean**, **Variance** and **Standard Deviation** on **Population Data** known to be a **Population parameters**. The population mean and population standard deviation are represented by the Greek letters μ and σ respectively. A parameter is a characteristic of a population.
Statistic
Calculating **Mean($\bar{x}$)**, **Variance** and **Standard Deviation** on **Sample Data** known to be a **Sample statistic**. A statistic is a characteristic of a sample.
If anyone ask and calculate statistic means, you have to calculate **$\bar{x}$, s2 (S Square) and S.**



Population Mean & Sample Mean

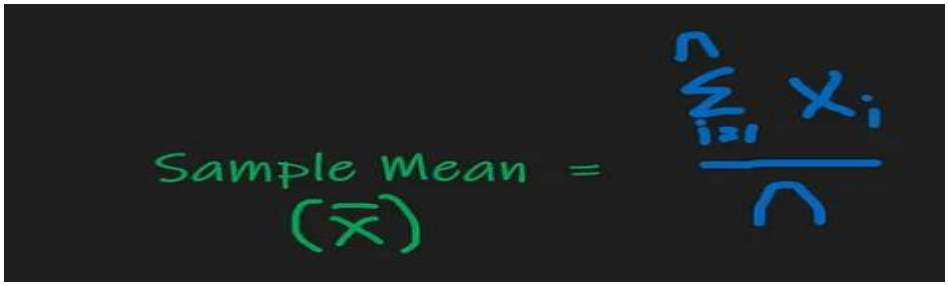
Population Mean

Mean gives the average of the data. If you calculate mean for population data is known as **Population Mean.** Population mean is a fixed one. . . it doesn't vary.



Sample Mean

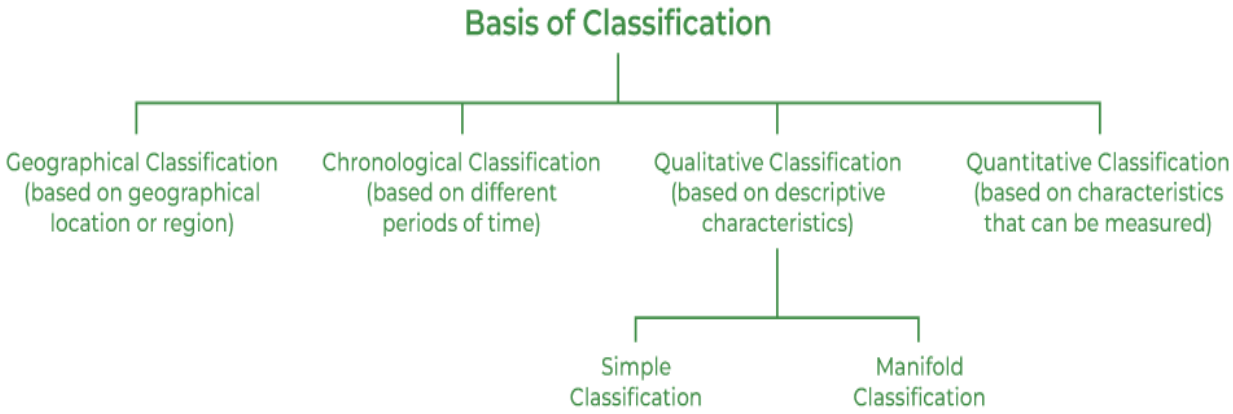
Calculation of mean using Sample data is known as **Sample Mean.** Sample mean vary as our data size/sample size increases. . .



Data presentation:

Classification of data:

The classification of statistical data is done after considering the scope, nature, and purpose of an investigation and is generally done on four bases; viz., geographical location, chronology, qualitative characteristics, and quantitative characteristics.



1. Geographical Classification

The classification of data on the basis of geographical location or region is known as **Geographical** or **Spatial Classification. For example,** presenting the population of different states of a country is done on the basis of geographical location or region.

States	Population (in '000)
Assam	31,205
Bihar	1,04,099
Goa	1,458
Gujarat	60,439
Haryana	25,351

2. Chronological Classification

The classification of data with respect to different time periods is known as **Chronological** or **Temporal Classification. For example,** the number of students in a school in different years can be presented on the basis of a time period.

Year	Number of Students (in '000)
2002	1,349
2007	2,457
2012	2,898
2017	3,145
2022	5,900

3. Qualitative Classification

The classification of data on the basis of descriptive or qualitative characteristics like region, caste,gender, education, etc., is known as **Qualitative Classification.** A qualitative classification can not be quantified and can be of two types; viz., **Simple**

4. Quantitative Classification

The classification of data on the basis of the characteristics, such as age, height, weight, income, etc., that can be measured in quantity is known as **Quantitative Classification. For example,** the weight of students in a class can be classified as quantitative classification.

Weight (in kg)	Number of Students
20-25	5
25-30	18
30-35	6
35-40	10
40-45	9

Frequency Distribution

A frequency distribution shows the frequency of repeated items in a graphical form or tabular form. It gives a visual display of the frequency of items or shows the number of times they occurred. Let's learn about frequency distribution in this article in detail.

What is Frequency Distribution?

Frequency distribution is used to organize the collected data in table form. The data could be marks scored by students, temperatures of different towns, points scored in a volleyball match, etc. After data collection, we have to show data in a meaningful manner for better understanding. Organize the data in such a way that all its features are summarized in a table. This is known as frequency distribution.

Let's consider an example to understand this better. The following are the scores of 10 students in the G.K. quiz released by Mr. Chris 15, 17, 20, 15, 20, 17, 17, 14, 14, 20. Let's represent this data in frequency distribution and find out the number of students who got the same marks.

Quiz Marks	No. of Students
15	2
17	3
20	3
14	2

We can see that all the collected data is organized under the column quiz marks and the number of students. This makes it easier to understand the given information and we can see that the number of students who obtained the same marks. Thus, frequency distribution in statistics helps us to organize the data in an easy way to understand its features at a glance.

Frequency Distribution Graphs

There is another way to show data that is in the form of graphs and it can be done by using a frequency distribution graph. The graphs help us to understand the collected data in an easy way. The graphical representation of a frequency distribution can be shown using the following:

- Bar Graphs: Bar graphs represent data using rectangular bars of uniform width along with equal spacing between the rectangular bars.
- Histograms: A histogram is a graphical presentation of data using rectangular bars of different heights. In a histogram, there is no space between the rectangular bars.
- Pie Chart: A pie chart is a type of graph that visually displays data in a circular chart. It records data in a circular manner and then it is further divided into sectors that show a particular part of data out of the whole part.
- Frequency Polygon: A frequency polygon is drawn by joining the mid-points of the bars in a histogram.

Types of Frequency Distribution

There are four types of frequency distribution under statistics which are explained below:

- **Ungrouped frequency distribution:** It shows the frequency of an item in each separate data value rather than groups of data values.
- **Grouped frequency distribution:** In this type, the data is arranged and separated into groups called class intervals. The frequency of data belonging to each class interval is noted in a frequency distribution table. The grouped frequency table shows the distribution of frequencies in class intervals.
- Relative frequency **distribution:** It tells the proportion of the total number of observations associated with each category.
- Cumulative frequency **distribution:** It is the sum of the first frequency and all frequencies below it in a frequency distribution. You have to add a value with the next value then add the sum with the next value again and so on till the last. The last cumulative frequency will be the total sum of all frequencies.

Frequency Distribution Table

A frequency distribution table is a chart that shows the frequency of each of the items in a data set. Let's consider an example to understand how to make a frequency distribution table using tally marks. A jar containing beads of different colors- red, green, blue, black, red, green, blue, yellow, red, red, green, green, green, yellow, red, green, yellow. To know the exact number of beads of each particular color, we need to classify the beads into categories. An easy way to find the number of beads of each color is to use tally marks. Pick the beads one by one and enter the tally marks in the respective row and column. Then, indicate the frequency for each item in the table.

Categories	Tally Marks	Frequency
Red	𝍸	5
Green	𝍸 |	6
Blue	||	2
Black	|	1
Yellow	|||	3

Thus, the table so obtained is called a frequency distribution table.

Types of Frequency Distribution Table

There are two types of frequency distribution tables: Grouped and ungrouped frequency distribution tables.

Grouped Frequency Distribution Table: To arrange a large number of observations or data, we use grouped frequency distribution table. In this, we form class intervals to tally the frequency for the data that belongs to that particular class interval.

For example, Marks obtained by 20 students in the test are as follows. 5, 10, 20, 15, 5, 20, 20, 15, 15, 15, 10, 10, 10, 20, 15, 5, 18, 18, 18, 18. To arrange the data in grouped table we have to make class intervals. Thus, we will make class intervals of marks like 0 – 5, 6 – 10, and so on. Given below table shows two columns one is of class intervals (marks obtained in test) and the second is of frequency (no. of students). In this, we have not used tally marks as we counted the marks directly.

Marks obtained in Test (class intervals)	**No. of Students (Frequency)**
0 – 5	3
6 – 10	4
11 – 15	5
16 – 20	8
Total	20

Ungrouped Frequency Distribution Table: In the ungrouped frequency distribution table, we don't make class intervals, we write the accurate frequency of individual data. Considering the above example, the ungrouped table will be like this. Given below table shows two columns: one is of marks obtained in the test and the second is of frequency (no. of students).

Marks obtained in Test	No. of Students
5	3
10	4
15	5
18	4
20	4
Total	20

Frequency Distribution Examples

- **Example 1: There are 20 students in a class. The teacher, Ms. Jolly, asked the students to tell their favorite subject. The results are as follows - Mathematics, English, Science, Science, Mathematics, Science, English, Art, Mathematics, Mathematics, Science, Art, Art, Science, Mathematics, Art, Mathematics, English, English, Mathematics.**
  Represent this data in the form of frequency distribution and identify the most-liked subject?
  Solution: 20 students have indicated their choices of preferred subjects. Let us represent this data using tally marks. The tally marks are showing the frequency of each subject.

Subject	Tally Marks	Number of Students
Art	𝍸	4
Mathematics	𝍸 ||	7
Science	𝍸	5
English	𝍸	4

  According to the above frequency distribution, mathematics is the most liked subject.

- **Example 2: 100 schools decided to plant 100 tree saplings in their gardens on world environment day. Represent the given data in the form of frequency distribution and find the number of schools that are able to plant 50% of the plants or more? 95, 67, 28, 32, 65, 65, 69, 33, 98, 96, 76, 42, 32, 38, 42, 40, 40, 69, 95, 92, 75, 83, 76, 83, 85, 62, 37, 65, 63, 42, 89, 65, 73, 81, 49, 52, 64, 76, 83, 92, 93, 68, 52, 79, 81, 83, 59, 82, 75, 82, 86, 90, 44, 62, 31, 36, 38, 42, 39, 83, 87, 56, 58, 23, 35, 76, 83, 85, 30, 68, 69, 83, 86, 43, 45, 39, 83, 75, 66, 83, 92, 75, 89, 66, 91, 27, 88, 89, 93, 42, 53, 69, 90, 55, 66, 49, 52, 83, 34, 36**
  Solution: To include all the observations in groups, we will create various groups of equal intervals. These intervals are called class intervals. In the frequency distribution, the number of plants survived is showing the class intervals, tally marks are showing frequency, and the number of schools is the frequency in numbers.

Number of plants survived	Tally Marks	Number of schools (frequency)
20 - 29	III	3
30 - 39	𝍸 𝍸 IIII	14
40 - 49	𝍸 𝍸 II	12
50 - 59	𝍸 III	8
60 - 69	𝍸 𝍸 𝍸 III	18
70 - 79	𝍸 𝍸	10
80 - 89	𝍸 𝍸 𝍸 𝍸 III	23
90 - 99	𝍸 𝍸 II	12
Total		100

So, according to class intervals starting from 50 – 59 to 90 – 99, the frequency of schools able to retain 50% or more plants are 8 + 18 + 10 + 23 + 12 = 71 schools. Thus, 71 schools are able to retain 50% or more plants in their garden.

Cumulative and Relative frequency distribution:

Cumulative Frequency

Cumulative frequency is the total of a frequency and all frequencies in a frequency distribution until a certain defined class interval. The running total of frequencies starting from the first frequency till the end frequency is the cumulative frequency. The total and the data are shown in the form of a table where the frequencies are divided according to class intervals. Let us learn more about cumulative frequency, plotting a cumulative frequency graph, and learn to read a cumulative frequency table along with solving examples.

Definition of Cumulative Frequency

In statistics, the frequency of the first-class interval is added to the frequency of the second class, and this sum is added to the third class and so on then, frequencies that are obtained this way are known as cumulative frequency (c.f.). A table that displays the cumulative frequencies that are distributed over various classes is called a cumulative frequency distribution or cumulative frequency table. There are two types of cumulative frequency - lesser than type and greater than type. Cumulative frequency is used to know the number of observations that lie above (or below) a particular frequency in a given data set. Let us look at a few examples that are used in many real-world situations.

Example 1: Robert is the sales manager of a toy company. On checking his quarterly sales record, he can observe that by the month of April, a total of 83 toy cars were sold.

Month	**Number of toy cars sold (Frequency)**	**Total number of toy cars sold (Cumulative Frequency)**
January	20	20
February	30	20 + 30 = 50
March	15	50 + 15 = 65
April	18	65 + 18 = 83

Note how the last cumulative total will always be equal to the total for all observations since all frequencies will already have been added to the previous total. Here, 83=20+30+15+1883=20+30+15+18

Example 2: A Major League Baseball team records its home runs in the 2020 session as given below.

Match	f (home runs)	cf (cumulative total)
Qualifying match	11	11
Quarterfinal match	8	11 + 8 = 19
Semifinal	10	19 + 10 = 29
Final	7	29 + 7 = 36

From the above table, it can be observed that the team made 29 home runs before playing in the finals.

Types of Cumulative Frequency

Cumulative frequency is the total frequencies showcased in the form of a table distributed in class intervals. There are two types of cumulative frequency i.e. lesser than and greater than, let us learn more about both types.

Lesser Than Cumulative Frequency

Lesser than cumulative frequency is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate starts from the lowest to the highest size. In other words, when the number of observations is less than the upper boundary of a class that's when it is called lesser than cumulative frequency.

Greater Than Cumulative Frequency

Greater than cumulative frequency is obtained by finding the cumulative total of frequencies starting from the highest to the lowest class. It is also called more than type cumulative frequency. In other words, when the number of observations is more than or equal to the lower boundary of the class that's when it is called greater than cumulative frequency.

Let us look at example to understand the two types.

Example: Write down less than type cumulative frequency and greater than type cumulative frequency for the following data.

Height (in cm)	Frequency (students)
140 – 145	2
145 – 150	5
150 – 155	3
155 – 160	4
160 – 165	1

Solution: We would have less than type and more than type frequencies as:

Height (in inches)	f	Height less than type (upper limits)	c.f	Height less than type (lower limits)	c.f
140 - 145	2	145	2	140	15
145 - 150	5	150	7	145	13
150 - 155	3	155	10	150	8
150 - 160	4	160	14	155	4
160 - 165	1	165	15	160	1

The following information can be gained from either graph or table

- Out of a total of 15 students, 8 students have a height of more than 150 cm
- None of the students are taller than 165 cm
- Only one of the 15 students has a height of more than 160 cm

Constructing a Cumulative Frequency Distribution Table

A cumulative frequency table is a simple visual representation of the cumulative frequencies for different values or categories. To construct a cumulative frequency distribution table, there are a few steps that can be followed which makes it simple to construct. Let us see what the steps are:

- **Step 1:** Use the continuous variables to set up a frequency distribution table using a suitable class length.
- **Step 2:** Find the frequency for each class interval.
- **Step 3:** Locate the endpoint for each class interval (upper limit or lower limit).
- **Step 4:** Calculate the cumulative frequency by adding the numbers in the frequency column.
- **Step 5:** Record all results in the table.

Example: During a 20-day long skiing competition, the snow depth at Snow Mountain was measured (to the nearest cm) for each of the 20 days. The records are as follows: 301, 312, 319, 354, 359, 345, 348, 341, 347, 344, 349, 350, 325,323, 324, 328,322, 332, 334, 337.

Solution:

Given measurements of snow depths are: 301, 312, 319, 354, 359, 345, 348, 341, 347, 344, 349, 350, 325,323, 324, 328,322, 332, 334, 337

Step 1: The snow depth measurements range from 301 cm to 359 cm. To produce the frequency distribution table, the data can be grouped in class intervals of 10 cm each.

In the Snow depth column, each 10-cm class interval from 300 cm to 360 cm is listed.

Step 2: The frequency column will record the number of observations that fall within a particular interval. The tally column will represent the observations only in numerical form.

Step 3: The endpoint is the highest number in the interval, regardless of the actual value of each observation.

For example, in the class interval of 311-320, the actual value of the two observations is 312 and 319. But, instead of using 219, the endpoint of 320 is used.

Step 4: The cumulative frequency column lists the total of each frequency added to its predecessor.

Using the same steps mentioned above, a cumulative frequency distribution table can be made as:

Snow Depth (x)	Tally	Frequency (f)	Endpoint	c.f
300 - 310	|	1	310	1
311 - 320	||	2	320	3
321 - 330	𝍸	5	330	8
331 - 340	|||	3	340	11
341 - 350	𝍸 ||	7	350	18
351 - 360	||	2	360	20

Constructing Cumulative Frequency Distribution Graph

The cumulative frequency distribution of grouped data can be represented on a graph. Such a representative graph is called a cumulative frequency curve or an ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results. In the world of statistics, graphs, in particular, are very important, as they help us to visualize the data and understand it better. So let us learn about the graphical representation of the cumulative frequency. There are two types of Cumulative Frequency

Curves (or Ogives): More than type Cumulative Frequency Curve and Less than type Cumulative Frequency Curve.

More Than Cumulative Frequency Curve

In the more than cumulative frequency curve or ogive, we use the lower limit of the class to plot a curve on the graph. The curve or ogive is constructed by subtracting the total from first-class frequency, then the second class frequency, and so on. The upward cumulation result is more than or greater than the cumulative curve. The steps to plot a more than curve or ogive are:

- **Step 1:** Mark the lower limit on the x-axis
- **Step 2:** Mark the cumulative frequency on the y-axis.
- **Step 3:** Plot the points (x,y) using lower limits (x) and their corresponding Cumulative frequency (y).
- **Step 4:** Join the points by a smooth freehand curve.

Less Than Cumulative Frequency Curve

In the mess than cumulative frequency curve or ogive, we use the upper limit of the class to plot a curve on the graph. The curve or ogive is constructed by adding the first-class frequency to the second class frequency to the third class frequency, and so on. The downward cumulation result is less than the cumulative frequency curve. The steps to plot a less than cumulative frequency curve or ogive are:

- **Step 1:** Mark the upper limit on the x-axis
- **Step 2:** Mark the cumulative frequency on the y-axis.
- **Step 3:** Plot the points (x,y) using upper limits (x) and their corresponding Cumulative frequency (y).
- **Step 4:** Join the points by a smooth freehand curve.

Example: Graph the two ogives for the following frequency distribution of the weekly wages of the given number of workers.

Weekly wages	**No. of workers**
0-20	4
20-40	5
40-60	6
60-80	3

Solution:

Weekly wages	**No. of workers**	**C.F. (Less than)**	**C.F. (More than)**
0-20	4	4	18 (total)
20-40	5	9 (4 + 5)	14 (18 - 4)
40-60	6	15 (9 + 6)	9 (14 - 5)
60-80	3	18 (15 + 3)	3 (9 - 6)

Less than curve or ogive:

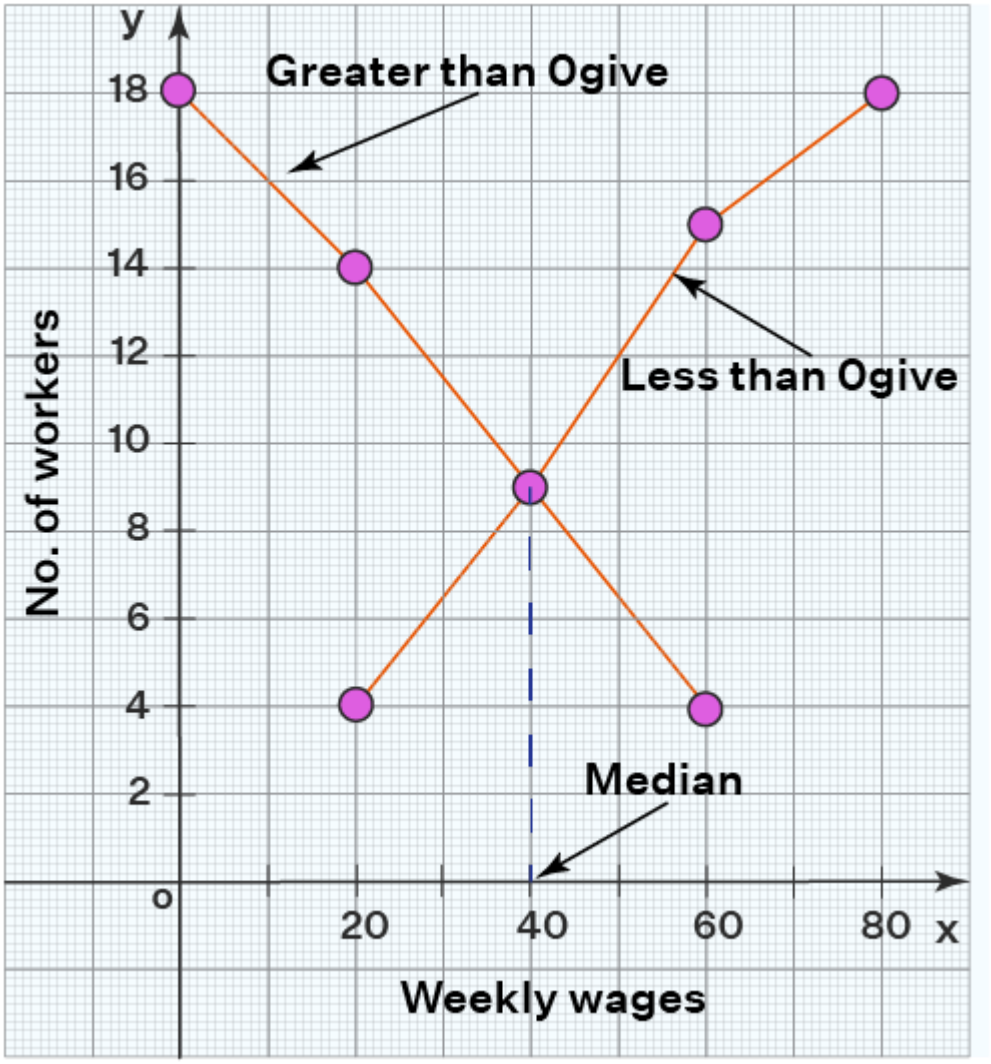
Mark the upper limits of class intervals on the x-axis and take the less than type cumulative frequencies on the y-axis. For plotting less than type curve, points (20,4), (40,9), (60,15), and (80,18) are plotted on the graph and these are joined by freehand to obtain the less than ogive.

Greater than curve or ogive:

Mark the lower limits of class intervals on the x-axis and take the greater than type cumulative frequencies on the y-axis. For plotting greater than type curve, points (0,18), (20,14), (40,9), and (60,3) are plotted on the graph and these are joined by freehand to obtain the greater than type ogive.

A perpendicular line on the x-axis is drawn from the point of intersection of these curves. This perpendicular line meets the x-axis at a certain point, this determines the median. Here the median is 40. The median of the given data could also be found from cumulative graphs. On drawing both the curves on the same graph, the point at which they intersect, the corresponding value on the x-axis, represents the median of the given data set.

The less than and greater than ogives shown in the graph below.



Relative Cumulative Frequency Graph

Relative cumulative frequency graphs are a type of ogive graphs that showcases the percentile of the given data. The ogive shows at what percent of the data is below a particular value. In other words, relative cumulative frequency graphs are ogive graphs that show the cumulative percent of the data from left to right. The two main aspects of this type of graph are, it shows the percentile and indicates the shape of the distribution. Percentiles is the data that is either in the ascending or descending order into 100 equal parts. It indicates the percentage of observations a value is above. Whereas a shape of the distribution helps in transforming observations using standard deviations to see how far specific observations are from the mean. One observation can be compared to another by standardizing the dataset. This particular aspect is widely used in statistics. Let us look at an example:

Example: A car dealer wants to calculate the total sales for the past month and wants to know the monthly sales in percentage after weeks 1, 2, 3, and 4. Create a relative cumulative frequency table and present the information that the dealer needs.

Week	No. of Cars Sold
1	10
2	17
3	14
4	11

Solution:

First total up the sales for the entire month:

10 + 17 + 14 + 11 = 52 cars

Then find the relative frequencies for each week by dividing the number of cars sold that week by the total:

- The relative frequency for the first week is: 10/52 = 0.19
- The relative frequency for the second week is: 17/52 = 0.33
- The relative frequency for the third week is: 14/52 = 0.27
- The relative frequency for the fourth week is: 11/52 = 0.21

To find the relative cumulative frequencies, start with the frequency for week 1, and for each successive week, total all of the previous frequencies

Week	Cars Sold	Relative Frequency	Cumulative Frequency
1	10	0.19	0.19
2	17	0.33	0.19 + 0.33 = 0.52
3	14	0.27	0.52 + 0.27 = 0.79
4	11	0.21	0.79 + 0.21 = 1

Note that the first relative cumulative frequency is always the same as the first relative frequency, and the last relative cumulative frequency is always equal to 1.

Examples on Cumulative Frequency

- **Example 1:** Create a cumulative frequency table showing the number of hours per week that Ryan plays video games, based on the given information.

  Ryan's Game Time

  Monday: 2 hrs
  Tuesday: 1 hr
  Wednesday: 2 hrs
  Thursday: 3 hrs
  Friday: 4 hrs
  Saturday: 2 hrs
  Sunday: 1 hr

  Solution: A cumulative frequency table for Ryan's game time can be made as follows:

Day	Frequency (Hours)	Cumulative Frequency (Hours)
Monday	2	2

Tuesday	1	2 + 1 = 3
Wednesday	2	3 + 2 = 5
Thursday	3	5 + 3 = 8
Friday	4	8 + 4 = 12
Saturday	2	12 + 2 = 14
Sunday	1	14 + 1 = 15

Thus, Ryan spends 15 hours of gaming in a week.

- **Example 2:** A weather forecaster highlights the lows over-night for the past 10 days in a small town in Wisconsin. The temperature readings are given in degrees Fahrenheit and are shown below. Use the data to make a frequency table. 41, 58, 41, 54, 49, 46, 52, 53, 55, 52

  Solution:

  Frequency is nothing but the number of times an event occurs in a given scenario.

  We will first choose a suitable class interval for the above data, then we will enter the frequency values to complete the table.

Interval	**Frequency**
40-44	2
45-49	2
50-54	4
55-59	2

- **Example 3:**

  The following represents scores that a class of 20 students received on their most recent Biology test. Plot a less than type Ogive.

  58, 79, 81, 99, 68, 92, 76, 84, 53, 57, 81, 91, 77, 50, 65, 57, 51, 72, 84, 89

  Solution

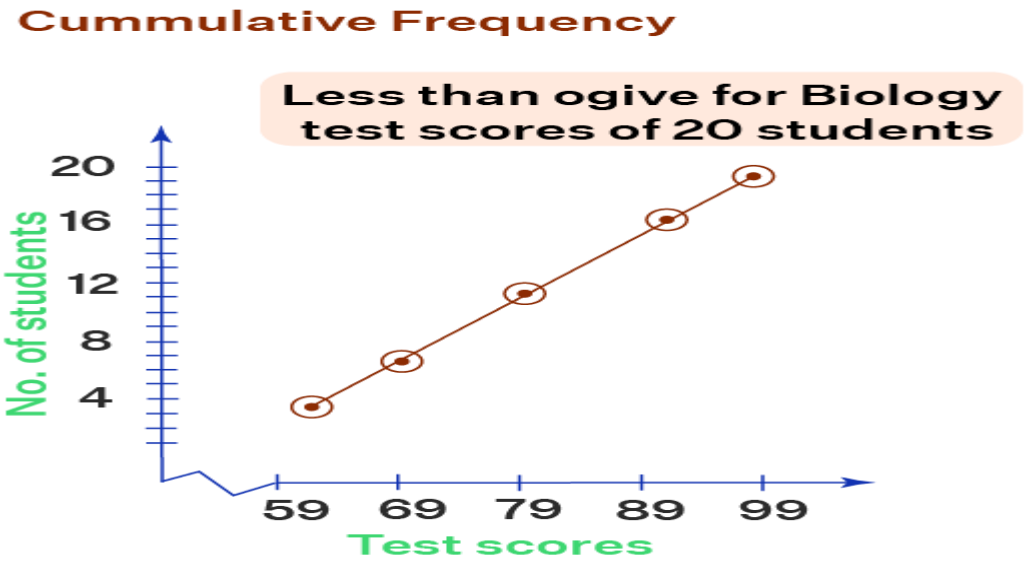
  The cumulative frequency distribution table can be created as:

Interval	**Frequency**	**Less than type Cumulative Frequency**
50-59	6	6
60-69	2	6 + 2 = 8
70-79	4	8 + 4 = 12
80-89	5	12 + 5 = 17

90-99	3	17 + 3 = 20

For plotting a less than type ogive the steps are given below:

- Mark the upper limit on the x-axis.
- Mark the cumulative frequency on the y-axis.
- Plot the points (x,y) using upper limits (x) and their corresponding cumulative frequency (y).
- Join the points using a freehand curve.



Relative frequency :

The number of times an event occurs is called a frequency. **Relative frequency** is an experimental one, but not a theoretical one. Since it is an experimental one, it is possible to obtain different relative frequencies when we repeat the experiments. To calculate the frequency we need

- Frequency count for the total population
- Frequency count for a subgroup of the population

We can find the relative frequency probability in the following way if we know the above two frequencies. The formula for a subgroup is;

Relative Frequency = Subgroup Count / Total Count

How to Calculate Relative Frequency?

The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes gives the value of relative frequency.

Let's understand the Relative Frequency formula with the help of an example

Let's look at the table below to see how the weights of the people are distributed.

Step 1: To convert the frequencies into relative frequencies, we need to do the following steps.

Step 2: Divide the given frequency bt the total N i.e 40 in the above case(Total sum of all frequencies).

Step 3 : Divide the frequency by total number Let's see how : 1/ 40 = 0.25.

Example: Let us solve a few more examples to understand the concepts better.

This is a frequency table to see how many students have got marks between given intervals in Maths.

Marks	**Frequency**	**Relative Frequency**

45 – 50	3	3 / 40 x 100 = 0.075
50 – 55	1	1 / 40 x 100 = 0.025
55 – 60	1	1 / 40 x 100 = 0.075
60 -65	6	6 / 40 x 100 = 0.15
65 – 70	8	8 / 40 x 100 = 0.2
70 – 80	3	3 / 40 x 100 = 0.275
80 -90	11	11 / 40 x 100 = 0.075
90 – 100	7	1 / 40 x 100 = 0.025

It is necessary to know the disparity between the theoretical probability of an event and the observed relative frequency of the event in test trials. The theoretical probability is a number which is calculated when we have sufficient information about the test. If each probable outcome in the sample space is equally likely, then we can consider the number of outcomes of a happening and the number of outcomes in the sample space to calculate the theoretical probability.

The relative frequency is dependent on the series of outcomes resulted in while doing statistical analysis. This frequency can be varied every time we repeat the experiment. The more tests we do during an experiment, the observed relative frequency of an event will get closer to the theoretical probability of the event.

Cumulative Relative Frequency

Cumulative relative frequency is the accumulation of the previous relative frequencies. To obtain that, add all the previous relative frequencies to the current relative frequency. The last value is equal to the total of all the observations. Because all the previous frequencies are already added to the previous total.

Relative Frequency Examples

Example 1: A die is tossed 40 times and lands 6 times on the number 4. What is the relative frequency of observing the die land on the number 4?

Solution: Given, Number of times a die is tossed = 40

Number of positive trial = 6

By the formula, we know,

Relative frequency = Number of positive trial/Total Number of trials

f = 6/40 = 0.15

Hence, the relative frequency of observing the die land on the number 4 is 0.15

Example 2: A coin is tossed 20 times and lands 15 time on heads. What is the relative frequency of observing the coin land on heads?

Solution: Total number of trials = 20

Number of positive trails = 15

By the formula, we know,

Relative frequency = Number of positive trial/Total Number of trials

f = 15/20 = 0.75
Hence, the relative frequency of observing the coin land on heads is 0.75

What is Descriptive Statistics?

Descriptive Statistics, as the name suggests, describes data. It is a method to collect, organize, summarize, display and analyze sample data taken from a population. Descriptive Statistics, unlike inferential statistics, is not based on probability theory. It paves the way to understand and visualize data better.

1. Mean/ Average

This measure of central tendency summarizes the data, by considering a value which is an estimate of the total data set. It helps us to ascertain the spread in variables between the minimum and maximum values.

Sample Mean
Population Mean
SampleData: 12,18,25,69,45
SampleMean: [(12+18+25+69+45)/5]=33.80
PopulationData:55,46,78,12,18,33,28,45,25,69,66
Population Mean: [(55+46+78+12+18+33+28+45+25+69+66)/11)] = 43.18

2. Median

- Median is the middle item in a data set arranged in ascending/descending order.
- If there are n observations then the Median = (n+1)/2 th observation.
- Computational Rule.
- If n is odd, then (n+1)/2 is an integer.
- If n is even, then use an average of n/2 and (n/2) +1th observation.

3. Mode

- Mode is the highest occurring observation.
- The greatest frequency can occur at two or more different values.
- If the data has only two modes, the data is bimodal.
- If the data has more than two modes, the data is multimodal.

4. Percentiles and Quartiles

- The P^{th} percentile in the ordered set is that value below which lies P% (P percent) of the observations in the set.
- The position of the P^{th} percentile is given by (n + 1) P/100, where n is the number of observations in the set.
- Quartiles are special names to percentiles.

Q1=25^{th}percentile
Q2=50thpercentile=median
Q3 = 75th percentile

1. Range

- The range of a data set is the difference between the largest and smallest data values.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.
- Range = $X_{max} - X_{min}$

2. Interquartile Range (IQR)

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

3. Variance

- The variance is a measure of variability that utilizes all the data.

- It is based on the difference between the value of each observation (x_i) and the mean (x for a sample, m for a population).

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$ < - Population variance

Sample variance - > $$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

4. Standard Deviation

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily comparable, than the variance, to the mean.
- If the data set is a sample, the standard deviation is denoted *s*.
- If the data set is a population, the standard deviation is denoted *s* (sigma).

5. Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.
- If the data set is a sample, the coefficient of variation is computed as follows:

$$\frac{s}{\bar{x}}(100)$$

- If the data set is a population, the coefficient of variation is computed as follows:

$$\frac{\sigma}{\mu}(100)$$

Standard Deviation:

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

The population standard deviation formula is given as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

Here,

- σ = Population standard deviation symbol
- μ = Population mean
- N = total number of observations

Similarly, the sample standard deviation formula is:

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

Here,

- s = Sample standard deviation symbol
- $\bar{x}$= Arithmetic mean of the observations
- n = total number of observations

Standard Deviation Examples

Example 1: There are 39 plants in the garden. A few plants were selected randomly and their heights in cm were recorded as follows: 51, 38, 79, 46, 57. Calculate the standard deviation of their heights.

Solution:

n = 5

Sample mean ($\bar{x}$)= (51+38+79+46+57)/5 = 54.2

Since, sample data is given, we use the sample SD formula.

$$SD = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(51-54.2)^2+(38-54.2)^2+(79-54.2)^2+(46-54.2)^2+(57-54.2)^2}{4}}$$

= 15.5

Answer: Standard deviation for this data is 15.5

Example 1: Using descriptive statistics, find the mean and mode of the given data.
{1, 4, 6, 1, 8, 15, 18, 1, 5, 1}
Solution: Total number of observations = 10
Sum of observations = 1 + 4 + 6 + 1 + 8 + 15 + 18 + 1 + 5 + 1 = 60
Mean = 60 / 10 = 6
Mode = Most frequently occurring observation = 1
Answer: Mean = 6, Mode = 1

Example 2: In a class of 50, 4 students were selected at random and their total marks in the final assessments are recorded, which are: 812, 836, 982, and 769. Find the variance and standard deviation of their marks.

Solution:

n = 4

Sample Mean ($\bar{X}$) = (812+836+982+769)/4 = 849.75

Here also, we have to calculate the sample standard deviation as the given data is just a sample.

$$\text{Variance} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{4}(X_i - 849.75)^2}{3}$$

$= [(812 - 849.75)^2 + (836 - 849.75)^2 + (982 - 849.75)^2 + (769 - 849.75)^2] /3$

= 8541.58

Using the SD formula,

SD = $\sqrt{8541.58}$ = 92.4

Answer: Variance is 8541.58 and standard deviation for this data is 92.4

Example 2: Find the sample variance of the following data

{7, 11, 15, 18, 36, 43}

Solution: Sample variance formula, $s^2 = \sum \frac{(X_i - \overline{X})^2}{n-1}$

Mean, $\overline{X}$ = 21.67, n = 6

$s^2 = [(7 - 21.67)^2 + (11 - 21.67)^2 + (15 - 21.67)^2 + (18 - 21.67)^2 + (36 - 21.67)^2 + (43 - 21.67)^2] / 6 - 1$

= 209.47

Answer: s^2 = 209.47

Example 3: Find the median and the mean deviation about the median for the given data

{9, 10, 12, 16, 17, 17, 18, 20}

Solution: n = 8

Median = $[(n / 2)^{th}$ term + $((n / 2) + 1)^{th}$ term$] / 2$

= $[(8 / 2)^{th}$ term + $((8 / 2) + 1)^{th}$ term$] / 2$

= $(4^{th}$ term + 5^{th} term$) / 2$

= $(16 + 17) / 2 = 16.5$

Mean deviation about median = $\sum_1^n \frac{|X-16.5|}{n}$

= [|9 - 16.5| + |10 - 16.5| + |12 - 16.5| + |16 - 16.5| + |17 - 16.5| + |17 - 16.5| + |18 - 16.5| + |20 - 16.5|] / 8

= 3.125

Answer: Median = 16.5, Mean deviation about median = 3.125

Graphical Statistics:

A **statistical graph** is a graph that organizes data, allowing a clearer visualization.
Two of your friends are excellent cooks, so they decide to start up a business to make some extra money during summer. They decide to sell artisan ice cream, but since they will be working in a small kitchen, they will not be able to sell a wide variety of ice cream flavors.
To decide which flavors they should focus on, you run a survey around your neighborhood asking for favorite ice cream flavors. You organize data into the following frequency table.

Flavor	**Frequency**
Chocolate	15
Vanilla	14
Strawberry	9
Mint-Chocolate	3
Cookie Dough	9

Table 1. ice cream flavors, statistical graphs.

As you are going back with your friends to communicate your findings, you realize they might be tired because of the kitchen set-up. Because of this, you first decide to make a friendlier display of data, so they do not have to look at raw numbers.
It is time to see what options you have for displaying your ice cream flavor survey.

Bar Charts

Bar charts are pretty straightforward. You line up the different categories of your survey and draw the bars depending on the frequency of each categorical variable. The higher the frequency, the taller the bar.

There are two ways of drawing bar charts: Using vertical bars and using horizontal bars.
The most common type of bar charts are those that use vertical bars. To draw a vertical bar chart, you first need to write the different categories on the horizontal axis and then the range of frequencies on the vertical axis. For your ice cream flavors example, this will look like this:

Figure 1. Empty bar chart

Next, you will need to draw bars whose height goes all the way up to the frequency of each variable. Usually, different colors are used, and the width of the bars is chosen such that the bars are not adjacent to each other.

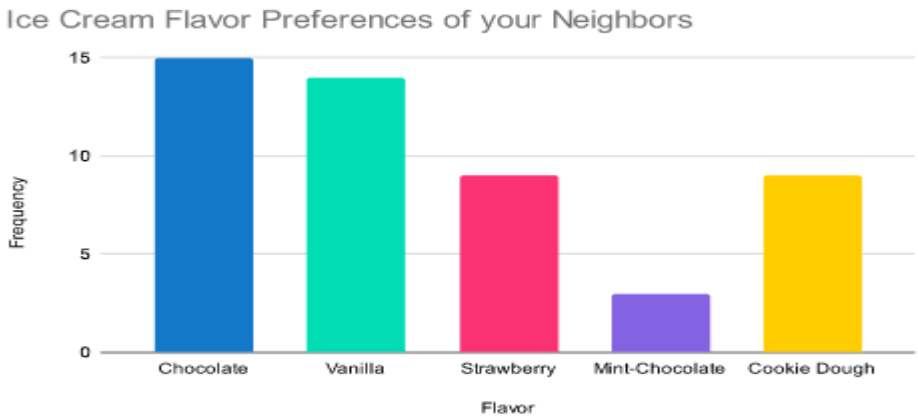


Figure 2. Vertical bar chart of the favorite flavors of ice cream of your neighbors

To draw a horizontal bar chart you follow the same idea, but now the variables are aligned vertically, while the frequencies are aligned horizontally.

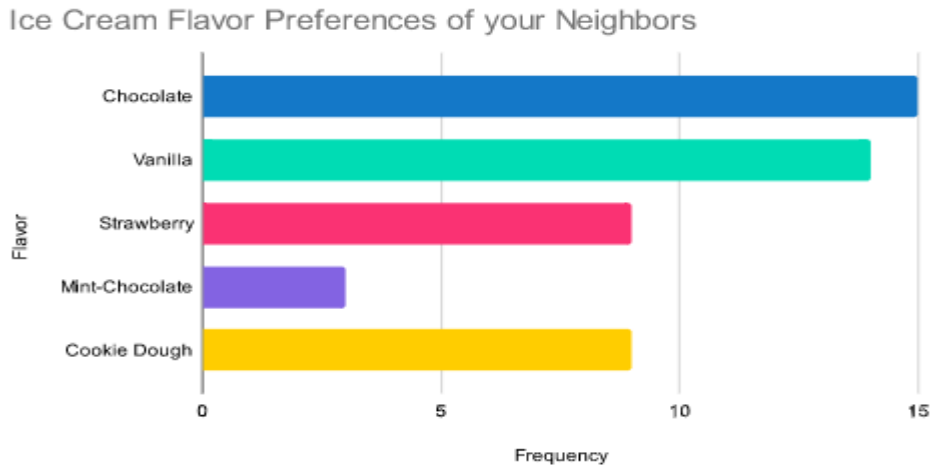


Figure 3. Horizontal bar chart of the favorite flavors of ice cream of your neighbors

Pie Charts

Pie charts are a very common way of displaying data. They picture the whole population as a circle, which is segmented into the different categories of your survey. The bigger the frequency of a category, the bigger the portion of the circle.

Because pie charts divide a circle into sectors, they are also known as **sector charts**.

To make a pie chart, you will need to do a **relative frequency table**, which is the same frequency table but with a column that shows the relative frequency of each category.

You can find the relative frequency by dividing the respective frequency by the total of inquiries (which is equal to the sum of all the frequencies).

To find the relative frequency of the chocolate flavor, you first need to note that your survey consists of 50 inquiries. Then, you need to divide the frequency of the chocolate flavor by this number, that is

15/50=0.3

Usually, you will need to write this as a percentage, so multiply it by 100. This means that the relative frequency is 30%.

This relative frequency corresponds to the percentage of the population that falls within each category. Here is a table with the relative frequency of the rest of the ice cream flavors.

Flavor	**Frequency**	**Relative Frequency**
Chocolate	15	30%
Vanilla	14	28%
Strawberry	9	18%
Mint-Chocolate	3	6%
Cookie Dough	9	18%

Table 2. ice cream flavors, statistical graphs.

Be sure that the relative frequencies add up to 100%.

Now that you know the relative frequencies of each category, you can proceed to draw the pie chart. Remember that the relative frequency tells you the percentage of the circle of each category.

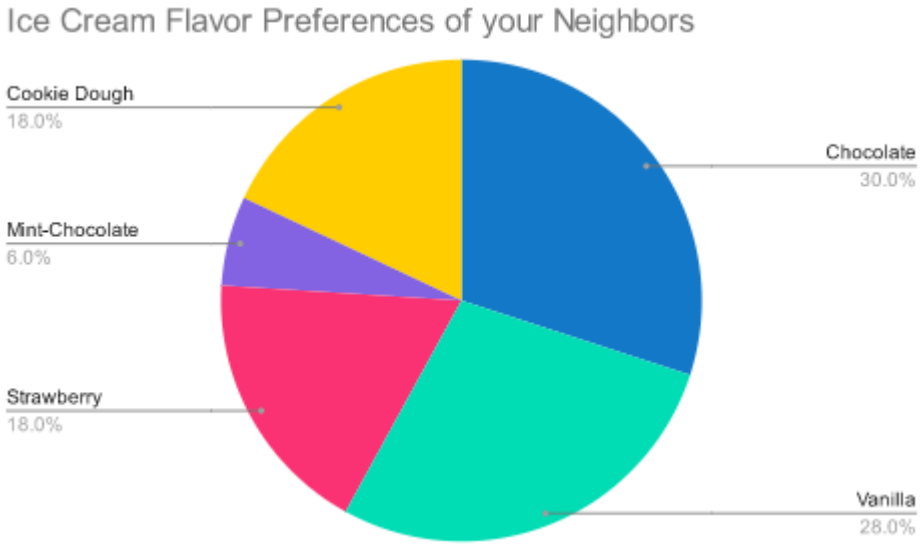


Figure 4. Pie chart of the favorite flavors of ice cream of your neighbors

Segmented Bar Charts

Segmented bar charts are practically a hybrid between a bar chart and a pie chart, closer to a pie chart. Instead of using a circle and dividing it into sectors, you divide a big bar into segments, where each segment represents a category.

Segmented bar charts are typically used when needing to compare two or more data sets. In the ice cream example, suppose you want to expand your survey to the next neighborhood, this way you can have a better picture of which ice cream flavors your friends should focus on. Here is a table of the survey on neighborhood B.

Flavor	**Frequency**	**Relative Frequency**
Chocolate	16	32%
Vanilla	12	24%
Strawberry	7	14%
Mint-Chocolate	5	10%
Cookie Dough	10	20%

Table 3. ice cream flavors, statistical graphs.

Since the goal of segmented bar charts is to compare two data sets, a table with the relative frequency of both neighborhoods will be very useful.

Flavor	**Relative Frequency B**	**Relative Frequency B**
Chocolate	30%	32%
Vanilla	28%	24%
Strawberry	18%	14%
Mint-Chocolate	6%	10%
Cookie Dough	18%	20%

Table 4. ice cream flavors, statistical graphs.

You can now draw the segmented bar chart. Usually, the two data sets are put next to each other for means of comparison.

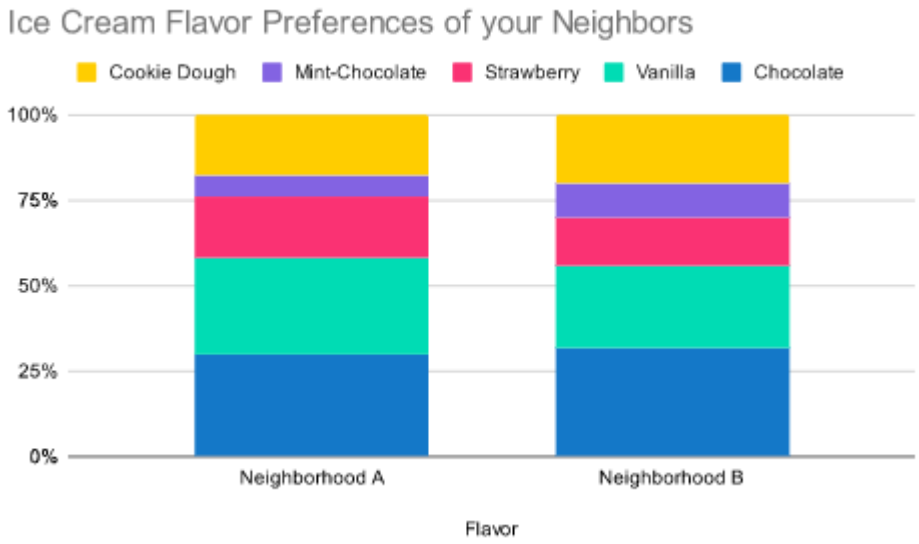


Figure 5. Segmented bar chart of the favorite flavors of ice cream of two neighborhoods

Segmented bar charts usually display the relative frequency of the data, so you will also need a table with relative frequencies to draw a segmented bar chart. You can also use segmented bar charts to represent the actual frequencies of your data, you just need to make sure that you use an adequate scale.