# Chapter 1

## Introduction to statistics

Statistics ıs the art of learning from data. It is concerning with the collection of data, their subsequent descriptıon and their analysis, which often leads to the drawing of conclusions,

Exp: to check the effect of a new tablet/drug on a group of patients

## Data vs Information:

**Data** is the raw material that is to be processed for information or for collection of details. It is unorganized data or facts that are to be processed. Data is plain fact and it has to be processed for further information. Data is alone enough to get details and find the meaning of something. Data is the computers language. Data is useless unless it is processed or has been made into something. Data has no meaning when it has not been interpreted. Data comes in figures, dates and numbers and is not processed.

Examples of Data

- Student Data on Admission Forms: When students get admission in a college. They fill admission form. This form contains raw facts (data of student) like name, father's name, address of student etc.
- Data of Citizens: During census, data of all citizens is collected.
- Survey Data: Different companies collect data by survey to know the opinion of people about their product.
- Students Examination data: In examination data about obtained marks of different subjects for all students is collected.

**Information** is processed data. The data that can be made useful is known as information. Information is basically the data plus the meaning of what the data was collected for. Data does not depend upon information but information depends upon data. It cannot be generated without the help of data. Information is something that is being conveyed. Information is meaningful when data is gathered and meaning is generated. Information cannot be generated without the help of data. Information is the meaning that has been formed with the help of data and that meaning makes sense because of the data that has been collected against the word. Information is processed and comes in a meaningful form.

## Descriptive statıstics:

The part of statistics concerned with the description and sumarization of data.

Exp: at the end of the experiment, the data should be described,

## İnferential statistics:

The part of statistics concerned wıth the drawing of conclusions from data. To be able to draw a conclusion from the data, we must take into account the possibility of chance.

To be able to draw logical conclusions from data, it is usually necessary to make some assumptions about the chances or probabilities of obtaınıng the different data values. The totality of these assumptions is reffered to as a probability model to describe the data, an understanding of statistical inference requires some knowledge of the theory of probability.

## Population:

The total collection of all the elements that we are interested is called a population (a collection of elements of interest).

## Sample:

A subgroup of the population that will be studied in detail is called a sample.

Probability model:

The mathematical assumptions relating to the lıkelihood of different data values.

Representative means sample:

The sample is chosen in such a way that all possible choices of the k members are equally likely (lottery)

## Stratified random sampling:

In this type of sample, first the population is stratified into sub-populations and then the correct number of elements is randomly chosen from each of the subgroups.

Exp: 2000 students in one school, 300 first-year class, 500 in second year class, 600 in $3^{rd}$ class, and 600 in the $4^{th}$ year class. We want to select 40 students, so we select 10 students from every class,

# Chapter 2

## Describing data Sets

In this chapter we learn methods for presenting and describing sets of data. We introduce different types of tables and graphs, which enable us to easily use key features of a data set.

### Frequency:

Is the number of times that a given value occurs in a data set

### Frequency table:

Present each distinct value with its frequency of occurrence.

Frequency table of students that enter to our department in different years

| year | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|
| Frequency | 75 | 50 | 82 | 95 | 98 |

### Relative Frequency table:

Present each distinct value with its percentage of frequency.

Relative Frequency table of students that enter to our department in different years

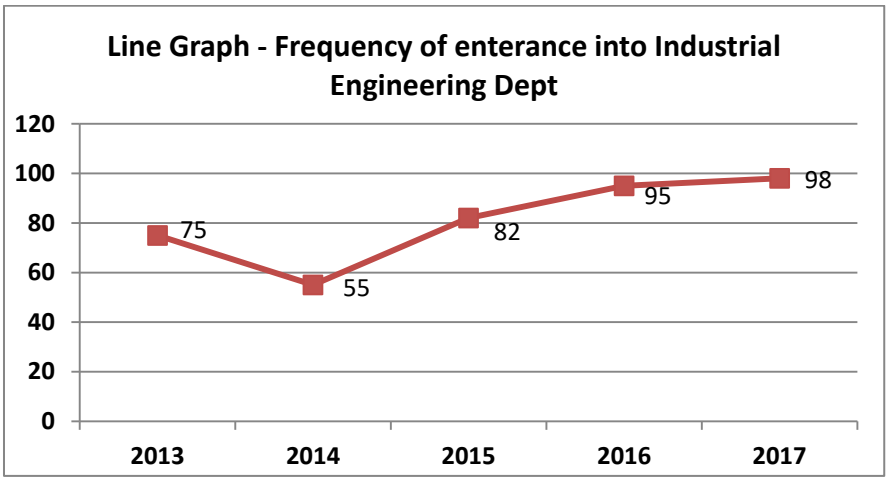
| year | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|
| Relative Frequency | 75/400=0.1875 | 0.125 | 0.205 | 0.2375 | 0.245 |

Note: summation of relative frequencies should be equal to 1.

### Line Graph:

Data from a frequency table can be graphically pictured by a line graph, which plots the successive values on the horizontal axis and indicates the corresponding frequency by the height of a vertical line.

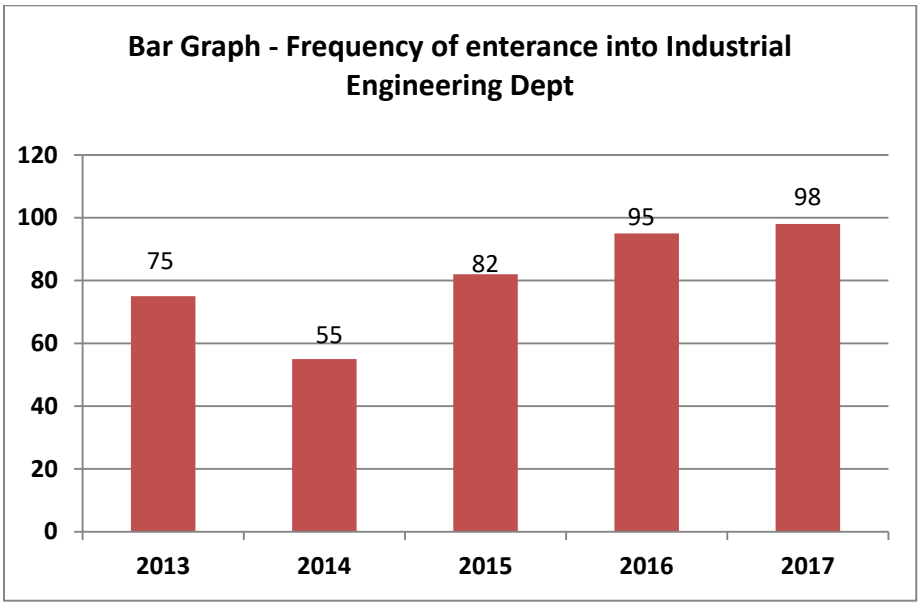
Example: draw line graph for frequency table



## Bar Graph:

Sometimes the frequencies are represented not by lines but rather by bars having some thickness. These graphs called bar graphs.

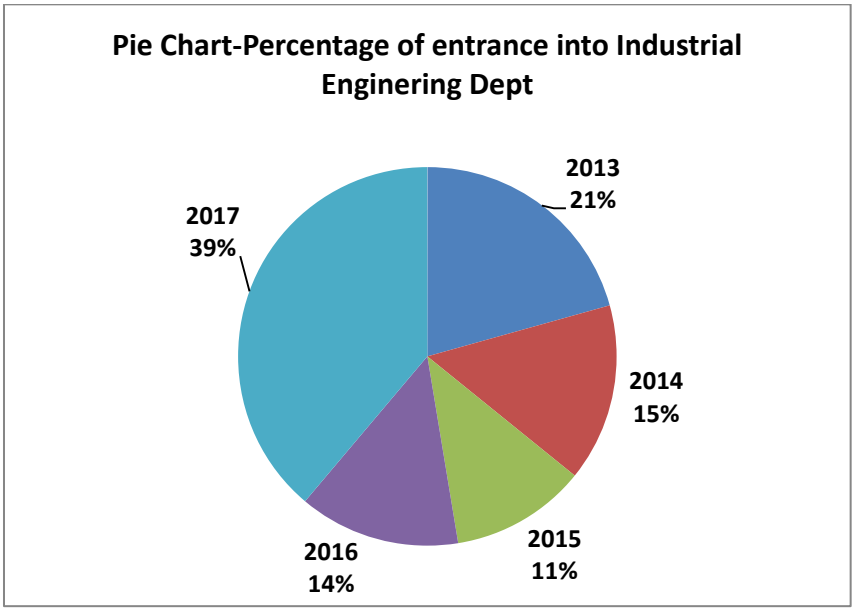
Example: draw Bar graph for frequency table




## Pie chart:

This chat is often used to plot relative frequency when the data are non-numeric. A circle is constructed and then is sliced up into distinct sectors, one for each different data value,

Example: draw Pie chart for frequency table



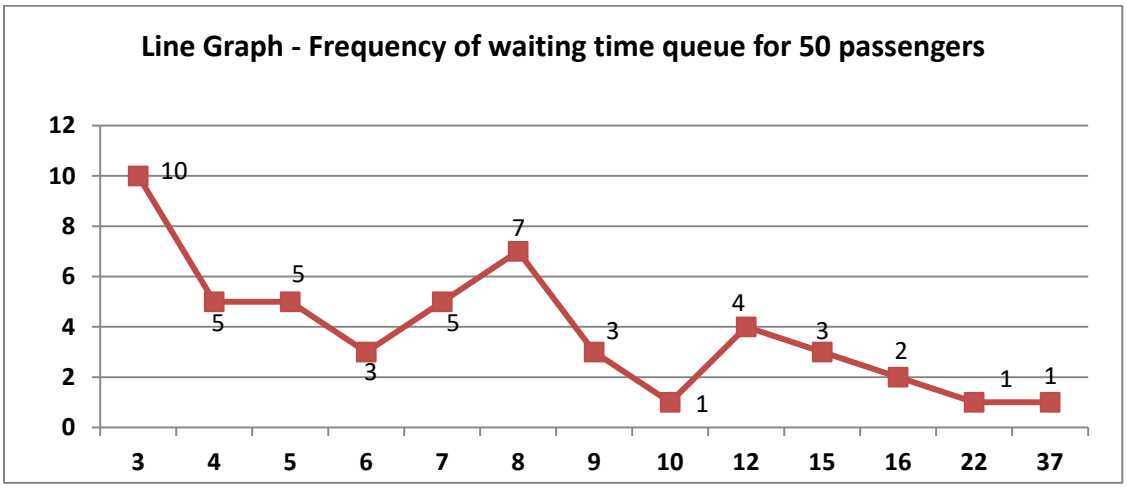
Example: the following data represent the waiting time queue for 50 passengers.

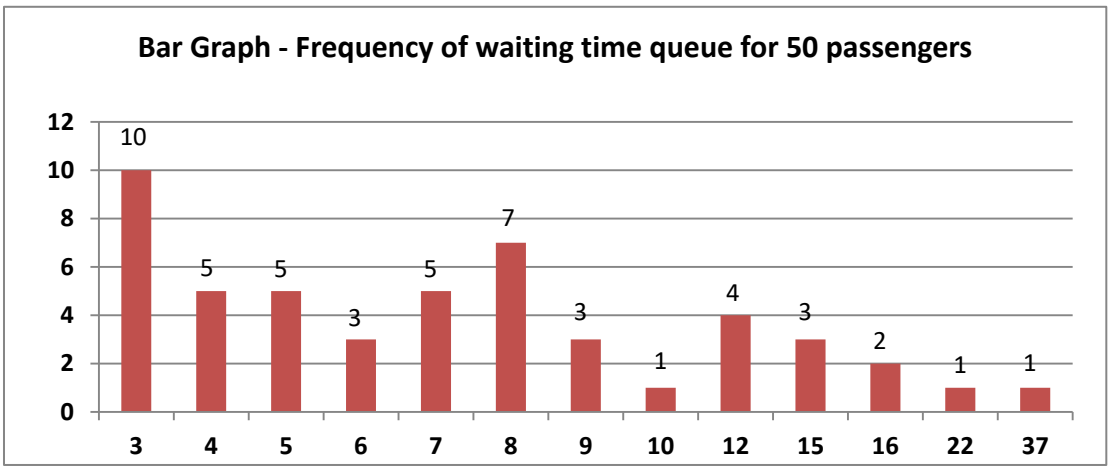
8, 4, 6, 3, 7, 3, 7, 5, 4, 8, 3, 7, 15, 16, 15, 8, 4, 4, 3, 3, 9, 5, 12, 8, 7, 5, 9, 3, 8, 9, 22, 10, 3, 37, 7, 6, 8, 3, 5, 16, 4, 15, 3, 12, 6, 8, 12, 12, 3, 5

A) Construct a frequency and relative frequency table for these data
B) Using a line graph, plot the data
C) Using a bar graph, plot the data
D) Suppose we define the following categories for waiting time
Waiting time $t \leq 6$, $6 < t \leq 12$, $12 < t \leq 20$, $t > 20$. Plot the relative frequencies using a pie chart.

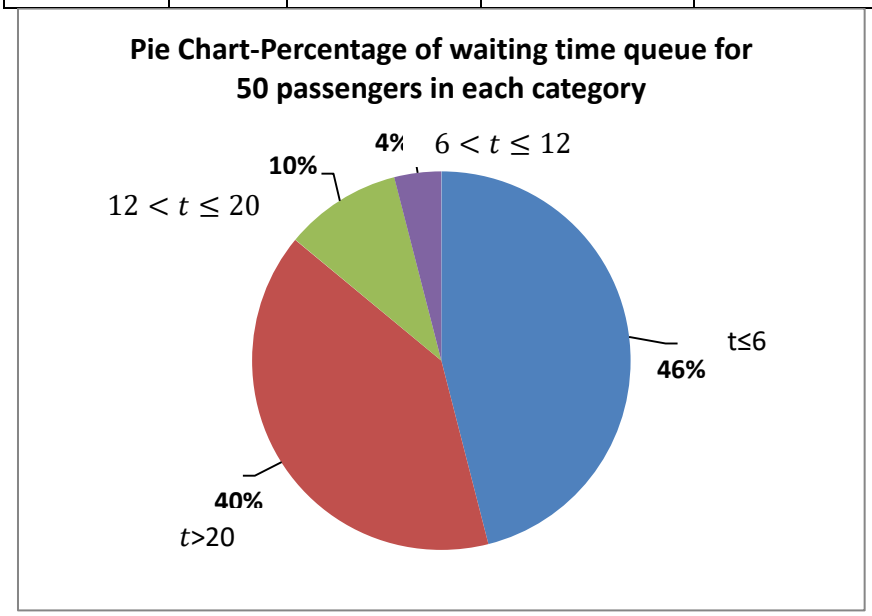
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 7 |
| 7 | 7 | 7 | 7 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 9 |
| 9 | 9 | 10 | 12 | 12 | 12 | 12 | 15 | 15 | 15 | 16 | 16 |
| 22 | 37 | | | | | | | | | | |

| Time | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 16 | 22 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 10 | 5 | 5 | 3 | 5 | 7 | 3 | 1 | 4 | 3 | 2 | 1 | 1 |
| Relative Frequency | 0.2 | 0.1 | 0.1 | 0.06 | 0.1 | 0.14 | 0.06 | 0.02 | 0.08 | 0.06 | 0.04 | 0.02 | 0.02 |



| Time | $t \leq 6$ | $6 < t \leq 12$ | $12 < t \leq 20$ | $t > 20$ |
|---|---|---|---|---|
| Frequency | 23 | 20 | 5 | 2 |
| RF | 0.46 | 0.4 | 0.1 | 0.04 |




## Histogram:

A graph, which the data are divided into class intervals, whose frequencies are shown in a bar graph.

- The end points of a class interval are called the class boundaries
- the class interval $a - b$ contains, $a \leq x < b$
- Number of intervals, $k = \sqrt{n}$ (round it up), where is the number of observations
- Class intervals=CI=Range/$\sqrt{n}$(use exact value of $\sqrt{n}$), where range=$x_{max} - x_{min}$

To construct a histogram from a data set:

1- Arrange the data in increasing order
2- Choose class intervals so that all data points are covered
3- Construct a frequency table/relative frequency table
4- Draw the bar graphs having heights determined by the frequencies in step 3

Relative frequency:

if f represents the frequency of Occurrence of some data value x then the relative frequency $\boldsymbol{f_n}$ where n is the total number of observations in data set.

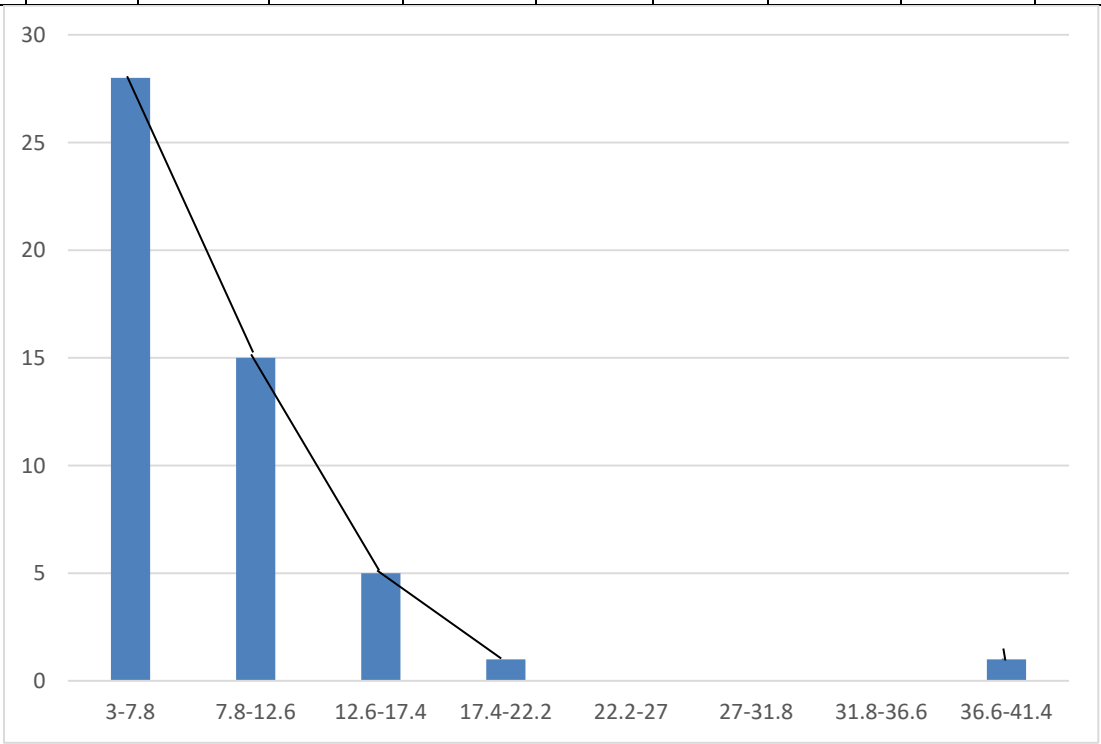
Example: By information of waiting time (previous example),

a) Find appropriate number of classes and class interval.
b) Construct frequency table and relative frequency table and histogram with using frequency table.
c) Is it symmetric or not. If not, what is its skewness?

Number of intervals, $k = \sqrt{n} = \sqrt{50} = 7.07 = 8$

Class intervals=CI=R/$\sqrt{n} = \frac{37-3}{\sqrt{50}} = 4.808 = 4.8$

| Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\sum$ |
|---|---|---|---|---|---|---|---|---|---|
| CI | 3-7.8 | 7.8-12.6 | 12.6-17.4 | 17.4-22.2 | 22.2-27 | 27-31.8 | 31.8-36.6 | 36.6-41.4 | |
| F | 28 | 15 | 5 | 1 | 0 | 0 | 0 | 1 | 50 |
| RF | 0.56 | 0.3 | 0.1 | 0.02 | 0 | 0 | 0 | 0.02 | 1 |




Because of long tail in right side, so our histogram is slewed to the right or positively skewed

Example: consider the following data:

4.1, 3.3, 2.4, 3.7, 2.1, 0.8, 2.1, 10, 3.6, 2.1, 5.1, 3.1, 2.2, 2.4, 2.2, 1.9, 1.8, 5.8, 2.9, 4.5, 5, 7.1

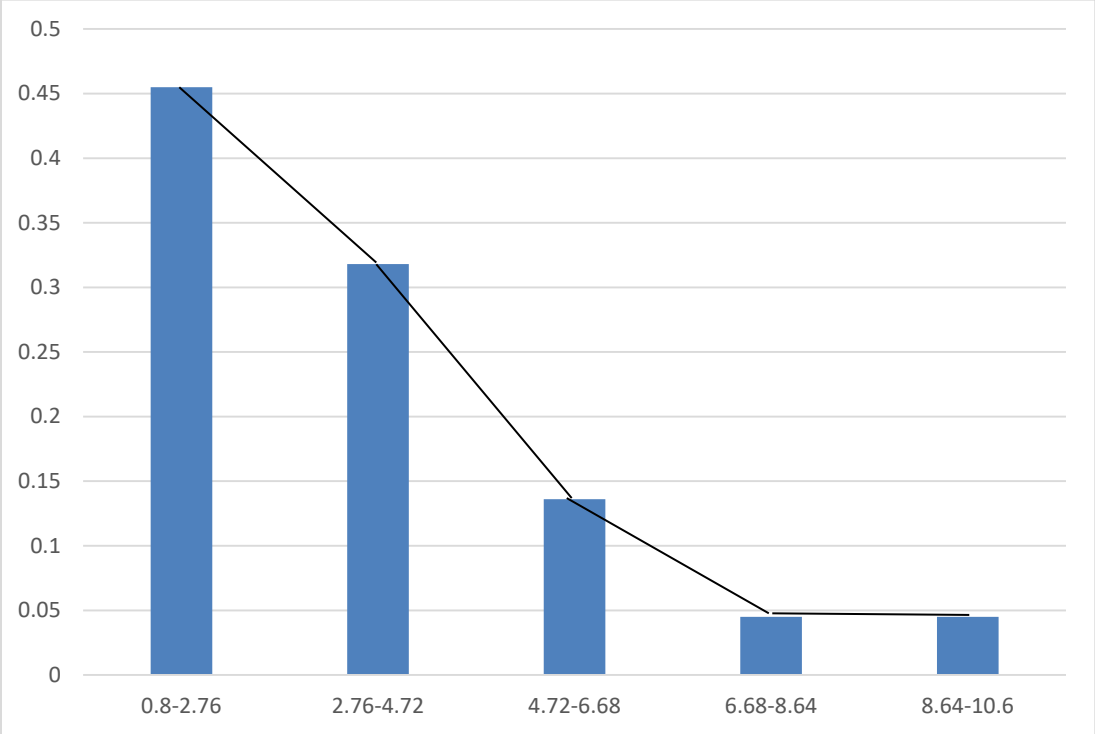
a) Find appropriate number of classes and class interval
b) Construct frequency table and draw histogram by using relative frequency
c) Is it symmetric or not. If not, what is its skewness?

0.8 1.8 1.9 2.1 2.1 2.1 2.2 2.2 2.4 2.4 2.9 3.1 3.3
3.6 3.7 4.1 4.5 5 5.1 5.8 7.1 10

Number of intervals, $k = \sqrt{n} = \sqrt{22} = 4.69 = 5$

Class intervals=CI=R/$\sqrt{n} = \frac{10-0.8}{\sqrt{22}} = 1.96$

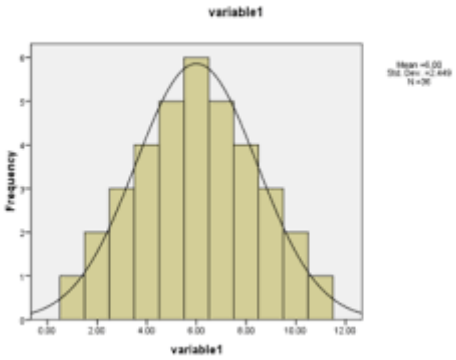
| Interval | 1 | 2 | 3 | 4 | 5 | $\sum$ |
|---|---|---|---|---|---|---|
| CI | 0.8-2.76 | 2.76-4.72 | 4.72-6.68 | 6.68-8.64 | 8.64-10.6 | |
| F | 10 | 7 | 3 | 1 | 1 | 22 |
| RF | 0.455 | 0.318 | 0.136 | 0.045 | 0.045 | 1 |



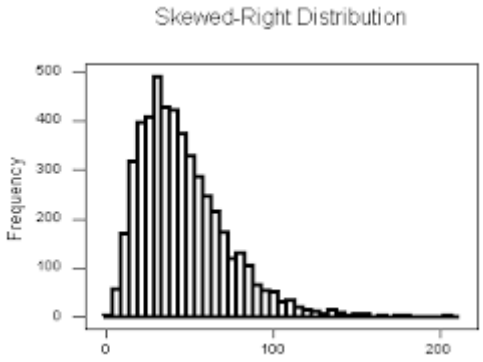

Because of long tail in right side, so our histogram is slewed to the right or positively skewed

**Symmetry and Skewness:**

- A histogram is perfectly symmetric if it's right half is a mirror image of left half

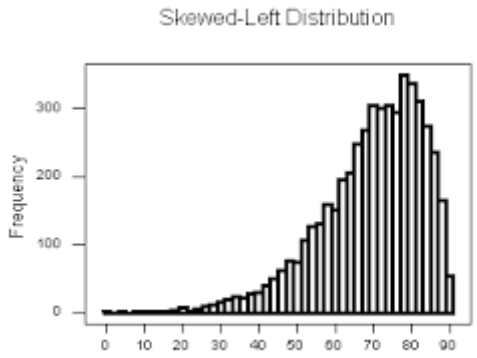



- Histograms that are not symmetric are referred to as skewed.
- A histogram with a long right-hand tail is said to be skewed to the right or positively skewed.




- **if the mean is bigger than the median that is the histogram is skewed right.**

- A histogram with a long left-hand tail is said to be skewed to the left or negatively skewed.




- **if the mean is smaller than the median that is the histogram is skewed left.**