# Chapter 01

## Basics and statistical vocabulary

- Statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us. They provide ways of gaining new insights into the behavior of many phenomena that you will encounter in your chosen field of specialization in engineering or science.

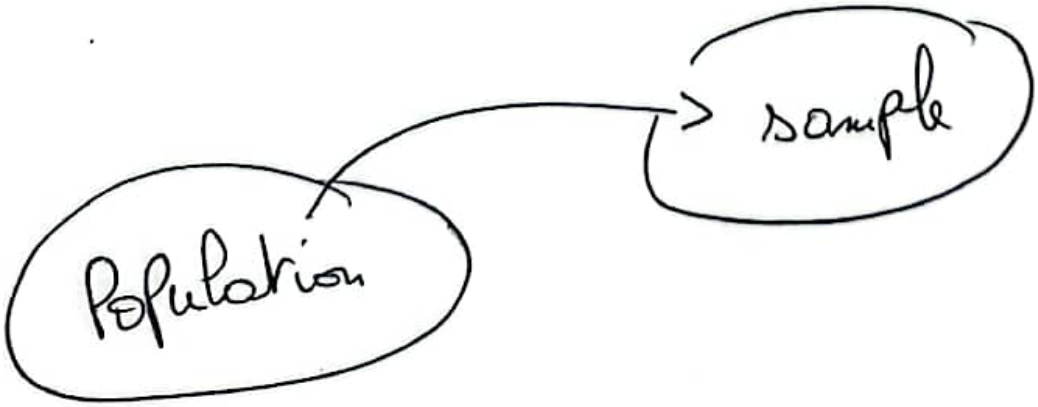
- The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence in the presence of uncertainty and variation

- Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

- We can collect data by gathering and counting. to-ling surveys, giving out questionnaires or by taking measurements. We display and analyse data so that We can describe the things, both physical and social, that we see and experience around us. We can also find answers to questions that might not be immediately obvious, and We can also identify questions for further investigation.

## (1) The population and the sample:

- In the language of statistics, one of the most basic concepts is sampling. In most statistical problems, a specified number of measurements or data - a sample- is drawn from a much larger body of measurements, called the population.



- When we use statistical language, we distinguish between the set of objects on which the measurements are taken and the measurements themselves. To experimenters, the objects on which mesurements are taken are called experimental units. the sample survey statistician calls them elements of the sample.

**Definition ①:** A population is the set of all measurements of interest to the investigator.

**Definition ②:** A sample is a subset of measurenat selected from the population of interest.

**Definition ③:** An experimental unit is the individua or object on which a variable is measured A single measurement or data value result When a variable is actually measured on an experimental unit.

## ② Variables and Types of data:

Now that you know that statistics can describe the whole population based on information gathered from a population sample we will move on to Exploratory Data Analysis (EDA).
Data we observe will be called the variables and their values variable variants.

**Definition ④:** A variable is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.

- Because the way of processing variables depends most on their Type, we will now explore how variabl are devided into different categories.
- Variables can be classified into one of two categories: qualitative or quantitative

**Definition ⑤:** Qualitative variables (or categorial) are described by words and are non-numerical, mesure a quality or characteristic on each experimental unit. Such as:

① Blood types, colours.

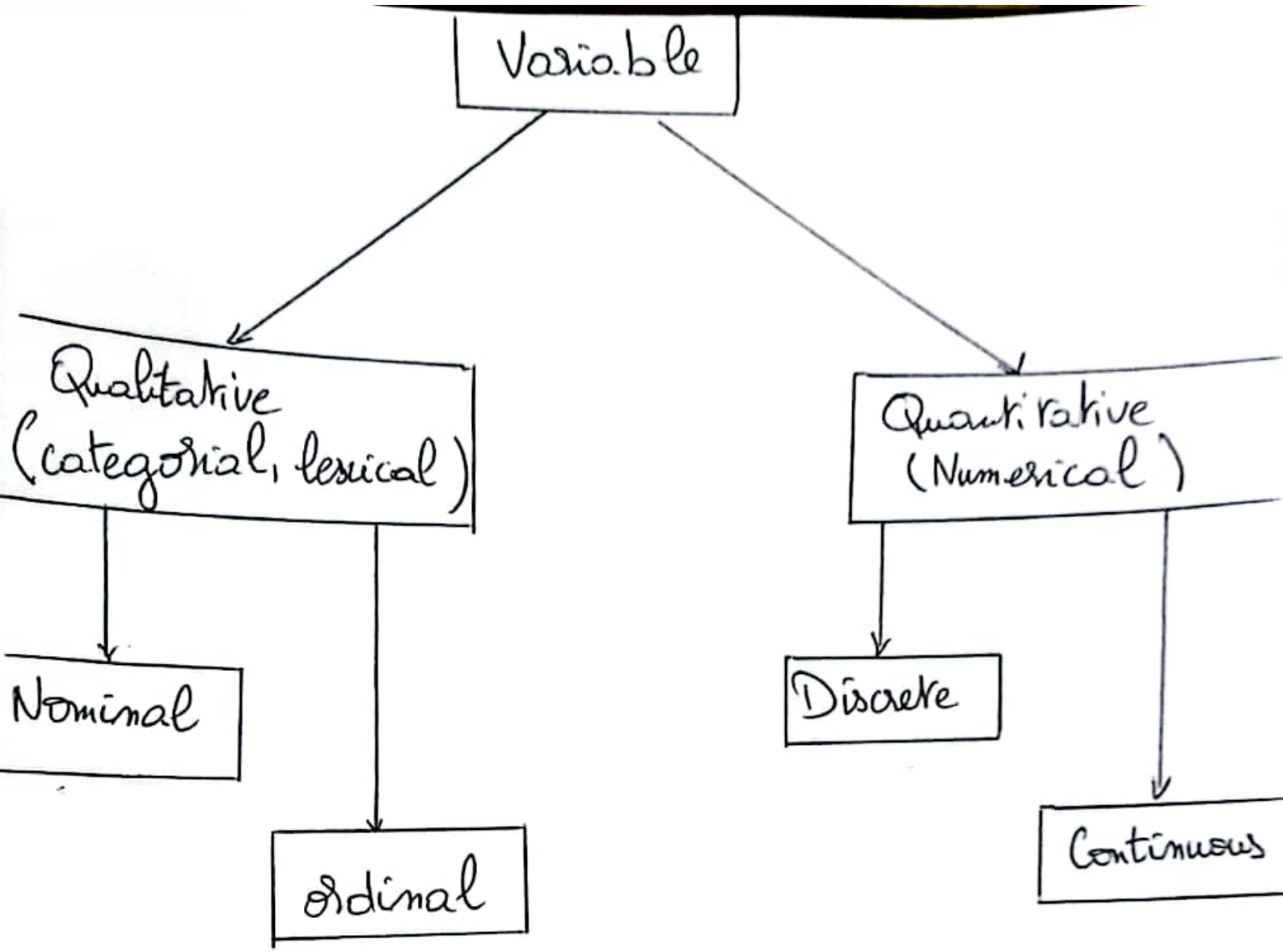
② Taste ranking: excellent, good, fair poor.

③ Color of an M & M's candy: brown, yellow, red, orange, green, blue.

**Definition ⑥:** Quantitative variables measure a numerical quantity or amount on each experimental unit. and are either discrete or continuous. Qualitative variables, often represented by the letter $x$, produce numerical data, such as those listed here:

$x$ = Volume of orange juice in a glass

$x$: Number of passengers on a flight.




The variables division is shown in the above diagram.

**Definition (7):** ho **Nominal variable** has equivalent variants. it is impossible to either compare them or sort them (sex, nationality, ...)

**Definition (7):** ordinal variable

Forms a transition between qualitative and quantitative variables: individual variant can be sorted and it is possible possible to compare one another (cloth sizes: S, M, L and XL)

---

**Definition (8):**

- As a general rule, discrete data are counted and cannot be made more precise (can assume only a finite or countable number of values).
- Whereas continuous data are measurements that are given to a chosen degree of accuracy (can assume the infinitely many values corresponding to the points on a line interval.

---

**Key point (1)**

- Discret data can take only certain values.
- Continuous data can take any value, possibly within a limited range.

Example ①: Identify each of the following variables as qualitative or quantitative.

① The most frequent use of your microwave oven (reheating, defrosting, warming, other).

② The number of consumers who refuse to answer a telephone survey.

③ The door chosen by a mouse in a maze experiment (A, B or C).

④ the winning time for a horse running in the Kentucky Derby.

⑤ The number of children in a fifth-grade class who are reading at or above grade level.

Solution : . Variables ① and ③ are both qualitative because only a quality or characteristic is measured for each individual. The categories for these two variables are shown in parentheses.

. The other three variables are quantitative. Variable 2, the number of consumers, is a discrete variable that can take on any of the values.

$x = 0, 1, 2, \ldots$

Similarly, variable 5, the number of children reading at or above grade level, can take on any of the values $x = 0, 1, 2, \ldots$ With a maximum value depending on the number of children in the class. Variable ⑦, the winning time for a Kentucky Derby horse, is the only continuous variable in the list. The winning time, if it could be measured with sufficient accuracy, could be 121 s, 121,8 seconds, 121,25 s, or any values between any two times we have listed.

## ③ Graphs for categorical data

- After the data have been collected, they can be consolidated and summarized to show the following information:
  - What values of the variable have been measured
  - How often each values has occurred.
- For this purpose, you can construct a "statistical table" that can be used to display the data graphically as a data distribution. The type of graph you choose depends on the type of variable you have measured.

- When the variable of interest is "qualitative", the statistical table is a list of the categories being considered along with a measure of how often each values occurred. You can measure "how often" in three different ways:

  - the "frequency", or number of measurements in each category.
  - the "relative frequency", or proportion of measurements in each category.
  - the "percentage" of measurements in each categ

For example, if you let n be the total number of measurements in the set, you can find the relative frequency and percentage using these relationship

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

$$\text{Percent} = 100 \times \text{Relative frequency}.$$

You will find that the sum of the frequencies is always n, the sum of the relative frequencies is 1, and the sum of the percentages is 100%.

The categories for a qualitative variable should be chosen so that:

- a measurement will belong to one and only one category.
- each measurement has a category to which it can be assigned.

- Once the measurements have been categorized and summarized in a "statistical table", you can use either a "pie chart" or a "bar chart" to display the distribution of the data.

**Key point** three steps to a data distribution

→ ① raw data.

→ ② statistical table

→ ③ graphs.

A pie chart is the familiar circular graph that shows how the measurements are distributed among the categories. A bar chart shows the same distribution of measurements in categories, with the height of the bar measuring how often a particular category was observed.

Example①: In a survey concerning public education 400 school administrators were asked to rate the quality of education in the Algeria. Their responses are summarized in Table 01. Construct a pie chart and a bar chart for this set of data.

Solution: To construct a pie chart, assign one sector of a circle to each category. the angle of each sector should be proportional to the proportion of measurements (or relative frequency) in that category. Since a circle contains 360°, you can use this equation to find the angle:

$$\text{Angle} = \text{Relative frequency} \times 360°$$

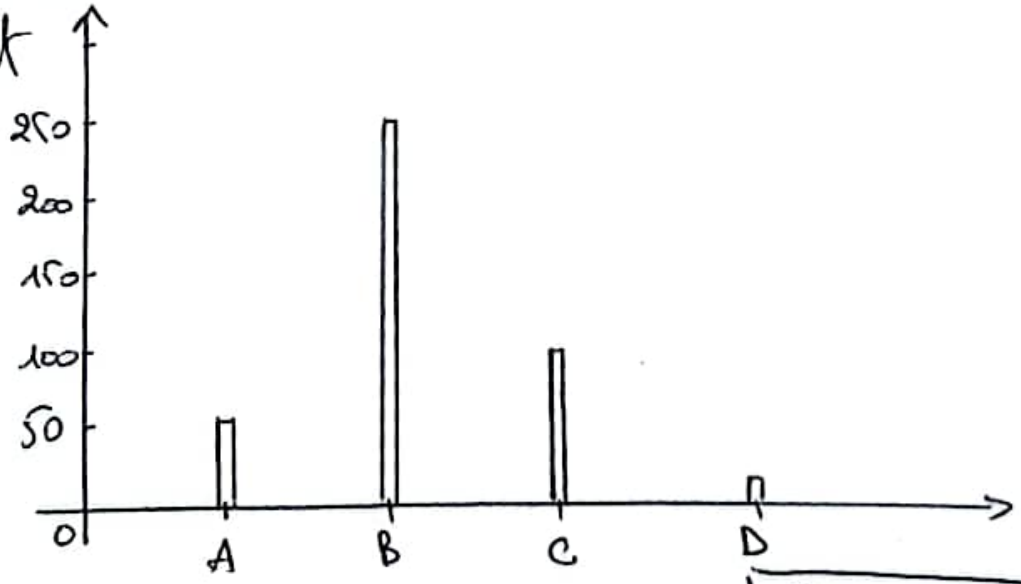
Table 01 Education Rating by 400 Educators

| Rating | Frequency |
|---|---|
| A | 35 |
| B | 260 |
| C | 93 |
| D | 12 |
| Total | 400 |

- Calculations for the Pie chart in Example.

| Rating | Frequency | Relative Frequency | Percent | Angle |
|---|---|---|---|---|
| A | 35 | 35/400 = 0,09 | 9% | 0,09 x 360 = 32,4° |
| B | 260 | 260/400 = 0,65 | 65% | 234,0° |
| C | 93 | 93/400 = 0,23 | 23% | 82,8° |
| D | 12 | 12/400 = 0,03 | 3% | 10,8° |
| Total | 400 | 1,00 | 100% | 360° |

- Bar Chart



Example ②: A snack size bag of peanut M and M's candies contains 21 candies with the colors listed in Table ①. The variable "color" is qualitative, so Table ③ lists the six categories along with a tally of the number of candies of each color. The last three columns of Table ③ give the three different measures of how often each category occurred.

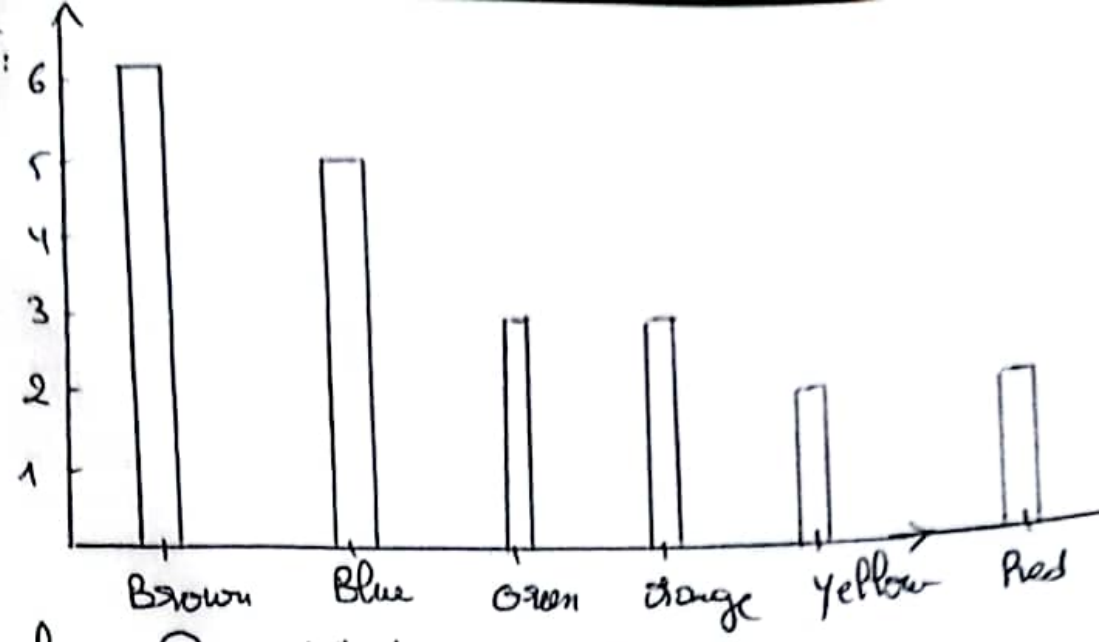
Table ②: Raw data: colors of 21 candies

| | | | |
|---|---|---|---|
| Brown | Green | Brown | Blue |
| Red | Red | Green | Brown |
| Yellow | Orange | Green | Blue. |
| Brown | Blue | Blue | Brown. |
| Orange | Blue | Brown | Orange - Yellow |

Table ③: Statistical Table: M & M's Data

| Category | Frequency | Relative Freq | Percent. |
|---|---|---|---|
| Brown | 6 | 6/21 | 28%. |
| Green | 3 | 3/21 | 14%. |
| Orange | 3 | 3/21 | 14%. |
| Yellow | 2 | 2/21 | 10%. |
| Red | 2 | 2/21 | 10%. |
| Blue | 5 | 5/21 | 24%. |
| Total | 21 | 1 | 100% |

- Bar chart:




## (4) Graphs for Quantitative data

Quantitative variables measure an amount or quantity on each experimental unit. If the variable can take only a finite or countable number of values, it is a discrete variable. A variable that ~~can~~ can assume an infinite number of values corresponding to points on a line interval is called continuous.

- Pie charts and Bar Charts:

The Pie chart displays how the Total quantity is distributed among the categories, and the Bar chart uses the height of the bar to display the amount in a particular category.

Key point: A relative frequency histogram resembles a bar chart, but it is used to graph quantitative rather than qualitative data.

**Definition**: A relative frequency histogram for a quantitative data set is a box graph in which the height of the box shows "how often" (measured as a proportion of relative frequency) measurements fall in particular class or subinterval. The classes or subintervals are plotted along the horizontal axis.

**Example**:

The data in Table ④ are the birth weights of 30 fullterm newborn babies

**Table ④**:

| | | | | |
|---|---|---|---|---|
| 7,2 | 7,8 | 6,9 | 6,2 | 8,2 |
| 8,0 | 8,2 | 5,6 | 8,6 | 7,1 |
| 8,2 | 7,7 | 7,5 | 7,2 | 7,7 |
| 5,8 | 6,9 | 6,8 | 8,5 | 7,5 |
| 6,1 | 7,9 | 9,4 | 9,0 | 7,8 |
| 8,5 | 9,0 | 7,7 | 6,7 | 7,7 |

the classes must be chosen so that each measurement falls into one and only one classe. We decided to use $\sqrt{30} \approx 5$ six intervals of equal length. Since the total span of the birth weights is

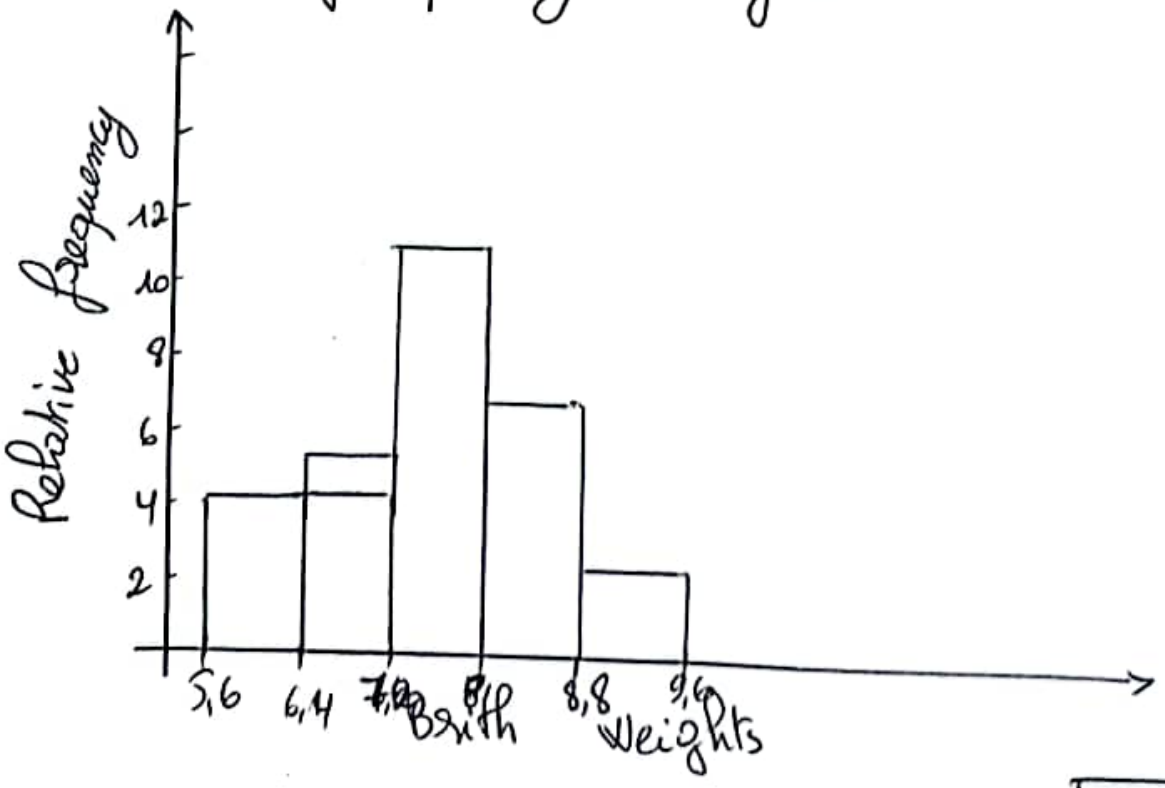
$$9,4 - 5,6 = 3,8$$

the minimum class width necessary to cover the range of the data is $(3,1 \div 5 = 0,76)$

- Relative Frequencies for the data of table ④

| Class | Class Boundaries | Class Frequency | Class Relative Frequency |
|---|---|---|---|
| 1 | $[5,6 - 6,4[$ | 4 | 4/30 |
| 2 | $[6,4 - 7,2[$ | 5 | 5/30 |
| 3 | $[7,2 - 8,0[$ | 11 | 11/30 |
| 4 | $[8,0 - 8,8[$ | 7 | 7/30 |
| 5 | $[8,8 - 9,6[$ | 3 | 3/30 |

- Relative frequency histogram



Key Point:

How do I Construct a Relative Frequency Histogram

1/ Choose the number of classes ($\sqrt{N}$), usually between 5 and 12. The more data you have, the more classes you should use

2/ Calculate the approximate class width by dividing the difference between the largest and smallest values by the number of classes.

3/ Round the approximate class width up to a convenient number.

4/ If the data are discrete, you might assign one class for each integer value taken on by the data. For a large number of integer values, you may need to group them into classes.

5/ Locate the class boundaries. The lowest class must include the smallest measurements. Then add the remaining classes using the left inclusion method.

6/ Construct a statistical table containing the classes, their frequencies, and their relative frequencies.

7/ Construct the histogram like a bar graph.