Descriptive Statistics

In statistics we are often presented with a large set of data obtained from some experiment or survey. From this data we would want to draw some conclusions about the underlying information it contained. These conclusions could be graphical or numerical. What type of analyses can be used depends on the type of data you have. Listed below are the most often used data types.

Categorical data:- The results of an opinion poll might be classified as Conservative, Labour, Liberal Democrat, Other. This type of data records the category that each person interviewed belongs to. **Such data are categorical data.**

Ordinal Data:- The sizes of clothes are often measured on a scale 1,2,3,4,5. Here we can write 4 > 2 as size 4 is larger than size 2, but we cannot perform the usual arithmetic operations such as addition, subtraction or multiplication. **Such data are ordinal data.**

Interval data:- Suppose daily temperatures are recorded using the centigrade scale, then we can write $20^{\circ}C > 15^{\circ}C$. Further we can write 20 – 15 =15 – 10. i.e., the temperature differences are equal. However, we cannot say $20^{\circ}C$ is twice as hot as $10^{\circ}C$, since if these temperatures were recorded on the Fahrenheit scale, they become $68^{\circ}F$ and $50^{\circ}F$ and the first is no longer twice the second. Thus, addition and subtraction are valid but not multiplication or division. **Such data are interval data**.

Ratio data:- Suppose the heights of a group of people are recorded. All the usual arithmetic operations can be performed on this type of data. **Such data are ratio data**.

Frequency distributions.

Suppose, in what follows that the data type you have is either, ordinal, interval or ratio so they can be ordered. If the dataset you have is reasonably large it is very difficult to digest the information it may contain. One way of obtaining a clearer picture of the information the data contains, is to group the data into classes and to count the number of data in each class.

Suppose our data set contained the weights of 80 oarsmen, recorded to 1 decimal place, and that the classes used are [75.0 – 84.9], [85.0 – 94.9], [95.0 – 104.9] and [105.0 – 114.9], with frequency counts of, 8, 29, 35, and 8 respectively. **The frequency distribution** could be presented in tabular form as follows.

Class	75.0 – 84.9	85.0 – 94.9	95.0 – 104.9	105.0 – 114.9
Frequency	8	29	35	8

We now define some standard terms related to frequency distributions.

Class limits:- These are the smallest and largest values that can go in a given class. For the first class these limits would be 75.0 and 84.9.

Class boundaries:- In the above example the class limits were 75.0 and 84.9, but we are told that the data are recorded to 1 decimal place. This being the case, the class boundaries,

which are the smallest and largest values which, after rounding, are 74.95 and 84.95 respectively.

Class frequency:-This is the number of observations in each class.

Class interval:- The class interval is the difference between the class boundaries and gives the range of possible values in each class. If the class interval is the same for all classes then we call it the **class interval of the distribution**.

Class mark:- The class mark is the midpoint of the class, i.e., the midpoint between the class boundaries.

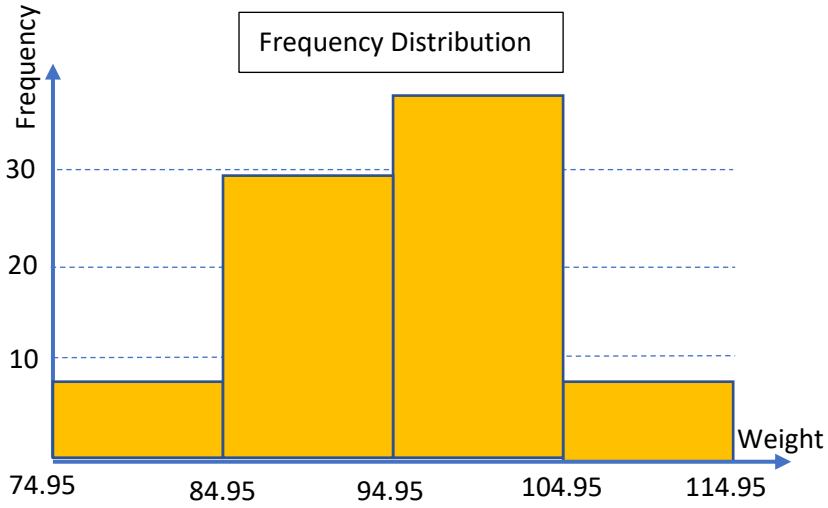
The table below lists all of the above for the data being considered.

Class Frequency	8	29	35	8
Class limits	75.0 – 84.9	85.0 – 94.9	95.0 – 104.9	105.0 – 114.9
Class Boundaries	74.95-84.95	84.95-94.95	94.95-104.95	104.95-114.95
Class interval	10.0	10.0	10.0	10.0
Class mark	79.95	89.95	99.95	109.95

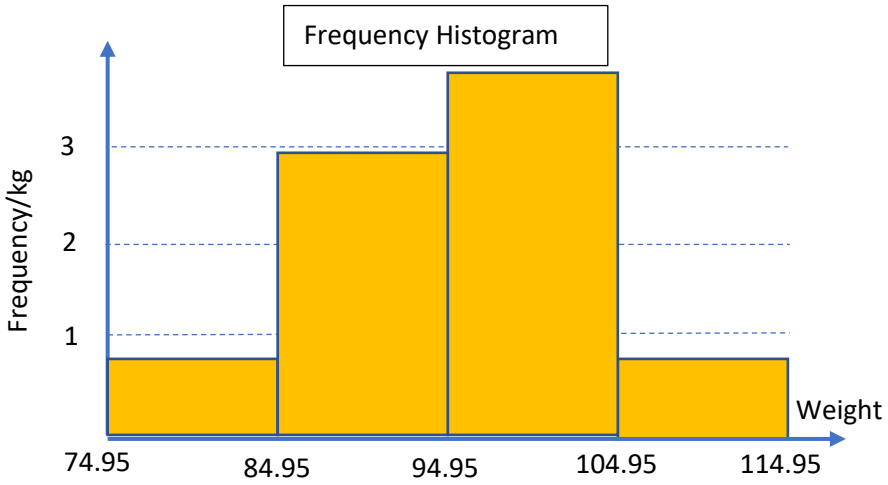
NB For a data point with value 84.95 it is customary to include it in the class (74.95-84.95] rather than in the class (84.95-94.95], similarly for other class boundary values.

Plots of frequency distributions

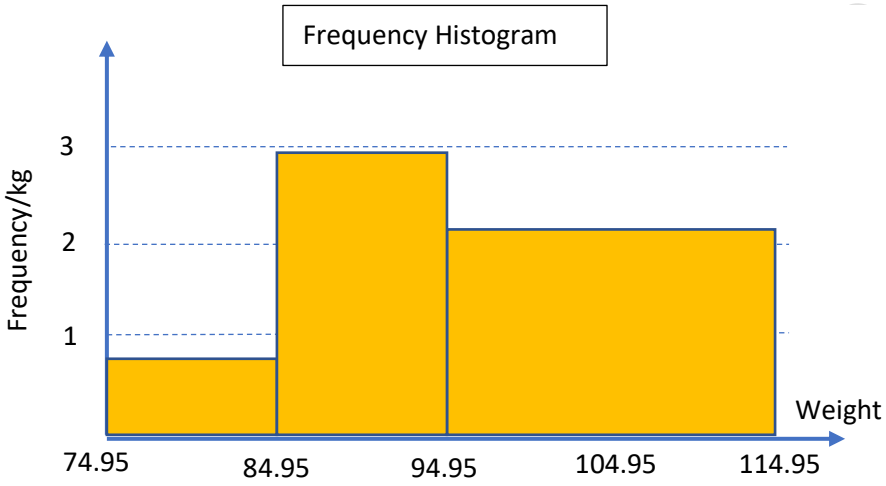
The plot of the above frequency distribution is shown alongside. The height of the bars is the frequency of the class and their width is the class interval of the distribution. Arguably the class interval is too wide as much of the finer detail may be lost.



Another, slightly more useful plot is the **frequency histogram**. The construction is exactly the same as that for the frequency distribution, but the vertical scale is

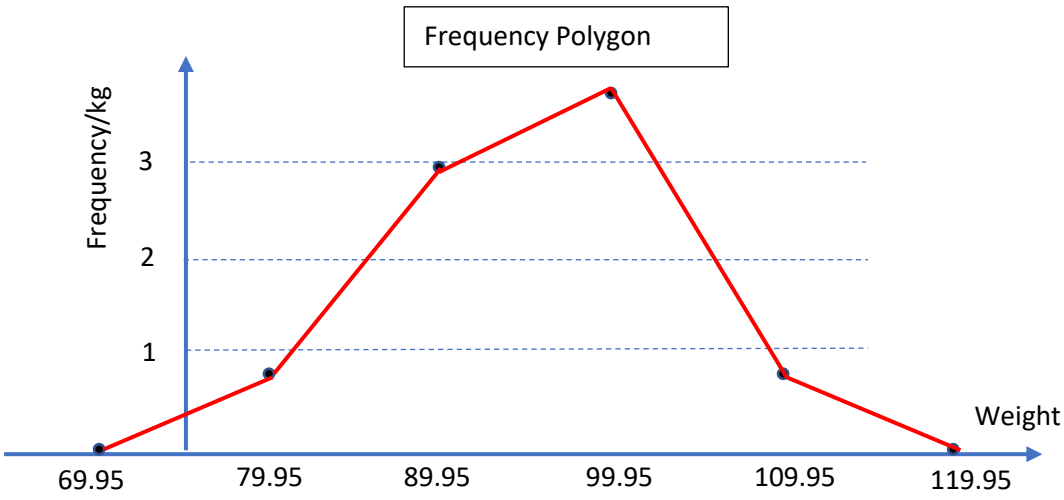


chosen in such a way that the area of each bar equates to the frequency. For the first class, of width 10, and frequency 8, the height of the corresponding histogram bar is $\frac{8}{10}$, the area being $\frac{8}{10} \times 10 = 8$, which is the frequency. The frequency histogram visually is the same but with the vertical scale changed to 0,1,2,3 and the units are frequency/Kg. The strength of the histogram approach comes when the class widths are not constant. Suppose we combined the last two classes, making the class boundaries (94.95,114.95]. The combined frequency is 35 + 8 = 43 and the class width is 20, making the height of the histogram equal to $\frac{43}{20} = 2.15$. This is shown in the figure alongside which has retained a similar shape to the earlier histogram. This would not have happened with a frequency plot.



Frequency polygon

A frequency polygon is easily obtained from the frequency histogram. The midpoints of the tops of the rectangles are joined by straight lines. It is usual to construct an extra class with zero frequency at either end of the range of values and to join the midpoints of these classes with the midpoints in the adjacent classes. This procedure is shown below.

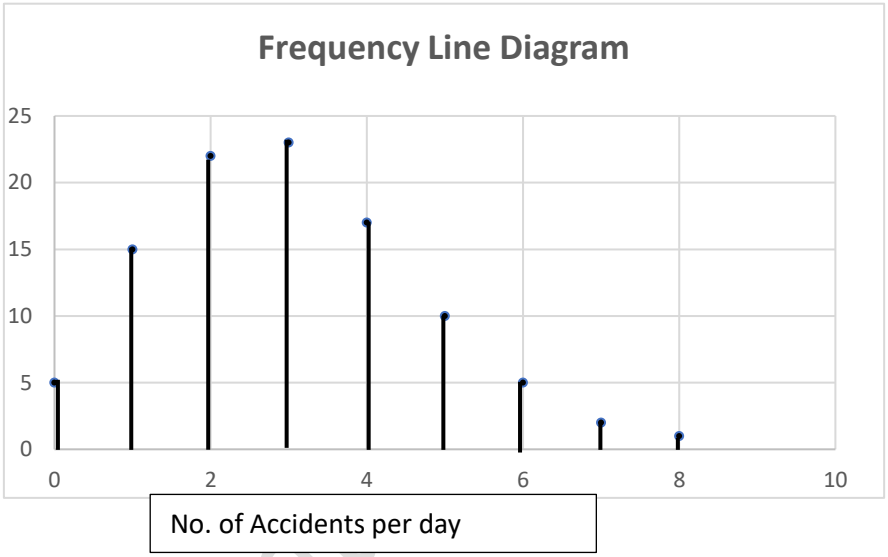


The above plots are appropriate when the data are continuous and need grouping. Suppose the data type is discrete, for example, the number of road accidents per day in a certain town. Over a period of 100 days the frequency distribution might be

No. of accidents	0	1	2	3	4	5	6	7	8
Frequency	5	15	22	23	17	10	5	2	1

Grouping and intervals are not really appropriate, and so a line diagram is used to represent this data as shown alongside.

The length of each line represents the frequency of the number of accidents in a day.



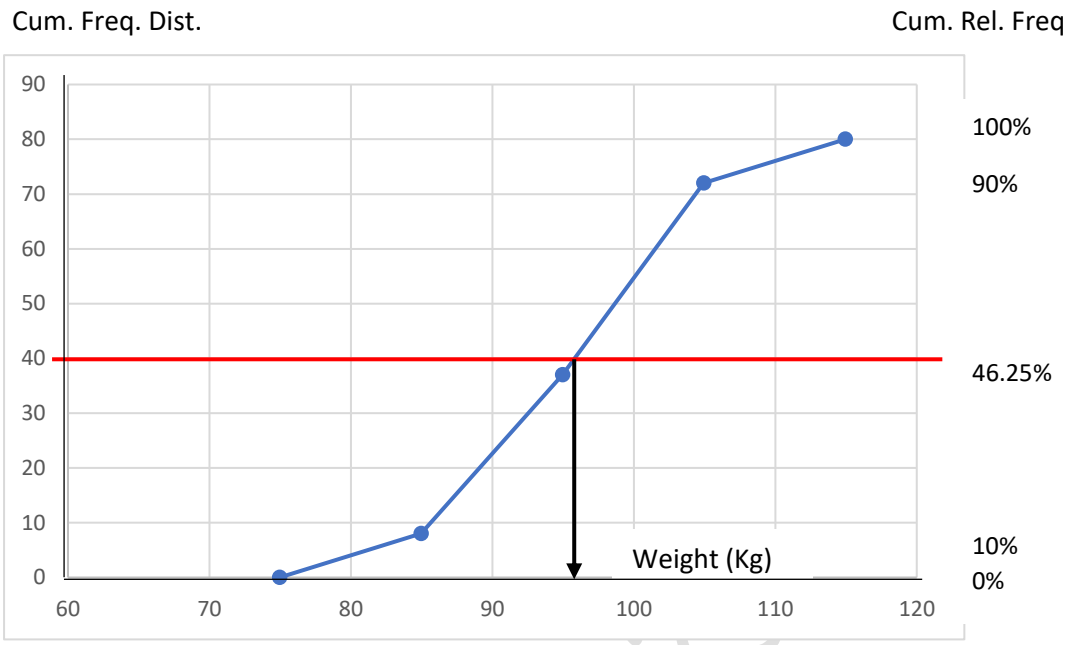
A closely related concept is **the relative frequency distribution**, which is the same as the frequency distribution but the vertical scale is the relative frequency of the outcome, relative frequency being the frequency divided by the total number of observations. For the above frequency distribution, the relative frequency of the outcome 2 is $\frac{22}{100} = 0.22$. This also extends to bar charts and histograms. A consequence of the definition of relative frequency is that the sum of all relative frequencies over intervals, or points (line diagrams) is one, consistent with one of the basic properties of a probability function.

Cumulative frequency distribution

This gives, for each class, the total number of observations less than or equal to the upper class boundary of interest. A table of the data from earlier is shown below with the appropriate calculations shown.

Class	Class frequency	Upper Class boundary	Cumulative Frequency Distribution	Cumulative Relative Freq. Distribution
75.0-84.9	8	84.95	8	10.0%
85.0-94.9	29	94.95	37	46.25%
95.0-104.9	35	104.95	72	90%
105.0-114.9	8	114.95	80	100%

To obtain a graphical representation of the cumulative frequency distribution, or cumulative relative frequency distribution the points are plotted at the **upper class boundaries** and joined by straight lines.



The figure is usually called a **cumulative frequency**, or **cumulative relative frequency, polygon** when the points are connected by straight lines. If the straight lines had been replaced by a **smooth curve** it would be called **an ogive**.

The cumulative relative frequency plot is probably the most important, as the vertical scale is the same for all datasets, and it also gives the means to estimate percentiles directly. The median, which is the middle point, can be obtained by drawing a horizontal line at 40 (freq) or 50%(rel freq) and where that cuts the polygon, projecting down gives the median weight, approximately 96 Kg.