"JUST THE MATHS"

UNIT NUMBER

18.1

STATISTICS 1
(The presentation of data)

by

A.J.Hobson

UNIT 18.1 - STATISTICS 1 - THE PRESENTATION OF DATA

18.1.1 INTRODUCTION

(i) The collection of numerical information often leads to large masses of data which, if they are to be understood, or presented effectively, must be summarised and analysed in some way. This is the purpose of the subject of **"Statistics"**.

(ii) The source from which a set of data is collected is called a **"population"**. For example, a population of 1000 ball-bearings could provide data relating to their diameters.

(iii) Statistical problems may be either **"descriptive problems"** (in which all the data is known and can be analysed) or **"inference problems"** (in which data collected from a **"sample"** population is used to infer properties of a larger population). For example, the annual pattern of rainfall over several years in a particular place could be used to estimate the rainfall pattern in other years.

(iv) The variables measured in a statistical problem may be either **"discrete"** (in which case they may take only certain values) or **"continuous"** (in which case they make take any values within the limits of the problem itself. For example, the number of students passing an examination from a particular class of students is a discrete variable; but the diameter of ball-bearings from a stock of 1000 is a continuous variable.

(v) Various methods are seen in the commercial presentation of data but, in this series of Units, we shall be concerned with just two methods, one of which is tabular and the other graphical.

18.1.2 THE TABULATION OF DATA

(a) Ungrouped Data
Suppose we have a collection of measurements given by numbers. Some may occur only once, while others may be repeated several times.

If we write down the numbers as they appear, the processing of them is likely to be cumbersome. This is known as **"ungrouped (or raw) data"**, as, for example, in the following table which shows rainfall figures (in inches), for a certain location, in specified months over a 90 year period:

TABLE 1 - Ungrouped (or Raw) Data

18.6	13.8	10.4	15.0	16.0	22.1	16.2	36.1	11.6	7.8
22.6	17.9	25.3	32.8	16.6	13.6	8.5	23.7	14.2	22.9
17.7	26.3	9.2	24.9	17.9	26.5	26.6	16.5	18.1	24.8
16.6	32.3	14.0	11.6	20.0	33.8	15.8	15.2	24.0	16.4
24.1	23.2	17.3	10.5	15.0	20.2	20.2	17.3	16.6	16.9
22.0	23.9	24.0	12.2	21.8	12.2	22.0	9.6	8.0	20.4
17.2	18.3	13.0	10.6	17.2	8.9	16.8	14.2	15.7	8.0
17.7	16.1	17.8	11.6	10.4	13.6	8.4	12.6	8.1	11.6
21.1	20.5	19.8	24.8	9.7	25.1	31.8	24.9	20.0	17.6

(b) Ranked Data
A slightly more convenient method of tabulating a collection of data would be to arrange them in rank order, so making it easier to see how many times each number appears. This is known as **"ranked data"**.

The next table shows the previous rainfall figures in this form

TABLE 2 - Ranked Data

7.8	8.0	8.0	8.1	8.4	8.5	8.9	9.2	9.6	9.7
10.4	10.4	10.5	10.6	11.6	11.6	11.6	11.6	12.2	12.2
12.6	13.0	13.6	13.6	13.8	14.0	14.2	14.2	15.0	15.0
15.2	15.7	15.8	16.0	16.1	16.2	16.4	16.5	16.6	16.6
16.6	16.8	16.9	17.2	17.2	17.3	17.3	17.6	17.7	17.7
17.8	17.9	17.9	18.1	18.3	18.6	19.8	20.0	20.0	20.2
20.2	20.4	20.5	21.1	21.8	22.0	22.0	22.1	22.6	22.9
23.2	23.7	23.9	24.0	24.0	24.1	24.8	24.8	24.9	24.9
25.1	25.3	26.3	26.5	26.6	31.8	32.3	32.8	33.8	36.1

(c) Frequency Distribution Tables

Thirdly, it is possible to save a little space by making a table in which each individual item of the ranked data is written down once only, but paired with the number of times it occurs. The data is then presented in the form of a **"frequency distribution table"**.

TABLE 3 - Frequency Distribution Table

Value	Frequency	Value	Frequency	Value	Frequency
7.8	1	15.8	1	21.1	1
8.0	2	16.0	1	21.8	1
8.1	1	16.1	1	22.0	2
8.4	1	16.2	1	22.1	1
8.5	1	16.4	1	22.6	1
8.9	1	16.5	1	22.9	1
9.2	1	16.6	3	23.2	1
9.6	1	16.8	1	23.7	1
9.7	1	16.9	1	23.9	1
10.4	2	17.2	2	24.0	2
10.5	1	17.3	2	24.1	1
10.6	1	17.6	1	24.8	2
11.6	4	17.7	2	24.9	2
12.2	2	17.8	1	25.1	1
12.6	1	17.9	2	25.3	1
13.0	1	18.1	1	26.3	1
13.6	2	18.3	1	26.5	1
13.8	1	18.6	1	26.6	1
14.0	1	19.8	1	31.8	1
14.2	2	20.0	2	32.3	1
15.0	2	20.2	2	32.8	1
15.2	1	20.4	1	33.8	1
15.7	1	20.5	1	36.1	1

(d) Grouped Frequency Distribution Tables

For about forty or more items in a set of numerical data, it usually most convenient to group them together into between 10 and 25 **"classes"** of values, each covering a specified range or **"class interval"** (for example, $7.5 - 10.5$, $10.5 - 13.5$, $13.5 - 16.5$,......).

Each item is counted every time it appears in order to obtain the **"class frequency"** and each class interval has the same **"class width"**.

Too few classes means that the data is over-summarised, while too many classes means that there is little advantage in summarising at all.

Here, we use the convention that the lower boundary of the class is included while the upper boundary is excluded.

Each item in a particular class is considered to be approximately equal to the "class mid-point"; that is, the average of the two **"class boundaries"**.

A **"grouped frequency distribution table"** normally has columns which show the class intervals, class mid-points, class frequencies, and **"cumulative frequencies"**, the last of these being a running total of the frequencies themselves. There may also be a column of **"tallied frequencies"**, if the table is being constructed from the raw data without having first arranged the values in rank order.

TABLE 4 - Grouped Frequency Distribution

Class Interval	Class Mid-point	Tallied Frequency	Frequency	Cumulative Frequency
7.5 – 10.5	9	////| ////| //	12	12
10.5 – 13.5	12	////| ////|	10	22
13.5 – 16.5	15	////| ////| ////|	15	37
16.5 – 19.5	18	////| ////| ////| ////	19	56
19.5 – 22.5	21	////| ////| //	12	68
22.5 – 25.5	24	////| ////| ////	14	82
25.5 – 28.5	27	///	3	85
28.5 – 31.5	30		0	85
31.5 – 34.5	33	////	4	89
34.5 – 37.5	36	/	1	90

Notes:

(i) The cumulative frequency shows, at a glance, how many items in the data are less than a specified value. In the above table, for example, 82 items are less than 25.5

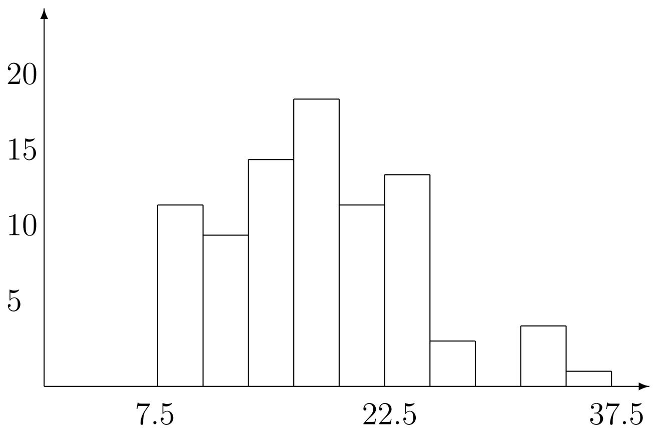
(ii) It is sometimes more useful to use the ratio of the cumulative frequency to the total number of observations. This ratio is called the **"relative cumulative frequency"** and, in the above table, for example, the percentage of items in the data which are less than 25.5 is

$$\frac{82}{90} \times 100 \simeq 91\%$$

18.1.3 THE GRAPHICAL REPRESENTATION OF DATA

(a) The Histogram

A **"histogram"** is a diagram which is directly related to a grouped frequency distribution table and consists of a collection of rectangles whose height represents the class frequency (to some suitable scale) and whose breadth represents the class width.

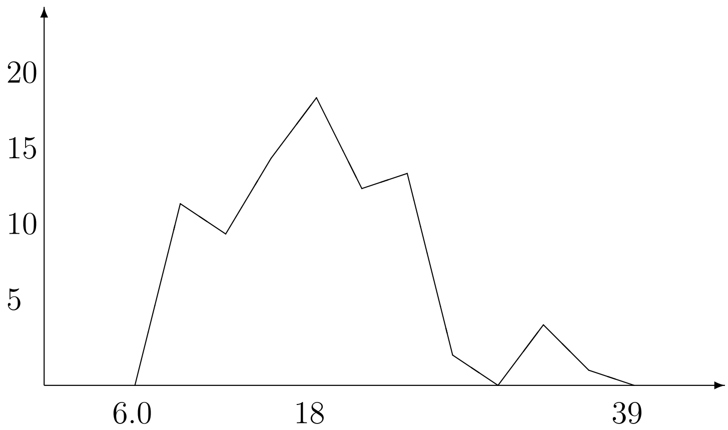


The histogram shows, at a glance, not just the class intervals with the highest and lowest frequencies, but also how the frequencies are distributed.

In the case of examination results, for example, there is usually a group of high frequencies around the central class intervals and lower ones at the ends. Such an ideal situation would be called a **"Normal Distribution"**.

(b) The Frequency Polygon

Using the fact that each class interval may be represented, on average, by its class mid-point, we may plot the class mid-points against the class frequencies to obtain a display of single points. By joining up these points with straight line segments and including two extra class mid-points, we obtain a **"frequency polygon"**.



Notes:

(i) Although the frequency polygon officially plots only the class mid-points against their frequencies, it is sometimes convenient to read-off intermediate points in order to estimate additional data. For example, we might estimate that the value 11.0 occurred 11 times when, in fact, it did not occur at all.

We may use this technique only for continuous variables.

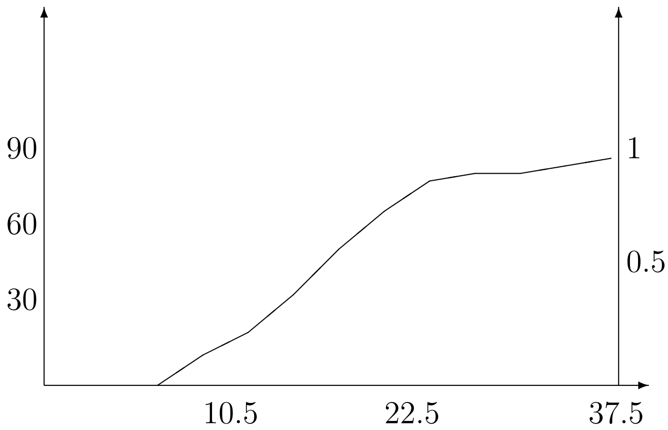
(ii) Frequency polygons are more useful than histograms if we wish to compare two or more frequency distributions. A clearer picture is obtained if we plot them on the same diagram.

(iii) If the class intervals are made smaller and smaller while, at the same time, the total number of items in the data is increased more and more, the points of the frequency polygon will be very close together. The smooth curve joining them is called the **"frequency curve"** and is of greater use for estimating intermediate values.

(c) The Cumulative Frequency Polygon (or Ogive)

The earlier use of the cumulative frequency to estimate the number (or proportion) of values less than a certain amount may be applied graphically by plotting the upper class-boundary against the cumulative frequency, then joining up the points plotted with straight line segments. The graph obtained is called the **"cumulative frequency polygon"** or **"ogive"**.

We may also use a second vertical axis at the right-hand end of the diagram showing the relative cumulative frequency. The range of this axis will always be 0 to 1.



18.1.4 EXERCISES

1. State whether the following variables are discrete or continuous:
   (a) The denominators of a set of 10,000 rational numbers;
   (b) The difference between a fixed integer and each one of any set of real numbers;
   (c) The volume of fruit squash in a sample of 200 bottles taken from different firms;
   (d) The minimum plug-gap setting specified by the makers of 30 different cars;
   (e) The points gained by 80 competitors in a skating championship.

2. The following marks were obtained in an examination taken by 100 students:

Marks	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
No. of Students	2	3	7	7	8

Marks	50 – 55	55 – 60	60 – 65	65 – 70	70 – 75
No. of Students	25	18	12	10	8

   (a) Draw a histogram for this data;
   (b) Draw a cumulative frequency polygon;
   (c) Estimate the mark exceeded by the top 25% of the students;
   (d) Suggest a pass-mark if 15% of the students are to fail.

3. In a test of 35 glue-laminated beams, the following values of the “spring constant” in kilopounds per inch were found:

   SPRING CONSTANT × 100

6.72	6.77	6.82	6.70	6.78	6.70	6.62
6.75	6.66	6.66	6.64	6.76	6.73	6.80
6.72	6.76	6.76	6.68	6.66	6.62	6.72
6.76	6.70	6.78	6.76	6.67	6.70	6.72
6.74	6.81	6.79	6.78	6.66	6.76	6.72

   (a) Arrange this data in rank order and hence construct a frequency distribution table;
   (b) Construct a grouped frequency distribution table with class intervals of length 0.02 starting with $6.61 - 6.63$ and include columns to show the class mid-points and the cumulative frequency.

4. The following data shows the diameters of 25 ball-bearings in cms.

0.386	0.391	0.396	0.380	0.397
0.376	0.384	0.401	0.382	0.404
0.390	0.404	0.380	0.390	0.388
0.381	0.393	0.387	0.377	0.399
0.400	0.383	0.384	0.390	0.379

   Construct a grouped frequency distribution table with class intervals $0.375 - 0.380$, $0.380 - 0.385$ etc. and construct

   (a) the histogram;
   (b) the frequency polygon;
   (c) the cumulative frequency polygon (ogive).

18.1.5 SELECTED ANSWERS TO EXERCISES

1. (a) discrete
   (b) continuous
   (c) continuous
   (d) discrete
   (e) discrete
2. (c) 63% is the mark exceeded by the top 25% of students
   (d) 43% needs to be the pass-mark if 15% of the students are to fail.