Section 8 (Cont'd)
Frequency Distributions

In this section, we continue with our analysis of statistics by considering how data[1] collected through experiment can be organised into groups (called classes) and presented in a manner that permits rapid analysis. This manner of presentation is called a frequency distribution and is an efficient method of presenting data only if the distribution is properly constructed. The rules necessary for it to be properly constructed are discussed here.

Raw data

Raw data are collected data which have not been organised numerically. An example is the set of 100 people's ages randomly selected from a census.

Arrays

An array is an organisation of raw numerical data in ascending or descending order of magnitude. The difference between the largest and smallest numbers is called the *range*. For example if the oldest person was 64 years and the youngest was 12 years then the range would be $65-12=53$ years.

Frequency distributions

When summarizing large amounts of raw data it is often useful to distribute the data into *classes* or *categories* and to determine the number of individuals belonging to each class, called the *class frequency*.

Age (Years)	Number of People
12-17	1
18-23	6
24-29	20
30-35	15
36-41	13
42-47	20
48-53	12
54-59	8
60-65	5

1 It is important to note that data is plural and refers to more than one element. The singular of data is datum.

A table of the data in their classes together with the corresponding class frequencies is called a *frequency table* or a *frequency distribution*. The table below is a distribution of the ages recorded to the nearest year of the 100 people in the sample from the census.

The first class or category is those those people who lie within the age range of 12 to 17 years indicated by 12-17. Since 1 person lies within this category the corresponding class frequency is 1.
Data organised and summarised as above are often called *grouped data*. The grouping of the data generally destroys much of the organised detail of the data but the advantage gained from this procedure is the overall picture and the establishment of relationships between subsets of the data.

Class Intervals & Class Limits
A symbol defining a class such as 12-17 in the above table is called a *class interval*. The end numbers are called *class limits* such as 12 & 17 above; The smaller number, 12, is called the *lower class limit* and 17, the larger number, is called the *upper class limit*. A class interval which has no upper class limit or no lower class limit is called an *open class interval*. For example referring to the age groups above the class interval "65 years or over" is an open class interval.

Class Boundaries
If years are recorded to the nearest year, the class interval 12-17 theoretically includes all measurements from 11.5 years to 17.5. These numbers are called *class boundaries* or *true class limits*; the lower is the *lower class boundary* and the higher *the upper class boundary.* Sometimes class boundaries are used to symbolise classes.
For example the various classes in the previous table could have been indicated by 11.5-17.5, 17.5-23.5, …… etc. To avoid ambiguity in using such notation, class boundaries should not coincide with actual observations. So if 17.5 years was an observation we could not decide which class it belonged to 11.5-17.5 or 17.5-23.5.

Size or Width of a Class Interval
The size or width of a class interval is the difference between the lower and upper class boundaries and is also referred to as the *class width, class size* or *class strength.* If all class intervals of a frequency distribution have equal widths, this common width is denoted by c. In such cases c is equal

to the difference between two successive lower limits or two successive upper limits. For the data we've being considering the class interval *c* is 17.5-11.5 = 6.

The Class Mark

The class mark is the midpoint in the class interval and is obtained by adding the lower and upper class limits and dividing by 2. Thus the class mark of the interval 12-17 is (12+17)/2=14.5. The class mark is also called the *class midpoint*.

From now on, all observations belonging to a class interval are assumed to coincide with the class mark. Thus all ages in the interval 12-17 years are considered to be 14.5 years. More importantly, if the classes are constructed properly for the given data, then the class marks will correspond to the averages of each class. Later, we'll use this property as a basis for a method used to analyse frequency distributions.

General Rules For Forming Frequency Distributions.

1. Determine the largest and smallest numbers in the raw data and thus find the range.
2. Divide the range into a convenient number of class intervals having the same size. If this is not possible then use class intervals of differing sizes and open class intervals, if required. Class intervals are usually chosen so that the class marks or midpoints correspond with observed data. This tends to lessen errors which may creep in in further mathematical analysis.
3. Class boundaries should not correspond with observed data.
4. Determine the number of observations falling into each interval; i.e. find the class frequencies. This is best done by using a *tally* or *score sheet.*

Histograms & Frequency Polygons

These are two graphical representations of frequency distributions.

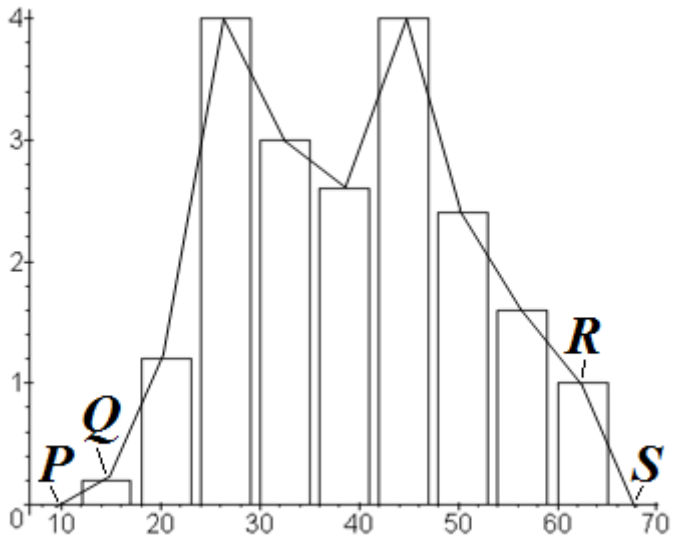
1. A *histogram or frequency histogram* consists of a set of rectangles having
   a) bases on a horizontal axis (the *x* axis) with centres at the class marks and lengths equal to the class interval sizes,
   b) areas proportional to class frequencies

   If the class intervals all have equal size, the heights of the rectangles

are proportional to the class frequencies and it is customary to take the heights numerically equal to the class frequencies. If class intervals do not have equal size, these heights may be adjusted.

2. A *frequency polygon* is a line graph of class frequency plotted against class mark. It can be obtained by connecting midpoints of the tops of the rectangles in the histogram.

The histogram and frequency polygon corresponding to the frequency distribution of the ages are shown on the same axes below. It is customary to add the extensions *PQ* and *RS* to the next lower and higher class marks which have a corresponding class frequency of zero. In such case the sum of the areas of the rectangles in the histogram equals the total area bounded by the frequency polygon and the *x* axis.



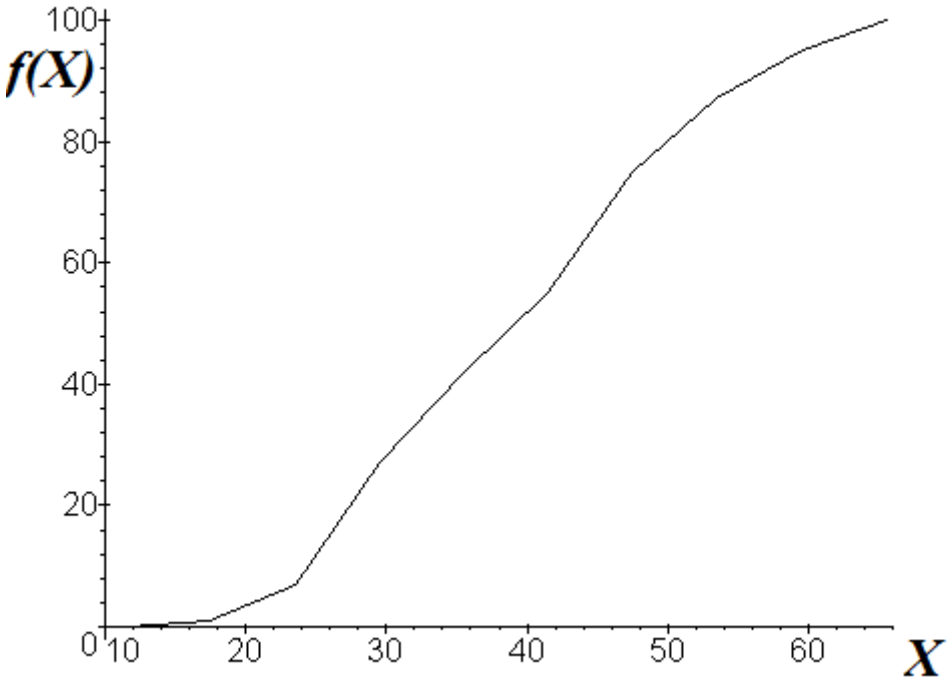
N.B.
While the above histogram is theoretically correct, it is more convenient for distributions with a common classwidth, *c*, to associate the height with the frequency rather than the more correct but esoteric area. As all frequency distributions for this course will have a common class width, then we shall use the height of the histogram to convey the class frequency from this point onwards.

Cumulative Frequency Distributions - Ogives.
The total frequency of all values less than the upper class boundary of a given class interval is called the *cumulative frequency* up to and including the class interval. In our example, the cumulative frequency up to the class interval 36-41 is 1 + 6 + 20 + 15 + 13 = 55 telling us that 55 of 100

people have ages less than 41.5 years. A table representing such a distribution is called a *cumulative frequency distribution, cumulative frequency table* or, more euphemistically, a *cumulative distribution.*

Age (Years)	Number of People
<11.5	0
<17.5	1
<23.5	7
<29.5	27
<35.5	42
<41.5	55
<47.5	75
<53.5	87
<59.5	95
<65.5	100

A graph showing cumulative frequency less than any upper class boundary plotted against the upper class boundary is called a *cumulative frequency polygon* or *ogive* and is shown below for our age sample.



The *relative cumulative frequency* or *percentage cumulative frequency* is the cumulative frequency divided by the total frequency. For example, the relative cumulative frequency of ages less than 47.5 years is 75%

signifying that 75% of the people in the sample have ages less than 47.5 years.

Frequency Curves. Smoothed Ogives

The data in our example can be considered to be drawn from a much larger population. As the number of observations in such a large population is thus very high then we would expect to still have finite non zero number of observations in smaller and smaller class intervals. As a curve can be considered to be made up of a very large number of straight line segments then we would expect for very large populations that the frequency polygon would be replaced by *frequency curves* and the relative frequency polygon by *relative frequency curves*.

It is not totally beyond the realms of reality to expect that such curves can be approximated by smoothing the frequency polygons and relative frequency polygons of the sample with this approximation improving as the sample size increases. For this reason a frequency curve is sometimes called a *smoothed frequency polygon*.

In a similar manner *smoothed ogives* are obtained by smoothing the cumulative frequency polygons or ogives.