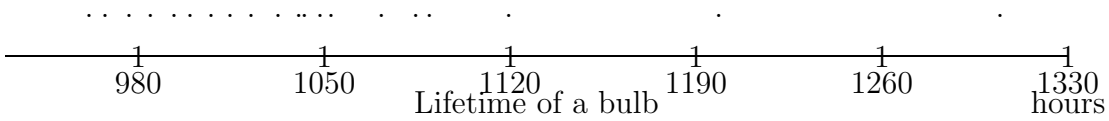
MA181 INTRODUCTION TO STATISTICAL MODELLING
GRAPHICAL REPRESENTATION OF DATA

Dotplot For all but the smallest samples, some form of graphical representation of the data is often useful in giving guidance as to how they should be analysed, or perhaps how they should nut. With less than 30 observations of a continuous random variable, plotting their values on a line probably provides the most valuable picture for, while with such a small sample one would not expect to be able to determine with any great accuracy the distribution from which the data have arisen, such a simple diagram, known as a *dotplot*, may be sufficient to show that, for instance, the assumption of a symmetric distribution would be unwise. As an example, consider these 20 measurements made of the lifetime of an electric light bulb (in hours):

975, 992, 1063, 1197, 1041, 1088, 960, 1039, 1031, 983

1147, 1006, 1070, 1013, 1118, 1021, 1051, 966, 998, 1302.

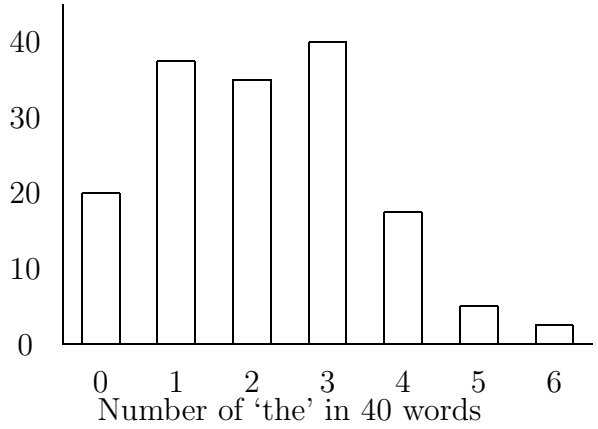
Plotting these values on a line produces the following diagram, which shows the data clustering on the left but more spread out on the right, suggesting that the distribution of the lifetime of the bulb is probably not symmetric, let alone bell-shaped, but skewed to the right, that is with a short left-hand tail and a long right-hand one.



When data have been rounded, it is quite possible for several observation to take the same value. In this case, equal points on the dotplot may be stacked, one above the other.

With 30 observations or more, it becomes feasible to draw a two dimensional picture of the sample distribution. The traditional diagrams used foe this purpose are the bar chart and the histogram, depending on whether the observations are of a discrete or a continuous random variable. In fact, the bar chart can prove quite successful even for small samples. More recently, a number of alternative diagrams have been suggested, due largely to the influence of the American statistician John Tukey. Two of these are mentioned later.

Bar chart Data from a discrete random variable can be represented by a *bar chart*, in which the possible values of the variable are marked along the horizontal axis (over the range of values for which observations occur) and a line, or bar, is drawn at each, parallel to the vertical axis, with length proportional to the frequency of the observations at that value. This is illustrated by the following diagram representing the number of occurrences of the word "the " in blocks of 40 words from Milton's *Paradise Lost*. The values 0-6 were observed, and these are shown marked with their respective bars. The height of any bar, and so the frequency of any observed value, can be read from the scale on the left. The discreteness of the random variable measured is clearly indicated by the separation of the bars.



Histogram A continuous random variable takes values over an interval (or the whole of) the real line, and an adequate representation of such a variable needs to make this clear, its picture should cover the x-axis. That being said, it is important to realise that any continuous variable can be measured only to a certain accuracy, in other words the observations made are, strictly speaking, discrete. Nonetheless, it is usual to create a diagram that displays the continuous nature of the random variable measured rather than the discreteness of the observations.

The *histogram*, a diagram that has been used for many generations, consists of a number of rectangles based on the x-axis, each with an *area* proportional to the frequency of the observations lying in the interval it covers. It is usual for most, if not all, of these intervals, known as *class intervals*, to have the same width except where the data becomes sparse, in which regions wider intervals are sometimes used. The choice of interval width, and so the number of rectangles drawn, requires some

care, too many rectangles and the outline of the distribution of the data is dominated by random fluctuations, too few and important features in the outline are smoothed out. Except with a very large data set, a histogram should contain 5-20 rectangles, with the exact number chosen being guided by the problems just mentioned and the desirability of placing *class boundaries*, the boundaries between class intervals, at convenient points. These should not be at values taken by any of the observations, otherwise problems arise in determining into which intervals those particular observations should be placed; it is usually best to set them midway between observations, or possible values of observations.

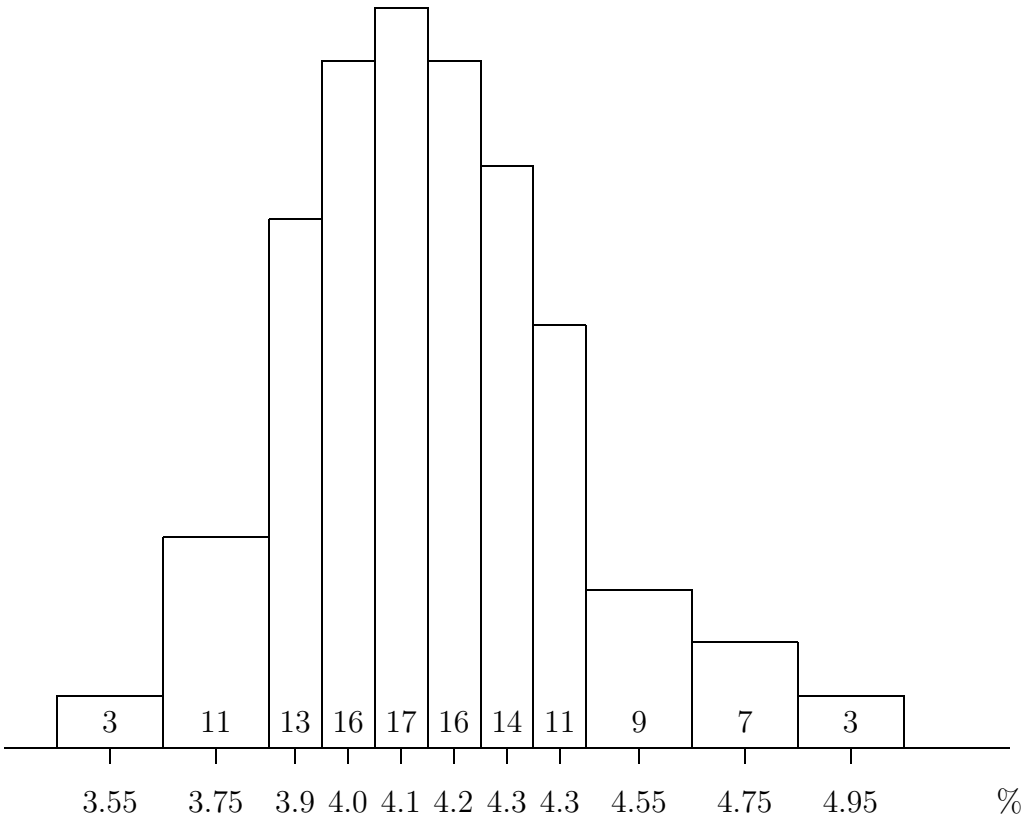
Consider the histogram drawn below of the percentage of butterfat from 120 three-year-old Ayrshire cows selected at random from a Canadian stock record book. The observations are of a continuous random variable, each one being given correct to one decimal place, so we assume that a reading like 3.7% is a rounded value of measurement lying in the interval 3.65-3.75%. When drawing a histogram of data, we could choose class intervals of 0.1%, which would result in 16 rectangles, or of width 0.2%, resulting in only eight rectangles. The former leads to raggedness in the extremes of the picture, while the latter seems to be throwing away just a little too much detail. A compromise is possible, in which intervals of width 0.1% are used in the center of the diagram and of width 0.2% in the tails. This is how the graph has been drawn, with 11 intervals and an outline that probably gives a good approximation to that of the distribution of the random variable measured. It is worth noting the following features:

(i) The area, and not the height, of each rectangle is proportional to the number of observations contained in the class interval at its base. Only if all the intervals are of the same width are the heights of the rectangles proportional to the frequencies. As a consequence, it is not possible here to mark the frequencies on a vertical scale along side the histogram. Instead, they are given inside the rectangles toward the bottom.

(ii) The class boundaries have all been set midway between pairs of possible values, i.e. at 3.45%, 3.65%, 3.85%, 3.95%, etc. In this way, no observation arises with the possibility of being placed in more than one class interval.

(iii) There is more than one way of marking relevant positions on the x-axis. One of these is to mark the class boundaries. However,

these often involve more decimal digits than the data themselves, so giving a cumbersome appearance. Perhaps the simplest is to mark just the midpoints, or *class marks*, of the intervals and to calculate anything else as and when required. This is the method adopted in the figure.



Stem-and-leaf plot The *stem-and-leaf plot* can be drawn for either a discrete or a continuous random variable. It looks somewhat like a barchart or a histogram on its side, except that the bars or rectangles are constructed from the very digits of data themselves. In this respect, for a sample from a continuous random variable, the diagram is strictly a representation of the observed rounded data.

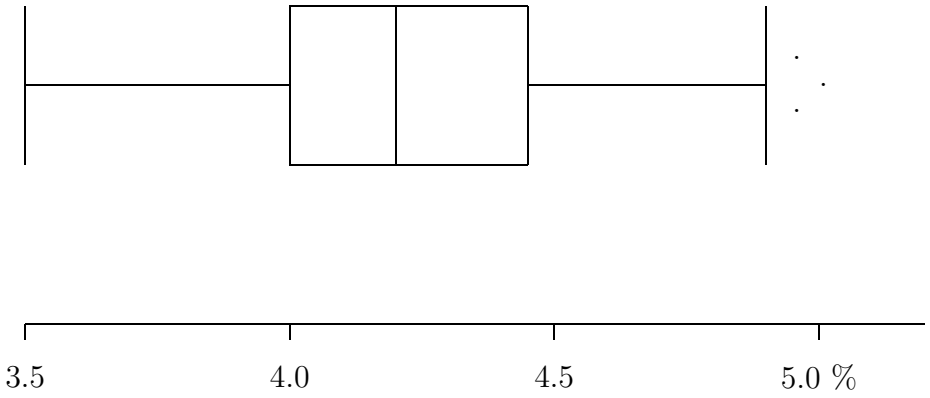
The stem of a stem-and-leaf plot consists of the most significant digits of the observations, while the leaves consist of the least significant, with a vertical line separating the two parts. In some cases, this line will represent the decimal point in the data.

The stem-and-leaf plot below shows the total scores of the 60 golfers who played all four rounds in the 1985 British Open golf tournament. The totals range from 282 to 301, the first two digits forming the stem and the last digit the leaf. The column on the left gives the cumulative counts of the data from both ends except that the number in brackets indicates the row in which the sample median lies and gives the count just for that row.

2	28	23
10	28	44444555
19	28	666667777
(13)	28	8888899999999
28	28	001111
22	29	22222333
14	29	444444455
5	29	77
3	29	8
2	30	01

Boxplot A *boxplot* (or *box-and-whisker plot* is a graph that illustrates the location and spread of a set of observations, highlighting any that are extreme in their values. The central box extends from the lower quartile, Q_1, of the data to the upper quartile, Q_3, with the median being marked by an internal line across the box. The distance $H = Q_3 - Q_1$ is known as the interquartile range. Whiskers are drawn from each end of the box extending as far as $1,5H$, or as far as the furthest observation within that range. Any observations lying even further out, known as outliers, are drawn as separate dots.

The boxplot below illustrates the butterfat data described above. The median is at 4,15, $Q_1 = 4,00$ and $Q_2 = 4,35$, so that $H = 0.35$. The lower whisker, which could extend down to 3.475, needs go no lower than the smallest observation, 3.5, while the upper whisker, which could extend to 4.875, stop at the observation 4.8. Consequently, three points (4.9,4.9 and 5.0) are outliers in the right extremity of the graph.



Percentage of butterfat from 120 cows