Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 2 Spring 2012

Exploratory data analysis: continued

Stem-and-leaf plots

A stem-and-leaf plot is like a dot plot rotated through 90°. It is a good way of getting a quick overview of the data when working by hand.

Split each number into two parts:

a	b
stem	leaf

The stems are chosen so that the number of different stems is between 5 and 20. The leaf is a single digit, possibly after rounding or cutting. If there are too few stems, use each one twice, one with leaves 0–4 and the next with leaves 5–9.

A vertical line is drawn (with a ruler), with the stems written on its left, in order from smallest at the top to biggest at the bottom, equally spaced. Work through the list of data. If the current value has stem a and leaf b, then write the leaf b on the row labelled by stem a, on the right-hand side of the line. Be sure to keep the leaves aligned vertically, to give the correct visual impression of how many leaves there are on each stem. When doing this by hand, there is no need to rearrange the leaves on each stem into order, but statistical computing packages may do this. Always give a key indicating what 'stem | leaf' represents.

Example Here are the heights (in metres) of 16 acacia trees of the same age.

> 12.2, 11.9, 11.9, 11.6, 10.7, 12.2, 11.2, 11.7, 10.2, 10.4, 0.9, 10.6, 10.6, 10.8, 9.2, 8.6.

Here is the stem-and-leaf plot.

0	9
1	
2	
3	
4	
5	
6	
7	
8	6
9	2
10	7 2 4 6 6 8
11	9 9 6 2 7
12	2 2

10 | 7 represents 10.7 metres.

To get a rough picture of the data, it may be faster to draw a stem-and-leaf plot by hand than to type the data into a computer. It is fairly easy to find the median and quartiles from a stem-and-leaf plot. Count in from each end until you have reached the middle, or a quarter of the way, respectively.

Frequency tables, bar charts and histograms

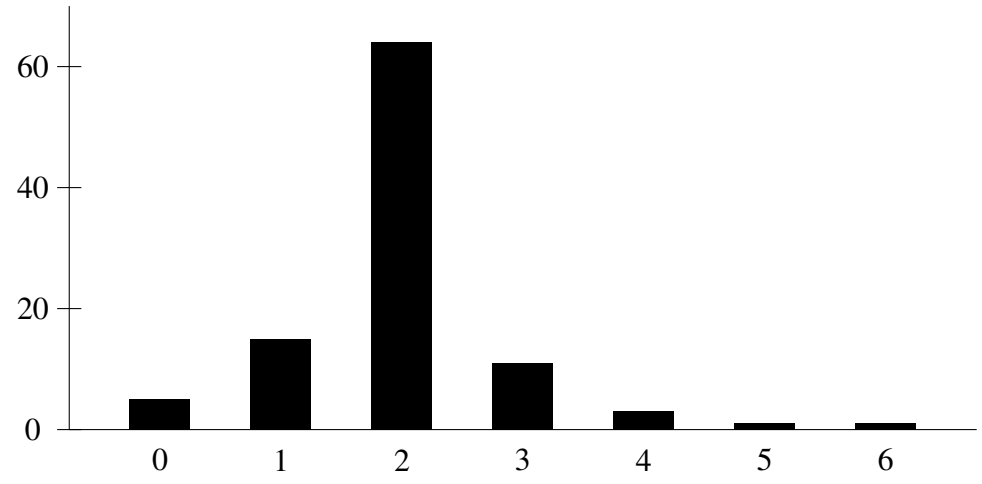
Qualitative data, or quantitative data which has not many distinct values, or has been grouped into intervals, may be shown in a frequency table.

Example The numbers of children in each of 100 families were recorded, giving the data in the first two columns of the table below. For example, 15 families had one child each.

number	frequency	relative frequency	cumulative relative frequency
0	5	0.05	0.05
1	15	0.15	0.20
2	64	0.64	0.84
3	11	0.11	0.95
4	3	0.03	0.98
5	1	0.01	0.99
6	1	0.01	1.00
$\geqslant 7$	0	0.00	1.00
total	100	1.00	

The numbers in the 'frequency' column can be shown on a bar chart. There is one bar for each category. The height of each bar is the corresponding frequency. The bars are not joined. Such a chart can also be drawn for qualititative data.

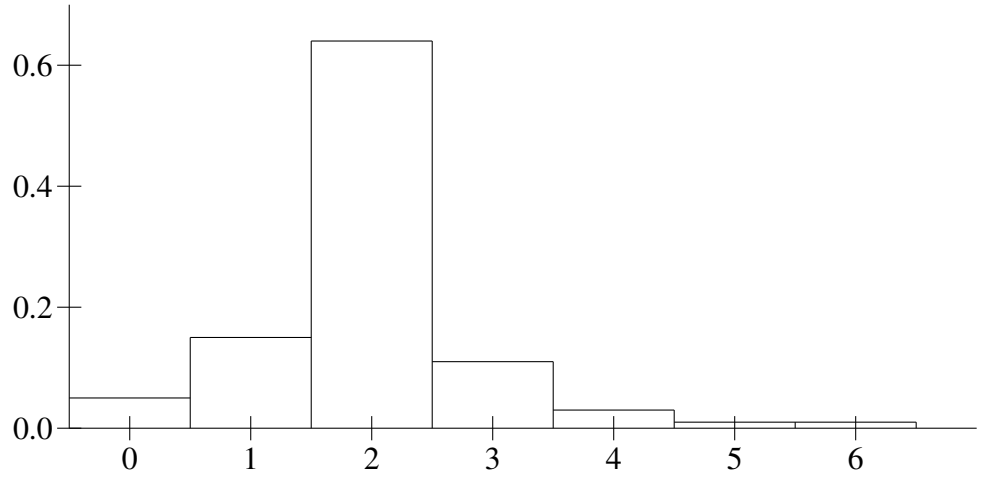
Here is a bar chart for the children data.



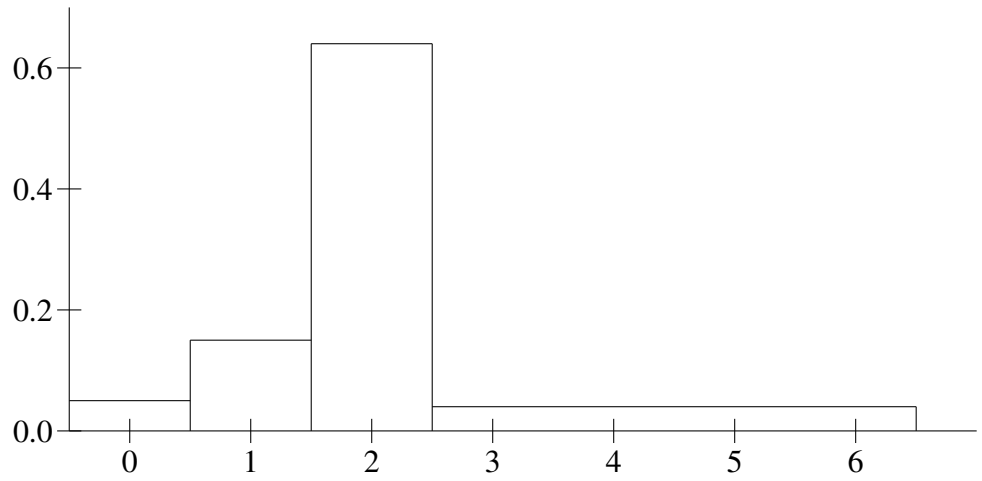
Each number in the 'relative frequency' column is obtained by dividing the frequency by the total. Thus these numbers sum to 1. This column is rather like the

probability mass function of a random variable. For numerical data, the values in this column can be plotted on a histogram. Now each box is centred on the relevant numerical value at the bottom, and its height is chosen so that the area of the box is equal to the relative frequency. Except where there are zero values of the relative frequency, the boxes touch each other.

Here is a histogram for the children data.



Sometimes, adjacent values with small frequencies are grouped together. In the children examples, the values 3, 4, 5 and 6 together have relative frequency 0.16. The rectangle covering all of these has its base of length 4; in order for its area to be equal to 0.16, its height has to be 0.04. This is shown below.



There is an alternative convention for histograms, namely that all values in the interval $[3,4)$ are shown in a rectangle whose base is that interval.

For quantitative or ordinal data, the frequency table contains a fourth column, headed 'cumulative relative frequency'. The entry in row i is the sum of the relative frequencies in row i and all rows with smaller labels. This is very like the cumulative distribution function of a random variable.

Interpreting the diagrams

The first thing to ask is: where is the centre? For example, do these data indicate tall people or short people?

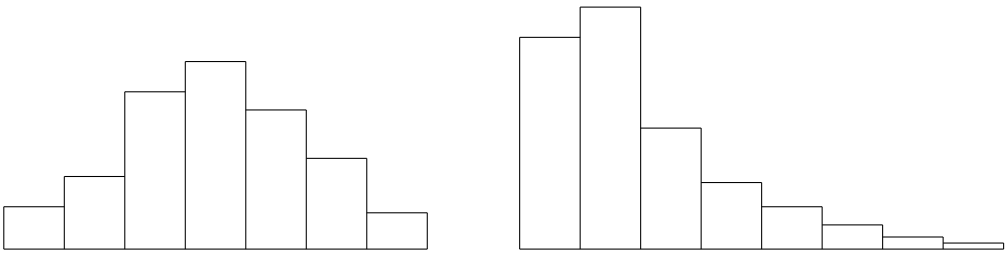
Is the diagram flat or peaked? If it is flat, the data are evenly distributed throughout the range, but if it is peaked the data are clustered in one region.

flat peaked

Are the data spread out over a large range, or clustered in a narrow one?

Does the diagram have one peak, or more? If it has one peak, the data are called *unimodal*. Marks on examinations are typically unimodal in Arts subjects, but may have two peaks in Mathematics.

Are the data more-or-less symmetric about some value, or are they skew? If there is a long tail to the right, the data are said to be *right-skewed* or *positively skewed*. Data on people's incomes are usually positively skewed, because a few people earn huge amounts of money. If there is a long tail to the left, the data are *left-skewed* or *negatively skewed*.



almost symmetric long right tail

Are there outliers? What caused them? Here are some reasons for outliers.

(i) There may be a mistake in recording the data.

(ii) Something unusual may have happened. For example, one acacia tree seems to have died; in the VW data, the two values of 30 may be values for driving around town while the rest are for long-distance driving.

(iii) This is simply an extreme value: after all, *some* value has to be the biggest!

Should we exclude outliers from further analysis? Yes in case (i); maybe in case (ii); no in case (iii).

Are there abrupt changes in the data? These may indicate changes in conditions, such as rain starting during harvest, or they may indicate mistakes: for example, one person recording all values in lb, then another person records all values in kg.

Indications of skewness

$$Q_3 - Q_2 > Q_2 - Q_1 \quad \text{and} \quad \bar{x} > \text{median} \qquad \text{indicate positive skew.}$$

$$Q_3 - Q_2 < Q_2 - Q_1 \quad \text{and} \quad \bar{x} < \text{median} \qquad \text{indicate negative skew.}$$

$$Q_3 - Q_2 = Q_2 - Q_1 \quad \text{and} \quad \bar{x} = \text{median} \qquad \text{indicate symmetry.}$$

The *third sample moment about the mean* is defined to be

$$m_3 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^3.$$

This is affected by the units we measure in. To correct for this, we use the *coefficient of skewness*, which is defined to be

$$\frac{m_3}{s^3}.$$

Positive m_3 indicates positive skew.

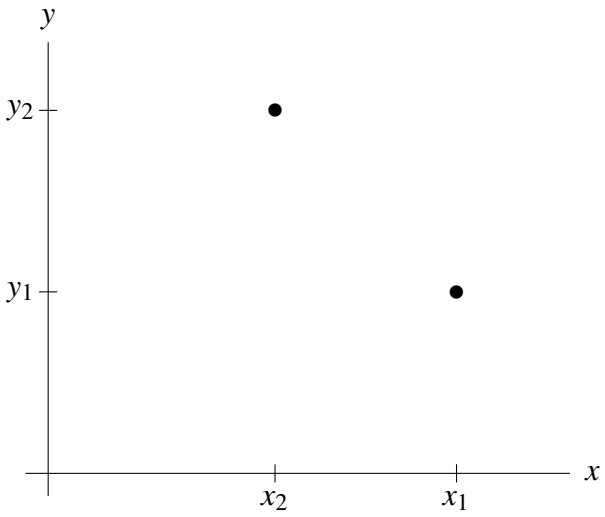
Negative m_3 indicates negative skew.

Symmetry implies that $m_3 = 0$, but it is possible to have $m_3 = 0$ without the data being symmetric.

Two or more variables measured on the same items

Suppose that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are two quantitative variables measured on the same n items (for example, the height and diameter of n trees).

We can plot $y_1, \ldots, y_n$ against $x_1, \ldots, x_n$ on an ordinary graph. This is called a *scatterplot*.



Scatterplots, and other means of displaying two or more variables, are covered in two Minitab practicals but not in lectures.

Definition Let $x_1, \ldots, x_n$ have sample mean $\bar{x}$ and sample standard deviation s_x, and let $y_1, \ldots, y_n$ have sample mean $\bar{y}$ and sample standard deviation s_y. The *sample covariance* s_{xy} is defined by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

and the *sample correlation coefficient* r is defined by

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y}.$$

For hand calculation, we rewrite s_{xy} as

$$\frac{n \sum_{i=1}^{n} (x_i y_i) - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n(n-1)}.$$

Theorem 2 *(a)* $-1 \leqslant r \leqslant +1$.

(b) $r = +1$ *if* $y_i = ax_i + b$ *with* $a > 0$.

(c) $r = -1$ *if* $y_i = ax_i + b$ *with* $a < 0$.

Proof (a) If t is any real number then

$$[(x_i - \bar{x}) + t(y_i - \bar{y})]^2 \geqslant 0,$$

so

$$\sum_{i=1}^{n} [(x_i - \bar{x}) + t(y - \bar{y})]^2 \geqslant 0.$$

$$\begin{aligned} \text{The left-hand side} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} 2t(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n} t^2 (y_i - \bar{y})^2 \\ &= (n-1)s_x^2 + 2t(n-1)s_{xy} + t^2(n-1)s_y^2 \\ &= (n-1)f(t), \end{aligned}$$

where $f(t) = s_x^2 + 2ts_{xy} + t^2 s_y^2$. This quadratic function of t is always positive or zero, so its discriminant (the part we think of as "$b^2 - 4ac$") is negative or zero: that is,

$$4s_{xy}^2 - 4s_x^2 s_y^2 \leqslant 0.$$

We can divide by 4 without changing the sign, so $s_{xy}^2 \leqslant s_x^2 s_y^2$. The quantity $s_x^2 s_y^2$ is positive, so again we can divide by it without changing the sign, so

$$\frac{s_{xy}^2}{s_x^2 s_y^2} \leqslant 1.$$

Taking square roots, we obtain

$$-1 \leqslant \frac{s_{xy}}{s_x s_y} \leqslant 1.$$

(b) and (c) Since $y_i = ax_i + b$ for $i = 1, \ldots, n$, Theorem 1(a) shows that $\bar{y} = a\bar{x} + b$, so

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} (x_i - \bar{x})(ax_i + b - a\bar{x} - b) \\ &= \sum_{i=1}^{n} a(x_i - \bar{x})^2 \\ &= a \sum_{i=1}^{n} (x_i - \bar{x})^2, \end{aligned}$$

so

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{a}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = as_x^2. \end{aligned}$$

Theorem 1(b) shows that $s_y = |a|\, s_x$. Therefore

$$r = \frac{s_{xy}}{s_x s_y} = \frac{as_x^2}{s_x\,|a|\,s_x} = \frac{a}{|a|}.$$

If $a > 0$ then $|a| = a$ and so $r = +1$. If $a < 0$ then $|a| = -a$ and so $r = -1$. ■

The sample correlation coefficent r gives some information about the scatterplot.

If $r = 1$, the points lie on a line with positive slope.

If r is positive but less than 1, then y_i tends to increase with x_i, but the relationship is not linear.

If $r = 0$, then there is no linear pattern (there may be a non-linear one) and y_i does not consistently either increase or decrease as x_i increases.

If r is negative but greater than -1, then y_i tends to decrease as x_i increases, but the relationship is not linear.

If $r = -1$, the points lie on a line with negative slope.