Health Sciences M.Sc. Programme

Applied Biostatistics

Week 2: Frequency distributions

Types of data

Data can be summarised to help to reveal information they contain. We do this by calculating numbers from the data which extract the important material. These numbers are called **statistics**. A statistic is anything calculated from the data alone.

It is often useful to distinguish between three types of data: qualitative, discrete quantitative and continuous quantitative. **Qualitative** data arise when individuals may fall into separate classes. These classes may have no numerical relationship with one another at all, e.g. sex: male, female; types of dwelling: house, maisonette, flat, lodgings; eye colour: brown, grey, blue, green, etc. **Quantitative** data are numerical, arising from counts or measurements. If the values of the measurements are integers (whole numbers), like the number of people in a household, or number of teeth which have been filled, those data are said to be **discrete**. If the values of the measurements can take any number in a range, such as height or weight, the data are said to be **continuous**. In practice there is overlap between these categories. Most continuous data are limited by the accuracy with which measurements can be made. Human height, for example, is difficult to measure more accurately than to the nearest millimetre and is more usually measured to the nearest centimetre. So only a finite set of possible measurements is actually available, although the quantity 'height' can take an infinite number of possible values, and the measured height is really discrete. However, the methods described below for continuous data will be seen to be those appropriate for its analysis.

We shall refer to qualities or quantities such as sex, height, age, etc. as **variables**, because they vary from one member of a sample to another. A qualitative variable is also termed a **categorical variable** or an **attribute**. We shall use these terms interchangeably.

Frequency distributions

When data are purely qualitative, the simplest way to deal with them is to count the number of cases in each category. For example, in the analysis of the census of a psychiatric hospital population, one of the variables of interest was the patient's principal diagnosis. To summarise these data, we count the number of patients having each diagnosis. The results are shown in Table 1. The count of individuals having a particular quality is called the **frequency** of that quality. For example, the frequency of schizophrenia is 474. The proportion of individuals having the quality is called the **relative frequency** or **proportional frequency**. The relative frequency of schizophrenia is 474/1467 = 0.32 or 32%. The set of frequencies of all the possible categories is called the **frequency distribution** of the variable.

Table 1. Principle diagnosis of patients in Tooting Bec Hospital

Diagnosis	Number of patients
Schizophrenia	474
Affective disorders	277
Organic brain syndrome	405
Subnormality	58
Alcoholism	57
Other and not known	196
Total	1467

Table 2. Likelihood of discharge of patients in Tooting Bec Hospital

Discharge:	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
unlikely	871	0.59	871	0.59
possible	339	0.23	1210	0.82
likely	257	0.18	1467	1.00
Total	1467	1.00	1467	1.00

Table 3. Parity of 125 women attending antenatal clinics at St. George's Hospital

Parity	Frequency	Relative frequency (percent)	Cumulative frequency	Relative cumulative frequency (percent)
0	59	47.2	59	47.2
1	44	35.2	103	82.4
2	14	11.2	117	93.6
3	3	2.4	120	96.0
4	4	3.2	124	99.2
5	1	0.8	125	100.0
Total	125	100.0	125	100.0

In this census we assessed whether patients were 'likely to be discharged', 'possibly to be discharged' or 'unlikely to be discharged'. The frequencies of these categories are shown in Table 2. Likelihood of discharge is a qualitative variable, like diagnosis, but the categories are ordered. This enables us to use another set of summary statistics, the cumulative frequencies. The **cumulative frequency** for a value of a variable is the number of individuals with values less than or equal to that value. Thus, if we order likelihood of discharge from 'unlikely', through 'possibly' to 'likely' the cumulative frequencies are 871, 1210 (= 871 + 339) and 1467. The **relative cumulative frequency** for a value is the proportion of individuals in the sample with values less than or equal to that value. For the example they are 0.59 (= 871/1467), 0.82 and 1.00. Thus we can see that the proportion of patients for whom discharge was not thought likely was 0.82 or 82%.

Table 4. FEV1 (litres) of 57 male medical students

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

Table 5. Frequency distribution of FEV1 in 57 male medical students

FEV1	Frequency	Relative frequency (percent)
2.0	0	0.0
2.5	3	5.3
3.0	9	15.8
3.5	14	24.6
4.0	15	26.3
4.5	10	17.5
5.0	6	10.5
5.5	0	0.0
Total	57	100.0

As we have noted, likelihood of discharge is a qualitative variable, with ordered categories. Sometimes this ordering is taken into account in analysis, sometimes not. Although the categories are ordered these are not quantitative data. There is no sense in which the difference between ‘likely’ and ‘possibly’ is the same as the difference between ‘possibly’ and ‘unlikely’.

Table 3 shows the frequency distribution of a quantitative variable, parity. This shows the number of previous pregnancies for a sample of women booking for delivery at St. George’s Hospital. Only certain values are possible, as the number of pregnancies must be an integer, so this variable is discrete. The frequency of each separate value is given.

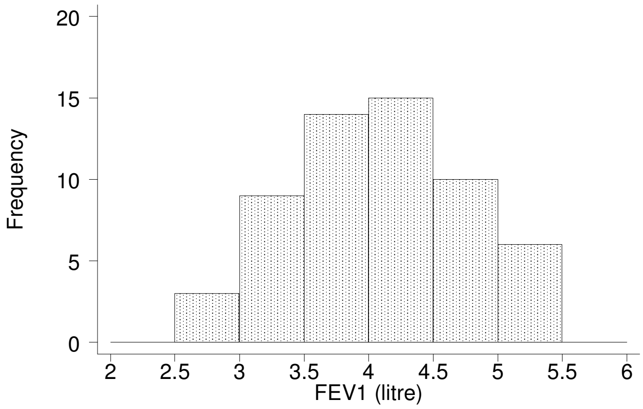
Table 4 shows a continuous variable, forced expiratory volume in one second (FEV1) in a sample of male medical students. As most of the values occur only once, to get a useful frequency distribution we need to divide the FEV1 scale into class intervals, e.g. from 3.0 to 3.5, from 3.5 to 4.0, and so on, and count the number of individuals with FEV1s in each class interval. The class intervals should not overlap, so we must decide which interval contains the boundary point to avoid it being counted twice. It is usual to put the lower boundary of an interval into that interval and the higher boundary into the next interval. Thus the interval starting at 3.0 and ending at 3.5 contains 3.0 but not 3.5. We can write this as ‘3.0 — ’ or ‘3.0 — 3.5^{-}’or ‘3.0 — 3.499’

If we take a starting point of 2.5 and an interval of 0.5 we get the frequency distribution shown in Table 5. Note that this is not unique. If we take a starting point of 2.4 and an interval of 0.2 we get a different set of frequencies.

Table 6. Tally system for finding the frequency distribution of FEV1

FEV1	Tally marks	Frequency
2.0—2.5$^-$		0
2.5—3.0$^-$	///	3
3.0—3.5$^-$	𝍸 ////	9
3.5—4.0$^-$	𝍸 𝍸 ////	14
4.0—4.5$^-$	𝍸 𝍸 𝍸	15
4.5—5.0$^-$	𝍸 𝍸	10
5.0—5.5$^-$	𝍸 /	6
5.5—6.0$^-$		0
Total		57

Figure 1. Histogram of FEV1: frequency scale



The frequency distribution can be calculated easily and accurately using a computer. Manual calculation is not so easy but must be done carefully and systematically. One way recommended by many texts is to set up a tally system, as in Table 6. We go through the data and for each individual make a tally mark by the appropriate interval. We then count up the number in each interval. In practice this is very difficult to do accurately, and it needs to be checked and double-checked.

Histograms and other frequency graphs

Graphical methods are very useful for examining frequency distributions. The most common way of depicting a frequency distribution is by a **histogram**. This is a diagram where the class intervals are on an axis and rectangles with heights or areas proportional to the frequencies erected on them. Figure 1 shows the histogram for the FEV1 distribution in Table 4. The vertical scale shows frequency, the number of observations in each interval.

Figure 2 shows a histogram for the same distribution, with frequency per unit FEV1 (or frequency density) shown on the vertical axis. The distributions appear identical and we may well wonder whether it matters which method we choose.

Figure 2. Histogram of FEV1: frequency per unit FEV1 or frequency density scale

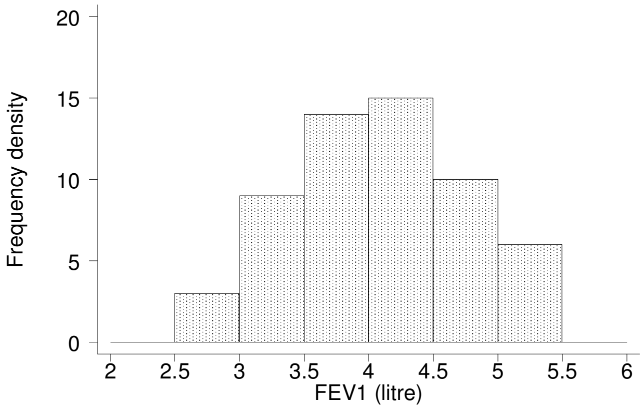
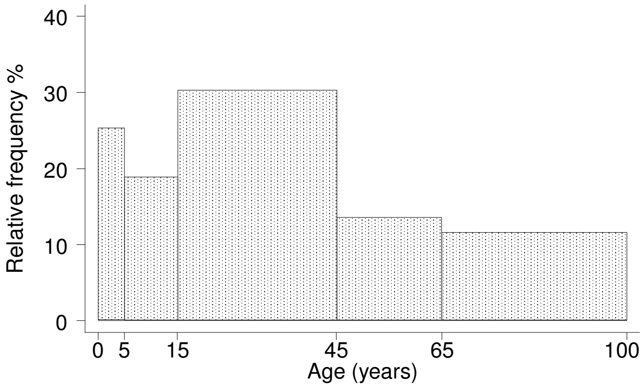


Table 7. Distribution of age in people suffering accidents in the home

Age group	Relative frequency (per cent)	Relative frequency per year (per cent)
0 — 4	25.3	5.06
5 — 14	18.9	1.89
15 — 44	30.3	1.01
45 — 64	13.6	0.68
65+	11.7	0.33

Figure 3. Age distribution of home accident victims: relative frequency scale



We see that it does matter when we consider a frequency distribution with unequal intervals, as in Table 6. If we plot the histogram using the heights of the rectangles to represent relative frequency in the interval we get Figure 3, whereas if we use the relative frequency per year we get Figure 4. These histograms tell different stories. Figure 3 suggests that the most common age for accident victims is between 15 and 44 years, whereas Figure 4 suggests it is between 0 and 4. Figure 4 is correct, Figure 3 being distorted by the unequal class intervals. It is therefore preferable in general to use the frequency per unit rather than per class interval when plotting a histogram. The frequency for a particular interval is then represented by the area of the rectangle on that interval. Only when the class intervals are all equal can the frequency for the class interval be represented by the height of the rectangle.

Figure 4. Age distribution of home accident victims: relative frequency density

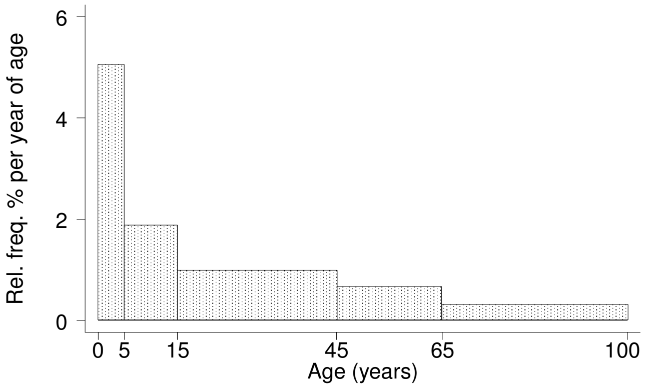


Figure 5. Histogram for a discrete variable, parity

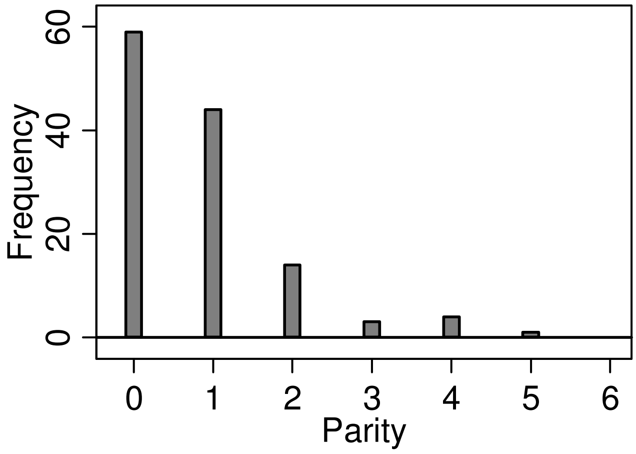


Figure 6. Frequency polygon for the FEV data

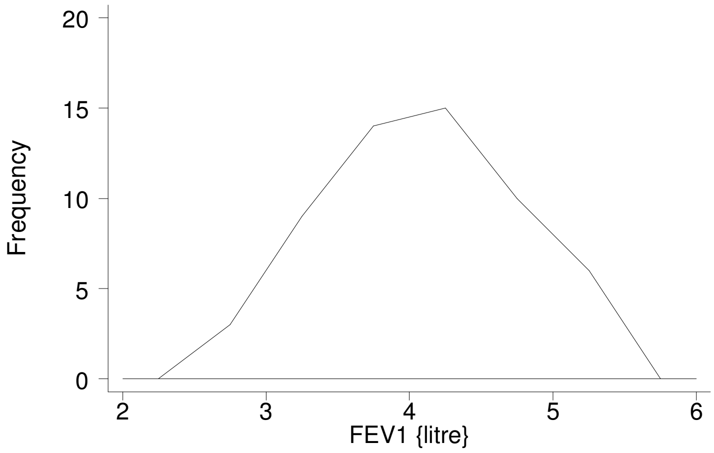


Figure 7. Frequency polygons for PEF comparing males and females

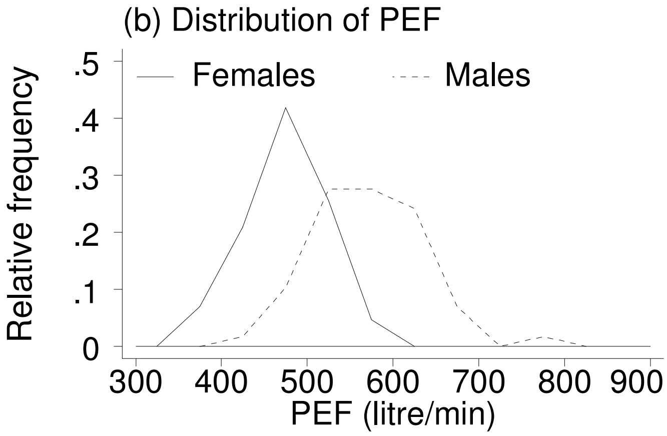


Figure 5 shows a histogram for a discrete variable, parity. Here we have shown the bars as distinct and separate, as only the integer values are possible.

Histograms are not the only graphical method to show a frequency distribution. Another which you may come across quite often is the frequency polygon

A **frequency polygon** joins up the mid-points of the tops of the bars. The bars are then removed to leave a graph like Figure 6. Frequency polygons are useful for showing more than one frequency distribution together. For example, Figure 7 shows the distribution of PEF is female and male students, enabling us to compare the two distributions easily.

The box and whisker plot is another frequency graph which is quite widely used, but we shall discuss it later.

Shapes of frequency distribution

Figure 1 shows a frequency distribution of a shape often seen in health data. The distribution is roughly symmetrical about its central value and has frequency concentrated about one central point. The most common value is called the **mode** of the distribution and Figure 1 has one such point, as do Figure 4 and Figure 5. They are **unimodal**. Figure 8 shows a very different shape. Here there are two distinct modes one near 5 and the other near 8.5. This distribution is **bimodal**. We must be careful to distinguish between the unevenness in the histogram which results from using a small sample to represent a large population and those which result from genuine bimodality in the data. The trough between 6 and 7 in Figure 8 is very marked and might represent a genuine bimodality. In this case we have children some of whom may have a condition which raises the cholesterol level and some of whom do not. We actually have two separate populations represented with some overlap between them. However, almost all distributions encountered in medical statistics are unimodal.

Figure 8. Serum cholesterol in children from kinships with familial hypercholesterolemia

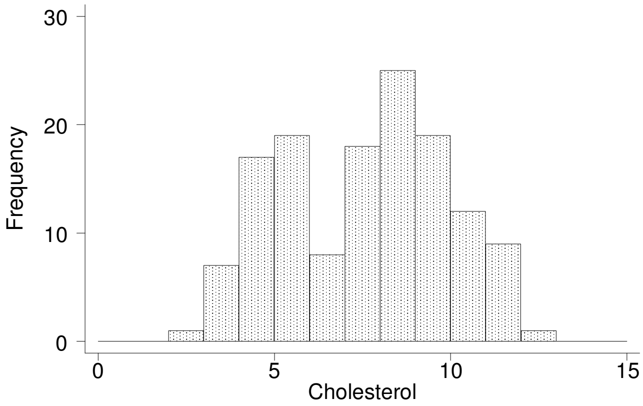


Figure 9. Serum triglyceride in cord blood from 282 babies

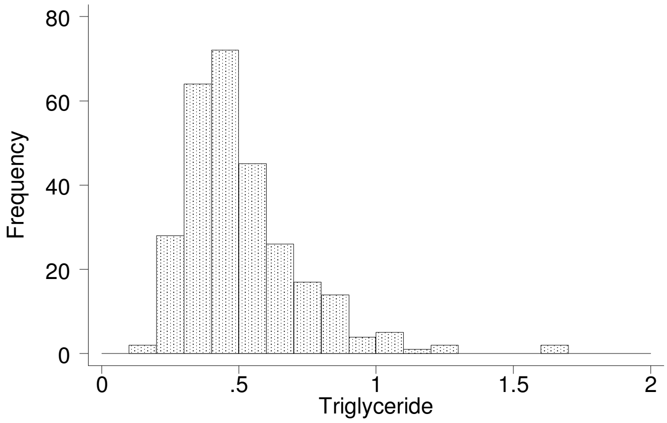


Figure 9 differs from Figure 1 in a different way. We have already noted that the distribution of FEV1 is symmetrical. The distribution of serum triglyceride is **skew**, that is, the distance from the central value to the extreme is much greater on one side than it is on the other. The parts of the histogram near the extremes are called the **tails** of the distribution. If the tail on the right is longer than the tail on the left as in Figure 9, the distribution is **skew to the right** or **positively skew**. Figures 4 and 5 also show distributions which are positively skew.

If the tails are equal the distribution is **symmetrical**, as in Figure 1. Most distributions encountered in health work are symmetrical or skew to the right, for reasons we shall discuss later.

If the tail on the left is longer than the tail on the right, the distribution is **skew to the left** or **negatively skew**. This is much more unusual in health data. Figure 10 shows an example, gestational age at birth. This is a rather artificially negative skew distribution, because some babies are delivered early because of obstetric intervention and none are allowed to be born later than 44 weeks for the same reason.

Figure 10. Distribution of gestational age

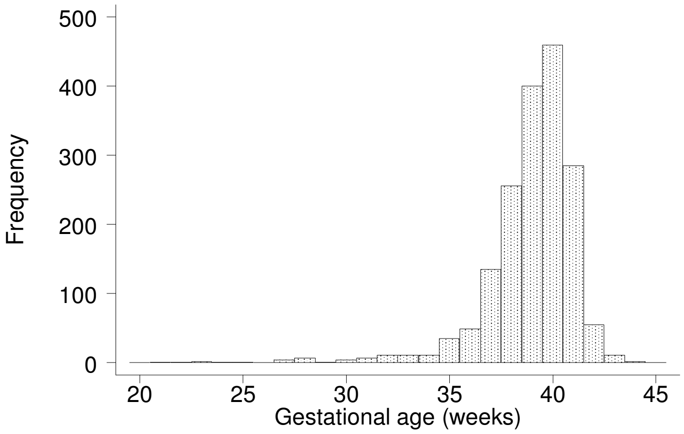
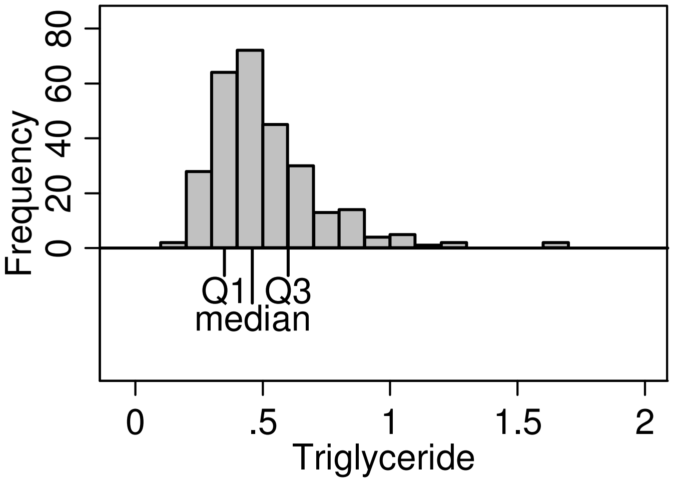


Figure 11. Histogram of serum triglyceride in cord blood, showing the positions of the three quartiles



Medians and quantiles

We often want to summarise a frequency distribution in a few numbers, for ease of reporting or comparison. The most direct method is to use quantiles. The **quantiles** are values which divide the distribution such that there is a given proportion of observations below the quantile. For example, the median is a quantile. The **median** is the central value of the distribution, such that half the points are less than or equal to it and half are greater than or equal to it. For the FEV1 data the median is 4.1, the 29th value in Table 4. If we have an even number of points, we choose a value midway between the two central values.

Other quantiles which are particularly useful are the **quartiles** of the distribution. The quartiles divide the distribution into four equal parts. The second quartile is the median. Figure 11 shows the three quartiles for the serum triglyceride data.

We often divide the distribution into 100 parts at 99 **centiles** or **percentiles**. The median is thus the 50th centile.

We use the quartiles in another graph to show a frequency distribution, the **box and whisker plot**. Figure 12 shows two examples. We draw a box whose height is the distance between the two quartiles and draw a line across at the median. We then draw lines stretching beyond the box to the minimum and maximum.

Figure 12. Two examples of box and whisker plots

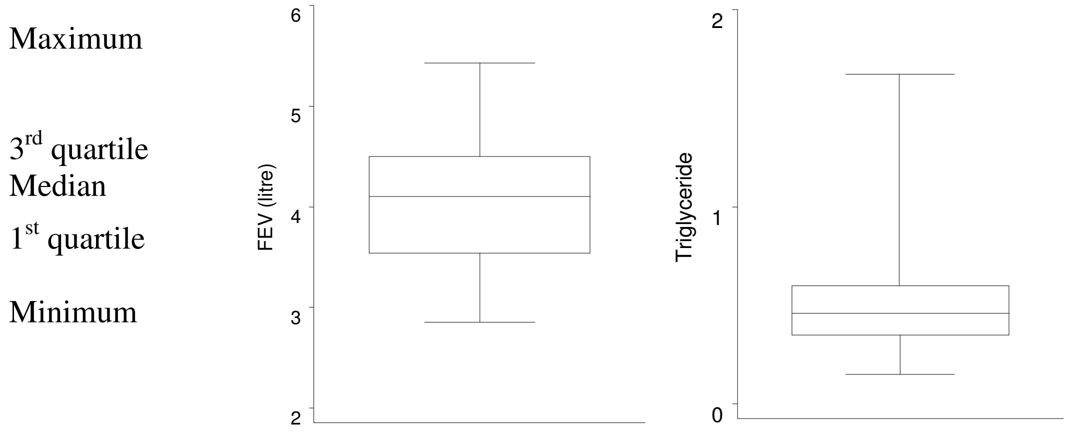


Figure 13. Four box and whisker plots to show the difference or similarity between four groups of subjects

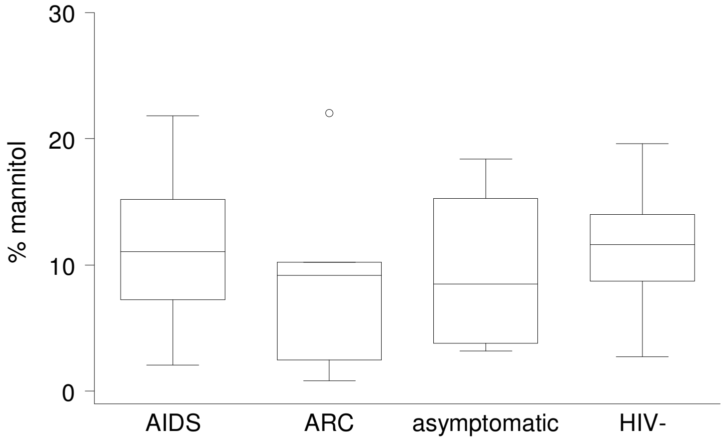


Figure 12 shows a symmetrical distribution, where the whiskers are of similar length, and a positively skew distribution, where the upper whisker is much longer than the lower.

This can be used to show several distributions together, as in Figure 13. This shows the distribution of mannitol absorption for four groups of subjects, classified by their HIV status and symptomatology. Points more than 1.5 box heights from the top or bottom of the box are often shown separately, as outlying points. This is the case for ARC. The numbers in the groups are small in this example, particularly for ARC, which makes them rather uneven.

Variability

The median is a measure of the central tendency or position of the middle of the distribution. We shall also need a measure of the spread, dispersion or variability of the distribution.

One obvious measure is the **range**, the difference between the highest and lowest values. This is a useful descriptive measure, but has disadvantages. First, it depends only on the extreme values and so can vary a lot from sample to sample. Second, it depends on the sample size. The larger the sample is, the further apart the extremes are likely to be. We can see this if we consider a sample of size 2. If we add a third member to the sample the range will only remain the same if the new observation falls between the other two, otherwise the range will increase. Third, it is mathematically difficult to deal with and is not suitable for use in analysis. It is a descriptive measure.

We can get round the second of these problems by using the **interquartile range**, the difference between the first and third quartiles.

The IQR is less variable than the range, but it is still is quite variable from sample to sample. It is also difficult to use in analysis. Although a useful descriptive measure, it is not the one preferred for purposes of comparison.

The interquartile range is also often presented as the first quartile and third quartile, rather than the difference between them.

J. M. Bland, 8 August 2006