Suggestions from members for possible future items in the Bulletin are always welcome. One of the most frequent requests we have received in the past has been for an introductory guide to data analysis. One basic reason for producing such a guide is that many of the people collecting wader measurements have not been taught statistical methods. Self instruction is a good way to learn but there is no book on statistics which starts from very basic principles and deals in a clear way with the problems frequently encountered by wader measurers. [The original introduction to the series noted the intention of the British Trust for Ornithology to produce such a guide. This was published in 1986 as Statistics for Ornithologists *(authors: Jim Fowler and Louis Cohen) and is recommended as further reading].*

To meet the undoubted need and popular demand we invited Dr Jeremy Greenwood to write a series of articles on basic statistics [. . .]. He is a particularly appropriate author because he makes considerable use of statistics both in biological research and in teaching biology students. He is also blessed with the faculty of clear concise writing. [. . .]

The [1978] *Editors*

Introductory statistics 1

JEREMY J.D. GREENWOOD

Department of Biological Sciences, The University, Dundee, Scotland, UK

Citation: Greenwood, J.J.D. 1978. Introductory statistics 1. *Wader Study Group Bull.* 24: 16–19.

Introduction

Statistics provides techniques for handling numerical data and for reducing them to a state of order, so that scientific conclusions may be drawn from them. With the present emphasis among wader workers on biometrical and population data, few readers of this *Bulletin* will need convincing that statistics has something to offer them. However many people find that, although there are plenty of books that offer a whole set of statistical tools, these books do not adequately explain which tool is appropriate to which job or allow the reader to feel that he understands the principles underlying the use of each tool.

These notes are designed to fill that gap. I hope that they will be useful to those with no previous knowledge of statistics, allowing them to carry out elementary analyses correctly and confidently, as well as providing a secure base from which to go on to more advanced methods. My approach will be to explain the methods and the ideas behind them in common-sense terms. I shall assume no mathematical ability beyond that of elementary arithmetic. In places it may be necessary to use some unfamiliar terminology but this will always be fully explained – not only in verbal terms but also by illustration through an example.

This first note is concerned with the *description of data.* The two other major areas of applied statistics – the estimation of population characteristics and the testing of hypotheses – will be considered later.

Data care

Let us assume you look after your equipment well and take care in making your measurements. Most people do. Equal care should be taken to conserve the integrity of the data. It should be recorded clearly and unambiguously – and any calculations you do should be recorded equally carefully. Though it is sensible to keep a permanent record separate from your field notebook, the latter is always the best source of the data for the calculations, since the permanent record will almost always contain transcription errors.

Never throw away original data – you may always need it for checking.

In any data, the last digit used should indicate the precision of the measurement. A record of 21 cm implies measurement to the nearest 1 cm; one of 21.0 cm implies measurement to the nearest 0.1 cm. The question of how many significant figures should be preserved during the calculations is not one that can be answered shortly. It is determined by the number required in the answer and by the extent to which rounding off carried out early in the calculation affects the accuracy of the final answer. Since modern calculators remove so much of the arithmetic labour, there is generally no reason for not working to the limits of accuracy of the calculator.

Graphs

The human brain is a marvellous mechanism for reducing the chaos of a multitude of incoming signals to a meaningful order. This is especially true when the input is visual, which is why graphs are so valuable. Properly produced graphs allow a great deal of information to be recorded in such a way that the brain can obtain a good overall impression of the whole set. Thus graphs are not just an ideal way of presenting one's results to other people but also a valuable aid during analysis of this data. Most analyses benefit by the data being summarised graphically at an early stage, so that one gets an overall impression, which is useful in deciding what calculations one needs to carry out.

Graphs should be drawn carefully, with clearly and unambiguously labelled axes. The type of graph to be used depends on the type of data and the use to which one wishes to put it. A way to learn how best to present data graphically is to look critically at published graphs. Ask yourself what the author is trying to get over. Is he successful in getting it over clearly? If he is not, ask yourself why – and remember not to repeat his mistakes when drawing your own graphs.

Good graphs contain a lot of information and convey that information clearly. Unfortunately, it is generally true that the more information in a graph then the less clear is the overall message. A balance between information content and clarity has to be struck. You must decide where and how to strike it.

The arithmetic mean

Faced with a set of data most people almost automatically calculate the mean. Why? Because the mean summarises in a single figure a great deal of the information in the whole data set. That single figure is comprehensible more quickly than the whole data set and this is a great advantage.

There are, in fact, several sorts of means. There are even measures of the "average" that are not, in the strict sense, means.

The usual mean is the arithmetic mean. Everyone knows how to calculate it, but here is a formula, for the sake of introducing some commonly-used terminology:

$$\bar{x} = \frac{\Sigma(x)}{n}$$

We commonly use x to indicate a single value of the measurement with which we are dealing.

$\Sigma(x)$ means the sum of all the x values in the set.

n is commonly used to indicate the number of items (x values) in the set.

$\bar{x}$ is the mean.

Thus the formula states: "The mean is equal to all the values added together and divided by the total number of values".

Variation

One thing the mean does not indicate is the amount of variation in the data. Consider these two sets:

A: 2, 4, 5, 6, 8.

B: 4, 5, 5, 5, 6.

They have the same mean but set A is clearly more variable than set B.

It is tempting to express such variation by quoting the smallest and the largest values in the set – what ornithologists commonly, though a little incorrectly, refer to as the range. However, the range contains information about only two of the values in the set. Furthermore, the range is affected by sample size; large samples are more likely to contain very large and very small values than are small ones. It is obviously much better to express variation in terms of a measure that includes information from all the data and which is independent of sample size. Such measures are the variance and the standard deviation. They are easy to calculate – indeed, many pocket calculators perform the operation automatically – so there is now never any excuse for quoting ranges rather than variances or standard deviations.

Variance

Variation can most obviously be measured in terms of the extent to which each individual datum is different from the mean. We may measure the difference of an individual datum as $(x - \bar{x})$. For statistical reasons that need not concern us, it is more useful to consider the squares of these deviations from the mean rather than the deviations themselves – i.e. the values $(x - \bar{x})^2$. If all these are added together, we obtain a quantity commonly known as "the sum of the squares", $\Sigma(x - \bar{x})^2$.

For most practical purposes the variance is calculated as

$$\frac{((x - \bar{x})^2)}{(n - 1)}$$

i.e. we divide the sum of the squares by $(n - 1)$.

If we divided by n instead of $(n - 1)$ the variance would be the mean of the values of $(x - \bar{x})^2$. It is, indeed, convenient to think of it as such a mean. I shall briefly explain in a future note why one does in fact divide by $(n - 1)$ and not by n.

A numerical example

A sample of five Blackbirds' nests and clutch sizes of 3, 4, 4, 5, 6. What is the variance of clutch size?

$\bar{x} = (3 + 4 + 4 + 5 + 6) / 5 = 4.4$

$(x - \bar{x})^2 = (3 - 4.4)^2 + (4 - 4.4)^2 + (4 - 4.4)^2 + (5 - 4.4)^2 + (6 - 4.4)^2$

$= (-1.4)^2 + (-0.4)^2 + (-0.4)^2 + 0.6^2 + 1.6^2$

$= 1.96 + 0.16 + 0.16 + 0.36 + 2.56$

$= 5.20$

Variance $= 5.20/5 - 1 = 1.3$

Easier arithmetic

Calculating the sum of squares directly (i.e. as $\Sigma(x - \bar{x})^2$) may seem rather long-winded. It is often easier to calculate:

$$\Sigma x^2 - (\Sigma x)^2/n$$

In this expression, Σx^2 states that one takes each value of x, squares it, and adds up all the squares. This is quite different from adding up all the values and squaring the sum – which is $(\Sigma x)^2$.

This expression is algebraically identical to $\Sigma(x - \bar{x})^2$, so it gives the same arithmetic result. Turning back to the Blackbird clutches:

$\Sigma x = 3 + 4 + 4 + 5 + 6 = 22$

$\Sigma x^2 = 9 + 16 + 16 + 25 + 36 = 102$

$\Sigma x^2 - \Sigma(x)^2/n = 102 - 22^2/5 = 102 - 96.8 = 5.2$

Standard deviation

Suppose we measured a set of wings in millimetres. The deviations from the mean would also be in millimetres. But their square, and therefore the variance, would be in square millimetres.

It is rather odd to measure the variation of lengths in terms of areas! Indeed, while the variance is a convenient measure for mathematicians, it lacks any common-sense interpretation. This is not true of the standard deviation, which is the square root of the variance. Its common-sense interpretation may be expressed in two ways:

a. In a large sample, 68% of the data lie within one standard deviation on either side of the mean.

b. In a large sample, 95% of the data lie within 1.96 standard deviation on either side of the mean.

The standard deviation is usually symbolised as s. Correspondingly, the variance is symbolised as s^2.

Thus, using the second formula for the sum of squares given above, we may give a formula for the standard deviation:

$$s = \sqrt{\frac{[\Sigma x^2 - (x)^2/n]}{(n-1)}}$$

Accuracy in calculation

For a large data set both Σx^2 and $(\Sigma x)^2/n$ may be very large. However, the difference between them will be small if the standard deviation is small. Thus any rounding-off of the final digits of the Σx^2 or $(\Sigma x)^2/n$ values may have comparatively large effect on the value of the difference between them and thus on the standard deviation.

The importance of this may be seen by considering the set of wing-lengths 11.7, 11.8, 11.8, 11.9, 12.0, 12.0, 12.0, 12.1, 12.1, and 12.2 cm. For this set, $\Sigma x = 119.6$, $\Sigma x^2 = 1430.64$, and $(\Sigma x)^2/n = 1430.42$. Thus the standard deviation is 0.158 cm. But suppose one had decided to work only to four significant figures, on the grounds that the original data were only accurate to three significant figures. Then $\Sigma x^2 = 1431$ and $(\Sigma x)^2/n = 1430$, giving a standard deviation of 0.333 – a highly inaccurate figure.

Beware of calculating machines that carry so few significant figures that they introduce such rounding-off errors during the calculation of a standard deviation.

Reducing the size of the numbers

Suppose that one has a set of wing-lengths, all over 100 mm. The calculations will be considerably eased if one takes 100 off each value before carrying out the calculations. Of course, one must remember to add it on again when one comes to work out the means.

Does one need to make any adjustment when one works out the variance or the standard deviation? No: the amount of variation is not changed by subtracting the same amount from each value in the set; the values of $(x - \bar{x})$ remain unaffected; and the values of Σx^2 and of $(\Sigma x)^2/n$ are reduced by exactly the same amount, so that the difference between them is unaffected.

The effects on the values of Σx^2 and $(\Sigma x)^2/n$ are another advantage of this sort of reduction. Subtracting 100 from each x value causes each of these values to be reduced by 10000n, which will be a very large amount if n is large. Thus such reduction may help to prevent the rounding off errors that might otherwise occur when using a calculator of somewhat limited capacity.

The Normal distribution

The distribution of data about the mean may take all sorts of shapes – it may be bimodal (with two peaks), grossly asymmetrical, and so forth. Commonly, however, biological data have a distribution close to the "Normal distribution". This is a particular statistical distribution that forms the basis of many statistical procedures. It is symmetrical and when drawn as a frequency distribution graph looks like the vertical section of a bell (Figure 1).

Most ornithological data are close enough to Normality for one to use standard statistical methods on them. If your data happen to be grossly non-Normal in distribution then the usual methods may be inappropriate. If in doubt, consult a competent statistician.

The common-sense interpretation of the standard deviation given earlier apply only to Normal data.

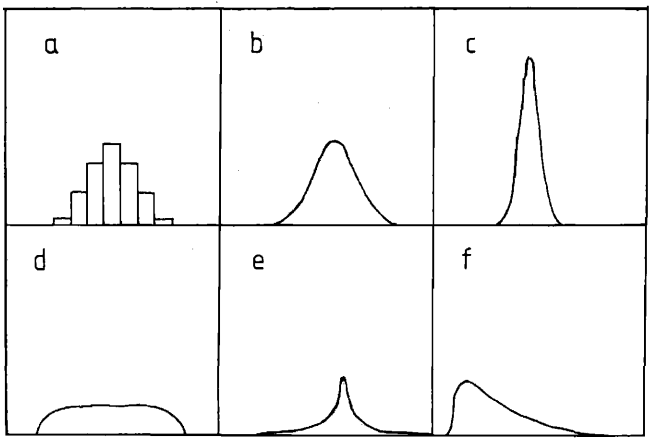

Figure 1. Some frequency distributions. In each case the horizontal axis represents the values of a variable, such as wing-length, and the vertical axis represents the number of individuals with each value of the variable, such as the number of birds of each wing-length.
a. A histogram of Normally distributed data.
b. Smoothed version of **a.**
c. Another smoothed Normal curve, with a smaller standard deviation than **b.**
d. & **e.** Symmetrical but not Normal distributions.
f. An asymmetrical distribution.