Introductory statistics 2

JEREMY J.D. GREENWOOD

Department of Biological Sciences, The University, Dundee, UK

Citation: Greenwood, J.J.D. 1979. Introductory statistics 2. *Wader Study Group Bull.* 25: 24–27.

Introduction

In this part, I wish to consider how one may reduce errors in calculation, what we can conclude about a population from a sample, and how precise such conclusions may be.

Errors in calculation

Silly errors are all too easy to make. The first step in avoiding it is to take care, but this is not enough. All calculations must be checked.

The best check is to give someone else the raw data and ask him to repeat the analysis. The second best is to repeat the analysis oneself using a different method – e.g. if you added the numbers down the column the first time, add them up the column the second time. Repeating the calculation in exactly the same way is not a good check, since one is likely to make the same error again.

A table for calculating mean and standard deviation

One way of reducing errors is to lay out ones calculations clearly and regularly. A good example of how a regular layout helps in calculation is the following method for calculating Σx and Σx^2.

The first two columns in Table 1 represent a simple frequency distribution of wing-lengths in a sample. To ease calculations, we can adjust the wing-lengths by subtracting 110 from each, giving the values in column 3. Let us refer to the numbers of birds (frequencies) as f and the adjusted wing-lengths as x. In column 4 we write the values of x^2. In columns 5 and 6 we write the values of fx (i.e. column 2 × column 3) and of fx^2 (i.e. column 2 × column 4). The sum of column 2 is the total number of birds: Σn in the usual terminology. The sum of column 5 is the sum of the x values weighted according to the number of birds with each value: x in the usual terminology. Similarly, the sum of column 6 is Σx^2.

We can now proceed as usual to calculate the mean and standard deviation from n, Σx, and Σx^2 in the usual way, not forgetting to add on the 110 mm to the calculated mean.

Table 1. Tabular layout of calculations for mean and standard deviation.

	Column number				
1	2	3	4	5	6
Wing length (mm) symbol	No. of birds f	Adjusted wing length x	x^2	fx	fx^2
112	1	2	4	2	4
113	0	3	9	0	0
114	2	4	16	8	32
115	5	5	25	25	125
116	4	6	36	24	144
117	2	7	49	14	98
118	1	8	64	8	64
Totals	**15**			**81**	**467**

$\bar{x} = \Sigma x/n = 81/15 = 5.40$; mean = 5.40 + 110 = 115.40 mm

$s^2 = \Sigma x^2 - (\Sigma x)^2/n - 1 = (467 - 81^2/15)/14 = 2.114$ mm

$s = \sqrt{2.114} = 1.45$

Samples and populations

When we study the characteristics of a sample of birds we are not really interested in the sample as such. We want to know something about the population from which it comes. We use the sample because we believe it to represent the whole population.

Clearly, this is only true if the sample is unbiased. Ornithologists are generally well aware of possible biases, so I shall not discuss the problem further. I shall assume that the samples under discussion are unbiased.

Estimating population characteristics

Suppose that the true mean wing-length in the whole population from which the birds in Table 1 were taken was 115.6 mm. We are not surprised that the mean of a sample of 15 is not identical with this: the vagaries of chance have caused "too many" small birds to be present in the sample and "too few" large ones. The sample mean can only be an estimate of the population mean. However, statisticians assure us that $\Sigma x/n$ is the best estimate of the population mean that can be obtained from a sample.

What about the variance and the standard deviation? It turns out that the best estimate of the population variance is $(x - \bar{x})^2/(n-1)$, which is why we have been using this formula or its equivalent: $(\Sigma x^2 - (\Sigma x)^2/n)/(n-1)$. Scarcely surprisingly, the best estimate of the population standard deviation is the square root of the variance estimate.

When introducing the variance, I suggested that one could think of it as an average value of the squared deviations from the mean. That being so, one might expect to divide the sum of squares, $\Sigma(x - \bar{x})^2$, by n rather than by n – 1. However, it turns out that if we do so we get a biased estimate of the population variance. Dividing by n – 1 gives us an unbiased estimate.

How precise are the estimates?

Assuming it to be unbiased, a large sample is more likely to provide a precise estimate of the population mean than a small sample. The precision of the estimate will also depend on how much variation there is between individuals for the character in question. If there is a little variation (i.e. the standard deviation is small), then even a small sample may provide a precise estimate of the mean. But if the population is highly variable, a large sample will be needed to give the same precision.

So a large sample from a population with a small standard deviation will give a precise estimate of the population mean. A small sample from a population with a large standard deviation will give an imprecise estimate. This is a rather vague conclusion. Fortunately, the precision of the estimate can be measured.

The measure of precision – or, rather, imprecision – is the "standard error of the mean". The larger it is, the more imprecise is the estimate of the mean. It is usually symbolised as s_x and can be calculated from the standard deviation and the sample size: $s_x = s/\sqrt{n}$

For the sample in Table 1: $s_x = 1.45/\sqrt{15} = 0.37$ mm.

Notice how this formula reflects what we already know about precision: when n is small or s is large, then the standard error is large.

Confidence limits: the idea

Though standard errors are useful to statisticians they do not convey much to the non-statisticians, beyond the general idea that a large standard error means that ones estimate is imprecise. it would be more useful to be able to say something like: "The best estimate of the population mean is 101 mm and it definitely lies between 92 mm and 102 mm".

Unfortunately, we can never be that definite. We can, however, say something similar: "The best estimate of the population mean is 101 mm and the chances are 95% that it lies between 95 mm and 107 mm". In this case, 95 mm and 107 mm are the "95% confidence limits of the mean".

It is important to remember that it is not definite that the true mean lies between the 95% confidence limits. There is a 5% chance (1 in 20) that it lies outside them. To put it another way, if one assumes that the true mean really does lie between the 95% confidence limits, one is wrong in one case in 20, on average.

Confidence limits: calculation

Confidence limits are easy to calculate, using the standard error and a statistic known as Student's t. The lower confidence limit is $_x - t.s_x$ and the upper one is $_x + t.s_x$.

The value of t is obtained from a table, of which Table 2 is a condensed version. (Fuller versions are to be found in any statistics book or set of statistical tables.) Of the various columns in such a table, we require the one for 95% confidence limits. The row we should use depends on the number of "degrees of freedom". For setting confidence limits to a mean this number is n – 1.

To make this explicit, let us turn again to the data of Table 1. The sample size is 15, so there are 14 degrees of freedom. The corresponding 95% value of t is 2.15. We have already calculated that the standard error of the mean for these data is 0.374 mm. Thus t.sx = 2.15 × 0.374 = 0.80 mm: the 95% confidence limits of the mean are 115.40 – 0.80 = 114.60 mm and 115.40 + 0.80 = 116.20 mm.

Other confidence limits

Table 2 contains a column headed 99%. Use of t values from this column gives 99% confidence limits: the chances are 99% that the true mean lies between these limits. They are, of course, wider than the 95% limits, because the level of confidence that the true mean lies between them is higher.

It is possible to set limits at any level of confidence (except 100%). By far the most commonly used level is 95%.

The precision of the standard deviation

Just as a standard error and confidence limits can be worked out for the estimate of the mean, so they can be worked out for the estimate of the standard deviation. The standard error of the standard deviation is usually symbolised as ss. For samples of 15 or more from normally distributed populations it is given by:

$s_s = 0.70711\ s_x$

In the example we have been using:

$s_s = 0.70711 \times 0.374 = 0.264$ mm.

Just as for the mean, confidence limits may be worked out for the standard deviation by multiplying this standard error by the value of the Student's *t* for n – 1 degrees of freedom. To stay with the same example: t.s = 2.15 × 0.264 = 0.57, so 95% confidence limits of the standard deviation are 0.88 mm and 2.02 mm (1.45 – 0.57 and 1.45 + 0.57).

Summary statistics

When presenting a summary of a set of data, what should we give? I believe that one should always give at least three pieces of information – mean, standard deviation, and sample size.

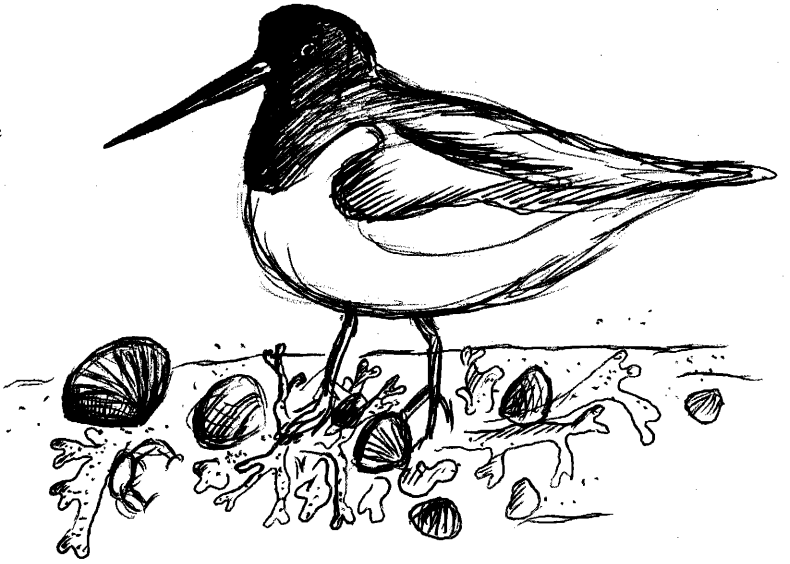

Table 2. Values of Student's t.

Percentage points for confidence limits		95%	99%
Percentage points for significance tests		5%	1%
Degrees of freedom:	1	12.70	63.70
	2	4.30	9.93
	3	3.18	5.84
	4	2.78	4.60
	5	2.57	4.03
	6	2.45	3.70
	8	2.31	3.36
	10	2.23	3.17
	12	2.18	3.06
	14	2.15	3.00
	16	2.12	2.92
	18	2.10	2.88
	20	2.09	2.85
	30	2.04	2.75
	100	1.98	2.62
	many	1.96	2.58

In some tables, the headings are in terms of probability values rather than percentages: 0.05 = 5%, 0.01 = 1%, etc.

The mean tells the reader the average value. The standard deviation tells him how variable the birds in the population are. The sample size allows him to work out how precise your estimates of the mean and standard deviation are likely to be and to carry out all the statistical calculations he is likely to want to perform on your data.

It is also helpful to the reader to provide him with a measure of precision by giving confidence limits (or standard errors) rather than leaving him to work them out for himself.

t table: an explanation

You will see that Table 2 is headed with two sets of "percentage points", those for confidence limits and those for significance tests. The latter are simply the complements of the former. Most published tables simply have the points for significance tests in their headings: for 95% confidence limits we need the column headed 5% in such tables. If in any doubt about which column to use, remember that the right one for 95% confidence limits is the one in which the values for the higher numbers of degrees of freedom are close to 2.

Introductory statistics 3

JEREMY J.D. GREENWOOD

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

Citation: Greenwood, J.J.D. 1979. Introductory statistics 1. *Wader Study Group Bull.* 26: 19–22.

Estimating the difference between two means

Consider the summary data of Table 3. It is fairly clear that the males in the population from which the sample was drawn are larger than the females on average: the sample means are 115.40 mm (males) and 112.14 mm (females). Since these are the best estimates of the population means, it seems reasonable to say that our best estimate of the mean difference between male and female winglengths in the population is 115.40 – 112.14 = 3.26 mm.

Thus we can say that, on the evidence available, males have wings 3.26 mm longer than females in this population.

The standard error of the difference

Just as we can measure how precise are our estimates of the means, we can measure how precise is the estimate of the difference. The difference also has a standard error associated with it. This may be calculated using a formula that looks horribly complicated but is easy to use. If n_1 and n_2 are the two sample sizes and s_1^2 and s_2^2 are the two estimated variances, then s.e. of difference:

$$s_{diff} = \sqrt{\left[\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}\right]\left[\frac{n_1 + n_2}{n_1 n_2}\right]}$$

For the data of Table 3:

$$s_{diff} = \sqrt{\left[\frac{(15 - 1)\ 1.45^2 + (21 - 1)\ 1.60^2}{15 + 21 - 2}\right]\left[\frac{15 + 21}{15 \times 21}\right]}$$

$$= 0.521 \text{ mm}$$

Confidence limits of the difference

Confidence limits can be calculated in the usual way, using Student's t with $n_1 + n_2 - 2$ degrees of freedom. For the data of Table 3, $n_1 + n_2 - 2 = 34$. The value of Student's t for 95% confidence limits and 34 degrees of freedom is 2.03. Thus $t.s_{diff} = 2.03 \times 0.521 = 1.06$ mm: the 95% confidence limits of the difference are 2.20 mm and 4.32 mm.

Thus we can say that the best estimate of the difference