INTRODUCTORY STATISTICS 5

by Jeremy J. D. Greenwood

Planning your research project

Some years ago, an eminent research director wrote "As a statistician, I still have to spend far too much of my time trying to find ways of making use of data for which the methods of collection were inefficient, if not invalid..... Papers are still being published in reputable journals where the methods of data collection and statistical analysis are deplorable if not impossible..... We still encounter the post-graduate degree student in despair, because he has spent two and a half years collecting his data, and he now finds that even his supervisor has very little idea about how these data can be analysed or presented, or even whether the method of data collection is relevant to the problem he is undertaking." The position is no better today. What is worse, it is scarcely any better among professionals than among amateurs. There are many reasons for this, few of which place those in positions of power in the biological establishment in a creditable light. Perhaps in practice the most important is that too often people carry out their research without any prior thought as to the way in which their data may be analysed.

Research projects should generally pass through a succession of stages:
1. Getting a preliminary idea on the problem to be tackled.
2. Formulating precisely the ornithological questions to be asked.
3. Gathering some preliminary data to get a rough idea of both the practical problems involved and the sort of data that are likely to emerge from a fuller investigation. This may allow a more precise approach to stage 2.
4. Careful planning of the investigations in the light of stages 2 and 3 and in the light of the statistical techniques available for extracting answers to the ornithological questions from the data one may expect to obtain.
5. Doing the field work.
6. Analysing the data according to the methods decided upon in stage 4, modified in the light of stage 5.
7. Communicating the results to others.

All too often, stages 2, 3 and 4 are rushed over, if not totally ignored. The result is the inefficient collection of data that turn out to be unanalysable, or even irrelevant.

Stage 4 is where any investigator needs the advice of colleagues. Most ornithologists would do well at this stage to discuss their project with at least two people - another ornithologist and a statistician. It is easy to find the first, but what about the second? It is important to choose the right statistician. Someone who simply has a degree in maths is no good: a theoretical statistician is no better. One wants someone who understands the practical problems of data gathering and analysis. Most university departments of statistics contain several such people, as do agricultural research stations; mathematics and biology departments of universities and technical colleges sometimes contain one or two. I have found that they are generally happy to advise the researcher, amateur or preofessional, provided:

(a) They are approached at the planning stage and not after the data have been gathered;
(b) The ornithological questions being asked have been clearly thought out and precisely formulated;
(c) Their advice is followed.

How much data should one collect?

As much as is necessary to estimate the parameter under investigation to the required degree of accuracy.

But how much is that? The answer, of course, depends on the parameter in question, the nature of the data from which it is being estimated, and the required degree of accuracy. It is always possible to estimate the required sample size in advance if one has a rough idea of the results one might expect. This is one of the areas where a statistician's advice is useful at stage 4 of the investigation.

I can illustrate the way in which one approaches this question through an example. Suppose one wished to estimate the mean weight of a population to a precision such that the confidence limits fell about 5% on each side of the mean. The first thing to do is to take a preliminary sample - the size of which depends on how much effort one can devote to preliminary work. Suppose one took a preliminary sample which yielded values of $\bar{x}$ = 110 gm and s = 15 gm. If 110 gm is a good estimate of the population mean weight, one requires that the confidence limits are about 5.5 gm (i.e. 5%) on each side of the mean. That is to say:

$$\text{required} \quad t.s_{\bar{x}} = 5.5 \text{ gm}$$

$$\therefore \text{required} \quad t.s/\sqrt{n} = 5.5 \text{ gm}$$

$$\therefore \text{required} \quad n = (t.s/5.5)^2$$

We have an estimate of s (15 gm). When n is greater than about 15, Student's t for 95% confidence limits is approximately 2. Thus:

$$\text{required} \quad n = (2 \times 15/5.5)^2 = 30$$

In this situation, one would plan to take slightly over 30 birds in ones main sample. This would achieve the required degree of accuracy. To take more would be a waste of time.

One might think that 30 was rather a small sample and that the results would not be very trustworthy. But the confidence limits are the measure of trustworthiness and if one only requires them to be within about 5% of the mean, then 30 would be an adequate sample. In practice, one would take rather more than 30, just to make reasonably sure of achieving ones aim of 5% precision, but there would be no reason to take many more.

It might turn out that to achieve the required degree of precision one needed to take a sample of impossibly large size. In this case, one must either decide to accept a lower level of precision or to abandon the project. There is no point in going ahead in the hope that a smaller sample than appears to be needed will give an estimate with the required degree of precision, just by good luck.

Estimation or significance testing?

In part 4 of this series I explained the limitations of significance tests. Perhaps three points may usefully be reiterated here. First, significance tests for which the null hypothesis is unreasonable are silly. Second, the effort involved in estimating a parameter and its confidence limits is usually no greater than the effort involved in carrying out the relevant significance test. Thirdly, estimation provides much more information than a significance test, especially a test that leads to the null hypothesis being accepted. Significance tests are used far more often than they should be: you are advised to set a new fashion.

Picking out interesting differences

It is not uncommon for people to gather a set of data, inspect it for interesting patterns or peculiarities, and then apply some statistical technique to determine the "significance" of what has been picked out. For example, one might study the frequency of some behaviour in ten different species and note that it was conspicuously higher in species number 9. There is a temptation to carry out tests to compare the frequency in that species with the frequencies in each of the others in turn. In doing so, one would obtain "significant" results in many cases and one might conclude that species number 9 was aberrant. But one would obtain similar results if one simply took ten samples of the same species and "tested" the sample with the highest frequency against each of the others. This is because picking out a species by examining the data and then "testing" it against the others violates the principles on which the standard tests are based. An even worse error is to compare the species that, on inspection of the data, seems aberrant with all the others combined.

Of course, if one has some advance reason for expecting one of the species in particular to show unusually high frequencies of the behaviour in question, it is quite valid to compare that species individually with the others. It is when the reason for making the comparison springs from the data themselves that the comparison is invalid.

If one has studied 10 different species and has no advance reason for expecting one in particular to be peculiar, one may make a general comparison by carrying out an analysis of variance. If there is significant variation between species, then one may go ahead and use special techniques for picking out which (if any) of the species is most responsible for the variation. But such tests are tricky and professional advice is essential.

Use appropriate methods

It is important, of course, to use statistical methods appropriate to the question in hand. In this series I have mentioned only certain areas of elementary statistics. Do not be tempted to apply the methods blindly to any problem to which you think they can be bent. If in any doubt, seek advice.

Associations

I have not considered, for example, the question of examining associations between variables. When one is interested in an association between two metrical characteristics - e.g. the relationship between weight and wing-length in a population - then correlation analysis or regression analysis are generally applied. There is a good deal of confusion amongst biologists as to which is the appropriate analysis to use under particular circumstances. The matter is too complex to go into here, except to say that in most cases where regression has been applied in ornithology it would have been more appropriate to use correlation.

There are two dangerous traps into which the unwary may fall when looking at association. The first may be illustrated by example. Suppose that one censused the population of lapwings nesting in a number of fields in two successive years. One might wish to see if the population density in the first year affected the population change between years. To do so, one would perhaps work out the population change for each field and examine its correlation with numbers in the first year. This sounds reasonable. But think it out: if N_1 and N_2 are the numbers in the first and second years, then one is looking at the correlation between $(N_2 - N_1)$ and N_1. Because N_1 is a component of both variables, one is likely to get a "significant" correlation (negative, in this case) even if there is no biological relationship at all. When examining associations involving derived variables (the difference, in this case) it is always important to ensure that none of the primary variables (N_1 and N_2, in this case) appear twice.

Finally, the demonstration of a significant correlation does not indicate which of the variables caused the other - or even that they are causally related to each other. The size of the world population over the last 50 years is strongly correlated with the number of members in the RSPB but any causal connection is surely remote! In less absurd cases it is easy to slip into the trap of assuming that correlation proves causation. Beware!

Finally

Never forget the ornithology involved in one's statistical analysis. If an analysis produces a result which is absurd in the light of one's ornithological knowledge, then one has probably made an error of calculation or used the wrong technique.

Dr. J.J.D. Greenwood, Department of Biological Sciences, The University, Dundee, Scotland.