MATHEMATICS AND THE STUDY OF LIFE

R.H. Crozier*

The young R.A. Fisher is said to have visited a local museum and come across a labelled skeleton of a fish. Gazing at the bones of the head (there are many), Fisher decided that biology was far too hard, and he henceforth retreated into the simpler world of mathematics. His subsequent efforts brought him much renown, and honours as disparate as a knighthood and emeritus status at the University of Adelaide (where he died). Despite his early experiences with fish heads, Sir Ronald yet had a great influence on biology, and not only through his great influence on statistics, but also through the various mathematical models he made which secured his position as one of the three patron saints of population genetics.

Most biologists look at the relationship between mathematics and biology somewhat differently from Fisher, perhaps partly because of his efforts! However, mathematics has an ever-increasing importance in biology, with applications that can be broadly grouped into two overlapping types (as in all science): description and hypothesis testing (statistics), and predictive or explanatory models.

Statistics: description

One of the unhappy certainties of life is that you can never be *certain* of anything. Descriptive statistics help us place limits on such uncertainty. Thus, a biologist may measure a sample of fish in order to determine the average weight of the population from which the sample was drawn. However, even neglecting measurement error, the average weight of the sample will *not* be the same as the average weight of the population. However, we can estimate the most probable value for the mean weight, and limits within which the true mean probably lies. These confidence limits themselves are a probability statement, with biologists usually favouring 95% limits (that is, a range with a probability of 0.95 of containing the true mean).

Statistics: hypothesis testing

Science is not a body of facts but a body of hypotheses. We add hypotheses to this body depending on how well they agree with data collected to test them. This testing is, as with statistical description, a matter of quantifying uncertainty. We can never *prove* a hypothesis (nor disprove it, but let us ignore this point as being *too* downbeat), but we can make statements about how well the data accord with it; or, in other words, about how plausible the hypothesis is.

* *Dr. Ross Crozier is a Senior Lecturer in the School of Zoology at the University of New South Wales.*

A very rough, and perhaps idiosyncratic, way of looking at this use of statistics is to divide it into: the detection of non-obvious trends in large data sets, and the testing of the likelihood that obvious trends in small data sets are in fact spurious.

An example of the examination of large data sets is the sifting of health statistics to uncover correlations between, say, the occurrence of different cancers with various factors such as air pollution, smoking, various foods, and so forth. Such sifting has to be done carefully, as the effects of such factors as race and class have to be removed. And finally, of course, we are still left with a probability statement, such as that there is only a 1% chance that the observed association could arise by accident alone, and so we conclude that, say, smoking and lung cancer are in fact probably associated.

If we flipped a coin four times and found the same side landing uppermost every time, we would have a striking trend and might suspect rigging and want our money back from the School. However, the probability that a true coin would do this is quite high (1/8), so that here we have an apparent departure from randomness that is in fact spurious.

Not all such departures are spurious. Examine the following data of chromosome numbers of ants collected in logs or under stones on a certain mountain-top (modesty forbids me to name the collector):

	in logs	under stones
18 chromosomes	6	0
22 chromosomes	0	7

Of course these data pass the Intra-Ocular Traumatic Test for interest (they reach up and hit you between the eyes), but furthermore the probability that they could arise by random assortment of ants and habitats is only 0.0012 (by Fisher's Exact test — yes, that name again). We can thus conclude that there is (almost) certainly an association between chromosome number and habitat for these ants at this site.

Models

A model is a mathematical interpretation of how a system operates. From it, we can (often) derive predictions that can then be tested against data to see how well we really understand the system of interest. A simple example would be an engineering-type model of limb function: given the strength of the various muscles and the lengths of the various bones, how high should the animal be able to jump? The resulting predictions can be readily tested by experiment.

However, the true complexity of life becomes apparent when we try to take account of random processes. I will now (briefly) outline two such *stochastic* models and their effects on biology. We shall consider two competing models about the forces acting on genetic variability. The "selectionist" model holds that most, perhaps all, of the observed genetic variability in natural populations is maintained by (natural) selection, whereas under the "neutralist" model the observed levels of variation simply reflect a balance between the origin of new genes through mutation and their ultimate loss or total success through random genetic drift (change in frequency resulting from each generation being an imperfect copy of its parents because of sampling error).

If we consider that the kind of selection most likely to be maintaining genetic variability is selection in favour of heterozygotes (individuals possessing two different genes for the same trait), as happens for sickle-cell anaemia in man, then the neutralists have a strong case. Let A stand for one gene, and a for the other, and let $p = q = 0.5$ be their frequencies. Further, suppose the homozygotes have only a 1% loss of vigour or chance of survival compared to the heterozygote. We thus have:

	genotypes		
	AA	Aa	aa
fitness	.99	1	.99
frequency	p^2	2pq	q^2

The average fitness of the whole population at the zygote stage is .995. This is all well and good, but what if there are 2,000 such genetic loci? In that case, the surviving fraction becomes not .995 but $(.995)^{2000}$, and to leave an average of two progeny each female would have to lay an average of 45,000 eggs! It's just not on.

You would therefore expect organisms to be rather uniform genetically, but it is now known from enzyme screenings that they are often in fact highly variable. Some flies have been found to have about half of their loci polymorphic; taking a conservative estimate of 5,000 genetic loci, this suggests 2,500 polymorphic ones! Given our calculations above, a simple role for selection in maintaining this wealth of variation seems unlikely.

However, the selectionists, spurred on by the traditional evolutionists' belief that just about every feature of organisms is adaptive, have been unconvinced by such figurative finesse and have sought to demonstrate that, in fact, selection *is* acting to maintain most if not all observed protein polymorphisms.

Such testing has not been easy. Anyone *can* easily show apparent influences of selection on any enzyme polymorphism, but unfortunately genetic loci occur in groups — on chromosomes. Thus, the effects apparent in studying one locus may simply reflect changes occurring at a nearby, linked, locus. Many have tried to circumvent this problem in laboratory tests by using substrates specific for the locus of interest, but usually with equivocal results. Attempts to solve the dispute by testing the form of gene frequency distributions, for example, have also bogged down as each side casts nets of elegant equations to secure the elusive quarry of truth.

And yet, the balance has shifted towards the traditionalists, the selectionists. This has largely resulted from two factors: if it is so easy to demonstrate the influence of selection, then, even if you don't *know* if it is acting on the locus under study each time, there must be a great many loci which *are* influenced by selection, and so the original 45,000-egg argument must be wrong. Secondly, coupled with this, selectionists have devised models for the action of selection that don't require so many deaths to maintain much polymorphism. However, the issue is still open — for one thing, these models make further predictions, for example about the likely effects of inbreeding, that have not been matched very well by the data. There is probably still time for a reader to enter the lists and solve this problem — the argument has been going for some 12 years!

I thank my colleagues M. Archer, P. Dixon, A. Mazanov and A. Woods for comments on this manuscript.