Chapter One

Planning and preliminary considerations

If you do not carefully plan ecological fieldwork your effort will be wasted. However, some unfocused preliminary work can be useful to help to form your ideas. It is impossible to plan a study without background information on the natural history of the organism and the nature of the habitat. While much can be gained from reading, undertaking computer searches, and talking to others with experience, it is often essential to do some preliminary sampling and try out alternative techniques to assess their relative merits. During a pilot study a record should be kept of the cost of each part of the sampling routine, normally expressed in man-hours. If you have a clear appreciation of the cost, you will not attempt a study that is beyond your resources. Having completed preliminary studies, the objectives of your study should be clearly stated and a sampling program designed. Your objectives should be written down and you should be able to describe them verbally in clear and simple terms. If you are unable to do this, you probably have confused and muddled objectives and you certainly will not be able to convey your ideas to others.

Armed with the information obtained during the preliminary investigations and a clear objective, the full sampling plan can be created. It is important to remember that ecological studies rarely go according to plan as the natural world is variable and the weather will frequently disrupt sampling. Further, all sampling methods have a certain range of population density over which they are appropriate and may become unsuitable if the population should greatly decline or rise. Therefore build contingencies into the plan. While some studies focus on the ecology of a single species or even race (autecology), others investigate whole communities or some part of them (synecology). Both types of study require the collection of data about populations and require consideration of the same general principles described below.

The need for sampling

As it is rarely possible to count or collect all the individuals in a population, the size of the population and its attributes such as mean size and age must be estimated by sampling. Much of the planning of an ecological study is concerned with ensuring that the samples accurately reflect the population or community as a whole.

The scale of the study

The size of the area to be studied needs to be determined. This area must be of sufficient size to adequately reflect the true nature of the target population or community. However, if it is too large you will waste resources and may not be able to complete the study.

Whether the area of study should be a single habitat (e.g. field, pond, woodland, or particular rock pool) or selected representatives of the habitat type from a wider geographic area will depend on whether an intensive or an extensive study is planned. Extensive studies have a low intensity of sampling per unit area or through time. They are frequently used to provide information on distribution and abundance for conservation or management programs. Intensive studies involve the repeated observation of the population of an organism with the intention of producing accurate estimates of population parameters. An intensive study would be needed, for example, to produce a life-table (see page 108).

Safety

During the planning stage you should identify potential hazards and plan how to work safely. You will find it useful to refer to Nichols (1983). When working on or close to water, or in remote localities, it is important not to work alone. Pay particular attention to risks associated with adverse weather and exposure, becoming lost, and the effects of heat and excess sun. You may also need to take preservatives such as formalin or other chemicals into the field. If so, consider carefully how they will be transported and how harmful exposure such as splashing formalin into your eyes will be avoided and dealt with. If you plan to handle wild animals ensure you have the correct protective equipment to avoid bites and scratches. Finally, get acquainted with the features and risks of the study area. Are there biting insects that will make concentration and careful measurement at certain times of day impossible? Are there parasitic mites in the vegetation just waiting for a soft-skinned mammal to pass? If these aggravations and dangers are recognized you will be able to dress defensively, use repellents and work at the best times of day so that field work is a pleasure rather than an ordeal.

Care for the environment

As your plans are developed always consider the potential harm caused to the habitat and its inhabitants by your study and minimize general disturbance.

Keep the number of organisms handled, removed and killed to a minimum, and take care not to trample on plants that are easily damaged. If animals are to be returned, ensure that they are released close to their point of capture and in a manner that will give them a good chance of survival. Consider carefully if any of the proposed methods are cruel and likely to cause unnecessary distress. In particular, if animals are to be killed, plan carefully how this will be done. Finally, ensure that you know the local regulations concerning the protection of habitats and endangered species and always obtain the consent of licensing authorities, landowners, etc.

Taxonomy

Are there any taxonomic difficulties? It is essential that you ensure that you can identify the target species and maintain a consistent taxonomy. As a study progresses taxonomic ability frequently improves. It may be necessary to revisit early samples to reassess species identifications. When possible, you should retain samples. Studies have failed when it became apparent that two or more species had been confused during the early period of sampling. Problems can also occur if only some life stages can be identified.

In community studies, you may need to identify or discriminate between large numbers of forms. Will this be feasible? The appropriate degree of taxonomic discrimination must be decided upon. It may not be essential to identify to the species level. For example, a study of stream invertebrates for the effects of pollution can be based on the presence of families and the calculation of a quality index such as the BMWP score (see Chapter 10). In community studies it is common to identify different groups to different taxonomic levels. For example, in freshwater studies stoneflies and dragonflies may be identified to species level while the larvae of small flies are recorded at the family level. Such mixed taxonomic resolution is often appropriate and the only practical option. However, you may need to consider what effect this will have on the analysis of species richness or diversity.

Sample sorting and species identification are often the most labor-intensive parts of a study and it may be useful to process a trial sample during a pilot study to assess the effort required.

Recording, labeling, and note taking

It is essential that you keep good records of your observations while in the field. Observations may be entered onto a tape recorder, palm-sized computer, digital camera, or onto paper. The durability of pencil on paper is second to none and unlike all the other devices listed above, it will even survive total immersion in water and will not require batteries. Always use notebooks of good quality paper that are designed to withstand water. If you use any form of electronic device always extract the data immediately upon returning from the field, ensure the data is not corrupted, and make backups and hard copy as appropriate. As a backup, notebooks should be photocopied, scanned or digitally photographed. To protect valuable data it may be useful to transmit it by e-mail to another locality. Data protection is not the only reason why information should be transcribed as soon as possible; field notes are often rather poorly written and will be better understood while events are still fresh in your memory. Never return to the field without having secured your data from the previous trip.

When taking notes remember to date the page, record who is present, and other general information such as the weather. When using photography, you will need to record the details of each photograph. Digital cameras are invaluable aids for ecologists. They allow you to instantly check that a good photograph has been taken and frequently allow details of the image to be entered as text or even a voice record.

Samples must be securely labeled. Usually you should place a pencil-and-paper label inside the sample and label the outside with a permanent marker. Paper of the correct quality must be used, as some papers will disintegrate when wet. Remember that a standard permanent marker is water insoluble, but the label might be lost if the sample is preserved in alcohol which leaks out. When labeling, use a numbering system that will not become ambiguous if part of the label is lost. For example, avoid roman numerals such as ii and iii. It is frequently advisable to check the labeling (and preservative) soon after returning from the field.

Data security and processing

A sampling plan needs to consider the processes involved in data acquisition, organization, analysis, and presentation. Smooth and rapid progress along this chain is aided by the use of computers, but only if the data can be easily transferred between software. If different software products are used, then compatibility must be considered. Both software and hardware capability need to be considered for each stage of the study. During data acquisition, it is important to assess the data storage and processing requirements. Automatic data collection devices, such as digital temperature recorders, can collect prodigious amounts of data if set to record at short intervals. Remember that data has only been acquired when it has been processed into a usable form. Data processing and input rates are often different. Data that takes many weeks to enter into a computer is often analyzed and plotted in a few seconds. Portable computers and palm-sized input devices allow field observations to be immediately stored in digital form and this may offer a way of streamlining the data-input process. Great care needs to be taken to ensure that data is not subsequently lost.

Ecological data is frequently arranged in spreadsheet software such as EXCEL. These programs are particularly appropriate for data that are naturally arranged in a grid, such as the species recorded from a number of samples. In many cases, appropriate statistical tests and plots can be undertaken within the spreadsheet program. If the data will need to be exported to a more specialized program for statistical analysis, you should ensure during the planning stage that the spreadsheet program is able to export the data in a suitable form.

Effect of the time of year on sampling

You must ensure that your proposed study will be undertaken at a suitable time of year. It is often important to sample when numbers are high. For plants you may need to work when they are in flower so that they can be easily identified. For seasonal insects such as butterflies the adults may only be abundant for a relatively short period of time each year. For annual species the life stage or age group that is chosen for study will determine the sampling period. The life stage of choice may be chosen because it is particularly easy to sample or possibly because it has a particularly important ecological impact. For insect pests, the best stage for sampling may be that most closely correlated with the amount of damage. For aquatic organisms, timing is often determined by the reproductive cycle. For example, in temperate waters, benthic surveys (the study of organisms living on or in the seabed) carried out in autumn will show a population dominated by recent recruits. The same survey carried out in the spring or early summer will show the resident community that has survived both competition for space and the rigors of the winter. While the life history stage to be sampled will depend on the objectives, the stage must not be one whose numbers change greatly with time and it must be present for a period of sufficient extent to allow the survey to be completed. Finally, the easier the stage is to sample and count the better.

Extensive surveys can result in different areas holding populations that differ in their development. The timing of blooms and larval production in marine plankton, for example, can vary by 1 month over 1 degree of latitude. The timing of the flowering of plants also shows clear latitudinal and altitudinal gradients. These differences may be used to advantage by allowing the different habitats to be sampled in succession.

Effect of the time of day on sampling

When observing birds and mammals the time of day when observations are taken is clearly important. However, it is not always realized that this can also be the case for invertebrates. The diurnal rhythms of insects may cause them to move from one part of the habitat to another. Many grassland insects move up and down the vegetation at certain times of the day or night and aquatic insects may emerge at a particular time. During the day quite a proportion of active insects may be airborne. Similarly, plankton also show diurnal rhythms in their vertical distribution and may concentrate in surface waters at particular times of day or night. Fish that remain hidden by day become active and vulnerable to trapping at night. There is a marked periodicity of host-seeking behavior in many blood-sucking invertebrates, a fact with which many of us are familiar, as just after sunset is often the worst time for mosquito attack. It may be that sampling problems can be overcome, or additional information gained, if work is undertaken at night, dusk, or dawn, rather than during conventional working hours.

Types of population estimate

If populations are to be estimated you will need to decide in what form the population size will be expressed. Estimates of population size can be expressed as absolute number, population intensity, relative number, or an index. This list of possible measures is ordered from the most difficult to the easiest to achieve. The objectives of the study will determine if a simple index of abundance will suffice or if greater effort must be made.

Absolute population is defined as the number of organisms per unit area or volume. It is almost impossible to construct a budget or to study mortality factors without the conversion of population estimates to absolute figures. While some organisms and habitats allow absolute population estimates to be obtained at reasonable cost, others, such as insects living amongst boulders in fast-flowing streams, are impossible to estimate.

Population intensity is the number of organisms per unit of habitat, e.g. per leaf, per plant, or per host. When the level of the population is being related to habitat availability or plant or host damage, it is more meaningful than an absolute estimate. It is also valuable when comparing the densities of natural enemies and their prey. However, the number of habitat units per unit area is potentially variable and may need to be assessed. Remember that a high intensity population may reflect either high absolute numbers or a shortage of hosts or habitats.

Relative estimates give an abundance estimate that cannot be related to a unit area or habitat and will only allow comparisons between similarly collected samples. Such comparisons can be either between different localities or at one locality through time. For many organisms they are the only estimates that can be achieved at a realistic cost. They are especially useful in extensive work on species distributions, monitoring changes in species richness, environmental assessments, and recording patterns of animal activity. The methods employed are either the catch per unit effort type or various forms of trapping, in which the number of individuals caught depends on a number of factors besides population density. There is no hard and fast line between relative and absolute methods, for absolute methods of sampling are seldom 100% efficient and relative methods can sometimes be corrected in various ways to give density estimates. Relative methods are important in applied research such as fisheries or game management where most of the available information may be derived from fishing or hunting returns.

Population indices are produced when the organisms themselves are not counted, but their products (e.g. frass, webs, exuviae, tubes, nests, pollen, pellets, fur) or effects (plant damage, footprints) are recorded. Such methods are frequently used for the study of rare or shy mammals such as otters or cats that may be infrequently observed.

Defining the habitat unit

Organisms only use some of the available space. Insects, for example, may only be found in association with their food plant and plants frequently require a particular soil and aspect. It is therefore essential to identify where the study organisms are to be found and the unit of habitat that should be sampled. The habitat unit could, for example, be the fleece of a sheep, a bag of grain, a rock in a stream, a single tree, or areas of grassland or muddy seabed. In a study of lichens it might even be individual gravestones. The identification of the habitat unit will often require both a literature review and some preliminary nonquantitative sampling.

If the sampling unit is a habitat feature such as a shoot, tree, or stone, it should meet the following criteria:

1 All habitat units in the study area must have an equal chance of selection for sampling.

2 The number and size of the habitat units should not change over the course of the study or any changes should be easily measured. For example, if the sampling unit is a young shoot you should take care that the number of shoots remains constant or any changes in number can be counted.

3 The proportion of the population using the sample unit as a habitat must remain constant. Many species move between different habitats and thus the proportion of the population in a particular habitat can change through time.

4 The sampling unit must lend itself to conversion to unit areas.

5 The sampling unit must be easily delineated in the field.

6 The sampling unit should be of such a size as to provide a reasonable balance between the variance in the number of organisms present and the cost.

7 The sampling unit must not be too small in relation to the organism's size, as this will increase edge-effect errors.

8 The sampling unit for mobile animals should approximate to the average ambit of an individual.

Quadrat sampling

When quadrat or unit area sampling, a square sampling unit is usually used. While other shapes such as a circle may reduce bias from edge effects, if the total habitat is to be divided into numbered sampling units (for random number selection), then circular units are impractical because of the gaps. Clearly, the larger the sampling unit is in relation to the size of the organism, the proportionally smaller the boundary edge effect. Anticipate edge effects by using a convention, e.g. only individuals crossing the top and left-hand boundaries are counted.

As a general principle, a higher level of reproducibility is obtained (for the same cost) by taking many small units rather than a few large ones. The main disadvantage of sampling by small units is the number of zeros that may result at low densities; this may produce unwanted restrictions on your analysis, e.g. you cannot take the logarithm of zero. Small sampling units may also enable precision to be increased by distinguishing between favorable and unfavorable microhabitats.

Having defined the habitat unit for study it is also essential to obtain information on its spatial extent during any pilot studies. For example, a grab lowered from a boat collects an unseen unit of seabed. If the objective is to study the benthic fauna of sand substrates, then the distribution of these substrates needs to be known before the sampling plan can be formulated. Similarly, a study of the insects of oak trees can only be planned if you have an appreciation of the distribution of the trees. Different sampling methods may be needed for preliminary and final sampling to ensure that the habitat distribution is understood. When benthic sampling, a pilot study might use dredges to obtain a general idea of animal presence and substrate distribution followed by grab sampling to estimate density in the selected habitat.

Subdividing the habitat unit

Animals and plants do not use their entire habitat equally, so during preliminary studies you should consider if the habitat units should be subdivided for sampling. The habitat must be considered in relation to the needs of the organisms under study. In a study of woodland birds, for example, the various levels of the trees – upper, middle and lower canopy – would be considered as potentially different divisions. In herbage or grassland, if leaves or another small sampling unit is being taken, the upper and lower parts of the plants might be treated separately. Subdivisions may be orientated with respect to an environmental gradient. If none is apparent, it is often convenient to divide the area into regions of regular shape, but this may be neither possible nor desirable. Aspect can be important. In the northern hemisphere north-facing rocks in the intertidal zone offer the coolest, most shady habitat available, and thus tend to favor red algae and its associated fauna. Conversely, species towards the northern limit of their range may prefer south-facing slopes. In aquatic habitats, depth, substrate, orientation with respect to flow and the degree of exposure are important. When sampling plants for insects, the amount of subdivision that is necessary varies greatly between species. While some insects may be distributed randomly over the foliage, others may be aggregated at a certain level and this may change with the seasons.

A hierarchical design is one in which, for example, a number of plants are sampled from each of a number of plots from within a number of fields. If a certain number of samples is collected randomly within each level of subdivision, this is often termed **nested random sampling** and is analyzed by a **nested analysis of variance** (see Underwood (1997), Zar (1996) or Sokal and Rohlf (1995) for an introduction to these methods in biology). There may be two or three hierarchical levels, rarely more.

Statistical considerations

Underwood (1997) gives a detailed description of the use of analysis of variance in ecological experimentation, and standard statistical textbooks such as those by Zar (1996) and Sokal and Rohlf (1995) are invaluable.

The number of samples

Never undertake an ecological study without considering how many samples will be required to meet your objective. When estimating mean abundance, the sample number cannot be accurately determined, but a working approximation can be obtained using an estimate or reasonable guess of the mean and variance of the population. These estimates may be derived from preliminary sampling or from published studies. Within a homogeneous habitat, in which the organisms under study are normally distributed, the number of samples (n) required is approximated by:

$$n = \frac{t^2 s^2}{(D\bar{x})^2} \quad (1.1)$$

where $\bar{x}$ = the mean number of organisms per sample, s^2 = sample variance, D = the proportional precision of the mean (i.e. to obtain an estimate within ±5% of the true value of the mean, then $D = 0.05$), and t = the "Student's t" obtained from standard statistical tables, but it is approximately 2 for $n > 10$ at the 5% level.

As an example, consider a study to estimate the density of small bivalve molluscs on a beach. During a preliminary visit 10 core samples were taken, giving a mean and variance of 5 and 10 animals per core. If it is desired to estimate the mean to a precision of ±10% then the number of core samples required is estimated as:

$$n = 2^2 \times 10/(0.1 \times 5)^2 = 40/0.25 = 160$$

Note that in the above calculation t is approximated as 2, which is quite accurate enough for such an estimate.

The distributions of many organisms are clumped and the assumption that the animals are normally distributed may be inappropriate. Southwood and Henderson (2000) have described methods of estimation when this is the case. However, provided you only use the estimated sample number as a guide, the assumption of normality will not produce a highly misleading estimate.

Another type of sampling program concerns the measurement of the frequency of occurrence of a particular organism or event, for example the frequency of occurrence of galls on a leaf. If it was found in a preliminary survey that 25% of the leaves of oak trees bear galls, the probability is 0.25. The number of samples (N) is given by:

$$N = \frac{t^2 p(1-p)}{D^2} \tag{1.2}$$

where p = the probability of occurrence (i.e. 0.25 in the above example), and D = the proportional precision of the mean (i.e. to obtain an estimate within ±5% of the true value of the mean, then $D = 0.05$).

Worked examples of the calculation of sample number for a variety of sampling programs are given by Greenwood (1996).

The number of samples per habitat unit (e.g. tree, rock, field, or puddle)

There are two aspects to be considered, firstly whether and exactly how the habitat unit should be subdivided and secondly the number of samples within each unit that should be taken for maximum efficiency. If it is decided that the habitat should be subdivided you should be aware that if the distribution of the population throughout the habitat is biased towards certain subdivisions, but the samples are taken randomly, **systematic errors** arise. This can be overcome either by sampling so that the differential number of samples from each subdivision reproduces in the samples the gradient in the habitat, or by regarding each part separately and correcting at the end.

Ideally, the within-subdivision sample variability should be much less than that between subdivisions. To determine the optimum number of samples per habitat unit (e.g. oak tree) (n), the variance of within-unit samples (s_s^2) must be compared with the variance of the between-unit samples (s_p^2) and set against the cost of sampling within the same unit (c_s) or of moving to another unit and sampling within it (c_p):

$$n_s = \sqrt{\frac{s_s^2}{s_p^2} \times \frac{c_p}{c_s}} \tag{1.3}$$

If the interunit variance, s_p^2 is the major source of variance and unless the cost of moving from unit to unit (e.g. oak tree to oak tree) is very high, n will be in the order of 1 or less (which means 1 in practice).

Often a considerable saving in cost without loss of accuracy in the estimation of the population, but with loss of information on the sampling error, may be obtained by taking randomly a number of subsamples that are bulked before sorting and counting. This is especially true where the extraction process is complex, as with soil samples.

The pattern of sampling

Again, it is important to consider the object of the program carefully. If the aim is to obtain estimates of the mean density, then it is desirable to minimize variance. However, if the distribution pattern is of prime interest then there is no virtue in a small variance. In order to obtain an unbiased estimate of the population, the sampling data should be collected at **random**, that is so that every habitat unit in the universe has an equal chance of selection. In the simplest form – the **unrestricted random sample** – the samples are selected by the use of random numbers from the whole area (universe) being studied (random number tables are in many statistical works, may be generated using a computer program or, failing these sources, the last two figures in the columns of numbers in most telephone books provide a substitute). The position of the sample site is based on two random numbers giving the distances along two coordinates; the point of intersection is taken as the center or a specified corner of the sample. If the size of the sample is large compared with the total area then the area should be divided into plots that will be numbered and selected using a single random number. Such a method eliminates any personal choice, as human bias in selecting sampling sites may lead to large errors.

However, a random choice method is not an efficient way to minimize the variance, since the majority of the samples may turn out to come from one area of the field. The method of **stratified random sampling** is therefore to be preferred for most ecological work; here the area is divided up into a number of equal sized subdivisions or strata and one sample is randomly selected from each strata. Alternatively, if the strata are unequal in size, the number of units taken in each part is proportional to the size of the part; this is referred to as self-weighting. Such an approach maximizes the accuracy of the estimate of the population.

When it is apparent that there are systematic variations in the density of the study organism across the study area, **stratified sampling** should be used. In general the strata are chosen to minimize within-stratum variance. Individual strata need not form continuous patches within the study area. When the habitat is stratified, biological knowledge can often be used to eliminate strata in which few animals would be found. Such a restricted universe will give a greater level of precision for the calculation of a mean than an unrestricted and completely random sample with a wide variance.

The other approach is the **systematic sample**, taken at a fixed interval in space (or time). In general, such spatial data cannot be analyzed statistically; however, it is often the method of choice in studies on the distribution of species or communities.

Biologists often use methods for site selection for random sampling that are less precise than the use of random numbers, such as throwing a stick or quadrat and sampling where it lands. Such methods are not strictly random; their most serious objection is that they allow the intrusion of a personal bias; quite frequently marginal areas tend to be undersampled as we are not inclined to throw towards a boundary. Bias may intrude due to causes other than personal selection. For example, grains of wheat that contain the older larvae or pupae of the grain weevil, *Sitophilus granarius*, are lighter than uninfested grains and may rise to the surface where samples are normally gathered. Similarly, the distribution and behavior of parasitized or sick animals may be such that they are more vulnerable to capture or observation.

Accuracy, precision, and completeness

Accuracy measures how close an estimate is to the real value and **precision** measures the reproducibility of the estimate. A decision on the accuracy and precision of population estimates is taken by considering both the objectives of the study and the variability of the system under study. For example, many species of insect pest exhibit 10- or even 100-fold population change within a single season and recruitment in fish stocks can vary 10-fold between years. For such species an estimate of population density with a standard error of about 25% of the mean, enabling the detection of a doubling or halving of the population, is sufficiently accurate for stock assessment studies. For life-table studies, a higher level of accuracy, frequently set at 10%, will be necessary.

Conservationists often seek to build species inventories. Here completeness will replace accuracy as a measure of quality. While for large mammals or birds the aim might be to record all resident species, for high diversity groups such as beetles the objective may be set at only 5–10% of the total fauna. Community studies often aim to generate summarizing statistics such as measures of diversity or species richness that can be used to compare localities or changes through time. The accuracy and precision of these estimates must be carefully considered if changes are to be detected at the desired resolution. It is often worthwhile to undertake numerical simulations using a computer as these can give an indication of the effort required to achieve a set level of accuracy. Such studies often demonstrate that the available resources will not allow the desired accuracy to be achieved. You will then need to decide if a lower accuracy is acceptable or the study objectives should be modified.

However, there is always much to be gained from field observations so never allow pessimistic contemplation of statistical niceties and accuracy to inhibit data collection. Excessive concerns about accuracy can always be used by those who favor the warmth of the hearth to being in the field. It should always be borne in mind that the law of diminishing returns applies to the reduction of sampling errors. In the long run, more knowledge of the ecology of an animal or plant may be gained by studying other areas, by making other estimates or by taking further samples than by straining for a very high level of accuracy in each operation. The one activity that will reward great care and effort is the sorting and picking of organisms from samples – remember that it is inevitable that the actual species list and individual abundances will lie above the recorded values, as some will almost inevitably be missed. You must always ensure that this task is given sufficient time.

Errors and confidence limits

The statistical errors of estimated parameters can usually be calculated and are referred to as the fiducial limits (the estimate (x) being expressed as $x \pm y$ where $\pm y$ is the fiducial limits). These are sometimes incorrectly referred to as "confidence" limits, but the distinction between the two terms is in practice unimportant. The fiducial limits are calculated for a given probability level, normally 0.05, which means that there are only 5 chances in 100 that the range given by the fiducial limits does not include the true value If more samples are taken the limits will be narrower, but the estimate may not move closer to the actual value. Biologists are often concerned that some of the assumptions about sampling efficiency may be incorrect. It is sound practice to compare estimates obtained using methods that make **different** assumptions. If the estimates are of the same order of magnitude, then you can have much greater confidence that the result is not misleading. Laughlin (1976) has suggested that ecologists may be satisfied with a higher probability level (say 0.2) and thus narrower fiducial limits for estimates based on more than one independent method.

The objective of a study is often expressed in terms of a hypothesis capable of scientific analysis. When a statistical analysis is anticipated you will need to create a null hypothesis for testing. For example, in a study of anthill distribution the null hypothesis might be that the mean anthill density is the same in open areas and in scrubland areas. A type I error is said to occur when we reject this null hypothesis when it is true. For our example we will have concluded that the density of anthills was significantly different when it was not. The opposite situation when we accept a null hypothesis when it is wrong is called a type II error.

The normal distribution and transformations

The most important and commonly used of the theoretical distributions is the normal or Gaussian distribution. This has a probability curve that is a symmetric bell-shape. The probability density of a normal variable $P(x)$ is:

$$P(x)\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{1.4}$$

where μ is the mean and σ the standard deviation.

This distribution is symmetric about the mean and the shape is determined by the standard deviation, σ. It is important to ecologists because most of the common statistical tests assume the data to be normally distributed. While many distributions obtained from observation, for example the heights of men, are approximately normal, the spatial distribution of individuals is seldom, if ever, normal. The distribution of a field population could approach normality only if the individuals were randomly scattered and the population was so dense or the size of sampling unit so large that considerable numbers were present in each sample. Therefore, in contrast to other distributions described below, the normal distribution is not of interest to ecologists as a means of describing dispersion. Its importance arises solely from the fact that for many statistical methods to be applied the frequency distribution must be normal. Although analysis of variance is quite robust to deviations from normality, data whose frequency distribution is considerably skewed and with the variance closely related to the mean cannot be analyzed without the risk of errors. It is sometimes possible to use a simple transformation to "normalize" the data.

Simple transformations

Ecological data is usually skewed because species distributions hold a few highly abundant species and a long tail of low abundance forms. Such distributions are usually transformed by taking logarithms or square roots. For example, if the square root transformation were applied to 9, 16, and 64 they would become 3, 4, and 8, and it will be observed that this tends to reduce the spread of the larger values. The interval between the second and third observations (16 and 64) is on the first scale nearly seven times that between the first and second observations; when transformed, the interval between the second and third observations is only four times that between the first and second. It is thus easy to visualize that a transformation of this type would tend to "push" the long tail of a skew distribution in, so that the curve becomes more symmetrically bell-shaped.

It must be stressed that transformation does however lead to difficulties, particularly in the consideration of the mean and other estimates (see below). It should not be undertaken routinely, but only when the conditions for statistical tests are grossly violated.

The distribution of individuals in natural populations is such that the variance is not independent of the mean. If the mean and variance of a series of samples are plotted, they tend to increase together. Where sampling and other errors are fairly large it will usually be found adequate to transform the data from a regularly distributed population by taking the square of each number, while for a slightly contagious (clumped or aggregated) population take the square root of each value, or for a distinctly contagious population take the logarithm.

In order to overcome difficulties with zero counts in log transformations a constant (normally 1) is generally added to the original count (x); this is expressed as "log ($x + 1$)". This transformation can produce serious distortions and should be avoided when possible. It is customary to transform percentages to angles (arcsin transformation). In some fields, such as multivariate analysis (see Chapter 12), extreme transformations such as the 4th root have been advocated. Such transformations should be avoided.

In summary, the use of transformations can lead to problems and should not be routinely undertaken. The biological interpretation of estimates based on transformed data is often difficult. If data have been transformed, the means of the untransformed and transformed values should be presented.