1

Sampling, data types

(A most influential first step...)

Biological objects are real things apprehensible by the senses: plant and animal individuals, their communities, organs, cells or other organizational units, and so on. Selection among these objects is largely determined by the objectives of our study, by the investigator's purposes and preferences, and are limited by financial and temporal constraints and other, more practical than scientific factors. Whenever one is not satisfied with a mere informal description of these objects but wishes to give a deep scientific evaluation as well, i.e., intends to do something what this book is all about, then several important questions must be answered before selecting the objects. Just to mention a few: Does the way of selecting the objects satisfy certain criteria imposed by the methods to be used in the main study? Are the observed and recorded data suitable for analysis by any method at all? Do we not restrict ourselves to a very narrow methodological sequence in forthcoming stages of the study? Is the selection in accordance with the objectives of the survey? and so on...

That is, if our work is not finished by a descriptive phase, then we launch a methodological series whose first step is the most decisive. Any error at the outset may completely destroy and invalidate many years of work. Theses and papers can be refused by the referees if it turns out that the final conclusions are not supported by adequate field work or the results cannot be taken seriously because sampling was biased. There is another type of unwise thinking as well. It happens very often that the biologist has finished her/his field work and gathered large amounts of data. When the field book is already filled up with numbers the researcher tries to find a statistical method that 'fits best' the data. This attempt proves unsuccessful in many cases, and it is often too late to realize that the entire work should have been started in a completely different way.

Now is time to emphasize that knowledge of the fundamentals of *sampling* theory is essential in multivariate studies as well. We can save lots of time and effort if our study plan clarifies all problems and outlines all tasks associated with data collection in advance. This

chapter has been written to help the investigator make correct decisions at this early but very important phase of work.

1.1 Sampling: basic terms

The biologist may obtain information on the objects of interest via simple observations. Numerical data are not recorded in this case, the results of the survey are summarized in the human brain mostly for the investigator himself. Sometimes these results are communicated verbally to others. For example, a phytosociologist goes to the field and examines the plant communities by visual means, thus forming a preliminary picture on the features of the vegetation which strongly depends on his previous experience. This is what vegetation scientists call the *reconnaissance* (Cain & Castro 1959). In admitting the importance of such precursory actions, we have to be aware that without data no evaluation of observations is possible in future stages of the study. One fundamental requirement of sampling is that data are recorded in a format *suitable for further processing.*

The first question in order is: how one defines the word *sample*? Before answering this question, which is not as straightforward as it appears, we must run through a short discussion of other terms. In statistics, the set of all the possible data that can be derived for a collection of objects is called the *population*. This terminology is somewhat misleading and is therefore a potential source of confusions in the field of biology, because the word population has long been reserved with a different meaning in genetics or demography. In order to solve this ambiguity, the set of all the possible data will be called the *sampling universe* in this book. For example, the height data of all the trees in a forest, the body weight measurements of all the individuals of a given species or, which is less obvious at first glance, the species identities of all fish in a lake constitute such universes. The sampling universe cannot be defined by other terms; the investigator decides based on special considerations of the subject whether an element is a member or a non-member of the universe.

In theory it is possible, but in practice it is very rarely feasible, that all the values of the sampling universe are determined by the surveyor. In fact, this process is called the c*omplete enumeration* and should be clearly distinguished from sampling. (Those, who can perform complete enumeration in their studies can jump to the next section.) During sampling, only a part or a subset of the universe becomes known, and this subset is called the *sample* (Fig. 1.1). The sample will serve as a basis for all analyses, therefore conclusions on the universe will completely rely on the sample. Consequently, it is extremely important to be familiar with the fundamentals of sampling methods.

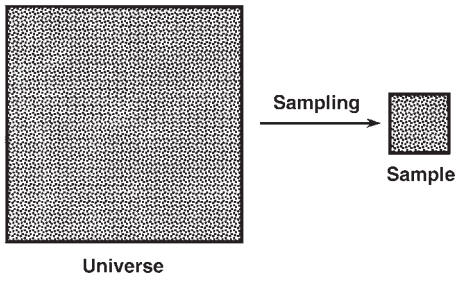

Figure 1.1. A simple scheme of sampling.

When the environmental biologist takes five test tubes of water from different points in Lake Erie in order to measure its pH, and calls each tube a 'sample', then how can we interpret the above definition of this term? True, there is a strong discrepancy between scientific jargon and everyday language, and it can only be resolved if the theoretical and empirical aspects of sampling are clearly distinguished from each other. From a theoretical viewpoint, we can say that the five pH values constitute a sample from all the possible pH measurements (which is an infinite number, see Subsection 1.2.2) that can be taken in Lake Erie in the given point of time. The terminological problem is solved by saying that each pH value was measured in *sampling units* (rather than 'samples'), that is, a test tube is considered as a sampling unit. There is no chance for confusion if we measure tree height in a forest or determine species identity of fish in a lake, because the sampling units are the individuals themselves which would never be called by anyone as 'samples' (this topic is elaborated in Subsection 1.2.2). In a theoretical sense, sampling is the derivation of a subset of all possible data; in a technical sense, sampling is a selection or placement of sampling units in the real representation of the universe.

In considering the manner a sample is selected, we must make another distinction. In biology, there are many instances of 'sampling' when the investigator himself decides whether a given individual should be included in the sample or not. Based on his previous experience, the phytosociologist often decides that some 'apparently degraded' or 'atypical' sections of the vegetation under study are to be partly ignored or completely excluded from the evaluation. For taxonomical studies, selection of specimens in 'good shape' is the general pratice, and individuals that appear less 'developed' are simply discarded. This type of data collection is called the *preferential* sampling: the investigator prefers certain parts of the universe on account of the others. It may also happen that, despite all efforts, some parts of the universe are not accessible (the area is fenced, there is no time to extend sampling to the whole universe, and so on). There is one thing in common in such situations: the sample will not represent the whole universe statistically, therefore conclusions derived from the sample *cannot be generalized to the universe!*

What are then the criteria of representativeness? How can we achieve the ideal case that our results and conclusions are satisfactory with respect to the entire statistical universe? The answer is fairly simple: the sampling procedure must involve a *random element*. This ensures that all parts of the universe have an equal chance of being included in the sample. As we see later, the technical realization of randomness is not as simple as it appears.

1.2 Sampling alternatives

Henceforth, sampling will be understood as a procedure that yields a representative sample. When making our plan for sampling we must consider three basic choices (cf. Kenkel, Juhász-Nagy & Podani 1989), as detailed below.

1.2.1 Estimation versus pattern detection

The first choice relates to the objectives of the survey. In many studies, the sample is taken for the purpose of estimating some statistical *parameter* of the universe. Such a parameter is the

average (more precisely, the expectation or mean) of a measurable or countable trait (e.g., body height and weight, number of individuals) or of a community characteristic as measured by an appropriate function (e.g., species/individual diversity). The literature on sampling is almost exclusively devoted to this kind of sampling objectives, because unbiased estimation is a fundamental condition of all significance tests and traditional biometric analyses. (In simple terms, unbiasedness means that the average of estimates derived from many samples equals the true value of the parameter sought.) Further requirement is that 'sampling error' be *minimized*, which is usually achieved by minimizing the variance of measurements.

We must note that this book will focus on problems in which estimation is not the final objective of the survey, although it may appear in the first phase of data collection. The book is devoted to exploratory data analysis as applied to reveal biological patterns or to summarize information thereof. In the widest sense of the word, pattern may be a classification, a background gradient or underlying continuous trend, or a spatial or temporal variation of biological objects. In order to get the deepest insight into such patterns, we will never need to minimize any 'error', since there is not much to reveal from a homogeneous sample. To the contrary, the sample must be taken so as to *maximize variation* (e.g., expressed in terms of variance or other meaningful function) of its elements. Ideally, the wider the morphological variation represented in the sample, the more we can learn about the morphological variability of the population of a species. A community survey will be more informative if the diversity of species assemblages is maximized in the sample. Needless to say, these two contrasting objectives, i.e., parameter estimation and pattern detection, require completely different sampling strategies.

1.2.2 Discrete versus continuous sampling universe

The examples mentioned in Section 1.1 (forest trees and water 'samples' in test tubes) illustrate the next, just as important contradiction. The individual trees are well-distinguishable, naturally distinct units (although we know that this is not always the case, but this is irrelevant to this discussion), and may be directly applied as sampling units. There are a finite number of trees in the forest (say N), so that the number of possible tree samples, with at least one element, that can be taken in this forest will be 2^N-1, again a finite number. Each tree may be included in or excluded from the sample, so that for N trees we have $2\times2\times2\times \ldots. \times2 = 2^N$ different possibilities, but one of these, the sample containing no trees at all, is ignored. Whether the samples with one or just a few individuals are useful for our purposes is a different matter. The forest composed of tree individuals is a good example of a *discrete* sampling universe. Most books devoted to sampling theory are confined to this situation, discussing in much detail the possible ways of selecting the discrete elements, but forgetting the next type of universe.

If one wishes to measure the pH of the water of Lake Erie, there is no natural, distinct sampling unit as in the above case. The sampling universe, in technical sense, is the whole water body of the lake and constitutes a *spatial continuum.* From this continuum, we have to take an arbitrarily delineated piece, the sampling unit in form of a the test tube of water, in order to be able to bring it to the laboratory for chemical analysis. The size of the sampling unit is determined by the investigator, but regardless the size, the number of possible units that

can be taken from the lake is infinite. Consequently, the number of samples that can be obtained will also be infinite. The situation is similar in a vegetation survey when the percentage cover of species is recorded. Plots (often called quadrats no matter if the shape is square or not) of arbitrary size and shape are placed in the community, and such a plot can be taken in an infinite number of ways. Taking blood 'samples' is another example: the full blood tissue represents the sampling universe from which a few cm^3 is removed to estimate certain parameters (e.g., hemoglobin %, number of monocytes).

1.2.3 Univariate versus multivariate cases

In the simplest situation, one measures a single property of the individuals, for example tree diameter in forestry. When we evaluate the spatial pattern of a single species by different quadrat or grid methods, the situation is also *univariate.* Various aspects of univariate sampling have been discussed in the literature in much detail, and we need not waste time and space here (see the literature review at the end of this chapter). We are interested in *multivariate* situations, where several properties (characters, traits, variables) are observed or measured simultaneously.

1.3 Main characteristics of sampling

The three choices introduced in Subsections 1.2.1 to 1.2.3 yield eight different combinations. For our purposes, these combinations are not equally important. Hereafter we shall focus on two of these, so that the subject matter of the book can be summarized very briefly in terms of the main criteria of sampling:

- *Pattern detection, multivariate case, discrete universe.* The study objects are natural units, such as individuals of a population, discrete habitats (lakes, islands), etc.
- *Pattern detection, multivariate case, continuous universe.* The objects are sampling units obtained by delineating certain pieces of the universe, such as soil, water and air 'samples' (understood in the colloquial sense of the word), and sampling units used in community analysis (e.g., points and linear units or two- and three-dimensional shapes).

Now we can turn to the discussion of the four basic characteristics of sampling strategies. In the discrete case, only the first two are meaningful; in the continuous case, all the four must be considered. It is therefore essential to know at the outset what choices are relevant to our actual survey.

1.3.1 Sample size

The first distinction relates to the theoretical and empirical sample sizes. The empirical sample size is the *number of sampling units* denoted, say, by m. Since we have clarified already that a water or a soil 'sample' is considered as a sampling unit, sample size and sampling unit size cannot be confused with each other (see Subsection 1.3.3). Unfortunately, the English language biological literature is not always consistent in this regard, and 'sample size' is frequently understood as sampling unit size, which leads to serious confusion.

Since we have defined the sample to be a subset of all possible data (Section 1.1), in theoretical sense, the number of values will be its size. A multivariate investigation is extended to several, say n variables, so that the theoretical sample size will be $n \times m$.

What factors should govern us when defining empirical and theoretical sample size in our study? As far as *m,* the empirical sample size, is concerned, the rule is that it should be as large as possible. The more units that are involved, the more information becomes available on the sampling universe. Costs, time and other logistic considerations will obviously impose a strong limitation on the value of *m*. When preparing the sampling survey one may also consider the maximum capacity of computer programs that will be used subsequently for data analysis, although a large sample can always be reduced later, whereas a small sample can never be increased afterwards.

The number of variables (n) should be as high as necessary for a meaningful description of the objects. Too many variables will of course be redundant because of their correlations, whereas some variables may turn out to be irrelevant. The 'difficulty' is that the investigator usually does not know in advance which variables are unnecessary or superfluous. In any case, one should always avoid using variables that are direct functions of each other. For example, the data should not include body height, body width and their ratio simultaneously; we should keep only two of them. Also, the length of leaf blade, the petiole and the sum of the two should not appear together. In many cases, the number of variables arises automatically by the end of sampling, such as the number of vascular plant species in a phytosociological study. It is advisable to consider all of them, and if necessary for some reason, the number of variables can always be reduced subsequently. Further comments concerning the number of variables are given in Subsections 1.4.3-1.4.7.

The proportion of empirical sample size and the number of variables deserves attention as well. If n is much larger than m, then it means almost always that the variables will be highly correlated (or associated, see Chapter 3). That is, if we have to decide upon the number of variables then it need not go far beyond m. In the reverse situation, when $m >> n$, it is worth searching for the possibility to include more meaningful variables, but this is not an absolute requirement. There is no general rule as to the relative magnitude of n and m although there are exceptions. In canonical variates analysis (Section 7.5), the number of variables cannot exceed the number of objects (observations), because singularity problems will hinder the computations. The same is true for calculating the generalized distance (Equation 3.94).

1.3.2 Deriving a sample from the universe

In a discrete universe, the individuals themselves represent the sampling units, so that the elements of the sample are *selected* by the surveyor. It is analogous to the *arrangement* of sampling units in the spatial or temporal continuum (continuous case). This subsection discusses methods that ensure that the selection or arrangement of units will guarantee the representativeness of the sample.

In *simple random sampling,* all individuals of the discrete universe, or any points in the continuum, have the same chance of being included in the sample. The elements of the sample are therefore independently chosen.

This condition is not always easily satisfied in practice. In the discrete case, all elements of the universe can be numbered, yielding the so-called the *sampling frame,* and then a random number generator will help us select the sample randomly. Such numbering is clearly impossible in most field studies, where randomness may be achieved by locating random points on the map of the study area, by identifying these points in the field and taking the closest individuals as sampling units (Fig. 1.2a). The 'random walk' method also applies in such cases: starting from a single random point we move at random distances in random directions (Fig. 1.2b),

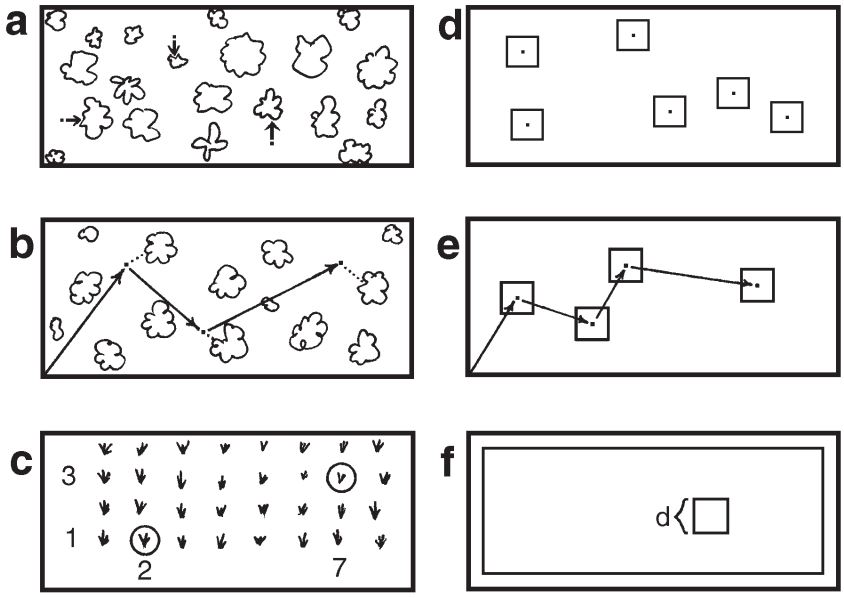

Figure 1.2. Implementation of simple random sampling in the field. **a:** random point and nearest individual, **b:** random point and nearest individual combined with random walking, **c:** random integers applied to regularly arranged elements of the universe, **d**: random point method to allocate quadrats, **e:** placement of quadrats by the random walk method, **f:** the edge effect is manifested in a *d*/2 wide outer strip for quadrats of side length *d*.

thus allocating random sample points. In a cornfield or a regular tree plantation, we have another alternative to ensure randomness. Columns and rows are selected using randomly generated numbers, and the resulting column and row indices will identify the elements of the sample (Fig. 1.2c).

In the continuous case, no sampling frame can be defined, but the map of the area or the random walk method can be used to allocate sampling points. A sampling unit, e.g., a quadrat in a plant community, can be arranged around randomly taken points, as shown in Figs 1.2d-e. There is one problem, the *edge effect,* which appears only in the continuous case. Quadrats that would overlap the boundary of the study area, i.e., part of them would fall outside, should be discarded. As a consequence, within a narrow strip (of width depending on the size of our quadrats), the points will have a lower chance of being sampled (Fig. 1.2f). The closer a point is to the boundary of the study area, the lower its chance of being included in a quadrat. Equality of chance is ensured for points falling inside the area only. The edge effect becomes more pronounced when the sampling unit size is increased. This fact has to be considered when evaluating the results. Complete correction of the edge effect is difficult, if possible at all, in multivariate pattern detection in a continuous universe. This problem seems serious in plant ecology, but the randomness of a water ‘sample’ from a lake is hardly influenced.

Random sampling can be performed in two or more steps if the units of the universe are agglomerated into clusters, for example. Random selection of aggregates (e.g., cell colonies) is the first step, and in the second phase we do a random selection within the aggregates previously chosen (e.g., cells within the colonies). Such a strategy always assumes the existence of a hierarchy amongst the units: small subsets of the universe are included in larger subsets, and so on. Hence its name, *nested* sampling or the term *subsampling* which better reflects the

subordinated relationships. Nested sampling is commonly applied to estimation problems, although it may also be useful for multivariate surveys (e.g., Green 1979, p. 36). Mixed strategies are also possible: for certain variables (e.g., plant species of a community) random quadrats are used, whereas for others (environmental variables such as soil reaction, lime content) nested arrangement is chosen, with many replicates within each quadrat.

A variant of the above strategy is *stratified random* sampling. This is the best method if there are external criteria by which the universe can be subdivided *a priori* into subsets ('strata, layers'). In each stratum, simple random sampling is performed so as to maintain the proportionality among strata in the final sample.

In stratified random sampling, attention is focused upon the proportions and the criteria that separate the layers. The criteria must be in fact external and cannot be a function of some variables included in the sampling. In vegetation science, for example, vegetation strata can be distinguished on the basis of the microtopography or soil properties of the study area. Use of the presence/absence of a species should be avoided, however, if that species is included among the sampled variables. Proportionality is ensured by applying different sample sizes in the strata. In the simplest case, these sample sizes are proportional to the physical size (e.g., area) of the strata. Other types of proportionality also exist, but these are mostly confined to estimation problems (for example, subsample sizes may be inversely proportional to the variance of strata in order to minimize total sampling variance).

Random sampling, although theoretically well-founded, is usually impractical. It may be problematic to ensure full randomness and the definition of the sampling frame is impossible if the universe is very large in size. A further problem is that the sample can give an uneven representation of the universe, as illustrated by Fig. 1.3a. Statistical correctness does not always coincide with our expectations: for randomly placed quadrats large pieces of the study area may completely remain untouched (and other pieces may remain intact for another set of

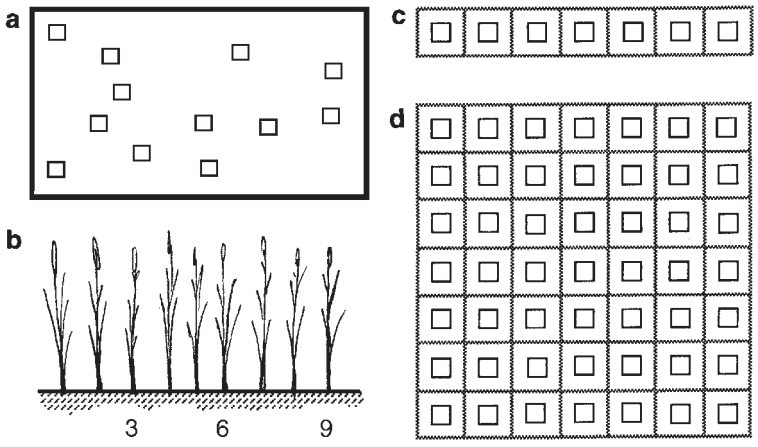

Figure 1.3. Random arrangement may be uneven, excluding large areas from sampling (**a**). Systematic methods ensure a more even distribution, e.g., by choosing every *k*-th idividual in the discrete case (**b**). The transect (**c**) and the grid (**d**) are examples for systematic arrangements in the continuous case.

random quadrats). A solution is to increase sample size heavily (which is not always feasible) or to apply the so-called *systematic* sampling. In this, only a single sampling unit, the *pivot element* is selected at random, and the others are obtained automatically by applying regular gaps, i.e., the *sampling interval* between the units.

In the discrete case, the size of sampling interval is a *k* integer. For instance, in a cornfield we decide in advance that every third plant will be included in the sample in every third row. The pivot element is preferably a random individual selected from 3x3 plants in one corner of the field. Then, in both directions, every third plant is included until we reach the boundary of the universe (Fig. 1.3b). Sample size is therefore a function of *k* and the size of the universe. In the continuous case, the sampling interval ('*spacing*') is a distance in space or time. The pivotal sampling unit is allocated at random, and the others are taken at fixed *k* distances. For transects (Fig. 1.3c), the units are arranged linearly, whereas grids (Fig. 1.3d) are two-dimensional. Transects are used preferably if the effect of an underlying factor, a gradient is to be revealed (e.g., a humidity gradient in riparian vegetation). Temporal transects are also used, for example, when examining the insect material collected by light-traps at regular time intervals. Grids ensure the even distribution of quadrats in an area and are therefore popular in vegetation mapping. In special cases, the units are contiguous and cover only a small area. Pattern detection procedures of ecology utilize contiguous grids to evaluate scale-dependence of structural parameters (univariate case: e.g., Greig-Smith 1983, multivariate case: Juhász-Nagy 1976, 1984, 1993). The grid is a mere starting point to generate increasing 'block sizes' by amalgamating the basic units ('*space series analysis*', see Subsection 1.5.2).

Systematic sampling may perform very poorly on rare occasions when the spatial distribution of the universe shows periodic trends, and at the same time, this regularity coincides with the sampling interval (Greig-Smith 1983). Assume, for instance, that the vegetation of more or less regularly spaced dunes is sampled by a transect, and the value of *k* is roughly the same as

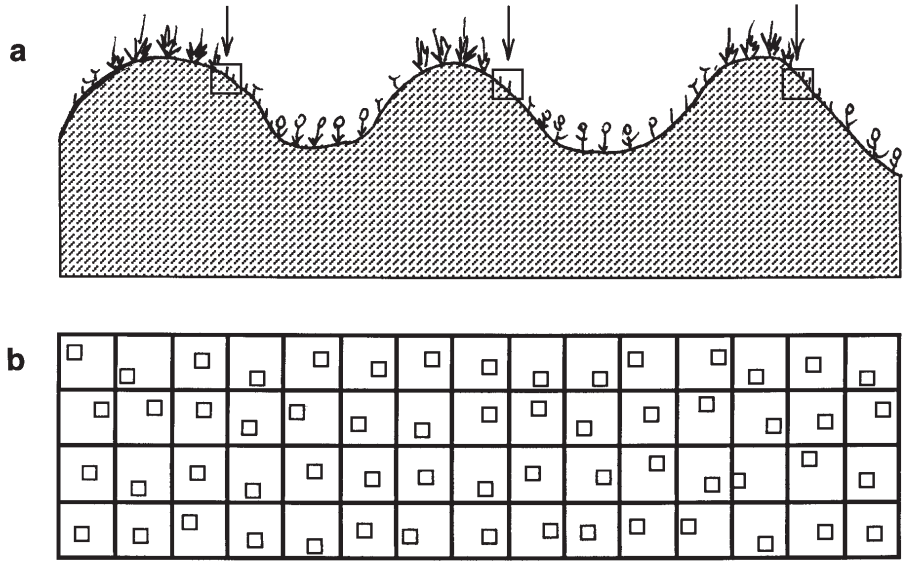

Figure 1.4. The coincidence of systematic arrangement and natural regularity (**a**). Semi-systematic strategy for the two-dimensional continuum (**b**).

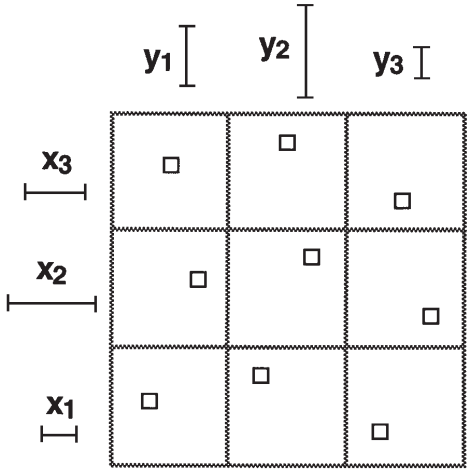

Figure 1.5. A special case of semi-systematic arrangement with fixed coordinates for a 3 by 3 grid. The origin in each block is the lower left corner. x_1 is applied to all sampling units in the first row, y_1 specifies the vertical coordinates of all units in the first column, and so on.

the distance between two dunes (Fig. 1.4a). Depending on the location of the pivot element, all the units will be placed on similar positions on the dunes. The sample thus obtained will not represent faithfully the vegetation of the entire study site, because there may be substantial differences between troughs and tops, southern and northern slopes, etc.

The potential bias due to systematic sampling may be eliminated by a mixed strategy, the *semi-systematic* method. The universe is subdivided into blocks of equal size (by a grid, for example) and then within each block a fixed number of sampling units are placed at random (Fig. 1.4b). Advantages of the random and systematic designs are thus combined.

The terminology regarding this sampling strategy is confusing and ambiguous. Greig-Smith 1983, and Southwood 1978, for example, understand this strategy as stratified, whereas for others (Orlóci & Kenkel 1985, Green 1979) the stratified method is the one as described earlier in this subsection. Although there is some similarity between the stratified and semi-systematic strategies (in both cases randomization is achieved in a subdivided universe), it is useful to maintain the terminological distinction. Stratification best reflects the case when the universe is subdivided according to an external criterion, i.e., not necessarily into similar units. The term semi-systematic, on the other hand, refers to the regular and artificial subdivision of the universe..

A variant of the semi-systematic strategy ('stratified unaligned systematic sampling', Quenouille 1949, Greig-Smith 1983, preferably called the *unalingned semi-systematic sampling*) rejects complete randomization within the blocks, ensuring a more even spread of sampling units over the area. For each row and column of the grid, a fixed random coordinate is generated to specify the positions of the sampling units (x_1, x_2, x_3 and y_1, y_2, y_3, respectively, in Fig. 1.5).

Smartt & Grainger (1974) found that this very special strategy gave the best estimates on the relative proportions of vegetation types in a biogeographical survey. Advantages of the method in a multivariate context need to be explored, however.

1.3.3 Sampling unit size

In a continuous universe, the sampling unit has to be artificially delineated by the investigator. The first question is: what size should this unit have? Practical considerations, such as ease in

the field, are decisive in many cases, but there are other aspects that require attention. First of all, distinction between two types of the continuous universe is important. One type is represented by *communities* in which organisms are arranged in the spatial continuum, and the other by their *media*, such as water, soil and air.

In community analysis, an obvious requirement is that the sampling units should not be too small compared to the organisms. It cannot be too large either, because of limitations in cost and effort. Nevertheless, a wide range of sizes remains within which the investigator can make a choice. This choice depends primarily on the objective of sampling and, in turn, of the entire study. For the sake of completeness, note that in estimation-oriented studies the rule of 'the smaller the better' (Elliott 1977) may be followed. If the time and money available restrict the size of the piece of the continuum to be included in the sample (the product of sample size and samplig unit size is fixed), then many small units are preferable to a few large ones, because this reduces sampling variance. The success of this reduction also depends on the spatial pattern of the objects, but this problem relates to estimation and need not be detailed here; see Green (1979, p. 131-133). In pattern detection, there is no interest in variance reduction, and the product of sample and sampling unit size is irrelevant. Instead, pilot sampling is necessary before the main survey so that one may establish the range of sizes within which the data obtained are expected to give as much 'information' on the pattern in the universe as possible.

> How can sampling unit size be optimized? Phytosociologists and plant ecologists have been looking for the answer for decades. Early attempts used the species/area curve, with various modifications, in order to find the 'optimum' plot size at which the number of species does not increase any further. The problem is that species number is a simple diversity measure (a textural variable) that provides no information at all on the area at which the largest amount of information can be extracted on the structure and pattern of the community. For the same reason, related textural parameters cannot be successful either.
>
> The work of Juhász-Nagy (1967-1993) helps orient us in this complicated area. He has shown that species/individual diversity measures should be replaced by species combination/quadrat (or florula-) diversity and related information theory functions, and their behavior must be examined as a function of quadrat size. These methods require relatively large sample sizes, however, especially if the number of species in the community is high. We may use simpler techniques, such as the expected resemblance versus area curve, which has peaks if resemblance is a distance-like measure (see Chapter 3), just as the information theory measures (Podani 1984b). In the main stage of the sampling survey, one should use the quadrat size where these functions reach their extreme values, maxima or minima (Fig. 1.6). These considerations are restricted to cases of presence/absence, however, and no general method has been suggested yet for the optimization of sampling unit size for multivariate abundance, cover or biomass data. A possible, but laborious solution is to try several sizes in different phases of the study, or in the whole survey, if possible, and to examine the effect of size upon the results and conclusions. Those who cannot afford such pilot studies must rely on various rules of thumb and other, more or less subjective propositions proliferating in the literature. For different plant community types 'adequate' quadrat sizes are found, for example, in Mueller-Dombois & Ellenberg (1974 p. 48), Westhoff & Maarel (1978), Gauch (1982, p. 55), Knapp (1984 p. 111), etc.

The procedures just mentioned are confined to communities of sessile organisms such as plants or benthic animals. For most animal assemblages, due to the mobility of the organisms, quite different and highly specialized sampling strategies are required. One such method is the line transect method of ornithologists in which the width and length of 'strips' and the date of

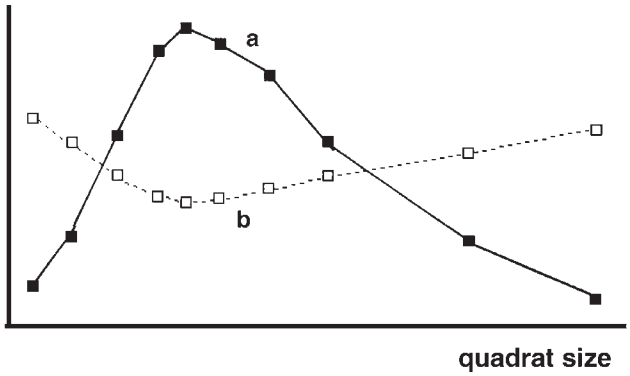

Figure 1.6. Sampling unit sizes nearly optimal for pattern detection can be determined by examining the dependence of species combination/quadrat (florula) diversity (**a**) and expected resemblance (**b**) on quadrat size. The scale units are immaterial on both axes and are therefore omitted.

sampling are the most important. In case of animals, usually the traditions and practical considerations become decisive, because it is often difficult to define 'optimal' sizes for data analysis. Nevertheless, there are studies demonstrating the utility of the Juhász-Nagy methods in the analysis of planktonic crustacean assemblages (Dévai et al. 1971).

The objective of the main study is not always classification and subsequent description of community types, as in phytosociology. Green (1979) lists many examples of using multivariate methods to evaluate community change caused by environmental deterioration. Such community monitoring draws conclusions from observed structural changes and therefore should also optimize sampling unit size. However, this size may change over time, especially if changes are substantial, and one may well question the existence of any optimum! The same is true for successional surveys utilizing permanent quadrats. When sampling unit size is fixed, spatial and temporal factors behind the changes cannot be distingusihed. This is sufficient to illustrate that – in theory – use of several sizes is usually unavoidable.

The universe of the medium-type poses fewer problems. Sampling unit size is simply a technical question that relates to the analytical methods and equipment available, as well as to their precision (e.g., in pH measurements in soil or concentration measurements in polluted air). A detailed discussion is beyond the scope of this book.

1.3.4 Shape of sampling units

When determining the shape of units, we must bear in mind the contrast between estimation and pattern detection. For estimation purposes, elongated shapes are usually optimal because sampling variance may be substantially decreased this way. In community studies, however, long shapes have an obvious disadvantage; such a sampling unit includes individuals that fall far apart in the field. Therefore, subsequent estimates of structural parameters, e.g., pairwise interspecific association (Pielou 1977, Greig-Smith 1983) or the simultaneous association of many species (e.g., Podani 1984a) can be misleading. Furthermore, elongated sampling units have much higher chance to overlap community or environmental boundaries than isodiametric units (squares, circles, Fig. 1.7). (As noted earlier, for many authors 'quadrats' are not always square in shape, as the term might suggest, and there are references to 'circular quadrats'

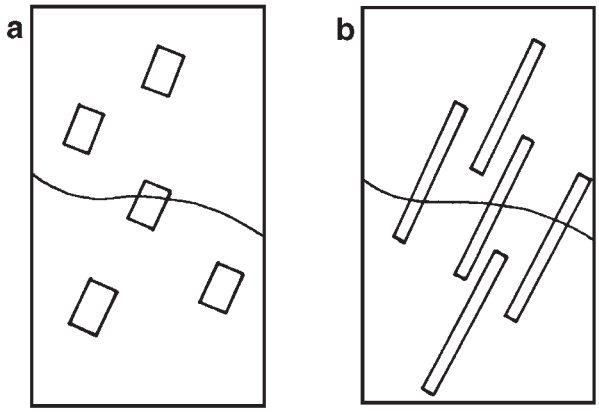

Figure 1.7. Square sampling units have a smaller chance to overlap with heterogeneities within the universe (**a**). Elongated shapes, on the other hand, can easily confuse pieces of the universe that do not 'belong together' (**b**).

in the literature!) In conclusion, for pattern analysis purposes in the multivariate case the isodiametric shape is preferred.

A further advantage of squares, and especially circles, is that the edge effect within the unit is minimized (this edge effect should not be confused with the one shown in Fig. 1.2f). For elongated shapes, the organisms have higher chance to fall upon the periphery of the unit and it is difficult to decide the threshold for including an individial. Southwood (1978) proposed to consider such individuals only on half of the periphery (e.g., on the left and upper side of a square). This convention, although arbitrary, seems fairly reasonable in sampling surveys no matter whether the objective is estimation or pattern detection.

For elongated shapes, there is a fifth property: *spatial orientation*. There is a contrast between random arrangement and uniform (Fig. 1.7b) orientation of rectangles. A constant orientation may coincide with an underlying gradient. Whenever we insist upon the use of rectangular shapes, random orientation is essential unless we know the underlying factors and we have a good reason to use a fixed angle of orientation.

1.3.5 On 'plotless' sampling, a brief overview

The discussion of sampling would be incomplete without mentioning a popular method of plant ecology, the so-called plotless sampling. This name indicates that the units are no longer two-dimensional: quadrats are 'reduced' to lines or sample points. The investigator records species identity of individuals touched upon by a line or a point (usually for estimation purposes), or the distance from the nearest individual is measured (in order to evaluate spatial pattern of a species, see Mueller-Dombois & Ellenberg 1974, pp. 93-118, Greig-Smith 1983, pp. 47-53, for details).

Plotless sampling rarely provides data for multivariate analysis. An interesting exception is presented by Williams et al. (1969). They recorded the species identity of the first, second, third, ... and *n*-th nearest neighbors to the random point. As a result, there was a sequence of species names associated with each sample point (considered as a sampling unit). Then, these sampling units – within a relatively small area – were classified at increasing values of *n* to evaluate spatial pattern of the community. Podani & Czárán (1997) combine this approach with Juhász-Nagy's information theory methods in the analysis of mapped spatial point patterns.

1.4 Data: measurement scales and other properties

The sampling units are selected in order to record and code, in the form of *data,* the variables (characters, features) that describe them. Without doing this, no sampling is done, because there would be no sample (a subset of data) at all! After discussing sampling strategies, a most logical step is therefore to introduce the possible data types.

Data are usually collected by measuring or counting. The observations, however, do not provide data directly applicable for analysis if they are not in numerical format. The characters observed must often be coded, i.e., a number is to be associated with each potential state of every character (so that *random variables* are defined in the statistical sense). Knowledge of data (variable) types is essential for an appropriate selection of data analytical methods, and I emphasize that the choice of the data type will largely determine the analytical techniques that can be used subsequently. The terms 'qualitative, quantitative, semiquantitative and numerical', which often appear in everyday speech and also in the scientific jargon, are vague and should be avoided. The discussion below provides a more straightforward typification whose understanding is essential to follow the discussion in the next chapters.

1.4.1 Scale types

It is generally acknowledged that the potential values of a variable may be expressed on four different scale types (Anderberg 1973).

1) For the *nominal* scale the distinguishability of states is the requirement. The only statement we can make is whether two states are identical or not, so that the = and ≠ operators are valid. For example, leaf shape (with states such as elliptical, obovate, lanceolate, etc.) is a well-known nominal character. The elliptical state may be coded with 1, and the lanceolate with 3, which is obviously arbitrary, and other coding may equally apply. Thus, the difference between the values is not meaningful, and operations other than the above two cannot be performed on such data. Taxonomists often refer to this type as *multistate* characters, but this is a little misleading. Nominal variables can often have only *two* states, and there are other two-state (i.e., binary) variables for which coding is not arbitrary at all (see below), which is a potential source of confusion.

Nominal variables coded by the non-negative integers 0, 1, 2, ... can be directly processed by two resemblance measures (Equations 3.103-104) and by block-clustering (Chapter 8). However, the majority of methods cannot treat such variables properly, unless some manipulations are performed. For example, a nominal variable with p different states can be replaced by p binary variables (Gordon 1981).

> For the morphological example mentioned above, each leaf shape will become a separate variable such that the presence of the given shape is coded by 1, the absence by 0. This means that this dichotomized variable will receive much weight in the analysis if the resemblance coefficients to be used treats presence and absence symmetrically (see Subsection 3.2.1). To illustrate this, let us consider a hypothetical example with two variables and 10 sampling units:

	Sampling units
Variable 1	1 2 1 3 4 3 2 5 1 2
Variable 2	1 1 0 0 0 0 1 1 1 0

After dichotomizing the first variable, we obtain six binary variables, five new and one original:

Sampling units
1 0 1 0 0 0 0 0 1 0
0 1 0 0 0 0 1 0 0 1
0 0 0 1 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0
- - - - - - - - - -
1 1 0 0 0 0 1 1 1 0

If we calculate the simple matching coefficient (Equation 3.6) to express similarity among units (columns), then the similarity between the first two will be 4/6 whereas for the first and the third we obtain 5/6. This larger value yields because units 1 and 3 agree in a character described in terms of five new binary variables. However, if similarity is expressed by Formula 3.23 both pairs will have the same similarity (1/6), intuitively a more acceptable situation. This is sufficient to illustrate that coding and the methods of data analysis must be compatible, otherwise the results are misleading.

2) For the next scale type, in addition to their distinguishability, the states can be ordered logically, and operators < and > are also valid. This is the so-called *ordinal* scale. A typical example is the Mohs hardness scale of solid materials. Differences are still meaningless in this case, the pair of talc and gypsum (the first two in the sequence) and the pair of corundum and diamond (the last two) do not imply the same difference in hardness. In phytosociology, the many variants of abundance/dominance (AD) scale are of this type (see Braun-Blanquet 1965, van der Maarel 1979, Kent & Coker 1992, Table 1.1)

This data type is the most difficult to evaluate. For many data analytical methods, the ordinal scale must be simplified to the nominal (so that we lose the sequential information), and for others expansion is necessary to the next type, the interval scale. This latter operation requires additional information, however (e.g., replacement of AD values by percentages following some convention, e.g., 'category means' as in column 3, Table 1.1), which is not free from arbitrariness. Sneath & Sokal (1973) proposed to replace a p-state ordinal variable with p-1 binary variables. If a given score is the k-th in this sequence, then the first k-1 binary variables will take the value of 1 and the others will be 0. If we consider the first variable of the above example to be an ordinal character, then the Sokal-Sneath method will provide the following four binary variables for the same sample data:

0 1 0 1 1 1 1 1 0 1
0 0 0 1 1 1 0 1 0 0
0 0 0 0 1 0 0 1 0 0
0 0 0 0 0 0 0 1 0 0

Such a conversion always overemphasizes the given character, regardless the coefficient used (in contrast with the previous scale type, for which careful choice of the resemblance coefficient solves the problem). Other possibilities are offered by various rank coefficients developed specifically for ordinal variables (Section 3.4).

3) The *interval* scale represents a great 'advancement' as compared to the previous two. The differences between states are meaningful, which has enormous consequences as far as the possible data analysis methods are concerned. Typical examples are the Celsius and Fahrenheit temperature scales. The difference between 10 and 20 ^{o}C is the same as between 20

Table 1.1. Ordinal scales from phytosociology. Note that the + 'score' cannot be processed and needs to be replaced by a small number, such as 0.1 (column 3). These scales were proposed before the computer age, so the authors did not anticipate the problems associated with the analysis of such variables. It is perhaps better, although more laborious, to record percentage cover data in the field.

State	Braun-Blanquet's scale	Converted into ratio scale as %	Domin's scale
+	cover less than 1 %	0.01	Single individual, negligible cover
1	1-5% cover	3	1-2 individuals, negligible cover
2	6-25% cover	15	More individuals, cover < 1%
3	26-50% cover	38	1-4% cover
4	51-75% cover	63	4-10% cover
5	76-100% cover	88	11-25% cover
6			26-33% cover
7			34 - 50% cover
8			51-75% cover
9			76-90% cover
10			91-100% cover

and 30 ^{o}C. We cannot say, however, that an object having a temperature of 30 ^{o}C is three times as warm as another object of 10 ^{o}C. This is because, in mathematical sense, the scale has no natural zero point (the melting point of ice is an arbitrary, though practical, zero point of the Celsius scale).

> Variables measured on the interval scale are interpretable by most of the methods, but care is needed especially in data transformation. The logarithmic and square root transformations, for example, cannot be used because of the arbitrary zero point.

4) Variables measured on the *ratio* scale have all the properties of the previous scales, and the presence of a natural mathematical zero point allows the interpretation of the ratio of values as well. In other words, the operation of division also applies. Measuring temperature on the Kelvin scale facilitates comparisons that were illogical on the Celsius scale. Variables obtained by measuring length, weight and area or by counting belong to this type, and may be subjected to any kind of data transformation.

1.4.2 A special data type: the binary variable

In addition to the admissible operations that characterize the scales, the number of states that a variable can take is also important. In many cases, the number of possible values is infinite which needs no further comments here. We have mentioned several times the 'opposite' situation when the number of states is only two, the *binary* variables. Species presence and absence in sampling units taken in communities, and presence/absence of morphological characters of taxa are common examples from biology.

Binary variables are usually coded by 0 and 1. One must be careful when deciding which state of the variable be coded by 0, and which by 1. As mentioned earlier, coding determines whether the resemblance functions used afterwards are meaningful or not.

> Coding is immaterial when the resemblance coefficient treats 0 and 1 symmetrically. Examples are information theory functions 3.112 and 3.115, Euclidean distance and related

measures, the simple matching coefficient, the *PHI*, Yule, Rogers - Tanimoto and Anderberg I-II indices (Chapter 3), i.e., those considering the *a* and *d* values (Section 3.2) of the 2×2 contingency table symmetrically, so that reverse coding does not influence the result (Subsection 3.2.1).

For another group of coefficients (e.g., Sörensen, Jaccard, Baroni-Urbani - Buser I-II, i.e., those that attribute unequal importance to *a* and *d*), reverse coding almost always produces different results. In this case, a logical choice is that the state, which means in some sense 'more' than the other, should be coded by 1 and the other by 0. On the ordinal, interval and the ratio scales it is usually easy to follow this convention. For two-state nominal variables, however, coding remains arbitrary, so that these coefficients do not apply. Whenever we are uncertain about a method that we do not know, it is advisable to do a short test by analyzing a sample data set with the two versions of coding and examine if the results differ.

1.4.3 Mixed data

The overwhelming majority of multivariate methods require that all the variables be measured on the same scale, except that ratio and interval scales can often appear together in most cases. There are situations, especially in taxonomy, when other combinations of scale types occur. The simultaneous presence of nominal and interval variables, or of the multistate and binary characters restricts the number of potentially applicable methods significantly. Coefficients 3.103-104 can treat such mixed data, thus allowing subsequent ordination and classification of objects. If we insist upon other techniques for which these coefficients cannot be used, then some variables have to be removed from the data, or scale conversion methods discussed by Anderberg (1973) must be applied to bring all variables to the same scale. Without giving detailed information on these methods, we note that conversion is straightforward in the direction of ratio → interval → ordinal → nominal scale. Each step in this sequence implies loss of information, and the investigator decides whether the loss is negligible or not (in vegetation science, for example, shift from AD values to presence/absence scores is most common). In the opposite direction incorporation of some external information is necessary.

In the previous paragraph, the mixed data type was understood in a sense that there is difference either in the measurement scale or in the number of states among the variables. This is consistent with the general usage of the term 'mixed', but it is to be pointed out that other kinds of mixtures also exist. It is common in ecology that sampling units are characterized by both biological variables (e.g., number of individuals of plant species) and chemical and physical features of the environment. Joint treatment of these two groups in a pooled data set, although mathematically feasible, is usually illogical. Some methods, such as canonical correlation analysis (Section 7.2), are designed specifically to these cases and are capable of revealing the relationships between the two groups of variables as well.

Even if all variables are measured on the same mathematical scale, the actual physical scales may be very different (e.g., temperature, pH, elemental concentrations in g/cm^3 or ppm and so on in the same study), another kind of mixtures. This problem will be discussed in Subsection 1.4.6 devoted to the commensurability of variables.

1.4.4 The problem of missing values

Multivariate techniques require that the data table be full. In other words, all the variables should be known for all the sampling units included in the study. It happens very often, how-

ever, that a few or more values are missing. In taxonomic studies some specimens are not intact, which is most critical in paleontological collections. In ecological surveys, field conditions may not allow complete observations, which cannot be retrieved at later time. Needless to say that in such cases the missing values cannot be represented by zeros, since they are considered implicitly as real data scores by all methods.

> The use of formulae 3.103-104 for measuring similarity between objects may circumvent the problem for some methods (e.g., cluster analysis and multidimensional scaling). If a given score is missing for either or both objects being compared, then the variable will be ignored for that pair. Too many eliminations will diminish the reliability of results, however, so that objects or variables with too many missing values should be removed. Other possibility is the estimation of missing values from the known scores (Beale & Little 1975, Gordon 1981). If object *Q* has a missing value for variable *i*, then estimates can be made according to either procedure discussed below.
>
> 1) Based on the known values, identify in the sample the object that is the most similar to *Q* (using a dissimilarity function, Chapter 3). Then, the missing value is estimated by the known value of this most similar object.
>
> 2) Perform cluster analysis (Chapters 4-5) based on the known data. Then, identify the group to which *Q* belongs, and the average of the known values for variable *i* in this group will serve as an estimated value for *Q*.
>
> 3) Compute the product moment correlation (Equation 3.70) among variables using the known values. Select the variable most correlated with variable *i*. Perform linear regression between these two variables, and using the regression coefficients estimate the missing value for *Q*. (A more complicated version of this procedure uses partial regression coefficients using more than one correlated variable).

As seen, even the simplest procedure appears quite cumbersome in practice, and there is neither guarantee nor confirmation that the estimation is successful. Whenever possible, missing data should therefore be avoided.

The investigator must be careful not to create missing data of his own. The most 'dangerous' field in this regard is taxonomy. Avoid the use of morphological characters whose presence depends on another character. For example, imagine an insect taxon in which some species have wings and others do not. Then, any characters referring to wing features will be obviously 'missing' (more precisely, not existing) for the wingless species (Sneath & Sokal 1973). In this case, wing features should be expressed in terms of different states of a single nominal 'wing character', with the state 0 indicating the wingless condition, 1 referring to a given combination of characters, and so on. 'Nominalization' of such characters, however, is not always possible.

1.4.5 Negative values and constants

Some measurements may provide negative values (e.g., temperature in ^{o}C or ^{o}F). Modification of raw data by standard deviation and centring (Subsection 3.2.1) will also yield negative numbers. Treat these cases carefully, because these values may cause serious computing problems during data analysis. Most computer programs abort when the calculation of the logarithm of a negative number (and of zero) is attempted. Sometimes the sum of negative and positive numbers may be exactly 0, and division by zero, if attempted when standardizing by

the total, will also halt computations. Therefore, it is recommended to modify the data so that all negative values are replaced by nonnegative numbers. For example, a constant can be added to all temperature values, which will not influence the results in the majority of methods. Care is needed when evaluating standardized data, because the use of resemblance coefficients that operate on the sums of values should be avoided.

'Variables' taking the same value for all the objects are uninformative and should be discarded. These *constant* or *invariant* characters will not influence the results, and their zero variance may cause computational problems as well (for instance, in principal components analysis, Chapter 7).

1.4.6 Weighting and commensurability

The investigator may often have the opinion that certain variables are 'more important' than the others, and this distinction should be somehow manifested in the results. Attributing more importance to characters arbitrarily by the data analyst, i.e., *external weighting* is not possible in most methods. If one insists, weighting 2, 3 or more times is possible by including the variable in question twice, three or many times in the data. Note, however, that such weighting may be removed automatically by many data analysis techniques, whereas for others the unit correlation of these repeated characters will cause problems. Dichotomization of nominal and ordinal characters using the method described above is another case of external weighting whose final effect, as we have seen, depends on the choice of the analytical method.

> Most methods covered in this book do not allow external weighting, the exception being cladistic data analysis in which the characters are by no means equal in importance (Chapter 6). In accordance with the objectives of cladistics, some characters are considered to convey much more information than others for phylogenetic reconstruction. This controversial proposition is reviewed by Farris (1969), Fitch (1984, p. 238) and Maddison & Maddison (1992, pp. 197-198).

Usually, biological data imply some sort of weighting by themselves (*implicit weighting*). For example, percentage cover values recorded in a forest community will certainly be much higher for trees and dominant grasses of the herb layer than for sparsely distributed orchids and lilies. These *a priori* differences may reach several magnitudes, and remain unchanged if classifications and ordinations start, for example, from Euclidean distances calculated among objects (Equation 3.47). As a consequence, the results are dominated by trees and grasses. An appropriate choice of methods, especially data transformation and standardization (Section 2.3) may compensate for this weighting (i.e., all species will become equally important) and may even reverse the situation.

Implicit weighting is closely related to the question of *commensurability* (Orlóci 1978). In the example above, the cover of trees and herbs is always commensurable, no matter how large the differences are. All variables express the same thing: the area occupied by individuals of a certain species within a sampling unit. In a data set containing different physico-chemical measurements, the variables are not always commensurable because of differences in the units of measurement. In an ecological survey, pH values may range from 4 to 8, whereas the concentration of a heavy metal in the soil falls between 100 and 200 ppm. Therefore, a relatively small change in heavy metal concentration (say, 5 ppm) will be more influential in subsequent

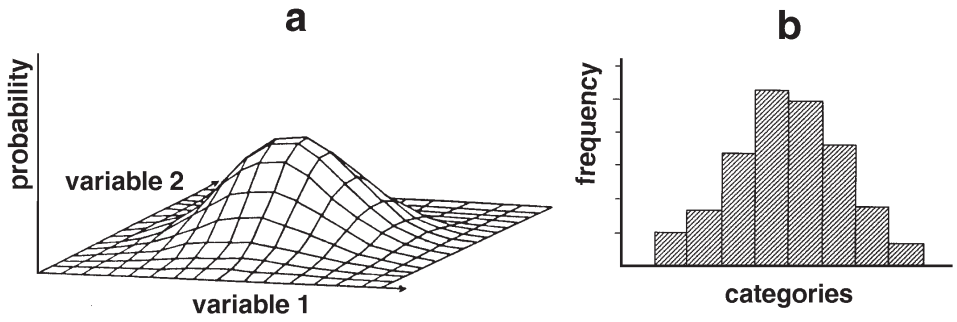

Figure 1.8. The probability density diagram of two-dimensional normal distribution (**a**) and an empirical frequency histogram (**b**).

data analysis than the maximum change of pH, which is undesirable, unless the data are standardized (Section 2.3).

1.4.7 The distribution of variables

Some multivariate techniques require no specific assumptions regarding the underlying distribution of variables (i.e., how probable are the possible values in the sampling universe). Almost all methods of cluster analysis (Chapters 4-5) and non-metric multidimensional scaling (Subsection 7.4.2) are cases in point. Contrary to some views, principal components analysis (Section 7.1) assumes nothing on the distribution of variables either (Chatfield & Collins 1980, p. 58), although the results are most easily interpretable if the distribution is multivariate normal (illustrated for two variables in Figure 1.8a). For discriminant (canonical variates) analysis and canonical correlation analysis, multivariate normality is a fundamental condition, especially if evaluation of results involves significance tests. If this condition is not satisfied, then the analysis still runs on the computer, but the numerical results should always be treated with caution.

Interpretable results may be obtained very often even if certain conditions are not satisfied. For example, ordination diagrams may successfully depict a two-group classification structure in two artificial dimensions that replace the orginal variables. In these cases, we say that the method is *robust.* However, significance tests (Subsections 7.2.1 and 7.5) and probability ellipses enhancing graphical interpretation of ordination scattergrams (Subsection 9.5.2) are invalid. If we wish to apply these methods, the distribution of variables needs to be examined previously (using, for example, frequency histrograms, Figure 1.8b). The variable with strongly non-normal distribution should be removed or transformed (Subsection 2.3.2) before analysis. Nonetheless, multivariate normality is not guaranteed, even though the variables taken separately are distributed normally (Reyment 1991).

1.5 Advanced topics

1.5.1 Space series analysis

When discussing sampling unit size, I pointed out that several sizes may be reasonable in pivot sampling or in the main survey, because there is no overall optimum due to a compex spatio-temporal change, for example. To ensure that sampling unit size is the only influential factor, the other characteristics of sampling (sample size, arrangement and sampling unit size) must

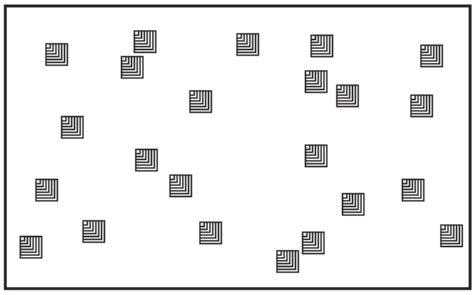

Figure 1.9. Random arrangement of nested sampling units increasing in size, facilitating space series analysis of data.

be held constant (Figure 1.9). Then, the effect of increasing plot sizes upon the data, the resemblance values, classifications and ordinations may be examined. In other words, the *scale-dependence* of results becomes analyzable and interpretable. Such a sampling strategy will allow data processing methods analogous to time series analysis, and may be called the *space series analysis* (see e.g., Podani 1984a, 1992). If we examine the vast literature of plant ecology, then we find that space series analysis is present, implicitly or explicitly, in several areas, such as diversity estimation (Pielou 1975), and is a fundamental strategy in pattern analyses of populations (Greig-Smith 1983) and in the evaluation of compositional (florula, faunula) diversity (Juhász-Nagy 1976 1984, Juhász-Nagy & Podani 1983). Space series analysis is not restricted to changes of sampling unit size; the other three characteristics of sampling may also be changed successively do define a series, as explained below.

Increase of *sample size* is the simplest series used to establish the sample size necessary to reach estimation at a given precision. This topic is treated in standard statistics: recall the well-known relationship between sample size and standard error. In addition, Orlóci (1991) and Orlóci & Pillar (1989) proposed changing sample size to stabilize resemblance and eigenstructures in ecological investigations. Decrease of sample size by eliminating less important variables is also useful to evaluate stability or invariance properties of multivariate analyses (e.g., Orlóci & Mukkattu 1973, Podani 1989d). The *arrangement* of sampling units in a continuous universe may also constitute a space series (Podani 1984a). The starting point is now the systematic arrangement from which we derive different stages of semi-systematic arrangements to end up with the fully randomized design (Figure 1.10). This can be accomplished effectively through computer simulation only (see Subsection 1.5.2 below). Stepwise elongation of *shape,* whilst the area of units remains fixed, is another possibility for space series analysis (Nosek 1976, Podani 1984b, Bartha & Horváth 1987).

I shall emphasize at several places in this book that space series analysis is not restricted to changes defined in the real, topographic space. This operation is also meaningful in pratically all kinds of abstract spaces associated with a multivariate survey, and is almost the only possible *modus operandi* whenever one wishes to evaluate the effect of unavoidable subjective decisions (related to data analysis methodology) upon the final results.

1.5.2 Computer simulated sampling

Completion of space series analysis, i.e., changing the sampling characteristics in the field, is a laborious work and cannot always be achieved. In fact, application of several combinations of sampling characteristics is impossible without completely destroying the plants. A partial solution to this problem is computer simulated sampling. Following the pioneering work by

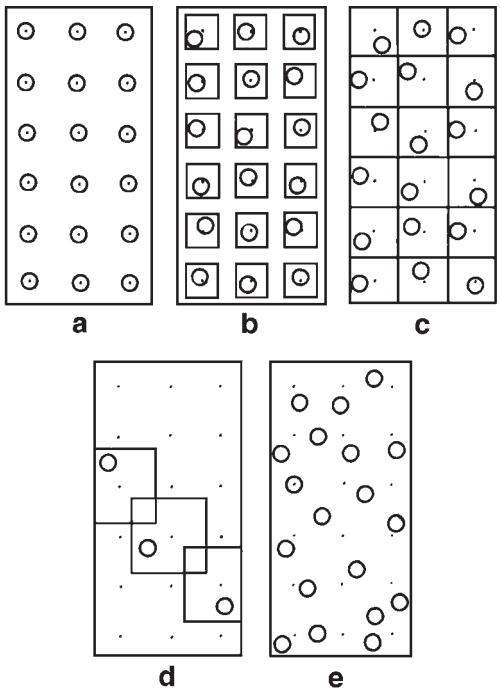

Figure 1.10. A possible series of arrangements of sampling units. **a**: systematic, **b**: semi-systematic with separate blocks, **c:** semi-systematic with contiguous blocks, **d**: transitional stage with partly overlapping blocks (only three shown), **e:** fully random arrangement with block size exceeding the study area (from Podani 1984a).

Palley & O'Regan (1961) and Arvanitis & O'Regan (1967) on forest parameter estimates, Sz•cs (1979) has founded the principles of sampling simulation in plant communities.

> The vegetation of the study area is transformed into a point pattern by photographic or other means. The data are then transferred into computer memory in the form of digitized coordinates. Another possibility is offered by a fine grid in each cell of which the presence of species is recorded, and these data are stored in the memory. The different sampling designs can be simulated by computer programs (e.g., **SYN-TAX 5**: Podani 1993, **MULTI-PATTERN**: Erdei & Tóthmérész 1993). The package developed by Arvanitis & Reich (1989) is useful for demonstration and educational purposes. A thorough review of the topic is presented in Podani (1987). This approach is especially straightforward for relatively small study areas, but memory limitations and great sampling effort do not allow its extension to large sites. The possibility of applying such methods to satellite data stored in pixel format needs to be investigated.

1.5.3 Resampling from the sample ('bootstrapping')

The origin of this term is the idiom 'pull yourself up by your own bootstraps', whose meaning becomes clear using a simple example taken from conventional statistics where the method was originally suggested (Efron 1982). Let us take a random sample of n elements from the universe and calculate some statistic, such as the mean or the variance. This statistic stands by itself, there is no basis for comparisons, for example. So, why not to take k random samples of n units from this sample with replacement and examine the k estimates thus obtained? (Replacement implies that the sample itself is considered as a representation of the universe in which each element is equally probable.) Resampling is done most easily by the computer, so that the bootstrapping method is a special case of computer simulated sampling. The statistic

calculated from such a sample is called the bootstrap estimate. Hundreds or thousands of such estimates can be used to draw an empirical frequency histogram in which the position of the original estimate can be identified. In this way, the potential bias of the statistic, its standard error, confidence interval and significance can be derived from a single sample (Manly 1991).

> Bootstrapping has been extensively used in evaluating and comparing results of multivariate analyses. Examples are found in Greenacre (1984), Knox (1989) and Knox & Peet (1989) for correspondence analysis, in Stauffer et al. (1985) for principal components analysis and in Pillar (1999) for clustering. The method is indispensable in cladistic analysis to quantify, for example, the influence of the choice of characters upon the resulting cladograms (Felsenstein 1985, Sanderson 1989, Hillis & Bull 1993, and many others).

1.6 Literature overview

The first impression one gets when examining dozens of books devoted to multivariate analysis is that the problems of data collection are ignored. Data are considered as given, and no space is 'wasted' on sampling theory. In the vast ecological literature, examples are Williams (1976), Legendre & Legendre (1983), Pielou (1984) and Digby & Kempton (1987). Other sources also disregard the details, and the subject matter is settled by giving some references that are thought to be important (e.g., Ludvig & Reynolds 1988, Jongman et al. 1987), or the topic is condensed into a short discussion (Orlóci 1978). There are more extensive presentations as well, but unfortunately they can be very misleading. For instance, Kershaw & Looney (1985) discuss the random arrangement, sample size, sampling unit size and shape in the context of estimation. This is correct as far as the population level problems treated in the book are concerned, but is completely irrelevant in a multivariate perspective. Recommendations such as 'the most suitable quadrat on theoretical grounds being the smallest possible, relative to the type of vegetation and to the practicability of the enumeration of such a quadrat size' (Kershaw & Looney 1985, p. 27) are difficult to adapt to actual surveys and irrelevant to pattern detection purposes. Greig-Smith's (1983) excellent book, published three times, also falls into this trap, even though the author himself points out in the first line of the chapter on sampling that 'the value of quantitative data ... depends on the sampling procedure used to obtain them.' Notwithstanding that the book devotes 144 pages to multivariate methods, the discussion on sampling is confined to estimation, measurement errors and precision. The author is aware of the problem, however, when he points out that 'the most useful procedure for sampling overall composition may not be the most satisfactory for examination of variation within the area' (p. 19). Actually, Greig-Smith was among the first who emphasized the importance of the compatibility between sampling and data analysis (Austin & Greig-Smith 1968).

Green (1979) also focuses attention on estimation and hypothesis testing, and the relationships between sampling and multivariate analysis receive much less emphasis (yet, the book is essentially 'multivariate in nature'). In the discussion of particular techniques, however, it becomes obvious that distinction needs to be made between the estimation and pattern detection objectives. The book is very useful, even though the topic of sampling is scattered over several chapters due to its peculiar structure.

Although the unbalanced treatment of sampling theory by Kershaw & Looney, Greig-Smith and Green is partly understandable, the book by Gauch (1982) is more disappointing, confining itself to generalities: earlier proposals are reproduced without criticism and without examining whether they are applicable in a multivariate context, the central topic of the book. Chapter 2 is a collection of useless statements such as 'one the whole, a rectangle that is two to four times as long as it is wide is ordinarily most accurate' or 'the number of samples' (i.e., sampling units)

'desired is affected by the accuracy of individual samples'. To suggest species/area curves in order to determine optimal quadrat size, as mentioned earlier, is simply misleading. Many other books, oriented towards phytosociology seem just as well confused over sampling problems and their relevance in multivariate studies. Knapp (1984) and Kent & Coker (1992) do not supersede the presentation by Gauch (1982), although the latter book devotes more than 120 pages to multivariate studies. The textbooks that consider the relevé method of phytosociology as the only possibility in vegetation surveys, thus proposing a largely preferential technique, need not even be mentioned here.

From a biostatistical viewpoint, Sampford (1962) and Cochran (1977) are the most reliable and comprehensive, but only if our objective is estimation. The proceedings edited by Cormack et al. (1979) contain 14 papers, all devoted to sampling procedures optimized for estimation purposes. Southwood (1978) can also be recommended to those dealing with estimating population parameters only.

Ecological, mostly hydrobiological, sampling is the subject matter of the book edited by Frontier (1983). The methods of selecting the sample are extensively treated with many examples, but – again – with emphasis on estimation and statistical hypothesis testing. No surprise then that the only chapter on multivariate methods considers the increase of precision as the most important requirement. Nevertheless, the book abounds in useful information and many details given are not discussed here.

The taxonomic literature, on the whole, is even more careless about sampling. Cole (1969), Dunn & Everitt (1982) and Stuessy (1990), for example, do not even mention the word 'sample', clearly showing that selection of the individuals for a taxonomic study completely relies on the personal, i.e., preferential, judgment of the investigator. Sneath & Sokal (1973), on the other hand, devotes more space to the selection of OTUs (operational taxonomic units) to be included in the study. To these authors the most important issues are: 1) how taxonomic resemblance is influenced by sampling? 2) how will an OTU represent a given taxon? The 'exemplar' method is considered of central importance; this assumes that it is sufficient to include a single specimen for each taxon in the study – if within-taxon variability is lower than among the taxa. Clearly, there is a vicious circle: if the objective of the study is the recognition of taxa, the exemplars cannot be selected in advance. The method works when the existence of taxa already recognized needs confirmation. There is at least one study (Moss 1968) examining whether taxonomic classification is affected by the choice of specimens. The conclusion in this case favoured the 'lazy' taxonomist: the choice is not very influential, but one should not draw far-reaching conclusions from this isolated survey.

In cladistic analysis, especially if based on molecular data (Sections 6.3-4), the phenomenon of within-taxon polymorphism requires careful determination of sample size, as empasized by Baverstock & Moritz (1990). The vicious circle mentioned above is eluded by a two-step strategy: the closely related taxa are evaluated first, then the geographically most isolated populations are compared to each cladistic line already defined. In this way, the relative magnitude of within-taxon genetic polymorphism and among taxa variation is estimated. The first situation requires larger sample sizes (Archie et al. 1989). If the pilot study reveals that among taxa variation is the highest, then the few replicates or even the exemplars will suffice. Contrary to Baverstock & Moritz's work, the majority of the cladistic literature do not discuss sampling (e.g., Duncan & Stuessy 1984, Forey et al. 1992).

As to the second topic of this chapter, data types, my review is more optimistic. This issue is apparently less critical, and a brief summarization is given in most books devoted to multivariate analysis. Nevertheless, there exists a terminological confusion over the meaning of

'quantitative', 'qualitative' and 'numerical'. It is always useful to stick to the scale types listed in Section 1.4. As mentioned above, the best discussion of scale types and their conversion is still in Anderberg (1973, pp. 26-69).

Although the selection of objects is mostly disregarded in taxonomy, the choice among characters that describe the taxa receives much more attention. Sneath & Sokal (1973, pp. 90-109, 147-157) presented an overview. which is still useful. Swofford & Olsen (1990, pp. 414-422) can be recommended as an introduction to the very specific character types of cladistics. Orlóci (1978, pp. 6-13) examines in detail the problem of selecting variables in the context of vegetation science.

1.7 Imaginary dialogue

Q: *I have the impression that you simply do not like preferential sampling, for example, the relevé method of phytosociology. What shall an investigator do, however, if he has already spent many years in the field and obtained the data relying largely on his own subjective decisions? Can anyone apply multivariate analysis methods if the (sometimes very) restrictive conditions on randomness were inapplicable for some reason?*

A: The answer is straightforward: data gathered through a preferential 'filter' of the biologist are extremely useful in their own right. Do not forget that an overwhelming proportion of biological knowledge has accumulated in such a way. We are lucky that the exploratory and summarization functions of multivariate analysis are at least partly independent on the circumstances of sampling. Interpretability of a classification is not affected even though the objects classified were selected subjectively. It is of course true that this classification is valid to these objects only, and cannot be generalized in any way. On the contrary, the traditional statistical methods oriented towards estimation and hypothesis testing are completely unreliable if the sampling is preferential.

Q: *If this is true, then why this vexation over sampling theory? Why should anyone bother with the characteristics of sampling when, forgetting the relatively few instances of hypothesis testing in multivariate analysis, most of the methods impose no specific constraints on the way the input data were collected?*

A: Admitting that a significant part of biological knowledge has derived from more or less biased observations and data collection methods, as well as the fortunate circumstance that multivariate methods are not dependent directly on sampling, does not mean that this topic should be forgotten. Investigators must answer at least two questions to themselves: 1) does the sampling strategy harmonize with the final objectives of the study? and 2) do they wish to generalize their conclusions or, alternatively, to accept the fact that the results will only be valid to a narrow range of their objects? Persons thinking these issues over seriously will never be too hasty in this very first, and very important phase of work.

Q: *The answer to the next question, I guess, will be related to the previous issues: can you pool sampling units taken by different persons, perhaps in different times, into one sample? Is it important that all units in a sample be of the same 'quality', size and shape?*

A: You sense correctly that multivariate methods do not have such assumptions: some result will be obtained even though the sample was collected by many people, under quite different

sampling conditions and so on. This result may or may not be meaningful, but one will never be able to estimate how this result was influenced by differences between samplers. Whenever possible, all sampling characteristics should remain the same throughout the entire study. Even phytosociological quadrats should be of constant size at the outset.

Q: *Oh yes, but it was you who pointed out in Subsection 1.3.3 that the 'optimum' size, i.e., the one best reflecting variation of pattern in the community may change over time during succession or degradation. So, the optimum may also be different for the community types that are to be recognized!*

A:Yes, you are completely right! This problem can be treated by *successive approximation*, a proposition attributed originally to Poore and 'revitalized' by Orlóci (1991). In this, space series analysis and subsequent data analysis are integral parts. For phytosociological quadrat size, the concept implies that the 'optimum' is applicable to a rough distinction of the types only. When this is done, we seek optima in each community separately and then revise the classification using these different sizes, find new types, and so on, until a stable result is obtained. Successive approximation is, in this case, a stepwise operation in which slight modifications to sampling or data analysis conditions produce the final result. I admit that this strategy requires much more effort than a conventional 'start-and-finish' approach.

Q: *You mention 'estimation' very frequently in this chapter, far too often emphasizing that this is of secondary importance in a multivariate study. Are you 100% sure?*

A: Estimation appears in fact at several stages of the study, so your concern is understandable. Estimation of the data values themselves is the first of these. For example, just recall plant cover in quadrats, which is never measured error-free. Any measurement of weight, length or concentration is also an estimation with a precision that depends on our tools. An investigation with estimation objectives was understood as one in which these estimated values serve as a basis for estimating some parameter of the statistical population (or universe), and the study ends or continues merely with hypothesis testing related to this parameter. A pattern analysis study just begins, however, at this point. It is true, on the other hand, that distances and similarities calculated from the data are also estimates and, by the same token, ordinations and classifications are also 'estimates' (in the widest sense of the word). The entire survey can be repeated giving you another 'estimate' of the ordination or classification you are looking for!

Q: *The chapter seems to suggest that all research plans concentrate upon the methodological sequence: data → resemblance → classification or ordination → evaluation of results. Is this always the case?*

A: Not at all, but this book will focus on problems that are usually characterized by this sequence. This is the main axis of the scheme of Figure 0.1. Some steps may be skipped, of course, as in molecular systematics, if distances are estimated directly without deriving any data (by DNA hybridization experiments, for example). Character-based cladistical methods (Chapter 6) depart most often from the above sequence. I can also imagine a situation when the observations directly produce a classification or ordination, so that evaluation of results remains the only computational task.