CHAPTER 1

Sampling

Sampling

Choosing a way to sample and collect data can be bewildering. If you find it hard to decide exactly how it should be done then seek help. Questions about sampling are among the questions that are most frequently asked to biometricians and the time to ask for assistance is while the sampling scheme is being designed. Remember: if you go wrong with data analysis it is easy to repeat it, but if you collect data in inappropriate ways you can probably not repeat it, and your research will not meet its objectives.

Although there are some particular methods that you can use for sampling, you will need to make some choices yourself. Sample design is the art of blending theoretical principles with practical realities. It is not possible to provide a catalogue of sampling designs for a series of situations – simply too much depends on the objectives of the survey and the realities in the field.

Sampling design has to be based on specific research objectives and the hypotheses that you want to test. When you are not clear about what it is that you want to find out, it is not possible to design an appropriate sampling scheme.

Research hypotheses

The only way to derive a sampling scheme is to base it on a specific research hypothesis or research objective. What is it that you want to find out? Will it help you or other researchers when you find out that the hypothesis holds true? Will the results of the study point to some management decisions that could be taken?

The research hypotheses should indicate the 3 basic types of information that characterize each piece of data: **where** the data were collected, **when** the data were collected, and **what** type of measurement was taken. The where, when and what are collected for each **sample unit**. A sample unit could be a sample plot in a forest, or a farm in a village. Some sample units are natural units such as fields, farms or forest gaps. Other sample units are subsamples of natural units such as a forest plot that is placed within a forest. Your **sampling scheme** will describe how sample units are defined and which ones are selected for measurement.

The objectives determine **what** data, the **variables** measured on each sampling unit. It is helpful to think of these as response and explanatory variables, as described in the chapter on data preparation. The response variables are the key quantities that your objectives refer to, for example 'tree species richness on small farms'. The explanatory variables are the variables that you expect, or hypothesize, to influence the response. For example, your hypothesis could be that 'tree species richness on small farms is influenced by the level of market integration of the farm enterprise *because* market integration determines which trees are planted and retained'. In this example, species richness is the response variable and level of market integration is an explanatory variable. The hypothesis refers to small farms, so these should be the study units. The 'because…' part of the hypothesis adds much value to the research, and investigating it requires additional information on whether species were planted or retained and why.

Note that **this manual only deals with survey data.** The only way of proving **cause-effect** relationships is by conducting well-designed **experiments** – something that would be rather hard for this example! It is common for ecologists to draw conclusions about causation from relationships founding surveys. This is dangerous, but inevitable when experimentation is not feasible. The risk of making erroneous conclusions is reduced by: (a) making sure other possible explanations have been controlled or allowed for; (b) having a mechanistic theory model that explains why the cause-effect may apply; and (c) finding the same relationship in many different studies. However, in the end the conclusion depends on the argument of the scientist rather than the logic of the research design. Ecology progresses by scientists finding new evidence to improve the inevitably incomplete understanding of cause and effect from earlier studies.

When data are collected is important, both to make sure different observations are comparable and because understanding change – trends, or before and after an intervention – is often part of the objective. Your particular study may not aim at investigating trends, but investigating changes over time may become the objective of a later study. Therefore you should also document when data were collected.

This chapter will mainly deal with **where** data are collected. This includes definition of the **survey area**, of the **size and shape of sample units and plots** and of **how sample plots are located** within the survey area.

Survey area

You need to make a clear statement of the survey area for which you want to test your hypothesis. The survey area should have explicit geographical (and temporal) boundaries. The survey area should be at the ecological scale of your research question. For example, if your research hypothesis is something like 'diversity of trees on farms decreases with distance from Mount Kenya Forest because seed dispersal from forest trees is larger than seed dispersal from farm trees', then it will not be meaningful to sample trees in a strip of 5 metres around the forest boundary and measure the distance of each tree from the forest edge. In this case we can obviously not expect to observe differences given the size of trees (even if we could determine the exact distance from the edge within the small strip). But if the 5 m strip is not a good survey area to study the hypothesis, which area is? You would have to decide that on the basis of other knowledge about seed dispersal, about other factors which dominate the process when you get too far from Mt Kenya forest, and on practical limitations of data collection. You should select the survey area where you expect to **observe the pattern** given the **ecological size** of the phenomenon that you are investigating.

If the research hypothesis was more general, for example 'diversity of trees on East African farms decreases with distance from forests because more seeds are dispersed from forest trees than from farm trees', then we will need a more complex strategy to investigate it. You will certainly have to study more than one forest to be able to conclude this is a general feature of forests, not just Mt Kenya forest. You will therefore have to face questions of what you mean by a 'forest'. The sampling strategy now needs to determine how forests are selected as well as how farms around each forest are sampled.

A common mistake is to restrict data collection to only part of the study area, but assume the results apply to all of it (see Figure 1.1). You can not be sure that the small window actually sampled is representative of the larger study area.

An important idea is that **bias** is avoided. Think of the case in which samples are only located in sites which are easily accessible. If accessibility is associated with diversity (for example because fewer trees are cut in areas that are more difficult to access), then the area that is sampled will not

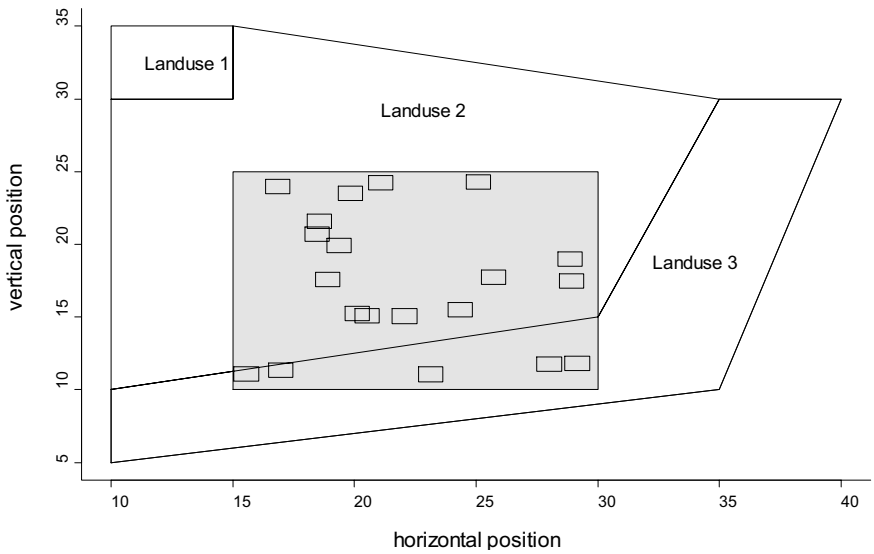

Figure 1.1 When you sample within a smaller window, you may not have sampled the entire range of conditions of your survey area. The sample may therefore not be representative of the entire survey area. The areas shown are three types of landuse and the sample window (with grey background). Sample plots are the small rectangles.

be representative of the entire survey area. An estimate of diversity based only on the accessible sites would give biased estimates of the whole study area. This will especially cause problems if the selection bias is correlated with the factors that you are investigating. For example, if the higher diversity next to the forest is caused by a larger proportion of areas that are difficult to access and you only sample areas that are easy to access, then you may not find evidence for a decreasing trend in diversity with distance from the forest. In this case, the dataset that you collected will generate estimates that are biased since the sites are not representative of the entire survey area, but only of sites that are easy to access.

The sample plots in Figure 1.1 were selected from a sampling window that covers part of the study area. They were selected using a method that allowed any possible plot to potentially be included. Furthermore, the selection was random. This means that inferences based on the data apply to the sampling window. Any particular sample will not give results (such as diversity, or its relationship with distance to forest) which are equal to those from measuring the whole sampling window. But the sampling will not predispose us to under- or overestimate the diversity, and statistical methods will generally allow us to determine just how far from the 'true' answer any result could be.

Size and shape of sample units or plots

A sample unit is the geographical area or plot on which you actually collected the data, and the time when you collected the data. For instance, a sample unit could be a 50 × 10 m^2 quadrat (a rectangular sample plot) in a forest sampled on 9th May 2002. Another sample unit could be all the land that is cultivated by a family, sampled on 10th December 2004. In some cases, the sample plot may be determined by the hypothesis directly. If you are interested in the influence of the wealth of farmers on the number of tree species on their farm, then you could opt to select the farm as the sample plot. Only in cases where the size of this sample plot is not practical would you need to search for an alternative sample plot. In the latter case you would probably use two sample units such as farms (on which you measure wealth) and plots within farms (on which you measure tree species, using the data from plots within a farm to estimate the number of species for the whole farm to relate to wealth).

The **size** of the quadrat will usually influence the results. You will normally find more species and more organisms in quadrats of 100 m^2 than in quadrats of 1 m^2. But 100 dispersed 1 m^2 plots will probably contain more species than a single 100 m^2 plot. If the aim is not to find species but understanding some ecological phenomenon, then either plot size may be appropriate, depending on the scale of the processes being studied.

The **shape** of the quadrat will often influence the results too. For example, it has been observed that more tree species are observed in rectangular quadrats than in square quadrats of the same area. The reason for this phenomenon is that tree species often occur in a clustered pattern, so that more trees of the same species will be observed in square quadrats. When quadrats are rectangular, then the orientation of the quadrat may also become an issue. Orienting the plots parallel or perpendicular to contour lines on sloping land may influence the results, for instance. As deciding whether trees that occur near the edge are inside or outside the sample plot is often difficult, some researchers find circular plots superior since the ratio of edge-to-area is smallest for circles. However marking out a circular plot can be much harder than marking a rectangular one. This is an example of the trade off between what may be theoretically optimal and what is practically best. Balancing the trade off is a matter of practical experience as well as familiarity with the principles.

As size and shape of the sample unit can influence results, it is best to stick to one size and shape for the quadrats within one study. If you want to compare the results with other surveys, then it will be easier if you used the same sizes and shapes of quadrats. Otherwise, you will need to convert results to a common size and shape of quadrat for comparisons. For some variables, such **conversion** can easily be done, but for some others this may be quite tricky. Species richness and diversity are statistics that are influenced by the size of the sample plot. Conversion is even more complicated since different methods can be used to measure sample size, such as area or the number of plants measured (see chapter on species richness). The average number of trees is easily converted to a common sample plot size, for example 1 ha, by multiplying by the appropriate scaling factor. This can not be done for number of species or diversity. Think carefully about conversion, and pay special attention to conversions for species richness and diversity. In some cases, you may not need to convert to a sample size other than the one you used – you may for instance be interested in the average species richness per farm and not in the average species richness in areas of 0.1 ha in farmland. Everything will depend on being clear on the research objectives.

One method that will allow you to do some easy conversions is to split your quadrat into sub-plots of smaller sizes. For example, if your quadrat is 40 × 5 m^2, then you could split this quadrat into eight

5×5 m^2 subplots and record data for each subplot. This procedure will allow you to easily convert to quadrat sizes of 5×5 m^2, 10×5 m^2, 20×5 m^2 and 40×5 m^2, which could make comparisons with other surveys easier.

Determining the size of the quadrat is one of the tricky parts of survey design. A quadrat should be large enough for differences related to the research hypothesis to become apparent. It should also not be too large to become inefficient in terms of cost, recording fatigue, or hours of daylight. As a general rule, several small quadrats will give more information than few large quadrats of the same total area, but will be more costly to identify and measure. Because differences need to be observed, but observation should also use resources efficiently, the type of organism that is being studied will influence the best size for the quadrat. The best size of the quadrat may differ between trees, ferns, mosses, butterflies, birds or large animals. For the same reason, the size of quadrat may differ between vegetation types. When studying trees, quadrat sizes in humid forests could be smaller than quadrat sizes in semi-arid environments.

As some rough indication of the size of the sample unit that you could use, some of the sample sizes that have been used in other surveys are provided next. Some surveys used 100×100 m^2 plots for differences in tree species composition of humid forests (Pyke et al. 2001, Condit et al. 2002), or for studies of forest fragmentation (Laurance et al. 1997). Other researchers used transects (sample plots with much longer length than width) such as 500×5 m^2 transects in western Amazonian forests for studies of differences in species composition for certain groups of species (Tuomisto et al. 2003). Yet other researchers developed methods for rapid inventory such as the method with variable subunits developed at CIFOR that has a maximum size of 40×40 m^2, but smaller sizes when tree densities are larger (Sheil et al. 2003). Many other quadrat sizes can be found in other references. It is clear that there is no common or standard sample size that is being used everywhere. The large range in values emphasizes our earlier point that there is no fixed answer to what the best sampling strategy is. It will depend on the hypotheses, the organisms, the vegetation type, available resources, and on the creativity of the researcher. In some cases, it may be worth using many small sample plots, whereas in other cases it may be better to use fewer larger sample plots. A pilot survey may help you in deciding what size and shape of sample plots to use for the rest of the survey (see below: pilot testing of the sampling protocol). Specific guidelines on the advantages and disadvantages of the various methods is beyond the scope of this chapter (an entire manual could be devoted to sampling issues alone) and the best advise is to consult a biometrician as well as ecologists who have done similar studies.

Simple random sampling

Once you have determined the survey area and the size of your sampling units, then the next question is where to take your samples. There are many different methods by which you can place the samples in your area.

Simple random sampling involves locating plots randomly in the study area. Figure 1.2 gives an example where the coordinates of every sample plot were generated by random numbers. In this method, we randomly selected a horizontal and vertical position. Both positions can be calculated by multiplying a random number between 0 and 1 with the range in positions (maximum – minimum), and adding the result to the minimum position. If the selected position falls outside the area (which is possible if the area is not rectangular), then a new position is selected.

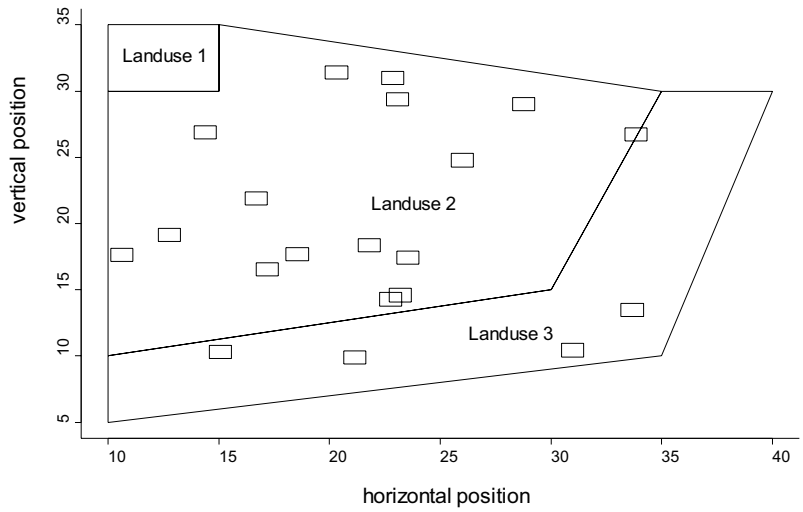

Figure 1.2 Simple random sampling by using random numbers to determine the position of the sample plots. Using this method there is a risk that regions of low area such as that under Landuse 1 are not sampled.

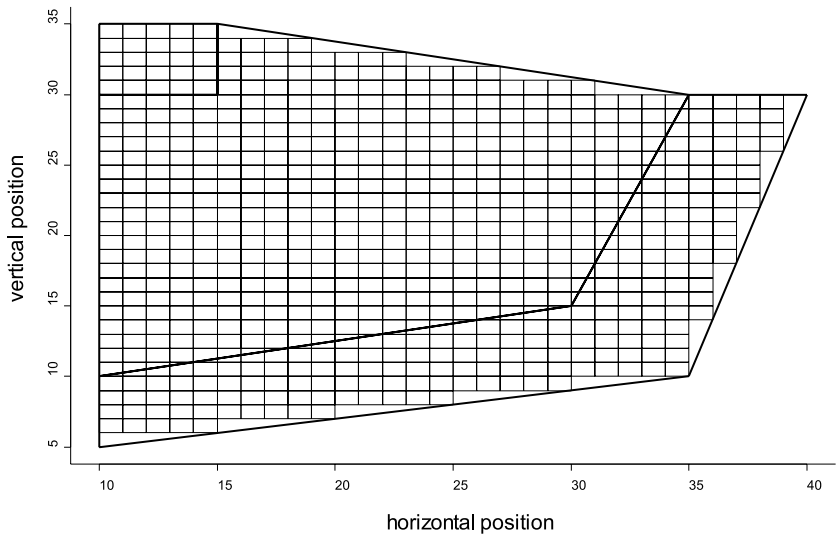

Figure 1.3 For simple random sampling, it is better to first generate a grid of plots that covers the entire area such as the grid shown here.

Simple random sampling is an easy method to select the sampling positions (it is easy to generate random numbers), but it may not be efficient in all cases. Although simple random sampling is the basis for all other sampling methods, it is rarely optimal for biodiversity surveys as described next. Simple random sampling may result in selecting all your samples within areas with the same environmental characteristics, so that you can not test your hypothesis efficiently. If you are testing a hypothesis about a relationship between diversity and landuse, then it is better to stratify by the type of landuse (see below: stratified sampling). You can see in Figure 1.2 that one type of landuse was missed by the random sampling procedure. A procedure that ensures that all types of landuse are included is better than repeating the random sampling procedure until you observe that all the types of landuse were included (which is not simple random sampling any longer).

It may also happen that the method of using random numbers to select the positions of quadrats will cause some of your sample units to be selected in positions that are very close to each other. In the example of Figure 1.2, two sample plots actually overlap. To avoid such problems, it is theoretically better to first generate the population of all the acceptable sample plots, and then take a simple random sample of those. When you use random numbers to generate the positions, the population of all possible sample plots is infinite, and this is not the best approach. It is therefore better to first generate a **grid** of plots that covers the entire survey area, and then select the sample plots at random from the grid.

Figure 1.3 shows the grid of plots from which all the sample plots can be selected. We made the choice to include only grid cells that fell completely into the area. Another option would be to include plots that included boundaries, and only sample the part of the grid cell that falls completely within the survey area – and other options also exist.

Once you have determined the grid, then it becomes relatively easy to randomly select sample plots from the grid, for example by giving all the plots on the grid a sequential number and then randomly selecting the required number of sample plots with a random number. Figure 1.4 shows an example of a random selection of sample plots from the grid. Note that although we avoided ending up with overlapping sample plots, some sample plots were adjacent to each other and one type of landuse was not sampled.

Note also that the difference between selecting points at random and gridding first will only be noticeable when the quadrat size is not negligible compared to the study area. A pragmatic solution to overlapping quadrats selected by simple random sampling of points would be to reject the second sample of the overlapping pair and choose another random location.

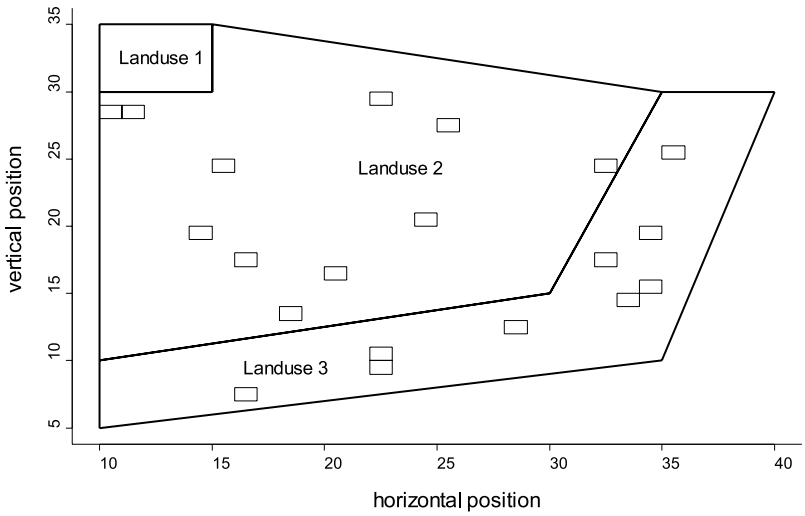

Figure 1.4. Simple random sampling from the grid shown in Figure 1.3.

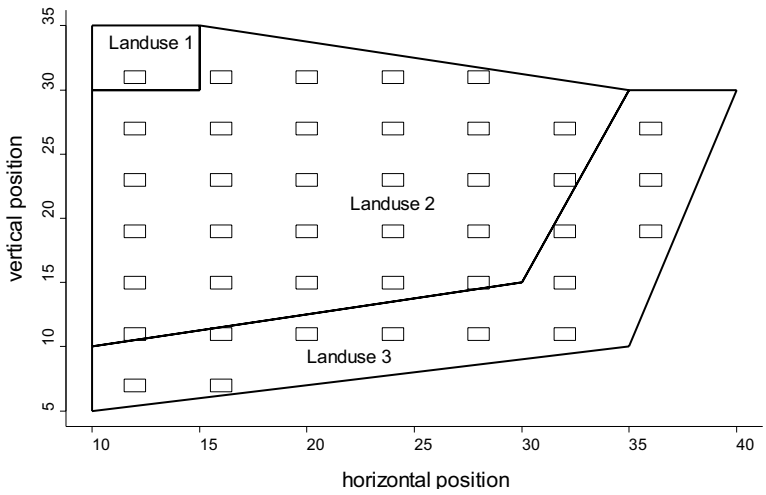

Figure 1.5 Systematic sampling ensures that data are collected from the entire survey area.

Systematic sampling

Systematic or regular sampling selects sample plots at regular intervals. Figure 1.5 provides an example. This has the effect of spreading the sample out evenly through the study area. A square or rectangular grid will also ensure that sample plots are evenly spaced.

Systematic sampling has the advantage over random sampling that it is easy to implement, that the entire area is sampled and that it avoids picking sample plots that are next to each other. The method may be especially useful for finding out where a variable undergoes rapid changes. This may particularly be interesting if you sample along an environmental gradient, such as altitude, rainfall or fertility gradients. For such problems systematic sampling is probably more efficient – but remember that we are not able in this chapter to provide a key to the best sampling method.

You could use the same grid depicted in Figure 1.5 for simple random sampling, rather than the complete set of plots in Figure 1.3. By using this approach, you can guarantee that sample plots will not be selected that are too close together. The grid allows you to control the minimum distance between plots. By selecting only a subset of sample plots from the entire grid, sampling effort is reduced. For some objectives, such combination of simple random sampling and regular sampling intervals will offer the best approach. Figure 1.6 shows a random selection of sample plots from the grid depicted in Figure 1.5.

If data from a systematic sample are analysed as if they came from a random sample, inferences may be invalidated by correlations between neigbouring observations. Some analyses of systematic samples will therefore require an explicitly spatial approach.

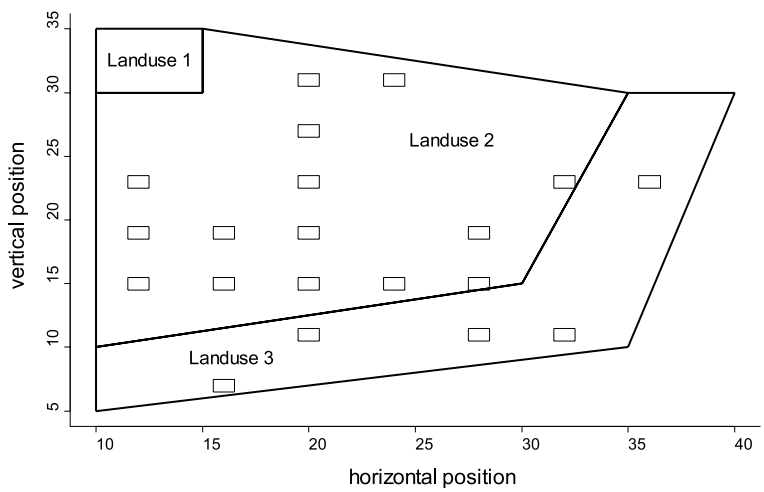

Figure 1.6 Random selection of sample plots from a grid. The same grid was used as in Figure 1.5.

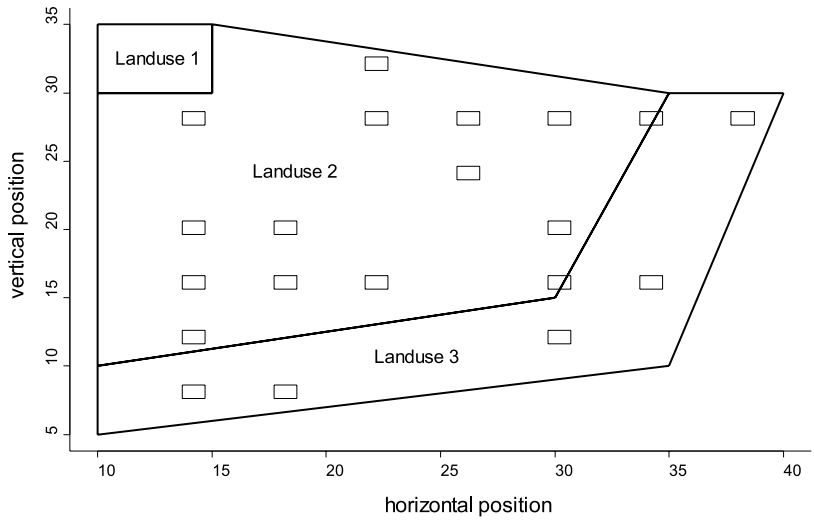

Figure 1.7. Systematic sampling after random selection of the position of the first sample plot.

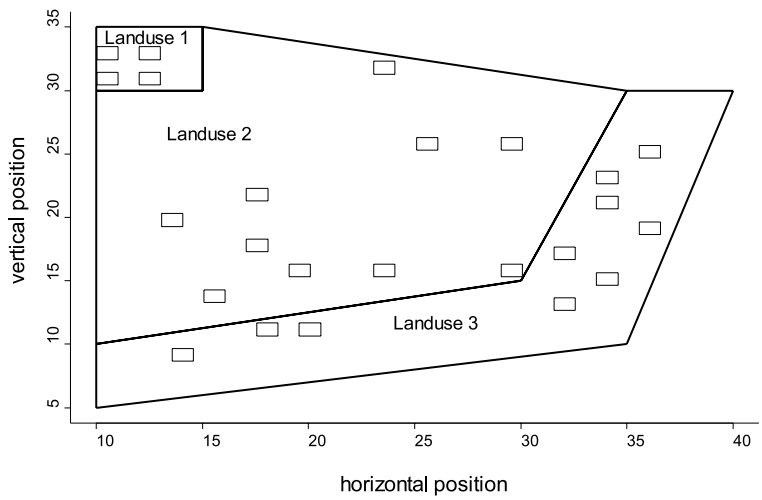

Figure 1.8 Stratified sampling ensures that observations are taken in each stratum. Sample plots are randomly selected for each landuse from a grid.

Another problem that could occur with systematic sampling is that the selected plots coincide with a periodic pattern in the study area. For example, you may only sample in valley bottoms, or you may never sample on boundaries of fields. You should definitely be alert for such patterns when you do the actual sampling. It will usually be obvious if a landscape can have such regular patterns.

Systematic sampling may involve no randomization in selecting sample plots. Some statistical analysis and inference methods are not then suitable. An element of randomization can be introduced in your systematic sampling by **selecting the position of the grid at random**. Figure 1.7 provides an example of selecting sample plots from a sampling grid with a random origin resulting in the same number of sample plots and the same minimum distance between sample plots as in Figure 1.6.

Stratified sampling

Stratified sampling is an approach in which the study area is subdivided into different **strata**, such as the three types of landuses of the example (Landuse 1, Landuse 2 and Landuse 3, figures 1.1-1.9). Strata do not overlap and cover the entire survey area. Within each stratum, a random or systematic sample can be taken. Any of the sampling approaches that were explained earlier can be used, with the only difference that the sampling approach will now be applied to each stratum instead of the entire survey area. Figure 1.8 gives an example of stratified random sampling with random selection of maximum 10 sample plots per stratum from a grid with random origin.

Stratified sampling ensures that data are collected from each stratum. The method will also ensure that enough data are collected from each stratum. If stratified sampling is not used, then a rare stratum could be missed or only provide one observation. If a stratum is very rare, you have a high chance of missing it in the sample. A stratum that only occupies 1% of the survey area will be missed in over 80% of simple random samples of size 20.

Stratified sampling also avoids sample plots being placed on the boundary between the strata so that part of the sample plot is in one stratum and another part is in another stratum. You could have noticed that some sample plots included the boundary between Landuse 3 and Landuse 2 in Figure 1.7. In Figure 1.8, the entire sample plot occurs within one type of landuse.

Stratified sampling can increase the precision of estimated quantities if the strata coincide with some major sources of variation in your area. By using stratified sampling, you will be more certain to have sampled across the variation in your survey area. For example, if you expect that species richness differs with soil type, then you better stratify by soil type.

Stratified sampling is especially useful when your research hypothesis can be described in terms of differences that occur between strata. For example, when your hypothesis is that landuse influences species richness, then you should stratify by landuse. This is the best method of obtaining observations for each category of landuse that will allow you to test the hypothesis.

Stratified sampling is not only useful for testing hypotheses with categorical explanatory variables, but also with **continuous** explanatory variables. Imagine that you wanted to investigate the influence of rainfall on species richness. If you took a simple random sample, then you would probably obtain many observations with near average rainfall and few towards the extremes of the rainfall range. A stratified approach could guarantee that you take plenty of observations at high and low rainfalls, making it easier to detect the influence of rainfall on species richness.

The main disadvantage of stratified sampling is that you need information about the distribution of the strata in your survey area. When this information is not available, then you may need

to do a survey first on the distribution of the strata. An alternative approach is to conduct systematic surveys, and then do some gap-filling afterwards (see below: dealing with covariates and confounding).

A modification of stratified sampling is to use **gradient-oriented transects** or **gradsects** (Gillison and Brewer 1985; Wessels et al. 1998). These are transects (sample plots arranged on a line) that are positioned in a way that steep gradients are sampled. In the example of Figure 1.8, you could place gradsects in directions that ensure that the three landuse categories are included. The advantage of gradsects is that travelling time (cost) can be minimized, but the results may not represent the whole study area well.

Sample size or the number of sample units

Choosing the sample size, the number of sampling units to select and measure, is a key part of planning a survey. If you do not pay attention to this then you run two risks. You may collect far more data than needed to meet your objectives, wasting time and money. Alternatively, and far more common, you may not have enough information to meet your objectives, and your research is inconclusive. Rarely is it possible to determine the exact sample size required, but some attempt at rational choice should be made.

We can see that the sample size required must depend on a number of things. It will depend on the complexity of the objectives – it must take more data to unravel the complex relationships between several response and explanatory variables than it takes to simply compare the mean of two groups. It will depend on the variability of the response being studied – if every sample unit was the same we only need to measure one to have all the information! It will also depend on how precisely you need to know answers – getting a good estimate of a small difference between two strata will require more data than finding out if they are roughly the same.

If the study is going to compare different strata or conditions then clearly we need observations in each stratum, or representing each set of conditions. We then need to plan for repeated observations within a stratum or set of conditions for four main reasons:

1. In any analysis we need to give some indication of the precision of results and this will depend on variances. Hence we need enough observations to estimate relevant variances well.
2. In any analysis, a result estimated from more data will be more precise than one estimated from less data. We can increase precision of results by increasing the number of relevant observations. Hence we need enough observations to get sufficient precision.
3. We need some 'insurance' observations, so that the study still produces results when unexpected things happen, for example some sample units can not be measured or we realize we will have to account for some additional explanatory variables.
4. We need sufficient observations to properly represent the study area, so that results we hope to apply to the whole area really do have support from all the conditions found in the area.

Of these four, 1 and 2 can be quantified in some simple situations. It is worth doing this quantification, even roughly, to make sure that your sample size is at least of the right order of magnitude.

The first, 1, is straightforward. If you can identify the variances you need to know about, then make sure you have enough observations to estimate each. How well you estimate a variance is determined by its degrees of freedom (*df*), and a minimum of 10 *df* is a good working rule. Get help finding the degrees of freedom for your sample design and planned analysis.

The second is also straightforward in simple cases. Often an analysis reduces to comparing means between groups or strata. If it does, then the

mathematical relationship between the number of observations, the variance of the population sampled and the precision of the mean can be exploited. Two approaches are used. You can either specify how well you want a difference in means to be estimated (for example by specifying the width of its confidence interval), or you can think of the hypothesis test of no difference. The former tends to be more useful in applied research, when we are more interested in the size of the difference than simply whether one exists or not. The necessary formulae are encoded in some software products.

An example from R is shown immediately below, providing the number of sample units (n) that will provide evidence for a difference between two strata for given significance and power of the t-test that will be used to test for differences, and given standard deviation and difference between the means. The formulae calculated a fractional number of 16.71 sample units, whereas it is not possible in practice to take 16.71 sample units per group. The calculated fractional number could be rounded up to 17 or 20 sample units. We recommend interpreting the calculated sample size in relative terms, and concluding that 20 samples will probably be enough whereas 100 samples would be too many.

```
Two-sample t test power calculation

              n = 16.71477
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

Sample size in each stratum

A common question is whether the survey should have the same number of observations in each stratum. The correct answer is once again that it all depends. A survey with the same number of observations per stratum will be optimal if the objective is to compare the different strata and if you do not have additional information or hypotheses on other sources of variation. In many other cases, it will not be necessary or practical to ensure that each stratum has the same number of observations.

An alternative that is sometimes useful is to make the number of observations per stratum proportional to the size of the stratum, in our case its area. For example, if the survey area is stratified by landuse and one category of landuse occupies 60% of the total area, then it gets 60% of sample plots. For the examples of sampling given in the figures, landuse 1 occupies 3.6% of the total area (25/687.5), landuse 2 occupies 63.6% (437.5/687.5) and landuse 3 occupies 32.7% (225/687.5). A possible proportional sampling scheme would therefore be to sample 4 plots in Landuse 1, 64 plots in Landuse 2 and 33 plots in Landuse 3.

One advantage of taking sample sizes proportional to stratum sizes is that the average for the entire survey area will be the average of all the sample plots. The sampling is described as **self-weighting**. If you took equal sample size in each stratum and needed to estimate an average for the whole area, you would need to weight each observation by the area of each stratum to arrive at the average of the entire area. The calculations are not very complicated, however.

Some researchers have suggested that taking larger sample sizes in larger strata usually results in capturing more biodiversity. This need not be the case, for example if one landuse which happens to occupy a small area contains much of the diversity. However, most interesting research objectives require more than simply finding the diversity. If the objective is to find as many species as possible, some different sampling schemes could be more effective. It may be better to use an adaptive method where the position of new samples is guided by the results from previous samples.

Simple random sampling will, in the long run, give samples sizes in each stratum proportional to the stratum areas. However this may not happen in any particular selected sample. Furthermore, the strata are often of interest in their own right, and more equal sample sizes per stratum may be more appropriate, as explained earlier. For these reasons it is almost always worth choosing strata and their sample sizes, rather than relying on simple random sampling.

Dealing with covariates and confounding

We indicated at the beginning of this chapter that it is difficult to make conclusions about cause-effect relationships in surveys. The reason that this is difficult is that there may be confounding variables. For example, categories of landuse could be correlated with a gradient in rainfall. If you find differences in species richness in different landuses it is then difficult or impossible to determine whether species richness is influenced by rainfall or by landuse, or both. Landuse and rainfall are said to be confounded.

The solution in such cases is to attempt to break the strong correlation. In the example where landuse is correlated with rainfall, then you could attempt to include some sample plots that have another combination of landuse and rainfall. For example, if most forests have high rainfall and grasslands have low rainfall, you may be able to find some low rainfall forests and high rainfall grasslands to include in the sample. An appropriate sampling scheme would then be to stratify by combinations of both rainfall and landuse (e.g. forest with high, medium or low rainfall or grassland with high, medium or low rainfall) and take a sample from each stratum. If there simply are no high rainfall grasslands or low rainfall forests then accept that it is not possible to understand the separate effects of rainfall and landuse, and modify the objectives accordingly.

An extreme method of breaking confounding is to **match sample plots**. Figure 1.9 gives an example.

The assumption of matching is that confounding variables will have very similar values for paired sample plots. The effects from the confounding variables will thus be filtered from the analysis.

The disadvantage of matching is that you will primarily sample along the edges of categories. You will not obtain a clear picture of the overall biodiversity of a landscape. Remember, however, that matching is an approach that specifically investigates a certain hypothesis.

You could add some observations in the middle of each stratum to check whether sample plots at the edges are very different from sample plots at the edge. Again, it will depend on your hypothesis whether you are interested in finding this out.

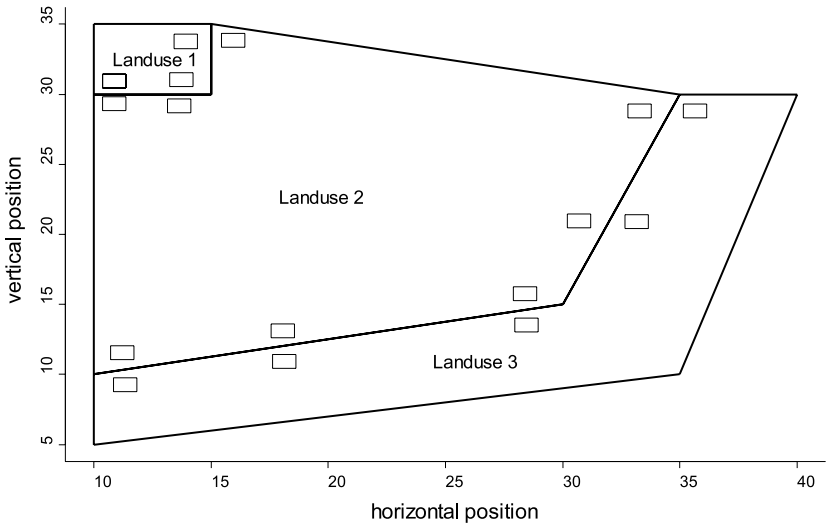

Figure 1.9 Matching of sample plots breaks confounding of other variables.

Pilot testing of the sampling protocol

The best method of choosing the size and shape of your sample unit is to start with a **pilot phase** in your project. During the pilot phase all aspects of the data collection are tested and some preliminary data are obtained.

You can evaluate your sampling protocol after the pilot phase. You can see how much variation there is, and base some modifications on this variation. You could calculate the required sample sizes again. You could also opt to modify the shape, size or selection of sample plots.

You will also get an idea of the time data collection takes per sample unit. Most importantly, you could make a better estimation of whether you will be able to test your hypothesis, or not, by already conducting the analysis with the data that you already have.

Pilot testing is also important for finding out all the non-statistical aspects of survey design and management. These aspects typically also have an important effect on the overall quality of the data that you collect.

References

Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, Foster RB, Nuñez P, Aguilar S, Valencia R, Villa G, Muller-Landau HC, Losos E, and Hubbell SP. 2002. Beta-diversity in tropical forest trees. *Science* 295: 666–669.

Feinsinger P. 2001. *Designing field studies for biodiversity conservation.* Washington: The Nature Conservancy.

Gillison AN and Brewer KRW. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management* 20: 103-127.

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates. (recommended as first priority for reading)

Hayek LAC and Buzas MA. 1997. *Surveying natural populations.* New York: Columbia University Press.

Laurance WF, Laurance SG, Ferreira LV, Rankin-de Merona JM, Gascon C and Loverjoy TE. 1997. Biomass collapse in Amazonian forest fragments. *Science* 278: 1117-1118.

Pyke CR, Condit R, Aguilar S and Lao S. 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science* 12: 553-566.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press.

Sheil D, Ducey MJ, Sidiyasa K and Samsoedin I. 2003. A new type of sample unit for the efficient assessment of diverse tree communities in complex forest landscapes. *Journal of Tropical Forest Science* 15: 117-135.

Sutherland WJ. 1996. *Ecological census techniques: a handbook.* Cambridge: Cambridge University Press.

Tuomisto H, Ruokolainen K and Yli-Halla M. 2003. Dispersal, environment and floristic variation of western Amazonian forests. *Science* 299: 241-244.

Underwood AJ. 1997. *Experiments in ecology: their logical design and interpretation using analysis of variance.* Cambridge: Cambridge University Press.

Wessels KJ, Van Jaarsveld AS, Grimbeek JD and Van der Linde MJ. 1998. An evaluation of the gradsect biological survey method. *Biodiversity and Conservation* 7: 1093-1121.

Examples of the analysis with the command options of Biodiversity.R

See in chapter 3 how Biodiversity.R can be loaded onto your computer.

To load polygons with the research areas:

```
area <- array(c(10,10,15,35,40,35,5,35,35,30,30,10),
 dim=c(6,2))

landuse1 <- array(c(10,10,15,15,30,35,35,30), dim=c(4,2))

landuse2 <- array(c(10,10,15,15,35,30,10,30,30,35,30,15),
 dim=c(6,2))

landuse3 <- array(c(10,10,30,35,40,35,5,10,15,30,30,10),
 dim=c(6,2))

window <- array(c(15,15,30,30,10,25,25,10),dim=c(4,2))
```

To plot the research area:

```
plot(area[,1], area[,2], type="n", xlab="horizontal position",
 ylab="vertical position", lwd=2, bty="l")

polygon(landuse1)

polygon(landuse2)

polygon(landuse3)
```

To randomly select sample plots in a window:

```
spatialsample(window, method="random", n=20, xwidth=1,
 ywidth=1, plotit=T, plothull=T)
```

To randomly select sample plots in the survey area:

```
spatialsample(area, method="random", n=20, xwidth=1, ywidth=1,
 plotit=T, plothull=F)
```

To select sample plots on a grid:

```
spatialsample(area, method="grid", xwidth=1, ywidth=1,
 plotit=T, xleft=10.5, ylower=5.5, xdist=1, ydist=1)

spatialsample(area, method="grid", xwidth=1, ywidth=1,
 plotit=T, xleft=12, ylower=7, xdist=4, ydist=4)
```

To randomly select sample plots from a grid:

```
spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xleft=10.5, ylower=5.5, xdist=1, ydist=1)

spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xleft=12, ylower=7, xdist=4, ydist=4)
```

To select sample plots from a grid with random start:

```
spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xdist=4, ydist=4)
```

To randomly select maximum 10 sample plots from each type of landuse:

```
spatialsample(landuse1, n=10, method="random", plotit=T)

spatialsample(landuse2, n=10, method="random", plotit=T)

spatialsample(landuse3, n=10, method="random", plotit=T)
```

To randomly select sample plots from a grid within each type of landuse. Within each landuse, the grid has a random starting position:

```
spatialsample(landuse1, n=10, method="random grid", xdist=2,
 ydist=2, plotit=T)

spatialsample(landuse2, n=10, method="random grid", xdist=4,
 ydist=4, plotit=T)

spatialsample(landuse3, n=10, method="random grid", xdist=4,
 ydist=4, plotit=T)
```

To calculate sample size requirements:

```
power.t.test(n=NULL, delta=1, sd=1, sig.level=0.05, power=0.8,
 type="two.sample")

power.t.test(n=NULL, delta=0.5, sd=1, sig.level=0.05,
 power=0.8, type="two.sample")

power.anova.test(n=NULL, groups=4, between.var=1, within.var=1,
 power=0.8)

power.anova.test(n=NULL, groups=4, between.var=2, within.var=1,
 power=0.8)
```

To calculate the area of a polygon:

```
areapl(landuse1)
```