Distance sampling: estimating animal density

All too often we hear in the news statements like "there are, at most, a few hundred individuals left of this endangered species" or "there is small hope for the persistence of this population given that so few are left". How do scientists count animals to make such statements? **Tiago Marques** explains the concepts and pitfalls of distance sampling—one of the most widely used methods for estimating animal populations.

Counting critters

Ideally, if feasible, we would count all the animals from a population of interest. However, not much imagination is needed to realise that it is not that easy to count all the polar bears in the Arctic or all the blue whales in the Pacific Ocean. The areas involved are large, animals move around and, even if it were possible, no one would be willing to pay the bill. These are extreme examples, but they illustrate the fact that, most often, one has to rely on some sampling approach to the problem. That is, based on observations on a fraction, sometimes quite small, of the population of interest, usually covering only a subset of the entire region of interest, we wish to draw inferences about the total population. Note that even for geographically restricted scenarios, such as, for example, puffins nesting in the Isle of May, counting them all is not an option.

Plot sampling

There are many possible methods that allow us to estimate animal abundance or density based on sampling. Arguably one of the most widely used is distance sampling[1]. But to introduce the ideas we first consider a more standard method known as strip or plot sampling: a number of plots, with total area a, are randomly allocated in a region of size A, and the total number n of animals within these plots recorded. Density in the covered area is, by definition, the number of animals per unit area,

$$D = \frac{n}{a},$$

and since a random design was used this is also a density estimate valid for the wider region. An estimate of abundance in the wider region we are interested in is obtained by multiplying the density estimate by the area of the region, that is,

$$\hat{N}_A = A\frac{n}{a},$$

(the use of $\hat{}$ represents an estimator of a given quantity rather than the quantity itself). The key assumption is that all animals within the plots selected for sampling are detected.

Distance sampling

The fact that puts us into the realms of distance sampling is to recognise that even having selected a sample of areas to cover, one usually cannot count all the animals within those areas. So we need to account for the proportion of the animals that were in the sampling units but which we failed to detect. But once we have done that, and hence estimated density in the covered area, extrapolation to the wider region is done based on the design properties—just as for plot sampling.

The idea is simple. Using a random design we allocate a number of sampling units, called transects, which cover a proportion of the survey region over which the population of interest exists. Transects can be either line transects or point transects, which correspond to rectangular and circle shaped survey plots, respectively. These transects are then surveyed. This is done by having an observer either moving along the line, in the case of line transects, or standing at the point for a short period of time, in the case of point transects, and recording the distances to all the detected animals. More specifically, it is the perpendicular distance from the line in the case of lines and the radial or animal-to-observer distance in the case of points that he measures. We focus here on line transects for simplicity but the ideas apply to point transects once the different geometry is taken into account. The key issue is to estimate the probability of detecting an animal given that the animal is in the area covered by the transects. This quantity is usually referred to as the detection probability and is denoted by P. Why is P important to estimate the abundance in the covered area? A simple example makes it clear. Let us say that we walk across a large patch of forest and we see eight bears. You then meet the biologist that works in the area and he says "From my experience, on average, I can only detect a third of the bears that are there". Then, intuitively, you know that there were probably around 24 bears in total and you missed around 16 (at least if you are as good as the biologist at detecting bears, which might not be the case, as observer experience is usually important!). Distance sampling is a way that allows you to formally estimate P. Then, an abundance estimate for the covered region of area a is given by

$$\hat{N}_A = \frac{n}{\hat{P}}.$$

To turn that into a density estimate you just divide the abundance estimate by a. A more elaborate example follows.

A motivating example

Imagine we have a density of 0.02 tortoises per m^2, corresponding to 20 000 tortoises in a square region of side 1000 m (I know there are too many tortoises for it to be plausible but bear with me for the sake of illustration!). The location of these tortoises is shown in Figure 1. The middle panel of figure 1 displays a single systematic random sample from 100 line transects, each 40 m long, surveyed, with the six detected tortoises shown as black dots. Tortoises further than w = 10 m away from the transect were ignored, leading to a covered area of 800 m^2 per transect (since our shaded area is 20 m × 40 m). In distance sampling we call w the truncation distance. The perpendicular distances smaller than w to all tortoises detected from the line were recorded. Pooled over the 100 transects, 1181 tortoises were detected. Although in real studies we do not have access to that information, here we know there were 1600 animals in the area covered by transects. The histogram in the rightmost panel of Figure 1 represents the perpendicular distances from all these to the nearest transect. Overlaid on this histogram, with shaded bars, are the distances to the detected animals. We can see that animals further away from the line are harder to see, which makes sense. From this figure we can get the notion that about 25%

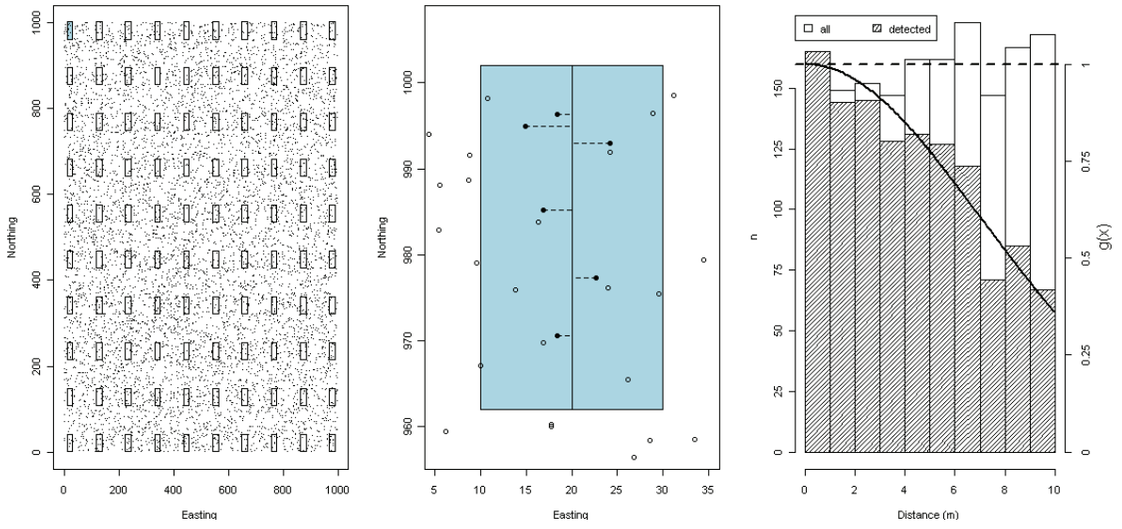

Figure 1. Distance sampling details. Left: survey area, with simulated tortoises shown as dots and areas covered as transects as rectangles. Middle: the top left transect from the left panels is magnified, with the distances to the detected tortoises shown as dashed lines. Right: histogram of the distances to all animals in the covered areas, as well as the distances to the detected animals. The detection function used to simulate the detections is shown as a solid line and the dashed line represents the average number of detections per bin if detection was certain at all distances (i.e. the total number of animals per bin)

of the tortoises in the covered areas are missed and that 75% of the tortoises are detected. This corresponds to the quotient of the area under the solid line, the detection function, by the area under the dashed line. A density estimate would therefore be around

$$\frac{1181}{0.75}/(100\times 20\times 40)=0.02$$

tortoises per m^2. In the histogram of Figure 1, the solid line represents the true detection function, used to simulate the detected animals. In practice, we use the distances to the detected animals, together with some assumptions (see details below), to estimate the detection probability and hence abundance or density. In the next section I describe, somewhat more formally, how this is done.

Estimating the detection function or a probability density function

Although it might be simpler to think about the function we are estimating as the detection function, in practice that is not how we implement the methods. The detection function represents a probability and, hence, is strictly non-negative. Whatever strictly non-negative function you might have, you get a probability density function (PDF) by dividing it by the integral of the said function over the relevant domain. Therefore we can rescale the detection function so that it becomes a PDF. Further, although full details can be found elsewhere[1], it can be shown that $P = 1/f(0)w$, where $f(0)$ represents the pdf of the detected distances, evaluated at distance 0. Why is this relevant? Because estimating PDFs is a job for which many tools exist and so we can take advantage of all these to deal with our problem. The standard approach, implemented in the widely used software for analysing distance sampling data, Distance[2], is to consider some parametric key functions, such as the half-normal, the hazard-rate or the uniform, and add to these some cosine or polynomial adjustment terms to improve the model fit to the data. Parameter estimates are obtained by maximum likelihood methods. Using standard model selection tools we can find a model that fits our data well and then proceed to estimate density by the standard formula

$$\hat{D} = \frac{n\hat{f}(0)}{2L}.$$

Bias and assumptions

A density estimate is only useful if it is unbiased, at least asymptotically, and so, in the long run, we hope to be, on average, about right. For distance sampling estimates to be approximately unbiased, we require a number of assumptions to hold, for which we list here the most important ones: (i) the probability of detecting an animal on the transect is 1; (ii) the survey can be seen as a snapshot in time (in practice it is enough that animal speed is small compared to observer speed); (iii) distances are measured without errors. For the methods to work well we also require that a large number of transects are used, and that the location of these is random and independent of the animals' locations. Failure of these assumptions will lead to bias, and therefore we want to use field methods that minimise their failure (see Buckland *et al.*[1] for further details). The basic methods can become considerably more complex as extensions are introduced to deal with the failure of these assumptions or to address particular aspects of different surveys, such as dealing with populations that occur in clusters or using covariates other than distance in modelling the detection function.

Variance

I conclude this toolkit by returning to the taster example—with a warning. Although the numbers reported in the news might be interesting, they are often hard to interpret because the variance associated with them is absent. Even unbiased density and abundance estimates without the corresponding variance or confidence interval, i.e. some measure about their precision, are effectively useless. The variance in an estimate measures how much one could expect things to vary if we repeated the same survey again. The confidence interval conveys the same message, but in the form of plausible upper and lower bounds for the estimated quantity. Therefore, always wonder...

If you read that a researcher has estimated that there are 1000 individuals left in an endangered species, what does that mean? Somewhere between 950 and 1050, or somewhere between 200 and 2000? The true relevance of the reported number lies in the answer to this question.

References

1. Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L. and Thomas, L. (2001) *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford: Oxford University Press.

2. Thomas, L., Laake, J. L., Strindberg, S., Marques, F. F. C., Buckland, S. T., Borchers, D. L., Anderson, D. R., Burnham, K. P., Hedley, S. L., Pollard, J. H., Bishop, J. R. B. and Marques, T. A. (2005) Distance 5.0, beta 4. Research Unit for Wildlife Population Assessment, University of St Andrews, UK.

Tiago Marques is at the Centre for Research into Ecological and Environmental Modelling at the University of St Andrews.