2

Assumptions and modelling philosophy

2.1 Assumptions

This section provides material for a deeper understanding of the assumptions required for the successful application of distance sampling theory. The validity of the assumptions allows the investigator assurance that valid inference can be made concerning the density of the population sampled. The existing theory covers a very broad application area and makes it difficult to present a simple list of all the assumptions that are generally true for all applications. Three primary assumptions are emphasized, but first two initial conditions are mentioned.

First, it is assumed that a population comprises objects of interest that are distributed in the area to be sampled according to some stochastic process with rate parameter D (= expected number per unit area). In particular, it is not necessary (in any practically significant way) that the objects be randomly (i.e. Poisson) distributed, although this is mistakenly given in several places in the literature. Rather, it is critical that the lines or points be placed randomly **with respect to the distribution of objects**. Random line or point placement justifies the extrapolation of the sample statistics to the population of interest. The area to be sampled must be defined, but its size need not be measured if only object density (rather than abundance) is to be estimated. Further, the observer must be able to recognize and correctly identify the objects of interest. This requirement seems almost trite, but in rich avian communities, the problem can be substantial. The distances from the line or point to the identified objects must be measured without bias.

Second, the design and conduct of the survey must pay due regard to good survey practice, as outlined in Chapter 7. If the survey is poorly designed or executed, the estimates may be of little value. Sound theory and analysis procedures cannot change this.

Three assumptions are critical to achieving reliable estimates of density from line or point transect sampling. These assumptions are given roughly in order of importance from most to least critical. The effects of partial failure of these assumptions and corresponding theoretical extensions are covered at length in later sections. All three assumptions can be relaxed under certain circumstances.

2.1.1 Assumption 1: objects on the line or point are detected with certainty

It is assumed that all objects at zero distance are detected, that is $g(0) = 1$. In practice, detection on or near the line or point should be nearly certain. Design of surveys must fully consider ways to assure that this assumption is met; its importance cannot be overemphasized.

It is sometimes possible to obtain an independent estimate of the probability of detection on the centreline in a line transect survey, for example by assigning two (or more) independent observers to each leg of search effort. Chapter 6 summarizes methods which have been developed for estimating $g(0)$, and Chapter 3 shows how the estimate can be incorporated in the estimation of density. It is important to note that $g(0)$ cannot be estimated from the distances y_i alone, and attempts to estimate $g(0)$ with low bias or adequate precision when it is known to be less than unity have seldom been successful. This issue should be addressed during the design of surveys, so that observation protocol will assure that $g(0) = 1$ or that a procedure for estimating $g(0)$ is incorporated into the design.

In fact, the theory can be generalized such that density can be computed if the value of $g(y)$ is known for **some** value of y. However, this result is of little practical significance in biological sampling unless an assumption that $g(y) = 1$ for some $y > 0$ is made (Quang and Lanctot 1991).

If objects on or near the line or point are missed, the estimate will be biased low (i.e. $E(\hat{D}) < D$). The bias is a simple function of $g(0)$: $E(\hat{D}) - D = -[1 - g(0)] \cdot D$, which is zero (unbiased) when $g(0) = 1$. Many things can be done in the field to help ensure that $g(0) = 1$. For example, video cameras have been used in aerial and underwater surveys to allow a check of objects on or very near the line; the video can be monitored after completion of the field survey. Trained dogs have been used in ground surveys to aid in detection of grouse close to the line.

Although we stress that every effort should be made to ensure $g(0) = 1$, the practice of 'guarding the centreline' during shipboard or aerial line transect surveys can be counterproductive. For example, suppose that most search effort is carried out using 20 × or 25 × tripod-

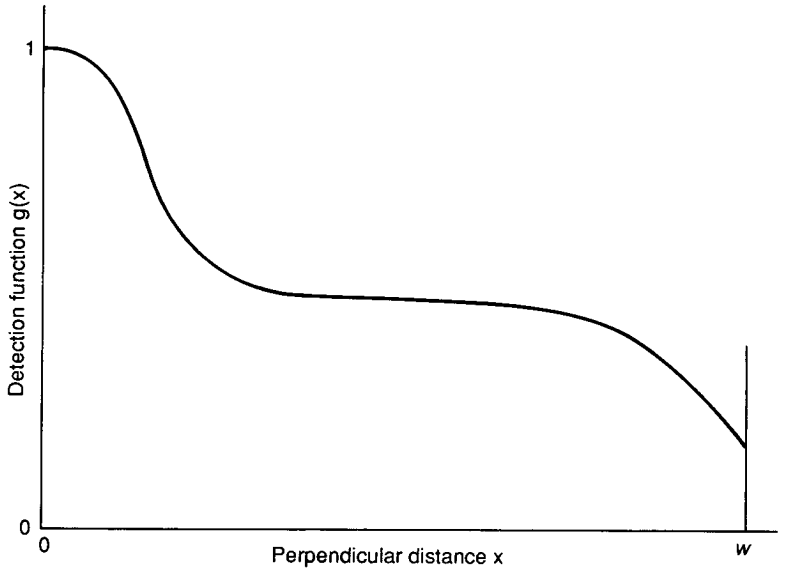

Fig. 2.1. Hypothetical detection function illustrating the danger of assigning an observer to 'guard the centreline'. This problem is most common in shipboard and aircraft surveys involving more than one observer.

mounted binoculars on a ship, but an observer is assigned to search with the naked eye, to ensure animals very close to the ship are not missed. If $g(0)$ in the absence of this observer is appreciably below 1, then the detection function may be as illustrated in Fig. 2.1. This function violates the **shape criterion** described later, and no line transect model can reliably estimate density in this case. The problem may be exacerbated if animals are attracted to the ship; the observer guarding the centreline may only detect animals as they move in toward the bow. Polacheck and Smith (unpublished) argued that if effort is concentrated close to the centreline, large bias can arise. Thus, field procedures should ensure both that $g(0) = 1$ and that the detection function does not fall steeply with distance from the line or point.

2.1.2 Assumption 2: objects are detected at their initial location

In studies of mobile animals, it is possible that an animal moves from its original location for some distance prior to being detected. The

measured distance is then from the random line or point to the location of the detection, not the animal's original location. If such undetected movements prior to detection were random (see Yapp 1956), no serious problem would result, provided that the animal's movement is slow relative to the speed of the observer. If movement is not slow, its effect must be modelled (Schweder 1977), or field procedures must be modified (Section 7.6). However, movement may be in response to the observer. If the movement is away from the transect line being traversed by the observer, the density estimator is biased low, whereas if the movement is toward the observer (e.g. some songbirds and marine mammals), the estimator of density will be biased high. Substantial movement away from the observer can often be detected in a histogram of the distance data (Fig. 2.2). However, if some animals move a considerable perpendicular distance and others remain in their original location, then the effect may not be detectable from the data. Ideally, the observer on a line transect survey would try to minimize such movement by looking well ahead as the area is searched. Field procedures should try to ensure that most detections occur beyond the likely range of the effect of the observer on the animals. In point transect surveys, one must be careful not to disturb animals as the sample point is approached, or perhaps wait a while upon reaching the point.

The theory of distance sampling and analysis is idealized in terms of dimensionless points or 'objects of interest'. Surveys of dead deer, plants or duck nests are easily handled in this framework. More generally, movement independent of the observer causes no problems, unless the object is counted more than once on the same unit of transect sampling effort (usually the line or point) or if it is moving at roughly half the speed of the observer or faster. Animals such as jackrabbits or pheasants will flush suddenly as an observer approaches. The measurement must be taken to the animal's original location. In these cases, the flush is often the cue that leads to detection. Animal movement after detection is not a problem, as long as the original location can be established accurately and the appropriate distance measured. Similarly, it is of no concern if an animal is detected more than once on different occasions of sampling the same transect. Animals that move to the vicinity of the next transect in response to disturbance by the observer are problematic. If the observer unknowingly records the same animal several times while traversing a transect, due to undetected movement ahead of him, bias can be large.

The assumption of no movement before detection is not met when animals take evasive movement prior to detection. A jackrabbit might hop several metres away from the observer into heavy cover and wait. As the observer moves closer, the rabbit might eventually flush. If the

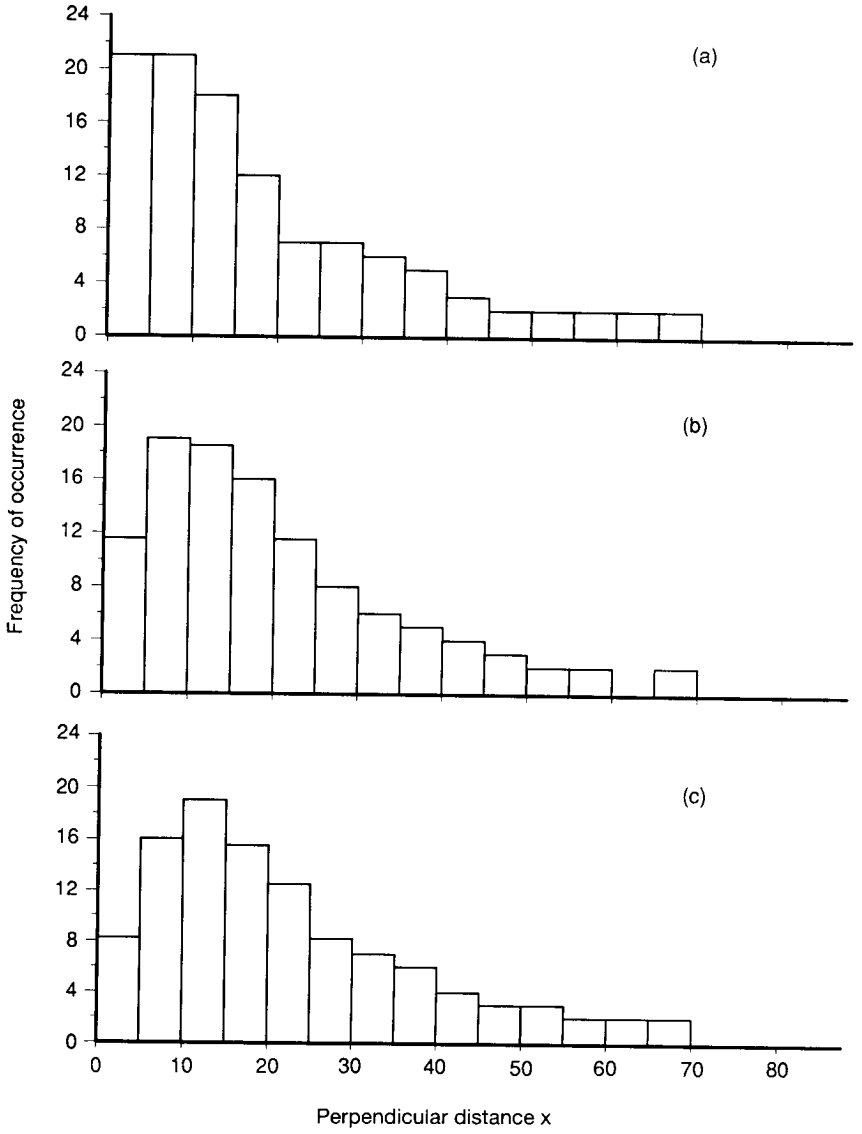

Fig. 2.2. Three histograms of perpendicular distance data, for equally spaced cutpoints, illustrating the effect of evasive movement prior to detection. Expected values are shown for the case where relatively little movement away from the observer was experienced prior to detection (a), while (b) and (c) illustrate cases where movement prior to detection was more pronounced. Data taken from Laake (1978).

new location is thought to be the original location and this distance is measured and recorded, then the assumption is violated. If this condition happens in only, say, 5% of the observations, then the bias is likely to be trivial. If a substantial portion of the population moves further from the line prior to detection, this movement will often be apparent from examination of the histogram of the distance data. If evasive movement occurs prior to detection, the estimator will be biased low ($E(\hat{D}) < D$) (Fig. 2.2b or c). Less frequently, members of a species will be attracted to the observer (Bollinger *et al.* 1988; Buckland and Turnock in press). If animals move toward the observer prior to being detected, a positive bias in estimated density can be expected ($E(\hat{D}) > D$). However, in this case, the movement is unlikely to be detected in the histogram, even if it is severe. It seems unlikely that methods will be developed for the reliable estimation of density for cases where a high proportion of the objects moves in response to the observer prior to detection without making some very critical and untestable assumptions (e.g. Smith 1979), unless relevant and reliable ancillary data can be gathered (Turnock and Quinn 1991; Buckland and Turnock in press).

2.1.3 Assumption 3: measurements are exact

Ideally, recorded distances (and angles, where relevant) are exact, without measurement errors, recording errors or heaping. For grouped data, detected objects are assumed to be correctly assigned to distance categories. Reliable estimates of density may be possible even if the assumption is violated. Although the effect of inaccurate measurements of distances or angles can often be reduced by careful analysis (e.g. grouping), it is better to gather good data in the field, rather than to rely on analytical methods. It is important that measurements near the line or point are made accurately. Rounding errors in measuring angles near zero are problematic, especially in the analysis of ungrouped data, and for shipboard surveys. If errors in distance measurements are random and not too large, then reliable density estimates are still likely, especially if the sample size is large (Gates *et al.* 1985). Biased measurements pose a larger problem (e.g. a strong tendency to overestimate the distances using ocular judgements), and field methods should be considered to minimize this bias.

For duck nests and other stationary objects, distances can be measured with a steel tape or similar device, but distances are often merely paced or estimated, taken with a rangefinder or estimated using binocular reticles. These approximate methods compromise the quality of the data, but are often forced by practical considerations. A useful alternative is to take grouped data in, say, 5–7 distance categories, such that the width of the

categories increases toward w (i.e. $[c_1 - 0] \leq [c_2 - c_1] \leq [c_3 - c_2] \leq \ldots$). Thus, careful measurement is required only near the cutpoints c_i.

(a) *Heaping* Often, when distances are estimated (e.g. ocular estimates, 'eyeballing'), the observer may 'round' to convenient values (e.g. 5, 10, 50 or 100) when recording the result. Thus, a review of the n distance values will frequently result in many 'heaped' values and relatively few numbers such as 3, 4, 7, 8 or 11. Heaping is common in sighting angles, which are often strongly heaped at 0, 15, 30, 45, 60 and 90 degrees. A histogram of the data will often reveal evidence of heaping. Often some judicious grouping of the data will allow better estimates of density, i.e. the analysis can often be improved by proper grouping of the distance data. Cutpoints for grouping distances from the line or point should be selected so that large 'heaps' fall approximately at the midpoints of the groups. For line transects, sighting distances and angles should not be grouped prior to conversion into perpendicular distances. Heaping can be avoided in the field by measuring distances, rather than merely estimating them. The effects of heaping can be reduced during the analysis by smearing (Butterworth 1982b). Heaping at perpendicular distance zero can result in serious overestimation of density. This problem is sometimes reduced if a model is used that always satisfies the **shape criterion** (Section 2.3.2), although accurate measurement is the most effective solution.

(b) *Systematic bias* When distances are estimated, it is possible that the errors are systematic rather than random. For example, there is sometimes a strong tendency to underestimate distances at sea. Each distance may tend to be over- or underestimated. In surveys where only grouped data are taken, the counts may be in error because the cutpoints c_i are in effect $c_i + \delta_i$ where δ_i is some systematic increment. Thus, n_1 is not the count of objects detected between perpendicular distances 0 and c_1, it is the count of objects detected between 0 and $c_1 + \delta_1$. Little can be done to reduce the effect of these biased measurements in the analysis of the data unless experiments are carried out to estimate the bias; a calibration equation then allows the biased measurements to be corrected. Again, careful measurements are preferable to rough estimates of distances.

(c) *Outliers* If data are collected with no fixed width w, it is possible that a few extreme outliers will be recorded. A histogram of the data will reveal outliers. These data values contain little information about the density and will frequently be difficult to fit (Fig. 1.9). Generally, such extreme values will not be useful in the final analysis of density,

and should be truncated. It is often recommended that the 5–10% of the largest observations be routinely truncated prior to analysis.

2.1.4 Other assumptions

Other aspects of the theory can be considered as assumptions. The assumption that detections are (statistically) independent events is often mentioned. If detections are somewhat dependent (e.g. 'string' flushes of quail), then the theoretical variances will be underestimated. However, we recommend that empirically based estimates of sampling variances be made, thus alleviating the need for this assumption. That is, if var(n) is estimated from independent replicate lines or points, then the assumption of within line or (point) independence is not problematic, provided the dependence is over short distances relative to the distance between replicate lines or points. Independence of detection of individual animals is clearly violated in clustered populations. This is handled by defining the cluster as the object of interest and measuring the ancillary variable, cluster size. This solution can be unsatisfactory for objects that occur in loose, poorly defined clusters, so that the location and size of the cluster may be difficult to determine or estimate without bias. The assumption of independence is a minor one in a properly designed survey, unless the clusters are poorly defined.

Statistical inference methods used here (e.g. maximum likelihood estimators of parameters, theoretical sampling variance estimators, and goodness of fit tests) assume independence among detections. Failure of the assumption of independence has little effect on the point estimators, but causes a bias (underestimation) in theoretical variance estimates (Cox and Snell 1989). The assumption of independence can fail because objects do not have a random (Poisson) distribution in space and this pattern could result in a dependency in the detections. Non-random distribution, by itself, is not necessarily a cause of lack of independence. If the transects are placed at random and a robust estimator of the sampling variance is used, then the assumption of independence can be ignored. At least in practice, it is not at all important that the objects be randomly distributed on the study area. Similarly, it is of little concern if detection on either side of the line or around the point is not symmetric, provided that the asymmetry is not extreme, such that modelling $g(y)$ is difficult.

A more practically important consideration relates to the shape of the detection function near zero distance. This shape can often be judged by examining histograms of the distance data using different groupings. Distance sampling theory performs well when a 'shoulder' in detectability exists near the line or around the point. That is, detectability is certain

near the line or point and stays certain or nearly certain for some distance. This will be defined as the '**shape criterion**' in Section 2.3. If detectability falls sharply just off the line or point, then estimation tends to be poor, even if the true model for the data is known. Thus if data are to be analysed reliably, the detection function from which they come should possess a shoulder; to this extent, the **shape criterion** is an assumption.

Some papers imply that an object should not be counted on more than one line or point. This, by itself, is not true as no such assumption is required. In surveys with $w = \infty$, an object of interest (e.g. a dead elk) can be detected from two different lines without violating any assumptions. As noted above, if in line transect sampling an animal moves ahead of the observer and is counted repeatedly, abundance will be overestimated. This is undetected movement in response to the observer; double counting, by itself, is not a cause of bias if such counts correspond to different units of counting effort. Bias is likely to be small unless repeated counting is common during a survey. Detections made behind the observer in line transect sampling may be utilized, unless the object is located before the start of a transect leg, in which case it is outside the rectangular strip being surveyed.

These assumptions, their importance, models robust to partial violations of assumptions, and field methods to meet assumptions adequately will be addressed in the material that follows.

2.2 Fundamental models

This section provides a glimpse of the theory underlying line and point transect sampling. This material is an extension of Section 1.6.

2.2.1 Line transects

In strip transect sampling, if strips of width $2w$ and total length L are surveyed, an area of size $a = 2wL$ is censused. All n objects within the strips are enumerated, and estimated density is the expected number of objects per unit area:

$$\hat{D} = n/2wL$$

In line transect sampling, only a proportion of the objects in the area a surveyed is detected. Let this unknown proportion be P_a. If P_a can be estimated from the distance data, the estimate of density could be written as

$$\hat{D} = n/2wL\hat{P}_a \tag{2.1}$$

Now, some formalism is needed for the estimation of P_a from the distances. The unconditional probability of detecting an object in the strip (of area $a = 2wL$) is

$$P_a = \frac{\int_0^w g(x)dx}{w} \tag{2.2}$$

In the duck nest example of Chapter 1, $g(x)$ was found by dividing the estimated quadratic equation by the intercept (77.05), to give

$$\hat{g}(x) = 1 - 0.0052x^2$$

Note that $\hat{g}(8) = 0.66$, indicating that approximately one-third of the nests near the edges of the transect were never detected. Then

$$\hat{P}_a = \frac{\int_0^8 (1 - 0.0052x^2)dx}{8}$$

$$= 0.888$$

Substituting the estimator of P_a from Equation 2.2 into $\hat{D}$ from Equation 2.1 gives

$$\hat{D} = \frac{n}{2L\int_0^w \hat{g}(x)dx} \tag{2.3}$$

because the w and $1/w$ cancel out. Then the integral $\int_0^w g(x)dx$ becomes the critical quantity and is denoted as μ for simplicity. Thus,

$$\hat{D} = n/2L\hat{\mu}$$

There is a very convenient way to estimate the quantity $1/\mu$. The derivation begins by noting that the probability density function (pdf) of the perpendicular distance data, conditional on the object being detected, is merely

$$f(x) = \frac{g(x)}{\int_0^w g(x)dx} \tag{2.4}$$

This result follows because the expected number of objects (including those that are not detected) at distance x from the line is independent of x. This implies that the density function is identical in shape to the detection function; it can thus be obtained by rescaling, so that the function integrates to unity.

By assumption, $g(0) = 1$, so that the pdf, evaluated at zero distance, is

$$f(0) = \frac{1}{\int_0^w g(x)dx}$$

$$= 1/\mu$$

The parameter $\mu = \int_0^w g(x)dx$ is a function of the measured distances. Therefore, we will often write the general estimator of density for line transect sampling simply as

$$\hat{D} = \frac{n \cdot \hat{f}(0)}{2L} \tag{2.5}$$

$$= \frac{n}{2L\hat{\mu}}$$

This estimator can be further generalized, but the conceptual approach remains the same. $\hat{D}$ is valid whether w is bounded or unbounded (infinite) and when the data are grouped or ungrouped. Note that either form of Equation 2.5 is equivalent to $\hat{D} = n/2wL\hat{P}_a$ (Equation 2.1).

For the example, an estimate of Gates' (1979) effective strip width is $\hat{\mu} = w\hat{P}_a = 8(0.888) = 7.10$ ft, and $\hat{D} = 534/(2 \times 1600 \times 7.10)$ nests/mile/ft = 124 nests/square mile.

The density estimator expressed in terms of an estimated pdf, evaluated at zero, is convenient, as a large statistical literature exists on the subject of estimating a pdf. Thus, a large body of general knowledge can be brought to bear on this specific problem.

2.2.2 Point transects

In traditional circular plot sampling, k areas each of size πw^2 are censused and all n objects within the k plots are enumerated. By definition, density is the number per unit area, thus

$$\hat{D} = \frac{n}{k\pi w^2}$$

In point transect sampling, only a proportion of the objects in each sampled area is detected. Again, let this proportion be P_a. Then the estimator of density is

$$\hat{D} = \frac{n}{k\pi w^2 \hat{P}_a} \tag{2.6}$$

The unconditional probability of detecting an object that is in one of the k circular plots is

$$P_a = \int_0^w \frac{2\pi r g(r) dr}{\pi w^2} \tag{2.7}$$

$$= \frac{2}{w^2} \int_0^w r g(r) dr$$

Substituting Equation 2.7 into Equation 2.6 and cancelling the w^2 terms, the estimator of density is

$$\hat{D} = \frac{n}{2k\pi \int_0^w r\hat{g}(r) dr} \tag{2.8}$$

Defining $\nu = 2\pi \int_0^w r g(r) dr$

then

$$\hat{D} = n/k\hat{\nu}$$

Clearly, ν is the critical quantity to be estimated from the distance data for a point transect survey.

2.2.3 Summary

The statistical problem in the estimation of density of objects is the estimation of μ or ν. Then the estimator of density for line transect sampling is

$$\hat{D} = n/2L\hat{\mu}$$

where $\mu = \int_0^w g(x)dx$

The estimator of density for point transect surveys can be given in a similar form:

$$\hat{D} = n/k\hat{\nu}$$

where $\nu = 2\pi \int_0^w rg(r)dr$

This, then, entails careful modelling and estimation of $g(y)$. Good statistical theory now exists for these general problems. Finally, we note that the estimator of density from strip transect sampling is also similar:

$$\hat{D} = n/2wL$$

where $P_a = 1$ and, by assumption, n is the count from a complete census of each strip.

2.3 Philosophy and strategy

The true detection function $g(y)$ is not known. Furthermore, it varies due to numerous factors (Section 1.7). Therefore, it is important that strong assumptions about the shape of the detection function are avoided. In particular, a flexible or 'robust' model for $g(y)$ is essential.

The strategy used here is to select a few models for $g(y)$ that have desirable properties. These models are selected *a priori*, and without particular reference to the given data set. This class of models excludes those that are not robust, have restricted shapes, or have inefficient estimators. Because the estimator of density is closely linked to $g(y)$, it is of critical importance to select models for the detection function

carefully. Three properties desired for a model for $g(y)$ are, in order of importance, **model robustness**, a **shape criterion**, and **efficiency**.

2.3.1 Model robustness

The most important property of a model for the detection function is **model robustness**. This means that the model is a general, flexible function that can take the variety of shapes that are likely for the true detection function. In general, this property excludes single parameter models; experience has shown that models with two or three parameters are frequently required. Most of the models recommended have a variable number of parameters, depending on how many are required to fit the specific data set. These are sometimes called semiparametric models.

The concept of **pooling robustness** (Burnham *et al.* 1980) is included here under model robustness. Models of $g(y)$ are pooling robust if the data can be pooled over many factors that affect detection probability (Section 1.7) and still yield a reliable estimate of density. Consider two approaches: stratified estimation $\hat{D}_{st}$ and pooled estimation $\hat{D}_p$. In the first case, the data could be stratified by factors affecting detectability (e.g. three observers and four habitat types) and an estimate of density made for each stratum. These separate estimates could be combined into an estimate of average density $\hat{D}_{st}$. In the second case, all data could be pooled, regardless of any stratification (e.g. the data for the three observers and four habitat types would be pooled) and a single estimate of density computed, $\hat{D}_p$. A model is **pooling robust** if $\hat{D}_{st} \doteq \hat{D}_p$. **Pooling robustness** is a desirable property. Only models that are linear in the parameters satisfy the condition with strict equality, although general models that are **model robust**, such as those recommended in this book, approximately satisfy the **pooling robust** property.

2.3.2 Shape criterion

Theoretical considerations and the examination of empirical data suggest that the detection function should have a 'shoulder' near the line or point. That is, detection remains nearly certain at small distances from the line or point. Mathematically, the derivative $g'(0)$ should be zero. This **shape criterion** excludes functions that are spiked near zero distance. Frequently, a histogram of the distance data will not reveal the presence of a shoulder, particularly if the histogram classes are large (Fig 2.3), or if the data include several large values (a long tail). Generally, good models for $g(y)$ will satisfy the **shape criterion** near zero distance. The **shape criterion** is especially important in the analysis of

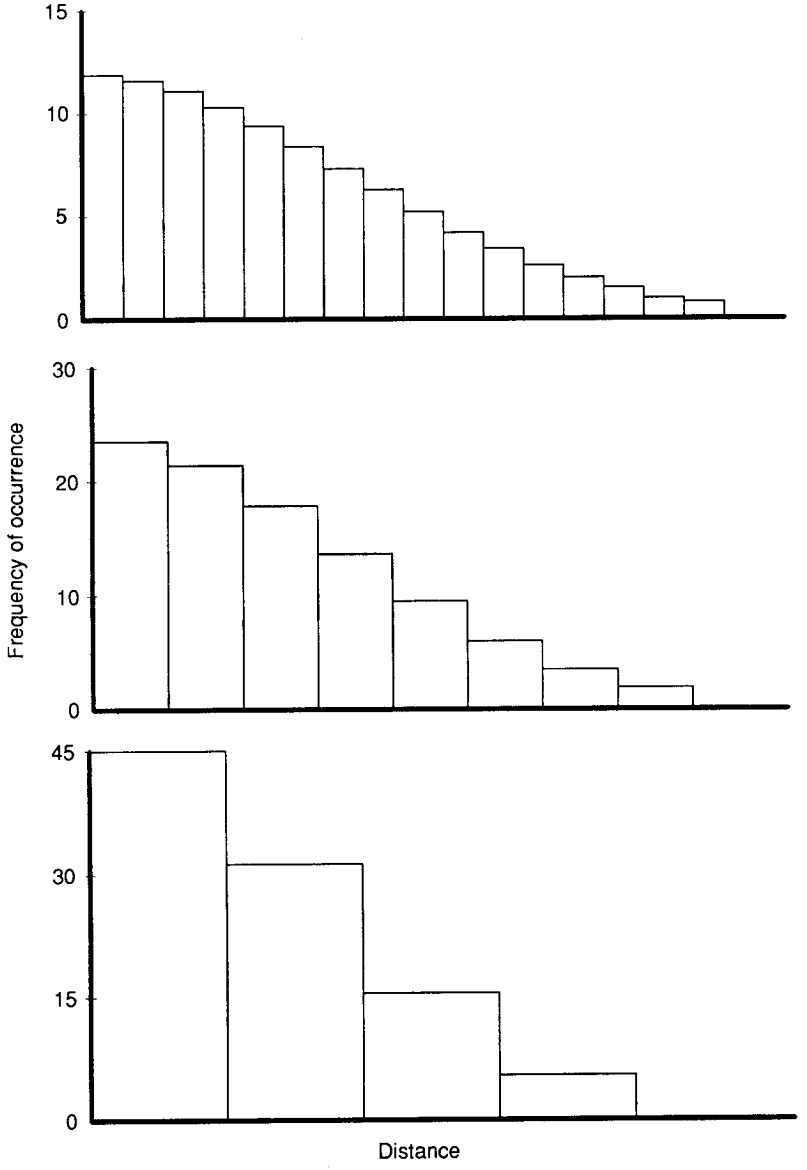

Fig. 2.3. Data ($n = 100$) from the half-normal model with $\sigma = 33.3$ and $w = 100$ shown with three different sets of group interval. As the group interval increases, the data appear to become more spiked. Adapted from Burnham *et al.* (1980).

data where some heaping at zero distance is suspected. This occurs most frequently when small sighting angles are rounded to zero, and gives

rise to histograms that show no evidence of a shoulder, even when the true detection function has a substantial shoulder.

2.3.3 Efficiency

Other things being equal, it is desirable to select a model that provides estimates that are relatively precise (i.e. have small variance). We recommend maximum likelihood methods, which have many good statistical properties, including that of asymptotic minimum variance. Efficient estimation is of benefit only for models that are model robust and have a shoulder near zero distance; otherwise, estimation might be precise but biased.

2.3.4 Model fit

Ideally, there would be powerful statistical tests of the fit of the model for $g(y)$ to the distance data. The only simple omnibus test available is the χ^2 goodness of fit test based on grouping the data. This test compares the observed frequencies n_i (based on the grouping selected) with the estimated expected frequencies under the model, $\hat{E}(n_i)$, in the usual way:

$$\chi^2 = \sum_{i=1}^{u} \frac{[n_i - \hat{E}(n_i)]^2}{\hat{E}(n_i)}$$

is approximately χ^2 with $u - m - 1$ degrees of freedom, where u is the number of groups and m is the number of parameters estimated. In isolation, this approach has severe limitations for choosing a model for $g(y)$, given a single data set (Fig. 2.4).

Generally, as the number of parameters in a model increases, the bias decreases but the sampling variance increases. A proper model should be supported by the particular data set and thus have enough parameters to avoid large bias but not so many that precision is lost (the **Principle of Parsimony**). Likelihood ratio tests (Lehmann 1959; Hogg and Craig 1970) are used in selecting the number of model parameters that are appropriate in modelling $f(y)$. The relative fit of alternative models may be evaluated using Akaike's Information Criterion (Akaike 1973; Sakamoto *et al.* 1986; Burnham and Anderson 1992). These technical subjects are presented in the following chapters.

2.3.5 Test power

The power of the goodness of fit test is quite low and, therefore, of little use in selecting a good model of $g(y)$ for the analysis of a particular

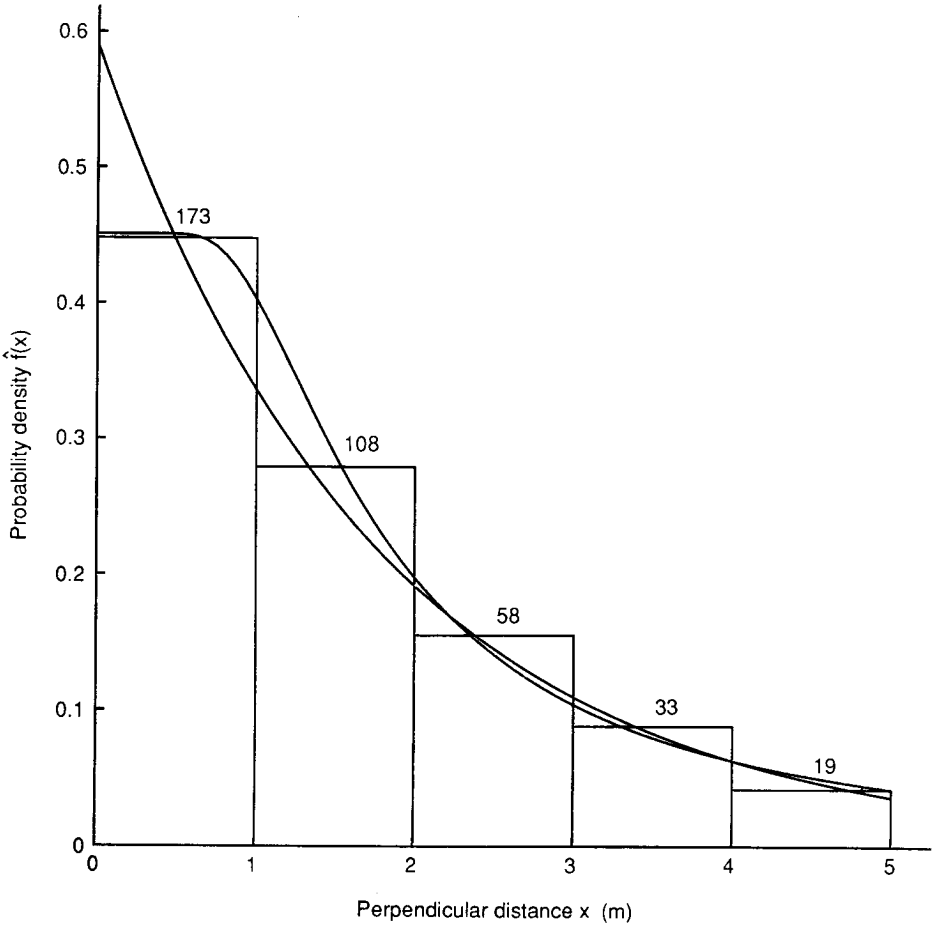

Fig. 2.4. The distance data are often of little help in testing the relative fit among models. Here, fits of the negative exponential model and the hazard-rate model to a line transect data set are shown. Both models provide an excellent fit ($\chi^2 = 0.49$, 3 df, $p = 0.92$, and $\chi^2 = 0.33$, 2 df, $p = 0.85$, respectively), even though the estimates of $f(0)$ are quite different ($\hat{f}(0) = 0.589$ and 0.450, respectively).

data set. In particular, this test is incapable of discriminating between quite different models near the line or point, the most critical region (Fig. 2.4). In addition, grouping data into fewer groups frequently diminishes the power of the test still further and may give the visual impression that the data arise from a spiked distribution such as the negative exponential, when the true detection function has a shoulder (Fig. 2.3).

While goodness of fit test results should be considered in the analysis of distance data, they will be of limited value in selecting a model. Thus,

a class of reliable models is recommended here, based on the three properties: **model robustness,** the **shape criterion** and **estimator efficiency**.

2.4 Robust models

Several models of $g(y)$ are recommended for the analysis of line or point transect data. These models, as implemented in program DISTANCE, have the three desired properties of **model robustness, shape criterion** and **estimator efficiency**. Following Buckland (1992a), the modelling process can be conceptualized in two steps. First, a '**key function**' is selected as a starting point, possibly based on visual inspection of the histogram of distances, after truncation of obvious outliers. Often, a simple key function is adequate as a model for $g(y)$, especially if the data have been properly truncated. Two key functions should probably receive initial consideration: the uniform and the half-normal (Fig. 2.5a). The uniform key function has no parameters, whereas the half-normal key has one unknown parameter to be estimated from the distance data. In some cases, the hazard-rate model (Fig. 2.5b) could be considered as a key function, although it requires that two key parameters be estimated.

Second, a flexible form, called a '**series expansion**', is used to adjust the key function, using perhaps one or two more parameters, to improve the fit of the model to the distance data. Conceptually, the detection function is modelled in the following general form:

$$g(y) = \text{key}(y)\,[1 + \text{series}\,(y)]$$

The key function alone may be adequate for modelling $g(y)$, especially if sample size is small or the distance data are easily described by a simple model. Theoretical considerations often suggest a series expansion appropriate for a given key. Three series expansions are considered here: (1) the cosine series, (2) simple polynomials, and (3) Hermite polynomials (Stuart and Ord 1987: 220–7). All three expansions are linear in their parameters. Thus, some generally useful models of $g(y)$ are:

Key function	*Series expansion*
Uniform, $1/w$	Cosine, $\sum_{j=1}^{m} a_j \cos\left(\frac{j\pi y}{w}\right)$
Uniform, $1/w$	Simple polynomial, $\sum_{j=1}^{m} a_j \left(\frac{y}{w}\right)^{2j}$

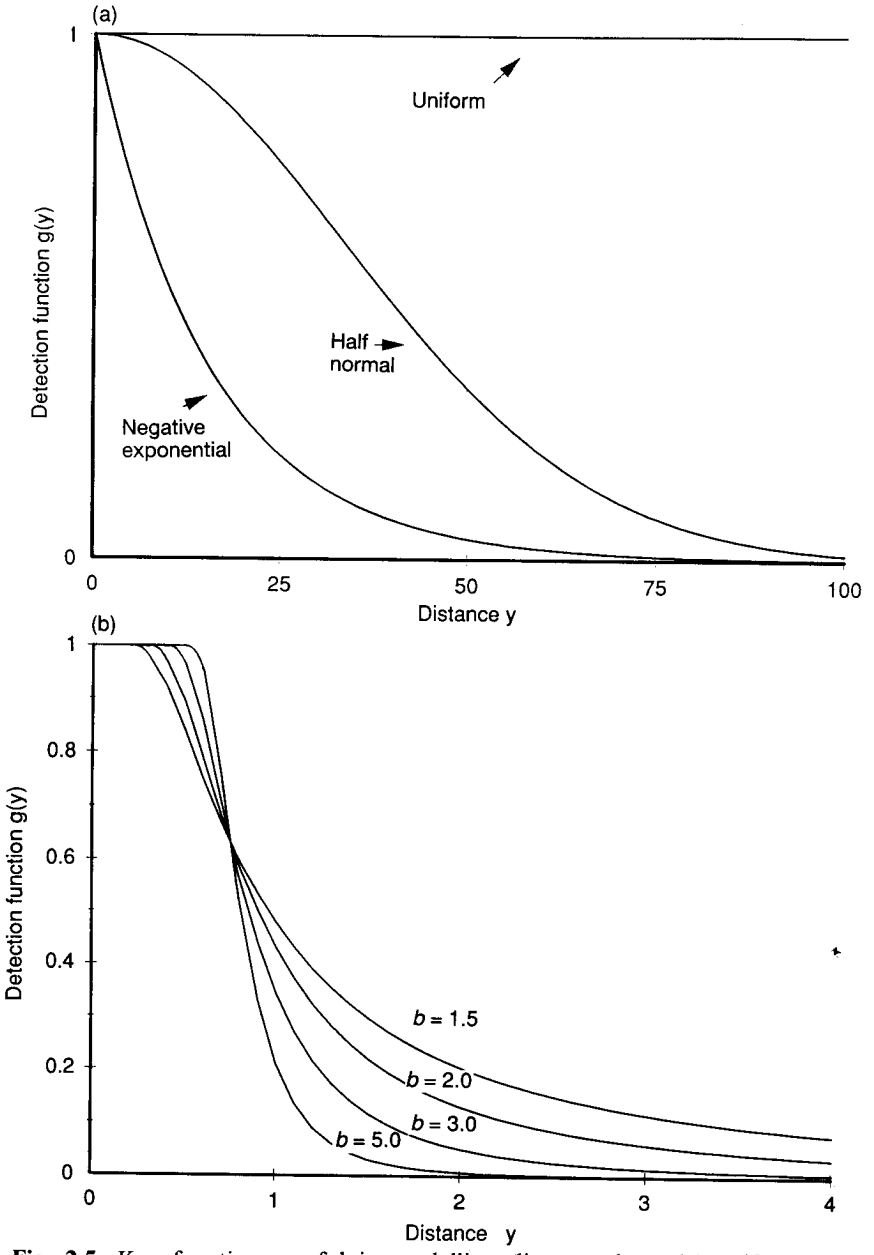

Fig. 2.5. Key functions useful in modelling distance data: (a) uniform, half-normal and negative exponential, and (b) hazard-rate model for four different values of the shape parameter b.

Half-normal, $\exp(-y^2/2\sigma^2)$	Cosine, $\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi y}{w}\right)$
Half-normal, $\exp(-y^2/2\sigma^2)$	Hermite polynomial, $\sum_{j=2}^{m} a_j H_{2j}(y_s)$ where $y_s = y/\sigma$
Hazard-rate, $1 - \exp(-(y/\sigma)^{-b})$	Cosine, $\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi y}{w}\right)$
Hazard-rate, $1 - \exp(-(y/\sigma)^{-b})$	Simple polynomial, $\sum_{j=2}^{m} a_j \left(\frac{y}{w}\right)^{2j}$

The uniform + cosine expression is the Fourier series model of Crain *et al.* (1979) and Burnham *et al.* (1980). This is an excellent omnibus model and has been shown to perform well in a variety of situations. The uniform + simple polynomial model includes the models of Anderson and Pospahala (1970), Anderson *et al.* (1980), and Gates and Smith (1980).

It may be desirable to use the half-normal key function with either a cosine expansion or Hermite polynomials. Because histograms of distance data often decline markedly with distance from the line, the half-normal may often represent a good choice as a key function. Similarly, the uniform key and one cosine term will often provide a good standard for possible further fitting with series adjustment terms. Theoretical reasons suggest the use of the Hermite polynomial in conjunction with the half-normal key, especially for the untruncated case. This is a minor point, and the reader should think of this as only an alternative form of a polynomial.

The final two models listed above use Buckland's (1985) hazard-rate as a two parameter key function and use cosine and simple polynomial expansions for additional fitting, if required. The hazard-rate model is a derived model in contrast to the others, which are proposed shapes. That is, the shape of this family of models is the result of *a priori* assumptions about the detection process. The hazard-rate model has been shown to possess good properties, especially for data that are genuinely spiked (as distinct from spuriously spiked, as a result of rounding). In addition, this model can have a marked shoulder that can be nearly flat for some distance from the line or point. Even for data appearing to be spiked, this model can fit a flat shoulder, yet provide a good fit.

These series-expansion models are non-parametric in the sense that the number of parameters used is data-dependent. The estimation theory for these models, including rules to select the number of parameters to

use, is given in the following chapter. Typically, given suitable distance truncation, an adequate model for $g(y)$ will include only one or two parameters, sometimes three. Sometimes the key function by itself will be adequate, with no terms in the series expansion. We emphasize that truncation will often be required as part of the modelling, especially if the data are ungrouped. Outlier observations provide relatively little information about density, but are often difficult to model, so that proper truncation should always be considered in modelling $g(y)$. Program DISTANCE allows the combination of any of the key functions with any of the series expansions as a model for $f(y)$. Some models have appeared in the literature that assume $g(y) = 1$ for some considerable distance from the line or point; the models suggested above do not impose this assumption.

Only these general models are emphasized as state-of-the-art, general approaches at this time. Program DISTANCE allows any key function to be used with any series expansion; however, the combinations listed above should be satisfactory for general use. Further effort directed at model evaluation and development might now be better directed at survey design and data collection techniques to meet critical assumptions.

The exponential + simple polynomial is available for the salvage analysis of poorly collected data where there is strong reason to believe that the distance data are truly spiked. It has the form:

$$\exp(-y/\lambda) \cdot [1 + \sum_{j=1}^{m} a_j (y/\lambda)^{2j}]$$

Use of this approach should be accompanied by adequate justification and we recommend its use only in unusual circumstances. Every consideration should be given to the use of the hazard-rate model for distance data that appear spiked because this model enforces the shape criterion, offers greater flexibility in fitting a spike, and gives a more realistic (larger) variance when the data are inadequate for reliable modelling.

2.5 Some analysis guidelines

Distance sampling represents a broad area and includes many types of application and degree of complexity of design and data. Thus, specific 'cookbook' procedures for data analysis cannot be given safely. Instead, we will suggest a useful strategy that could be considered when planning

the analysis of a data set. In this section we will consider only a simple survey and will not address stratification and other complications given in later chapters.

2.5.1 Exploratory phase

The exploratory phase of the analysis involves the preparation of histograms of the distance data under several groupings. Sometimes it is effective to partition the data into 10–20 groups to get a fine-grained picture of the distance data. Examination of such histograms can provide insight into the presence of heaping, evasive movement, outliers and the occasional gross error. Prominent heaps can be mitigated by judicious grouping or splitting prior to further analysis. Evasive movement is problematic, but it is important to know that movement is present (movement toward the line or point generally cannot be detected from the distance data alone). Some truncation of the distance data is nearly always suggested, even if no obvious outlier is noticed. We frequently recommend that 5–10% of the largest observations be truncated. A more refined rule of thumb is to truncate the data when $g(x) \doteq 0.15$ for line transects or 0.10 for point transects. If the data were taken as grouped data in the field, then options for further truncation are more limited. Some liberal truncation is generally recommended. Empirical estimates of $\text{var}(n)$ can be computed and compared with the variance under the Poisson assumption (i.e. $\widehat{\text{var}}(n) = n$). One can examine the stability of the ratio $\widehat{\text{var}}(n)/n$ over various design features. If the data are from a clustered population, plots of s or $\log_e(s)$ vs x or r should be made and examined. Of course, data entry errors and other anomalies should be screened and corrected. This analysis phase is open-ended but the analyst is encouraged to begin to understand the data and possible violations of the assumptions. Chatfield (1988, 1991) offered some general practical advice relevant here. Program DISTANCE allows substantial exploratory options.

2.5.2 Model selection

Model selection cannot proceed until proper truncation and, where relevant, grouping have been tentatively addressed. Thus, this phase begins once a data set has been properly prepared. Several robust models should be considered (e.g. those in Section 2.4). The following chapters will introduce and demonstrate the use of likelihood ratio tests, goodness of fit tests and Akaike's Information Criterion (AIC; Akaike 1973) as aids in objective model selection. Here it might be appropriate to remind the analyst that it is the fit of the model to the distance data near the

line or point that is most important (unless there is thought to be heaping at zero distance). Usually analysis will suggest additional exploratory work, so that the process is iterative. For example, it may become apparent that the fit of one or more models could be improved by selecting a different truncation point w, or by grouping ungrouped data, or by changing the choice of group intervals for grouped data. Further, if data are available over several years, taken in the same habitat type by the same observer, then it might be prudent to pool the data for the estimation of $f(0)$ or $h(0)$, but to use the year-specific sample sizes n_i, where i is year, to estimate annual abundance. The validity of this approach must then be assessed, for example using a likelihood ratio test to determine whether a common value for $f(0)$ or $h(0)$ can be assumed.

2.5.3 Final analysis and inference

At some point the analyst selects a model believed to be the best for the data set under consideration. In some cases, there may be several competing models that seem equally good. In most cases, there will be a subset of models that can be excluded from final consideration because they perform poorly relative to other models. Often, if two or three models seem to fit equally well to a data set, estimation of density D and mean cluster size $E(s)$ under these models will be quite similar (see examples in Chapter 8).

Once a single model has been selected, the analyst can address further issues. Thus, one might consider bootstrapping to obtain improved estimates of precision, or carry out a Monte Carlo study to understand further the effect of some assumption failure (e.g. overestimation of a significant proportion of detection distances in an aerial survey, due to the aircraft flying too low at times). Finally, estimates of density or abundance and their precision are made, and qualifying statements presented, such as discussion of the effects of failures of assumptions.

The above guidelines give a broad indication of how the analyst might proceed. They will be developed in the following chapters, both to give substance to the theory required at each step, and to show how the philosophy for analysis is implemented in real examples.