3

Statistical theory

3.1 General formula

The analysis methods for distance sampling described here model measured or estimated distances from a line or point so that density of objects in a study area may be estimated. Conceptually, object density varies spatially, and lines or points are placed at random or systematically in the study area to allow mean density to be estimated.

Suppose in a given survey that objects do not occur in clusters and that distances are only recorded out to a distance w from the line or point, or equivalently that recorded distances are truncated at distance w. Suppose further that the true density is D objects per unit area. Let the area covered by the survey within distance w of the line or point be a, and let the probability of detection for an object within this area, unconditional on its actual position, be P_a. Then the expected number of objects detected within distance w, $E(n)$, is equal to the expected number of animals in the surveyed area, $D \cdot a$, multiplied by the probability of detection, P_a, so that

$$D = \frac{E(n)}{a \cdot P_a}$$

If objects occur in clusters, so that $E(n)$ is the expected number of clusters, then the above equation should be multiplied by $E(s)$, the expectation of cluster size for the population:

$$D = \frac{E(n) \cdot E(s)}{a \cdot P_a}$$

Although the result is then perfectly general, it is convenient to modify the definitions of a and P_a to show explicitly two components of the general formula that are implicit in the above form of the equation.

First, if a is defined to be $2wL$ for a line transect survey, where L is the total transect length, or $a = k\pi w^2$, where k is the number of circular plots in a point transect survey, then the area surveyed within a distance w of the line or point can be expressed as $c \cdot a$. Usually $c = 1.0$, but if for example only one side of a line transect is counted, then $c = 0.5$. Similarly, if only an angle of ϕ radians ($\phi < 2\pi$) is counted in a point transect survey, then $c = \phi/2\pi$. This factor is required for example in the cue counting method (Section 6.10), in which the sector counted is of angle ϕ.

Second, a basic assumption of the standard line and point transect methods is that the probability of detection at zero distance $g(0)$ is unity. In surveys of inconspicuous objects or, for example, of whales, this assumption may be unreasonable. It may be possible to estimate $g(0)$, in which case it is convenient to rescale the detection function $g(y)$ such that $g(0) = 1.0$, and to define the probability of detection on the line or at the point to be g_0. The unconditional probability of detection of an object (or cluster) in the surveyed area can then be factorized into $g_0 \cdot P_a$. This yields the general equation

$$D = \frac{E(n) \cdot E(s)}{c \cdot a \cdot P_a \cdot g_0} \tag{3.1}$$

Estimation of g_0 is generally problematic, so that if at all possible, surveys should be designed such that all or almost all objects on or close to the line or point are detected. Further discussion of this issue is reserved for Chapter 6.

Replacing parameters in Equation 3.1 by their estimators gives

$$\hat{D} = \frac{n \cdot \hat{E}(s)}{c \cdot a \cdot \hat{P}_a \cdot \hat{g}_0} \tag{3.2}$$

The variance of $\hat{D}$ may be approximated using the delta method (Seber 1982: 7–9). Assuming correlations between the four estimation components are zero, the variance estimate is then:

$$\widehat{\text{var}}(\hat{D}) = \hat{D}^2 \cdot \left\{ \frac{\widehat{\text{var}}(n)}{n^2} + \frac{\widehat{\text{var}}[\hat{E}(s)]}{[\hat{E}(s)]^2} + \frac{\widehat{\text{var}}(a \cdot \hat{P}_a)}{(a \cdot \hat{P}_a)^2} + \frac{\widehat{\text{var}}[\hat{g}_0]}{[\hat{g}_0]^2} \right\} \tag{3.3}$$

The assumption of no correlation is a mild one in the sense that estimation is usually done in a way that ensures it holds. Because P_a is estimated conditional on n, no correlation term exists between n and

$\hat{P}_a$ if we can assume that $E(\hat{P}_a|n) = E(\hat{P}_a)$, independent of n. This assumption holds if $(\hat{P}_a|n)$ is unbiased for P_a:

$$\begin{aligned}\text{cov}[n,(\hat{P}_a|n)] &= E[n \cdot \hat{P}_a|n] - E(n) \cdot E(\hat{P}_a|n) \\ &= E_n[E[\{n \cdot (\hat{P}_a|n)\}|n]] - E(n) \cdot P_a \\ &= E_n[n \cdot E(\hat{P}_a|n)] - E(n) \cdot P_a \\ &= E(n) \cdot P_a - E(n) \cdot P_a \\ &= 0\end{aligned}$$

When sample size is adequate, $(\hat{P}_a|n)$ is approximately unbiased.

Similarly, $E(s)$ may be estimated conditional on n and the detection distances, rendering $\hat{E}(s)$ uncorrelated with n or $\hat{P}_a$. Estimation of g_0, if required, is usually based on additional, independent data.

Although area $a \to \infty$ as $w \to \infty$, the product $a \cdot \hat{P}_a$ remains finite, so that all three equations hold when there is no truncation. To estimate $a \cdot P_a$, a form must be specified, explicitly or implicitly, for the detection function $g(y)$, which represents the probability of detection of an object or object cluster at a distance y from the line or point. The simplest form is that of the Kelker strip: the truncation point w is selected such that it is reasonable to assume that $g(y) = 1.0$ for $0 \leqslant y \leqslant w$. More generally it seems desirable that the detection function has a 'shoulder'; that is, $g'(0)$ should be zero, so that the detection function is flat at zero. This is the shape criterion defined by Burnham *et al.* (1980). The detection function should also be non-increasing, and have a tail that goes asymptotically to zero.

The relationship between the detection function and the probability density function of distances, $f(y)$, is different for line and for point transects. We use the notation y to represent either x, the perpendicular distance of an object from the centreline in line transect sampling, or r, the distance of an object from the observer in point transect sampling.

For line transect sampling, the relationship between $g(x)$ and $f(x)$ is particularly simple. Intuitively, because the area of a strip of incremental width dx at distance x from the line is independent of x, it seems reasonable that the density function should be identical in shape to the detection function, but rescaled so that it integrates to unity. This result may be proven as follows. Suppose for the moment that w is finite.

$$\begin{aligned}f(x)dx &= \text{pr}\{\text{object is in } (x, x+dx) \mid \text{object detected}\} \\ &= \frac{\text{pr}\{\text{object is in } (x, x+dx) \text{ and object is detected}\}}{\text{pr}\{\text{object is detected}\}}\end{aligned}$$

$$= \frac{\text{pr}\{\text{object is detected} | \text{object is in } (x, x+dx)\} \cdot \text{pr}\{\text{object is in } (x, x+dx)\}}{P_a}$$

$$= \frac{g(x) \cdot \dfrac{dx \cdot L}{w \cdot L}}{P_a}$$

Thus $$f(x) = \frac{g(x)}{w \cdot P_a}$$

It is convenient to define $\mu = w \cdot P_a$, so that

$$f(x) = \frac{g(x)}{\mu}$$

Because $\int_0^w f(x)dx = 1$, it follows that $\mu = \int_0^w g(x)dx$. Figure 3.1 illustrates the result that P_a, the probability of detecting an object given that it is within w of the centreline, is $\mu = \int_0^w g(x)dx$ (the area under the curve) divided by $1.0w$ (the area of the rectangle); that is, $w \cdot P_a = \mu$, which is well-defined even when w is infinite.

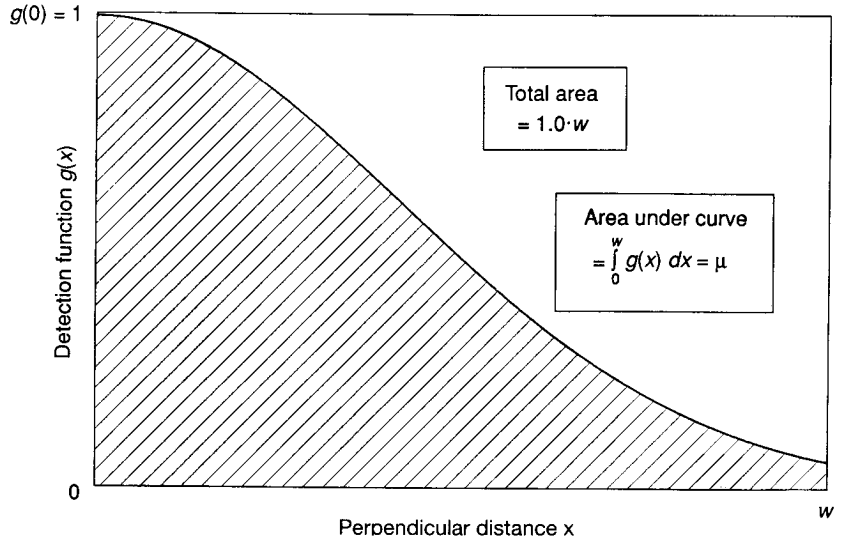

Fig. 3.1. The unconditional probability that an animal within distance w of the line is detected is the area under the detection function μ divided by the area of the rectangle $1.0\ w$.

The parameter μ is often termed the effective strip width (or more strictly, the effective strip half-width); if all objects were detected out to a distance μ on either side of the transect, and none beyond, then the expected number of objects detected would be the same as for the actual survey.

Let the total length of transect be L. Then the area surveyed is $a = 2wL$, and $a \cdot P_a = 2\mu L$. Since $\mu = g(x)/f(x)$ and $g(0) = 1.0$ after rescaling, if necessary, by the factor g_0, then $\mu = 1/f(0)$, and so for line transects, Equation 3.1 becomes:

$$D = \frac{E(n) \cdot f(0) \cdot E(s)}{2cLg_0} \tag{3.4}$$

The parameter $f(0)$ is statistically well-defined and is estimable from the perpendicular distances $x_1, \ldots, x_n$ in a variety of ways.

The derivation for point transects is similar, but the relationship between $g(r)$ and $f(r)$ is less simple. The area of a ring of incremental width dr at distance r from the observer is proportional to r. Thus we might expect that $f(r)$ is proportional to $r \cdot g(r)$; using the constraint that $f(r)$ integrates to unity, $f(r) = 2\pi r g(r)/\nu$, where $\nu = 2\pi \int_0^w r g(r) dr$. A more rigorous proof follows.

$$f(r)dr = \text{pr}\{\text{object is in the annulus } (r, r + dr) \mid \text{object detected}\}$$

$$= \frac{\text{pr}\{\text{object is in } (r, r + dr) \text{ and object is detected}\}}{\text{pr}\{\text{object is detected}\}}$$

$$= \frac{\text{pr}\{\text{object is detected} \mid \text{object is in } (r, r + dr)\} \cdot \text{pr}\{\text{object is in } (r, r + dr)\}}{P_a}$$

$$= \frac{g(r) \cdot \dfrac{2\pi r dr}{\pi w^2}}{P_a}$$

so that $$f(r) = \frac{2\pi r \cdot g(r)}{\pi w^2 \cdot P_a}$$

To be a valid density function, $\int_0^w f(r)dr = 1$, so that

$$f(r) = \frac{2\pi r \cdot g(r)}{\nu}, \text{ with } \nu = 2\pi \int_0^w r g(r) dr = \pi w^2 \cdot P_a$$

This result also holds for infinite w. Analogous to μ, ν is sometimes called the effective area of detection.

Let there be k points, so that the surveyed area is $a = k\pi w^2$. The probability of detection of an object, given that it is within a distance w of the observer, is now $P_a = \nu/(\pi w^2)$, so that $a \cdot P_a$ becomes $k\nu$; again this holds as $w \to \infty$. Since $\nu = 2\pi r g(r)/f(r)$ and $g(0) = 1.0$ (after rescaling if necessary), then for point transects Equation 3.1 becomes:

$$D = \frac{E(n) \cdot h(0) \cdot E(s)}{2\pi c k g_0} \tag{3.5}$$

where $h(0) = \lim_{r \to 0} f(r)/r = 2\pi/\nu$

Note that $h(0)$ is merely the derivative of the probability density $f(r)$ evaluated at $r = 0$; alternative notation would be $f'(0)$. It is thus estimable from the detection distances $r_1, \ldots, r_n$. Whereas $f(x)$ and $g(x)$ have identical shapes in line transect sampling, for point transects, $g(r)$ is proportional to $f(r)/r$. The constant of proportionality is $1/h(0)$.

Results equivalent to Equations 3.2 and 3.3 follow in the obvious way. Note that for both line and point transects, behaviour of the probability density function at zero distance is critical to object density estimation.

Burnham *et al.* (1980: 195) recommended that distances in point transect sampling should be transformed to 'areas' before analysis. Thus, the ith recorded area would be $u_i = \pi r_i^2$, $i = 1, \ldots, n$. If $f_u(u)$ denotes the probability density function of areas u_i, it may be shown that $f_u(u) = f(r)/(2\pi r) = g(r)/\nu$. The advantage of this transformation should now be apparent; the new density is identical in form to that for perpendicular distances in line transect sampling (where $f(x) = g(x)/\mu$), so line transect software may be used to analyse the data. Further, if r is allowed to tend to zero, then $f_u(0) = h(0)/(2\pi)$, and the development based on areas is therefore equivalent to that based on distances. This seems to suggest that modelling of areas rather than distances is preferable. However, as noted by Buckland (1987a), the transformation to area appreciably alters the shape of the detection function, and it is no longer clear that a model for area should satisfy the shape criterion. For example the half-normal model for distances, which satisfies the shape criterion, transforms to the negative exponential model for areas, which does not satisfy the shape criterion, whereas the hazard-rate model of Hayes and Buckland (1983) retains both its parametric form and a shoulder under the transformation, although the shoulder becomes narrower. We now recommend modelling the untransformed distance data, because line transect detection functions may then be more safely carried across to point transects, thus allowing the focus of analysis to be the detection function in both cases.

3.2 Hazard-rate modelling of the detection process

There are many possible models for the detection function that may fit any given data set well. If all give similar estimates, then model selection may not be critical. However, when the observed data exhibit little or no 'shoulder', it is not uncommon that one model yields an estimated density around double that for another model. Although the development of robust, flexible models allows workers to obtain good fits to most of their data sets, it does not guarantee that the resulting density estimators have low bias. There is some value therefore in attempting to model the detection process, to provide both some insight to the likely form of the detection function and a parametric model that might be expected to fit real data well. Hazard-rate methods have proved particularly useful for this purpose, and have been developed by Schweder (1977), Butterworth (1982a), Hayes and Buckland (1983) and Buckland (1985) for line transect sampling, and by Ramsey *et al.* (1979) and Buckland (1987a) for point transect sampling. We consider only continuous hazard-rate models at this stage; discrete hazard-rate models are described in Chapter 6.

3.2.1 Line transect sampling

At any one point in time, there is a 'hazard' that an object will be detected by the observer, which is a function of the distance r separating the object and observer. If the object is on the line, the observer will be moving directly towards it, so that r decreases quite quickly. The farther the object is from the line, the slower the rate of decrease in distance r, so that the observer has more time to detect the object at larger distances. Hazard-rate analysis models this effect, and also allows the hazard to depend on the angle of the object from the observer's direction of travel.

Suppose an object is at perpendicular distance x from the transect line, and let the length of transect line between the observer and the point of closest approach to the object be z, so that r, the distance between the observer and the object, satisfies $r^2 = x^2 + z^2$ (Fig. 1.5). Suppose also that the observer approaches from a remote point on the transect so that z may be considered to decrease from ∞ to 0, and assume for simplicity that the object cannot be detected once the observer has passed his/her point of closest approach. Let

$$h(z, x)dz = \text{pr}\{\text{object sighted while observer is in } (z, z - dz) \mid \text{not sighted while observer is between } \infty \text{ and } z\}$$

and

$p(z, x) = \text{pr}\{\text{object not sighted while observer is between } \infty \text{ and } z\}$

where both probabilities are conditional on the perpendicular distance x. Solving the forwards equations

$$p(z - dz, x) = p(z, x)\{1 - h(z, x)dz\}$$

for $p(z, x)$, and setting $p(\infty, x) = 1$, yields

$$p(z, x) = \exp\left\{-\int_z^\infty h(v, x)dv\right\}$$

so that

$$\begin{aligned} g(x) &= 1 - p(0, x) \\ &= 1 - \exp\left\{-\int_0^\infty h(z, x)dz\right\}, \quad 0 \leqslant x < \infty \end{aligned}$$

Changing the variable of integration from z to r gives

$$g(x) = 1 - \exp\left\{-\int_x^\infty \frac{r}{\sqrt{(r^2 - x^2)}} k(r, x)dr\right\}$$

where $k(r, x) \equiv h\{\sqrt{(r^2 - x^2)}, x\}$

Time could be incorporated in the model, but for a continuous hazard-rate process, there is little value in doing so provided that the speed of the observer is not highly variable. Otherwise the development has been general up to this point. To progress further, it is necessary to restrict the form of the hazard. A plausible hazard should satisfy the following conditions:

1. $k(0, 0) = \infty$;
2. $k(\infty, x) = 0$;
3. $k(r, x)$ is non-increasing in r for any fixed x.

For example suppose that the hazard belongs to the family defined by:

$$\int_x^\infty \frac{r}{\sqrt{(r^2 - x^2)}} k(r, x)dr = (x/\sigma)^{-b} \quad \text{for some } \sigma \text{ and } b \qquad (3.6)$$

Hayes and Buckland (1983) give two hazards from this family. In the first, the hazard of detection is a function of r alone:

$$k(r, x) = cr^{-d}, \; r \geqslant x$$

so that $b = d - 1$ and $\sigma = \left\{ \dfrac{c\Gamma[(d-1)/2]\Gamma(0.5)}{2\Gamma(d/2)} \right\}^{1/(d-1)}$

The second hazard function allows the hazard of detection to be greater for objects directly ahead of the observer than for objects to the side. In practice this may arise if an object at distance r is more likely to flush when the observer moves towards it, or if the observer concentrates search effort in the forward direction. The functional form of the second hazard is:

$$k(r, x) = c \cdot r^{-d} \cos\theta, \qquad \text{where } \sin\theta = x/r$$

so that $b = d - 1$ and $\sigma = \left\{ \dfrac{c}{d-1} \right\}^{1/(d-1)}$

The family of hazards defined by Equation 3.6, to which the above two belong, yields the detection function

$$g(x) = 1 - \exp[-(x/\sigma)^{-b}] \tag{3.7}$$

This is the hazard-rate model derived by Hayes and Buckland (1983) and investigated by Buckland (1985), although the above parameterization is slightly different from theirs, and has better convergence properties. The parameter b is a shape parameter, whereas σ is a scale parameter. The model should provide a good representation of the 'true' detection function when the hazard process is continuous, sighting (or auditory) conditions are homogeneous, and visibility (or sound) falls off with distance according to a power function, although it appears to be robust when these conditions are violated. It may be shown that $g'(0) = 0$ for $b > 0$, which covers all parameter values for which the detection function is a decreasing function. Hence the above two hazards which are sharply 'spiked' (the derivative of the hazard with respect to r, evaluated at $r = 0$, is infinite) give rise to a detection function that always satisfies the shape criterion of Burnham *et al.* (1980). For untruncated data the detection function integrates to a finite value only if $b > 1$. For truncated data, the model has a long tail and a narrow shoulder if $b < 1$, and convergence problems may be encountered for

extreme data sets. These problems may be avoided and analyses are more robust when the constraint $b \geq 1$ is imposed (Buckland 1987b). Equation 3.7 is plotted for a range of values for the shape parameter likely to be encountered in real data sets in Fig. 2.5b.

Although in this book we describe the model of Equation 3.7 as 'the' hazard-rate model, any detection function may be described as a hazard-rate model in the sense that a (possibly implausible) hazard exists from which the detection function could be derived. Equation 3.7 is sometimes referred to as the complementary log-log model, a label which is both more accurate and more cumbersome.

3.2.2 Point transect sampling

For point transect sampling, the hazard-rate formulation is simpler, since there is only one distance, the sighting distance r, to model. The probability of detection is no longer a function of distance moved along the transect, but of time spent at the point. Define the hazard function $k(r, t)$ to be such that

$$k(r, t)dt = \text{pr}\{\text{an object at distance } r \text{ is detected during } (t + dt) | \text{it is not detected during } (0, t)\}$$

Then the detection function becomes:

$$g(r) = 1 - \exp\left[-\int_0^T k(r, t)dt\right]$$

where T is the recording time at each point. If the observer is assumed to search with constant effort during the recording period, then $k(r, t) = k(r)$, independent of t, so that

$$g(r) = 1 - \exp[-k(r)T] \tag{3.8}$$

If the hazard is assumed to be of the form $k(r) = cr^{-d}$, then

$$g(r) = 1 - \exp[-(r/\sigma)^{-b}]$$

where $b = d$ and $\sigma = (cT)^{1/b}$. The effect of increasing the time spent at each point is therefore to increase the scale parameter. This widens the shoulder on the detection function, making it easier to fit. Scott and Ramsey (1981) plotted the changing shape of a detection function as time spent at the point increases from four to 32 minutes. The disadvantages of choosing T large are that assumptions are more likely to

be violated (Section 5.9), and after a few minutes, the number of new detections per minute will become small.

The parametric form of the above detection function is identical to that derived for line transects (Equation 3.7). Moreover, if sightings are squared prior to analysis, the parametric form remains unaltered, so that maximum likelihood estimation is invariant to the transformation to squared distances. This property is not shared by other widely used models for the detection function. Burnham *et al.* (1980: 195) suggested squaring distances, to allow standard line transect software to be used for analysing point transect data, but we now advise against this strategy (Buckland 1987a).

3.3 The key function formulation for distance data

Most formulations proposed for the probability density of distance data from line or point transects may be categorized into one of two groups. If there are theoretical reasons for supposing that the density has a given parametric form, then parametric modelling may be carried out. Otherwise, robust or non-parametric procedures such as Fourier series, splines, kernel methods or polynomials might be preferred. In practice it may be reasonable to assume that the true density function is close to a known parametric form, yet systematic departures can occur in some data sets. In this instance, a parametric procedure may not always give an adequate fit, yet a non-parametric method may be too flexible, perhaps giving very different fits to two related data sets from a single study, due to small random fluctuations in the data. An example of the latter occurs when a one term Fourier series model is selected for one data set and a two term model for a second. The second data set might be slightly larger, or show a slightly smaller shoulder; both increase the likelihood of rejecting the one term fit. Bias in estimation of $f(0)$ can be a strong function of the number of Fourier series terms selected (Buckland 1985), so that comparisons across data sets may be misleading. The technique described by Buckland (1992a, b) and summarized below incorporates knowledge of the likely shape of the density function, whether theoretical or from past experience, and allows polynomial or Fourier series adjustments to be made, to ensure a good fit to the data.

Simple polynomials have been used for fitting line transect data by Anderson and Pospahala (1970). However, low order simple polynomials may have unsuitable shapes. By taking the best available parametric form for the density, $\alpha(y)$, and multiplying it by a simple polynomial, this shortcoming is removed. We call $\alpha(y)$ the **key function**. If it is a

good fit, it needs no adjustment; the worse the fit, the greater the adjustments required.

When the key function is the untruncated normal (or half-normal) density, Hermite polynomials (Stuart and Ord 1987: 220–7) are orthogonal with respect to the key, and may therefore be preferred to simple polynomials. Hermite polynomials are usually fitted by the method of moments, leading to unstable behaviour when the number of observations is not large or when high order terms are included. Buckland (1985) overcame these difficulties by using numerical techniques to obtain maximum likelihood estimates of the polynomial coefficients. These procedures have the further advantage that the Gram-Charlier type A and the Edgeworth formulations yield identical curves; for the method of moments, they do not (Stuart and Ord 1987: 222–5).

Let the density function be expressed as

$$f(y) \doteq \frac{\alpha(y)}{\beta} \cdot \left[1 + \sum_{j=1}^{m} a_j \cdot p_j(y_s)\right]$$

where $\alpha(y)$ is a parametric key, containing k parameters (usually 0, 1 or 2);

$$p_j(y_s) = \begin{cases} y_s^j, \text{ if a simple polynomial is desired, or} \\ H_j(y_s), \text{ the } j\text{th Hermite polynomial, } j = 1, \ldots, m, \text{ or} \\ \cos(j\pi y_s), \text{ if a Fourier series is desired;} \end{cases}$$

y_s is a standardized y value (see below);

$$a_j \begin{cases} = 0 \text{ if term } j \text{ of } p_j(y_s) \text{ is not used in the model, or} \\ \text{is estimated by maximum likelihood;} \end{cases}$$

β is a normalizing function of the parameters (key parameters and series coefficients) alone.

It is necessary to scale the observed distances. For the simple polynomial formulation, estimation is invariant to choice of scale, but the operation is still necessary to avoid numeric problems when fitting the model. If the key function is parameterized such that a single key parameter, σ say, is a scale parameter, y_s may be found as y/σ for each observation. If the parameters of the key function are fully integrated into the estimation routine, σ can be estimated by maximum likelihood (see below). Otherwise the key function may be fitted by maximum likelihood in the absence of polynomial adjustments, and subsequent

fitting of polynomial terms can be carried out conditional on those key parameter estimates. For the Fourier series formulation, analyses are conditional on w, the truncation point, and $y_s = y/w$. In practice, it is simpler to use this standardization for all models, a strategy used in DISTANCE.

For line transect sampling, the standard form of the Fourier series model is obtained by setting the key function equal to the uniform distribution, so that $\alpha(y) = 1/w$. Used in conjunction with simple polynomials, this key gives the method of Anderson and Pospahala (1970). The standard form of the Hermite polynomial model arises when the key function is the half-normal. Point transect keys are found by multiplying their line transect counterparts by y (or, equivalently, $2\pi y$). The key need not be a valid density function. For the half-normal line transect key, define $\alpha(y) = \exp[-(y/\sigma)^2/2]$, and absorb the denominator of the half-normal density, $\sqrt{(\pi\sigma^2/2)}$, into β. In general, absorb any part of the key that is a function of the parameters alone into β.

For line transect sampling, the detection function is generally assumed to be symmetric about the line. Similarly for point transect sampling, detection probability is assumed to be independent of angle. The detection function may be envisaged as a continuous function on $(-w, +w)$; for line transects, negative distances would correspond say to sightings to the left of the line and positive to the right, and for point transects, this function can be thought of as a section through the detection 'dome', passing through the centre or point. The function is assumed to be symmetric about zero (although analyses are robust to this assumption). Hence only cosine terms are used for the Fourier series model, and only polynomials of even order for polynomial models, so that the detection function is an even function. In the case of the Hermite polynomial model, the parameter of the half-normal key corresponds to the second moment term, so that the first polynomial to be tested is of order four if terms are tested for inclusion in a sequential manner. The first adjustment to the half-normal fit therefore adjusts for kurtosis. It may be that kurtosis for the true detection function is close to that for the normal distribution, but a higher order moment may be very different. In this case it may be more profitable to test for inclusion of terms in a stepwise manner: select all terms of even order up to an arbitrary order, say 10, and include at the first step that term which gives the greatest increase in the value of the likelihood. Next include the term that gives the greatest improvement when fitted simultaneously with the first term selected. Continue until a likelihood ratio test indicates that no significant improvement in the fit has been achieved.

When the key function is not normal and testing is sequential, it is less clear which polynomial term should be tested first. Any key

will contain a parameter, or a function of parameters, that corresponds to scale, so a possible rule is to start with the term of order four, whatever the key. An alternative rule is to start with the term of order $2 \cdot (k + 1)$, where k is the number of key parameters. We advise against the use of keys with more than two parameters. For stepwise testing, all even terms down to order two and up to an arbitrary limit can be included.

3.4 Maximum likelihood methods

We concentrate here on the likelihood function for the detection distances, y_i, $i = 1, \ldots, n$, conditional on n. If the full data set was to be modelled in a comprehensive way, then the probability of realizing the data $\{n, y_1, \ldots, y_n, s_1, \ldots, s_n\}$ might be expressed as

$$\Pr(n, y_1, \ldots, y_n, s_1, \ldots, s_n) = \Pr(n) \cdot \Pr(y_1, \ldots, y_n, s_1, \ldots, s_n | n)$$
$$= \Pr(n) \cdot \Pr(y_1, \ldots, y_n | n) \cdot \Pr(s_1, \ldots, s_n | n, y_1, \ldots, y_n)$$

Thus, inference on the distances y_i can be made conditional on n, and inference on the cluster sizes s_i can be conditional on n and the y_i. This provides the justification for treating estimation of D (with $g_0 = 1$) as a series of three univariate problems.

Rao (1973) and Burnham *et al.* (1980) present maximum likelihood estimation methods for both grouped and ungrouped distance data. Applying those techniques to the key formulation of Section 3.3 yields the following useful results.

3.4.1 Ungrouped data

Define $\mathcal{L}(\underline{\theta}) = \prod_{i=1}^{n} f(y_i)$

where y_i is the ith recorded distance, $i = 1, \ldots, n$

$\theta_1, \ldots, \theta_k$ are the parameters of the key function

$\theta_{k+j} = a_j$, $j = 1, \ldots, m$, are the parameters (coefficients) of the adjustment terms.

Then $\log_e[\mathcal{L}(\underline{\theta})] = l = \sum_{i=1}^{n} \log_e[f(y_i)] = \sum_{i=1}^{n} \log_e[f(y_i) \cdot \beta] - n \cdot \log_e \beta$

Hence $$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \log_e[f(y_i)]}{\partial \theta_j}$$

$$= \sum_{i=1}^{n} \left\{ \frac{1}{f(y_i) \cdot \beta} \cdot \frac{\partial [f(y_i) \cdot \beta]}{\partial \theta_j} \right\} - \frac{n}{\beta} \cdot \frac{\partial \beta}{\partial \theta_j}, j = 1, \ldots, k+m$$

where

$$\frac{\partial [f(y_i) \cdot \beta]}{\partial \theta_j} = \begin{cases} \alpha(y_i) \cdot \left[\sum_{j'=1}^{m} a_{j'} \cdot \frac{\partial p_{j'}(y_{is})}{\partial y_{is}} \right] \cdot \frac{\partial y_{is}}{\partial \theta_j} + \left[1 + \sum_{j'=1}^{m} a_{j'} \cdot p_{j'}(y_{is}) \right] \cdot \frac{\partial \alpha(y_i)}{\partial \theta_j}, \\ \qquad\qquad 1 \leqslant j \leqslant k \quad (3.9) \\ \alpha(y_i) \cdot p_{j-k}(y_{is}), \text{ for all } j > k \text{ for which } a_{j-k} \text{ is non-zero} \end{cases}$$

$$\frac{\partial p_{j'}(y_{is})}{\partial y_{is}} = \begin{cases} j \cdot p_{j-1}(y_{is}), \text{ with } p_0(y_{is}) = 1, \\ \quad \text{for simple and Hermite polynomials} \\ -j\pi \cdot \sin(j\pi y_s), \text{ for the Fourier series model} \end{cases}$$

When $k = 1$ and $y_{is} = y_i/\theta_1$, $\dfrac{\partial y_{is}}{\partial \theta_1} = -y_i/\theta_1^2$; when $k = 1$ and $y_{is} = y_i/w$,

$$\frac{\partial y_{is}}{\partial \theta_1} = 0$$

The equations $\partial l/\partial \theta_j = 0, j = 1, \ldots, k+m$, may be solved using for example Newton–Raphson or a simplex procedure. To change between simple and Hermite polynomials, it is merely necessary to redefine $p_j(y_s)$, $j = 1, \ldots, m$; to change between polynomial and Fourier series adjustments, the derivative of $p_j(y_s)$ with respect to y_s must also be redefined. If a different key $\alpha(y)$ is required, the only additional algebra needed to implement the method is to find $\partial \alpha(y)/\partial \theta_j$ and $\partial y_s/\partial \theta_j$, $1 \leqslant j \leqslant k$; β and $\partial \beta/\partial \theta_j$ are evaluated by numerical integration.

The Fisher information matrix per observation may be estimated by the Hessian matrix $H(\hat{\underline{\theta}})$, with jhth element

$$H_{jh}(\hat{\underline{\theta}}) = \frac{1}{n} \cdot \left[\sum_{i=1}^{n} \frac{\partial \log_e[f(y_i)]}{\partial \hat{\theta}_j} \cdot \frac{\partial \log_e[f(y_i)]}{\partial \hat{\theta}_h} \right]$$

This may be formed from quantities already calculated. If a function of the parameters, $g(\underline{\theta})$, is to be estimated by $g(\hat{\underline{\theta}})$, then

$$\widehat{\text{var}}\{g(\hat{\underline{\theta}})\} = \frac{1}{n} \cdot \left[\frac{\partial g(\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right]' [H(\hat{\underline{\theta}})]^{-1} \left[\frac{\partial g(\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right]$$

3.4.2 Grouped data

Suppose the observations y are grouped, the ith group spanning the interval (c_{i1}, c_{i2}), $i = 1, \ldots, u$. In general, the data may be truncated at either or both ends. For line and point transects, it is usual that $c_{i1} = 0$ (no left truncation) and $c_{i2} = c_{i+1,1}$, $i = 1, \ldots, u-1$. The likelihood function is now multinomial. Let the group frequencies be $n_1, \ldots, n_u$, with cell probabilities

$$\pi_i = \int_{c_{i1}}^{c_{i2}} f(y)\, dy$$

Then

$$\mathcal{L}(\underline{\theta}) = \frac{n!}{n_1! \ldots n_u!} \prod_{i=1}^{u} \pi_i^{n_i}, \text{ with } n = \sum_{i=1}^{u} n_i$$

$$\log_e[\mathcal{L}(\underline{\theta})] = l = \sum_{i=1}^{u} n_i \cdot \log_e(\pi_i) + \text{a constant}$$

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{u} \frac{n_i}{\pi_i} \cdot \frac{\partial \pi_i}{\partial \theta_j}, \quad j = 1, \ldots, k+m$$

Define $P_i = \pi_i \cdot \beta$, so that

$$\frac{\partial \pi_i}{\partial \theta_j} = \frac{1}{\beta} \cdot \left[\frac{\partial P_i}{\partial \theta_j} - \frac{\partial \beta}{\partial \theta_j} \cdot \pi_i\right]$$

Then if P_i and $\partial P_i/\partial \theta_j$, $j = 1, \ldots, k+m$, $i = 1, \ldots, u$, can be found,

$$\beta = \sum_{i=1}^{u} P_i \qquad \text{and} \qquad \frac{\partial \beta}{\partial \theta_j} = \sum_{i=1}^{u} \frac{\partial P_i}{\partial \theta_j}$$

Given parameter estimates, the P_i may be evaluated by numerically integrating the numerator, $f(y) \cdot \beta$, of the density function:

$$P_i = \pi_i \cdot \beta = \int_{c_{i1}}^{c_{i2}} f(y) \cdot \beta dy$$

Similarly,

$$\frac{\partial P_i}{\partial \theta_j} = \int_{c_{i1}}^{c_{i2}} \frac{\partial \{f(y) \cdot \beta\}}{\partial \theta_j} dy$$

and may be found using numerical integration on

$$\frac{\partial [f(y) \cdot \beta]}{\partial \theta_j} = \begin{cases} \alpha(y) \cdot \left[\sum_{j'=1}^{m} a_{j'} \cdot \frac{\partial p_{j'}(y_s)}{\partial y_s} \right] \cdot \frac{\partial y_s}{\partial \theta_j} \\ + \left[1 + \sum_{j'=1}^{m} a_{j'} \cdot p_{j'}(y_s) \right] \cdot \frac{\partial \alpha(y)}{\partial \theta_j}, \quad 1 \leqslant j \leqslant k \\ \alpha(y) \cdot p_{j-k}(y_s), \text{ for all } j > k \text{ for which } a_{j-k} \text{ is non-zero} \end{cases} \tag{3.10}$$

The implications of changing between simple and Hermite polynomials or between polynomial and Fourier series adjustments, and of changing the key function, are identical to the case of ungrouped data.

Again, a robust iterative procedure is required to maximize the likelihood. Variances follow as for ungrouped data, except that the information matrix per observation, $I(\underline{\theta})$, now has jhth element

$$I_{jh}(\underline{\theta}) = \sum_{i=1}^{u} \frac{1}{\pi_i} \cdot \frac{\partial \pi_i}{\partial \theta_j} \cdot \frac{\partial \pi_i}{\partial \theta_h}, \; j, h = 1, \ldots, k + m$$

All of these quantities are now available, and so a function of the parameters $g(\underline{\theta})$ is estimated by $g(\hat{\underline{\theta}})$ with variance

$$\widehat{\text{var}}\{g(\hat{\underline{\theta}})\} = \frac{1}{n} \cdot \left[\frac{\partial g(\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}} \right]' [I(\hat{\underline{\theta}})]^{-1} \left[\frac{\partial g(\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}} \right]$$

If data are analysed both grouped and ungrouped, and the respective maxima of the likelihood functions are compared, the constant combinatorial term in the likelihood for grouped data should be omitted. As the number of groups tends to infinity and interval width tends to zero,

the likelihood for grouped data tends to that for ungrouped data, provided the constant is ignored.

3.4.3 Special cases

Suppose no polynomial or Fourier series adjustments are required. The method then reduces to a straightforward fit of a parametric density. The above results hold, except the range of j is now from 1 to k, and for ungrouped data Equation 3.9 reduces to

$$\frac{\partial[f(y_i) \cdot \beta]}{\partial\theta_j} = \frac{\partial\alpha(y_i)}{\partial\theta_j}, j = 1, \ldots, k$$

For grouped data, Equation 3.10 reduces to the above, with the suffix i deleted from y.

For the Hermite polynomial model, it is sometimes convenient to fit the half-normal model as described in the previous paragraph, and then to condition on that fit when making polynomial adjustments. For the standard Fourier series model, the key is a uniform distribution on (0, w), where w is the truncation point, specified before analysis. In each of these cases, the adjustment terms are estimated conditional on the parameters of the key. Thus Equation 3.9 reduces to

$$\frac{\partial[f(y_i) \cdot \beta]}{\partial\theta_j} = \alpha(y_i) \cdot p_{j-k}(y_{is}), \text{ for non-zero } a_{j-k} \text{ and } k < j \leq k + m$$

Equation 3.10 reduces similarly, but with suffix i deleted; otherwise results follow through exactly as before, but with j restricted to the range $k + 1$ to $k + m$. This procedure is necessary whenever the uniform key is selected. For keys that have at least one parameter estimated from the data, the conditional maximization is useful only if simultaneous maximization across all parameters fails to converge.

A third option that is sometimes useful is to refit the key, conditional on polynomial or Fourier series adjustments. Equation 3.9 then becomes

$$\frac{\partial[f(y_i) \cdot \beta]}{\partial\theta_j} = \alpha(y_i) \cdot \left[\sum_{j'=1}^{m} a_{j'} \cdot \frac{\partial p_{j'}(y_{is})}{\partial y_{is}}\right] \cdot \frac{\partial y_{is}}{\partial\theta_j}$$

$$+ \left[1 + \sum_{j'=1}^{m} a_{j'} \cdot p_{j'}(y_{is})\right] \cdot \frac{\partial\alpha(y_i)}{\partial\theta_j}, 1 \leq j \leq k$$

and similarly for Equation 3.10, but minus the suffix i throughout. The range of j is from 1 to k; otherwise results follow exactly as before.

3.4.4 The half-normal detection function

If the detection function is assumed to be half-normal, and the data are both ungrouped and untruncated, the above approach leads to closed form estimators and a particularly simple analysis for both line and point transect sampling. Suppose the detection function is given by $g(y) = \exp(-y^2/2\sigma^2)$, $0 \leq y < \infty$. We consider the derivation for line transects ($y = x$) and point transects ($y = r$) separately.

(a) *Line transects* With no truncation, the density function of detection distances is $f(x) = g(x)/\mu$, where

$$\mu = \int_0^\infty g(x)dx = \int_0^\infty \exp(-x^2/2\sigma^2)dx = \sqrt{\frac{\pi\sigma^2}{2}}$$

Given n detections, the likelihood function is

$$\mathcal{L} = \prod_{i=1}^{n} \{g(x_i)/\mu\} = \left\{ \prod_{i=1}^{n} \exp(-x_i^2/2\sigma^2) \right\} / \mu^n$$

so that $l = \log_e(\mathcal{L}) = -\sum_{i=1}^{n} \{x_i^2/2\sigma^2\} - n \cdot \log_e\{\sqrt{(\pi\sigma^2/2)}\}$

Differentiating l with respect to σ^2 (i.e. $k = 1$, $\theta_1 = \sigma^2$ and $m = 0$ in terms of the general notation) and setting the result equal to zero gives:

$$\frac{dl}{d\sigma^2} = \sum_{i=1}^{n} x_i^2/2\sigma^4 - n/2\sigma^2 = 0$$

so that $\hat{\sigma}^2 = \sum_{i=1}^{n} x_i^2/n$

Then $\hat{f}(0) = 1/\hat{\mu} = \sqrt{\{2/(\pi\hat{\sigma}^2)\}}$

By evaluating the Fisher information matrix, we get

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n}$$

from which

$$\text{var}\{\hat{f}(0)\} = \frac{1}{n\pi\sigma^2} = \frac{\{f(0)\}^2}{2n}$$

Equation 3.4, with each of $E(s)$, c and g_0 set equal to one, yields

$$D = \frac{E(n) \cdot f(0)}{2L}$$

from which

$$\hat{D} = \frac{n \cdot \hat{f}(0)}{2L} = \left[2\pi L^2 \sum_{i=1}^{n} x_i^2/n^3 \right]^{-0.5}$$

The methods of Section 3.7 yield an estimated variance for $\hat{D}$.

Quinn (1977) investigated the half-normal model, and derived an unbiased estimator for $f(0)$.

(b) *Point transects* The density function of detection distances is given by $f(r) = 2\pi r \cdot g(r)/\nu$. For the half-normal detection function,

$$\nu = 2\pi \int_0^w r \cdot g(r)dr = 2\pi \int_0^w r \cdot \exp(-r^2/2\sigma^2)dr$$

$$= [-2\pi\sigma^2 \cdot \exp(-r^2/2\sigma^2)]_0^w = 2\pi\sigma^2\{1 - \exp(-w^2/2\sigma^2)\}$$

Because there is no truncation, $w = \infty$, so that $\nu = 2\pi\sigma^2$. Note that if we substitute $w = \sigma$ into this equation, then the expected proportion of sightings within σ of the point transect is $2\pi\sigma^2\{1 - \exp(-0.5)\}/\nu = 39\%$. This compares with 68% for line transects; thus for the half-normal model, nearly 70% of detections occur within one standard deviation of the observer for line transects, whereas less than 40% occur within this distance for point transects. This highlights the fact that expected detection distance is greater for point transects than for line transects, a difference which is even more marked if the detection function is long-tailed.

If n detections are made, the likelihood function is given by

$$\mathcal{L} = \prod_{i=1}^{n} \{2\pi r_i \cdot g(r_i)/\nu\} = \left\{ \prod_{i=1}^{n} r_i \cdot \exp(-r_i^2/2\sigma^2) \right\} / (\sigma^{2n})$$

so that $l = \log_e(\mathcal{L}) = \sum_{i=1}^{n} \{\log_e(r_i) - r_i^2/2\sigma^2\} - n \cdot \log_e(\sigma^2)$

This is maximized by differentiating with respect to σ^2 and setting equal to zero:

$$\frac{dl}{d\sigma^2} = \sum_{i=1}^{n} r_i^2/2\sigma^4 - n/\sigma^2 = 0$$

so that $\hat{\sigma}^2 = \sum_{i=1}^{n} r_i^2/2n$

It follows that $\hat{h}(0) = 2\pi/\hat{\nu} = 1/\hat{\sigma}^2$. Equation 3.5, with each of $E(s)$, c and g_0 set equal to one, gives

$$D = \frac{E(n) \cdot h(0)}{2\pi k}$$

so that

$$\hat{D} = \frac{n \cdot \hat{h}(0)}{2\pi k} = \frac{n^2}{\pi k \sum_{i=1}^{n} r_i^2}$$

The maximum likelihood method yields var$[\hat{h}(0)]$. The half-normal detection function has just one parameter (σ^2), so that the information matrix is a scalar. It yields

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n}$$

and

$$\text{var}\{\hat{h}(0)\} = \frac{2}{n\sigma^4} = \frac{2\{h(0)\}^2}{n}$$

Estimation of var(n) and var($\hat{D}$) is covered in Section 3.7.

3.4.5 Constrained maximum likelihood estimation

The maximization routine used by DISTANCE allows constraints to be placed on the fitted detection function. In all analyses, the constraint $\hat{g}(y) \geq 0$ is imposed. In addition, $\hat{g}(y)$ is evaluated at ten y values, y_1 to y_{10}, and the non-linear constraint $\hat{g}(y_i) \geq \hat{g}(y_{i+1})$, $i = 1, \ldots, 9$, is enforced. The user may override this constraint, or replace it by the weaker constraint that $\hat{g}(0) \geq \hat{g}(y_i)$, $i = 1, \ldots, 10$. If the same data set is analysed by DISTANCE and by TRANSECT (Laake *et al.* 1979), different estimates may be obtained; TRANSECT does not impose constraints, and in addition does not fit the Fourier series model by maximum likelihood.

DISTANCE warns the user when a constraint has caused estimates to be modified. In these instances, the analytic variance of $\hat{f}(0)$ or $\hat{h}(0)$ may be unreliable, and we recommend that the bootstrap option for variance estimation is selected.

3.5 Choice of model

The key + adjustment formulation for line and point transect models outlined above has been implemented in DISTANCE, so that a large number of models are available to the user. Although this gives great flexibility, it also creates a problem of how to choose an appropriate model. We consider here criteria that models for the detection function should satisfy, and methods that allow selection between contending models.

3.5.1 Criteria for robust estimation

Burnham *et al.* (1979, 1980: 44) identified four criteria that relate to properties of the assumed model for the detection function. In order of importance, they were model robustness, pooling robustness, the shape criterion and estimator efficiency.

(a) *Model robustness* Given that the true form of the detection function is not known except in the case of computer simulations, models are required that are sufficiently flexible to fit a wide variety of shapes for the detection function. An estimator based on such a model is termed model robust. The adoption of the key + series expansion formulation means that any parametric model can yield model robust estimation, by allowing its fit to be adjusted when the data dictate. A model of this type is sometimes called 'semiparametric'.

(b) *Pooling robustness* Probability of detection is a function of many factors other than distance from the observer or line. Weather, time of day, observer, habitat, behaviour of the object, its size and many other factors influence the probability that the observer will detect it. Conditions will vary during the course of a survey, and different objects will have different intrinsic detectabilities. Thus the recorded data are realizations from a heterogeneous assortment of detection functions. A model is pooling robust if it is robust to variation in detection probability for any given distance y. A fuller definition of this concept is given by Burnham *et al.* (1980: 45).

(c) *Shape criterion* The shape criterion can be stated mathematically as $g'(0) = 0$. In words, it states that a model for the detection function should have a shoulder. The restriction is reasonable given the nature of the sighting process. Note that the hazard-rate derivations of Section 3.2 gave rise to detection functions which possess a shoulder for all parameter values, even though sharply spiked hazards with infinite slope at zero distance were assumed. If the shape criterion is violated, robust estimation of object density is problematic if not impossible.

(d) *Estimator efficiency* Estimators that have poor statistical efficiency (i.e. that have large variances) should be ruled out. However, an estimator that is highly efficient should be considered only if it satisfies the first three criteria. High estimator efficiency is easy to achieve at the expense of bias, and the analyst should be satisfied that an estimator is unbiased, or at least that there is no reason to suppose it might be more biased than other robust estimators, before selecting on the basis of efficiency.

3.5.2 The likelihood ratio test

The requirement for adjustment terms to a given key function can be judged using likelihood ratio tests. Suppose that a fitted model has m_1 adjustment terms (Model 1). A likelihood ratio test allows an assessment of whether the addition of another m_2 term improves the adequacy of a model significantly. The null hypothesis is that Model 1, with m_1 adjustment terms, is the true model, whereas the alternative hypothesis is that Model 2 with all $m_1 + m_2$ adjustment terms is the true model. The test statistic is

$$\chi^2 = -2\log_e(\mathcal{L}_1/\mathcal{L}_2)$$

$$= -2[\log_e(\mathcal{L}_1) - \log_e(\mathcal{L}_2)]$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are the maximum values of the likelihood functions for Models 1 and 2 respectively. If Model 1 is the true model, the test statistic follows a χ^2 distribution with m_2 degrees of freedom.

The usual way to use likelihood ratio tests for line and point transect series expansion models is to fit the key function, and then fit a low order adjustment term. The adjustment would normally be a polynomial of order four or the first term in a cosine series. If it provides no significant improvement as judged by the above test, the fit of the key alone is taken. If the adjustment term does improve the fit, the next term is added (usually the polynomial of order six, or the second term of a cosine series), and a likelihood ratio test is again carried out. The process is repeated until the test is not significant, or until a maximum number of terms has been attained. This method is therefore sequential, and is the default method used by DISTANCE. The conventional significance level is 5% ($\alpha = 0.05$), so that the most recently added term is retained if the likelihood ratio statistic exceeds $\chi^2_{0.05} = 1.96^2 = 3.84$. Unless sample sizes are large, the test has rather low power, and the sequential method risks biased estimation of density and underestimation of variance. We suggest that $\alpha = 0.15$ be adopted, in which case the value 3.84 is replaced by $\chi^2_{0.15} = 2.07$.

Terms may be added in a stepwise manner, as in regression. For forward stepping, that term not yet in the model for which the χ^2 statistic from the likelihood ratio test is largest is included next, provided its test statistic is significant at the selected level. For backward stepping, the term already in the model with the smallest test statistic is dropped, unless it is significant at the α% level.

The likelihood ratio test requires that Model 1 is a special case of Model 2. The models are said to be nested or hierarchical. The following procedure is similar in character, but allows the user to select between non-hierarchical models.

3.5.3 Akaike's Information Criterion

Akaike's Information Criterion (AIC) provides a quantitative method for model selection, whether or not models are hierarchical (Akaike 1973). It treats model selection within an optimization rather than a hypothesis testing framework. Burnham and Anderson (1992) illustrated the application of AIC, and Akaike (1985) presented the theory underlying the method. AIC is defined as

$$\text{AIC} = -2 \cdot \log_e(\mathcal{L}) + 2q$$

where $\log_e(\mathcal{L})$ is the log-likelihood function evaluated at the maximum likelihood estimates of the model parameters and q is the number of parameters in the model. The first term, $-2 \cdot \log_e(\mathcal{L})$, is a measure of how well the model fits the data, while the second term is a penalty for the addition of parameters (i.e. model complexity). For a given data set, AIC is computed for each candidate model and the model with the lowest AIC is selected. Thus, AIC attempts to identify a model that fits the data well and does not have too many parameters (the principle of parsimony). For the special case of nested models and $m_2 = 1$, model selection based on AIC is exactly equivalent to a likelihood ratio test with $\chi^2_\alpha = 2.0$, which corresponds to $\alpha = 0.157$, close to the value of 0.15 recommended above for the likelihood ratio test.

For analyses of grouped data, DISTANCE omits the constant term from the multinomial likelihood when it calculates the AIC. This ensures that the AIC tends to the value obtained from analysis of ungrouped data as the number of groups tends to infinity, where each interval length tends to zero.

3.5.4 Goodness of fit

Goodness of fit can be a useful tool for model selection. Suppose the n distance data from line or point transects are split into u groups, with sample sizes $n_1, n_2, \ldots, n_u$. Let the cutpoints between groups be defined by $c_0, c_1, \ldots, c_u = w$ ($c_0 > 0$ corresponds to left truncation of the data). Suppose a model with q parameters is fitted to the data, so that the area under the estimated density function between cutpoints c_{i-1} and c_i is $\hat{\pi}_i$. Then

$$\chi^2 = \sum_{i=1}^{u} \frac{(n_i - n \cdot \hat{\pi}_i)^2}{n \cdot \hat{\pi}_i}$$

has a χ^2 distribution with $u - q - 1$ degrees of freedom if the fitted model is the true model.

Although a significantly poor fit need not be of great concern, it provides a warning of a problem in the data or the selected detection model structure, which should be investigated through closer examination of the data or by exploring other models and fitting options. Note that it is the fit of the model near zero distance that is most critical; none of the model selection criteria of goodness of fit statistics, AIC and likelihood ratio tests give special emphasis to this region. A possible criterion for selecting between models is to calculate the χ^2 goodness of fit statistic divided by its degrees of freedom for each model, and to

select the model which gives the smallest value. A disadvantage of this approach is that the value of the χ^2 statistic depends on arbitrary decisions about the number of groups into which the data are divided and on where to place the cutpoints between groups. For several reasons, we prefer the use of AIC. However, a significant goodness of fit statistic is a useful warning that the model might be poor, or that an assumption might be seriously violated.

3.6 Estimation for clustered populations

Although the general formula of Section 3.1 incorporates the case in which the detections are clusters of objects, estimation of the expected cluster size $E(s)$ is often problematic. The obvious estimator, the average size of detected clusters, may be subject to size bias; if large clusters are detectable at greater distances than small clusters, mean size of detected clusters will be biased upwards.

3.6.1 Truncation

The simplest solution is to truncate clusters that are detected far from the line. The truncation distance need not be the same as that used if the detection function is fitted to truncated perpendicular distance data; if size bias is potentially severe, truncation should be greater. To be certain of eliminating the effects of size bias, the truncation distance should correspond roughly to the width of the shoulder of the detection function. Then $E(s)$ is estimated by $\bar{s}$, the mean size of the n clusters detected within the truncation distance. Generally, a truncation distance v corresponding to an estimated probability of detection $\hat{g}(v)$ in the range 0.6 to 0.8 ensures that bias in this estimate is small. Variance of $\bar{s}$ is estimated by:

$$\widehat{\text{var}}(\bar{s}) = \frac{\sum_{i=1}^{n} (s_i - \bar{s})^2}{n(n-1)}$$

where s_i denotes the size of cluster i. This estimator remains unbiased when the individual s_i have different variances.

3.6.2 Weighted average of cluster sizes and stratification

Truncation may prove unsatisfactory if sample size is small. Quinn (1979) considered both post-stratifying detections by cluster size and

pooling across cluster size in line transect sampling. He showed that estimation of the detection function, and hence of abundance of clusters, is not compromised by pooling the data. He noted the size bias in detected clusters, and proposed the estimator

$$\hat{E}(s) = \frac{\sum_v n_v s_v \hat{f}_v(0 \mid s = s_v)}{\sum_v n_v \hat{f}_v(0 \mid s = s_v)}$$

where summation is over the recorded cluster sizes. Thus there are n_v detections of clusters of size s_v, and the effective strip width for these clusters is $1/f_v(0 \mid s = s_v)$. The estimate is therefore the average size of detected clusters, weighted by the inverse of the effective strip width at each cluster size. For point transect sampling, $\hat{h}_v(0 \mid s = s_v)$ would replace $\hat{f}_v(0 \mid s = s_v)$. As Quinn noted, if data are pooled with respect to cluster size, the $f_v(0 \mid s = s_v)$ are not individually estimated. He suggested that the effective strip width might be assumed to be proportional to the logarithm of cluster size, so that

$$\hat{E}(s) = \frac{\sum_v n_v s_v / \log_e(s_v)}{\sum_v n_v / \log_e(s_v)}$$

This method is used in the procedures developed by Holt and Powers (1982) for estimating dolphin abundance in the eastern tropical Pacific. If it is adopted, the recommendation of Quinn (1985) should be implemented: plot mean perpendicular distance as a function of cluster size to assess the functional relationship between cluster size and effective strip width. The method should not be used in conjunction with truncation of clusters at larger distances, because cluster size is then underestimated. The purpose of truncation is to restrict the mean cluster size calculation to those clusters that are relatively unaffected by size bias, so effective strip width of the retained clusters cannot be assumed proportional to the logarithm of cluster size. Clusters beyond the truncation distance are larger than average when size bias is present, so that the above weighted mean, if applied after truncating distant clusters, corrects for the effects of size bias twice.

Quinn (1985) examined further the method of post-stratifying by cluster size. He showed that the method necessarily yields a higher coefficient of variation for abundance of clusters than the above method in which data are pooled across cluster size, but found that the result

does not extend to estimates of object abundance. For his example, he concludes that the method of pooling is superior for estimating cluster abundance, and the method of post-stratification for estimating object abundance. This conclusion is likely to be true more generally. To apply the method of post-stratification, cluster size intervals should be defined so that sample size is adequate to allow estimation of $f(0)$ in each stratum. Stratification strategies relevant to this issue are discussed in more detail in Section 3.8.

3.6.3 Regression estimators

The solution of plotting mean distance y against cluster sizes was proposed by Best and Butterworth (1980), who predicted mean cluster size at zero distance, using a weighted linear regression of cluster size on distance. This suffers from the difficulty that, if the detection function has a shoulder, mean cluster size is not a function of distance until distance exceeds the width of the shoulder. Sample size is seldom sufficient to determine that a straight line fit is inadequate, so that estimated mean cluster size at zero distance is biased downwards. Because this is assumed to be an unbiased estimate of mean size of all clusters in the population, population abundance is underestimated. A solution to this problem is to replace detection distance y_i for the ith detection by $\hat{g}(y_i)$ in the regression, where $\hat{g}(y)$ is the detection function estimated from the fit of the selected model to the pooled data, and to predict mean cluster size when detection is certain ($\hat{g}(y) = 1.0$). Thus if there are n detections, at distances y_i and of sizes s_i, if $E_d(s|y)$ denotes the expected size of detected clusters at distance y, and $E(s)$ denotes the expected size of all clusters, whether detected or not (assumed independent of y), we have:

$$\hat{E}_d(s|y) = a + b \cdot \hat{g}(y)$$

where a and b are the intercept and slope respectively of the regression of s on $\hat{g}(y)$. Then

$$\hat{E}(s) = \hat{E}_d(s|y = 0) = a + b$$

and $$\widehat{\text{var}}[\hat{E}(s)] = \left[\frac{1}{n} + \frac{(1-\bar{g})^2}{\sum_{i=1}^{n}\{\hat{g}(y_i) - \bar{g}\}^2}\right] \cdot \hat{\sigma}^2$$

with $\hat{\sigma}^2$ = residual mean square

and $$\bar{g} = \frac{\sum_{i=1}^{n} \hat{g}(y_i)}{n}$$

A further problem of the regression method occurs when cluster size is highly variable, so that one or two large clusters might have large influence on the fit of the regression line. Their influence may be reduced by transformation, for example to $z_i = \log_e(s_i)$. Suppose a regression of z_i on $\hat{g}(y_i)$ yields the equation $\hat{z} = a + b \cdot \hat{g}(y)$. Thus at $\hat{g}(y) = 1.0$, mean log cluster size is estimated by $a + b$ and $E(s)$ is estimated by

$$\hat{E}(s) = \exp(a + b + \widehat{\text{var}}(\hat{z})/2)$$

where $$\widehat{\text{var}}(\hat{z}) = \left[1 + \frac{1}{n} + \frac{(1 - \bar{g})^2}{\sum_{i=1}^{n} \{\hat{g}(y_i) - \bar{g}\}^2} \right] \cdot \hat{\sigma}^2$$

$\hat{\sigma}^2$ is the residual mean square, and $\bar{g}$ is as above.
Further,

$$\widehat{\text{var}}\{\hat{E}(s)\} = \exp\{2(a + b) + \widehat{\text{var}}(\hat{z})\} \cdot \{1 + \widehat{\text{var}}(\hat{z})/2\} \cdot \widehat{\text{var}}(\hat{z})/n$$

3.6.4 Use of covariates

The pooling method, with calculation of a weighted average cluster size, may be improved upon theoretically by incorporating cluster size as a covariate in the model for the detection function. Drummer and McDonald (1987) considered replacing detection distance y in a parametric model for the detection function by y/s^{γ}, where s is size of the cluster recorded at distance y and γ is a parameter to be estimated. Although their method was developed for line transect sampling, it can also be implemented for point transects. Ramsey *et al.* (1987) included covariates for point transect sampling by relating the logarithm of effective area searched to a linear function of covariates, one of which could be cluster size; this is in the spirit of general linear models. The same approach might be applied to effective strip width in line transect sampling, although the logarithmic link function might no longer be appropriate. These methods are discussed further in Section 3.8. Quang (1991) developed a method of modelling the bivariate detection function $g(y,s)$ using Fourier series.

3.6.5 Replacing clusters by individual objects

The problems of estimating mean cluster size can sometimes be avoided by taking the sampling unit to be the object, not the cluster. Even when detected clusters show extreme selection for large clusters, this approach can yield an unbiased estimate of object abundance, provided all clusters on or near the line are detected. The assumption of independence between sampling units is clearly violated, so robust methods of variance estimation that are insensitive to failures of this assumption should be adopted. Use of resampling methods allows the line, line segment or point to be the sampling unit instead of the object, so that valid variance estimation is possible. Under this approach, results from goodness of fit tests, likelihood ratio tests and AIC should not be used for model selection, since they will yield many spurious significant results. One solution is to select a model based on an analysis of clusters, then to refit the model, with the same number of adjustment terms, to the data recorded by object. If the number of clusters detected is small, if cluster size is highly variable, or if mean cluster size is large, the method may perform poorly.

3.6.6 Some basic theory for size-biased detection of objects

We present here some basic theoretical results when detection of clusters is size-biased. In this circumstance it is necessary to distinguish between the probability distribution of cluster sizes in the population from which the sample is taken from the distribution of s in the sample. Some of these results are in the literature (e.g. Quinn 1979; Burnham *et al.* 1980; Drummer and McDonald 1987; Drummer 1990; Quang 1991).

Let the probability distribution of cluster sizes in the region sampled be $\pi(s)$, $s = 1, 2, 3, \ldots$. This distribution applies to all the clusters, not to the detected sample. If there is size bias, then the sample of detected clusters has a different probability distribution, say $\pi^*(s)$, $s = 1, 2, 3, \ldots$. Consider first line transect sampling. Let the conditional detection function be $g(x|s)$ = probability of detection at perpendicular distance x given that the cluster is of size s, and let the detection function unconditional on cluster size be $g(x)$. Denote the corresponding probability density functions (pdf) by $f(x|s)$ and $f(x)$ respectively. The conditional pdf at $x = 0$ is

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

where w need not be finite. Note also the results

$$g(x|s) = \frac{f(x|s)}{f(0|s)} \qquad \text{and} \qquad g(x) = \frac{f(x)}{f(0)}$$

which are useful in derivations of results below.

For any fixed s, cluster density is given by the result for object density in the case without clusters:

$$D(s) = \frac{E[n(s)] \cdot f(0|s)}{2L}$$

where $n(s)$ is the number of detections of clusters of size s and $D(s)$ is the true density of clusters of size s. We need not assume that c and g_0 from Equation 3.1 equal one; however, the complication of g_0 varying by s is not considered here.

The key to deriving results is to realize that

$$\pi(s) = \frac{D(s)}{D}$$

and

$$\pi^*(s) = \frac{E[n(s)]}{E(n)}$$

where

$$D = \sum_{s=1}^{\infty} D(s) \qquad \text{and} \qquad n = \sum_{s=1}^{\infty} n(s)$$

By substituting the results for D and $D(s)$ into the first equation and using the result of the second, we derive

$$\pi^*(s) = \left[\frac{f(0)}{f(0|s)}\right] \cdot \pi(s)$$

Note that well-defined marginal probabilities and distributions exist; for example,

$$g(x) = \sum_{s=1}^{\infty} g(x|s) \cdot \pi(s)$$

from which

$$f(0) = \frac{1}{\int_0^w g(x)dx} \quad \text{and} \quad D = \frac{E(n) \cdot f(0)}{2L}$$

Given the above it is just a matter of using algebra to derive results of interest. Some key results are:

$$\pi^*(s) = \frac{\left[\int_0^w g(x|s)dx\right] \cdot \pi(s)}{\sum_{s=1}^{\infty}\left[\int_0^w g(x|s)dx\right] \cdot \pi(s)} = \frac{\frac{\pi(s)}{f(0|s)}}{\sum_{s=1}^{\infty}\frac{\pi(s)}{f(0|s)}}$$

By definition, $E(s) = \sum_{s=1}^{\infty} s \cdot \pi(s)$, so that

$$E(s) = \frac{\sum_{s=1}^{\infty} f(0|s) \cdot s \cdot \pi^*(s)}{\sum_{s=1}^{\infty} f(0|s) \cdot \pi^*(s)} = \frac{\sum_{s=1}^{\infty} f(0|s) \cdot s \cdot E[n(s)]}{\sum_{s=1}^{\infty} f(0|s) \cdot E[n(s)]}$$

The validity of Quinn's (1979) estimator, given in Section 3.6.2, is now apparent. The marginal pdf satisfies

$$f(0) = \sum_{s=1}^{\infty} f(0|s) \cdot \pi^*(s) = \frac{1}{\sum_{s=1}^{\infty}\frac{\pi(s)}{f(0|s)}}$$

These results are consistent with the formulae for density of individuals either as

$$D = \frac{E(n) \cdot f(0) \cdot E(s)}{2L}$$

or as

$$D = \frac{\sum_{s=1}^{\infty} E[n(s)] \cdot f(0|s) \cdot s}{2L}$$

A bivariate approach involves modelling the bivariate detection function $g(x, s)$, perhaps using generalized linear or non-linear regression. Adopting a univariate approach, we can estimate $E(s)$ in a linear regression framework. The key quantity needed here for theoretical work is the conditional probability distribution of detected cluster size given that detection was at perpendicular distance x, symbolized $\pi^*(s|x)$. Alternative representations are

$$\pi^*(s|x) = \frac{g(x|s) \cdot \pi(s)}{\sum_{s=1}^{\infty} g(x|s) \cdot \pi(s)} \equiv \frac{g(x|s) \cdot \pi(s)}{g(x)}$$

or

$$\pi^*(s|x) = \frac{f(x|s) \cdot \pi^*(s)}{\sum_{s=1}^{\infty} f(x|s) \cdot \pi^*(s)} \equiv \frac{f(x|s) \cdot \pi^*(s)}{f(x)}$$

If $T(s)$ represents any transformation of s, then we can compute conditional (on x) properties of $T(s)$, for example

$$E[T(s)|x] = \frac{\sum_{s=1}^{\infty} T(s) \cdot g(x|s) \cdot \pi(s)}{\sum_{s=1}^{\infty} g(x|s) \cdot \pi(s)} = \sum_{s=1}^{\infty} T(s) \cdot \pi^*(s|x)$$

$$= \frac{\sum_{s=1}^{\infty} T(s) \cdot f(x|s) \cdot \pi^*(s)}{\sum_{s=1}^{\infty} f(x|s) \cdot \pi^*(s)}$$

In particular, to evaluate the reasonableness of the regression estimator of $\log_e(s)$ on $\hat{g}(x)$, we can plot $E[\log_e(s)|x]$ against $\hat{g}(x)$ or otherwise explore this relationship, including computing $\text{var}[\log_e(s)|x]$.

Similar formulae exist for point transect sampling; in fact, many are the same. The relationship between the conditional detection function $g(r|s)$ and the corresponding pdf of distances to detected clusters is now

$$f(r|s) = \frac{r \cdot g(r|s)}{\int_0^w r \cdot g(r|s)dr}$$

and

$$h(0|s) = \lim_{r \to 0} \frac{f(r|s)}{r} = \frac{1}{\int_0^w r \cdot g(r|s)dr}$$

from which

$$D(s) = \frac{E[n(s)] \cdot h(0|s)}{2\pi k}$$

Univariate results are

$$g(r) = \sum_{s=1}^{\infty} g(r|s) \cdot \pi(s)$$

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

$$h(0) = \frac{1}{\int_0^w r \cdot g(r)dr}$$

and

$$D = \frac{E(n) \cdot h(0)}{2\pi k}$$

Using these formulae we can establish that

$$\pi^*(s) = \frac{h(0)}{h(0|s)} \cdot \pi(s)$$

This is obtained from

$$\frac{D(s)}{D} = \frac{E[n(s)] \cdot h(0|s)}{E(n) \cdot h(0)}$$

The conditional and unconditional $h(\cdot)$ functions are related by

$$h(0) = \sum_{s=1}^{\infty} h(0|s) \cdot \pi^*(s) = \frac{1}{\sum_{s=1}^{\infty} \frac{\pi(s)}{h(0|s)}}$$

Also, analogous to the line transect case with $h(\cdot)$ in place of $f(\cdot)$, we have

$$\pi^*(s) = \frac{\left[\int_0^w r \cdot g(r|s)dr\right] \cdot \pi(s)}{\sum_{s=1}^{\infty}\left[\int_0^w r \cdot g(r|s)dr\right] \cdot \pi(s)} = \frac{\frac{\pi(s)}{h(0|s)}}{\sum_{s=1}^{\infty} \frac{\pi(s)}{h(0|s)}}$$

and

$$E(s) = \frac{\sum_{s=1}^{\infty} h(0|s) \cdot s \cdot \pi^*(s)}{\sum_{s=1}^{\infty} h(0|s) \cdot \pi^*(s)} = \frac{\sum_{s=1}^{\infty} h(0|s) \cdot s \cdot E[n(s)]}{\sum_{s=1}^{\infty} h(0|s) \cdot E[n(s)]}$$

In general, all the results for point transects can be obtained from the analogous results for line transects by making the following replacements: $r \cdot g(r)$ for $g(x)$, and $r \cdot g(r|s)$ for $g(x|s)$. In particular, note what happens to $\pi^*(s|r)$ in point transect sampling:

$$\pi^*(s|r) = \frac{r \cdot g(r|s) \cdot \pi(s)}{\sum_{s=1}^{\infty} r \cdot g(r|s) \cdot \pi(s)} \equiv \frac{r \cdot g(r|s) \cdot \pi(s)}{r \cdot g(r)}$$

$$\equiv \frac{g(r|s) \cdot \pi(s)}{\sum_{s=1}^{\infty} g(r|s) \cdot \pi(s)} \equiv \frac{g(r|s) \cdot \pi(s)}{g(r)}$$

This has exactly the same form as for line transects. An alternative expression is

$$\pi^*(s|r) = \frac{f(r|s) \cdot \pi^*(s)}{\sum_{s=1}^{\infty} f(r|s) \cdot \pi^*(s)} \equiv \frac{f(r|s) \cdot \pi^*(s)}{f(r)}$$

(defined to give continuity at $r = 0$), which looks structurally like the result for line transects. However, here the probability density function necessarily differs in shape from that for line transects, whereas the detection function of the previous expression might plausibly apply to both point and line transects.

3.7 Density, variance and interval estimation

3.7.1 Basic formulae

Substituting estimates into Equation 3.4, the general formula for estimating object density from line transect data is

$$\hat{D} = \frac{n \cdot \hat{f}(0) \cdot \hat{E}(s)}{2cL\hat{g}_0}$$

From Equation 3.3, the variance of $\hat{D}$ is approximately

$$\widehat{\text{var}}(\hat{D}) = \hat{D}^2 \cdot \left\{ \frac{\widehat{\text{var}}(n)}{n^2} + \frac{\widehat{\text{var}}[\hat{f}(0)]}{[\hat{f}(0)]^2} + \frac{\widehat{\text{var}}[\hat{E}(s)]}{[\hat{E}(s)]^2} + \frac{\widehat{\text{var}}[\hat{g}_0]}{[\hat{g}_0]^2} \right\}$$

Equivalent expressions for point transect sampling are

$$\hat{D} = \frac{n \cdot \hat{h}(0) \cdot \hat{E}(s)}{2\pi c k \hat{g}_0}$$

and

$$\widehat{\text{var}}(\hat{D}) = \hat{D}^2 \cdot \left\{ \frac{\widehat{\text{var}}(n)}{n^2} + \frac{\widehat{\text{var}}[\hat{h}(0)]}{[\hat{h}(0)]^2} + \frac{\widehat{\text{var}}[\hat{E}(s)]}{[\hat{E}(s)]^2} + \frac{\widehat{\text{var}}[\hat{g}_0]}{[\hat{g}_0]^2} \right\}$$

If $g_0 = 1$ (detection on the line or at the point is certain) or $E(s) = 1$ (no clusters), the terms involving estimates of these parameters are eliminated from the above equations. Generally, the constant $c = 1$, further simplifying the equations for $\hat{D}$.

To estimate the precision of $\hat{D}$, the precision of each component in the estimation equation must be estimated. Alternatively, resampling or empirical methods can be used to estimate $\text{var}(\hat{D})$ directly; some options are described in later sections. If precision is estimated component by component, then methods should be adopted for estimating mean cluster size and probability of detection on the line that provide variance estimates, $\widehat{\text{var}}[\hat{E}(s)]$ and $\widehat{\text{var}}[\hat{g}_0]$. Estimates of $f(0)$ or $h(0)$ and corresponding variance estimates are obtained from DISTANCE or similar software, using maximum likelihood theory. If objects are distributed randomly, then sample size n has a Poisson distribution, and $\widehat{\text{var}}(n) = n$. Generally, biological populations show some degree of aggregation, and Burnham *et al.* (1980: 55) suggested multiplication of the Poisson variance by two if no other approach for estimating $\text{var}(n)$ was available. If data are recorded by replicate lines or points, then a better method is to estimate $\text{var}(n)$ from the observed variation between lines and points. This method is described in the next section.

Having obtained $\hat{D}$ and $\widehat{\text{var}}(\hat{D})$, an approximate $100(1 - 2\alpha)\%$ confidence interval is given by

$$\hat{D} \pm z_\alpha \cdot \sqrt{\{\widehat{\text{var}}(\hat{D})\}}$$

where z_α is the upper $\alpha\%$ point of the $N(0,1)$ distribution. However, the distribution of $\hat{D}$ is positively skewed, and an interval with better coverage is obtained by assuming that $\hat{D}$ is log-normally distributed. Following the derivation of Burnham *et al.* (1987: 212), a $100(1 - 2\alpha)\%$ confidence interval is given by

$$(\hat{D}/C, \hat{D} \cdot C)$$

where

$$C = \exp[z_\alpha \cdot \sqrt{\{\widehat{\text{var}}(\log_e \hat{D})\}}]$$

and

$$\widehat{\text{var}}(\log_e \hat{D}) = \log_e \left[1 + \frac{\widehat{\text{var}}(\hat{D})}{\hat{D}^2} \right]$$

This is the method used by DISTANCE, except z_α is replaced by a slightly better constant that reflects the actual finite and differing degrees of freedom of the variance estimates.

The use of the normal distribution to approximate the sampling distribution of $\log_e(\hat{D})$ is generally good when each component of $\widehat{\text{var}}(\hat{D})$ (e.g. $\widehat{\text{var}}(n)$ and $\widehat{\text{var}}[\hat{f}(0)]$) is based on sufficient degrees of freedom (say 30 or more). However, sometimes the empirical estimate of var(n) in particular is based on less than 10 replicate lines, and hence on few degrees of freedom. When component degrees of freedom are small, it is better to replace z_α by a constant based on a t-distribution approximation. In this case we recommend an approach adapted from Satterthwaite (1946); see also Milliken and Johnson (1984) for a more accessible reference.

Adapting the method of Satterthwaite (1946) to this distance sampling context, z_α in the above log-based confidence interval is replaced by the two-sided alpha-level t-distribution percentile $t_{df}(\alpha)$ where df is computed as below. The coefficients of variation $\text{cv}(\hat{D})$, $\text{cv}(n)$, $\text{cv}[\hat{f}(0)]$ or $\text{cv}[\hat{h}(0)]$, and, where relevant, $\text{cv}[\hat{E}(s)]$ and $\text{cv}(\hat{g}_0)$ are required, together with the associated degrees of freedom. In general, if there are q estimated components in $\hat{D}$, then the computed degrees of freedom are

$$df = \frac{[\text{cv}(\hat{D})]^4}{\sum_{i=1}^{q} \frac{[\text{cv}_i]^4}{df_i}} = \frac{\left[\sum_{i=1}^{q} [\text{cv}_i]^2 \right]^2}{\sum_{i=1}^{q} \frac{[\text{cv}_i]^4}{df_i}}$$

This value may be rounded to the nearest integer to allow use of tables of the t-statistic.

For the common case of line transect sampling of single objects using k replicate lines, the above formula for df becomes approximately

$$df = \frac{[\text{cv}(\hat{D})]^4}{\dfrac{[\text{cv}(n)]^4}{k-1} + \dfrac{\{\text{cv}[\hat{f}(0)]\}^4}{n}}$$

(The actual degrees of freedom for var$[\hat{f}(0)]$ are n minus the number of parameters estimated in $\hat{f}(x)$.) This Satterthwaite procedure is used by program DISTANCE, rather than just the first order z_α approximation. It makes a noticeable difference in confidence intervals for small k, especially if the ratio cv(n)/cv$[\hat{f}(0)]$ is greater than one; in practice, it is often as high as two or three.

3.7.2 Replicate lines or points

Replicate lines or points may be used to estimate the contribution to overall variance of the observed sample size. In line transects, the replicate lines may be defined by the design of the survey; for example if the lines are parallel and either systematically or randomly spaced, then each line is a replicate. Surveys of large areas by ship or air frequently do not utilize such a design for practical reasons. In this case, a 'leg' might be defined as a period of search without change of bearing, or all effort for a given day or watch period. The leg will then be treated as a replicate line. When data are collected on an opportunistic basis from, for example, fisheries vessels, an entire fishing trip might be considered to be the sampling unit.

Suppose the number of detections from line or point i is n_i, $i = 1, \ldots, k$, so that $n = \sum n_i$. Then for point transects (or for line transects when the replicate lines are all the same length), the empirical estimate of var(n) is

$$\widehat{\text{var}}(n) = k \sum_{i=1}^{k} \left(n_i - \frac{n}{k} \right)^2 / (k-1)$$

For line transects, if line i is of length l_i and total line length $= L = \sum_{i=1}^{k} l_i$, then

$$\widehat{\text{var}}(n) = L \sum_{i=1}^{k} l_i \left(\frac{n_i}{l_i} - \frac{n}{L} \right)^2 / (k-1)$$

Encounter rate n/L is often a more useful form of the parameter than n alone; the variance of encounter rate is $\widehat{\text{var}}(n)/L^2$. There is a similarity here to ratio estimation in finite population sampling, except that we take all

line lengths l_i, and hence L, to be fixed (as distinct from random) values. Consequently, the variance of a ratio estimator does not apply here, and our $\widehat{\text{var}}(n/L)$ is a little different from classical finite sampling theory.

If the same line or point is covered more than once, and an analysis of the pooled data is required, then the sampling unit should still be the line or point. That is, the distance data from repeat surveys over a short time period of a given line or point should be pooled prior to analysis. Consider point transects, in which point i is covered t_i times, and in total, n_i objects are detected. Then

$$\widehat{\text{var}}(n) = T \sum_{i=1}^{k} t_i \left(\frac{n_i}{t_i} - \frac{n}{T} \right)^2 / (k-1)$$

where

$$T = \sum_{i=1}^{k} t_i$$

The formula for line transects becomes

$$\widehat{\text{var}}(n) = T_L \sum_{i=1}^{k} t_i \cdot l_i \left(\frac{n_i}{t_i \cdot l_i} - \frac{n}{T_L} \right)^2 / (k-1)$$

where

$$T_L = \sum_{i=1}^{k} t_i \cdot l_i$$

Generally, t_i will be the same for every point or line, in which case the above formulae simplify. The calculations may be carried out in DISTANCE by setting SAMPLE equal to t_i for point i (point transects) or $t_i \cdot l_i$ for line i (line transects).

The above provides empirical variance estimates for just one component of Equation 3.2, which may then be substituted in Equation 3.3. A more direct approach is to estimate object density for each replicate line or point. Define

$$\hat{D}_i = \frac{n_i \cdot \hat{E}_i(s)}{c \cdot a_i \cdot \hat{P}_{a_i} \cdot \hat{g}_{0i}}, \quad i = 1, \ldots, k$$

Then for point transects (and line transects when all lines are the same length),

$$\hat{D} = \left\{ \sum_{i=1}^{k} \hat{D}_i \right\} / k \tag{3.11}$$

and

$$\widehat{\text{var}}(\hat{D}) = \left\{ \sum_{i=1}^{k} (\hat{D}_i - \hat{D})^2 \right\} / \{k(k-1)\} \tag{3.12}$$

For line transects with replicate line i of length l_i,

$$\hat{D} = \left\{ \sum_{i=1}^{k} l_i \hat{D}_i \right\} / L \tag{3.13}$$

and

$$\widehat{\text{var}}(\hat{D}) = \sum_{i=1}^{k} \{l_i(\hat{D}_i - \hat{D})^2\} / \{L(k-1)\} \tag{3.14}$$

In practice, sample size is seldom sufficient to allow this approach, so that resampling methods such as the bootstrap and the jackknife are required.

3.7.3 The jackknife

Resampling methods start from the observed data and sample repeatedly from them to make inferences. The jackknife (Gray and Schucany 1972; Miller 1974) is carried out by removing each observation in turn from the data, and analysing the remaining data. It could be implemented for line and point transects by dropping each individual sighting from the data in turn, but it is more useful to define replicate points or lines, as above. The following development is for point transects, or line transects when the replicate lines are all of the same length.

First, delete all data from the first replicate point or line, so that sample size becomes $n - n_1$ and the number of points or lines becomes $k - 1$. Estimate object density using the reduced data set, and denote the estimate by $\hat{D}_{(1)}$. Repeat this step, reinstating the dropped point or line and removing the next, to give estimates $\hat{D}_{(i)}$, $i = 1, \ldots, k$. Now calculate the **pseudovalues:**

$$\hat{D}^{(i)} = k \cdot \hat{D} - (k-1) \cdot \hat{D}_{(i)},\ i = 1, \ldots, k \qquad (3.15)$$

These pseudovalues are treated as k replicate estimators of density, and Equations 3.11 and 3.12 yield a jackknife estimate of density and variance:

$$\hat{D}_J = \left\{ \sum_{i=1}^{k} \hat{D}^{(i)} \right\} / k$$

and

$$\widehat{\text{var}}_J(\hat{D}_J) = \left\{ \sum_{i=1}^{k} (\hat{D}^{(i)} - \hat{D}_J)^2 \right\} / \{k(k-1)\}$$

For line transects in general, Equation 3.15 is replaced by

$$\hat{D}^{(i)} = \{L \cdot \hat{D} - (L - l_i) \cdot \hat{D}_{(i)}\} / l_i,\ i = 1, \ldots, k$$

and the jackknife estimate and variance are found by substitution into Equations 3.13 and 3.14:

$$\hat{D}_J = \left\{ \sum_{i=1}^{k} l_i \hat{D}^{(i)} \right\} / L$$

and

$$\widehat{\text{var}}_J(\hat{D}_J) = \sum_{i=1}^{k} \{l_i(\hat{D}^{(i)} - \hat{D}_J)^2\} / \{L(k-1)\}$$

An approximate $100(1 - 2\alpha)\%$ confidence interval for density D is given by

$$\hat{D}_J \pm t_{k-1}(\alpha) \cdot \sqrt{\{\widehat{\text{var}}_J(\hat{D}_J)\}}$$

where $t_{k-1}(\alpha)$ is from Student's t-distribution with $k - 1$ degrees of freedom.

This interval may have poor coverage when the number of replicate lines is small; Buckland (1982) found better coverage using

$$\hat{D} \pm t_{k-1}(\alpha) \cdot \sqrt{\{\widehat{\text{var}}_J(\hat{D})\}}$$

where $\hat{D}$ is the estimated density from the full data set and

$$\widehat{\text{var}}_J(\hat{D}) = \sum_{i=1}^{k} \{l_i(\hat{D}^{(i)} - \hat{D})^2\} / \{L(k-1)\}$$

The jackknife provides a strictly balanced resampling procedure. However there seems little justification for assuming that the pseudovalues are normally distributed, and the above confidence intervals may be poor when the number of replicate lines or points is small. Further there is little or no control over the number of resamples taken; under the above procedure, it is necessarily equal to the number of replicate lines or points k, and performance may be poor when k is small. Thirdly a resample can never be larger than the original sample, and will always be smaller unless there are no sightings on at least one of the replicate lines or points. The bootstrap therefore offers greater flexibility and robustness.

3.7.4 The bootstrap

The bootstrap (Efron 1979) provides a powerful yet simple method for variance and interval estimation. Consider first the non-parametric bootstrap, applied in the most obvious way to a line transect sample. Suppose the data set comprises n observations, $y_1, \ldots, y_n$, and the probability density evaluated at zero, $f(0)$, is to be estimated. Then a bootstrap sample may be generated by selecting a sample of size n **with replacement** from the observed sample. An estimate of $f(0)$ is found from the bootstrap sample using the same model as for the observed sample. A second bootstrap sample is then taken, and the process repeated. Suppose in total B samples are taken. Then the variance of $\hat{f}(0)$ is estimated by the sample variance of bootstrap estimates of $f(0)$, $\hat{f}_i(0)$, $i = 1, \ldots, B$ (Efron 1979). The percentiles of the distribution of bootstrap estimates give approximate confidence limits for $f(0)$ (Buckland 1980; Efron 1981). An approximate $100(1 - 2\alpha)\%$ central confidence interval is given by $[\hat{f}_{(j)}(0), \hat{f}_{(k)}(0)]$, where $j = (B + 1)\alpha$ and $k = (B + 1)(1 - \alpha)$ and $\hat{f}_{(i)}(0)$ denotes the ith smallest bootstrap estimate (Buckland 1984). To yield reliable confidence intervals, the number of bootstrap samples B should be at least 200, and preferably in the range 400–1000, although around 100 are adequate for estimating standard errors. The value of B may be chosen so that j and k are integer, or j and k may be rounded to the nearest integer values, or interpolation may be used between the ordered values that bracket the required percentile. Various modifications to the percentile method have been proposed, but the simple method is sufficient for our purposes.

The parametric bootstrap is applied in exactly the same manner, except that the bootstrap samples are generated by taking a random sample of size n from the fitted probability density, $\hat{f}(y)$.

If no polynomial or Fourier series adjustments are made to the fit of a parametric probability density, the above implementation of the bootstrap (whether parametric or non-parametric) yields variance estimates for $\hat{f}(0)$ close to those obtained using the information matrix. Since the bootstrap consumes considerably more computer time (up to B times that required by an analytical method), it would not normally be used in this case. When adjustments are made, precision as measured by the information matrix is conditional on the number of polynomial or Fourier series terms selected by the stopping rule (e.g. a likelihood ratio test). The Fourier series model in particular gives analytical standard errors that are strongly correlated with the number of terms selected (Buckland 1985). The above implementation of the bootstrap avoids this problem by applying the stopping rule independently to each bootstrap data set so that variation arising from estimating the number of terms required is accounted for (Buckland 1982).

In practice the bootstrap is usually more useful when the sampling unit is a replicate line or point, as for the jackknife method. The simplest procedure is to sample with replacement from the replicate lines or points using the non-parametric bootstrap. Unlike the jackknife, the sample need not be balanced, but a degree of balance may be forced by ensuring that each replicate line or point is used exactly B times in the B bootstrap samples (Davison *et al.* 1986). Density D is estimated from each bootstrap sample, and the estimates are ordered, to give $\hat{D}_{(i)}$, $i = 1, \ldots, B$. Then

$$\hat{D}_B = \left\{ \sum_{i=1}^{B} \hat{D}_{(i)} \right\} / B$$

and

$$\widehat{\text{var}}_B(\hat{D}_B) = \left\{ \sum_{i=1}^{B} (\hat{D}_{(i)} - \hat{D}_B)^2 \right\} / (B - 1)$$

while a $100(1 - 2\alpha)\%$ confidence interval for D is given by $[\hat{D}_{(j)}, \hat{D}_{(k)}]$, with $j = (B + 1)\alpha$ and $k = (B + 1)(1 - \alpha)$ as above. (Note that the estimates do not need to be ordered if a confidence interval is not required.) The estimate based on the original data set, $\hat{D}$, is usually used in preference to the bootstrap estimate $\hat{D}_B$, with $\text{var}(\hat{D})$ estimated by $\widehat{\text{var}}_B(\hat{D}_B)$.

If an automated model selection procedure is implemented, for example using AIC, the bootstrap allows model selection to take place in each individual replicate. Thus the variability between bootstrap estimates of density reflects uncertainty due to having to estimate which model is appropriate. In other words, the bootstrap variance incorporates a component for model misspecification bias. By applying the full estimation procedure to each replicate, components of the variance for estimating the number of adjustment terms and for estimating $E(n)$, $E(s)$ and g_0 (where relevant) are all automatically incorporated. An example of such an analysis is given in Chapter 5.

A common misconception is that no model assumptions are made when using the non-parametric bootstrap. However, the sampling units from which resamples are drawn are assumed to be independently and identically distributed. If the sampling units are legs of effort, then each leg should be randomly located and independent of any other leg. In practice, this is seldom the case, but legs should be defined that do not seriously violate the assumption. For example, in marine line transect surveys, the sampling effort might be defined as all effort carried out in a single day. The overnight break in effort will reduce the dependence in the data between one sampling unit and the next, and the total number of sampling units should provide adequate replication except for surveys of short duration. It is wrong to break effort into small units and to bootstrap on those units. This is because the assumption of independence can be seriously violated, leading to bias in the variance estimate. If transect lines are designed to be perpendicular to object density contours, each line should be a sampling unit; subdivision of the lines may lead to overestimation of variance. In the case of point transects, if points are positioned along lines, then each line of points should be considered a sampling unit. If points are randomly distributed or evenly distributed throughout the study area, then individual points may be taken as sampling units. If a single line or point is covered more than once, and an analysis of the pooled data is required, the sampling unit should still be the line or point; it is incorrect to analyse the data as if different lines or points had been covered on each occasion. Analysis of such data is addressed in Section 3.7.2.

3.7.5 A finite population correction factor

We denote the size of the surveyed area, within distance w of the line or point, by a. If the size of the study area is A, a known proportion a/A is sampled. Moreover, a finite population of objects, N, exists in the area. Thus the question arises of whether a finite population cor-

rection (fpc) adjustment should be made to sampling variances. We give here a few thoughts on this matter.

Assume that there is no stratification (or that we are interested in results for a single stratum). Then for strip transect or plot sampling, fpc = $1 - a/A$. The adjusted variance of $\hat{N}$ is

$$\text{var}(\hat{N}) \cdot (1 - a/A)$$

where $\text{var}(\hat{N})$ is computed from infinite population theory. In distance sampling, not all the objects are detected in the sampled area a, so that the fpc differs from $1 - a/A$. Also, no adjustment is warranted to $\text{var}(\hat{P}_a)$ because this estimator is based on the detection distances, which conceptually arise from an infinite population of possible distances, given random placement of lines or points, or different choices of sample period.

Consider first the case where objects do not occur in clusters, and the following simple formula applies:

$$\hat{N} = A \cdot \frac{n}{a \cdot \hat{P}_a}$$

for which

$$[\text{cv}(\hat{N})]^2 = [\text{cv}(n)]^2 + [\text{cv}(\hat{P}_a)]^2$$

The fpc is the same whether it is applied to coefficients of variation or variances. Heuristic arguments suggest that the fpc might be estimated by $1 - n/\hat{N}$ or by $1 - (a \cdot \hat{P}_a)/A$. These are clearly identical. In the case of a census of sample plots (or strips), $\hat{P}_a \equiv 1$ and the correct fpc is obtained. For the above simple case of distance sampling, $\text{cv}(\hat{N})$ corrected for finite population sampling is

$$[\text{cv}(\hat{N})]^2 = [\text{cv}(n)]^2 \cdot \left[1 - \frac{a \cdot \hat{P}_a}{A}\right] + [\text{cv}(\hat{P}_a)]^2$$

This fpc is seldom large enough to make any difference. When it is, then the assumptions on which it is based are likely to be violated. For the correction $1 - (a \cdot \hat{P}_a)/A$ to be valid, the surveyed areas within distance w of each line or point must be non-overlapping. Further, it must be assumed that an object cannot be detected from more than one

line or point; if objects are mobile, the fpc $1 - n/\hat{N}$ is arguably inappropriate.

If objects occur in clusters, correction is more complicated. First consider when there is no size bias in the detection probability. The above result still applies to $\hat{N}_s$, the estimated number of clusters. However, the number of individuals is estimated as

$$\hat{N} = A \cdot \frac{n}{a \cdot \hat{P}_a} \cdot \bar{s}$$

Inference about N is limited to the time when the survey is done, hence to the actual individuals then present. If all individuals were counted ($P_a = 1$), $\text{var}(\hat{N})$ should be zero; hence a fpc should be applied to $\bar{s}$ and conceptually, it should be $1 - (n \cdot \bar{s})/\hat{N} = 1 - (a \cdot \hat{P}_a)/A$. Thus for this case we have

$$[\text{cv}(\hat{N})]^2 = \left[[\text{cv}(n)]^2 + [\text{cv}(\bar{s})]^2\right] \cdot \left[1 - \frac{a \cdot \hat{P}_a}{A}\right] + [\text{cv}(\hat{P}_a)]^2$$

Considerations are different for inference about $E(s)$. Usually one wants the inference to apply to the population in the (recent) past, present and (near) future, and possibly to populations in other areas as well. If this is the case, $\text{var}(\bar{s})$ should not be corrected using the fpc.

Consider now the case of clusters with size-biased detection. The fpc applied to the number of clusters is as above. For inference about $\hat{N}$, the fpc applied to the variance of $\hat{E}(s)$ is still $1 - (n \cdot \bar{s})/\hat{N}$, which is now equal to

$$1 - \frac{\bar{s}}{\hat{E}(s)} \cdot \frac{a \cdot \hat{P}_a}{A}$$

Thus the adjusted coefficient of variation of $\hat{N}$ is given by

$$[\text{cv}(\hat{N})]^2 = [\text{cv}(n)]^2 \cdot \left[1 - \frac{a \cdot \hat{P}_a}{A}\right] + [\text{cv}(\hat{P}_a)]^2 + [\text{cv}\{\hat{E}(s)\}]^2 \cdot \left[1 - \frac{\bar{s}}{\hat{E}(s)} \cdot \frac{a \cdot \hat{P}_a}{A}\right]$$

in the case of size-biased detection of clusters.

We reiterate that these finite population corrections will rarely, if ever, be worth making.

3.8 Stratification and covariates

Two methods of handling heterogeneity in data, and of improving precision and reducing bias of estimates, are stratification and inclusion of covariates in the analysis. Stratification might be carried out by geographic region, environmental conditions, cluster size, time, animal behaviour, detection cue, observer, or many other factors. Different stratifications may be selected for different components of the estimation equation. For example, reliable estimation of $f(0)$ or $h(0)$ (or equivalently, effective strip width or effective area), and of g_0 where relevant, requires that sample size is quite large. Fortunately, it is often reasonable to assume that these parameters are constant across geographic strata. By contrast, encounter rate or cluster size may vary appreciably across strata, but can be estimated with low bias from small samples. In this case, reliable estimates can be obtained for each geographic stratum by estimating $f(0)$ or $h(0)$ from data pooled across strata and other parameters individually by stratum, although it may prove necessary to stratify by, say, cluster size or environmental conditions when estimating $f(0)$ or $h(0)$. In general, different stratifications may be needed for each component of Equation 3.2.

Post-stratification refers to stratification of the data after the data have been collected and examined. This practice is generally acceptable, but care must be taken. For example, if geographic strata are defined to separate areas for which encounter rate was high from those for which it was low, and estimates are given separately for these strata, there will be a tendency to overestimate density in the high encounter rate stratum, and underestimate density in the low encounter rate stratum. Variance will be underestimated in both strata. If prior to the survey, there is knowledge of relative density, geographic strata should be defined when the survey is designed, so that density is relatively homogeneous within each stratum. Survey effort should then be greater in strata for which density is higher (Section 7.2.3).

Variables such as environmental conditions, time of day, date or cluster size might enter the analysis as covariates rather than stratification factors. If the number of potential covariates is large, they might be reduced in some way, for example through stepwise regression or principal components regression. To carry out a covariate analysis, an appropriate model must be defined. For example, the scale parameter of a model for the detection function might be replaced by a linear function of parameters:

$$\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \cdots$$

where X_{1i} might be sea state (Beaufort) at the time of detection i, X_{2i} might be cluster size for the detection, and so on.

3.8.1 Stratification

The simplest form of a stratified analysis is to estimate abundance independently within each stratum. A more parsimonious approach is to assume that at least one parameter is common across strata, or a subset of strata, an assumption that can be tested. Consider a point transect survey for which points were located in V geographic strata of areas A_v, $v = 1, \ldots, V$. Suppose we assume there is no size bias in detected clusters, and abundance estimates are required by stratum. Suppose further that data are sparse, so that $h(0)$ is estimated by pooling detection distances across strata. From Equation 3.2 with $c = 1$, $g_0 = 1$ and $a \cdot P_a = 2\pi k/h(0)$, we obtain

$$\hat{D}_v = \frac{n_v \cdot \hat{h}(0) \cdot \bar{s}_v}{2\pi k_v} = \frac{\hat{h}(0) \cdot \hat{M}_v}{2\pi} \qquad \text{where } \hat{M}_v = \frac{n_v \bar{s}_v}{k_v}$$

for stratum v. Mean density $\hat{D}$ is then the average of the individual estimates, weighted by the respective stratum areas A_v:

$$\hat{D} = \frac{\sum_v A_v \hat{D}_v}{A} \qquad \text{with } A = \sum_v A_v$$

The variance of any $\hat{D}_v$ may be found from Equation 3.3. However, to estimate $\text{var}(\hat{D})$, care must be taken, since one component of the estimation equation is common to all strata in a given year. The correct equation is:

$$\widehat{\text{var}}(\hat{D}) = \hat{D}^2 \cdot \left\{ \frac{\widehat{\text{var}}(\hat{M})}{\hat{M}^2} + \frac{\widehat{\text{var}}[\hat{h}(0)]}{[\hat{h}(0)]^2} \right\}$$

where $$\hat{M} = \frac{\sum_v A_v \hat{M}_v}{A}$$

and $$\widehat{\text{var}}(\hat{M}) = \frac{\sum_v A_v^2 \cdot \widehat{\text{var}}(\hat{M}_v)}{A^2}$$

with $$\widehat{\text{var}}(\hat{M}_v) = \hat{M}_v^2 \cdot \left\{ \frac{\widehat{\text{var}}(n_v)}{n_v^2} + \frac{\widehat{\text{var}}(\bar{s}_v)}{\bar{s}_v^2} \right\}$$

Thus the estimation equation has been separated into two components, one of which ($\hat{M}_v$) is estimated independently in each stratum, and the other of which is common across strata. Population abundance is estimated by $\hat{N} = A \cdot \hat{D} = \sum A_v \hat{D}_v$, with $\widehat{\text{var}}(\hat{N}) = A^2 \cdot \widehat{\text{var}}(\hat{D})$. Further layers of strata might be superimposed on a design of this type. In the above example, each stratum might be covered by more than one observer, or several forests might be surveyed, and a set of geographic strata defined in each. Provided the principle of including each independent component of the estimation equation just once in the variance expression is adhered to, the above approach is easily generalized.

The areas A_v are weights in the above expressions. For many purposes, it may be appropriate to weight by effort rather than area. For example, suppose two observers independently survey the same area in a line transect study. Then density within the study area may be estimated separately from the data of each observer (perhaps with at least one parameter assumed to be common between the observers), and averaged by weighting the respective estimates by length of transect covered by the respective observers. Note that in this case, an average of the two abundance estimates from each stratum is required, rather than a total. If stratification is by factors such as geographic region, cluster size, animal behaviour or detection cue, then the strata correspond to mutually exclusive components of the population and the estimates should be summed, whereas if stratification is by factors such as environmental conditions, observer, time or date (assuming no migration), then each stratum provides an estimate of the whole population, so that an average is appropriate.

Note that the stratification factors for each component of estimation may be completely different provided the components are combined with care. As a general guide, stratification prior to estimation of $f(0)$ or $h(0)$ should only be carried out if there is evidence that the parameter varies between strata, and some assessment should be made of whether the number of strata can be reduced. This policy is recommended since estimation of the parameter is unreliable if sample size is not large. Encounter rate and mean cluster size on the other hand may be reliably estimated from small samples, so if there is doubt, stratification should be carried out. Further, if abundance estimates are required by stratum, then both encounter rate and mean cluster size should normally be estimated by stratum. If all parameters can be assumed common across strata, such as observers of equal ability covering a single study area at

roughly the same time, stratification is of no benefit. Also, in the special case that strata correspond to different geographic regions, effort per unit area is the same in each region, the parameter $f(0)$ or $h(0)$ can be assumed constant across regions, and estimates are not required by region, stratification is unnecessary. Proration of a total abundance estimate by the area of each region is seldom satisfactory. An example for which stratification was used in a relatively complex way to improve abundance estimation of North Atlantic fin whales is given in Section 8.5.

Further parsimony may be introduced by noting that $\text{var}(n_v)$ is a parameter to be estimated, and $b_v = \text{var}(n_v)/n_v$ is often quite stable over strata. Especially if the n_v are small, it is useful to assess the assumption that $b_v = b$ for all v. If it is reasonable, the number of parameters is reduced. The parameter b can be estimated by

$$\hat{b} = \frac{\sum_v \widehat{\text{var}}(n_v)}{n}$$

and $\text{var}(n_v)$ is then more efficiently estimated as $\widehat{\text{var}}_p(n_v) = \hat{b}/n_v$. This approach is described further in Section 6.3, and illustrated in Section 8.4. The same method might also be applied to improve the efficiency of $\widehat{\text{var}}(\bar{s}_v)$.

3.8.2 Covariates

Several possibilities exist for incorporating covariates. Ramsey *et al.* (1987) used the effective area, v, as a scale parameter in point transect surveys, and related it to covariates using a log link function:

$$\log_e(\text{v}) = \beta_0 + \sum_j \beta_j \cdot X_j$$

where X_j is the jth covariate. Computer programs for implementing this approach for the case of an exponential power series detection function are available from the authors.

Drummer and McDonald (1987) considered a single covariate X, taken to be cluster size in their example, and incorporated it into detection functions by replacing y by y/X^γ, where γ is a parameter to be estimated. Thus the univariate half-normal detection function

$$g(y) = \exp\left[-\frac{y^2}{2\sigma^2}\right]$$

becomes the 'bivariate' detection function:

$$g(y \mid X) = \exp\left[-\frac{y^2}{2(\sigma X^{\gamma})^2}\right]$$

The interpretation is now that $g(y \mid X)$ is the detection probability of a cluster at distance y, given that its size is X. Drummer and McDonald proposed the following detection functions as candidates for this approach: negative exponential, half-normal, generalized exponential, exponential power series and reversed logistic. They implemented the method for the first three, although their procedure failed to converge to plausible parameter values for the generalized exponential model for the data set they present. Their software (SIZETRAN) is available (Drummer 1991).