2015

Efficiency and robustness in distance sampling

Robert Graham Clark
University of Wollongong

Efficiency and robustness in distance sampling

Abstract

Distance sampling is a technique for estimating the abundance of animals or other objects in a region, allowing for imperfect detection. The impact of uncertainty about the detection parameters on the precision of the estimated abundance is important, both for deciding on the appropriate sample size and determining whether to use distance sampling or an alternative method. This paper derives the asymptotic penalty due to this uncertainty, and tabulates it for a variety of models. The penalty is typically between 2 and 4 but can be much higher, particularly for steeply declining detection rates where distance sampling is typically most strongly recommended. The asymptotic results are confirmed in a simulation study which also examines model-averaging, mis-specified detection function and simple strip expansion. The paper shows that distance sampling needs larger sample sizes than is commonly supposed, and so should not be regarded as the only acceptable approach to abundance estimation.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

15-15

Efficiency and Robustness in Distance Sampling

Robert Graham Clark

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Efficiency and Robustness in Distance Sampling

Robert Graham Clark

National Institute for Applied Statistics Research Australia,

University of Wollongong, Australia.

Abstract

Distance sampling is a technique for estimating the abundance of animals or other objects in a region, allowing for imperfect detection. The impact of uncertainty about the detection parameters on the precision of the estimated abundance is important, both for deciding on the appropriate sample size and determining whether to use distance sampling or an alternative method. This paper derives the asymptotic penalty due to this uncertainty, and tabulates it for a variety of models. The penalty is typically between 2 and 4 but can be much higher, particularly for steeply declining detection rates where distance sampling is typically most strongly recommended. The asymptotic results are confirmed in a simulation study which also examines model-averaging, mis-specified detection function and simple strip expansion. The paper shows that distance sampling needs larger sample sizes than is commonly supposed, and so should not be regarded as the only acceptable approach to abundance estimation.

Keywords: abundance estimation; detection function; distance sampling; maximum likelihood estimation; statistical ecology; variance penalty

1 Introduction

The number of animals in a region is often of ecological importance. Ecologists can reliably count animals that are close to them, but tend to miss those that are distant. Unfortunately, only a few animals will be very close to an observer's path in typical abundance studies. The total abundance in a region can be estimated by simply scaling up the number of animals in a narrow strip around the observers' paths, but this might be excessively variable. Practitioners of distance sampling avoid this hazard by including animals near and far, and then estimating how many animals they are likely to have missed. This estimation is accomplished by modelling how the probability of detection declines with distance from the observer.

This paper considers distance sampling in its usual conjunction with line transect sampling. Parallel transect lines are laid down across a region. Observers move along the transects, and record observations of animals (or plants, or groups of animals, or other objects) and their perpendicular distances from the transect. Empirically, more objects are detected near to transect lines than far from them in many studies, suggesting that detectability is a decreasing function of distance. The distance sampling methodology exploits this phenomenon, by modelling the detection rate as a function of distance. The number of detected objects can then be scaled to estimate the abundance N allowing for imperfect detection. Provided the assumptions of the method are met, the maximum range can be made fairly large, thereby increasing the sample size, while avoiding or reducing bias due to declining detection rate. Distance sampling is widely used in ecology: a Web of Science search found 276 articles on distance sampling in ecology journals in 2014 alone. The wide range of applications include wild horses in the Australian Alps (Walter and Hone, 2003), large herbivores in South Africa's Kruger National Park (Kruger et al., 2008) and odonata (dragonflies) in a rainforest locality in Papua New Guinea (Oppel, 2006). For a detailed description of the approach, see Buckland et al. (2004).

The major assumptions of the method are

i. Detection is perfect at zero distance.

ii. The detection function is of known form, with some unknown parameters requiring estimation. Alternatively, model-averaging may be used provided the detection function is assumed to be one of a known set of alternatives.

iii. Animals' distances to the nearest transect line are (at least approximately) uniformly distributed.

It is also assumed that there is no measurement error (for example, false positive detections), that there is no movement of objects which could lead to multiple chances of detection, and that detection events are independent. This paper will consider the classical distance sampling (CDS) scenario where there is only one observer. The same methodology can be used with multiple observers by pooling their detections. Mark recapture and mark recapture distance sampling (MRDS) are methods which more fully use data from multiple observers by matching their detections; MRDS, in particular, can be used to relax assumption (i). More recent approaches combine a spatial point process model with the detection model (see for example Johnson et al. 2010 and Oedekoven et al. 2014). Spatial models allow abundances to be estimated for subregions, and can exploit spatial trends in estimation, however inference may be sensitive to the assumed spatial model which must therefore be carefully constructed. This paper focuses on CDS, as most applications of line transect sampling remain single observer, and spatial models are not always feasible.

Robustness to the 3 assumptions above is explored in the literature. Robustness to (i) is improved by the use of MRDS. Assumption (ii) is dealt with by the use of flexible families of detection functions with two or more parameters, and the use of model averaging. Buckland et al. (2001) argue that (iii) is approximately satisfied provided transect lines are placed randomly or systematically. However, Barry and Welsh (2001) question the uniformity assumption and find that classical distance sampling estimators are biased in a design-based framework, that is, under repeated random placement of transect lines. The matter remains in contention (Melville and Welsh, 2001; Fewster et al., 2005; Melville et al., 2008). Melville and Welsh (2001) suggest an alternative approach where the detection function is estimated from a separate calibration study.

A natural alternative to distance sampling is the simple scaling up of observations in a strip about the transect. When strips are too wide, simple expansion is severely negatively biased due to non-detections of the most distant animals in the strip. When strips are sufficiently narrow, this bias becomes negligible, but the variance of the estimated abundance becomes large. Distance sampling aims to achieve the low variance of wide strips while avoiding the bias, by adjusting for non-detection. But there is a hidden cost - the effect of unknown detection parameters on the precision of the estimated abundance -

which reverses at least some of the benefit due to wider strips. This paper illuminates this cost both asymptotically and in small samples.

The most commonly used estimator of the abundance (N) is an empirical method of moments (MOM) estimator using a conditional maximum likelihood estimator of the detection parameters ($\boldsymbol{\theta}$). A simple expression for the maximum likelihood estimator (MLE) of N has also been derived assuming that detection function parameters are known (Borchers and Buckland, 2002, pp.17-19,138). Fewster and Jupp (2009) prove a Central Limit Theorem for the MLE of N in a more general setting which subsumes CDS, and show that the MOM and MLE estimators are asymptotically close.

Section 2 reviews CDS in more detail, and describes the example which motivated this research: the estimation of wild horse numbers in south-eastern Australia. Section 3 derives the MLE of N making use of the Stirling approximation for factorials. For the first time, it is shown that the resulting MLE is identical to the usual empirical method of moments estimator of N, and derive the limiting variance. This asymptotic variance is identical to that of Fewster and Jupp (2009), but the use of the Stirling approximation allows a simpler derivation. The limiting variance of $\hat{N}$ is expressed as the variance when the detection function is known multiplied by a penalty for unknown $\boldsymbol{\theta}$. This penalty is tabulated for various detection models. It can be substantial, and for many situations arising in practice is between 2 and 4, depending on how rapidly the detection function declines with distance. Section 4 is a simulation study to evaluate the small sample performance of various estimators including the MLE when the detection function is correctly and incorrectly specified, a model averaged estimator, and simple expansion estimators. Section 5 is a discussion.

2 Review of Classical Distance Sampling

2.1 Methodology for Classical Distance Sampling

The aim is to estimate the number of objects in a two-dimensional region which may be irregular. Transect lines are generally parallel and are oriented to roughly maximise within-transect variability and minimise between-transect variability. To simplify discussion, transects will be assumed to run in an east-west direction. The region has a maximal north-south extent of Y. Points are located by Cartesian coordinates (x, y), with $0 \leq y \leq Y$ the northerly coordinate and and x the easterly coordinate. Let l_y be the horizontal length of the region at vertical position y.

Observation may either be one-sided (only objects to the north, or only objects to the south, are observed) or two-sided. Only objects with perpendicular distance up to a pre-chosen limit w from a transect line have a chance of being observed. Two-sided observation is the more common case, but one-sided observation is sometimes necessary, for example if the observer can only see out one side of a vehicle.

Let A be the area of the region, N be the number of objects, $\bar{N} = N/A$ be the density of objects and n be the number of detections. It is assumed that the probability of observing an object at perpendicular distance u from a transect line is $g(u, \boldsymbol{\theta})$ when $0 \leq u \leq w$, where $\boldsymbol{\theta}$ is a vector of p parameters specifying the function within a family. It is assumed that $g(0, \boldsymbol{\theta}) = 1$ and that g is a non-increasing function of distance. The Distance software (Thomas et al., 2010), which implements both CDS and MRDS, allows four possible functional forms for $g()$, including the half-normal model, $g(u) = \exp\left(-\frac{u^2}{\theta^2}\right)$, and the hazard rate function, $g(u) = 1 - \exp\left(-(u/\theta_1)^{-\theta_2}\right)$. Both of these functions satisfy the *shoulder condition* that $g'(0) = 0$, with the hazard rate model giving greater flexibility in modelling the *shoulder width* (i.e. the range near 0 over which g is relatively flat). The Distance software also includes the uniform and negative exponential detection functions, and allows for more general functions defined by polynomial or trigonometric expansions based on the four basic functions. Fewster et al. (2005), citing Buckland et al. (1993), state that the minimum number of detections needed to reliably estimate $g()$ is 60-80.

A challenge in applying CDS is in identifying an appropriate detection function, $g(d; \boldsymbol{\theta})$ where d is distance and $\boldsymbol{\theta}$ are unknown parameters controlling the shape of the detection curve. In practice, the data often do not definitively identify the appropriate distance

function and the estimate of N may be sensitive to the assumed function (Buckland et al., 2001, p44). The use of "robust models", which have enough flexibility to model a range of typical shapes, is recommended by Buckland et al. (2001, p.46-49), with the hazard rate model given as a particularly useful example.

It is also possible to include other covariates affecting the detectability of objects in the distance function, such as characteristics of the animal or plant.

Let d_i, $i = 1, \ldots, N$, be the perpendicular distances from the objects to the nearest transect line, and let $\delta_i = 1$ for observed objects and $\delta_i = 0$ for the rest. The full observed data usually consists of $\boldsymbol{y_s} = (y_1, \ldots, y_n)$, the values of the vertical coordinates y_i where $\delta_i = 1$. Given the transect line positions, this implies knowledge of the perpendicular distances $\boldsymbol{d_s} = (d_1, \ldots, d_n)$. In CDS, $\boldsymbol{d_s}$ are treated as the complete data and $\boldsymbol{y_s}$ is ignored, whereas $\boldsymbol{y_s}$ are integral if spatial models are used. Let $\boldsymbol{d_{Uc}}$ refer to the distances of the N_c objects where $d_i \leq w$, where "c" stands for "covered". Buckland et al. (2004) assume that $\boldsymbol{d_{Uc}}$ are independent and identically distributed uniform $U(0, w)$. For example, this is the case if object locations $\boldsymbol{Y_U}$ are independent and uniformly distributed on $[0, Y]$. The assumption of uniformly distributed $\boldsymbol{Y_U}$ is often implausible (e.g. Courtois et al. 2013). However if lines are placed completely at random or are evenly spaced, and a buffer zone is used to avoid biases at the edge of the region, then distances will tend to be uniformly distributed when averaged over possible transect placements (Fewster et al., 2008, Figure 1).

Assuming $\boldsymbol{d_{Uc}} \overset{i.i.d.}{\sim} U(0, w)$, the distribution of d_i given $\delta_i = 1$ is easily derived as

$$g_{cond}\left(d_i; \boldsymbol{\theta}\right) = g\left(d_i; \boldsymbol{\theta}\right) / \int_0^w g\left(u; \boldsymbol{\theta}\right) du \tag{1}$$

Let $\bar{g}\left(\boldsymbol{\theta}\right) = w^{-1} \int_0^w g\left(u; \boldsymbol{\theta}\right) du$ be the unconditional probability of detection. The likelihood of $\boldsymbol{d_s}$ given n is

$$L_d = \prod_{i=1}^{n} g_{cond}\left(d_i; \boldsymbol{\theta}\right) = \left(\prod_{i=1}^{n} g\left(d_i; \boldsymbol{\theta}\right)\right) \left(\int_0^w g\left(u; \boldsymbol{\theta}\right) du\right)^{-n} = \left(\prod_{i=1}^{n} g\left(d_i; \boldsymbol{\theta}\right)\right) w^{-n} \bar{g}\left(\boldsymbol{\theta}\right)^{-n} \tag{2}$$

and the corresponding conditional log-likelihood is

$$l_d = \sum_{i=1}^{n} \log g\left(d_i; \boldsymbol{\theta}\right) - n \log\left(\bar{g}\left(\boldsymbol{\theta}\right)\right) - n \log w. \tag{3}$$

The parameters $\boldsymbol{\theta}$ can be obtained by setting the derivative of l_d with respect to $\boldsymbol{\theta}$ to 0. Prior to this, it is convenient to define

$$\boldsymbol{h}(u;\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} g(u;\boldsymbol{\theta})$$

and $\bar{\boldsymbol{h}}(\boldsymbol{\theta}) = w^{-1}\int_0^w \boldsymbol{h}(u;\boldsymbol{\theta})\,du$. Notice that $\boldsymbol{h}(u;\boldsymbol{\theta})$ is a p-vector where p is the number of parameters in $\boldsymbol{\theta}$. The partial derivative of $\bar{g}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is $\bar{\boldsymbol{h}}(\boldsymbol{\theta})$, subject to regularity conditions allowing the derivative operator to be taken within the integral. Setting the derivative of l_d to 0 gives the following estimating equation for $\boldsymbol{\theta}$:

$$0 = \sum_{i=1}^{n} g(d_i;\boldsymbol{\theta})^{-1}\boldsymbol{h}(d_i;\boldsymbol{\theta}) - n\bar{g}(\boldsymbol{\theta})^{-1}\bar{\boldsymbol{h}}(\boldsymbol{\theta}) \tag{4}$$

The most commonly used estimator of N_c is motivated by the fact that the expected number of animals observed under the model is $E[n] = N_c\bar{g}(\theta)$. This suggests an empirical method of moments estimator of N_c:

$$\hat{N}_{c(MOM)} = \frac{n}{\bar{g}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}\right)} \tag{5}$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}$ is the solution to (4) (CML stands for conditional maximum likelihood). This is the approach proposed by Buckland et al. (2004) and is widely used in practice.

The usual estimator of N is based on the fact that probability sampling is used to place transects, and hence in expectation over repeated random transect placement, $E_p[N_c] = NP$ where E_p refers to design expectation and P is the proportion of the region which is covered, i.e. which falls within distance w of a transect. The value of P is assumed to be known. Hence N is estimated by

$$\hat{N}_{MOM} = \frac{\hat{N}_{c(MOM)}}{P} = \frac{n}{P\bar{g}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}\right)}. \tag{6}$$

Ignoring edge effects, P may be approximated by $P \approx Sl_{tot}w/A$ where $l_{tot} = \sum_{i=1}^{m} l_{y_i}$ is the total length of all transects, and $S = 1$ for one-sided and 2 for two-sided observation.

The estimator in (6) is then equal to

$$\hat{N}_{MOM} = \frac{nA}{S\bar{g}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}\right) wl_{tot}}. \tag{7}$$

and the population density $\bar{N} = N/A$ is estimated by

$$\hat{\bar{N}}_{MOM} = \frac{\hat{N}_{MOM}}{A} = \frac{n}{S\bar{g}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}\right) wl_{tot}}. \tag{8}$$

This is estimator (2.35) on page 17 of Buckland et al. (2004), where P is calculated using (2.5).

2.2 Maximum Likelihood Estimation of N and N_c

Buckland et al. (2004) also discuss maximum likelihood estimation of N_c. The density of $\boldsymbol{d_s}$ given n is L_d in (2). Under the assumed model, δ_i are independent Bernoulli random variables with expected value $\bar{g}(\boldsymbol{\theta})$. Hence

$$n \sim \text{bin}\left(N_c, \bar{g}\left(\boldsymbol{\theta}\right)\right). \tag{9}$$

The likelihood is the product of the probability function of n and L_d:

$$L_c = \binom{N_c}{n} \prod_{i=1}^{n} \left\{ g\left(d_i; \boldsymbol{\theta}\right) / \int_0^w g\left(u; \boldsymbol{\theta}\right) du \right\} \tag{10}$$

Buckland et al. (2004) suggest that this could be maximised with respect to N_c and $\boldsymbol{\theta}$ to obtain a maximum likelihood estimator of N_c, but in practice the literature of applications of CDS (to the author's knowledge) uses the moments-based estimator (5).

In design expectation (i.e. under repeated random placement of transects), $E_p\left(N_c\right) = NP$. Based on this, Buckland et al. (2004) suggest that even if N_c is estimated by maximising (10), N should be estimated using $\hat{N} = \hat{N}_c/P$. In this approach, the estimation of N_c relies on the detection model, but the scaling up from N_c to N relies only on random placement of transects.

Another approach is to directly estimate N by maximum likelihood, without resort to a mixture of model-based and design-based reasoning. MLEs are asymptotically optimal

when the model is correctly specified, subject to regularity conditions, and this provides some guarantee that the best (or nearly best) estimator is being used. Borchers and Buckland (2002, section 7.2) derive the likelihood for N and $\boldsymbol{\theta}$. Under the same model assumptions, $n \sim \text{bin}\left(N, P\bar{g}\left(\boldsymbol{\theta}\right)\right)$ and the full likelihood is

$$\begin{aligned} L &= \binom{N}{n}\left(1-P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{N-n}\left(P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{n}\prod_{i=1}^{n}\left\{g\left(d_i;\boldsymbol{\theta}\right)/\int_0^w g\left(u;\boldsymbol{\theta}\right)du\right\} \\ &= \binom{N}{n}\left(1-P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{N-n}P^n w^{-n}\prod_{i=1}^{n}g\left(d_i;\boldsymbol{\theta}\right) \end{aligned} \tag{11}$$

Note that (10) is the special case of (11) where the region of interest is defined to be the covered area, so that $P = 1$.

When $\boldsymbol{\theta}$ is known, the MLE of N is the integer part of $n/(P\bar{g}\left(\boldsymbol{\theta}\right))$ (Borchers and Buckland, 2002, pp.17-19,138). The same reference also shows that if the factorial terms in (11) are approximated using Stirling's rule and N is treated as a continuous parameter, the MLE is then $\hat{N}_{known\theta} = n/(P\bar{g}\left(\boldsymbol{\theta}\right))$. It is straightforward to derive the variance of $\hat{N}_{known\theta}$ using the fact that $n \sim \text{bin}\left(N, P\bar{g}\left(\boldsymbol{\theta}\right)\right)$:

$$\text{var}\left(\hat{N}_{known\theta}\right) = P^{-2}\bar{g}\left(\boldsymbol{\theta}\right)^{-2}\text{var}(n) = NP^{-1}\bar{g}\left(\boldsymbol{\theta}\right)^{-1}\left(1-P\bar{g}\left(\boldsymbol{\theta}\right)\right) \tag{12}$$

In fact, it directly follows that when $\boldsymbol{\theta}$ is unknown, the MLE of N using Stirling's Rule is $n/P\bar{g}\left(\hat{\boldsymbol{\theta}}\right)$ where $\hat{\boldsymbol{\theta}}$ maximises L, although this corollary is not stated by Borchers and Buckland (2002) or elsewhere. Section 3 of this paper goes further, and shows that the full likelihood in (11) is maximised by $\hat{N}_{MOM}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}$ as defined in (4) and (6).

Fewster and Jupp (2009) derived a Central Limit Theorem for $\hat{N}_{MLE}$ and showed that $\hat{N}_{MLE} - \hat{N}_{MOM}$ is of asymptotic order $O_p(1)$ which is remarkably close.

It is worth noting here that the conditional likelihood (2) assumes that distances are independently distributed. In reality there may sometimes be spatial trends and local clumpiness in both animal locations and detectability. To partially allow for this, the variances of $\hat{N}_{c(MOM)}$ and $\hat{N}_{MOM}$ may be estimated using drop-a-group jackknife, or an ultimate cluster variance estimator, where groups or clusters are transects. For details, see Buckland et al. (2004).

2.3 Motivating Example: Brumbies in South Eastern Australia

Populations of wild horses (also called brumbies) are present in Kosciusko National Park and Victoria in South Eastern Australia. They are a non-native species with a possible impact on the ecology of these areas, and are also culturally important (Walter and Hone, 2003). It is therefore important to periodically monitor their abundance. Brumby population sizes have been estimated by aerial observation from a helicopter (e.g. Walter 2002, 2003; Walter and Hone 2003; Dawson 2009). The basic approach is for the helicopter to fly parallel east-west transect lines, evenly spaced at 2km intervals. Two observers in the helicopter record numbers of groups, and numbers of horses in each group, up to 200m on the left side. The perpendicular distance of each observed group is also measured. This is done using a bar protruding from the left of the plane with markings indicating distance. The bar is calibrated under the assumption that the helicopter flies at a constant height (usually 100m), so that distance can be measured. Distances are recorded in bands of 0-50m, 50-100m, 100-150m, and 150-200m. CDS is then used to calculate estimates of the abundance of groups.

The author of this paper was retained by the NSW Office of Environment and Heritage, who contribute funding to aerial monitoring of brumby numbers, to review past reports and selected literature and make recommendations regarding future sampling methods. All views in this paper are the authors' own and not those of this Office. This section relates to part of the review, while the remainder of the paper is subsequent research not associated with the Office.

We focus here on the study described in Walter and Hone (2003), as re-analysed by Laake et al. (2008), who compare distance sampling, mark-recapture and mark-recapture distance sampling estimates of abundance. Table 1 includes results from Table 6 and the text of page 1145 of Walter and Hone (2003) and Table 1 and the text of page 304 of Laake et al. (2008). The strip transect expansion estimates are just the number of observed groups (by either observer) divided by the strip length multiplied by either 50m or 200m, with no allowance for undercount. The values of the Akaike Information Criterion (AIC) are also shown for each detection model. Model averaged estimates are calculated using weights calculated from the AICs as described in Burnham and Anderson (2002). Mark recapture and mark recapture distance sampling estimates were also considered but are not shown in Table 1 as they are outside the scope of this paper.

The table shows that the CDS estimates vary quite a bit depending on the detection function assumed, with the negative exponential and hazard rate functions giving estimates around 30% higher than the other two functions. At the same time, the data provide little guidance as to which of these models applies, as the AICs all lie within 2 of each other. This uncertainty contributes to the high coefficient of variation (CV $\approx$ 30%) of the model-averaged CDS estimator. Moreover, the discrepancies between the various CDS estimators mean that the model-averaged estimator is sensitive to which models are included.

Table 1: Estimates of Density ($\bar{N} = N/A$) of Horse Groups

Method	Detection Function	AIC	$\widehat{Density}$ $\hat{\bar{N}} = \hat{N}/A$ (groups/km^2)	$CV\%\left(\hat{\bar{N}}\right)$
CDS	negative exponential	0.00	0.36	24.8
CDS	uniform (cosine)	0.96	0.28	18.8
CDS	half-normal	1.35	0.28	19.9
CDS	hazard rate	1.87	0.36	57.2
CDS	model-averaged	n/a	0.33	30.3
Strip Transect 0-50m	n/a	n/a	0.31	22.1
Strip Transect 0-200m	n/a	n/a	0.18	16.1

The CVs are high and are quite different depending on which detection function is used, with the two-parameter hazard rate model leading to much higher CV than the other detection models which are one-parameter. The model-averaging CV is in between. This suggests that the variance of the abundance estimator is higher due to the use of two rather than one parameters.

The strip transect expansion estimator using detections up 200m has much lower CV, but is much lower than the CDS estimator. This reflects the large negative bias that would be expected for this estimator, because it does not allow for declining detection with distance. The expansion estimator based on detections up to 50m is more plausible. It is close to the model-averaged CDS estimator, suggesting that the bias is much less when only the nearest 50m of detections are used. This is not surprising, because the detection rate would decline relatively little over this shorter range. What is surprising is that the 0-50m option gives similar CVs to the 0-200m model-averaged CDS method, even though it only uses one quarter of the distance range and 44% of the detected groups. This motivated the research in the current paper on the efficiency of CDS estimators.

3 Maximum Likelihood for Classical Distance Sampling

3.1 The Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator of N under Stirling's approximation for factorials will be derived in this section, where the likelihood is given by (11). As discussed in section 2.2, the estimation of N_c is the special case where $P = 1$. Stirling's rule $\log(x!) \approx x\log(x) - x$ implies that

$$\begin{aligned}
\log\binom{N}{n} &= \log\{N!/n!\,(N-n)!\} \\
&\approx N\log(N) - N - (n\log(n) - n) - \{(N-n)\log(N-n) - (N-n)\} \\
&= N\log(N) - (N-n)\log(N-n) - n\log(n). \qquad (13)
\end{aligned}$$

There is a positive probability that n is equal to 0 or N, in which case $\log(n)$ or $\log(N-n)$ are not defined. The limit of the right hand side of (13) can easily be shown to equal 0 in these cases, so the following refinement of (13) is used:

$$\log\binom{N}{n} \approx \left\{ \begin{array}{ccc} N\log(N) - (N-n)\log(N-n) - n\log(n) & \text{if} & 0 < n < N \\ 0 & \text{if} & n = 0 \text{ or } n = N \end{array} \right\}. \qquad (14)$$

L will be maximised with respect to N and $\boldsymbol{\theta}$ treating N as a continuous parameter. Stirling's approximation for $\log(x!)$ is very accurate even for small x as long as x is at least 2 or 3, and both $N - n$ and n would be well above this in practice. Theorem 1 states the MLEs and the approximate Fisher information. The proof is in in Appendix 1.

Theorem 1 *The model defined by (1) and (9) is assumed, and it is assumed that the likelihood (11) can be approximated using (14). Let Ω be an open set defining the set of feasible values of $\boldsymbol{\theta}$. If there is a unique $\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}$ satisfying (4) then it is the maximum likelihood estimator, and the MLE of of N on $(0,\infty)$ is $\hat{N}_{MOM}$ in (6).*

Let D be a random variable with density $g(d)/\int_0^w g(u)du$ for $0 \leq d \leq w$. The Fisher Information of $\left(N, \boldsymbol{\theta}^T\right)^T$ is approximately equal to

$$\mathcal{I}\left(N, \boldsymbol{\theta}\right) \approx \left[\begin{array}{cc} \mathcal{I}_{NN} & \mathcal{I}_{N\theta}^T \\ \mathcal{I}_{N\theta} & \mathcal{I}_{\theta\theta} \end{array} \right] \tag{15}$$

for large N, where

$$\begin{aligned} \mathcal{I}_{NN} &= N^{-1}\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{g}\left(\boldsymbol{\theta}\right) \\ \mathcal{I}_{\theta\theta} &= NPw^{-1}\int_0^w \boldsymbol{h}\left(u;\boldsymbol{\theta}\right)\boldsymbol{h}\left(u;\boldsymbol{\theta}\right)^T g\left(u;\boldsymbol{\theta}\right)^{-1} du + NP\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right)\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right)^T\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} \end{aligned} \tag{16}$$

$$= NP\bar{g}\left(\boldsymbol{\theta}\right)\mathrm{var}_D\left(\boldsymbol{h}\left(D;\boldsymbol{\theta}\right)/g\left(D;\boldsymbol{\theta}\right)\right) + NP\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right)\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right)^T\bar{g}\left(\boldsymbol{\theta}\right)^{-1}\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} \tag{17}$$

$$\mathcal{I}_{N\theta} = \left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \tag{18}$$

Surprisingly, the maximum likelihood estimators of $\boldsymbol{\theta}$ and N are identical to the usual CDS estimators $\hat{\boldsymbol{\theta}}_{\boldsymbol{CML}}$ and $\hat{N}_{MOM}$ defined by (4) and (6), even though (4) was based on a score equation conditional on n, and (6) was motivated by an empirical method of moments argument. This is reminiscent of the result in Borchers and Buckland (2002, exercise 6.4) that the MLE of the abundance in mark-recapture studies is equal to the Petersen estimator. The theorem also generalises the result of Buckland et al. (2004) for known $\boldsymbol{\theta}$.

Let V be the inverse of the Fisher Information matrix in (15). Subject to regularity conditions, maximum likelihood estimators are asymptotically normal with expectation equal to the true parameter values and variance-covariance matrix equal to the inverse of the Fisher Information matrix. Unfortunately these regularity conditions do not hold here, for example condition (M3) of the Central Limit Theorem in Lehmann (1999, pp.499-500) is not met, because n and $\boldsymbol{d_s}$ are dependent. Moreover, (15) is only the approximate

Fisher information based on a Taylor Series expansion, whereas the usual Central Limit Theorem requires the exact Fisher information. Clark (2015) uses an alternative method of proof to derive a Central Limit Theorem for the MLE. It turns out that V is indeed the limiting variance of $\left(\hat{N}, \hat{\boldsymbol{\theta}}^T\right)^T$. The proof in Clark (2015) is essentially a simplified version of the proof of the result in Fewster and Jupp (2009) which does not make the Stirling approximation. Theorem 1 is an advance on the result in Fewster and Jupp (2009), because it demonstrates the identity of $\hat{N}_{MLE}$ and $\hat{N}_{MOM}$, which greatly simplifies the calculation of the MLE in practice, and also provides a simpler derivation of V.

Let V be the inverse of the Fisher Information matrix in (15). Subject to regularity conditions, maximum likelihood estimators are asymptotically normal with expectation equal to the true parameter values and variance-covariance matrix equal to the inverse of the Fisher Information matrix. This would make V the asymptotic variance-covariance matrix of $\left(\hat{N}, \hat{\boldsymbol{\theta}}^T\right)^T$, and V_{11} the limiting variance of $\hat{N}$. Unfortunately these regularity conditions do not hold here, for example condition (M3) of the Central Limit Theorem in Lehmann (1999, pp.499-500) is not met, because n and $\boldsymbol{d_s}$ are dependent. Moreover, (15) is only the approximate Fisher information based on a Taylor Series expansion, whereas the usual Central Limit Theorem requires the exact Fisher information. Theorem 2 uses an alternative method of proof to derive a Central Limit Theorem for the MLE. It turns out that V is indeed the limiting variance of $\left(\hat{N}, \hat{\boldsymbol{\theta}}^T\right)^T$. The proof is similar to the proof of a Central Limit Theorem in Fewster and Jupp (2009) but is simpler due to the use of the Stirling approximation.

Before stating Theorem 2, it is convenient to express V in block form. We will henceforth mostly write $g(u)$,$\boldsymbol{h(u)}$, $\bar{g}$ and $\bar{\boldsymbol{h}}$ for readability, rather than $g(u;\boldsymbol{\theta})$ etc. Using a standard result on the inverse of a matrix in block form (e.g. 5.16a of Harville 1997), V is equal to

$$V = \begin{bmatrix} V_{11} & V_{21}^T \\ V_{21} & V_{22} \end{bmatrix} \tag{19}$$

where

$$\left.\begin{aligned} V_{22} &= \left\{\mathcal{I}_{\theta\theta} - \mathcal{I}_{N\theta}\mathcal{I}_{NN}^{-1}\mathcal{I}_{N\theta}^T\right\}^{-1} = N^{-1}P^{-1}\Delta^{-1}\bar{g}^{-1} \\ V_{11} &= \mathcal{I}_{\theta\theta}^{-1} + \mathcal{I}_{\theta\theta}^{-1}\mathcal{I}_{N\theta}^T V_{22}\mathcal{I}_{N\theta} \\ &= N\left(1 - P\bar{g}\right)P^{-1}\bar{g}^{-1} + NP^{-1}\bar{g}^{-3}\bar{\boldsymbol{h}}^T\Delta^{-1}\bar{\boldsymbol{h}} \\ V_{21} &= V_{21}^T = -V_{22}\mathcal{I}_{N\theta}\mathcal{I}_{NN}^{-1} \\ &= -\bar{g}^{-2}P^{-1}\Delta^{-1}\bar{\boldsymbol{h}} \end{aligned}\right\} \tag{20}$$

and Δ is the p by p matrix defined by

$$\Delta = \mathrm{var}_D \left[\boldsymbol{h}\left(D\right) / g\left(D\right) \right]. \tag{21}$$

Equations (19) and (20) can be rewritten as

$$V = \begin{bmatrix} N\tilde{V}_{11} & \tilde{V}_{21}^T \\ \tilde{V}_{21} & N^{-1}\tilde{V}_{22} \end{bmatrix} \tag{22}$$

where

$$\left.\begin{array}{rcl} \tilde{V}_{11} & = & \left(1 - P\bar{g}\right) P^{-1}\bar{g}^{-1} + P^{-1}\bar{g}^{-3}\bar{\boldsymbol{h}}^T\Delta^{-1}\bar{\boldsymbol{h}} \\ \tilde{V}_{22} & = & \bar{g}^{-1}P^{-1}\Delta^{-1} \\ \tilde{V}_{21} & = & -\bar{g}^{-2}P^{-1}\Delta^{-1}\bar{\boldsymbol{h}} \end{array}\right\} \tag{23}$$

Theorem 2 on the next page gives a central limit theorem for the maximum likelihood estimators of N and $\boldsymbol{\theta}$. The proof is in Appendix 2.

Theorem 2 *The following is assumed:*
(i) Let $N \to \infty$ with $\boldsymbol{\theta}$ held fixed.
(ii) The partial derivatives of $g_{cond}(d;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist, where $g_{cond}(d;\boldsymbol{\theta}) = g\left(d;\boldsymbol{\theta}\right) / \int_0^w g(u;\boldsymbol{\theta})du$ for $0 \le d \le w$ and $g_{cond}(d;\boldsymbol{\theta}) = 0$ otherwise.
(iii) The partial derivatives of $\int g_{cond}(d;\boldsymbol{\theta})dd$ with respect to $\boldsymbol{\theta}$ exist and can be obtained by differentiating under the integral sign.
(iv) $g\left(d;\boldsymbol{\theta}\right)$ are distinct for different values of $\boldsymbol{\theta}$, and $g(d;\boldsymbol{\theta}) > 0$ for $0 \le d \le w$.
(v) The set Ω of feasible values of $\boldsymbol{\theta}$ is open, and the true value of $\boldsymbol{\theta}$ belongs to Ω.
(vi) There is a unique $\boldsymbol{\theta} \in \Omega$ satisfying (4).
Then the limiting distribution of the MLEs $\hat{N}$ and $\hat{\boldsymbol{\theta}}$ is given by:

$$N^{1/2}\left\{ \begin{pmatrix} \hat{N}/N \\ \hat{\boldsymbol{\theta}} \end{pmatrix} - \begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix} \right\} \xrightarrow{d} N\left(0, \tilde{V}\right) \tag{24}$$

where the elements of $\tilde{V}$ are defined by (23).

Rescaling (24) appropriately, the asymptotic variance of $\left(\hat{N}_c, \hat{\boldsymbol{\theta}}^T\right)^T$ is V, the inverse approximate Fisher Information. The limiting variance of $\hat{N}$, V_{11} from (20), is of primary

interest. It can be expanded by elementary operations as

$$V_{11} = \text{var}\left(\hat{N}_{known\theta}\right) F \tag{25}$$

where $\text{var}\left(\hat{N}_{known\theta}\right)$ is the variance of $\hat{N}$ when $\boldsymbol{\theta}$ is known, as defined by (12), and

$$F = 1 + \bar{\boldsymbol{h}}^{\boldsymbol{T}} \Delta^{-1} \bar{\boldsymbol{h}} \bar{g}^{-2} \left(1 - P\bar{g}\right)^{-1}$$

is a penalty term attributable to $\boldsymbol{\theta}$ requiring estimation. The penalty F is always 1 or more, because Δ is a variance-covariance matrix, and so is positive semi-definite.

3.2 Numerical Values of the Asymptotic Variance for Selected Models

Values of Δ are obtained by rewriting as

$$\Delta = \frac{w^{-1} \int_0^w \boldsymbol{h}(u) \boldsymbol{h}(u)^T g(u)^{-1} du}{\bar{g}} - \frac{\bar{\boldsymbol{h}} \bar{\boldsymbol{h}}^{\boldsymbol{T}}}{\bar{g}^2}$$

and calculating by quadrature using the *integrate* function in the R Statistical Environment (R Core Team, 2013). Table 2 shows values of F numerically calculated for various hazard rate models. The parameter θ_2 determines the shape of the detection curve, with 1.1 giving a very narrow shoulder (i.e. steeply declining for small distances) and 3 giving a very wide shoulder. Hazard rate detection models for a number of values of θ_2 are illustrated in Figure 1 in the next section. The parameter θ_1 is calculated numerically to give the specified $\bar{g}$ in each row. Table 2 shows that F increases as θ_2 decreases, i.e. as the shoulder becomes narrower. For given $\bar{g}$, F decreases as the coverage rate P increases. This is because increasing P improves the precision of $\hat{N}$, but it improves the precision of $\hat{N}_{known\theta}$ even faster. F decreases as $\bar{g}$ increases, for fixed P.

The scenarios in the table which are most common in practice are $\bar{g}$ equal to 0.3 or 0.6, and $P = 0.3$. F varies from 2.0 to 7.0 in this subset of the table.

Table 3 shows similar results for half-normal detection models. The values of F are generally much closer to 1 than in Table 2, varying from 1.6 to 2.6 in the subset of the table where $\bar{g} \in \{0.3, 0.6\}$ and $P = 0.3$. F increases with P, as in Table 2. However, F increases with $\bar{g}$, opposite to the hazard rate case. It is unclear why this is the case.

A possible reason is that the hazard rate function remains very close to 1 for a shoulder region of distances close to 0, and is insensitive to $\boldsymbol{\theta}$ in this region, so that uncertainty in $\boldsymbol{\theta}$ has little asymptotic effect in this region. The width of this shoulder region is governed by both θ_1 and θ_2 and is wider when $\bar{g}$ is high. In contrast the half-normal function decreases smoothly even for small distances, even for high $\bar{g}$.

Table 2: Asymptotic penalty (F) for the hazard rate model for selected values of the coverage rate P, the shape parameter (θ_2) and the mean detection rate $\bar{g}$. The cutoff is $w = 1$ in all cases.

P	$\bar{g}$	θ_2					
		1.1	1.25	1.5	2	2.5	3
0.3	0.3	6.97	5.56	4.19	2.91	2.33	2.00
0.6	0.3	7.63	6.06	4.54	3.12	2.47	2.11
0.9	0.3	8.44	6.68	4.97	3.38	2.65	2.25
0.3	0.6	5.62	4.72	3.78	2.82	2.34	2.05
0.6	0.6	6.92	5.77	4.56	3.33	2.71	2.34
0.9	0.6	9.23	7.63	5.96	4.24	3.38	2.87
0.3	0.9	3.85	3.41	2.91	2.36	2.06	1.87
0.6	0.9	5.52	4.82	4.04	3.16	2.69	2.39
0.9	0.9	11.94	10.25	8.35	6.24	5.08	4.35

Table 3: Asymptotic penalty (F) for the half-normal model for selected values of the coverage rate P and the mean detection rate $\bar{g}$. The cutoff is $w = 1$ in all cases.

P	$\bar{g}$	Penalty (F)
0.3	0.3	1.55
0.6	0.3	1.61
0.9	0.3	1.69
0.3	0.6	1.90
0.6	0.6	2.15
0.9	0.6	2.60
0.3	0.9	2.54
0.6	0.9	3.44
0.9	0.9	6.90

4 Simulation Study

4.1 Design of Simulation Study

Distance data are simulated for abundances N such that the expected numbers of detections are $E[n] = 100, 200, \ldots, 1000$, with 4 detection functions. 10,000 simulations are used throughout. Distances d_i for $i = 1, \ldots, N$ are generated as independent uniforms $U(0, 2)$. The range of observation, w, is set to 1, so the covered region consists of half of the full region ($P = 0.5$). Objects are detected with probability $g(d_i; \boldsymbol{\theta})$ when $d_i \leq w$ for each i, with detection independent across objects.

Figure 1 shows the 4 distance functions used: hazard rate (hr) functions $g(d) = 1 - e^{-(d/\theta_1)^{-\theta_2}}$ with $\boldsymbol{\theta} = (0.405, 1.25)$, $\boldsymbol{\theta} = (0.448, 2)$ and $\boldsymbol{\theta} = (0.484, 3)$ corresponding to a very narrow, narrow, and wide shoulder respectively; and a half-normal (hn) function $g(d) = e^{-(d/\theta)^2/2}$ with $\theta = 0.502$. These parameters give a variety of shapes of the detection function, with all 4 having the same average detection rate of $\bar{g}(\boldsymbol{\theta}) = 0.6$. This means that the number of detections is approximately $n \approx P\bar{g}(\boldsymbol{\theta})N = 0.3N$. The hazard rate and half-normal detection functions are among those proposed in Buckland et al. (2001). The two-parameter hazard rate function meets the requirement of Buckland et al. (2001, p41) of being a flexible model, giving some robustness to mis-specified detection function; in particular, it allows the shoulder to be narrow or wide. The half-normal detection function is less flexible, but is also often used, and would generally be easier to fit from data as it has only one parameter. The values of $g(w)$ for all three functions are at least 0.11, and all but the wide hazard rate function are at least 0.14, roughly in line with the rule of thumb for choice of w in Buckland et al. (2001, p16). The figure also shows the asymptotic penalty F due to unknown $\boldsymbol{\theta}$ for each detection function. The penalty ranges from 2 to 5.4.

CDS estimates of $\boldsymbol{\theta}$ and N are calculated by maximum likelihood as described in Section 3, using several alternative models: the hazard rate model; the half-normal model; and the hazard rate or half-normal models with $\boldsymbol{\theta}$ assumed to be fully known. The last option is of course unrealistic in practice, but is included to show the impact of uncertainty about $\boldsymbol{\theta}$. A model-averaged estimator of N is also calculated. This is the weighted average of the values of $\hat{N}$ from the hazard rate and half-normal models, with weights proportional to $\exp(-AIC/2)$ where AIC is the Akaike Information Criterion for each model (Burnham

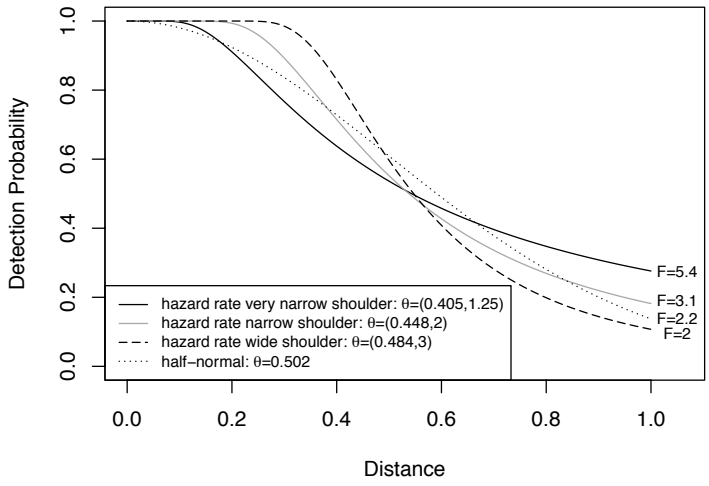

Figure 1: Detection Functions used to Generate Simulated Data. The variance penalty factors F from (25) due to unknown $\boldsymbol{\theta}$ are also shown.

and Anderson, 2002, section 2).

A strip transect expansion estimator $\hat{N} = n/P$ is also calculated. This is unbiased under the binomial model $n \sim bin(N, P)$, which incorrectly assumes perfect detection up to distance w. The expansion estimator is also applied to restricted datasets using only those distances up to a range of w' =0.1, 0.25 and 0.5, with $\hat{N} = n[d \leq w']/(Pw'/w)$. This is to test whether this simple approach works well when restricted to a narrow range, in order to reduce the bias due to uncorrected under-detection.

The variance of $\hat{N}$ is estimated by plugging in the maximum likelihood estimator of $\boldsymbol{\theta}$ into V_{11} from (20). The variance of the strip transect expansion estimator is estimated by making the assumption of perfect detection required by this estimator. Under this assumption, $\text{var}\,(n/P) = P^{-2}\text{var}(n) = P^{-1} - 1$.

All computations are carried out in the R statistical environment version 3.0.1 (R Core Team, 2013). The Distance package (Miller, 2014) is not used because of occasional non-convergence (this would generally not be an issue in practice, but is a problem in a large simulation). Instead, the log-likelihood l is maximised using the result that $\hat{N} = n/P\bar{g}\left(\hat{\boldsymbol{\theta}}\right)$ and then maximising the profile log- likelihood $l\left(\boldsymbol{\theta}, \hat{N}\left(\boldsymbol{\theta}\right)\right)$ with respect to $\boldsymbol{\theta}$. The maximisation is done via the *optim* function using the Nelder-Mead method for the two-parameter hazard rate function and the Brent method for the one-parameter half-normal function. Variances of the maximum likelihood estimators are estimated using the inverse negative hessian of l with respect to N and $\boldsymbol{\theta}$, calculated by the *hessian* function

in the numDeriv package (Gilbert and Varadhan, 2012). The complete simulation requires approximately 8 hours on a Macbook Pro with a 2.7GHz Intel Core i7 processor and 16GB of RAM. The program to conduct the simulation and produce the figures and table will be made available as online supplementary material.

4.2 Simulation Results

Table 4 compares the various estimators when the true detection function is the hazard rate function with very narrow shoulder. The CDS estimator using the correct (hazard rate) model has negligible bias, but quite high CV at 20%. Its variance is nearly 5.5 times higher than if $\boldsymbol{\theta}$ are known, quite close to the asymptotic penalty $F = 5.4$. If the half-normal model is incorrectly assumed, the relative bias is -9% but the relative root mean squared error (RRMSE) is substantially lower due to the much lower CV of this estimator. A simple strip expansion estimator using a short range of only 0.25 (out of a maximum of 1) performs similarly well. However, making the strip too narrow with a range of 0.1 leads to high CV, and making the strip too wide (0.5 or 1) leads to excessive bias. The model-averaged CDS estimator has slightly better performance than CDS with the hazard rate function.

Confidence interval non-coverage is somewhat high for CDS with hazard rate (13% vs nominal 5%). Coverage rates are best for the expansion estimator with short ranges (0.1 and 0.25). Model-averaging confidence intervals also have coverage close to nominal. The other estimators lead to poor coverage, presumably because of bias.

In Table 5, where a hazard rate with narrow shoulder applies, CDS using the correct distance function again has a high CV of 17%, although lower than when the shoulder is very narrow. The resulting confidence interval non-coverage is much improved at 7%. Otherwise, results are similar to Table 4. CDS using the incorrect half-normal model, and strip expansion with range of 0.25, are again the best options in terms of RRMSE.

When the true hazard rate detection function has a wide shoulder (Table 6), CDS using this model becomes the best option. Simple strip expansion with a range of 0.25, and model-averaging, have approximately the same RRMSE. CDS with half-normal detection function has about about 3 percentage points higher RRMSE.

The results in Table 7 with true detection function given by half-normal are very similar to those in 5 with the narrow-shoulder hazard rate. This is not surprising as these detection

functions are the most visually similar in Figure 1.

In Tables 4, 6 and 7, the simulation estimates of the efficiency are fairly close to the asymptotic value of F. They diverge by about 1/3 in Table 5. Figure 2 shows how the simulation estimates of F converge to the asymptotic values as n and N increase. For the hazard rate with very narrow shoulder and the halfnormal models, the asymptotic approximation is good even for $E[n] = 100$. For the other two models, the small sample penalties are quite a bit higher than the asymptotic value when $E[n] = 100$, but converge to the asymptote as $E[n]$ increases. The results provide a computational confirmation of the derivation of F in Section 3.

Table 4: Properties of estimators of N when expected sample size is 100 and distance function is hazard rate with very narrow shoulder. CDS is classical distance sampling, expansion refers to simple expansion of number of detections with distance less than or equal to some maximum range. Relative bias, coefficient of variation (CV), relative root mean squared error (RRMSE) and non-coverage of nominal 95% confidence intervals are shown. Efficiency is relative to the variance when $\boldsymbol{\theta}$ is known. Corresponding asymptotic value (F) shown in brackets.

Method	Rel.Bias(%)	CV(%)	RRMSE(%)	Noncoverage(%)	Efficiency
CDS with hazard rate model	1.4	19.5	19.6	13.0	5.47 (5.36)
CDS with halfnormal model	-8.7	11.5	14.4	16.8	1.90
expansion (range=0.1)	0.0	24.1	24.1	6.6	8.32
expansion (range=0.25)	-4.0	14.3	14.9	8.0	2.95
expansion (range=0.5)	-18.3	8.9	20.3	57.7	1.12
expansion (range=1)	-40.1	5.1	40.4	100.0	0.37
model-averaged CDS	-1.5	17.4	17.5	9.0	4.34

Table 5: Properties of estimators of N when expected sample size is 100 and distance function is hazard rate with narrow shoulder. CDS is classical distance sampling, expansion refers to simple expansion of number of detections with distance less than or equal to some maximum range. Relative bias, coefficient of variation (CV), relative root mean squared error (RRMSE) and non-coverage of nominal 95% confidence intervals are shown. Efficiency is relative to the variance when $\boldsymbol{\theta}$ is known. Corresponding asymptotic value (F) shown in brackets.

Method	Rel.Bias(%)	CV(%)	RRMSE(%)	Noncoverage(%)	Efficiency
CDS with hazard rate model	3.1	17.2	17.5	7.1	4.23 (3.13)
CDS with halfnormal model	2.0	12.3	12.5	5.1	2.17
expansion (range=0.1)	0.0	24.1	24.1	6.6	8.32
expansion (range=0.25)	-0.4	14.6	14.6	5.8	3.05
expansion (range=0.5)	-12.3	9.1	15.2	31.4	1.18
expansion (range=1)	-40.0	5.1	40.4	100.0	0.37
model-averaged CDS	3.6	14.7	15.1	3.2	3.08

Table 6: Properties of estimators of N when expected sample size is 100 and distance function is hazard rate with wide shoulder. CDS is classical distance sampling, expansion refers to simple expansion of number of detections with distance less than or equal to some maximum range. Relative bias, coefficient of variation (CV), relative root mean squared error (RRMSE) and non-coverage of nominal 95% confidence intervals are shown. Efficiency is relative to the variance when $\boldsymbol{\theta}$ is known. Corresponding asymptotic value (F) shown in brackets.

Method	Rel.Bias(%)	CV(%)	RRMSE(%)	Noncoverage(%)	Efficiency
CDS with hazard rate model	2.2	14.3	14.4	5.9	2.91 (2.23)
CDS with halfnormal model	11.6	12.9	17.3	10.5	2.38
expansion (range=0.1)	0.0	24.1	24.1	6.6	8.32
expansion (range=0.25)	0.0	14.6	14.6	5.7	3.06
expansion (range=0.5)	-7.4	9.3	11.8	15.5	1.23
expansion (range=1)	-40.0	5.1	40.3	100.0	0.37
model-averaged CDS	6.9	13.1	14.8	3.7	2.47

Table 7: Properties of estimators of N when expected sample size is 100 and distance function is halfnormal. CDS is classical distance sampling, expansion refers to simple expansion of number of detections with distance less than or equal to some maximum range. Relative bias, coefficient of variation (CV), relative root mean squared error (RRMSE) and non-coverage of nominal 95% confidence intervals are shown. Efficiency is relative to the variance when $\boldsymbol{\theta}$ is known. Corresponding asymptotic value (F) shown in brackets.

Method	Rel.Bias(%)	CV(%)	RRMSE(%)	Noncoverage(%)	Efficiency
CDS with hazard rate model	-2.3	17.6	17.8	15.5	4.45
CDS with halfnormal model	0.4	12.0	12.0	5.0	2.05 (2.05)
expansion (range=0.1)	-0.7	24.0	24.0	6.7	8.24
expansion (range=0.25)	-4.0	14.3	14.9	8.1	2.95
expansion (range=0.5)	-14.4	9.0	17.0	40.4	1.16
expansion (range=1)	-40.1	5.0	40.4	100.0	0.36
model-averaged CDS	-0.2	14.2	14.2	5.1	2.89

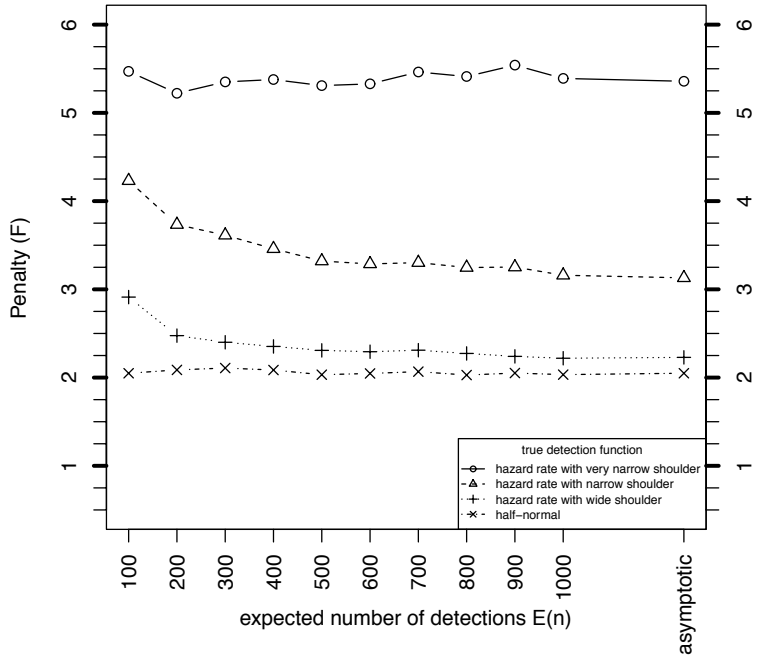

Figure 2: Variance of MLE of N when $\boldsymbol{\theta}$ is unknown relative to when it is known, for various sample sizes based on simulation, and asymptotically from (25), where $P = E[N_c/N] = 0.5$

5 Discussion

Like any statistical method, distance sampling works well in some situations and less well in others. It greatly reduces the bias due to imperfect detection associated with distance. On the other hand, it may introduce other biases due to mis-specified detection function or non-uniform distribution of distances. It also creates a new source of variability, due to the need to estimate detection function parameters.

This paper reveals the extent of this source of variability. Section 3 derives an expression for the asymptotic variance factor due to unknown detection parameters. Depending on the coverage rate and detection function, this penalty factor can be substantial: in some cases over 10, but more typically between 2 and 4. Simulation confirms the asymptotic result, with even greater penalties sometimes applying when the number of detections is small. The penalty is most severe when the detection function has a narrow or very narrow shoulder; this is precisely the scenario that distance sampling is intended for.

The choice of a statistical technique is a tradeoff between multiple biases, cost, simplicity and variance. The tradeoff is different in every application. Distance sampling removes one source of bias provided its assumptions are justified, but it does so at the price of inflated variances. If their variances are large enough, estimates are of little practical use even if their biases are small. Other techniques must then be considered even if they may be more biased. For example indirect measures such as counts of tracks, burrows or scats (e.g. Morrison 2009, p184) may give less volatile indicators of change over time, even though they do not provide abundance estimates. Simple strip expansion is another option which can give reduced mean squared error if the goldilocks strip width can be identified. Alternatively the detection function can be estimated from a separate calibration study (Melville and Welsh, 2001).

The choice of methodology for assessing abundance, as well as the determination of required sample size, should be informed by consideration of all relevant biases and by the likely achievable precision. The results in this paper will help in this process.

References

Barry, S. C., Welsh, A., 2001. Distance sampling methodology. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (1), 23–31.

Borchers, D. L., Buckland, S. T., 2002. Estimating Animal Abundance: Closed Populations. Springer-Verlag, London.

Buckland, S., Anderson, D., Burnham, K., Laake, J., Borchers, D., 2001. Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press, Oxford.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., Thomas, L., 2004. Advanced Distance Sampling. Oxford University Press, Oxford.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., et al., 1993. Distance Sampling: Estimating Abundance of Biological Populations. Chapman & Hall, London.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Clark, R. G., 2015. Efficiency and robustness in distance sampling. `http://niasra.uow.edu.au/publications/UOW185981.html`, National Institute for Applied Statistics Research Australia (NIASRA) Working Paper.

Courtois, E. A., Devillechabrolle, J., Dewynter, M., Pineau, K., Gaucher, P., Chave, J., 2013. Monitoring strategy for eight amphibian species in French Guiana, South America. PloS One 8 (6), e67486.

Dawson, M., 2009. 2009 Aerial Survey of Feral Horses in the Australian Alps. A report prepared for the Australian Alps Liaison CommitteeAvailable from http://www.australianalps.environment.gov.au/publications/research-reports/pubs/2009feralhorsealpssurvey.pdf.

Fewster, R., Jupp, P., 2009. Inference on population size in binomial detectability models. Biometrika 96 (4), 805–820.

Fewster, R. M., Laake, J. L., Buckland, S. T., 2005. Line transect sampling in small and large regions. Biometrics 61 (3), 856–859.

Fewster, R. M., Southwell, C., Borchers, D. L., Buckland, S. T., Pople, A. R., 2008. The influence of animal mobility on the assumption of uniform distances in aerial line-transect surveys. Wildlife Research 35 (4), 275–288.

Gilbert, P., Varadhan, R., 2012. numDeriv: Accurate Numerical Derivatives. R package version 2012.9-1.
URL `http://CRAN.R-project.org/package=numDeriv`

Grimmett, G., Stirzaker, D., 2001. Probability and Random Processes. Oxford University Press, Oxford.

Harville, D. A., 1997. Matrix Algebra from a Statistician's Perspective. Vol. 157. Springer, New York.

Johnson, D. S., Laake, J. L., Ver Hoef, J. M., 2010. A model-based approach for making ecological inference from distance sampling data. Biometrics 66 (1), 310–318.

Kruger, J., Reilly, B., Whyte, I., 2008. Application of distance sampling to estimate population densities of large herbivores in Kruger National Park. Wildlife Research 35 (4), 371–376.

Laake, J., Dawson, M. J., Hone, J., 2008. Visibility bias in aerial survey: mark–recapture, line-transect or both? Wildlife Research 35 (4), 299–309.

Lehmann, E. L., 1999. Elements of Large-Sample Theory. Springer, New York.

Melville, G., Welsh, A., 2001. Line transect sampling in small regions. Biometrics 57 (4), 1130–1137.

Melville, G. J., Tracey, J. P., Fleming, P. J., Lukins, B. S., 2008. Aerial surveys of multiple species: critical assumptions and sources of bias in distance and mark–recapture estimators. Wildlife Research 35 (4), 310–348.

Miller, D. L., 2014. Distance: a simple way to fit detection functions to distance sampling data and calculate abundance/density for biological populations. R package version 0.9.
URL `{http://CRAN.R-project.org/package=Distance}`

Morrison, M. L., 2009. Restoring Wildlife: Ecological Concepts and Practical Applications. Island Press.

Oedekoven, C., Buckland, S., Mackenzie, M., King, R., Evans, K., Burger Jr, L., 2014. Bayesian methods for hierarchical distance sampling models. Journal of Agricultural, Biological, and Environmental Statistics, 1–21.

Oppel, S., 2006. Using distance sampling to quantify odonata density in tropical rainforests. International Journal of Odonatology 9 (1), 81–88.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Thomas, L., Buckland, S. T., Rexstad, E. A., Laake, J. L., Strindberg, S., Hedley, S. L., Bishop, J. R., Marques, T. A., Burnham, K. P., 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology 47 (1), 5–14.

Walter, M., 2002. The Population Ecology of Wild Horses in the Australian Alps. Ph.D. thesis, University of Canberra.

Walter, M., 2003. The effect of fire on wild horses in the Australian Alps National parks. A report prepared for the Australian Alps Liaison CommitteeAvailable from http://australianalps.environment.gov.au/publications/research-reports/pubs/feral-horses-fire-effects.pdf.

Walter, M. J., Hone, J., 2003. A comparison of 3 aerial survey techniques to estimate wild horse abundance in the Australian Alps. Wildlife Society Bulletin, 1138–1149.

Appendix 1: Proof of Theorem 1

Applying (13), we approximate $l = \log(L)$ by

$$\begin{aligned} l &\approx N\log(N) - (N-n)\log(N-n) - n\log(n) + (N-n)\log\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right) \\ &\quad + n\log(P) - n\log(w) + \sum_{i=1}^{n}\log g\left(d_i;\boldsymbol{\theta}\right) \end{aligned} \tag{26}$$

The next step is to differentiate l to obtain the score function:

$$\begin{aligned} \frac{\partial l}{\partial N} &= N\cdot N^{-1} + 1\cdot\log(N) - (N-n)\cdot(N-n)^{-1} - \log(N-n) + \log\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right) \\ &= \log(N) - \log(N-n) + \log\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right) \qquad (27) \\ \frac{\partial l}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^{n} g\left(d_i;\boldsymbol{\theta}\right)^{-1}\boldsymbol{h}\left(d_i;\boldsymbol{\theta}\right) - (N-n)\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \qquad (28) \end{aligned}$$

The MLE is obtained by setting (27) and (28) to 0. Firstly, set (27) to 0 and exponentiate both sides:

$$1 = \hat{N}\left(\hat{N} - n\right)^{-1}\left(1 - P\bar{g}\left(\hat{\boldsymbol{\theta}}\right)\right)$$

which leads directly to

$$\hat{N} = n/\left\{P\bar{g}\left(\hat{\boldsymbol{\theta}}\right)\right\}. \tag{29}$$

Setting (28) to 0, and then substituting for $\hat{N}$ from (29) gives an estimating equation for $\boldsymbol{\theta}$:

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} g\left(d_i;\boldsymbol{\theta}\right)^{-1}\boldsymbol{h}\left(d_i;\boldsymbol{\theta}\right) - \left(\hat{N} - n\right)\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n} g\left(d_i;\boldsymbol{\theta}\right)^{-1}\boldsymbol{h}\left(d_i;\boldsymbol{\theta}\right) - \left(nP^{-1}\bar{g}\left(\boldsymbol{\theta}\right)^{-1} - n\right)\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n} g\left(d_i;\boldsymbol{\theta}\right)^{-1}\boldsymbol{h}\left(d_i;\boldsymbol{\theta}\right) - nP^{-1}\bar{g}\left(\boldsymbol{\theta}\right)^{-1}\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)^{-1} P\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n} g\left(d_i;\boldsymbol{\theta}\right)^{-1}\boldsymbol{h}\left(d_i;\boldsymbol{\theta}\right) - n\bar{g}\left(\boldsymbol{\theta}\right)^{-1}\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) \qquad (30) \end{aligned}$$

Results (29) and (30) give the Theorem result on the maximum likelihood estimators of N and $\boldsymbol{\theta}$.

The Fisher Information is given by the variance of the score vector, the elements of which are given by (27) and (28). It can be written as

$$\mathcal{I}(N,\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_{NN} & \mathcal{I}_{N\theta}^T \\ \mathcal{I}_{N\theta} & \mathcal{I}_{\theta\theta} \end{bmatrix} = \begin{bmatrix} \text{var}(\partial l/\partial N) & \text{cov}(\partial l/\partial N, \partial l/\partial \boldsymbol{\theta})^T \\ \text{cov}(\partial l/\partial N, \partial l/\partial \boldsymbol{\theta}) & \text{var}(\partial l/\partial \boldsymbol{\theta}) \end{bmatrix}.$$

The block elements of $\mathcal{I}$ are easily derived. Some preliminary notes:

(a) Let D be a random variable with density $g(d;\boldsymbol{\theta})/\int_0^w g(u;\boldsymbol{\theta})du$ for $0 \le d \le w$.

(b) The distances d_i follow the same distribution as D, conditional on n.

(c) $E\left(g(d_i;\boldsymbol{\theta})^{-1}\boldsymbol{h}(d_i;\boldsymbol{\theta})|n\right) = \int_0^w g(d;\boldsymbol{\theta})^{-1}\boldsymbol{h}(d;\boldsymbol{\theta})\,g(d;\boldsymbol{\theta})\,dd/\int_0^w g(u;\boldsymbol{\theta})du = \bar{\boldsymbol{h}}(\boldsymbol{\theta})/\bar{g}(\boldsymbol{\theta})$.

For the remainder of the proof, I will write $g(u)$, $\boldsymbol{h(u)}$, $\bar{g}$ and $\bar{\boldsymbol{h}}$ for readability, rather than $g(u;\boldsymbol{\theta})$ etc. Using (a), (b) and (c), we obtain:

$$\begin{aligned}
\mathcal{I}_{\theta\theta} &= \text{var}(\partial l/\partial\boldsymbol{\theta}) = \text{E}\{\text{var}(\partial l/\partial\boldsymbol{\theta}|n)\} + \text{var}\{\text{E}(\partial l/\partial\boldsymbol{\theta}|n)\} \\
&= \text{E}(n\text{var}(\boldsymbol{h}(D)/g(D))) + \text{var}\left\{n\bar{\boldsymbol{h}}\bar{g}^{-1} - (N-n)(1-P\bar{g})^{-1}P\bar{\boldsymbol{h}}\right\} \\
&= E[n]\text{var}(\boldsymbol{h}(D)/g(D)) + \text{var}\left\{n\bar{\boldsymbol{h}}\bar{g}^{-1}(1-P\bar{g})^{-1}(1-P\bar{g}+P\bar{g}) + \text{const.}\right\} \\
&= E[n]\text{var}(\boldsymbol{h}(D)/g(D)) + \text{var}\left\{n\bar{\boldsymbol{h}}\bar{g}^{-1}(1-P\bar{g})^{-1}\right\} \\
&= NP\bar{g}\text{var}(\boldsymbol{h}(D)/g(D)) + \text{var}(n)\,\bar{\boldsymbol{h}}\bar{\boldsymbol{h}}^{\boldsymbol{T}}\bar{g}^{-2}(1-P\bar{g})^{-2} \\
&= NP\bar{g}\text{var}(\boldsymbol{h}(D)/g(D)) + NP\bar{g}(1-P\bar{g})\,\bar{\boldsymbol{h}}\bar{\boldsymbol{h}}^{\boldsymbol{T}}\bar{g}^{-2}(1-P\bar{g})^{-2} \\
&= NP\bar{g}\text{var}(\boldsymbol{h}(D)/g(D)) + NP\bar{\boldsymbol{h}}\bar{\boldsymbol{h}}^{\boldsymbol{T}}\bar{g}^{-1}(1-P\bar{g})^{-1} \qquad (31)
\end{aligned}$$

$$\begin{aligned}
\mathcal{I}_{N\theta} &= \text{cov}(\partial l/\partial N, \partial l/\partial\boldsymbol{\theta}) = \text{Ecov}(\partial l/\partial N, \partial l/\partial\boldsymbol{\theta}|n) + \text{cov}\{\text{E}(\partial l/\partial N|n), \text{E}(\partial l/\partial\boldsymbol{\theta}|n)\} \\
&= 0 - \text{cov}\left\{\log(N-n), n\bar{g}^{-1}\bar{\boldsymbol{h}} + n(1-\bar{g})^{-1}\bar{\boldsymbol{h}}\right\} \\
&= -\bar{g}^{-1}(1-\bar{g})^{-1}\bar{\boldsymbol{h}}\text{cov}\{\log(N-n), n\} \qquad (32)
\end{aligned}$$

As $N \to \infty$ $n \xrightarrow{p} NP\bar{g}$, so the right hand side of (32) can be approximated using a first

order Taylor Series of n about $E[n] = NP\bar{g}$:

$$\begin{aligned}
\mathcal{I}_{N\theta} &\approx -\bar{g}^{-1}\left(1-P\bar{g}\right)^{-1} P\bar{\boldsymbol{h}}\cdot \\
&\quad \operatorname{cov}\left\{\log\left(N-NP\bar{g}\right)-\left(N-NP\bar{g}\right)^{-1}\left(n-NP\bar{g}\right), n\right\} \\
&= \bar{g}^{-1}\left(1-P\bar{g}\right)^{-1}\bar{\boldsymbol{h}}N^{-1}\left(1-P\bar{g}\right)^{-1}\operatorname{var}(n) \\
&= \bar{g}^{-1}\left(1-P\bar{g}\right)^{-2}\bar{\boldsymbol{h}}N^{-1}NP\bar{g}\left(1-NP\bar{g}\right) \\
&= \left(1-P\bar{g}\right)^{-1} P\bar{\boldsymbol{h}} \qquad (33)
\end{aligned}$$

The top-left element of $\mathcal{I}$, $\mathcal{I}_{NN}$, can also be approximated by a first order Taylor Series about $n = N\bar{g}$:

$$\begin{aligned}
\mathcal{I}_{NN} &= \operatorname{var}\left(\partial l/\partial N\right) = \operatorname{var}\left(\log\left(N-n\right)\right) \qquad (34) \\
&\approx \operatorname{var}\left\{-\log\left(N-NP\bar{g}\right)+\left(N-NP\bar{g}\right)^{-1}\left(n-NP\bar{g}\right)\right\} \\
&= N^{-2}\left(1-P\bar{g}\right)^{-2}\operatorname{var}(n) \\
&= N^{-2}\left(1-P\bar{g}\right)^{-2} NP\bar{g}\left(1-P\bar{g}\right) \\
&= N^{-1}\left(1-P\bar{g}\right)^{-1} P\bar{g} \qquad (35)
\end{aligned}$$

Appendix 2: Proof of Theorem 2

The required result is a limit as $N \to \infty$ with $\boldsymbol{\theta}$ held fixed. The proof has three parts: (I) the limiting distribution of $\hat{\boldsymbol{\theta}}$ conditional on n letting $n \to \infty$; (II) the limiting distribution of n as $N \to \infty$; (III) the limiting joint distribution of $\hat{\boldsymbol{\theta}}$ and n as $N \to \infty$; and (IV) the limiting joint distribution of $\hat{\boldsymbol{\theta}}$ and $\hat{N}$ as $N \to \infty$. The major challenge is to move from (I) and (II) to (III), as this switches from a limit in n to a limit in N.

(I) Limiting Distribution of $\hat{\boldsymbol{\theta}}$ conditional on n as $n \to \infty$

Conditional on n, $\boldsymbol{d_s}$ are i.i.d. with density

$$g_{cond}(d) = g\left(d;\boldsymbol{\theta}\right) / \int_0^w g\left(u;\boldsymbol{\theta}\right) du = g\left(d;\boldsymbol{\theta}\right) / \left(w\bar{g}\left(\boldsymbol{\theta}\right)\right)$$

on $(0, w)$. The estimator $\hat{\boldsymbol{\theta}}$ is the solution to estimating equation (30). This equation is in fact the score equation conditional on n (Buckland et al., 2004). We make use of the central limit theorem for MLEs (Lehmann, 1999, Theorem 7.5.1), which requires assumptions (M1) through (M6) and (M2') stated in that text. (M1) follows from assumption (iv). (M2) and (M2') follow from both (iv) and (v). (M3) follows from the model since d_i are i.i.d. given n. (M4) follows from (iv). (M5) is (ii) and (M6) is (iii). (These assumptions are satisfied for $\boldsymbol{d_s}$ given n, but not for the full data $(n, \boldsymbol{d_s})$, as noted in the discussion following Theorem 1.) Theorem 7.5.1 of Lehmann (1999) applies and so

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}_{cond}^{-1}\right) \tag{36}$$

conditional on n as $n \to \infty$, where

$$\begin{aligned}
\mathcal{I}_{cond} &= \mathrm{var}\left(\partial_{\boldsymbol{\theta}} \log g_{cond}\left(D;\boldsymbol{\theta}\right)\right) \\
&= \mathrm{var}\left[\partial_{\boldsymbol{\theta}} \left\{\log g\left(D;\boldsymbol{\theta}\right) - \log\left(w\bar{g}\left(\boldsymbol{\theta}\right)\right)\right\}\right] \\
&= \mathrm{var}\left[\boldsymbol{h}\left(D;\boldsymbol{\theta}\right) / g\left(D;\boldsymbol{\theta}\right)\right] = \Delta.
\end{aligned}$$

using the definition of Δ in (21). Limit (36) can be equivalently expressed as

$$\boldsymbol{t_1} \xrightarrow{d} \boldsymbol{T_1} \tag{37}$$

where $\boldsymbol{t_1} = \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ and $\boldsymbol{T_1} \sim N\left(\mathbf{0}, \Delta^{-1}\right)$. An equivalent expression is

$$P\left[\boldsymbol{t_1} \in A_1 | n\right] = P\left[\boldsymbol{T_1} \in A_1 | n\right] + r\left(n; A_1\right) \tag{38}$$

where A_1 is any measurable set and $r\left(n; A_1\right)$ is a remainder term satisfying $\lim_{n\to\infty} r\left(n; A_1\right) = 0$.

(II) Limiting Distribution of n as $N \to \infty$

Under the model assumptions, $n \sim \text{bin}\left(N, P\bar{g}\left(\boldsymbol{\theta}\right)\right)$. Let $t_2 = N^{-1/2}\left(n - NP\bar{g}\left(\boldsymbol{\theta}\right)\right)$. The Central Limit Theorem for binomial random variables means that

$$t_2 \xrightarrow{d} T_2 \tag{39}$$

where $T_2 \sim N\left(0, P\bar{g}\left(\boldsymbol{\theta}\right)\left(1 - P\bar{g}\left(\boldsymbol{\theta}\right)\right)\right)$.

Note also that

$$n/N \xrightarrow{p} P\bar{g}\left(\boldsymbol{\theta}\right) \tag{40}$$

from the Weak Law of Large Numbers (e.g. Grimmett and Stirzaker 2001, ch.5).

(III) Limiting Joint Distribution of $\hat{\boldsymbol{\theta}}$ and n as $N \to \infty$

Let A_1 and A_2 be measurable sets. Let $1(E)$ be equal to 1 when event E is true and 0 otherwise. Using the fact that t_2 is a function of n, we obtain:

$$\begin{aligned}
P\left[\boldsymbol{t_1} \in A_1, t_2 \in A_2\right] &= E_n\left\{P\left[\boldsymbol{t_1} \in A_1, t_2 \in A_2 | n\right]\right\} \\
&= E_n\left\{1\left(t_2 \in A_2\right) P\left[\boldsymbol{t_1} \in A_1 | n\right]\right\} \\
&= E_n\left\{1\left(t_2 \in A_2\right)\left(P\left[\boldsymbol{T_1} \in A_1\right] + r\left(n; A_1\right)\right)\right\} \\
&= P\left[\boldsymbol{T_1} \in A_1\right] E_n\left[1\left(t_2 \in A_2\right)\right] + E_n\left[1\left(t_2 \in A_2\right) r\left(n; A_1\right)\right] \\
&= P\left[\boldsymbol{T_1} \in A_1\right] P\left(t_2 \in A_2\right) + E_n\left\{1\left(t_2 \in A_2\right) r\left(n; A_1\right)\right\}
\end{aligned} \tag{41}$$

The Strong Law of Large Numbers (e.g. Grimmett and Stirzaker 2001, p329) implies that $n/N \xrightarrow{as} P\bar{g}\left(\boldsymbol{\theta}\right)$ as $N \to \infty$, which in turn means that

$$n \xrightarrow{as} \infty \tag{42}$$

as $N \to \infty$ (where "a.s." denotes almost sure convergence), which implies that $r(n; A_1) \xrightarrow{as} 0$ as $N \to \infty$. Hence

$$1(t_2 \in A_2)\, r(n; A_1) \xrightarrow{as} 0 \tag{43}$$

as $N \to \infty$. Since $r(n; A_1)$ is a difference between two probabilities, and $1(t_2 \in A_2)$ is either 0 or 1, it follows that

$$|1(t_2 \in A_2)\, r(n; A_1)| \leq 1 \tag{44}$$

Using the dominated convergence theorem (e.g. Grimmett and Stirzaker 2001, p180), (43) and (44) imply that

$$\lim_{N\to\infty} E_n \{1(t_2 \in A_2)\, r(n; A_1)\} = 0 \tag{45}$$

Using (45) and (39), taking the limit of (41) as $N \to \infty$ gives:

$$\lim_{N\to\infty} P[\boldsymbol{t_1} \in A_1, t_2 \in A_2] = P[\boldsymbol{T_1} \in A_1] \lim_{N\to\infty} P(t_2 \in A_2) + 0 = P[\boldsymbol{T_1} \in A_1]\, P(T_2 \in A_2) \tag{46}$$

Limit (46) applies for any measurable sets A_1 and A_2 and so can be written equivalently as

$$\begin{pmatrix} t_2 \\ \boldsymbol{t_1} \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} P\bar{g}(\boldsymbol{\theta})[1 - P\bar{g}(\boldsymbol{\theta})] & \mathbf{0^T} \\ \mathbf{0} & \Delta^{-1} \end{bmatrix}\right) \tag{47}$$

Let $\boldsymbol{t_3} = \boldsymbol{t_1}\sqrt{N/n} = \sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$. Using (40) and (47), Slutsky's Theorem (e.g. Grimmett and Stirzaker 2001, p318) means that

$$\begin{pmatrix} t_2 \\ \boldsymbol{t_3} \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} P\bar{g}(\boldsymbol{\theta})[1 - P\bar{g}(\boldsymbol{\theta})] & \mathbf{0^T} \\ \mathbf{0} & P^{-1}\bar{g}^{-1}\Delta^{-1} \end{bmatrix}\right) \tag{48}$$

Writing $y = n/N$, (48) can be re-written as

$$N^{1/2} \begin{pmatrix} y - P\bar{g}(\boldsymbol{\theta}) \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Psi) \tag{49}$$

where

$$\Psi = \begin{bmatrix} P\bar{g}(\boldsymbol{\theta})[1 - P\bar{g}(\boldsymbol{\theta})] & \mathbf{0_p^T} \\ \mathbf{0_p} & P^{-1}\bar{g}^{-1}\Delta^{-1} \end{bmatrix}.$$

and $\mathbf{0_p}$ is a p by 1 vector of zeroes.

(IV) Limiting Joint Distribution of $\hat{\boldsymbol{\theta}}$ and $\hat{N}$ as $N \to \infty$

Let

$$\boldsymbol{t_4} = \begin{pmatrix} \hat{N}/N \\ \hat{\boldsymbol{\theta}} \end{pmatrix} = \begin{pmatrix} y/Pg\left(\hat{\boldsymbol{\theta}}\right) \\ \hat{\boldsymbol{\theta}} \end{pmatrix} = f\left(y, \hat{\boldsymbol{\theta}}\right).$$

The final part of the proof uses the Delta method to obtain the limiting distribution of $\boldsymbol{t_4}$ from (49). Firstly note that

$$f\left(\begin{bmatrix} P\bar{g}\left(\boldsymbol{\theta}\right) \\ \boldsymbol{\theta} \end{bmatrix}\right) = \begin{bmatrix} 1 \\ \boldsymbol{\theta} \end{bmatrix}.$$

Secondly, we find the Jacobian matrix, i.e. the partial derivatives of $f\left(y, \hat{\boldsymbol{\theta}}\right)$, evaluated at $y = P\bar{g}\left(\boldsymbol{\theta}\right)$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$:

$$\begin{aligned} J &= \left.\frac{\partial f\left(y, \hat{\boldsymbol{\theta}}\right)}{\partial\left(y, \hat{\boldsymbol{\theta}}\right)}\right|_{y=P\bar{g}(\boldsymbol{\theta}), \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = \left.\begin{bmatrix} P^{-1}\bar{g}\left(\hat{\boldsymbol{\theta}}\right)^{-1} & \mathbf{0_p^T} \\ -yP^{-1}\bar{g}\left(\hat{\boldsymbol{\theta}}\right)^{-2}\bar{\boldsymbol{h}}\left(\hat{\boldsymbol{\theta}}\right) & I_p \end{bmatrix}\right|_{y=P\bar{g}(\boldsymbol{\theta}), \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \\ &= \begin{bmatrix} P^{-1}\bar{g}\left(\boldsymbol{\theta}\right)^{-1} & \mathbf{0_p^T} \\ -\bar{g}\left(\boldsymbol{\theta}\right)^{-1}\bar{\boldsymbol{h}}\left(\boldsymbol{\theta}\right) & I_p \end{bmatrix}. \end{aligned}$$

where I_p is the p by p identity matrix. The Delta Method (Lehmann, 1999, Theorem 5.4.6) then means that

$$N^{1/2}\left\{\boldsymbol{t_4} - \begin{pmatrix} 1 \\ \boldsymbol{\theta} \end{pmatrix}\right\} \xrightarrow{d} N\left(0, J^T\Psi J\right) \tag{50}$$

The variance in (50) can be expanded as:

$$\begin{aligned} J^T\Psi J &= \begin{bmatrix} P^{-1}\bar{g}^{-1} & -\bar{g}^{-1}\bar{\boldsymbol{h}}^T \\ \mathbf{0_p} & I_p \end{bmatrix}\begin{bmatrix} P\bar{g}\left(1-P\bar{g}\right) & \mathbf{0_p^T} \\ \mathbf{0_p} & P^{-1}\bar{g}^{-1}\Delta^{-1} \end{bmatrix}\begin{bmatrix} P^{-1}\bar{g}^{-1} & \mathbf{0_p} \\ -\bar{g}^{-1}\bar{\boldsymbol{h}} & I_p \end{bmatrix} \\ &= \begin{bmatrix} P^{-1}\bar{g}^{-1} & -\bar{g}^{-1}\bar{\boldsymbol{h}}^T \\ \mathbf{0_p} & I_p \end{bmatrix}\begin{bmatrix} \left(1-P\bar{g}\right) & \mathbf{0_p^T} \\ -P^{-1}\bar{g}^{-2}\Delta^{-1}\bar{\boldsymbol{h}} & P^{-1}\bar{g}^{-1}\Delta^{-1} \end{bmatrix} \\ &= \begin{bmatrix} P^{-1}\bar{g}^{-1}\left(1-P\bar{g}\right) + P^{-1}\bar{g}^{-3}\bar{\boldsymbol{h}}^{\boldsymbol{T}}\Delta^{-1}\bar{\boldsymbol{h}} & -P^{-1}\bar{g}^{-2}\bar{\boldsymbol{h}}^{\boldsymbol{T}}\Delta^{-1} \\ -P^{-1}\bar{g}^{-2}\Delta^{-1}\bar{\boldsymbol{h}} & P^{-1}\bar{g}^{-1}\Delta^{-1} \end{bmatrix}. \end{aligned} \tag{51}$$

Equations (50) and (51) are the required result.