

CREEM

Centre for Research into Ecological and Environmental Modelling
University of St Andrews

INTERNATIONAL WORKSHOPS

Introduction to Distance Sampling

St. Andrews
16th August – 19th August 2016

Introduction to Distance Sampling 2016

CREEM, University of St. Andrews

INDEX

	Page	
Lectures		
1	Introduction to Distance Sampling	1
2	Choosing Detection Functions	9
3	Further Ways to Think About Line Transects	17
4	Assessing Model Performance	21
5	Making Distance Sampling Work	27
6	Measures of Precision	33
7	Point Transect Sampling	48
8	Survey Design	59
9	Automated Survey Design	73
10	Stratification and Clustered Populations	86
11	Covariates in the Detection Function	96
12	Multipliers and Indirect Survey Methods	104
13	Field Methods	112
Exercises		
1	Line Transect Estimation	119
2	Line Transect Estimation with Distance (duck nests)	120
3	Line Transect Estimation with Distance	126
4	Variance Estimation Methods	129
5	Point Transects	133
6	Automated Survey Design	135
7a	Analysis of Stratified Data	141
7b	Analysis of Clustered Data	143
8	Covariates in detection function	146
9a	Analysis with Use of Multipliers	150
9b	Analysis of Cue Counts	152
Solutions		
1	Line Transect Estimation	153
2	Line Transect Estimation with Distance (duck nests)	155
3	Line Transect Estimation with Distance	156
4	Variance Estimation Methods	157
5	Point Transects	160
6	Automated Survey Design	162
7a	Analysis of Stratified Data	164
7b	Analysis of Clustered Data	165
8	Covariates in detection function	167
9a	Analysis with Use of Multipliers	169
9b	Analysis of Cue Counts	171

Schedule
Introduction to Distance Sampling
16th August – 19th August 2016

Tuesday 16th August

08:45	Registration
09:00	Participant introductions Review of: methods for estimating animal abundance; distance sampling; choosing a detection function
10:45	Coffee/tea
11:00	Review of: goodness of fit; distance sampling assumptions Analysis of line transect data in Distance Computer session: line transect exercises
12:45	Lunch
13:45	Precision of distance sampling estimates Computer session: assessing precision
15:45	Coffee/tea
16:00	Demonstration of data import Participants' data
18:00	Adjourn

Wednesday 17th August

09:00	Point transect sampling Computer session: point transect exercises
10:45	Coffee/tea
11:00	Survey design Participants' data
12:45	Lunch
13:45	Automated survey design Computer session: automated survey design
15:45	Coffee/tea
16:00	Stratification and cluster size complications Computer session: choice of exercises Participants' data
18:00	Adjourn

Thursday 18th August

09:00	Fitting detection function with covariates Computer session: covariates in detection function
10:45	Coffee/tea
11:00	Participants' data Indirect survey methods and use of multipliers
12:45	Lunch
14:00	Computer session: multipliers Participants' data
15:45	Coffee/tea
16:00	Participants' data
18:00	Adjourn

Friday 19th August

09:00	Field methods Description of reprints distributed to participants
10:45	Coffee/tea
11:00	Participants' data Special topics
12:45	Lunch
14:00	Participants' data
15:45	Coffee/tea
16:00	Participants' data
16:45	Workshop summary
17:00	Workshop closes

Introduction to Distance Sampling

- Overview of wildlife population assessment methods
- Plot sampling
- Distance sampling
 - Basic idea
 - Types of distance sampling

Wildlife Population Assessment

- How many are there?
- What are their trends?
- Why?
 - Vital rates (survival, fecundity, etc)
- What might happen if...?
 - Scenario planning
 - Risk assessment
 - Decision support

Rapid assessment methods and indices

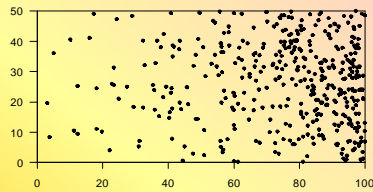
- Perhaps emphasis is just on trends
 - Questionnaire surveys
 - e.g. UK adder survey
 - Presence/absence
 - e.g. UK otter surveys
 - Index methods
 - e.g., Point counts for birds (US Breeding Bird Survey)
- **Warning!**
 - For estimating trends, must assume no trend in proportion detected

Methods of estimating abundance

- Complete census
- Plot sampling
- Distance sampling
- Mark-recapture
- Removal method

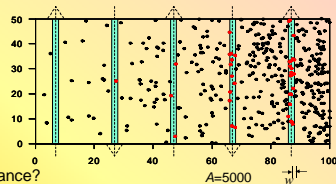
Complete census

- Let
 N = population size (abundance)
 A = size of study region = 5000
 D = animal density = N/A
- Method: count everything!
 $N = 412$
 $D = 412/5000 = 0.0824$
- Rarely possible in practice!



Plot sampling (or strip transect)

- Let
 k = number of strips = 5
 L = total line length = $50 \times 5 = 250$
 w = the strip half-width = 1
 a = area of region covered
 $= 2wL = 2 \times 1 \times 250 = 500$
 n = number of animals counted = 36
- From this, how do we estimate abundance?



Intuitive estimator of abundance

- I saw 36 animals
- I covered $500/5000 = 1/10^{\text{th}}$ of the study region
- So, I estimate there are $36/(1/10) = 36 \times 10 = 360$ animals

$$\hat{N} = \frac{n}{a/A} = \frac{nA}{a} = \frac{36 \times 5000}{500} = 360$$

(Hat "" means an estimate.)

Concept – Plot sampling

- Step 1: How many in covered region, N_a ?

Plot sampling: $N_a = n$

- Step 2: Given N_a , how many in study region, N

If transects placed at random: $\hat{N} = \frac{N_a}{a/A}$

- Overall: $\hat{N} = \frac{n}{a/A} = \frac{nA}{a} = \frac{nA}{2wL}$ for strip transects

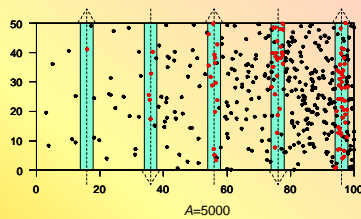
Distance (line transect) sampling

- An extension of plot sampling where not all animals in the covered region are detected

- Here $w = 2$ (strip can be wider, as don't have to see everything)
- $a = 1000$
- $n = 68$ (more animals seen)

- Let P_a = proportion of animals detected within covered region

- Imagine we know (or can estimate) $\hat{P}_a = 0.7$



Intuitive estimator of abundance

- I saw 68 animals
- The estimated proportion seen was 0.7
- So, I estimate the true number of animals in the strips was $68/0.7 = 97.1$
- I covered $1000/5000 = 1/5^{\text{th}}$ of the study region
- So, I estimate there are $97.1/(1/5) = 485.7$ animals

$$\hat{N} = \frac{n/\hat{P}_a}{a/A} = \frac{nA}{a\hat{P}_a} = \frac{68 \times 5000}{1000 \times 0.7} = 485.7$$

Concept – Distance sampling

- Step 1: How many in covered region, N_a ?

Distance sampling: $\hat{N}_a = n/\hat{P}_a$

- Step 2: Given N_a , how many in study region, N

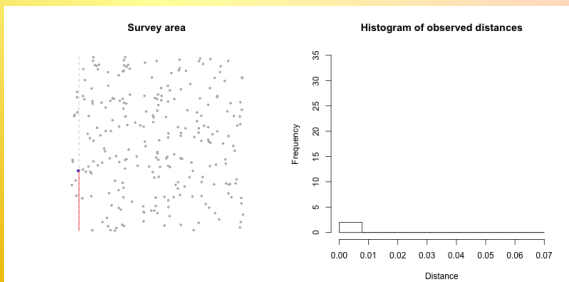
If transects placed at random: $\hat{N} = \frac{\hat{N}_a}{a/A}$

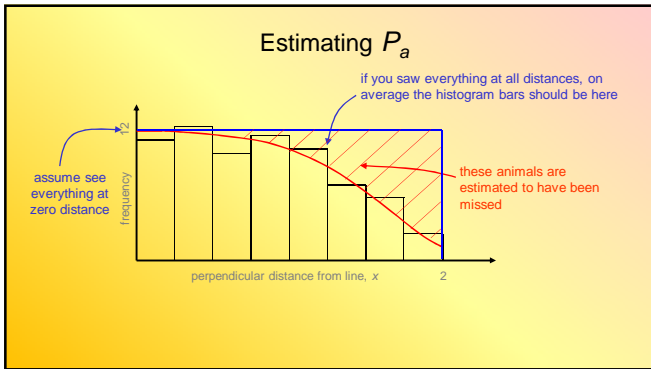
- Overall: $\hat{N} = \frac{n/\hat{P}_a}{a/A} = \frac{nA}{a\hat{P}_a} = \frac{nA}{2wL\hat{P}_a}$ ← for line transects

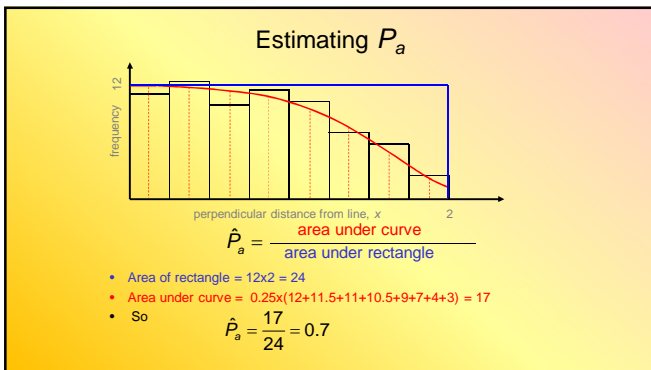
- So how do we estimate P_a ?

Estimating P_a

Record perpendicular distance, x , from transect line to each observed object

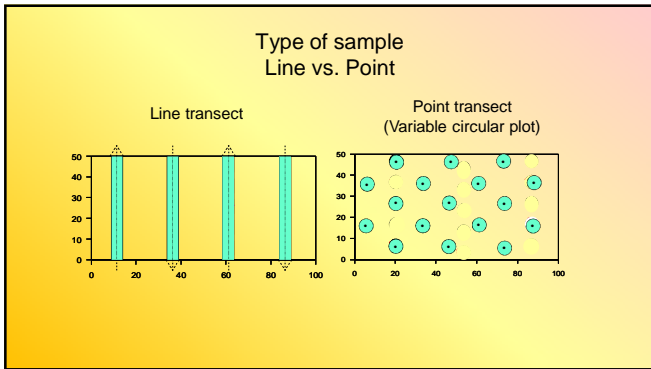


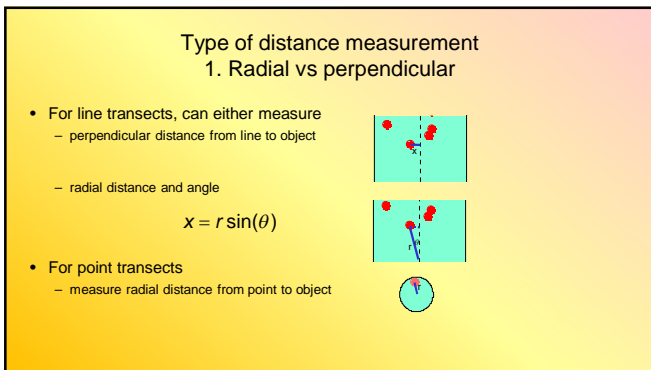


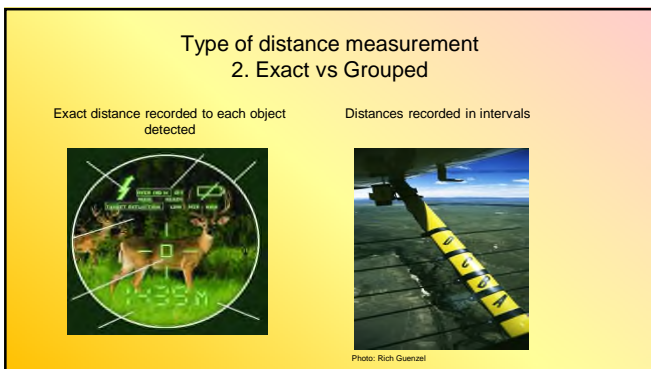


Types of distance sampling

(not exhaustive!)







Type of object

1. Individuals vs Clusters




Photo: Ron Marlow

Each object detected is a single individual

Each object detected is a cluster of individuals
- will need to estimate expected cluster size





Photo: Thomas Norris


Type of Object

2. Direct vs Indirect




Objects are animals (or plants) of interest ...

... or something they produce (an "indirect survey")

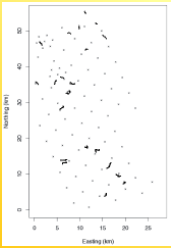


Another example is a cue count



Method of detection

Active vs Passive



84 hydrophones on sea floor of Atlantic Laysania Test and Evaluation Center in Bahamas. From Marques et al. (2009).

Observers actively search for animals and record distances




Photo: Ullas Karanth

Animals are trapped and generate their own distances ("passive distance sampling")




Photo: Steve Dawson

Recap of main ideas so far

- Distance sampling is an extension of plot sampling
 - In plot sampling, we see everything in the covered region

$$\hat{N} = \frac{n}{g/A} = \frac{nA}{a} = \frac{nA}{2wL} \quad \hat{D} = \frac{\hat{N}}{A} = \frac{n}{2wL}$$

strip transects

- In distance sampling, we do not see everything, and we estimate the proportion detected,

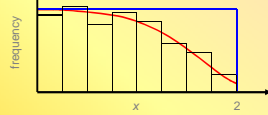
$$\hat{N} = \frac{n/\hat{P}_a}{g/A} = \frac{nA}{a\hat{P}_a} = \frac{nA}{2wL\hat{P}_a} \quad \hat{D} = \frac{\hat{N}}{A} = \frac{n}{2wL\hat{P}_a}$$

line transects

- How do we estimate P_a ?

$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}}$$

line transects



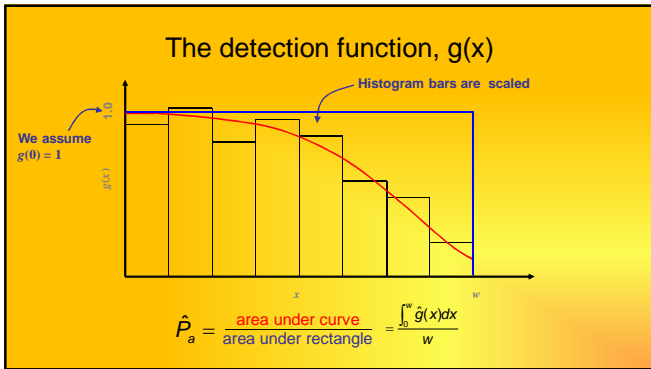
Choosing a Detection function

Overview

- Formal definition
- Criteria for a good detection function model
- Key functions and adjustment terms
- Fitting models in Distance
- Choosing the number of parameters
- Introduction to truncation

Formal definition

- The **detection function** describes the relationship between distance and the probability of detection
- Formally denoted by $g(x)$ (usually referred to as 'g of x')
- **$g(x)$ = the probability of detecting an animal, given that it is at distance x from the line**
- Key to the concept of distance sampling



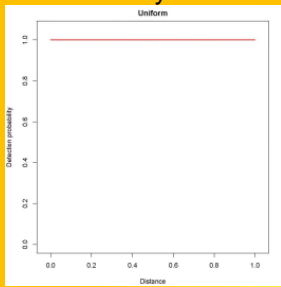
- ### Modelling $g(x)$
- $g(x)$ represents the **underlying** relationship between detection probability and distance
 - However, the true form of $g(x)$ is unknown to us
 - We need to **estimate** $g(x)$ by fitting a **model** to our data
 - i.e. we need to find a curve that will approximate the underlying relationship

- ### Criteria for robust estimation
- Four main criteria for a good model:
 1. **Model robustness** – use a model that will fit a wide variety of plausible shapes for $g(x)$
 2. **Shape criterion** – use a model with a ‘shoulder’ – i.e. $g'(0)=0$
 3. **Pooling robustness** – use a model for the average detection function, even when many factors affect detectability
 4. **Estimator efficiency** – use a model that will lead to a precise estimator of density

Key functions

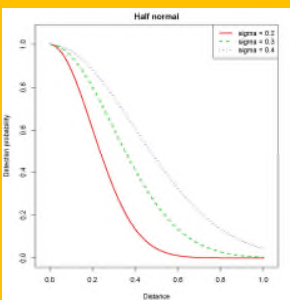
- The first step in constructing a model for $g(x)$ is to choose a **key function**
- This determines the basic model shape
- Four key functions available in Distance:
 1. Uniform
 2. Half normal
 3. Hazard rate
 4. Negative exponential

Key functions (cont.)



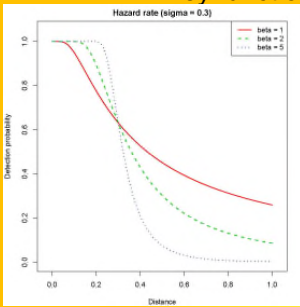
- Model formula:
 $g(x) = 1, x \leq w$
- Parameters = 0
- Shape criterion?
Yes
- Model robust?
No

Key functions (cont.)



- Model formula:
 $g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right), x \leq w$
- Parameters = 1
- Shape criterion?
Yes
- Model robust?
No

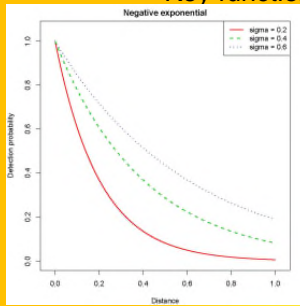
Key functions (cont.)



- Model formula:

$$g(x) = 1 - \exp\left[-\left(\frac{x}{\sigma}\right)^\beta\right], x \leq w$$
- Parameters = 2
- Shape criterion?
Yes
- Model robust?
Yes

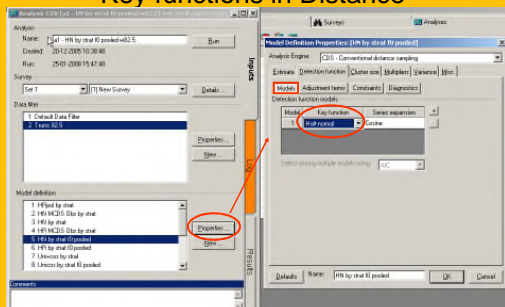
Key functions (cont.)



- Model formula:

$$g(x) = \exp\left(-\frac{x}{\sigma}\right), x \leq w$$
- Parameters = 1
- Shape criterion?
No
- Model robust?
No

Key functions in Distance



Adjustment terms

- Models can be made more robust by adding a series of **adjustment terms** to the key function
- Key function $\times (1 + \text{Series})$
- Series = $\alpha_1 \times \text{term}_1 + \alpha_2 \times \text{term}_2 + \dots$ etc.
- The α_i parameters must be estimated
- Resulting curve model is scaled so that $g(0)=1$
- The number of adjustment terms needs to be chosen

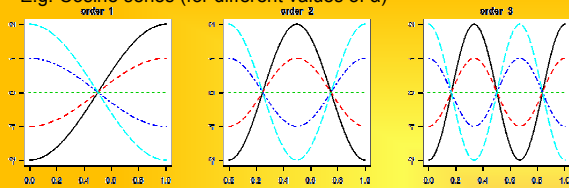
Adjustment terms

- Distance allows the selection of three types of series (one type per model)

Key function	Series adjustment
Uniform*	Cosine*
Half normal†	Hermite polynomial†
Hazard rate	Simple polynomial
Negative exponential	

How adjustment terms work

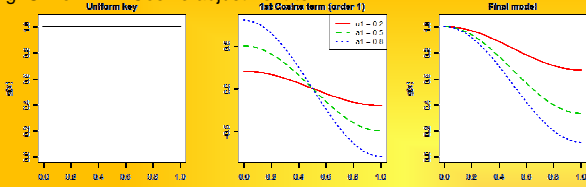
- E.g. Cosine series (for different values of α)



- (1st order only used for uniform)

How adjustment terms work

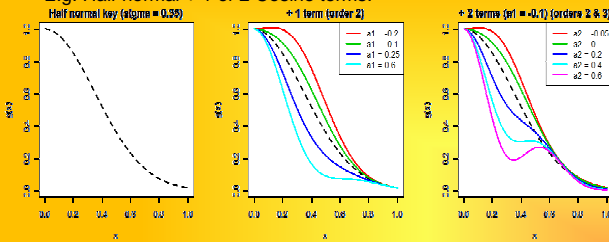
- E.g. Uniform + 1 Cosine adjustment term:



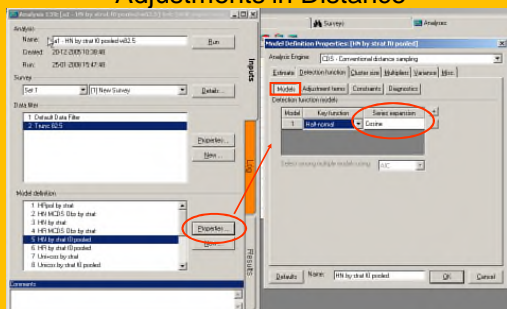
- The effect of the adjustment terms depends on the value of their parameters

How adjustment terms work

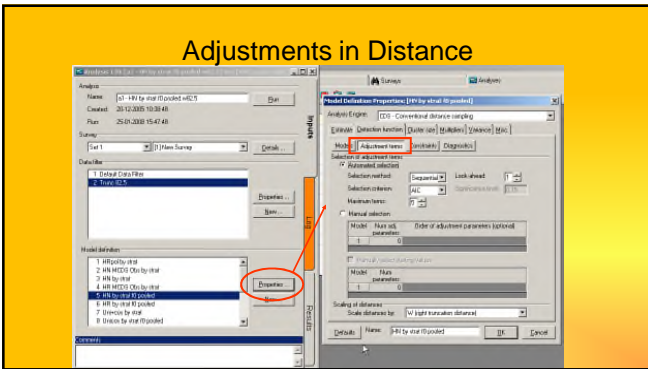
- E.g. Half normal + 1 or 2 Cosine terms:



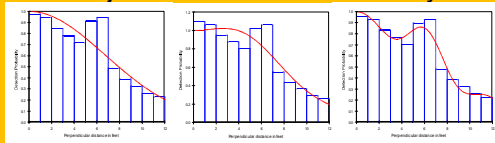
Adjustments in Distance



Adjustments in Distance



Adjustment terms – how many?



Half normal	Half normal	Half normal
0 adjustment terms	1 adjustment term	5 adjustment terms
1 parameter	2 parameters	6 parameters
$\hat{P}_a = 0.65$	$\hat{P}_a = 0.72$	$\hat{P}_a = 0.63$
$CV(\hat{P}_a) = 5.8\%$	$CV(\hat{P}_a) = 11.6\%$	$CV(\hat{P}_a) = 19.9\%$

Note: There is a monotonicity constraint in Distance that is switched on by default to prevent detection functions from increasing. The constraint had to be turned off to produce the third plot. The third plot is for demonstration only – it would not be a good detection function to choose (unless there was a biological reason why detection probability would increase at those distances).

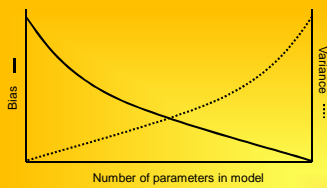
How many parameters?

- Models with too few parameters will not be flexible enough to describe the underlying relationship
- Adding parameters will improve the fit
- But models with too many parameters will be too flexible and will also describe the random noise in the data
- We generally require models with an intermediate number of parameters

How many parameters?

- This problem can also be expressed as a trade-off between bias and variance
- Models with too few parameters tend to produce estimates with low variance and high bias
- Models with too many parameters tend to produce estimates with low bias and high variance (note the increasing CV for the estimate of P_a on the previous slide)

How many parameters?



- Need an objective way of choosing the 'best' model...

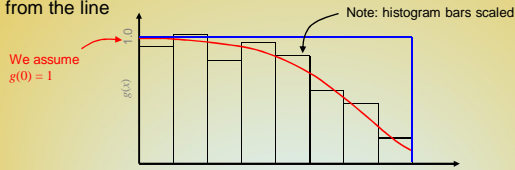
Truncation

- Need to choose the value of w (right truncation)
- Large distances contribute little to estimating the shape of $g(x)$ at small distances (i.e. the shoulder) and may lead to poor fit and high variance
- Typically we might truncate around 5% of observation for line transects (perhaps nearer 10% for point transects)
- Can truncate in the field or at the analysis stage

Three more ways to think about line transects

1. The detection function, $g(x)$

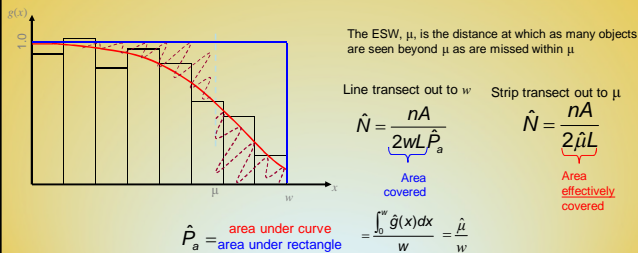
- $g(x)$ = probability of detecting an animal, given that it is at distance x from the line



$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{\int_0^w \hat{g}(x) dx}{1 \times w}$$

2. Effective strip (half) width, $\hat{\mu}$

- Instead of a line transect out to w , where proportion P_a objects are seen, think of a strip transect out to some distance $\hat{\mu}$.



The ESW, $\hat{\mu}$, is the distance at which as many objects are seen beyond $\hat{\mu}$ as are missed within $\hat{\mu}$.

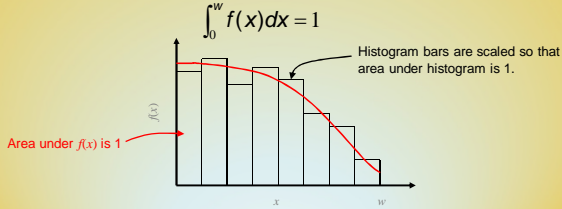
$$\hat{N} = \frac{nA}{2wL\hat{P}_a} \quad \hat{N} = \frac{nA}{2\hat{\mu}L}$$

Area covered Area effectively covered

$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{\int_0^w \hat{g}(x) dx}{w} = \frac{\hat{\mu}}{w}$$

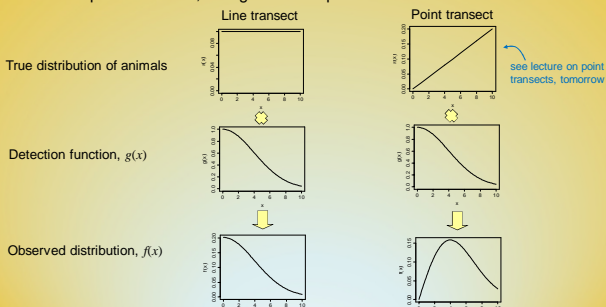
3. The probability density function, $f(x)$

- $f(x)dx$ = probability of observing an animal between distance x and $x+dx$, given it was observed somewhere in $(0,w)$
- $f(x)$ is called the probability density function (pdf) of the observed distances
- Because observations are between 0 and w , the area under $f(x)$ is 1.0



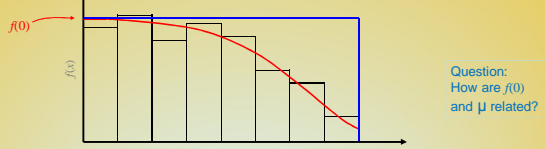
Why is $f(x)$ useful?

1. Useful for point transects, as it gives the expected distribution of detection distances



Why is $f(x)$ useful?

2. Gives another way to estimate P_a
 - Lots of statistical machinery to fit pdfs, so this is the way Distance does it.



$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{1}{\hat{f}(0)w} \quad \hat{N} = \frac{nA}{2wL\hat{P}_a} = \frac{nA}{2wL\left(\frac{1}{\hat{f}(0)w}\right)} = \frac{nA\hat{f}(0)}{2L}$$

Formulae – line transects

Three ways to think about line transects

1. Proportion seen or average probability of detection in covered region, P_a

$$\hat{N} = \frac{nA}{2wL\hat{P}_a} \quad \hat{D} = \frac{n}{2wL\hat{P}_a}$$

2. Effective strip (half-)width, ESW, μ . $P_a = \mu/w$

$$\hat{N} = \frac{nA}{2\hat{\mu}L} \quad \hat{D} = \frac{n}{2\hat{\mu}L}$$

3. Pdf of observed distances, $f(x)$, evaluated at 0 distance $f(0) = 1/\mu$

$$\hat{N} = \frac{n\hat{f}(0)A}{2L} \quad \hat{D} = \frac{n\hat{f}(0)}{2L}$$

Which method when?

- Strip transects
 - Populations that occur in large, loose clusters (e.g. walruses)
 - Stationary objects, at high density, and easily detected
- Line transects
 - Sparsely distributed populations for which sampling needs to be efficient (e.g. whales, deer)
 - Populations that occur in well-defined clusters, and at low or medium cluster density (e.g. dolphin or fish schools)
 - Populations that are detected through a flushing response (e.g. grouse, hares)
- Point transects
 - Populations at high density, especially if surveys are multi-species (e.g. songbirds)
 - Populations that occur in patchy habitat
 - Populations that occur in difficult terrain, or on land where access to walk predetermined lines is problematic (e.g. bird populations in rain forest or on arable farmland)

Notation – line transects

• Known constants and data:

k = number of lines

l_j = length of j^{th} line, $j=1, \dots, k$

$L = \sum l_j$ = total line length

n = number of animals or clusters detected

x_i = distance of i^{th} detected animal or cluster from the line, $i=1, \dots, n$

w = truncation distance for x

A = size of region of interest

a = area of "covered" region = $2wL$

s_i = size of i^{th} detected cluster, $i=1, \dots, n$

Notation – line transects

- Parameters and functions:

N = population size / abundance of animals

N_s = abundance of clusters

D = density = animals per unit area = N/A

D_s = density of clusters

$g(x)$ = detection function

$f(x)$ = probability density function (pdf) of observed distances

$f(0)$ = $f(x)$ evaluated at 0 distance

μ = effective strip (half-)width

P_a = probability of detecting an animal or cluster given it is in the covered area a

$E(s)$ = mean size of clusters in the population

Assessment of model performance

- Likelihood
- AIC
- Absolute measures of model fit

Likelihood

$f(x)$ = probability density function of x

$f(x) dx$ = Pr (animal was between x and $x+dx$ from the line, given it was detected between 0 and w) for small dx

When distances are exact, the **likelihood** is given by

$$L = \prod_{i=1}^n f(x_i) = f(x_1) \times f(x_2) \times \dots \times f(x_n)$$

x_i = distance of i^{th} detected animal from the line.

We fit $f(x)$ by finding the values for the parameters of $f(x)$ (or equivalently $g(x)$) that maximize L (or $\log_e(L)$).

Akaike's Information Criterion

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

L is the maximized likelihood (evaluated at the maximum likelihood estimates of the model parameters)

and q is the number of parameters in the model.

- Models need not be special cases of one another
- Select the model with smallest AIC
- Gives a relative measure of fit

Limitations of AIC

Cannot be used to select between models when:

- sample size n differs
- truncation distance w differs
- data are grouped, and cutpoints differ
- data are grouped in one analysis and ungrouped in the other

Goodness-of-Fit

- Chi-squared test for grouped (interval) data; if data are exact, we must specify interval cutpoints for this test
- Q-Q plots and related tests for exact data

Chi-squared tests

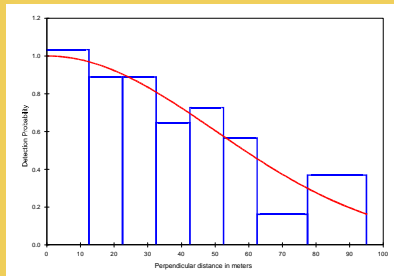
Define u distance intervals, with n_i detections in interval i , $i = 1, \dots, u$.

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

where $n = \sum n_i$
and $\hat{\pi}_i$ is the proportion of the area under the estimated pdf, $\hat{f}(x)$, that lies in interval i .

If the model is 'correct': $\chi^2 \sim \chi_{u-q-1}^2$
 $q =$ no. of parameters

Chaffinch line transect data



χ^2 goodness-of-fit test

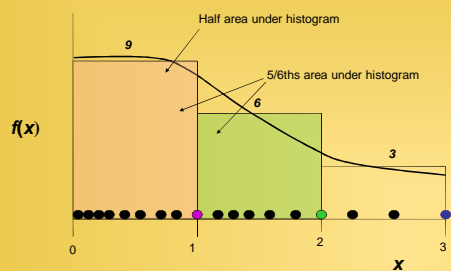
Cell i	Cut Points	Observed Values	Expected Values	Chi-square Values
1	0.000	12.5	16	15.32
2	12.5	22.5	11	11.63
3	22.5	32.5	11	10.62
4	32.5	42.5	8	9.33
5	42.5	52.5	9	7.87
6	52.5	62.5	7	6.37
7	62.5	77.5	3	6.96
8	77.5	95.0	8	4.91

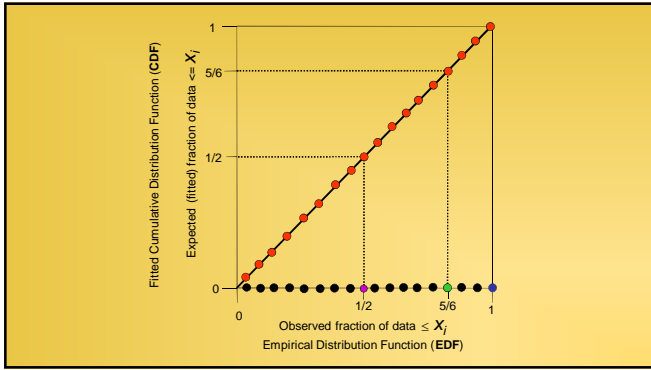
Total Chi-square value = 4.6970 Degrees of Freedom = 6.00

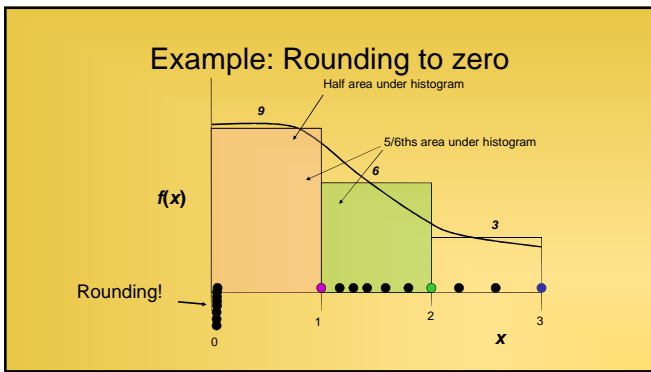
Probability of a greater chi-square value, $P = 0.58322$

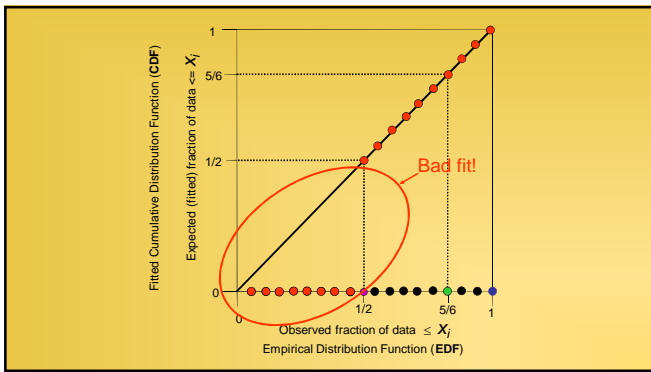
The program has limited capability for pooling. The user should judge the necessity for pooling and if necessary, do pooling by hand.

Q-Q Plots and Related Tests

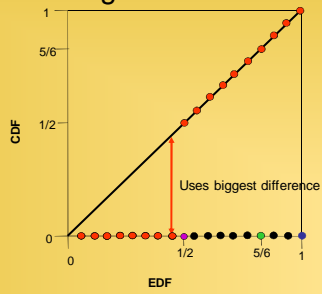




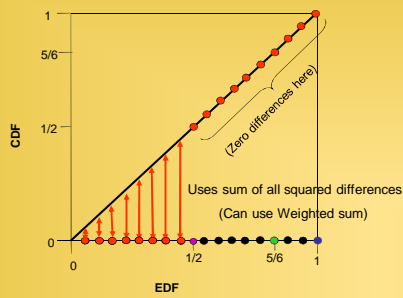




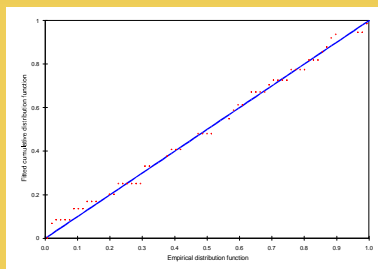
Kolmogorov-Smirnov test



Cramér-von Mises test



Chaffinch line transect Q-Q plot



K-S test and C-von M test

Kolmogorov-Smirnov test

D_n = 0.0573 p = 0.9703

Cramer-von Mises family tests

W-sq (uniform weighting) = 0.0368 0.900 < p <= 1.000

Relevant critical values:

W-sq crit(alpha=0.900) = 0.0000

C-sq (cosine weighting) = 0.0257 0.900 < p <= 1.000

Relevant critical values:

C-sq crit(alpha=0.900) = 0.0000

Q-Q Plot Summary

- Q-Q plots show goodness-of-fit at “high resolution” – without requiring grouping into intervals
- Kolmogorov-Smirnov test and Cramér-von Mises test are goodness-of-fit tests that do not require grouping
- Cramér-von Mises test can be weighted, to give higher weight to x near zero

Making Distance Sampling Work

- Assumptions and effect of violation
- Reliable distance sampling
- Pooling robustness
- Examples of imperfect data
- Analysis hints
- Chapter 2 of Introductory book

Recap of distance sampling

- There are two stages to estimating abundance
 - **Stage 1:** given n , how many objects are in the surveyed/covered region (of size a), N_a
 - Need to estimate P_a (or $f(0)$ or ESW, etc.)

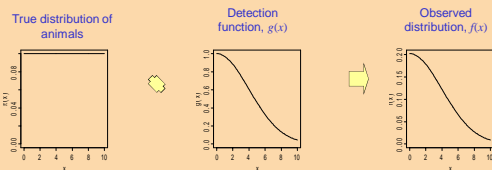
$$\hat{N}_a = \frac{n}{\hat{P}_a}$$
 - **Stage 2:** given \hat{N}_a , how many objects are in study region (of size A), N
 - 'Scale up' from what we see in the survey region to the whole study region

$$\hat{N} = \frac{N_a}{a/A}$$

Assumptions for estimating N_a (stage 1)

1. Animals distributed independently of line or point

- This ensures the true distribution of animals with respect to the line or point is known
- Violated by non-random line/point placement
- Substantial violation can produce substantial bias (e.g. roadside counts)
e.g. for line transects



Assumptions for estimating N_a (stage 1)

2. All animals on the line or point are detected i.e. $g(0)=1$

- It is a critical assumption - violation causes negative bias
- e.g. if $g(0)=0.8$, estimates of N are 80% of true N on average



Images courtesy of FreeDigitalPhotos.net

Assumptions for estimating N_a (stage 1)

3. Observation process is a 'snapshot'

Other ways to phrase this:

- Observers are moving much faster than the animals
- Animals do not move before they can be detected

Problems of independent/non-responsive movement

- An animal moving independently of the observer (compared to moving in response to the observer) produces positive bias; size of bias depends on relative rate of movement of observer and animal, and type of survey.
- Point transect methods in particular need to use 'snapshot' method.

Assumptions for estimating N_a (stage 1)

3. Observation process is a 'snapshot' (continued...)

Problems of responsive movement

- Responsive movement can cause large bias
- It can occur **within** a single line/point or **between** lines/points
- If animals are 'driven' from one line/point to the next ahead of the observer, positive bias will result.
- Note: movement independent of observer outwith 'snapshot' is fine – in this case, the same animal can be detected on multiple lines/transects

Assumptions for estimating N_a (stage 1)

4. Distances are measured accurately

- Random errors cause bias.
 - Bias is generally small for line transect estimators.
 - Can be large for point transect estimators.
 - Both are sensitive to systematic bias and to rounding to 0 distance (or angle).
- Can use grouped data collection.

5. Detections are independent

- Violation has little effect. (Model selection methods for $g(x)$, such as AIC, are somewhat affected)

Assumptions for estimating N given N_a (stage 2)

1. Lines or points are located according to a survey design with appropriate randomization

- We use properties of the survey design to extrapolate from the surveyed/covered region to the study region ('design-based')
- Non-random survey design means density in surveyed/covered region may not be representative of density in study region. Also variance may be biased.



Image courtesy of FreeDigitalPhotos.net

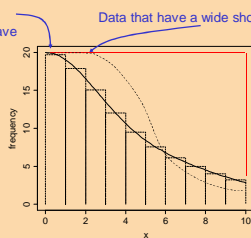
Reliable distance sampling (1)

1. Reliable estimation of P_a (or $f(0)$ or ESW, etc)

- In addition to the assumptions, we would like:

SHAPE CRITERION

Detection function should have a 'shoulder' (i.e. $g'(0)=0$)

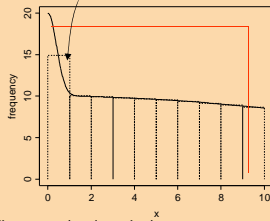


Data that have a wide shoulder are preferable

A wide shoulder makes it easier to estimate area under rectangle (or $f(0)$, etc)

(1) Reliable estimation of P_a

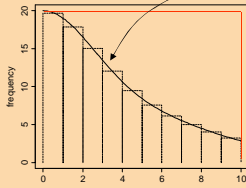
Good field methods will avoid a 'spike' like this



Avoid a) rounding distances (and angles) to zero,
b) 'guarding the trackline'

(1) Reliable estimation of P_a (cont.)

Flexible detection function model can fit the data (see later)



Sample size of observations (~60-80)
- less for detection functions with 'easy' shapes
- more for point transects and 'difficult shapes'.

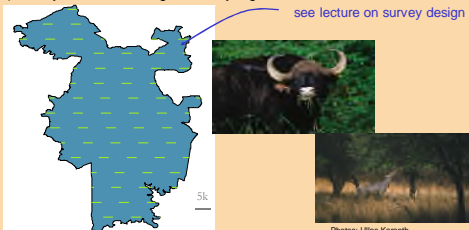
Reliable distance sampling (2)

2. Reliable estimation of N from N_a

- In addition to the assumption of randomized design, we would like a 'large' sample of lines or points (20 or more), evenly distributed through the study region

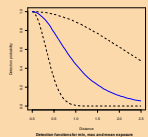
[see lecture on survey design](#)

e.g. surveys of tiger prey in India

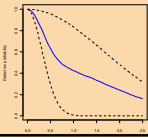


Pooling robustness

- Individuals can have quite different detection functions, but this produces little bias (up to a point!)
- **'Pooling robustness' = robust to pooling of multiple detection functions**
- e.g. Simulation study (in progress!) Truth = 1000 animals

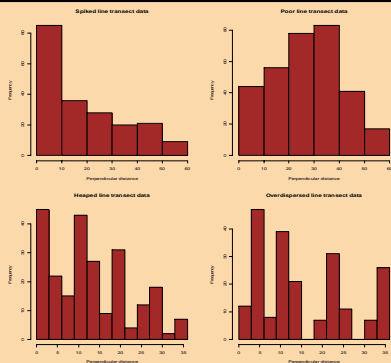


Scenario 1: animals have a gamma distribution of detection functions between min and max shown.
 Mean estimate from simulation: 984 animals (SE 2.3). Bias -1.6%



Scenario 2: half of animals have max detection function, half have minimum.
 Mean estimate from simulation: 976 animals (SE 2.7). Bias -2.4%

Non-ideal data



Analysis hints

- See Section 2.5 of introductory distance sampling book
- It is not a cookbook!
 - Do not simply use the programme defaults in Distance!



The art of model selection

Analysis hints (1)

Stage 1: Exploratory data analysis

- Goal is to understand patterns in distance data, and make preliminary decisions about analysis
- It is never too early to start looking at the data (can then rectify problems)
- Exact data: examine QQ-plots and histograms with lots of cutpoints (in Distance, use Model Definition | Detection Function | Diagnostics)
- Carry out preliminary analysis with a simple model (e.g. half normal, no adjustments). Examine histograms to assess if assumptions are violated
- Make preliminary decisions about truncation and whether to group exact data (Use Data Filter | Intervals)
- For clustered populations, look for evidence of size bias (see Clustered Populations lecture).

Analysis hints (2)

Stage 2: Model selection

- Decide whether to analyse the data as grouped or ungrouped
- Select appropriate truncation distance.
- Choose cutpoints if using grouped data.
- Select and fit a small number of key/adjustment combinations
- Check histograms, goodness-of-fit, AIC and summary tables and choose a model
- This is an iterative process – more exploratory work may be required.
- Check evidence of size-bias if population is in clusters

Analysis hints (3)

Stage 3: Final analysis and inference

- Select best model, or
- Perhaps use model averaging - bootstrap with more than one model selected if model choice is uncertain and influential
- Extract summary analyses and histograms for reporting

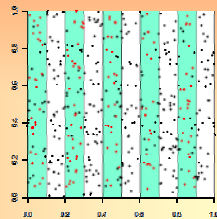
Measures of Precision

Overview

- How to quantify uncertainty
- Why variance is important
- Components of variation in Distance sampling
- Controlling variance
- Estimating variance
- Confidence Intervals
- References: Section 3.6 of introductory book (also see paper by Fewster et al. 2009)

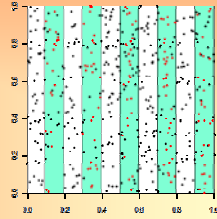
How do estimates behave?

- Consider an artificial population
 - $D = 500$ per unit² (no density gradient)
 - Design: 5 transects equally-spaced ($w=0.05$)
- Results:
 - $n = 140$
 - $\hat{f}(0) = 34.6$
 - $\hat{D} = 484.4$



How do estimates behave?

- Consider a duplicate survey
 - Same population model
 - Same survey design (with a new random start point)



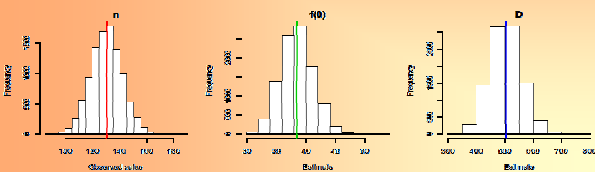
- Results:
 - $n = 139$
 - $\hat{f}(0) = 37.6$
 - $\hat{D} = 522.1$

How do estimates behave?

- Imagine repeating this process over and over, using the same survey design and a population drawn from the same density model
- Each survey will yield:
 - A different value for n
 - A different value for $\hat{f}(0)$
 - A different value for \hat{D}

How do estimates behave?

- What happens if we repeat this simulated survey 10,000 times?
- We end up with **distributions** for n , $\hat{f}(0)$ and \hat{D}



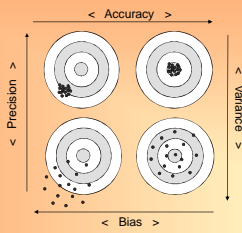
How do estimates behave?

- We are interested in the **hypothetical long-run** behaviour of our estimator

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

- How variable are the estimates?
 - E.g. what is the variance of the distribution for \hat{D} ?
- What is the average value of the estimates?
 - E.g. is the distribution for \hat{D} centred on the truth?

Bias vs. Variance



Low precision = high variance = high uncertainty

Quantifying uncertainty

- Different ways of measuring uncertainty:
 - Variance** = the average squared difference from the mean (the inverse of precision)

- If the estimator for D is unbiased, then

$$\text{Var}[\hat{D}] = E[(\hat{D} - D)^2]$$

- Standard error** = the standard deviation of an estimator (i.e. the square root of estimator variance)

$$\text{Se}[\hat{D}] = \sqrt{\text{Var}[\hat{D}]}$$

Quantifying uncertainty

3. **Coefficient of Variation (CV)** = the standard error divided by the mean (i.e. a standardised version of the standard error)

$$CV[\hat{D}] = \frac{Se[\hat{D}]}{E[\hat{D}]}$$

- Useful for comparing variances when the scale and/or the units of measurement differ
- E.g. consider two variables: X has mean = 100 and variance = 400, Y has mean = 1 and variance = 0.04

$$CV[X] = \frac{\sqrt{400}}{100} = \frac{20}{100} = 0.2 = 20\% \quad CV[Y] = \frac{\sqrt{0.04}}{1} = \frac{20}{100} = 0.2 = 20\%$$

Quantifying uncertainty

4. **Confidence Interval (CI)** = a range of plausible values for the truth

- Calculations are based on variance
- Different ways to calculate CIs, depending on the data, e.g.
 - Normal
 - Lognormal (available in Distance)
 - Bootstrap (available in Distance)
- More about CIs later...

Why is variance important?

- In a real survey, we use an estimator and the survey data to produce a single estimate for D
- If the estimator variance is low, then individual estimates are more likely to be close to the truth (assuming low bias)
- If estimator variance is high, then individual estimates are more likely to be far from the truth
- **For reliable results, we want estimators with LOW variance (and low bias!)**

Variance by components

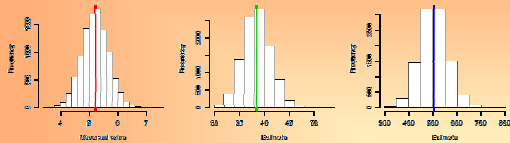
- We can break down the familiar distance sampling density estimator (for line transects with no clusters) into three components:

$$\hat{D} = \frac{n\hat{f}(0)}{2L} = \frac{1}{2} \times \frac{n}{L} \times \hat{f}(0)$$

Constant
(no variance)
Encounter rate
Detection
function

Variance by components

- We can calculate variance measures separately for each component



	n/L	$\hat{f}(0)$	\hat{D}
Mean	26.1	38.5	500.6
Se	2.27	2.71	56.34
CV	8.69 %	7.04 %	11.26 %

Variance by components

- The variance of \hat{D} is affected by the variance of its components
- If the variance of n is high, then the variance of n/L will be high and the variance of \hat{D} will be high
- Similarly, if the variance of $\hat{f}(0)$ is high then the variance of \hat{D} will be high
- So for reliable estimates, we want $Var[n/L]$ and $Var[\hat{f}(0)]$ to be low

Variance by components

- Distance provides several variance measures for each component

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval	
f(0)	1.5726	0.19139	12.17	1.2304	2.0102
p	0.42391	0.51589E-01	12.17	0.33165	0.54185
ESW	0.63587	0.77383E-01	12.17	0.49747	0.81277
n/L	0.80579E-01	0.25990E-01	32.25	0.40601E-01	0.15992
DS	0.63361E-01	0.21843E-01	34.47	0.31088E-01	0.12914
E(S)	2.0143	0.20292	10.07	1.6433	2.4692
D	0.12763	0.45838E-01	35.92	0.61445E-01	0.26510
N	10849.	3896.4	35.92	5223.0	22534.

Encounter rate variance

- The **encounter rate** = n/L = the number of detections per unit of distance
- The variance of n/L is related to the variance of n , and therefore to the variances of counts for individual transects

$$\text{Var}[n] = \text{Var}[n_1] + \dots + \text{Var}[n_k] \leftarrow \begin{array}{l} \text{assumes} \\ \text{independence} \end{array}$$

- Therefore, if counts from individual transects are highly variable the variance of n/L will also be high

Controlling variance

- We can use this knowledge of encounter rate variance to help design good surveys
- Three main ways we can reduce encounter rate variance:
 - Use systematic survey designs
 - Run transects parallel to density gradients
 - Use designs with several transects

Estimating variance – Analytic

- We can describe the relationship between the variance of \hat{D} and the variance of its components more formally using a useful approximation known as the **Delta method**

$$\{cv(\hat{D})\}^2 = \left\{cv\left(\frac{n}{L}\right)\right\}^2 + \{cv(\hat{f}(0))\}^2$$

- Rule: when two or more components are multiplied together, **squared CVs add**

Estimating variance – Analytic

- To estimate $var(n/L)$ we need to use data from the individual lines (or points)
- A **minimum of 20 replicate lines (or points)** is recommended for obtaining a reliable estimate of encounter rate variance
- The (new improved) formula used in Distance:

$$\left\{cv\left(\frac{n}{L}\right)\right\}^2 = \frac{k}{n^2(k-1)} \sum_{i=1}^k \ell_i^2 \left(\frac{n_i}{\ell_i} - \frac{n}{L}\right)^2$$

k = number of lines
 ℓ_i = effort for line i
 n_i = count for line i

Estimating variance – Analytic

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval
f(0)	1.5726	0.19139	12.17	1.2304 2.0102
p	0.42391	0.51589E-01	12.17	0.33165 0.54185
ESW	0.63587	0.77383E-01	12.17	0.49747 0.81277
n/L	0.80579E-01	0.25990E-01	32.25	0.40601E-01 0.15992
DS	0.63361E-01	0.21843E-01	34.47	0.31088E-01 0.12914
R(S)	2.0143	0.20292	10.07	1.6433 2.4692
D	0.12763	0.45838E-01	35.92	0.61445E-01 0.26510
N	10849.	3896.4	35.92	5223.0 22534.

Measurement Units
Density: Numbers/Sq. nautical mi
ESW: nautical miles

Component Percentages of Var(D)
Detection probability : 11.5
Encounter rate : 80.7
Cluster size : 7.9

Similarly, N and D always have the same CV

p and ESW are derived from 1/f(0), so these three share the same CV

Estimating variance – Analytic

- To find the **relative contributions** of each component we take the ratio of squared CVs

- E.g. $100\% \times \frac{\{cv(\hat{f}(0))\}^2}{\{cv(\hat{D})\}^2} =$ The percentage relative contribution made by $f(0)$

Component	Typical values	
	Line	Point
Encounter rate	70-80%	40-50%
Detection function	<30%	>50%

Estimating variance – Analytic

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval
$\hat{f}(0)$	1.5726	0.19139	35.92	1.2304 2.0102
p	0.42391	0.51589E-01	12.17	0.33165 0.54185
ESW	0.63589	0.77383E-01	12.17	0.49747 0.81277
m/L	0.89379E-01	0.21999E-01	32.25	0.49601E-01 0.15392
DS	0.63361E-01	0.21843E-01	34.47	0.31088E-01 0.12914
$\hat{f}(D)$	2.0143	0.20292	10.07	1.6423 2.4692
D	0.12743	0.45838E-01	35.92	0.61445E-01 0.26510
N	10849.	3896.4	35.92	5223.0 22534.

Measurement Units
Density: Numbers/Sq. nautical mi
ESW: nautical miles

Component Percentages of Var(D)
Detection probability : 11.5% ← $12.17^2 / 35.92^2 \times 100\%$
Encounter rate : 80.3% ← $32.25^2 / 35.92^2 \times 100\%$
Cluster size : 7.9%

Estimating variance – Bootstrap

- Works well if the original sample is **large and representative**
- The distribution of density estimates approximates the true distribution that we would (theoretically) get from duplicate surveys
- The variance of the bootstrap estimates can be used as an estimate of the true variance
- In Distance we **resample the individual transects**

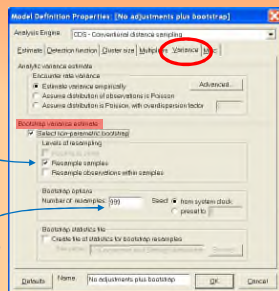
Estimating variance – Bootstrap

- For example, consider a survey with 12 replicate lines
 - Bootstrap sample 1:
 - Transects: 5, 12, 1, 7, 6, 11, 7, 6, 9, 7, 11, 2
 - Density estimate = D_1
 - Bootstrap sample 2:
 - Transects: 3, 4, 9, 1, 12, 7, 8, 11, 1, 3, 2, 12
 - Density estimate = D_2
- Do this B times and use the variance of the B density estimates as an estimate of $\text{var}(\hat{D})$

Estimating variance – Bootstrap

The usual option
(samples = transects)

The number of
bootstrap samples to
use (test on a small
number first to ensure
all is properly set up)



Confidence Intervals

- Confidence intervals (CIs) give us a **range of plausible values** for the truth
- Constructed using data from a single sample
- If we were to carry out multiple surveys and construct 95% CIs from each survey, we would expect 95% of those CIs to contain the true value
- To calculate CIs we need to know the **shape** of the distribution of estimates

Confidence Intervals - Analytic

mean = 1, se = 0.5

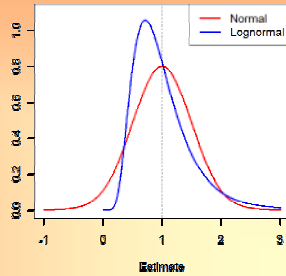
- Two choices:

1. Normal

- symmetrical
- easy to use
- allows negative values

1. Lognormal

- asymmetric (skewed)
- trickier to use
- typically higher interval limits
- does not allow negative values



Confidence Intervals - Analytic

- Distance uses 95% lognormal CIs

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval
F(0)	1.5726	0.19139	12.17	1.2304 - 2.0102
p	0.42391	0.51589E-01	12.17	0.33165 0.54185
ESW	0.63587	0.77383E-01	12.17	0.49747 0.81277
n/L	0.80579E-01	0.25990E-01	32.25	0.40601E-01 0.15992
DS	0.63361E-01	0.21843E-01	34.47	0.31088E-01 0.12914
E(S)	2.0143	0.20292	10.07	1.6433 2.4692
D	0.12763	0.45838E-01	35.92	0.61445E-01 0.26510
N	10849	3896.4	35.92	5223.0 22534.

$$\left(\frac{\hat{D}}{C}, \hat{D} \times C\right) \quad C = \exp\left[1.96\sqrt{\ln\left[1 + (cv(\hat{D}))^2\right]}\right]$$

Confidence Intervals – Bootstrap

- We can use the bootstrap estimates to construct CIs for the true density in two ways:
 - Parametric**
Use the lognormal CI method with the bootstrap estimate of variance instead of the analytic estimate
 - Non-parametric**
Place the bootstrap estimates in order of increasing size and use percentiles as the CI limits (e.g. for a 95% CI using 999 bootstrap estimates, take the 25th estimate as the lower limit and the 975th estimate as the upper limit)

Confidence Intervals – Bootstrap

- Both options are provided in Distance

Pooled Estimates:

	Estimate	%CV	#	df	95% Confidence Interval	
DS	0.20906E-01	38.14	999	19.43	0.96777E-02	0.45162E-01
D	0.39125E-01	42.24	999	23.93	0.12166E-01	0.41557E-01
N	27987.	42.24	999	23.93	0.16933E-01	0.90294E-01
					0.22020E-01	0.81797E-01
					12127.	64589.
					15752.	58511.

Note: Confidence interval 1 uses bootstrap SE and log-normal 95% intervals.
Interval 2 is the 2.5% and 97.5% quantiles of the bootstrap estimates.

Option 1 (parametric)
Option 2 (non-parametric)

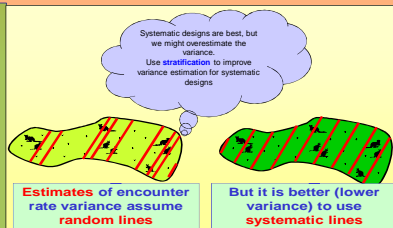
Producing a better estimate of variance when systematic samplers are used

- Fewster, RM, Buckland, ST, Burnham, KP, Borchers, DL, Jupp, PE, Laake, JL, and Thomas, L. 2009. Estimating the encounter rate in distance sampling. Biometrics 65: 225-236.

Systematic samples

Problem:

Systematic designs give the best variance, but the worst variance estimation!

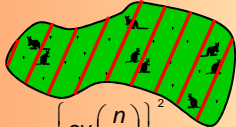


No unbiased estimator exists for estimating variance from a single systematic sample

Systematic samples advice

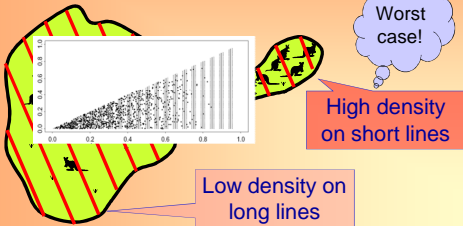
1. Usually, do nothing!

Variance estimation based on random lines will not be perfect, but adequate

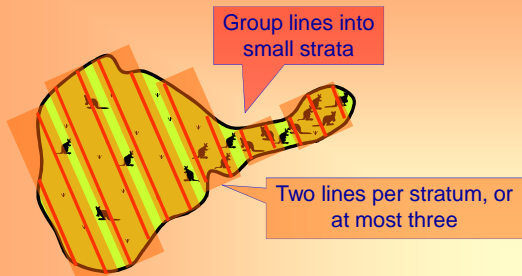


$$\left\{ cv \left(\frac{n}{L} \right) \right\}^2 = \frac{k}{n^2 (k-1)} \sum_{i=1}^k \ell_i^2 \left(\frac{n_i}{\ell_i} - \frac{n}{L} \right)^2$$

If there are **strong trends**, variance might be significantly overestimated



Post-stratification can give much better variance estimates



Post-stratification can give much better estimates of variance

Pool by-stratum variance estimates together, weighted by Total Effort in Stratum

Trends **within** strata are minor; Estimate encounter rate variance separately for each stratum

$$\hat{v}\text{ar}\left(\frac{n}{L}\right) = \frac{1}{L^2} \sum_{h=1}^H L_h^2 \hat{v}\text{ar}_h\left(\frac{n_h}{L_h}\right)$$

In Distance 7:

Click on the "Advanced..." tab

Choose this option

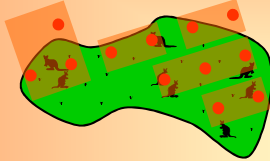
Successive pairs of lines will be grouped together, according to their ID in the sample layer (1 & 2, 3 & 4, etc). (If there are an odd number of lines, the last 3 will be grouped.)

Overlapping strata are even better, as you get a larger sample size of post-strata

Choose this option

Systematic point transect surveys

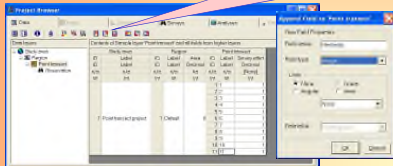
Less of an issue (no problem of different line lengths), but can similarly group into strata of two or three adjacent points for encounter rate variance if required.



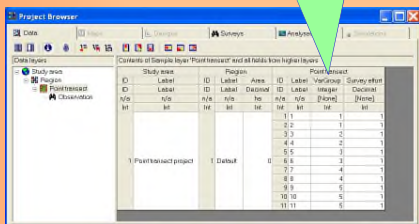
However, it's harder to do in Distance – need to manually post-stratify.

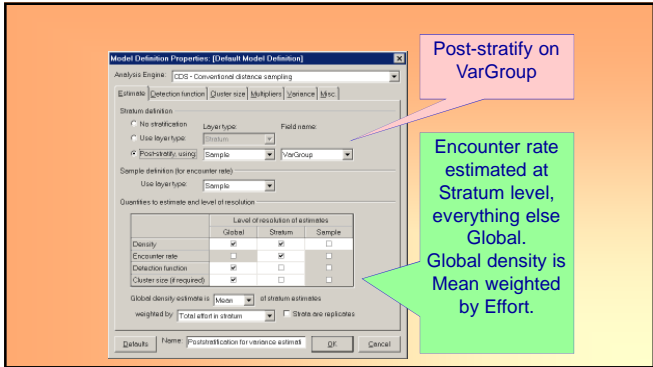
Can only do non-overlapping post-stratification this way.

Add new field VarGroup into the Point transect layer (i.e., the sample layer)



Enter values into VarGroup so that it groups together points 1 and 2, 3 and 4, etc

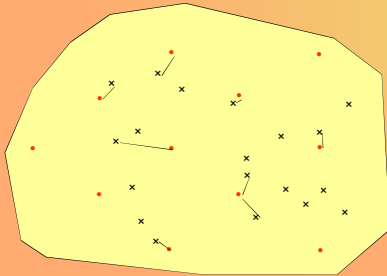




Post-stratify on VarGroup

Encounter rate estimated at Stratum level, everything else Global. Global density is Mean weighted by Effort.

Point transect sampling



Random points or systematic grid of points randomly placed; observer records distance to any detected animals

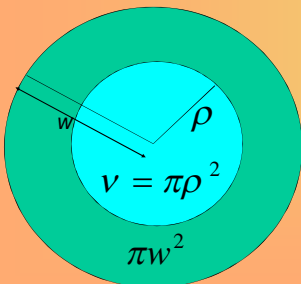
Point transect sampling

For k point counts with certain detection to distance w .

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

How does this change if detection is uncertain?

Effective radius and effective area



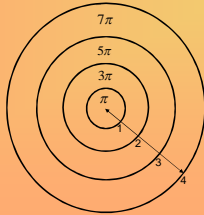
ρ = effective radius
 v = effective area

Covered area: $a = k\pi w^2$

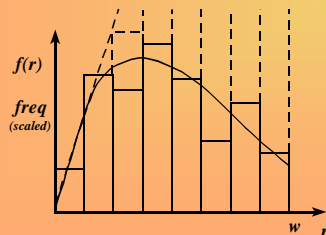
Proportion detected: $P_a = \frac{k\pi\rho^2}{k\pi w^2} = \frac{\rho^2}{w^2}$

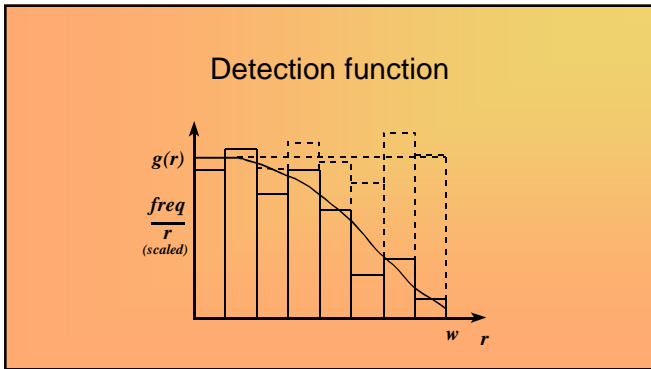
Estimated density: $\hat{D} = \frac{n}{a\hat{P}_a} = \frac{n}{k\pi w^2 \times \hat{\rho}^2 / w^2} = \frac{n}{k\pi\hat{\rho}^2}$

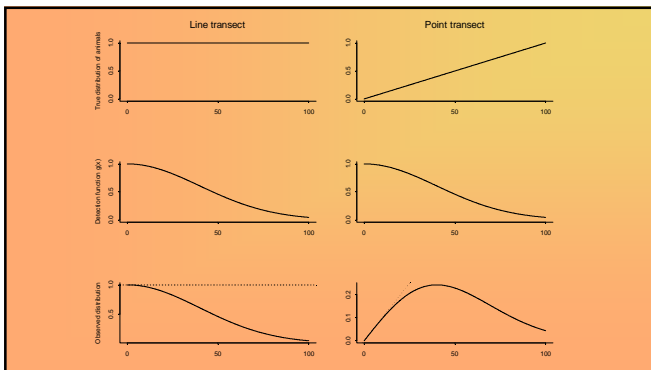
Area and hence number of birds increase linearly with distance:

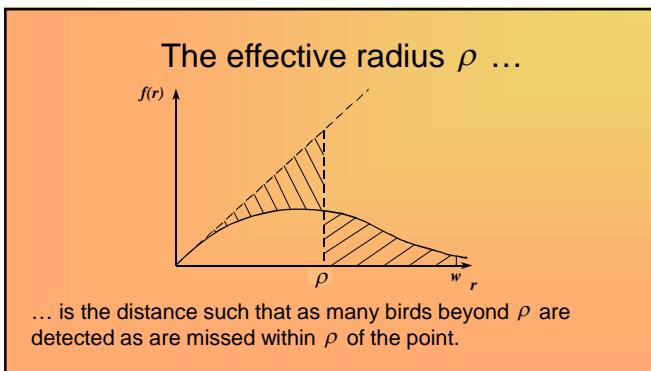


Probability density function









Area under curve:

$$\int_0^w f(r) dr = 1$$

Area of triangle:

$$\frac{\rho \times \rho f'(0)}{2} = \frac{\rho^2 h(0)}{2}$$

Hence $\hat{\rho}^2 = \frac{2}{h(0)}$ and $\hat{v} = \frac{2\pi}{h(0)}$

so that $\hat{D} = \frac{n\hat{h}(0)}{2\pi k}$

Notation: point transects

Known constants and data:

- k = number of points
- n = no. of animals or clusters detected
- r_i = distance of i^{th} detected animal or cluster from the point, $i = 1, \dots, n$
- w = truncation distance for r
- A = size of region of interest
- a = size of covered region = $k\pi w^2$
- s_i = size of i^{th} detected cluster, $i = 1, \dots, n$

Point transect notation (cont)

Functions:

- $g(r)$ = detection function
- $f(r)$ = probability density function (pdf) of detection distances
- $h(r) = f'(r)$ = slope of pdf $f(r)$
- $h(0)$ = slope of pdf evaluated at $r=0$

Point transect notation (cont)

Parameters:

- D = density = animals per unit area
- D_s = density of clusters
- N = population size = $D \cdot A$
- ρ = effective radius = $\sqrt{2/h(0)}$
- v = effective area (per point) = $2\pi / h(0)$
- P_a = prob. of detection of animal or cluster in the covered area a

Comparative study^a

1. Point transect, 5-minute counts (9.8 hrs)
2. Point transect, snapshot method (8.4 hrs)
3. Cue counting, 5 mins per point (10.0 hrs)
4. Line transect sampling (7.9 hrs)
5. Territory mapping

^aBuckland, S.T. 2006. Point-transect surveys for songbirds: robust methodologies. *The Auk* 123:345-357.

Focal species in Montrave study

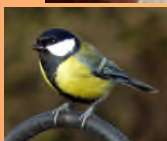
• Chaffinch
Fringilla coelebs



• Robin
Erithacus rubecula



• Great tit
Parus major



• Wren
Troglodytes troglodytes



Study area, Montrave Estate

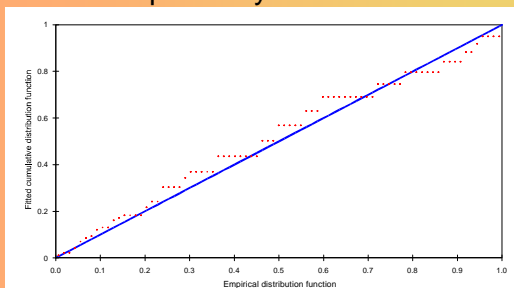


Parkland and mixed woodland
33.2 ha
 $k = 32$ points

The data

	Chaffinch	Great tit	Robin	Wren
M1 ($w=110m$) n :	74	44	57	132
M2 ($w=110m$) n :	63	18	50	117
M3 ($w=92.5m$) n :	627	177	785	765
Cue rate:				
Sample size	33	12	26	43
Mean	7.9	8.2	17.9	7.3
M4 ($w=95m$) n :	73	32	80	155
M5 territories:	25	7	28	43

Example analyses: chaffinch



K-S and C-von M tests

Kolmogorov-Smirnov test

D_n = 0.0978

p = 0.4205

Cramer-von Mises family tests

W-sq (uniform weighting) = 0.1194

0.400 < p <= 0.500

Relevant critical values:

W-sq crit(alpha=0.500) = 0.1188

W-sq crit(alpha=0.400) = 0.1464

C-sq (cosine weighting) = 0.0705

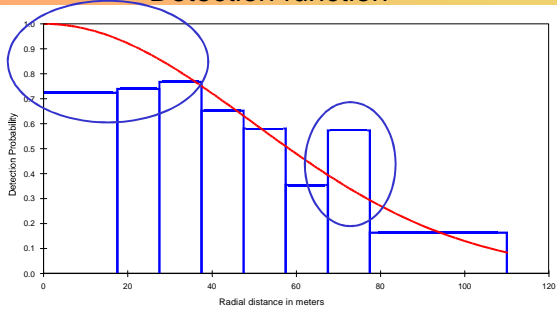
0.500 < p <= 0.600

Relevant critical values:

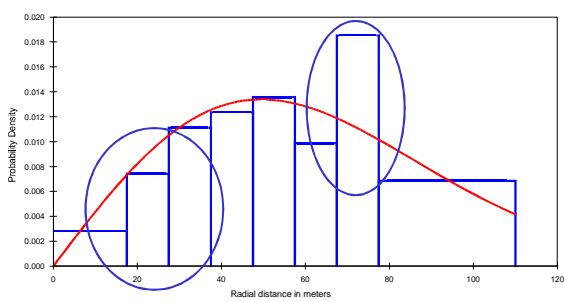
C-sq crit(alpha=0.600) = 0.0623

C-sq crit(alpha=0.500) = 0.0770

Detection function



Probability density function



Chi-square gof test

Cell i	Cut Points	Observed Values	Expected Values	Chi-square Values
1	0.000 - 17.5	4	5.36	0.385
2	17.5 - 27.5	6	7.29	0.229
3	27.5 - 37.5	9	9.42	0.019
4	37.5 - 47.5	10	10.57	0.031
5	47.5 - 57.5	11	10.77	0.005
6	57.5 - 67.5	8	10.15	0.454
7	67.5 - 77.5	15	8.95	4.096
8	77.5 - 110.	18	18.50	0.013

Total Chi-square value = 5.1918 Degrees of Freedom = 6.00

Probability of a greater chi-square value P = 0.51946

The program has limited capability for pooling. The user should judge the necessity for pooling and if necessary, do pooling by hand.

Estimation summary

Effort : 64.00000
 # samples : 32
 Width : 110.0000
 # observations: 81

Model 1
 Half-normal key, $k(y) = \text{Exp}(-y^{**2}/(2*A(1)**2))$

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval	
h(0)	0.44566E-03	0.69514E-04	15.60	0.32734E-03	0.60674E-03
p	0.37089	0.57851E-01	15.60	0.27242	0.50494
EDR	66.991	5.2246	7.80	57.373	78.220
n/K	1.2656	0.12697	10.03	1.0320	1.5522
D	0.89769	0.16648	18.55	0.62355	1.2924
N	38.000	5.5637	18.55	21.000	43.000

Estimation summary (cont.)

Measurement Units

Density: Numbers/hectares
 EDR: meters

Component Percentages of Var(D)

Detection probability : 70.7
 Encounter rate : 29.3

Estimated densities

Method	Chaffinch		Great Tit		European Robin		Winter Wren	
	\hat{d}	95% CL	\hat{d}	95% CL	\hat{d}	95% CL	\hat{d}	95% CL
Conventional point sampling	1.03	0.74-1.43	0.58	0.36-0.94	0.52	0.26-1.06	1.29	0.80-2.11
Snapshot	0.90	0.62-1.29	0.22	0.13-0.39	0.60	0.38-0.94	1.02	0.80-1.32
Cue-count	0.71	0.45-1.23	0.26	0.09-0.76	0.82	0.52-1.31	1.21	0.82-1.79
Line transect	0.64	0.46-0.90	0.26	0.16-0.42	0.69	0.47-1.00	1.07	0.87-1.31
Territory mapping	0.75		0.21		0.84		1.30	

Estimated effective detection radii (meters)

Method	Chaffinch		Great Tit		European Robin		Winter Wren	
	\hat{r}	95% CL	\hat{r}	95% CL	\hat{r}	95% CL	\hat{r}	95% CL
Conventional point sampling	67	58-78	62	51-74	74	52-104	71	57-90
Snapshot	67	57-78	64	54-75	65	54-77	75	69-83
Cue-count	74	70-79	65	58-71	51	47-57	66	63-69
Line transect [†]	59	48-72	63	47-84	60	44-83	75	65-86

[†]effective strip half-width shown for line transect method

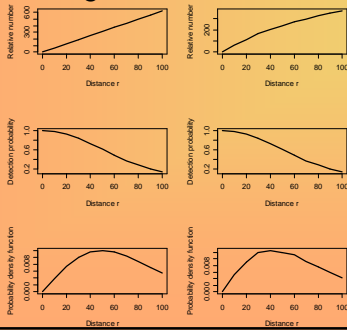
Estimated hours of fieldwork to obtain a 10% CV for estimated density

Method	Common chaffinch	Great tit	European robin	Winter wren
Conventional point sampling	28	60	131	61
Snapshot	29	70	44	14
Cue-count	56	352	57	40
Line transect	22	49	29	11

Simulation study, three investigations

1. All assumptions satisfied:
half-normal model, 1000 replicates
2. Overlapping points:
Point separation 100m, effective detection radius 106m
3. Edge effect (similar to Montrave study area),
no sampling in buffer zone, birds detected outside study
area boundary not recorded

Edge effect simulation



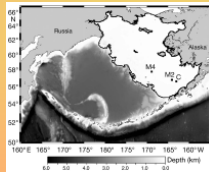
Simulation results – true density = 1

	Popn 1	Popn 2	Popn 3	Popn 3, w=80m
\bar{n}	353	354	41	32
mean	1.0029	1.0056	0.9509	0.9961
sd	0.0706	0.0815	0.1924	0.3160
se(mean)	0.0022	0.0026	0.0061	0.0100
mean(se)	0.0754	0.0750	0.2099	0.3557

Popn 1: all assumptions hold
 Popn 2: overlapping plots
 Popn 3: edge effect

Point transects with marine mammals

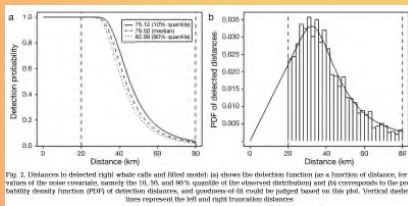
- Seafloor mounted acoustic recording packages deployed and listening for right whale “up-calls”
- Constitutes an example of cue counting
- Analysis incorporated
 - false-positive proportion in call classification,
 - ambient noise as covariate,
 - left truncation because of inexact distance estimation at small distances



Not a recommended allocation of survey effort; proof of concept

Right whale abundance estimates

- Detection probability of 0.29 (CV=2%) from fitted model
- Density estimate of 0.26 whales per 10000km² (CV=29%)
- Abundance in shelf region of Bering Sea: 25 (CI: 13-47)



See Marques, Munger, Thomas, Wiggins and Hildebrand (2011) Estimating North Pacific right whale density using passive acoustic cue counting. *Endangered Species Research* 13:163-172.

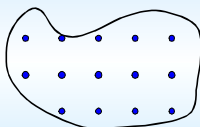
Survey design

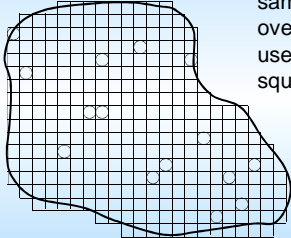
- What are your objectives?
- What precision do you need?
- What resources are required?
- Are sufficient resources available?
- Include training in the costings.
- Cost for statistical advice!!
- Conduct a pilot survey.

Line or point placement

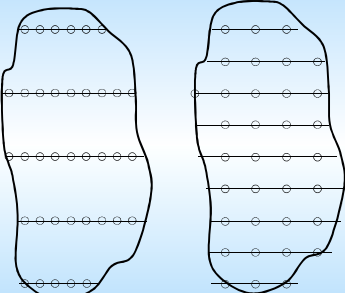
- Use randomly positioned lines or points, or a systematic grid of lines or points, randomly superimposed on the study area
- **Do not** use roads, tracks, etc.
- Stratify the study area if strong differences in habitat or density are apparent
- Aim to orientate lines perpendicular to density contours or to linear features (e.g. woodland edge)
- Many short lines are preferred to a few long lines

Point transect survey design





Simple random sampling without overlapping plots: use a grid of squares of side $2w$



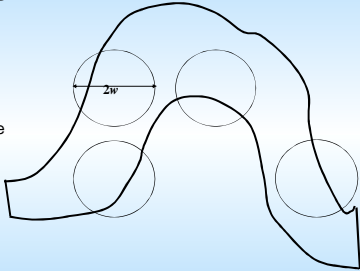
Points along lines:

Left-hand design: the lines should be taken as the sampling units,

Right-hand design: the individual points can be taken as the sampling units

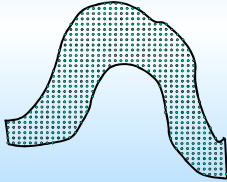
Edge Effects

- A problem if study area is small or narrow relative to w
- Issues
 - Coverage probability close to the edge
 - Animals detected outside the region boundary



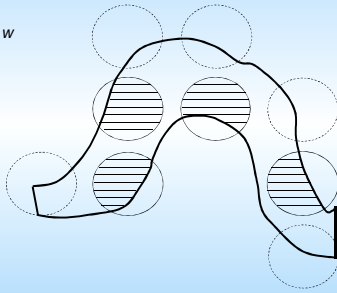
Coverage probability – probability for a given design that a point within the survey region will be sampled

To calculate the probability, repeatedly sample from the grid using a specific design (eg: 10 random points) and see how frequently a given point is included.

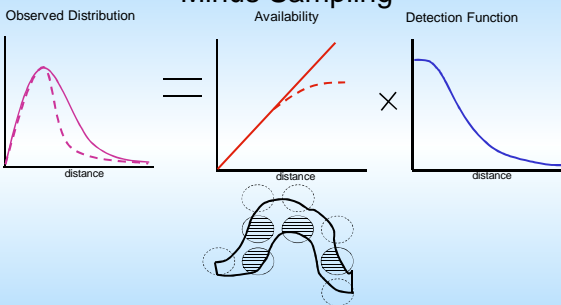


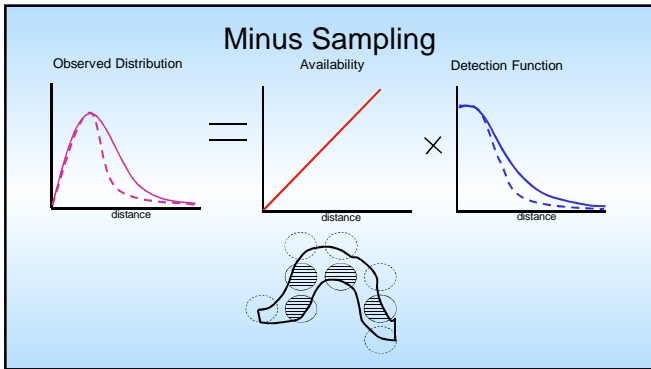
Minus Sampling

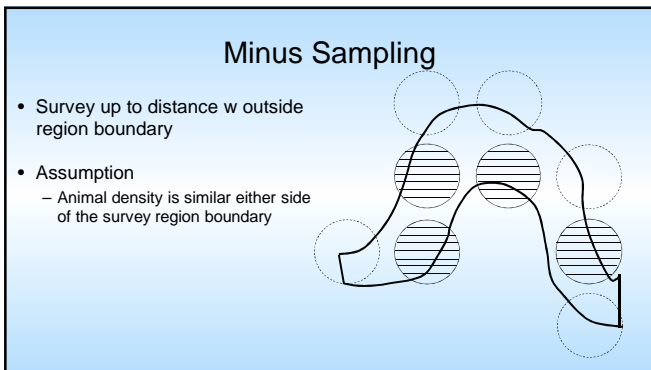
- Coverage probability is lower within w of the edge
- Assumption
 - Animal density within w of the survey region boundary is the same as for $> w$
- For data collection and analysis options, see 6.7 of "Introduction to Distance Sampling"

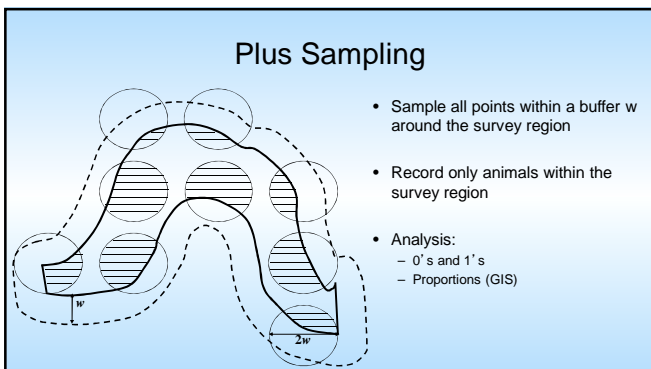


Minus Sampling

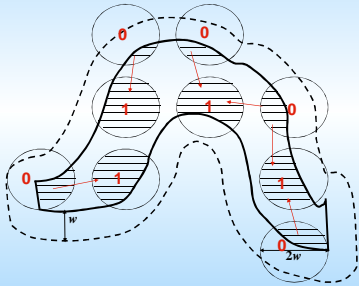






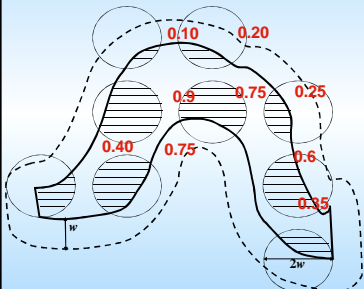


Plus Sampling



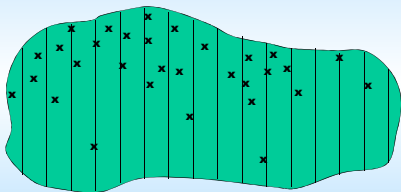
- Sample all points within a buffer w around the survey region
- Record only animals within the survey region
- Analysis:
 - 0's and 1's
 - Proportions (GIS)

Plus Sampling



- Sample all points within a buffer w around the survey region
- Record only animals within the survey region
- Analysis:
 - 0's and 1's
 - Proportions (GIS)

Line transect survey design



Simple random sampling without overlapping strips: use a grid of rectangles width $2w$ and length l

Systematic grid of short lines with adjustment to avoid partial lines at the edge

Surveyed area decreases with distance from transect

Conventional analysis can give valid density estimate.

Coverage probability lower at edge

- See Section 6.7 of Introduction to Distance Sampling, 2001

Extrapolate lines beyond boundary recording only animals within survey region

The diagram shows an irregularly shaped survey region with a solid black boundary. Inside the region, there are several vertical grey lines representing survey transects. These lines are extended beyond the boundary as red lines. Small red triangles at the ends of the extrapolated lines indicate where animals were recorded. The text indicates that only animals within the survey region are recorded, while the lines are extrapolated for analysis.

Use of a buffer zone to eliminate edge effects

The diagram shows a curved boundary and a rectangular survey region. A dashed line represents a buffer zone extending beyond the boundary. Inside the survey region, there are several 'x' marks representing animals. Outside the survey region but within the buffer zone, there are 'o' marks representing animals detected outside the region. The text explains that the line is extended beyond the boundary, but effort and recording are limited to the survey region.

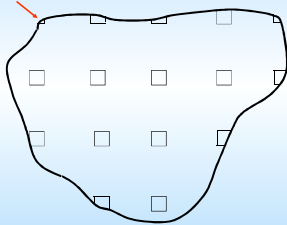
Extend the line beyond the boundary, but don't include the associated effort, and don't record animals detected outside the region (o)

A circuit design

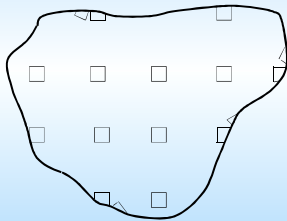
The diagram shows an irregularly shaped survey region with a solid black boundary. Inside the region, there is a grid of small squares representing circuits. Some squares are located near the boundary and are enclosed in dashed lines, indicating they overlap the boundary. A red arrow points from one of these dashed squares to a larger, detailed diagram of a square circuit with a smaller square inside it, representing a circuit design.

If we remove the circuits that overlap the boundary (dashed) we undersample the edge.

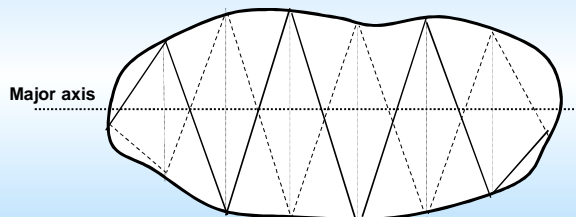
- We can include partial circuits at the edge
- Fine if travelling for a small amount of survey is acceptable



Or delete circuits mostly outside, and reflect and displace remaining edge circuits

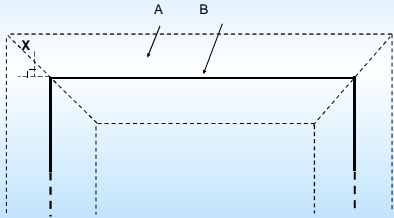


Saw-tooth or zigzag designs



Corners in saw-tooth and circuit designs:

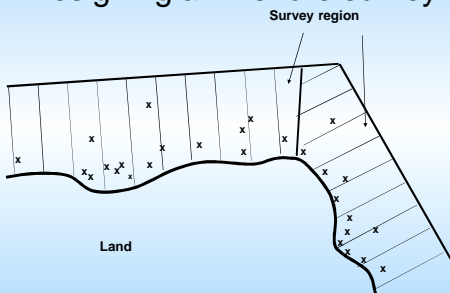
Animal X is detected in trapezium A, so is associated with line segment B



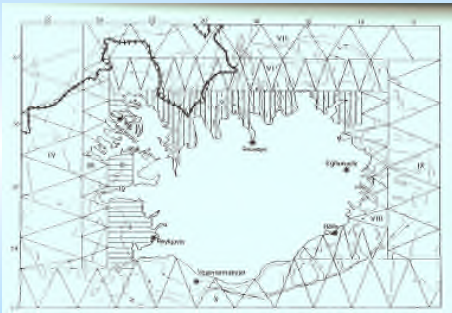
Right-angle corners:



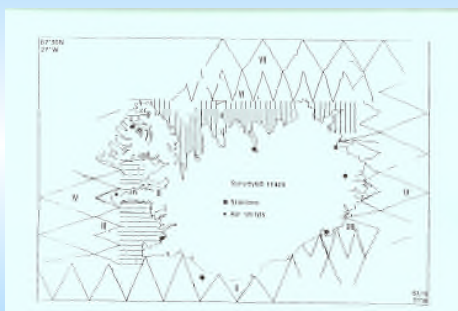
Designing an inshore survey



Iceland – aerial survey design, whale survey



Actual effort, Icelandic whale survey



Stratification (Geographic)

Why stratify?

1. To improve precision.
 - a. Estimate inter-stratum differences rather than have them contribute to variance.
 - b. Reduce overall variance by increasing effort in strata which contribute most to variance.
2. Because want estimates by sub-region/stratum.
3. For logistic reasons

Stratification (Geographic)

What to stratify?

1. Encounter rate: Density often varies spatially.
2. Detection function: May vary spatially. There are often sample size limitations on stratified estimation (too few detections in some strata).
3. Mean cluster size: May vary spatially. There may be sample size limitations on stratified estimation.

NB: If any of the above are estimated by pooling across strata, when in reality they differ between strata, within-stratum estimates are biased.

Stratification (Spatial)



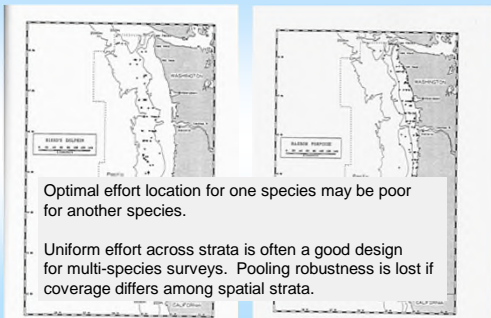
Most animals between 200m and 2000m contours, so put more effort into a shelf-edge stratum?

But:

Sample size too low in other strata?

Other species?

Stratification (Spatial)



Optimal effort location for one species may be poor for another species.

Uniform effort across strata is often a good design for multi-species surveys. Pooling robustness is lost if coverage differs among spatial strata.

Sample size

- Aim for at least 60-80 sightings for fitting the detection function
- and at least 20 lines or points for estimating encounter rate n/L or n/k
- Whether reliable estimates can be obtained from smaller samples depends on the data

Sample size – continued

More observations are required:

- if detection function is spiked
- if population is highly aggregated
- for point transect sampling

Increasing sample size using repeat counts

If a line is sampled three times,

- pool the distance data from the three visits
- enter survey effort as three times the line length.

If a point is sampled three times,

- enter survey effort as 3.

Determining total line length

Pilot study: n_0 animals (or clusters) counted from lines totalling L_0 in length.

Total line length required in main survey is

$$L = \left(\frac{q}{[cv_t(\hat{D})]^2} \right) \times \frac{L_0}{n_0}$$

where $cv_t(\hat{D})$ is the target cv (e.g. 10% is 0.1)

and ...

Determining line length (cont)

q is approximately $\frac{V(n)}{n} + \frac{nV[\hat{f}(0)]}{[\hat{f}(0)]^2}$

Pilot studies are typically too small to estimate q . If past similar data sets are not available, assume $q = 3$.

Line length example

A pilot study yields $n_0 = 20$ observations from lines of total length 5km. We require a CV of 10%, and assume $q = 3$.

$$L = \frac{3}{0.1^2} \times \frac{5}{20} = 75\text{km}$$

Estimated sample size is

$$n = L \times \frac{n_0}{L_0} = 75 \times \frac{20}{5} = 300$$

Determining line length (cont)

If pilot survey is sufficiently large, calculate line length for main survey as

$$L = \frac{L_0 [cv(\hat{D}_0)]^2}{[cv_t(\hat{D})]^2}$$

where

$cv(\hat{D}_0)$ is the cv of estimated density obtained from the pilot survey, and L is total line length in the main survey

Point transects: number of points

$$k = \left(\frac{q}{[cv_t(\hat{D})]^2} \right) \times \frac{k_0}{n_0}$$

or

$$k = \frac{k_0 [cv(\hat{D}_0)]^2}{[cv_t(\hat{D})]^2}$$

where k_0 points in the pilot survey yielded n_0 detections, or estimated density of \hat{D}_0

Automated Survey Design

Aim: Use geographic information system (GIS) within Distance to aid survey design and evaluate properties of different designs

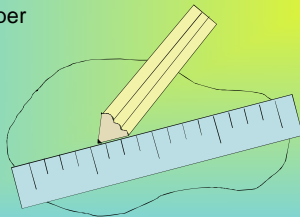
- Based on PhD thesis work by Samantha Strindberg
- Strindberg, S. and Buckland, S. T. 2004. Zigzag survey designs in line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics* **9**:443-461
- Thomas, L., Williams, R., and Sandilands, D. 2007. Designing line transect surveys for complex survey regions. *Journal of Cetacean Research and Management* **9**:1-13
- See Chapter 7 of Advanced book (Design of distance sampling surveys and Geographic Information Systems by Strindberg, Buckland and Thomas)

Contents

- Background and Terminology
- Point Transect Designs
- Line Transect Designs
- Design-based Abundance Estimates
- Survey Design in Distance
- ArcGIS

Background

- Pen and Paper



Automated survey design using GIS technology

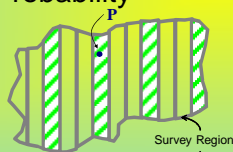
- Easily generate surveys based on randomised designs
- Print out maps or download to GPS
- Evaluate properties of different designs
 - optimise for any situation

Terminology

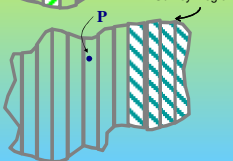
- **Sampler** – a sample unit
 - Strip (line transect)
 - Circle (point transect)
- **Design** – an algorithm for laying out samplers
- **Survey** – a single realisation of a design
- **Sampling strategy** – design & estimator
- **Coverage probability** – probability for a given design that a point within the survey region will be sampled

Example: Coverage Probability

– Uniform coverage probability, $\pi = 1/3$



– Uniform coverage probability, $\pi = 1/3$
– Uneven coverage for any given realisation



Which Design?

- **Uniformity** of coverage probability
- **Even-ness** of coverage within any given realisation
- **Overlap** of samplers
- **Cost** of travel between samplers
- **Efficiency** when density varies within the region
- **Edge** effects

Point Transect Designs

- Simple Random



versus

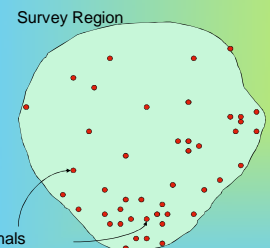
- Systematic Grid



Comparison

- **Uniform coverage** – both have uniform coverage probability
- Systematic has more **even coverage** for any given realisation
- Can have **overlap** of samplers in simple random design
- **Cost** of travel is similar
 - If this is important a cluster sampling design can be used

Density Variation



- Systematic generally more reliable
- If variation in density is predictable
 - Consider stratification
 - Or unequal coverage probability design
- If not predictable
 - Adaptive sampling
- Consider modelling the density surface, (i.e. a model-based estimate)

Example of Stratified Point Transect Design

Survey #0

Survey #1

Survey #2

Survey #3

Survey #4

Survey #5

Survey #6

Survey #7

Survey #8

Survey #9

Survey #10

Survey #11

Survey #12

Survey #13

Survey #14

Survey #15

Survey #16

Survey #17

Survey #18

Survey #19

Survey #20

Survey #21

Survey #22

Survey #23

Survey #24

Survey #25

Survey #26

Survey #27

Survey #28

Survey #29

Survey #30

Survey #31

Survey #32

Survey #33

Survey #34

Survey #35

Survey #36

Survey #37

Survey #38

Survey #39

Survey #40

Survey #41

Survey #42

Survey #43

Survey #44

Survey #45

Survey #46

Survey #47

Survey #48

Survey #49

Survey #50

Survey #51

Survey #52

Survey #53

Survey #54

Survey #55

Survey #56

Survey #57

Survey #58

Survey #59

Survey #60

Survey #61

Survey #62

Survey #63

Survey #64

Survey #65

Survey #66

Survey #67

Survey #68

Survey #69

Survey #70

Survey #71

Survey #72

Survey #73

Survey #74

Survey #75

Survey #76

Survey #77

Survey #78

Survey #79

Survey #80

Survey #81

Survey #82

Survey #83

Survey #84

Survey #85

Survey #86

Survey #87

Survey #88

Survey #89

Survey #90

Survey #91

Survey #92

Survey #93

Survey #94

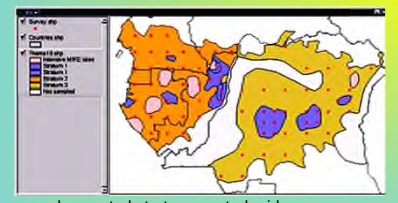
Survey #95

Survey #96

Survey #97

Survey #98

Survey #99



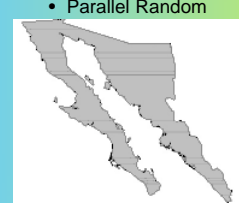
Example showing complex nested strata: a nested grid

Effort allocation set using formulae in Section 7.2.2.3 of *Introduction to Distance Sampling*
 (For more about this example, see Central Africa Pilot Project at <https://cites.org/eng/program/mike/pilot/index.shtml>)


Line Transect Designs

- Full Length Transects

- Parallel Random



- Systematic



Often used in aerial (and sometimes shipboard) surveys

Full Length Line Transects

- considerations

- **Coverage** for a given realisation is more critical as there tend to be fewer lines than points – lines are more expensive
- **Transit** (off-effort) time can be considerable
- Other full-width line transect designs include random line orientation, non-overlapping random parallel, etc.

Segmented Line Transect Designs

-Fixed Length Transects

- Systematic segmented trackline

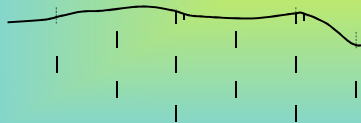


- Systematic segmented grid



Edge Effects

- Not a problem if you are willing to survey incomplete segments
- Otherwise you could reflect back incomplete segments already more than $\frac{1}{2}$ inside





This gives even coverage probability but is hard to reliably automate

Edge Effects (contd.)

Could push segments back in if they are already more than 1/2 inside



- Systematic segmented trackline
- Systematic segmented grid

Edge Effects (contd.)

... but this leads to uneven coverage probability near the edge

- Systematic segmented trackline
- Systematic segmented grid

N.B. Both use random orientation of transects in the northern stratum

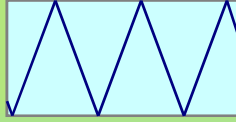
Fixed Length Line Transects

- considerations

- Systematic segmented grid seems superior
- Consider random orientation of lines, (in Distance, type -1 under angle in Effort Allocation tab)
- Random orientation of each segment may be even better, (not yet in Distance)
- Other designs (such as circuit samplers) are worth considering, (not yet in Distance)

Zig-zag Line Transect Designs

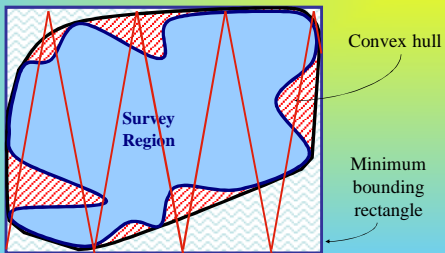
- Used commonly in shipboard surveys



- **Advantage** (over systematic parallel)
 - Improved efficiency
- **Disadvantages**
 - Design is difficult in complex regions
 - Coverage probability may be uneven

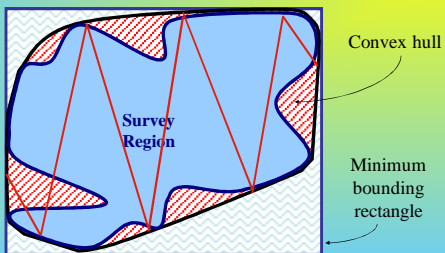
Design Difficulties

non-convex survey region



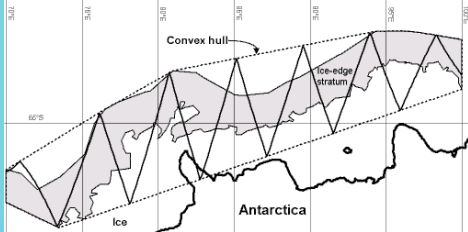
Design Difficulties

non-convex survey region



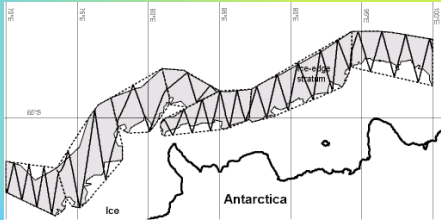
Dealing with Complex Survey Regions

- Example: Antarctic shipboard survey



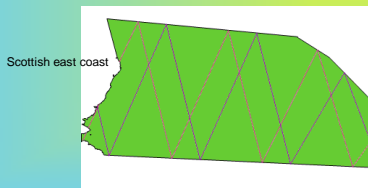
Dealing with Complex Survey Regions

- Example: Antarctic shipboard survey, (contd.)
- Study region divided into suitable strata to increase efficiency



Efficiency

- Example: SCANS II – ship survey in North Sea
- Cross survey region twice



Effort Allocation

- Example: SCANS II – aerial survey
- Distance outputs total track length for survey

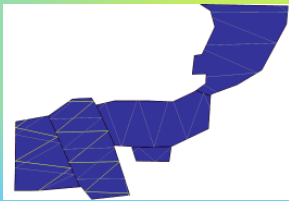
Considerations:

- Total effort available
- Required transit effort
- Rest periods
- Spare survey



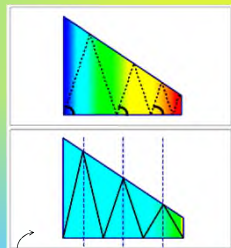
Stratification

- Example: SCANS II – aerial survey
- Stratification based on prior knowledge of animal density



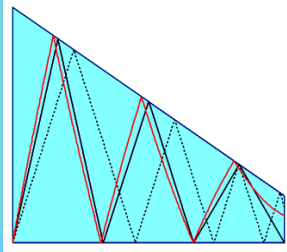
Coverage probability for zig-zag designs

- Equal angle zig-zag
- Equal spaced zig-zag



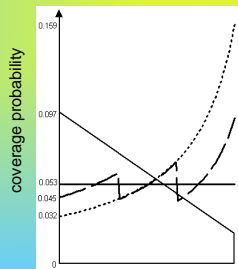
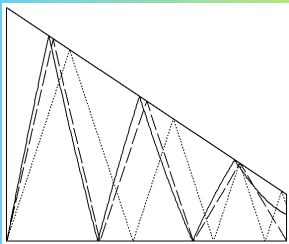
Coverage probability for zig-zag designs (contd.)

- Adjusted angle zig-zag



- Even coverage probability parallel to the design axis
- In practise, approximate curved path with a series of straight lines

Coverage Probability Comparison for Zig-zag Designs



Automated survey design in Distance

- Import data from GIS (or type it in!)
- Create coverage grid
- Create design
- Generate example surveys from design (run 2nd option)
- Assess even-ness of coverage probability via simulation (run 1st option)
- Finally can export GIS data, map or sampler coordinates

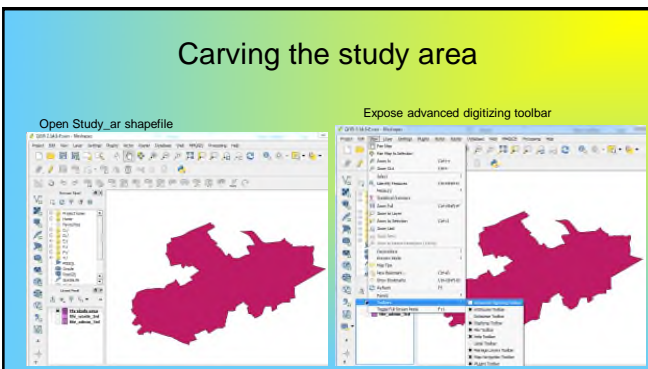
Main Points

- Line transects are preferable
- Parallel designs give uniform coverage
 - Systematic designs give more even coverage for any survey
- Zig-zag designs usually more practical
- Lines should be placed parallel to density gradient
 - Otherwise should be placed to maximise number of samplers

QGIS software to split study area into strata

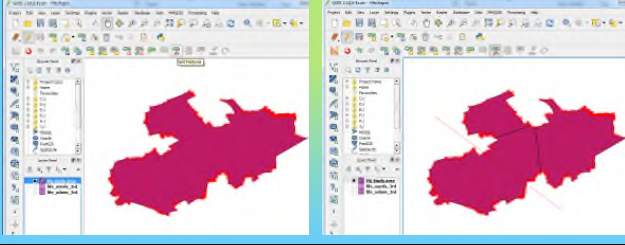
- Ensure Geographic Coordinate System is defined
- Project data on to a flat surface
 - Eg. Albers Equal Area Conical Projection

Carving the study area



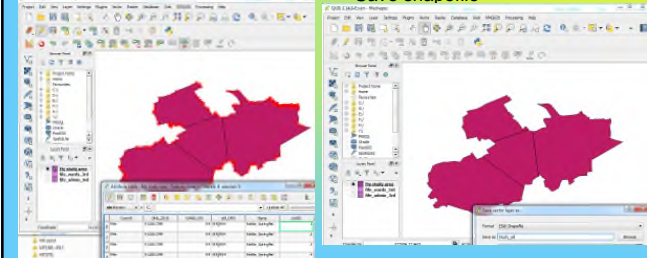
Carving study area (cont.)

- Toggle edit and choose split
- Right click start, left click finish



Save newly created strata shapefile

- Open attribute table, create/edit LinkID
- Save shapefile



Next Steps

- Create a new distance project
- Add correct number of strata
- Close project
- Replace the empty shapefiles created in the Distance project .dat folder with those manipulated in ArcView
 - Note: You may need to rename your files to match those created by Distance

Things to Remember!

- Arc ToolBox to define coordinate system and project shapefile
- 'Cut Polygon Features' in ArcView can be used to divide area into strata
 - Define LinkID in attributes table
 - Non-editing mode to create LinkID field
 - Editing mode to change LinkID values
- Distance help 'Importing Existing GIS Data'

Stratification and Clustered Populations

Stratification

- Why stratify?
- Stratification by:
 - Geographic area
 - Survey
 - Species / cluster size
- Limitations of Distance
- Section 3.7 in introductory book

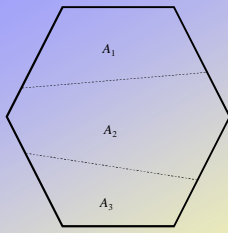
Stratification is used to:

- reduce variance and improve precision
- and for producing estimates in regions of interest

Stratify by:

- AREA or GEOGRAPHIC REGION
 - the study region is partitioned into smaller regions
- SURVEY
 - used when different surveys cover the same geographic area
- POPULATION/SPECIES/CLUSTER SIZE
 - same geographic region with different 'sub-stocks' in it

Area/Geographic stratification



Total size of study region
 $A = A_1 + A_2 + A_3$

Estimate density in each sub-region
 $\hat{D}_1, \hat{D}_2, \hat{D}_3$

Abundance in each sub-region is given by

$$\hat{N}_1 = A_1 \hat{D}_1$$

$$\hat{N}_2 = A_2 \hat{D}_2$$

$$\hat{N}_3 = A_3 \hat{D}_3$$

Total abundance is

$$\hat{N} = \hat{N}_1 + \hat{N}_2 + \hat{N}_3$$

$$= A_1 \hat{D}_1 + A_2 \hat{D}_2 + A_3 \hat{D}_3$$

Overall (Global in Distance) density is

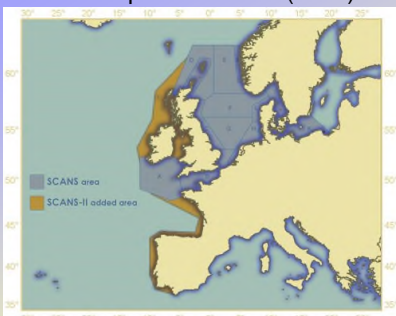
$$\hat{D} = \frac{\hat{N}}{A} = \frac{A_1 \hat{D}_1 + A_2 \hat{D}_2 + A_3 \hat{D}_3}{A_1 + A_2 + A_3}$$

$$= \left(\frac{A_1}{A}\right) \hat{D}_1 + \left(\frac{A_2}{A}\right) \hat{D}_2 + \left(\frac{A_3}{A}\right) \hat{D}_3$$

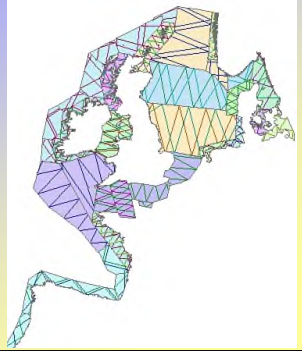
$$= \sum_{i=1}^3 \left(\frac{A_i}{A}\right) \hat{D}_i$$

Note form of equation

Example: SCANS II (2005)



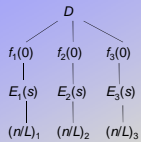
SCANS II
survey effort



Example of stratified data

Layer	Area	Length	...
1	10000	10000	...
2	20000	20000	...

Example: Full geographic stratification



Model Definition Properties: Full Stratification

Estimate: [Detection function] [Cluster size] [Abundance] [Density] [Size]

Stratification definition:

- Full stratification
- User layer type: Stratum
- Full density stratification

Sample definition (for secondary strata):

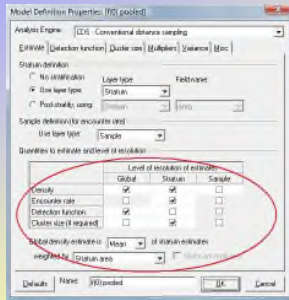
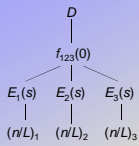
- User layer type: Sample

Quantities to estimate and level of resolution:

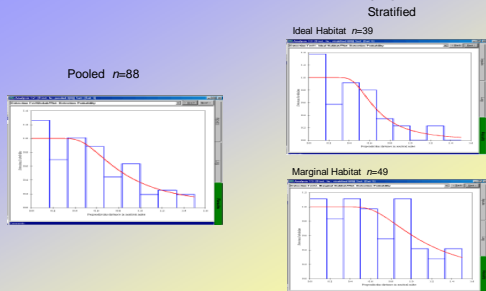
	Global	Stratum	Sample
Density	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Encounter rate	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Detection function	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Cluster size (if required)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Global density estimate is: Mean of stratum estimates

Example: $f(0)$ pooled



Pooled vs Stratified $f(0)$



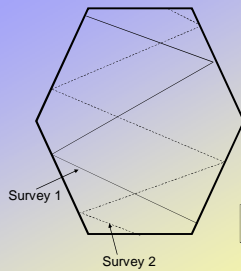
It's a Model Selection Problem

	Pooled	Stratum 1	Stratum 2	Stratum Sum
Log likelihood $\log_e(L)$	-180.490	-72.699	-104.676	-177.375
No. parameters (q)	2	2	2	4
AIC	364.980	149.398	213.352	362.75

Criterion for stratification of $f(0)$:
Fit separate $f(0)$ for each strata if

$$AIC_{pooled} > \sum_{strata} AIC_{stratum}$$

Non-geographic stratification Stratification by survey



Let L_i be effort for survey i

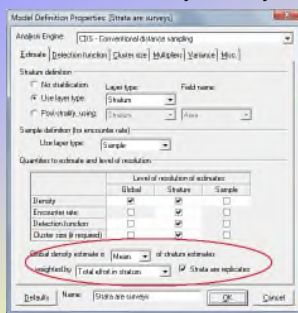
Global density is given by

$$\hat{D} = \left(\frac{L_1}{L_1 + L_2} \right) \hat{D}_1 + \left(\frac{L_2}{L_1 + L_2} \right) \hat{D}_2$$

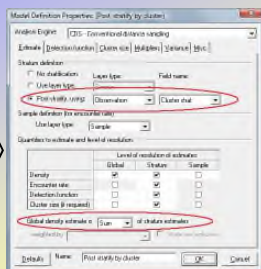
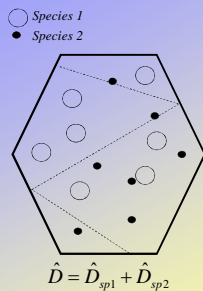
$$= \sum_{i=1}^2 \left(\frac{L_i}{L} \right) \hat{D}_i$$

This is the same form as before, but weighting factor now depends on effort

Stratification by survey

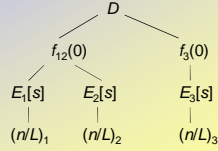


Stratification by species



Limitations in Distance

- Distance cannot currently do multilevel stratification in one run
- Two runs are necessary
 - Estimate $f(0)$, $E[s]$ and n/L by stratum
 - Combine strata 1 and 2 to estimate $f_{12}(0)$
- Care must be taken when calculating cv's because the density estimates for stratum 1 and 2 have an estimated $f(0)$ in common



Alternatives to stratification in Distance

- Small sample sizes can lead to low precision in stratum-specific estimates
- An alternative approach to reducing bias due to heterogeneity is Multiple Covariates Distance Sampling (MCDS)
 - Covariates, other than distance, are incorporated into the scale parameter of the detection function
 - MCDS can be used to fit the detection function at multiple levels e.g. stratum-specific density estimates can be obtained even if you don't have enough data to fit separate detection functions for each stratum
 - MCDS methods are covered in an upcoming lecture.

Clustered Populations

- What changes when animals occur in clusters
- Size bias
- Methods to deal with size bias
- How to implement these methods in Distance
- Section 3.5 in introductory book

Clustered populations

$$\hat{D} = \hat{D}_s \times \hat{E}(s)$$

← Density of clusters
← Mean cluster size

$$[cv(\hat{D})]^2 = \frac{\hat{V}(\hat{D})}{\hat{D}^2} \approx [cv(\text{encounter rate})]^2 + [cv(\text{detection function})]^2 + [cv(\text{cluster size})]^2$$

Mean cluster size estimation

Distance (x)

No Size Bias

- Mean of observed sizes does not change with distance

Distance (x)

Size Bias

- Smaller clusters less detectable at larger distances
- Mean observed cluster size **increases** with distance

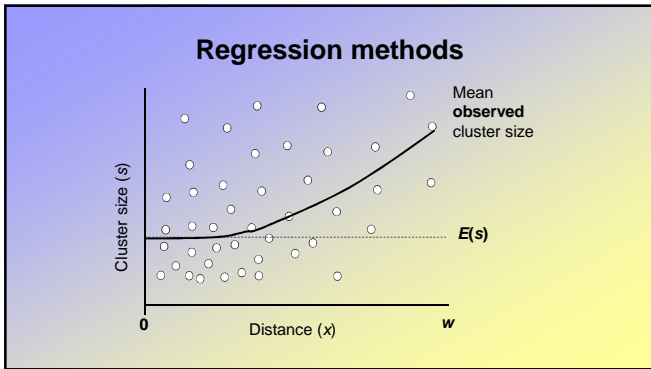
Effect of size bias on sample mean

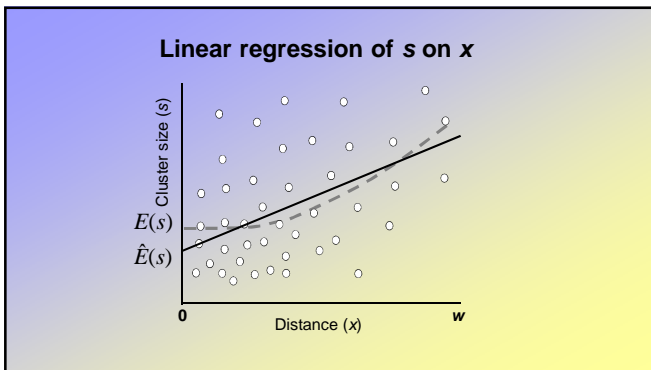
If size bias is present, $\hat{E}(s) = \bar{s}$ will be positively biased:

Distance (x)

Observed mean

True mean

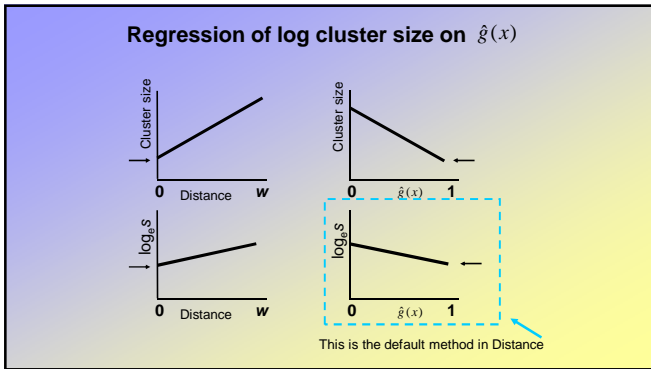


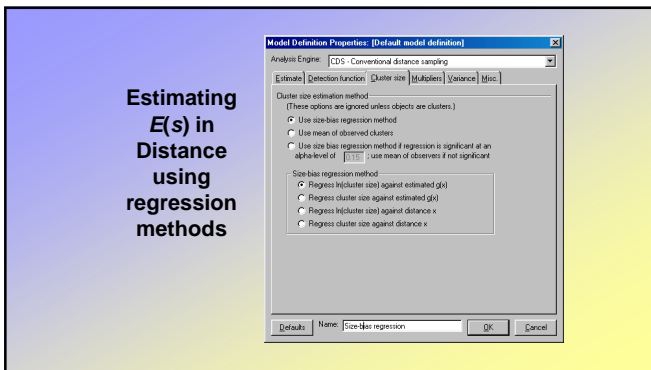


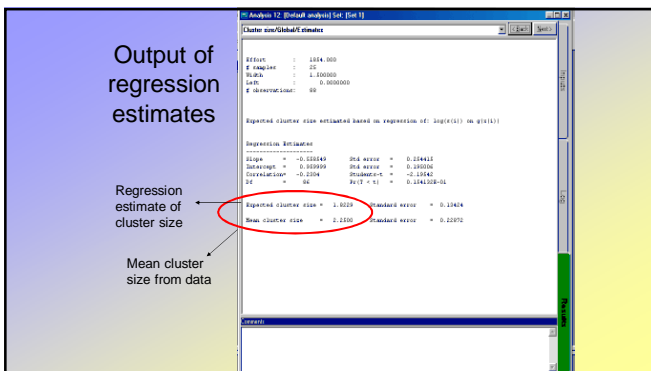
Problems with the linear regression method

- **Problem:** Relationship between s and x is not linear – no relationship when detection is certain (i.e. in the shoulder of the detection function).
- **Solution:** Linearize by regressing s on $\hat{g}(x)$

- **Problem:** Variance in s increases with $E(s)$ – large cluster sizes distort the fit.
- **Solution:** Regress **log of cluster size** on $\hat{g}(x)$







Multiple covariate distance sampling (MCDS)

Aim of MCDS

Model the effect of additional covariates on detection probability, in addition to distance, while assuming probability of detection at zero distance is 1

i.e. model $f(0)$ as a function of covariates

- Based on PhD thesis work by Fernanda Marques
- Chapter 3 of Advanced book
- Marques, T.A., L. Thomas, S.G. Fancy and S.T. Buckland. 2007. Improving estimates of bird density using multiple covariate distance sampling. *The Auk* 127: 1229-1243.
- Section 5.3.2.1 of Buckland et al. (2015) and
 - <http://www.creem.st-and.ac.uk/DS.M&A/amakihi/amakihi.html>

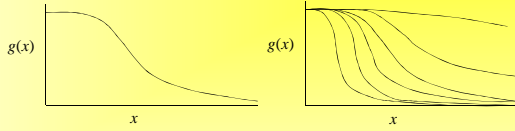
Contents

- Why additional covariates?
- Multiple covariate models
- Estimating abundance
- MCDS in Distance
- Complications
 - Clustered populations
 - Adjustment terms
 - Stratification
- MCDS analysis guidelines

Why additional covariates?

In conventional distance sampling (CDS) analysis all factors affecting detectability, except distance, are ignored

In reality, many factors may affect detectability



Sources of heterogeneity:
Object : species, sex, cluster size
Effort: observer, habitat, weather

Examples of heterogeneity 1

Effect of time of day on Rufous Fantail birds in Micronesia (point transects). Ramsey et. al. 1987, Biometrics 43:1-11

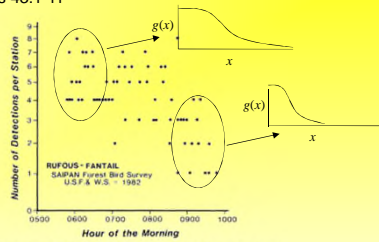


Figure 1. Station counts of Rufous Fantails on Saipan appear higher in the early morning hours than in the late morning ($n = 84, r = -.80$).

Examples of heterogeneity 2

Effect of sea state (and other covariates) on sea turtles in the Eastern Tropical Pacific (shipboard line transects). Beavers and Ramsey, 1998, J. Wildl. Manage. 63: 948-957

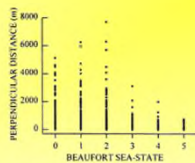


Fig. 2. Covariates of air temperature, sea surface temperature, and Beaufort sea-state plotted against unadjusted, ungrouped perpendicular sighting distances (m) of sea turtles in the eastern tropical Pacific, 1989-90.

Examples of heterogeneity 3

Effect of cluster size on beer cans. Otto and Pollock, 1990, Biometrics 46: 239-245

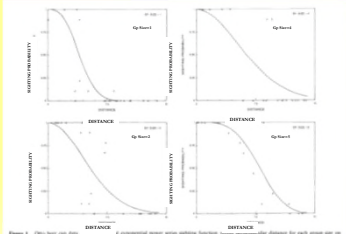


Figure 3. Otto beer can data - a mixture of exponential power series fitting function based on cluster size for each group size in column 1. The α gives the proportion of observations that signed the cluster.

Why worry about heterogeneity?

In CDS, we use models that are pooling robust, so why worry about heterogeneity?

- Pooling robustness works for all but extreme levels of heterogeneity
- Potential bias if density is estimated at a 'lower level' than detection function (e.g. density by geographic region, detection function global)
- Could potentially increase precision of detection function estimate
- Interest in sources of heterogeneity in their own right (e.g. group size)

Dealing with heterogeneity

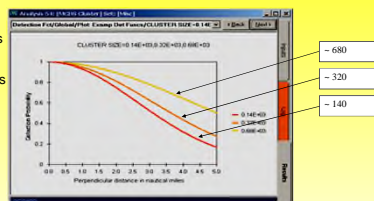
Stratification

Requires estimating separate detection function parameters for each stratum, so often not possible due to lack of data

Model as covariates in detection function

Allows a more parsimonious approach:

- can model effect of numerical covariates
- can 'share information' about detection function shape between covariate levels



Multiple covariate models

Recap of CDS models

$$g(x) = \text{Pr}[\text{animal at distance } x \text{ is detected}]$$

$$= k(x) \left[1 + \sum_{j=1}^m a_j p_j(x_s) \right] / c$$

Key function $k(x)$ j^{th} series adjustment term $a_j p_j(x_s)$ Scaling constant to ensure $g(0) = 1$

CDS models continued

Key functions	Shape parameter	Series adjustments
• Hazard rate $k(x) = 1 - \exp\left[-\left(\frac{x}{\sigma}\right)^{-b}\right]$		• Cosine $\cos(j\pi x_s)$
• Half-normal $k(x) = \exp\left(\frac{-x^2}{2\sigma^2}\right)$		• Polynomial x_s^j
• Neg. exp. $k(x) = \exp\left(\frac{-x}{\lambda}\right)$		• Hermite poly. $H_j(x_s)$
• Uniform $k(x) = 1$		x_s are scaled distances (see later)

Modelling with covariates

ignoring adjustments terms (for now)

$g(x, \mathbf{z}) = \text{Pr}[\text{animal at distance } x \text{ and covariates } \mathbf{z} \text{ is detected}]$

Assume the covariates affect the **scale** of the key function, not its **shape**. So choose key functions with a scale parameter

Let $\sigma(\mathbf{z}) = \exp\left(\beta_0 + \sum_{j=1}^J \beta_j z_j\right)$

e.g. Hazard rate $k(x, \mathbf{z}) = 1 - \exp\left[-\left(\frac{x}{\sigma(\mathbf{z})}\right)^{-b}\right]$

Half normal $k(x, \mathbf{z}) = \exp\left(\frac{-x^2}{2\sigma(\mathbf{z})^2}\right)$

k is used here to denote the "key" function

Modelling with covariates

Example: Dolphin tuna vessel data
 Model: half-normal, with no adjustments
 Covariate: cluster size, s

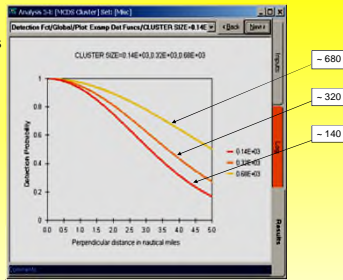
$$g(x, s) = \exp\left(\frac{-x^2}{2\sigma(s)^2}\right)$$

$$\sigma(s) = \exp(\beta_0 + \beta_1 s)$$

$$= \exp(\beta_0) \cdot \exp(\beta_1 s)$$

$$= A1 \cdot \exp(A2s)$$

From distance output
 $\hat{A}1 = 2.331$
 $\hat{A}2 = 0.00895$



Estimating abundance without covariates using Horvitz-Thompson estimator

$$\hat{N} = \sum_{i=1}^n \frac{1}{\Pr[\text{animal included}]} = \sum_{i=1}^n \frac{1}{\left[\frac{2L\hat{\mu}_i}{A}\right]} = \frac{nA}{2L\hat{\mu}}$$

Recall that $f(x)$ = pdf of observed x 's $= \frac{g(x)}{\int g(x)dx} = \frac{g(x)}{\mu}$

Because $g(0)=1$ by assumption, then $f(0) = 1/\mu$

So $\hat{N} = \frac{nA}{2L} \hat{f}(0)$

Estimating abundance with covariates

$$\hat{N} = \sum_{i=1}^n \frac{1}{\Pr[\text{animal included}]} = \sum_{i=1}^n \frac{1}{\left[\frac{2L\hat{\mu}(z_i)}{A}\right]} = \frac{A}{2L} \sum_{i=1}^n \frac{1}{\hat{\mu}(z_i)}$$

Now $f(x|z) = \frac{g(x,z)}{\int g(x,z)dx} = \frac{g(x,z)}{\mu(z)}$

Because $g(0,z)=1$ by assumption, then $f(0|z) = 1/\mu(z)$

So $\hat{N} = \frac{A}{2L} \sum_{i=1}^n \hat{f}(0|z_i)$

Note similarity to CDS estimator

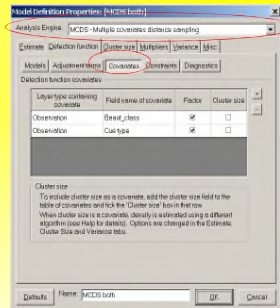
MCDS in Distance

In Model Definition, choose MCDS analysis engine

See Chapter 9 of online Users Guide

Covariate type:

- Factor covariates classify the data into distinct classes or levels. Can be numeric or text. One parameter per factor level.
- Non-factor (ie. continuous) covariates must be numeric. One parameter per covariate + 1 for the intercept.



Complications

1. Clustered populations

There are two approaches to estimating number of individuals when objects are in clusters:

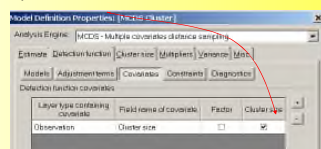
$$(1) \hat{N} = \sum_{i=1}^n \frac{1}{\Pr[\text{group included}]} \hat{E}[s] \quad (2) \hat{N} = \sum_{i=1}^n \frac{\text{group size}}{\Pr[\text{group included}]}$$

$$= \frac{A}{2L} \hat{E}(s) \sum_{i=1}^n \hat{f}(0 | z_i) \quad = \frac{A}{2L} \sum_{i=1}^n s_i \hat{f}(0 | z_i)$$

When cluster size is *not* a covariate, we use (1); when it is a covariate, we use (2)

Clustered populations (contd.)

To tell Distance that a covariate represents cluster size, tick the box:



When cluster size is a covariate:

- Distance does not estimate variance using analytic methods: the bootstrap must be used (Reflected in the Variance tab)
- There is no need for size bias regression methods (Cluster size tab changes)
- No stratification allowed (Estimate tab)

Complications

2. Adjustment terms

With adjustments: $g(x, z) = k(x, z) \left[1 + \sum_{j=1}^m a_j p_j(x_s) \right] / c$

Adjustment terms use *scaled* distances, x_s

- cosine adjustment of order 2: $\cos(2\pi x_s)$
- simple polynomial of order 4: x_s^4

Why scale?

- Avoid numerical problems
- Limits cosine adjustment to a small number of 'wiggles'

How to scale?

Adjustment terms (contd.)

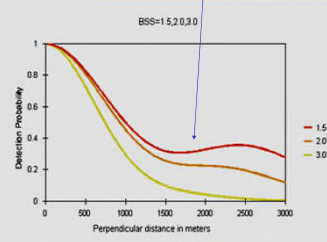
Scenario 1: Scale distances by w , the right truncation distance $x_s = x/w$

Note: no monotonicity constraint

Then covariates affect the scale of the key function, but adjustment terms are unaffected by covariates, so the overall shape varies with covariate value:

e.g. half-normal with 1 cosine adjustment of order 2

$$g(x|z) \propto \exp\left(\frac{-x^2}{2\sigma(z)^2}\right) \left[1 + a_2 \cos\left(\frac{2\pi x}{w}\right) \right]$$



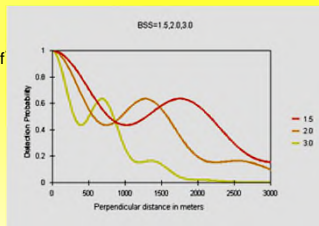
Adjustment terms (contd.)

Scenario 2: Scale distances by $\sigma(z)$, the estimated scale parameter $x_s = x/\sigma(z)$

Then covariates affect the scale of the key function, and the scale of the adjustment terms, so only the scale and not the shape of the overall function is affected:

e.g. half-normal with 1 cosine adjustment of order 2

$$g(x|z) \propto \exp\left(\frac{-x^2}{2\sigma(z)^2}\right) \left[1 + a_2 \cos\left(\frac{2\pi x}{\sigma(z)}\right) \right]$$

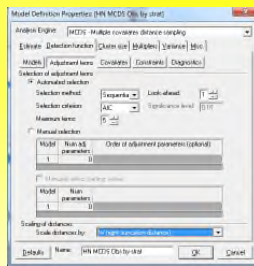


Adjustment terms (contd.)

The previous was an extreme example, to illustrate the difference between scaling factors.

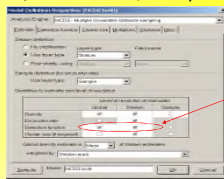
Generally:

- start with no adjustment terms and introduce covariate terms one by one
- check the fit with adjustments looks reasonable
- consider whether to scale by w or σ
- you may need fewer adjustment terms with MCDS than CDS analyses



Complications 3. Stratification

If we want stratum-level estimates of density/abundance we can fit the detection function with covariates globally, and estimate $f(0|\mathbf{z})$ by stratum:



Tick both boxes

- If estimating density by sample, could estimate $f(0|\mathbf{z})$ by stratum
- Global variance estimate for density/abundance must be calculated via the bootstrap

MCDS analysis guidelines

Choose covariates that are:

- independent of distance
- not strongly correlated with each other

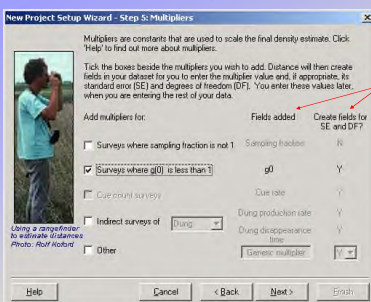
Specifying the model:

- factor covariates generally harder to fit
- avoid or limit automatic selection of adjustment terms
- if using adjustments, consider whether to scale by w or σ
- check convergence and monotonicity
- add only one covariate at a time
- where necessary, use starting values and bounds for parameters
- consider reducing the truncation distance, w , if more than 5% of the $P_{ij}(z_i)$ are <0.2 , or if any are less than 0.1

Multipliers and Indirect Methods

- Why and how we use multipliers
- Cue counting
- Indirect surveys
- Lure and trapping point transects
- Section 3.1.4 in introductory book

Multipliers in Distance



Adds appropriate fields to Global data layer

Multipliers

- If $g(0) < 1$, then the standard method of analysis will produce a density estimate that is proportional to the true density. Then true density (without clusters) is estimated using

$$\hat{D} = \frac{n\hat{f}(0)}{2L} \cdot \frac{1}{\hat{g}(0)}$$

These are called multipliers

- In some surveys, cues (whale blows, bird songs) are the object of detection rather than the animal itself.
- For instantaneous cues (whale blows, bird songs) animal density, D , is estimated by cue density D_c divided by cue rate r

$$\hat{D} = \frac{\hat{D}_c}{r}$$

Multipliers: examples

The multiplier, denoted by c , might be

- a known constant
 - sampling fraction $\neq 1$
- a parameter, or product of parameters, to be estimated
 - $\hat{g}(0) < 1$
 - some proportion of the population is surveyed
 - cue counting
 - indirect surveys

Examples: sampling fraction $\neq 1$

One-sided line transect sampling: $c = 0.5$ to represent the fraction of the strip surveyed

$$\hat{D} = \frac{n\hat{f}(0)}{2L} \cdot \frac{1}{0.5} = \frac{n\hat{f}(0)}{L}$$

In point transect sampling if one quarter of the circle was surveyed: $c = 0.25$

$$\hat{D} = \frac{n\hat{h}(0)}{2\pi k} \cdot \frac{1}{0.25}$$

Point transect sampling with each point visited five times: $c = 5$

Cue counting where c is the proportion of the circle covered by the observation sector (see later)

Examples: parameters to be estimated

- Surveys where $g(0) < 1$
- Surveys in which only a proportion of the population is surveyed:
 - $c = p$ where p is the proportion surveyed,
 - usually must be estimated,
 - e.g. desert tortoises, seabirds on land/at sea, whales with long dive times
- Cue counting where c is the cue rate
- Indirect surveys e.g. dung/nest surveys (see later)

Multipliers: variance

Remember the multiplier is denoted by c .

If c must be estimated (by \hat{c}) then this additional variance needs to be included in the density variance

For line transect sampling

$$\hat{D} = \frac{n\hat{f}(0)}{2L\hat{c}} \quad cv(\hat{D}) = \sqrt{\{cv(n)\}^2 + \{cv[\hat{f}(0)]\}^2 + \{cv(\hat{c})\}^2}$$

For point transect sampling

$$\hat{D} = \frac{n\hat{h}(0)}{2\pi k\hat{c}} \quad cv(\hat{D}) = \sqrt{\{cv(n)\}^2 + \{cv[\hat{h}(0)]\}^2 + \{cv(\hat{c})\}^2}$$

Cue counting: point transects

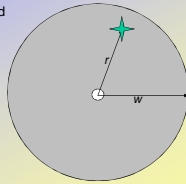
Point transect survey where distance to detected cue is recorded

Cue is single burst of song (instantaneous cue)

Valid even if birds are moving during the count

Cue density is

$$\hat{D}_{cues} = \frac{n\hat{h}(0)}{2\pi} \text{ cues per unit area}$$



And if you searched for time $T = \sum_{i=1}^k$ time spent at point i

$$\hat{D}_{cues/T} = \frac{n\hat{h}(0)}{2\pi T}$$

cues per unit area, per unit time

Note: the standard point transect estimator is

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

Cue Counting: animal density

We want animal density, not cue density per unit time, so

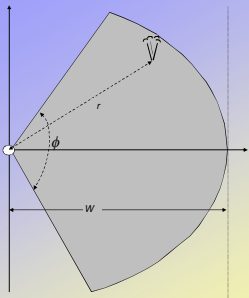
$$\hat{D}_{animals} = \frac{\hat{D}_{cues/T}}{\hat{\eta}}$$

Where $\hat{\eta}$ is the estimated number of cues per animal, per unit time

New component of variance

$$CV^2[\hat{D}_{animals}] \approx CV^2[\hat{D}_{cues/T}] + CV^2[\hat{\eta}]$$

Cue Counting: line transects



Fraction of circle searched: $\frac{\phi}{2\pi}$
 (ϕ in radians)

So that:

$$\hat{D}_{cues} = \left(\frac{n\hat{h}(0)}{2\pi} \right) \div \left(\frac{\phi}{2\pi} \right) = \frac{n\hat{h}(0)}{\phi}$$

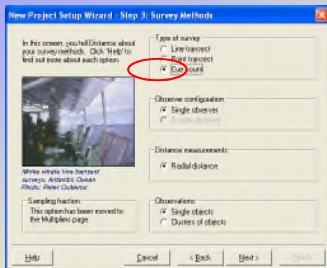
cues per unit area.

And if you searched for time T

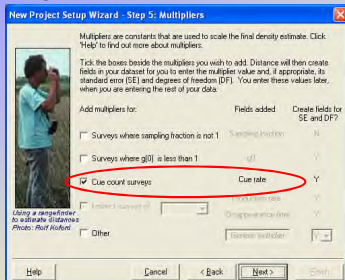
$$\hat{D}_{cues/T} = \frac{n\hat{h}(0)}{\phi T}$$

cues per unit area, per unit time.

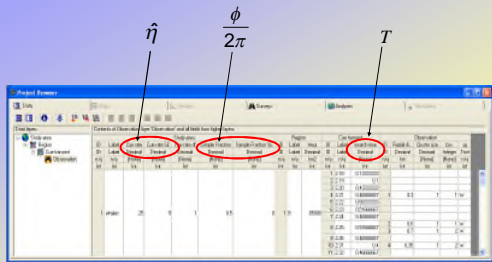
Setting up a cue counting project



Setting up multipliers for cue counting



Cue counting project: example data



Indirect surveys

- Useful when direct distance sampling of a population is difficult,
 - but estimating the density of some object produced by the animals is feasible
- Examples are dung surveys of deer, elephants, big cats and nest surveys of apes
- Production rate and the disappearance rate of the objects of interest need to be estimated
- Key difference between direct and indirect surveys
 - for direct surveys, an estimate of abundance at the time of the survey is obtained
 - for indirect surveys, the final estimate of abundance is an average over a time period corresponding to the mean time to decay of the object

Estimating animal density from indirect surveys

Example: a line transect survey of dung (the same procedure also applies to surveys of nests)

Use conventional methods to estimate the density of the object of interest, in this case we estimate dung density,

$$\hat{D}_d = \frac{n\hat{f}(0)}{2L} = \text{dung density}$$

- Divide dung density by \hat{d} = estimated mean time to decay (in days say)

$$\hat{G} = \frac{\hat{D}_d}{\hat{d}} = \text{dung production per day per unit area}$$

- Finally, divide by \hat{r} = estimated daily production of dung by one animal, (number of dung piles per day)

$$\hat{D} = \frac{\hat{G}}{\hat{r}} = \frac{\hat{D}_d}{\hat{d} \cdot \hat{r}} = \text{animal density}$$

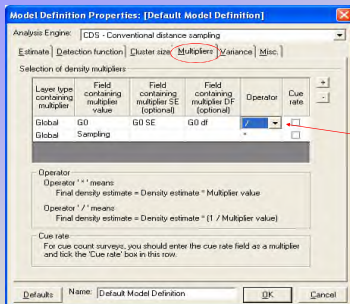
Estimating defecation rates

- Observe the animals in the wild in the study region, and record defecation rate
- Observe animals in captivity, in an environment as close as possible to that of the study region
- Put a known number of captive animals into a natural enclosure clear of dung
 - Leave them for a period that is less than the shortest decay time
 - Count, or estimate, the dung abundance at the end of the period
 - Defecation rate is then estimated from

$$\hat{r} = \frac{\text{number of dung piles}}{\text{number of animals} \times \text{number of days in enclosure}}$$
- Sample size is the number of animals, not the number of dung piles
- Similar considerations apply to nests

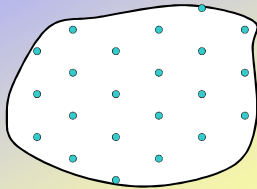
Estimating dung decay rates

- May vary spatially and seasonally and so carry out the decay rate study in the region and time leading up to the survey
- Define consistent criteria for determining whether dung has decayed
- Search for and mark fresh dung at a representative sample of sites at intervals of time which span the decay period of more persistent dung
- During the line transect survey, pay a single visit to each marked dung pile and record whether or not it has decayed (more visits may be required if the line transect survey is of long duration)
- Analyse the data using logistic regression with time between marking and the revisit as the explanatory variable (and possibly additional variables)
- Similar considerations apply to estimating nest decay rates



Trapping and lure point transects

These use just one trap (or lure) per sampling plot:

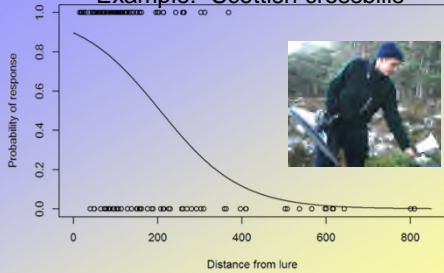


Estimating the detection function

We do not know the initial location of animals that are trapped or lured, so that distances from the point are unobserved.

We therefore need a sample of animals whose initial location is known. We then record whether each of these is trapped, or lured to the point.

Example: Scottish crossbills^a



^aBuckland, S.T., Summers, R.W., Borchers, D.L. and Thomas, L. 2006. Point transect sampling with traps or lures. *Journal of Applied Ecology* **43**, 377-384
Section 9.2.1 of Buckland et al. (2015) and <http://www.creem.st-and.ac.uk/DS.M&A/crossbills/crossbills.html>

Key Largo wood rats

- Over 4 years, 33 females and 22 males were radio collared
- More than 1000 trials (trap exposures) were conducted on these individuals
- Sex-specific random effects models were used to estimate detection probabilities as functions of distance of animal from trap
- Clearly these secretive animals are unlikely to be caught in traps even if the traps are atop the animal

Potts, J.M., S.T. Buckland, L. Thomas and A. Savage. 2012. Estimating abundance of cryptic but trappable animals using trapping point transects: a case study for Key Largo woodrats. *Meth. Ecol. and Evol.* 3:695-703.

Section 9.2.2 of Buckland et al. (2015)

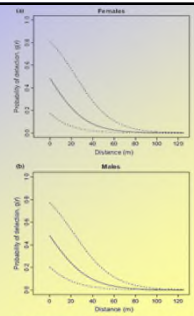
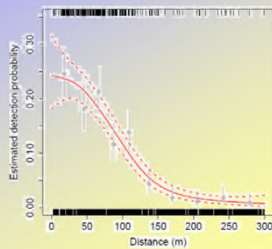


Fig. 8. Mean detection functions for adult females corresponding to the best model for the female (Panel A, $M = \text{mfm}, k = \text{mfm}$) and male (Panel B, $M = \text{mfm}, k = \text{mfm}$) woodrats. The data show that the mean Akaike's Information Criterion (AIC) value of all the models fitted to the data is significantly lower than the AIC value of the best model. The 95% CI for the AIC value of the best model is shown in brackets. The 95% CI for the AIC value of the best model is shown in brackets.

Baltic harbour porpoise

- Hydrophones (C-PODs) placed in the Baltic
- Visual tracking of porpoises by observers set up the "trials"
- Logistic regression permits estimation of detection probability of porpoises at different distances from the hydrophones

• See <http://www.sambah.org> for more details



Example detection function (data from Line Kyhn)

Advantage

- We do not assume that detection at the point is certain – we allow $g(0) < 1$

Disadvantage

- Trade assumptions for data
- We need to know the initial location of a number of animals, e.g. using radio-tagging or lure trials

Field Methods:

(given an adequate survey design has been used)

- Objectives of adequate field methods
 - $g(0)=1$
 - Reduce / avoid effect of movement
 - Get accurate and precise distances
- General recommendations
- A few special circumstances

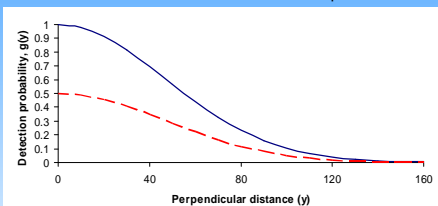
"Considerable potential exists for poor field procedures to ruin an otherwise good survey"

Goal: ensure key assumptions met

- $g(0)=1$
- no responsive movement prior to detection
- distances measured without error
- detection function has a wide shoulder

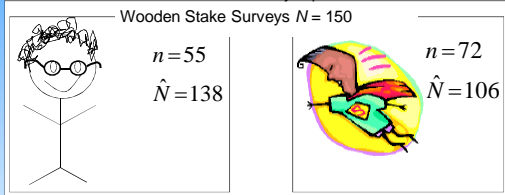
Make sure that $g(0)$ is 1

- Traditional data tells you nothing about $g(0)$
- Good field methods and common sense help to achieve it



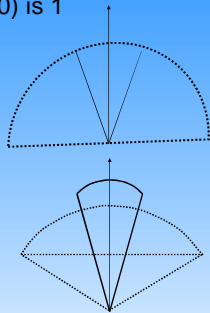
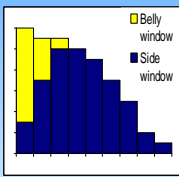
Make sure that $g(0)$ is 1

- Do not try to see everything
- But try to see everything on the line
 - More detections do not necessarily equate to better data



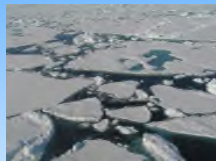
Make sure that $g(0)$ is 1

- Use multiple observers
- But avoid spiked data...



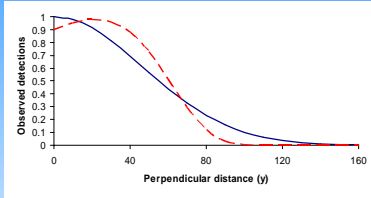
Warning – $g(0)$ is probably < 1 !

- Situation
 - Even with a well-defined search protocol and good observers, animals near the line may be missed
- Problems
 - Underestimation in density/abundance
 - Added variability (if $g(0)$ changes with survey period) reduces power
- Solutions
 - Independent observers to estimate $g(0)$
 - Technology (Video Camera, Infrared)
 - Change methods (go slower, lower)
 - Independent estimates of $g(0)$
 - trials on animals of known location



Avoid the effect of movement

- detect animals prior to responsive movement



- effect on data is not always obvious

Avoid the effect of movement

For points:

Snapshot method, waiting periods (before and after)
Use cues rather than individuals?

For lines:

Look ahead
Move slowly, carefully, quietly
– but if observer speed < 2-3 times average animal speed, see
Section 6.5 of introduction to distance sampling book

Get accurate and precise distances

Technological aids can be invaluable - use whenever possible

Avoid introducing more uncertainty by guessing



Get accurate and precise distances

If possible, mark the transect line

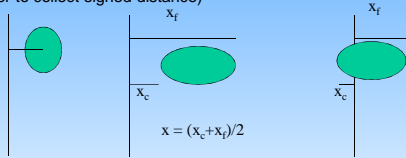


A clear definition of what you are measuring distance to helps to guard against spiked data and bias



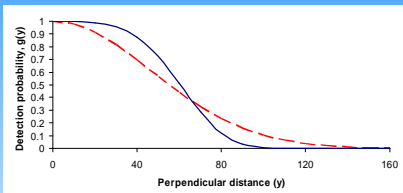
Get accurate and precise distances

- If size of animal/object is large compared to scale of measurements, define what measurement is to be made (e.g. from line to centre, tallest part, flower, etc)
- If measuring distances to clusters, get the distance to the "centre of the cluster"
- In practice, the mean between closest and furthest away distance might be enough (remember to collect signed distance)



General recommendations

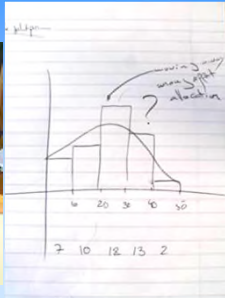
- Strive for wide shoulder in detection function



- Think about optimal effort allocation (ensure $g(0)$ while distributing effort)
- More than one observer?

General recommendations

- If possible, review data during survey



General recommendations

- Recording data should be easy, accurate and reliable
- Collect only relevant data
 - Perpendicular distance or distance and angle? (Angles for point transects?)
 - Cluster size
 - Effort (line length; no. of points); line or point ID
- Observer name, survey block, date, start time, end time, weather, environmental conditions, habitat, sex, species, age, etc...

General recommendations

- Make data collection as easy as possible e.g.:
 - dedicated field sheets
 - distance intervals for aerial surveys
 - tape recorder + voice activated microphone
 - separate person to record data
 - automated data entry (ship's GPS, etc.)
 - video
- Have a backup
 - backup recording method
 - backup of field data

General recommendations

- (most...) OBSERVERS ARE HUMAN...
 - Observing for long hours can be boring – plan breaks /rotations
 - Want to count what you see
 - have a ">w" category
 - for one-sided transects, have a category for negative values
 - Teach observers how to search
 - Emphasize effort on and near line
 - Look ahead
 - Look back if necessary
 - Do not assume observers know what to do
 - Go with observers to the field
 - Test and train observers – reward good observers?



Special circumstances: Multi-species surveys

- Problems
 - Species differences in detection
 - Identification of similar species
 - High density situations
- Solutions
 - Multiple observers
 - Training
 - Focus on key species



Animals at high density

- Consider strip transects
- Reduce truncation width
- Increase observation time (move more slowly)
- Multiple observers
- Streamline data collection

One-sided transects

- Avoid!
- Problems:
 - accurate line determination
 - movement into or out of survey strip
- Leads to heaping at zero distance

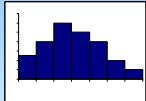


Some of what can go wrong, will likely go wrong

I spent all my money and have no data!



What do I do with this?



- Situation
 - Hi tech breakdown
 - No planning
 - Haven't thought about assumptions
- Problems
 - Data are lost
 - Poor quality data
- Solutions
 - Sometimes low-tech is better
 - Backups
 - Conduct a pilot survey
 - Train observers
 - Examine data during survey

Introduction to Distance Sampling

Exercise 1: Line transect estimation by hand

1) Plot a histogram of the following duck nest data, and fit a detection function by eye. From your histogram, estimate the proportion of nests within 2.4m of the line that are seen, P_a . Hence estimate nest density D (number of nests per square meter or per square kilometer – be careful of units!).

$n=534$ nests. $L=2575$ km.

Perpendicular distance band (meters)	0.0-0.3	0.3-0.6	0.6-0.9	0.9-1.2	1.2-1.5	1.5-1.8	1.8-2.1	2.1-2.4
Frequency	74	73	79	66	78	58	52	54

Having produced your fit to the histogram, to assist in producing your estimate of nest density, fill in these blanks.

Area of rectangle =

Area under your fitted detection function =

$$P_a = \frac{\text{area}_{\text{curve}}}{\text{area}_{\text{rectangle}}} =$$

$$\hat{N}_a = \frac{n}{P_a} =$$

$$\hat{D} = \frac{\hat{N}_a}{a} = \frac{\hat{N}_a}{2wL} =$$

Complete only part 1) of this exercise until instructed to go further.

2) Now use your histogram to estimate the effective half-width of search μ . Again estimate nest density D . How does it compare to your estimate from part (a)?

3) Rescale the y-axis to make your curve into the probability density function $f(x)$. Read off $f(0)$, and again estimate nest density D . How does it compare with your previous estimates?

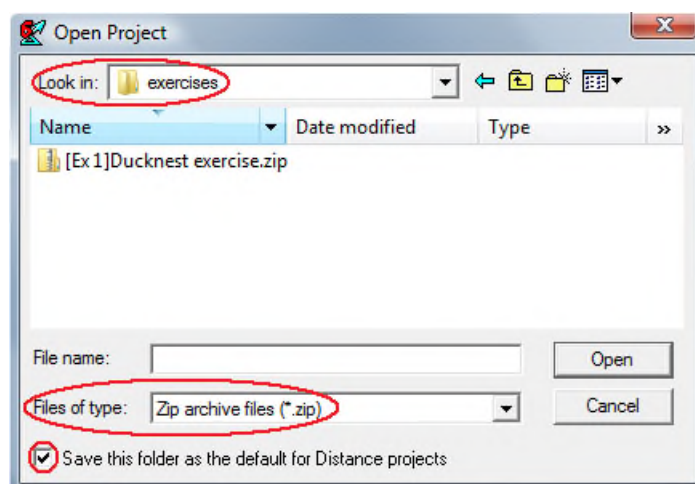
Introduction to Distance Sampling

Exercise 2: Line transect estimation using Distance: Ducknests

GETTING STARTED ON THE COMPUTER

Click on “**Start**”. A list will be displayed. Click on “**Programs**”, then “**Distance**”. Now click on “**Distance 7.0**”. (Or double-click the **Distance 7.0** icon on the desktop.) This opens Distance.

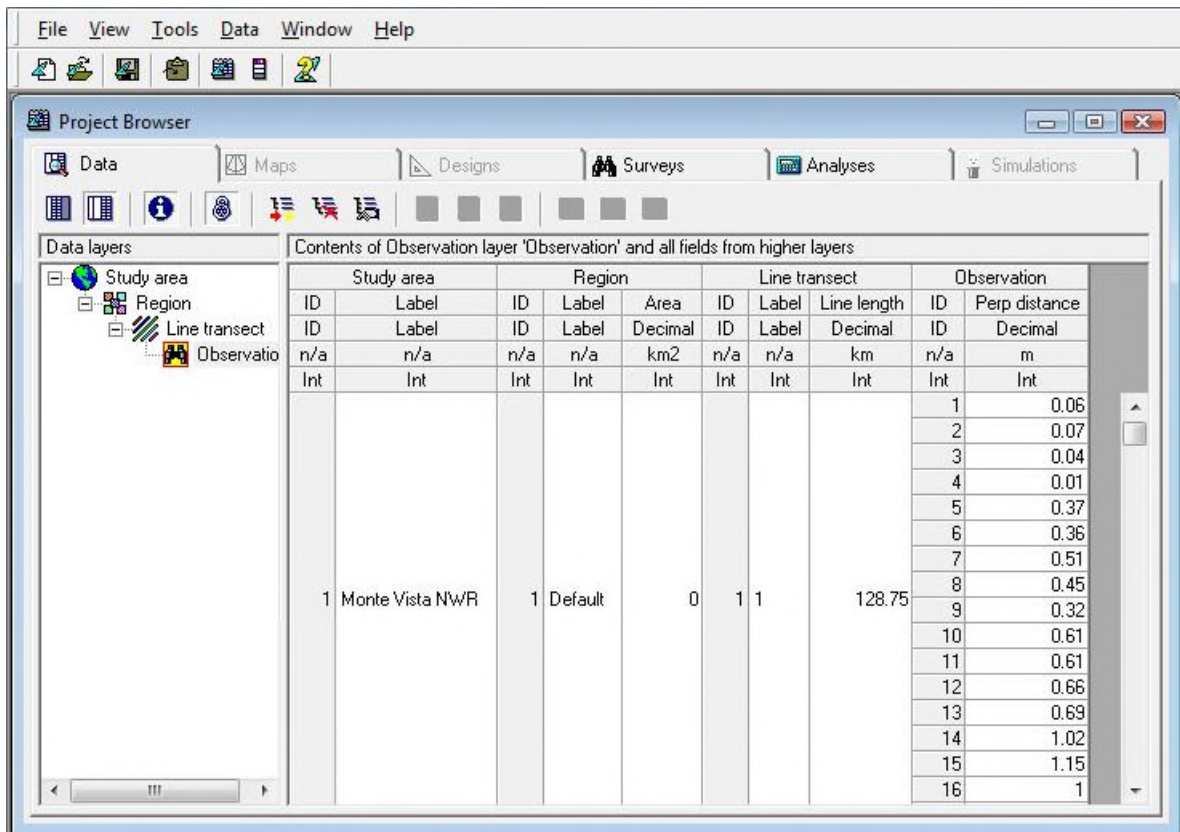
1. Refer to the data from the graph paper exercise (Exercise 1). These data have been set up as a Distance project, which have been archived in a compressed (.zip) file on your thumb-drives. You should copy the subdirectory containing the Distance projects for the workshops to the My Distance Projects folder under My Documents, or to a location of your choice (to make it easier to find for subsequent exercises).
- Select **File** followed by **Open project**. Under “**Files of type**” choose **Zip archive files (*.zip)**.
 - Next to “**Look in:**”, browse for the thumb drive directory (or wherever you have saved the exercises). **Note:** Distance includes some sample projects. The Sample Projects folder is the default folder Distance opens when instructed to open a new project, and contains a different duck nest data set, so make sure that you are looking in the right place!
 - You can change the default folder to one of your choice by checking the box next to “Save this folder as the default for Distance projects”. If you do this, the next time you open a project, Distance will look in the folder you specified – containing all the exercises relevant to this workshop. The Sample Projects folder will still exist in Distance, and you may want to look at those projects at a later date.



- Double click on **Ducknest exercise.zip**. Click **OK** to unpack the project into the current directory and open it. Next time you open the project, you can open the file *Ducknest exercise.dst* directly.

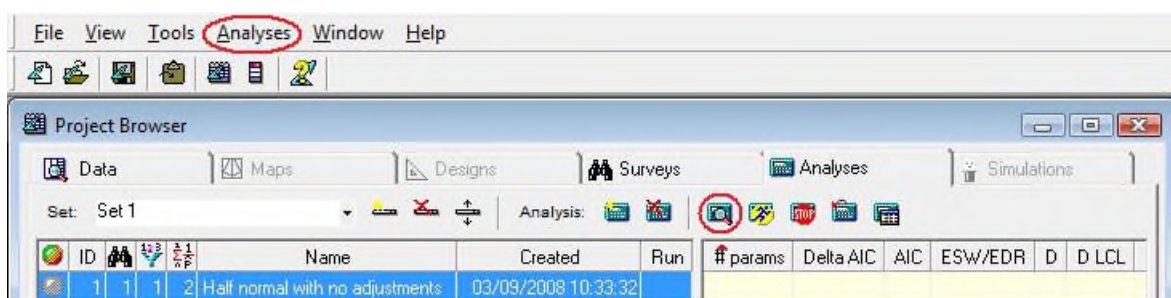
Examining the data

- Click on the **Data** tab of the **Project Browser** to show the **Data Explorer**. Look at the data structure and in particular how the distance data have been entered. (You will need to click on **Observation** in the left hand pane of the Data Explorer to see this.)

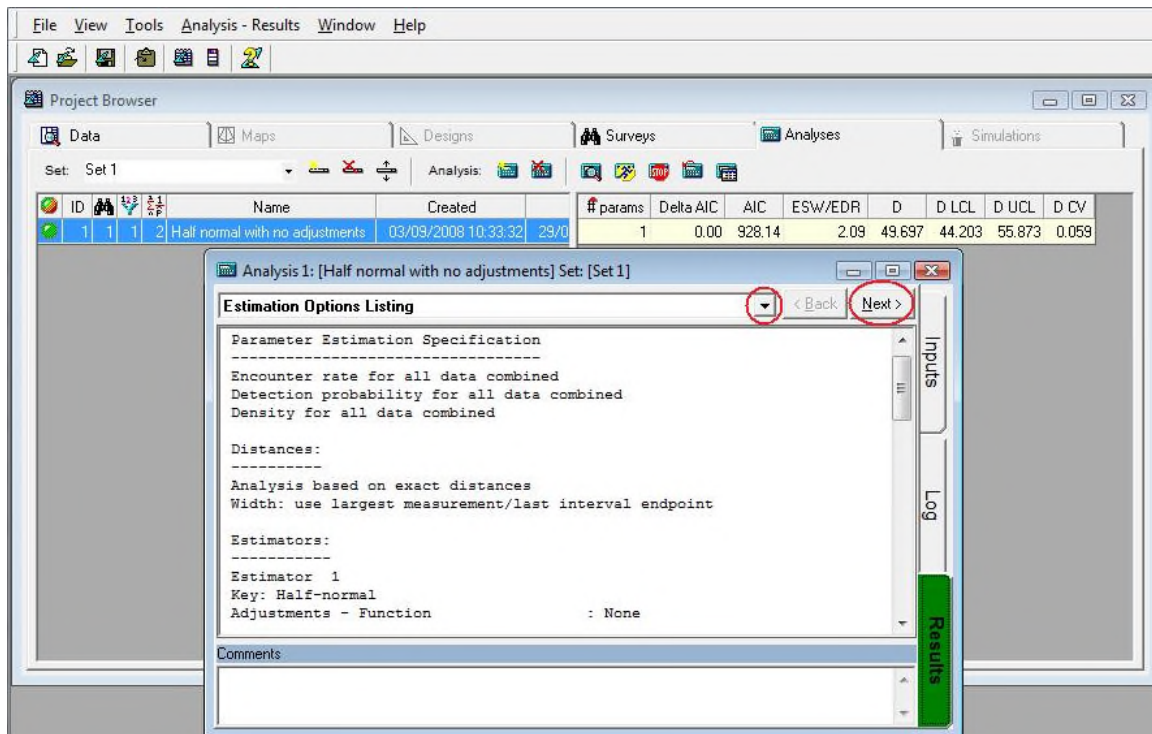


Studying the first analysis

- Now click on the **Analysis** tab of the **Project Browser** to show the **Analysis Browser**. You should see one analysis listed, called “Half-normal no adjustments.” Double-click on the grey status button for this analysis to open the **Analysis Inputs** tab for this analysis (you can do the same thing by clicking the 3rd button after “Analysis:” on the Analysis Browser menu bar, or by choosing **Analyses** then **Analysis Details...** from the menu bar at the top).

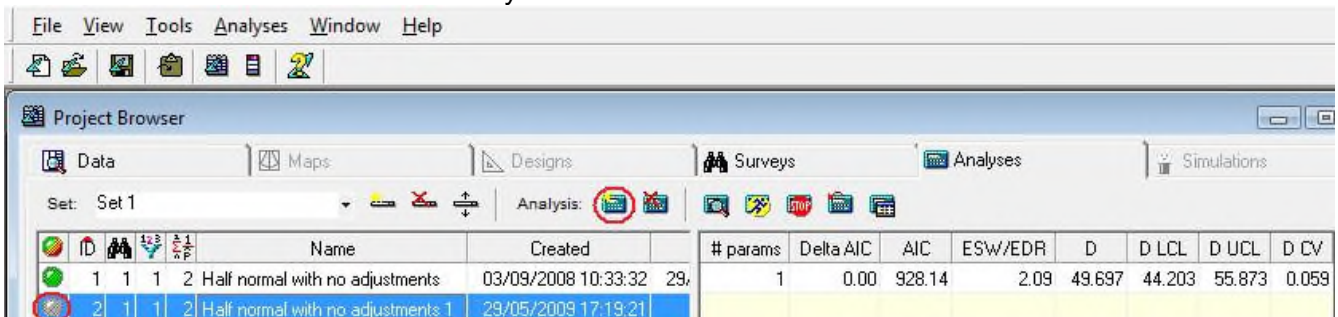


- A grey status icon indicates that this analysis has yet to be run. Click on the **Run** button in order to run the analysis. The **Results** tab should turn green.
- Click on the **Results** tab to see the results, and use the **Next >** button to move through the pages of results, looking at each page and trying to relate the analysis given here to the one you did by hand. (**Note:** These are the analysis details (Inputs/Log/Results) for one analysis – you can resize this window so that you can view details from multiple analyses when you have more than one analysis to compare).

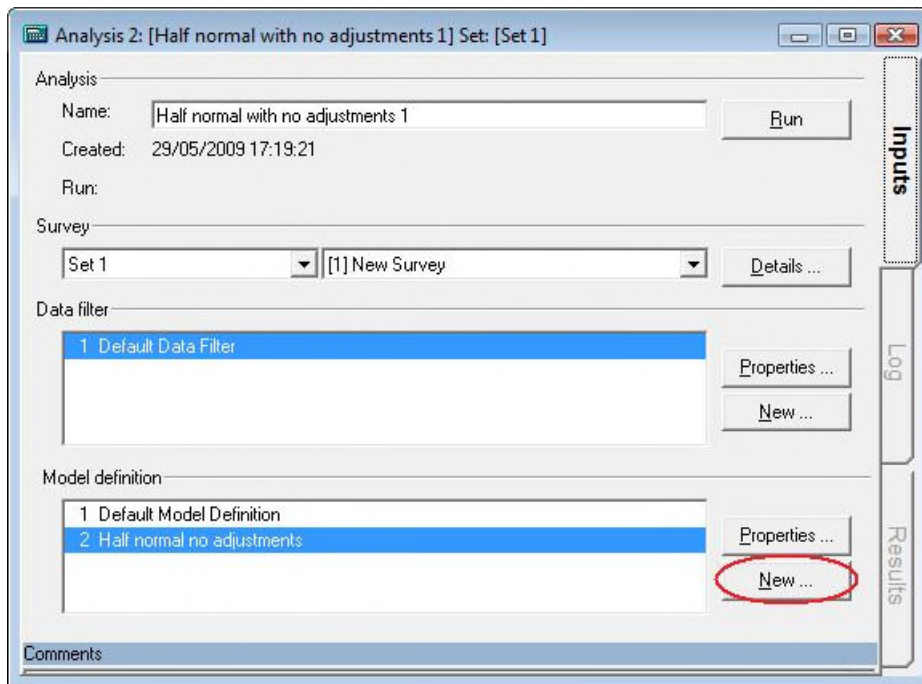


Creating a new analysis

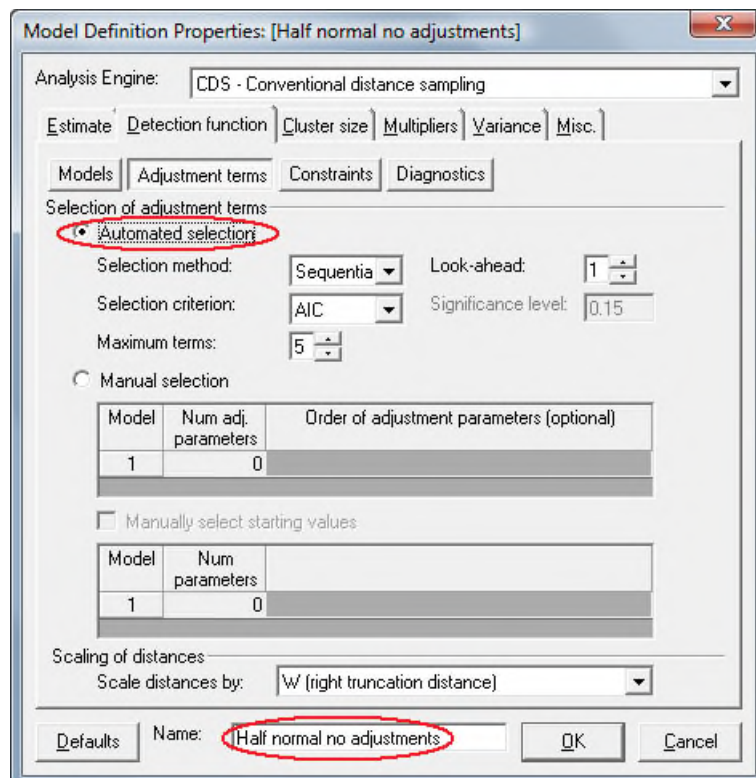
- Return to the Analysis Browser, and click on the first button after “Analysis:” on the Analysis Browser menu bar (“New Analysis”). Double-click on the status button to go to the Analysis Details window for this new analysis.



- Because the analysis is not run, you are taken to the **Inputs** tab. You will not need to edit the Survey or Data Filter for this example, but click on **New** in the **Model Definition** section.

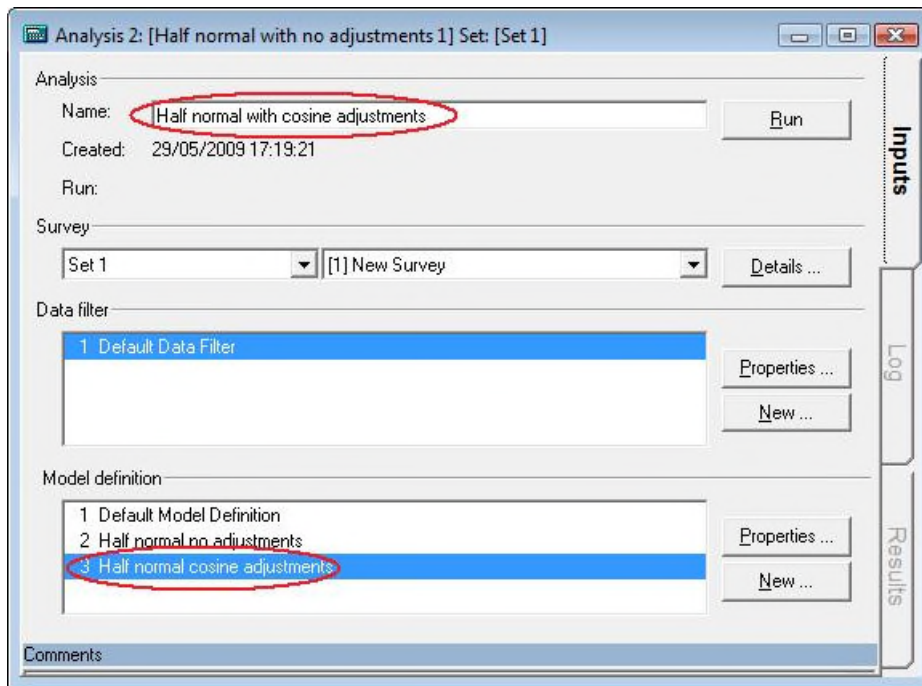


- Specify a half-normal key function with a cosine series adjustment, allowing selection of adjustment terms. When you have defined your new model, give it a suitable name (one that reflects the options you have set) and select **OK**.

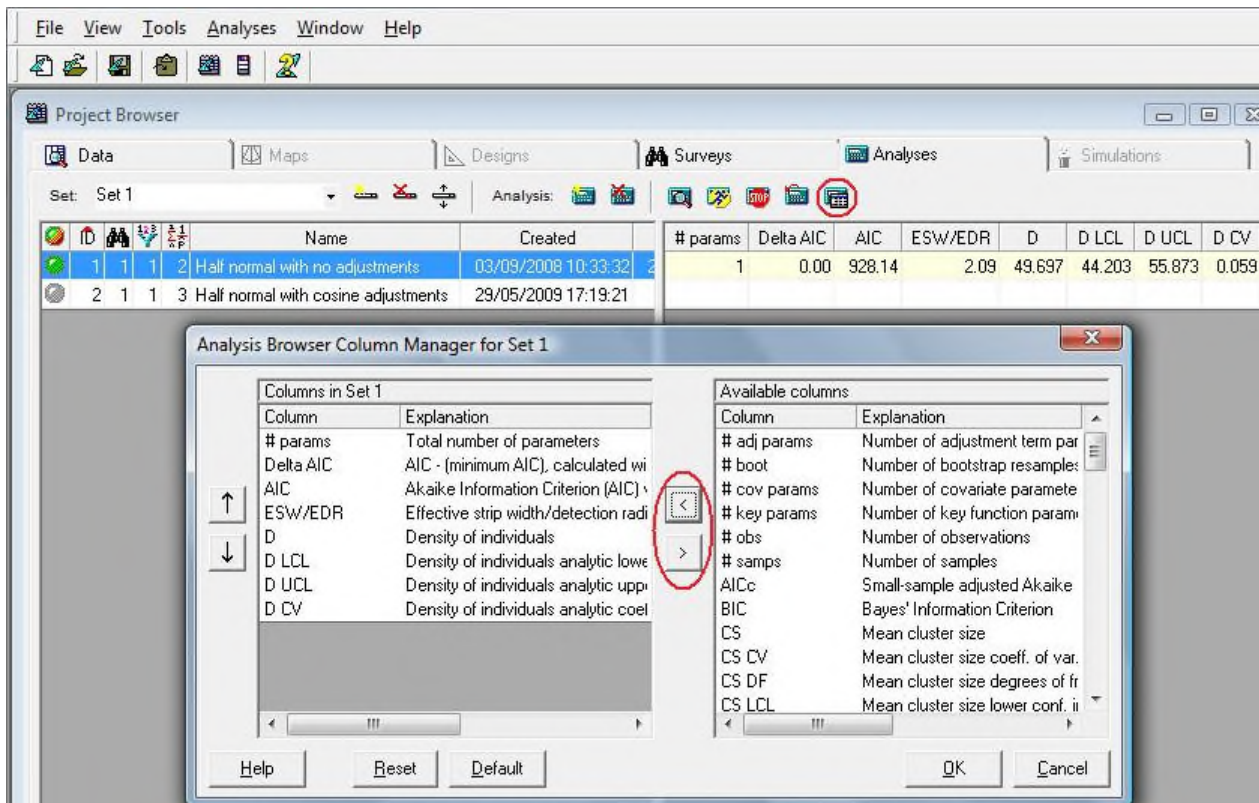


- Now give your analysis a suitable name, and click the run button. When the analysis finishes, it will automatically take you to the log tab if there were problems, or the results tab if the

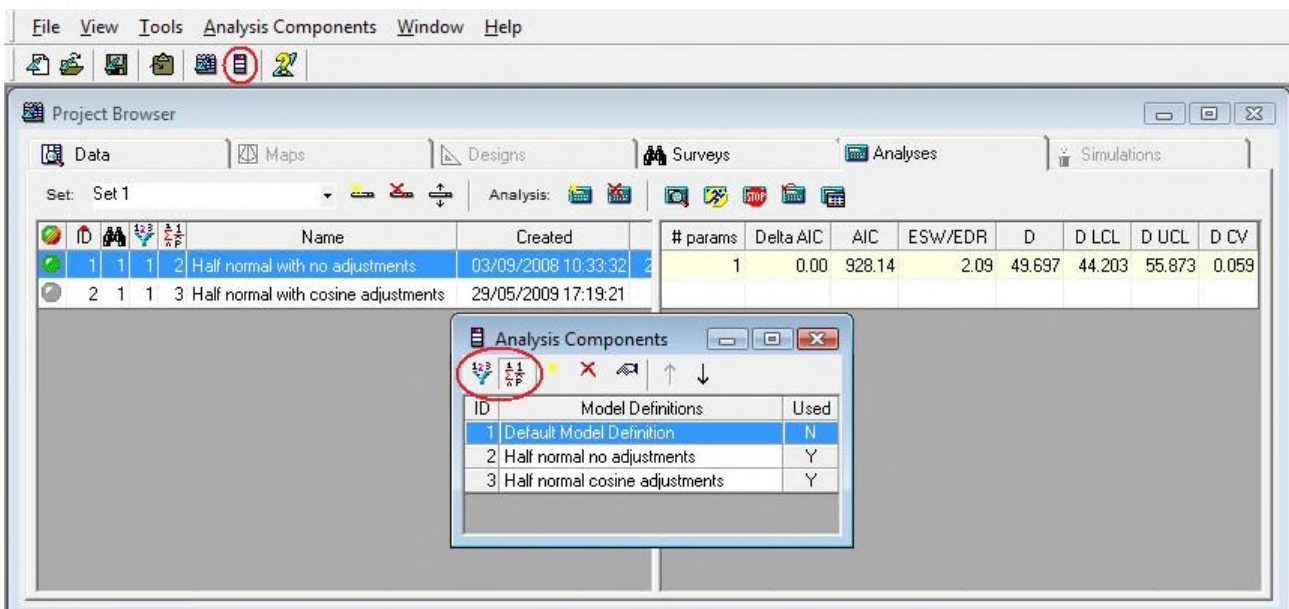
analysis ran without errors or warnings. From the results tab, you can investigate the result of your analysis.



- Create one more detection function model, this time specifying the hazard rate as key function, and Hermite polynomial as the adjustment. Compare the performance of the 3 models you fitted to this dataset. **Note:** when you create a new analysis (or model definition or data filter), Distance copies the settings from whichever analysis (or model definition or data filter) was highlighted at the time (the name is also copied). The default settings are not restored automatically.
- It is easiest to compare results from different analyses using the Analysis Browser. You can change the default columns in the browser using the **Column Manager** (furthest button on the right of the Analysis Browser menu bar).



- As you create more Data Filters and Model Definitions, you may find that you want to change their order, rename or delete them. A convenient way to do this is using the **Analysis Components** window – click the 6th button from the right on the main menu bar (“View Analysis Components”). In the Analysis Components window, clicking the first button lists the Data Filters and clicking the second button lists the Model Definitions.

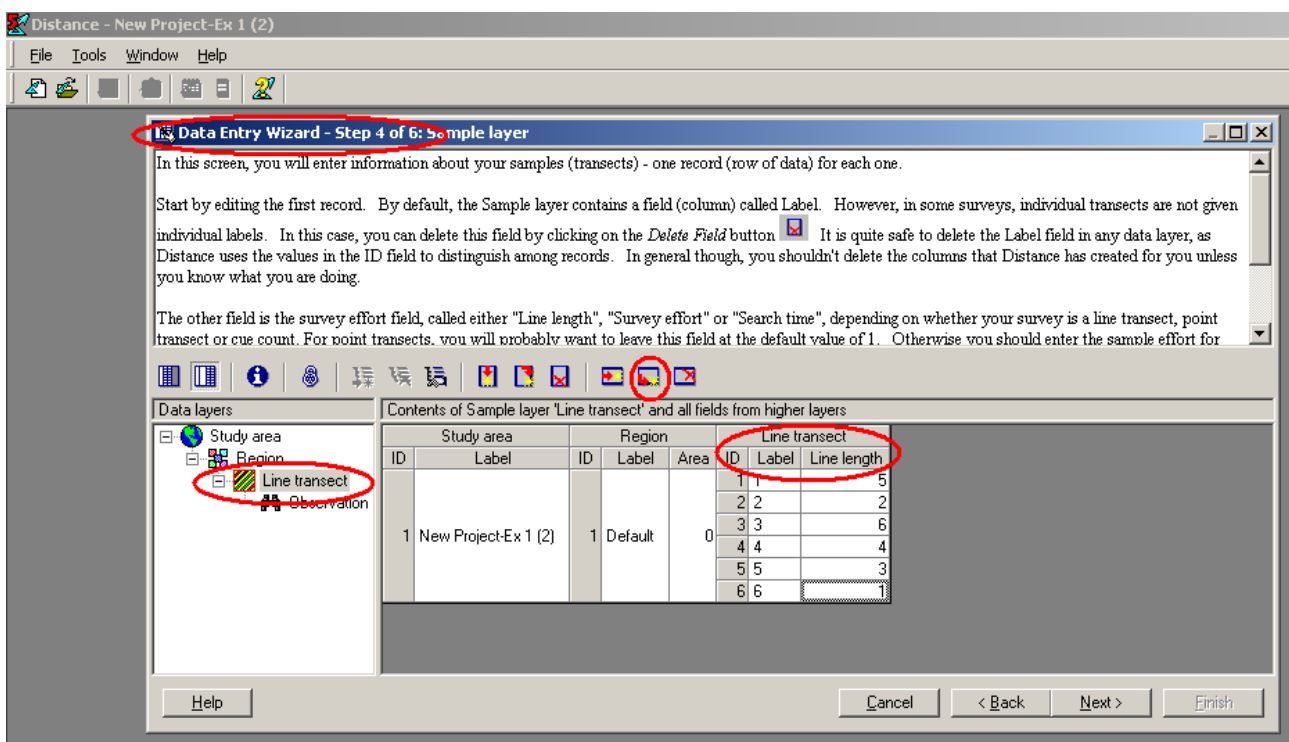


Introduction to Distance Sampling

Exercise 3: Line transect estimation using Distance

1(a). The line transect data immediately below were generated from a half-normal model.

- Open a new project (click on **File** then on **New project ...**), name it, and click on **Create**. Step through the New Project Setup Wizard (you should not need to change any of the defaults, except the units for density estimates to km² not the default hectares, but study each page) and click on **Finish**. This takes you to the Data Entry Wizard. Click **Next** until you get to the “line transect” page Step 4 of 6: Sample layer. Enter say the first 6 line labels (e.g. “line 1”, “line 2”, ...) and lengths (5, 2, ...). You need to click on the “append new record after current” button on the menu bar or type CTRL + Enter together before entering the information for each line.



- When you have finished, click on **Next** and enter the distances corresponding to each observation in a similar fashion (using CTRL + Enter between each observation). Once you have entered the distance data, go to the analysis browser, and carry out an analysis of these data using the half-normal detection function key.

Perpendicular distances in metres generated from a half-normal line transect model.

Line 1; length 5km
 7.9 10.2 12.4 3.8 4.8 8.5 13.4 5.8 7.5 11.5
 0.9 9.2 12.5 6.1
 Line 2; length 2km
 9.1 6.4 21.2
 Line 3; length 6km
 3.8 12.6 4.7 17.9 14.5 5.1 4.2 3.6
 Line 4; length 4km
 11.2 12.2 1.8 35.8 2.6 6.2 9.7 4.0 9.7
 Line 5; length 3km

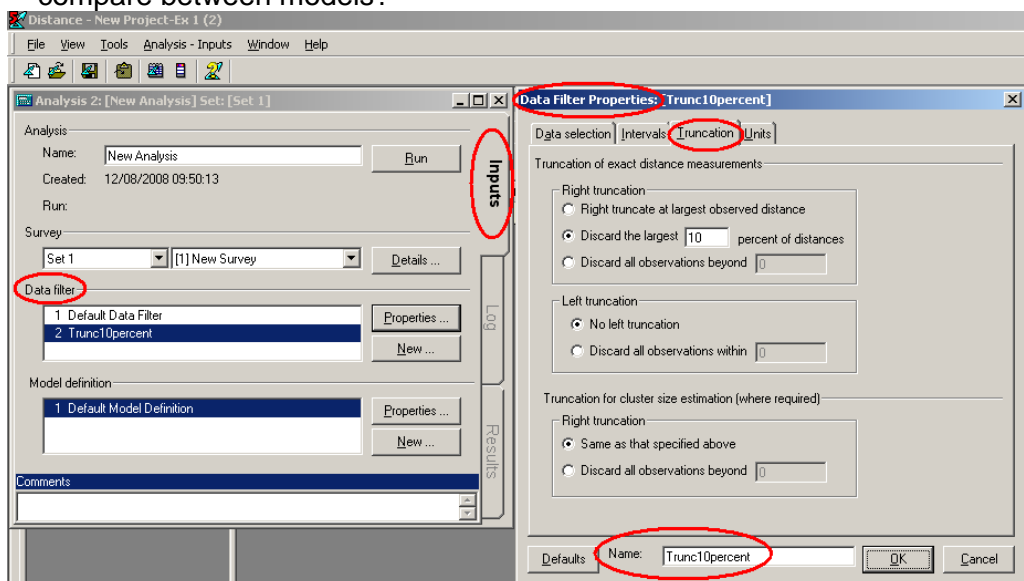
```

6.9 5.1 3.3
Line 6; length 1km
6.0 18.4 3.8 2.9
Line 7; length 4km
3.3 2.9 3.7 13.2 1.0 2.3 13.4 16.2 3.8 19.3
11.1
Line 8; length 4km
0.8 1.5 0.7 10.2 10.0 0.6 7.6 4.4
Line 9; length 5km
1.0 1.0 1.2 4.6 9.2 15.8 1.9 3.3 3.7 5.8
5.9 4.8 12.4 7.6 10.6 17.8 5.8
Line 10; length 7km
0.0 0.6 2.0 6.9 7.2 7.7 10.2 1.3 1.7 8.4
13.4 19.4 12.8 13.2 6.3 10.0 12.4 19.5 1.7 3.1
3.3 19.4 16.6
Line 11; length 3km
no detections
Line 12; length 4km
1.0 6.6 12.4 4.9 15.4

```

(b) The full data set is in project **Exercise3.zip** Choose Open project and select zip file type.

- Experiment with keys other than the half-normal (uniform, hazard-rate and negative exponential), to assess whether these data can be satisfactorily analysed using the wrong model.
- For each key, determine a suitable truncation point, and decide on whether, and which, adjustments are needed. Truncation points come under the data filter – click **New...** in the **Data Filter** section and create and name your own data filter, including truncation. In the example data filter below, the largest 10% of distances were truncated – you may want to truncate at a specific distance, depending on the data.
- Given that the true density was 79.8 animals / km² for these data, how do bias and precision compare between models?



Additional question

2. Below are perpendicular distance data (m) from line transect surveys of capercaillie (a large grouse) in Scotland. Total line length was 240km. The data are also in a text file **capercaillie.txt** in the Distance project directory. In the text file, column 1 is the transect number, column 2 is the transect length and column 3 is perpendicular distance. Columns are separated by tab characters. Create a new Distance project and either enter the data by hand or use the **Data Import Wizard** (Tools > Import Data Wizard) to import the data from the text file. Then decide on a suitable model for the detection function and estimate bird density.

CAPERCAILLIE, MONAUGHTY FOREST

n=112

28.0	17.0	15.0	14.0	18.0	0.0	38.0	6.0	50.0	65.0
75.0	1.0	70.0	28.0	40.0	40.0	40.0	15.0	40.0	30.0
5.0	55.0	60.0	40.0	24.0	30.0	0.0	50.0	55.0	10.0
40.0	10.0	30.0	34.0	24.0	30.0	15.0	20.0	14.0	48.0
0.0	30.0	2.0	52.0	11.0	48.0	28.0	38.0	25.0	35.0
45.0	0.0	16.0	12.0	2.0	14.0	12.0	24.0	70.0	50.0
49.0	40.0	80.0	18.0	27.0	30.0	30.0	60.0	58.0	14.0
0.0	56.0	40.0	19.0	21.0	0.0	38.0	20.0	28.0	30.0
20.0	16.0	0.0	69.0	40.0	46.0	50.0	40.0	70.0	67.0
28.0	12.0	12.0	22.0	40.0	48.0	48.0	15.0	12.0	0.0
15.0	20.0	17.0	30.0	30.0	32.0	48.0	20.0	10.0	20.0
42.0	30.0								

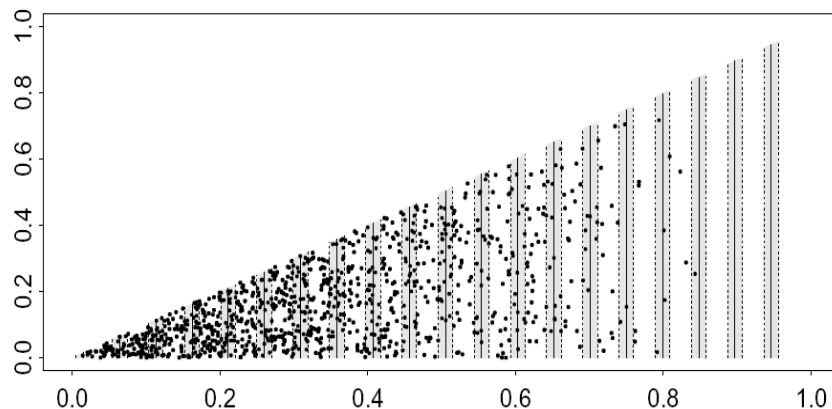
Introduction to Distance Sampling

Exercise 4: Variance estimation for systematic designs, the bootstrap method

In the lecture describing measures of precision we explained that systematic survey designs usually have the best variance properties, but obtaining good estimates of the variance is a difficult problem for statisticians. In this exercise we give an example of a situation where the systematic design gives a density estimate with much better precision than a random design. This means that the usual variance estimators used in Distance, which are based on a random design, give variance estimates that are far too high. The true variance is low, but the estimated variance is high. We will see how to implement a post-stratification scheme that enables us to get a better estimate of the variance. We also look at another case to see that the unstratified variance estimates provided by Distance are usually fine for a systematic design: things only go wrong when there are very strong trends in animal density, especially when the strong trends are associated with changes in line length (e.g. the highest densities always occur on the shortest lines, or vice versa).

We begin with the population and survey shown below. All the populations in this exercise are simulated on a computer: they are not real data. Note the characteristics: extreme trends with very high density on short lines and very low density on long lines. Additionally, the systematic design (search strips are shaded) covers a fairly large portion of the survey area. These are the danger signals that the usual Distance variance estimators might not work well, and a post-stratification scheme should be considered.

The survey region is a triangle, with dimensions 1km by 1km. The systematically placed search



strips are shaded above.

Basic variance estimation, with bootstrapping

1. Open the Distance project *Systematic_variance_2.zip*.
2. On the Analysis tab, click New Analysis. Rename it *Without post-stratification*.
3. Under Model definition, click Properties. Rename the new model: *No_adjustments_plus_bootstrap*.
4. Click the tab for Detection function, and click Adjustment terms. Select Manual selection so that no adjustment terms are fitted. Select the Constraints button, and click No constraints. These options will reduce the work that has to be done during bootstrapping.
5. Click the tab for Variance, and check the box for Bootstrap variance estimate: Select non-parametric bootstrap. The box Resample samples should be checked (this means resample transect lines). Leave the other settings at default, noting that there will be 999 bootstrap resamples conducted.
6. Click OK and then run the model. You can see the progress of the bootstrap in the bar at the top. Wait a few moments until the bootstrapping is completed.

7. Your analytic output should look like this:

	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	2044.6	27.70	20.74	1161.0	3600.6
N	1022.0	27.70	20.74	581.00	1800.0

- Because we have simulated these data, we know what the true values are. The true number of animals in the survey region is $N=1000$, and the true density is $D=2000 \text{ km}^{-2}$ (1000 animals in an area of size $A=0.5 \text{ km}^2$). The point estimates are good, but what do you think about the precision in the output above?
- Find the bootstrapped confidence intervals for D and N, and check whether they are similar to the confidence intervals above.
- What percentage of the total density variance is attributed to encounter rate estimation and what percentage to the detection function estimation?

Variance estimation for systematic designs using post-stratification

Recall we have a particular situation in which we have systematically placed transects, unequal in length. Furthermore there exists an east-west gradient in animal density juxtaposed such that the shortest lines are those that pass through the portion of the study area with the highest density. We examine a means by which we can use post-stratification to produce a better estimate of the variance in estimated density.

Post-stratification to improve variance estimation

The estimation of encounter rate variance in Exercise 4 used estimators that assumed the transect lines were randomly placed throughout the triangular region. In our case, the transects were not random, but systematic. In some circumstances, this can reduce the encounter rate variance a great deal. The data we are working with is an example of this. There are very high densities on the very shortest lines. In samples of lines collected using a completely random design, the sample by chance might not contain any of these very short lines, or it might contain several. The variance is therefore very high, because the density estimates will be greatly affected by how many lines fall into the short-line / high-density region: we will get very low density estimates if there are no short lines, but very high density estimates if there are several short lines. By contrast, in a systematic sample, we cover the region methodically and we will always get nearly the same number of lines falling in the high density region. The systematic density variance is therefore much lower than the random placement density variance.

Although there is no way of getting a variance estimate that is exactly unbiased for a systematic sample¹, we can greatly improve on the random-based estimate by using a post-stratification scheme. This works by grouping together pairs of adjacent lines from the systematic sample. Each pair of adjacent lines is grouped into a stratum. The strata will improve variance estimation, because the systematic sample behaves more like a stratified sample than a random sample.

Follow the steps below.

- Open the Distance project we used in the previous section (**Systematic_variance_2.dst**; it has the ".dst" extension because you uncompressed it minutes ago).
- Click the Analyses tab, and click the "New analysis" button to create a new analysis. Double click the grey ball and the Analysis Details Window should come up. Name the new

¹ because it is effectively a sample of size 1 – only the first line position was randomly chosen, and the rest followed on deterministically from there.

analysis something like *With post stratification*.

3. Under Model Definition, click New. Change the name at the bottom of the dialogue box to *Poststratified_no_adjustments_no_bootstrap*. (We don't want to conduct a bootstrap for our poststratified data, because it would involve some extra confusion and is not necessary.) In the Variance tab, click Advanced..., and select the option "Post-stratify, grouping adjacent pairs of samplers". Un-tick the option "Select non-parametric bootstrap".
4. Click OK and then Run to run the analysis. How does the variance and confidence limits compare with those you obtained in the previous section? What are the implications? Note what percentage of the overall variance now comes from encounter rate and from estimating the detection function, and compare this with the earlier percentages.
5. Now try the overlapping post-stratification option. A simulation study in Fewster et al. (2009) concluded that its performance was very similar to, but marginally better than the regular post-stratification. When the sample size of lines is small, it gives more post-strata and so is to be preferred for that reason. Create a new analysis, called say *With overlapping post stratification*, and then a new Model Definition for that analysis, in which you choose the Advanced variance option "Post-stratify, with overlapping strata made up of adjacent samplers". How does the variance compare with those you previous obtained? How do the degrees of freedom in the Estimation Summary – Encounter Rate page of output compare with that from the previous question?
6. (Optional) If you wish, you can try manual post-stratification. This is good practice if you need to do post-stratification for point transect studies. In this case you will have to add a new field to the sample layer, and then set up a new model definition in which you tell Distance to use post-stratification. Here goes:
 - a) Click the Data tab. Click the padlock button on the toolbar to unlock the data sheet for modification.
 - b) On the left-hand outline, click Line transect. The data sheet expands to 20 rows, each row corresponding to one line transect. This is the best format for the data sheet to be in when entering a new stratum number for each transect.
 - c) Click on the cell corresponding to Line transect Label 1. Several buttons on the tool-bar should become live. Click on the button corresponding to *Append field after current*. (The button has an arrow pointing sideways then downwards.)
 - d) You are prompted for Field name: enter VarGroup to indicate that you are grouping lines together for the purpose of variance estimation. Click Field type: Integer, and click OK.
 - e) You can now enter the line groupings for post-stratified variance estimation. Enter label 1 for lines 1 and 2; label 2 for lines 3 and 4; label 3 for lines 5 and 6; and so on, to finish with label 10 for lines 19 and 20. You have now defined 10 strata, each containing two adjacent transect lines from the systematic sample of lines.
 - f) After entering the column of VarGroup labels, click the padlock button again to lock the data sheet.
 - g) Now we will analyse the post-stratified data. Click the Analyses tab. Create a new analysis with a suitable name - .e.g, *Manual post stratification*
 - h) Create a new Model definition, with a suitable name. In the Estimate tab, click the button for Poststratify. Select Layer type: sample, and Field name: VarGroup. This means that we want to poststratify at the sample (transect) level, using our newly defined groupings VarGroup to delimit the strata.
 - i) For the levels of resolution, select the following:
 - Density: Global *and* Stratum
 - Encounter Rate: Stratum only
 - Detection function: Global only

- Cluster size (not required): Global only

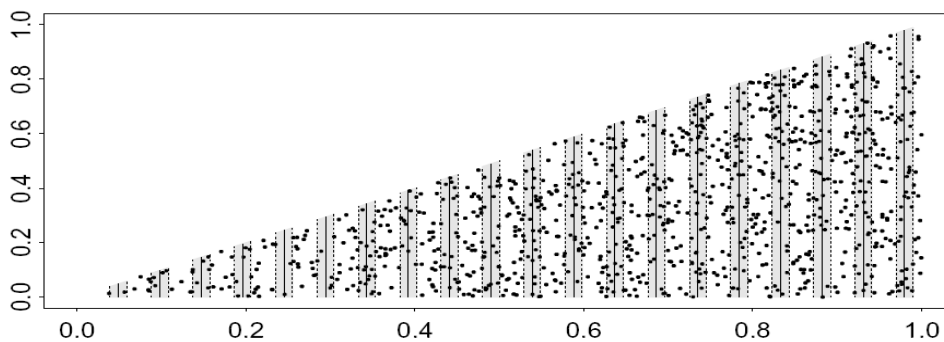
These settings ensure that it is only encounter rate variance that is affected by the post-stratification scheme; the detection function is still pooled over all observations as before.

- In the next field, enter Global density estimate is *Mean* of stratum estimates, and in the next field select *Weighted by Total effort in stratum*. Do **not** tick the box saying *Strata are Replicates*.
- Click OK and run the new model. The point estimates should be the same as the previous non-overlapping post stratification run.

Note: The precision of D and N are greatly improved in the post-stratified analyses. Note that we are not getting something for nothing: the second analysis is giving us an answer much closer to the true answer, while the first analysis was simply giving us the wrong answer. We have not *changed* the true variance by our post-stratification scheme: we are just getting a better *estimate* of the true variance. Because the data above were generated by simulation, we can use repeated simulated surveys to check that the second answer is indeed close to the true density variance over the repeats.

Systematic designs where post-stratification is not needed

The following simulated population does not exhibit strong trends across the survey region. Otherwise, the strip dimensions and systematic design are the same as for the previous example.



Open the project **Systematic_variance_1.zip**. Add the new data column `VarGroup` before conducting any analyses this time. With the augmented data, repeat the analyses you performed on the `Systematic_variance_2.zip` project. Find the relevant outputs. Has the post-stratification scheme been necessary in this case?

Introduction to Distance Sampling

Exercise 5: Point transect exercises

1. Simulated point transect data from 30 points are given in project **PTEExercise1.zip**. These data were generated from a half-normal detection function, and true density was 79.6 animals / ha. Experiment with keys other than the half-normal (uniform, hazard-rate and negative exponential), to assess whether these data can be satisfactorily analysed using the wrong model. For each key, determine a suitable truncation point, and decide on whether, and which, adjustments are needed. (Truncation points come under the data filter.) How do bias and precision compare between models?
2. The projects **Wren1.zip**, **Wren2.zip**, **Wren3.zip** and **Wren4.zip** contain winter wren data, collected at Montrave, Scotland in 2004. Each project corresponds to a different method of data collection. Thirty-two points were defined through 33.2 ha of parkland (Fig. 1), and detection distances were measured in metres with the aid of a laser rangefinder. Three types of point transect data were collected: 1. standard five-minute counts; 2. the 'snapshot' method; and 3. a cue count method. In addition, line transect data were collected (method 4), and territory mapping was conducted, which gave an estimate of 43 wren territories (1.30 territories ha⁻¹).
 - a) Select a single model for exploratory data analysis. Experiment with different truncation distances w , and select a suitable value for each method. Do you see potential problems with any of the data sets?
 - b) Try other models and other model options. Use plots, AIC values and goodness-of-fit test statistics to determine an adequate model.
 - c) Record your estimates of density for each method. Record also the corresponding confidence intervals. Compare your answers with those of others in the workshop.

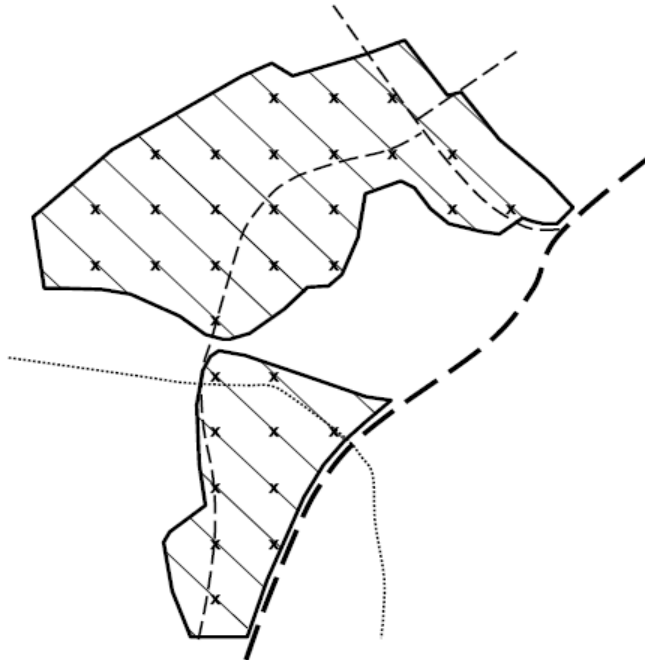


Figure 1: The study site at Montrave in Fife, Scotland. The dotted line is a small stream, the thin dashed lines are tracks, and the thick dashed line a main road. The 32 points are shown by crosses, and are laid out on a systematic grid with 100m separation. The diagonal lines are the transects used for method 4.

3. The Sample Projects directory contains two point transect projects, **Savannah Sparrow 1980.zip** and **Savannah Sparrow 1981.zip**. These were part of a large data set collected in Arapaho National Wildlife Refuge, Colorado. For both data sets, consider an appropriate truncation distance, decide on a suitable model for the detection function, and estimate density, both for each stratum individually and for the whole study area. You should include in your analysis an assessment of whether the detection function can be estimated from data pooled across strata, or whether separate estimates are needed per stratum. (This will be covered in the lecture discussing stratification if you don't already know how to do it.)

Introduction to Distance Sampling

Exercise 6: Automated Survey Design Exercises

1. Point transect survey of North-eastern Mexico

Reviewing the data

Extract and open the project **MexicoUnPrj** from the archive **MexicoUnPrj.zip**. This project contains data from 4 states in North-eastern Mexico. Let's begin by reviewing the data. On the top menu-bar select **File, Project Properties**. The **General** tab gives you information about the location of the project file and its associated data folder (MexicoUnPrj.dat). The **Geographic** tab gives you information about the default geo-coordinate system of the geographic data, and the default map projection. The geo-coordinate system is used to locate the geographic data (which is stored in decimal degrees of latitude and longitude) on the earth's surface. The projection is used to convert these data from the curved surface of the earth into a flat plane that can be used for displaying maps and designing surveys. The resulting projection has linear units, such as metres or kilometres. If you are planning a survey that will take place over a small geographic area, and you are inputting your data by hand, then you don't need to worry about geo-coordinate systems or projections and can set both these options to [None]. In this example, however, the survey area is quite large and the projection chosen will make some difference to the results. Click **Cancel** to close the Project Properties window without saving any changes you have made.

Click on the **Data** tab of the **Project Browser** to view the **Data Explorer**. In the left-hand pane, under Data Layers, you can see that there are four layers in the project: "Mex", "MexStrat", "Grid1" and "Grid2". You can tell the layer types by looking at the icons beside the names: Mex is a Global layer, MexStrat is a Stratum layer and Grid1 and Grid2 are Coverage layers. When you open a new project, the Global layer is selected by default, so the layer Mex is now selected. Click on the **Data Layer Properties** button on this tab (7th button from left) to find out more about this layer. The **Layer Properties** window opens, and under the **Geographic data** tab, you can see that the geographic data is stored in a shapefile called Mex.shp in the data folder, and that the shapes in this layer are Polygons (i.e. solid shapes). Click **Cancel** to return to the Data Explorer.

In the right-hand pane of the Data Explorer, you can see its' fields: ID, Label, Area and Shape. There is one record, with ID = 1. The Shape field holds the geographic information for that record. Because this layer holds polygons, the shape record has the word "Polygon" in it. Double click on this word to open the **Shape Properties** window. This is where you edit the geographic information inside Distance (an alternative is to edit the shapefile Mex.shp from outside of distance, using a GIS package such as ArcGIS). Click **Cancel** to return to the Data Explorer.

The coverage layers Grid1 and Grid2 contain a grid of points that will be used for determining probability of coverage for our survey designs. If you click on "Grid 1" in the left-hand pane of the Data Explorer, its records open in the right-hand pane, and you can see that it has 177 records. A better way to look at the grid points is to view them in a map. Click on the **Maps** tab in the **Project Browser**. Click on the **New Map** button (3rd button along) to create a new map. Double-click on the words "New Map" to edit the name of the map, and call it "Grid 1". To view the map, click the **View Map** button (5th button along), or double click on the map's ID. A **Map Window** opens.

The map starts life blank. You add layers to the map by clicking the **Add Layer to Map** button (7th button along). Click this button and select "Mex" from the list of layers. Then click the button **Add Layer** button again and select "Grid 1" from the list. Now you can see the grid points.

You could also add the points from Grid 2 to the same map. If you do this, you will see that the grid points for Grid 2 are much closer together than those for Grid 1. (Grid 2 was generated with a spacing of 20 km, while for Grid 1 the spacing was 50km.) The points for Grid 2 obscure those from Grid 1 – you can change the order of the map layers by clicking on a the legend "Grid 2" in the left-hand pane of the map and holding the mouse button down while you drag it down to below "Grid 1". Click the [X] button in the top right corner of the Map Window to close the map. (Say "Yes" if it asks you to save changes.)

In the **Data Explorer**, click on the MexStrat layer to see those data. You can see that there are 4 strata. If you want to see where they are, you could create a new map in the **Maps** tab and add the MexStrat layer to the map. If there are layers on the new map that you don't want, you can remove them with the **Delete Selected Layer** button in the **Map** window.

When you've finished exploring the data, move on to create a new design.

Creating a new design

Click on the **Designs** tab of the **Project Browser**. To create a new design, click the **New Design** button (1st one after the word "Design:"). A new record appears in the left-hand pane, called "New Design". Double click on the name, and edit it to call the new design "150 random points". If you need more room, click and drag to the right the vertical splitter that divides the Designs window into two. Click the **Show Details** button (3rd one after the word "Design") to open the **Design Details** window. Look under "Type of design" to see the sampler and design class; the default sampler is "Point" and the default design class is "Simple Random Sampling". Click the **Properties** button to set the properties for this design. The **Design Properties** window opens. The options you see on the design properties tabs depend on the type of design. In this example, choose the following options:

- Under Stratum layer, choose the stratum layer "MexStrat".
- Under Design coordinate system, make sure the box "Same coordinate system as stratum" is unchecked. The projection should say "Plate Carree" and the units "Metre".
- In the **Effort Allocation** tab, under Edge Sampling select the "Plus" option. Uncheck the box "Same effort for all strata". A list of the four strata in the MexStrat layer appears. Under "Allocation by stratum", click the "Percentage from" radio button, and enter "150" as the number of points. In this example, we will put most of our effort into the two Baja strata (perhaps because this is where we think most of the animals of interest live). Under "Effort %" enter 10 for Sinaloa, 10 for Sonora, 40 for Baja Sur (south) and 40 for Baja Norte (north).
- In the **Sampler** tab, select Kilometre for the point sampler radius units. Let's imagine we're surveying for a very vocal species and that our truncation distance will be 5km, so we enter 5 under radius (for this example we'll assume same sampler properties for all strata).
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. This is far too few for an accurate simulation, but will do for the purposes of demonstration. Under grid layer, choose "Grid 2", which is the one with the grid points closer together.

Now click OK to close the **Design Properties** window. The properties window closes and we are back with the design details.

Automated generation of new surveys

Click the **Run** button on the **Design Details** window. A window pops up offering you two choices: (1) Calculate coverage probability statistics, and (2) Generate a new Survey. Choose the second option, and give the new Survey a useful name like "150 points survey" and the new layer a name like "150 points". Then click OK. A **Survey Details** window opens, and the status bar at the top of the Distance window says "Running Survey". At this point you have to be patient while the survey runs. Distance is creating a set of randomly located survey points, based on the design. When it is finished, the **Survey Details Results** tab opens, and you can review some statistics about the new survey. Click the "Next >" button to see a map of the points – you should be able to see that there are more in the

Western strata (Baja) than the eastern. Click "Next>" again to see a list of the points, with latitude and longitude for each. (You could, for example, use this to make a survey plan to take into the field. To copy this text to another file, press the "Copy current window" button, 4th from the left on the top toolbar. Then open, say, a Word document and click Paste to copy it there. You can also copy the map of points by displaying the map and pressing the copy button, or choosing the menu item Survey – Results | Copy Map to Clipboard)

Click on [X] to close the **Survey Details** window, and click on the **Surveys** tab of the project browser. You can see that your new survey has been added there. If you select it and click the "Show Details" button (3rd from left after the word "Survey") you get back to the **Survey Details** window **Results** tab. Click on the **Inputs** tab and then **Properties ...** button. Under **Data Layers**, you can see that the new Sample data layer "150 points" has been entered as the lowest sample layer. Close the **Survey Properties** and **Survey Details** windows, and click on the **Data** tab of the **Project Browser**. You can see that the new sample data layer "150 points" has been added below the "MexStrat" data layer.

Design statistics

Go back to the **Design Details** window for your design, and click **Run** again. This time, choose the top option (Calculate probability of coverage statistics) and click OK. You have to wait while Distance generates multiple simulated surveys and uses these to work out the probability that each grid point will be covered by the survey. When it has finished, you can see the results in the **Results** tab, and a map of coverage probability by pressing the "Next >" button. In theory, this design should produce an even probability of coverage within stratum. However, you can see that there is considerable variation. Why is this? What would happen if you repeated the run with more simulation runs (say 500, or 1000)? (Before you spend a lot of time running simulations with this project, read the next section.)

Working with projected raw data

There is a technical problem with the way the geographic data are stored in MexicoUnPrj. Each time you view a map or run a survey design, the data have to be projected from the latitude and longitude format in which they are stored using the projection you have specified (Plate Caree in this case). This takes some computer time, so if you're doing lots of survey design work there's a trick to make things more efficient. The trick involves projecting the raw data files.

We used ArcGIS to project the shapefiles in MexicoUnPrj using the Plate Caree projection, and stored this new data in the project MexicoPrj. So rather than being stored in latitude and longitude, the data in MexicoPrj is stored in meters. Run a second instance of Distance, and then extract and open the project **MexicoPrj**. Look under the **Project Properties**, and you will see that the GeoCoordinate system and Projection are both set to [None], and that the units are meters. So, we've projected the raw data, and so long as we project all the data layers the same way we don't need to tell Distance anything about the coordinate systems used.

As a check that the data really are projected, go the Data Explorer and double-click on the global layer's Polygon record. The first value is something like $x = -12594701$ $y = 3230255$ – this gives the number of meters of that point on the polygon from some reference point on the earth. If you do the same thing in the MexicoUnPrj project, you'll see that the first value is something like $x = -113$ $y = 29$, which is the latitude and longitude of that point.

If you're going to do lots of experimenting with the Mexico data this evening, or at home, it's better to use the MexicoPrj project, as you'll find the probability of coverage simulations run quite a bit faster. Meanwhile, move on to the next exercise.

2. Entering geographic data into Distance, and generating Coverage grids

The purpose of this exercise is to show you how to enter geographic data by hand into Distance, and how to generate Coverage grid layers.

Create a new project and enter data

On the top menu-bar select **File, New Project** (or click the toolbar button). In the **Create New Project** dialog box give it the File Name "Trapezium" and then click on **Create**. The new project setup wizard starts up. Under "I want to", select the option to "design a new survey", and click **Next**. Then click **Finish**.

The **Project Browser** will open up, showing the **Data** tab. Click on the menu **File | Project properties**, and look under the **Geographic** tab to confirm that there is no geographic coordinate system for this project (i.e. non-earth referenced), and that the default units are metres. Click **OK** to close the **Project Properties** window.

In the **Data** tab of the **Project Browser**, you can see that Distance has created a global data layer called Study Area, with default fields ID, Label and Shape. Double click on the word "Polygon" to open the **Shape Properties** window to edit the new survey region. Click on the **Insert Point** button 4 times and fill in the following (X,Y) coordinates: (0,0), (0,100), (120,20) and (120,0). Click **OK** to return to the Data Explorer.

Generate a coverage grid layer

To generate a coverage grid layer click on the **Create New Data Layer** button (5th from left) in the **Data** tab of the **Project Browser**. Enter "TrapGrid" as your Layer Name and set the Parent Layer to "Study Area" and the Layer Type to "Coverage". You should now be able to click on the **Properties...** button. In the **Grid Properties** that pops up set the "Distance between grid points" to 2.5 and the "Units of distance" to "Metre". Once you press **OK** you should proceed to add the grid points to the layer. This may take a few moments.

Create a new map on the **Map** tab of the **Project Browser** and add your new global and coverage layers to take a look at them.

Creating a new design

Click on the **Designs** tab of the **Project Browser** and then the **New Design** button. Rename your design "equal angle zigzag" and then click the **Show Details** button to open the **Design Details** window. Select the "Line" sampler and set the design class to "Equal Angle Zigzag". Click the **Properties** button to set the following properties for this design:

- As the Trapezium survey region is non-earth referenced you don't need to make any changes on the **General Properties** tab.
- In the **Effort Allocation** tab, under "Effort determined by" select the Sampler angle option. In the "Allocation by stratum" section set the Line length units to be Metres. Make sure the "Update effort in real time" check box is ticked. As there is only one survey stratum it does not matter whether the "Same effort for all strata" check box is ticked or not. The "Absolute values" radio button is the only one available when effort is determined by sampler angle. Enter 75 in the "Angle" (measured in degrees) column of the table. The "Length" column should now read 463.644. The accuracy of this approximation of zigzag length depends on the shape of the survey region, but should be relatively accurate for convex survey regions.
- In the **Sampler** tab, set the width to 1 meter.
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. Under grid layer, choose previously created "TrapGrid".

Click **OK** to close the **Design Properties** window and return to the design details.

Design statistics

Run your design to work out the coverage probabilities - this design will take a while to run! In the second page of the **Design Details Results** tab that opens when its finished, take a look at the coverage probabilities map. Note how uneven these probabilities are and how they increase as the trapezium height decreases for the equal angle zigzag design.

Additional investigations

If you are particularly interested in zigzag surveying, you might want to come back to this exercise after completing exercise 3, and compare the coverage probabilities of the three different types of zigzag designs. You can do this when you get home. For now, skip ahead to exercise 3.

For work on your own:

Create two new designs - one for the equal spaced zigzag and one for the adjusted angle zigzag. To facilitate comparisons, you want to set properties for both that are somewhat equivalent to those for the equal angle design. You can see the mean trackline length for the equal angle design in its Results tab (about 460 metres). You can then set the effort allocation for the two new designs to be the same as this value. Make sure that the Coverage probability tab shows "Estimate by simulation" and that you have an appropriate Grid Field Name.

Try creating a few surveys for each design, so you can see how they differ. Then run the coverage probability simulations. How do the coverage probabilities for the 3 designs differ? You may need more simulations to see a strong difference between the equal spacing and adjusted angle design.

3. *Systematic parallel line aerial survey of marine mammals in St Andrews bay*

Reviewing the data

Open the project archived in **StAndrews.zip**. This project contains the survey region for an aerial survey of porpoise, common dolphins and seals in and around St Andrews bay. (For locals: the nearer St Andrews bay region has been extended in an easterly direction out past bell rock, as there are some pockets of deeper water out there that are of interest with regard to the distribution of cetaceans. The survey region has a chunk missing due to a no-fly zone around Buddo Ness, just below Carnoustie). To take a look at the survey region create a new map in the **Maps** tab and add the layer **StAndrews** to the map.

The small survey plane permits a total flight time of approximately 250 km (excluding the flight time to and from the landing strip in Fife Ness, just down the coast). A systematic line sampling design is going to be used. The survey plane permits easy movement between survey lines, but it would still be efficient to spend as much of the 250 km flight time on effort surveying rather than on movement between the sampler lines. The aim of this exercise is to decide on a systematic line spacing that gives about 200 km on-effort trackline with the total trackline length constrained to 250 km. To do this create a number of systematic line sampling designs with different line spacings, generate the design statistics for these designs and then the statistics for the total trackline length to the on-effort trackline length for different designs.

Before proceeding to the design stage you need to generate a coverage grid layer, as this will be needed to generate design statistics.

Generate a coverage grid layer

To generate a coverage grid layer click on the **Data** tab of the **Project Browser** and then the **Create New Data Layer** button (5th from left). Enter "Grid5" as your Layer Name and set the Parent Layer to "StAndrews" and the Layer Type to "Coverage". You should now be able to click on the **Properties...** button. In the **Grid Properties** that pops up set the "Distance between grid points" to 5 and the "Units of distance" to "Kilometre". (This is too far apart for estimating probability of coverage, but we know coverage is even for this design, so choosing a wide spacing makes the simulations run faster.) Once you press **OK** you should proceed to add the grid points to the layer. This may take a few moments.

Now create and generate a couple of designs with a spacing of your choice (some suggested spacings include 4.5, 5, 5.5 & 6 km)

Creating a new design

Click on the **Designs** tab of the **Project Browser** and then the **New Design** button. Rename your "New Design" something like "systematic line test" and then click the **Show Details** button to open the **Design Details** window. Select the "Line" sampler and set the design class to "Systematic Random Sampling". Click the **Properties** button to set the following properties for this design:

- On the **General Properties** tab under Stratum layer, the StAndrews stratum layer should be selected. Under Design coordinate system, the design coordinate system should be "Non-earth referenced". (The data have already been projected from the OSGB 1936 geo-coordinate system using the transverse mercator projection – the same trick we used for the MexicoPrj project.)
- In the **Effort Allocation** tab, under Edge Sampling select the "Minus" option. In the "Allocation by stratum" section set the Line length units to be Kilometres. Make sure the "Update effort in real time" check box is ticked. As there is only one survey stratum it does not matter whether the "Same effort for all strata" check box is ticked or not. Click the "Systematic line spacing" radio button and enter the line spacing in the "Spacing" column of the table. When you enter a 5 km spacing for instance the "Length" column should then read 226.203 and the "Samplers" column 8. The accuracy of this approximation of on-effort line length and total number of line samplers depends on the shape of the survey region, but should at least give you some indication of what to expect.
- In the **Sampler** tab, select Kilometre for the line sampler width units. Set the truncation width to 2 km.
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. This is too few to give accurate coverage probabilities, but sufficient for the on-effort and total trackline length statistics. Under grid layer, choose previously created "Grid 5". Make sure the Grid field name is the same as your design name.

Click OK to close the **Design Properties** window and return to the design details.

Design statistics

For each design run Distance generates multiple simulated surveys and uses these to work out the statistics for on-effort and total trackline length. Run your designs and in the **Design Details Results** tab that opens review the statistics to decide on suitable systematic line spacing.

Automated generation of new surveys

To see an example survey, go back to the **Design Details** window for your selected design click **Run** again this time choosing the "Generate a new Survey" option. The second page of the survey results displays a map of the survey region with the systematic lines superimposed. You can add this map to the **Map browser** and manipulate it there by clicking on the 6th button on the Survey map results page.

Introduction to Distance Sampling

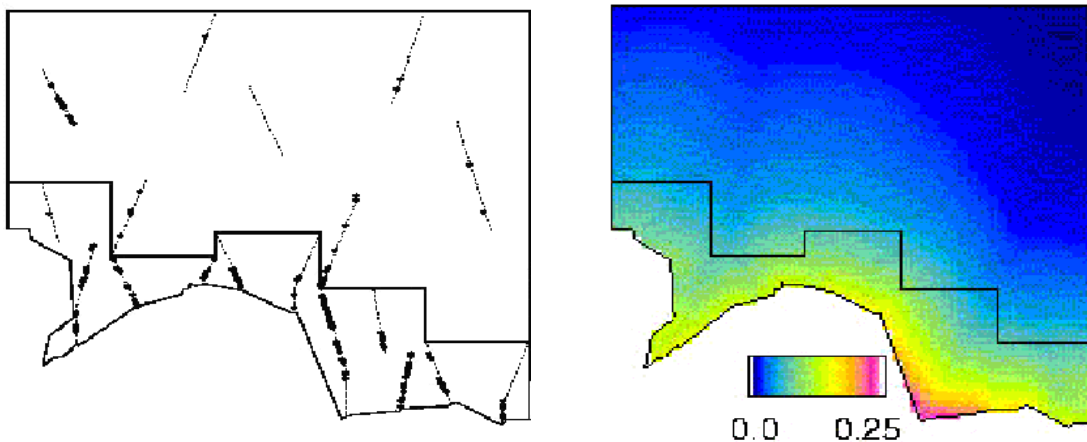
Exercise 7a: Analysis of Stratified Data

The Data

The Distance project **Stratify exercise** contains data from a stratified survey of Antarctic minke whales. The data are “exact” insofar as they are calculated directly from the estimates of radial distance and angle recorded by the observers. While angle boards and reticule binoculars were used for estimation of angles and distances when possible, the transitory nature of cues (usually blows) and the pitch and roll of the vessel, among other things, leads to in errors in estimating angles and distances. Angular errors are typically of the order of a degree or two; the coefficient of variation of distance estimation errors is typically of the order of 10%.

The two strata were surveyed by different vessels at the same time. Because the whales tend to be found in high densities against the ice edge, where they feed, densities in southern strata are typically higher than those in northern strata. In fact this is the primary reason for using a stratified survey design. It is also the reason for covering the southern strata more intensely; in this survey the transect length per unit area in the southern stratum, is more than 2.5 times that in the northern stratum.

Here are pictures of the sort of design used and a typical density gradient. The irregular bottom border is the ice-edge; the “steps” define the boundary between southern and northern strata; dotted lines are transects; solid dots are detections.



Analysis Exercises

Begin by opening the project from its archive **Stratify exercise.zip**. The project contains one analysis specification, called “Full geog stratification”. This is a fully stratified analysis of the data. Seven equal perpendicular distance intervals, truncation at 1.5 nautical miles (nm), and a hazard rate detection function form with no adjustment terms are used to estimate the detection function. As the focus of these exercises is stratification, do not investigate other perpendicular distance intervals and detection function forms; the given models are adequate. Use the **Analysis browser** to familiarise yourself with the details of this analysis specification.

1. Having done that, run the analysis “Full geog stratification”. Look at the results, and note the AIC statistics from each detection function fit.

2. To stratify $f(0)$ or not to stratify?: Create a new analysis identical to “Full geog stratification” by clicking the **New Analysis** icon in the **Analysis** tab of the **Project browser** after selecting the existing analysis. The new analysis will be a copy of the existing one.

Create a new model definition for this new analysis by going to the **Inputs** tab and highlighting the “haz rate+no adj full strat” model, then clicking the **New** tab. This will copy the existing model definition – modify the new model definition so that $f(0)$ is to be estimated from the pooled strata (click the **Detection function** x **Global** cell of the table on the **Estimate** tab of the **Model Definition Properties** window you get after clicking **New**). Give this new model definition a suitable name and then click **OK**.

Run the new analysis and look at the output. By comparing the AIC from this analysis with the sum of the AICs from the analysis “Full geog stratification”, and considering the fits of each detection function, decide whether or not to pool strata for estimation of $f(0)$.

If you have time, here’s a more difficult question.

3. Create an analysis without any stratification and estimate density using it. Why is the density estimate so much higher than those from 1. and 2. above?

Introduction to Distance Sampling

Exercise 7b: Analysis of Clustered Data

The Data

Cluster exercise.zip contains “exact” perpendicular distance and cluster size data from a survey of Antarctic minke whales (the same data as are in the project file stratify exercise.zip).

Open the **Cluster exercise.zip** project in Distance. Use the data explorer to familiarise yourself with the data (click the **Data** tab in the **Project Browser**, followed by the **Region**, then **Line Transect**, then **Observation** symbols in the left window). Ignore the “Cluster strat” data column for the moment, it is dealt with below.

Analysis Exercises

This exercise will allow you to explore some of the different methods of dealing with clustered data, as discussed in the lecture. The following methods will be used:

- Regression
- Truncation
- Post-stratification

The project contains one analysis specification, called “E(s) by ln(s)_g(x)”. Use the **Analysis browser** to familiarise yourself with the details of this analysis specification. This analysis uses a regression of the log of school size (s) against the estimated detection function to estimate mean school size (look under the **Cluster size** tab in **Model Definition Properties**). Seven equal perpendicular distance intervals, truncation at 1.5 nautical miles (nm), and a hazard rate detection function form with no adjustment terms are used to estimate the detection function. As the focus of these exercises is mean cluster size estimation, do not investigate other perpendicular distance intervals and detection function forms; the given models are adequate.

Using regression

- 1) Run the analysis “E(s) by ln(s)_g(x)”. Look at the results and the cluster size estimation pages in particular.
 - a) Is the regression method estimate of E(s) bigger than the observed mean cluster size?
 - b) What percentage of the variance of the density estimate is due to cluster size estimation?

Using truncation

- 2) Using the fitted detection function, decide on an appropriate point at which to truncate the data in order to use the mean observed cluster size as an estimate of E(s). Create a new analysis, identical to “E(s) by ln(s)_g(x)” except that it should use the truncation method to estimate E(s). To do this, click the “**New Analysis**” icon in the **Analysis browser** after selecting the existing analysis, then add a new **Data filter** in which the right truncation for cluster size estimation on the **Truncation** tab

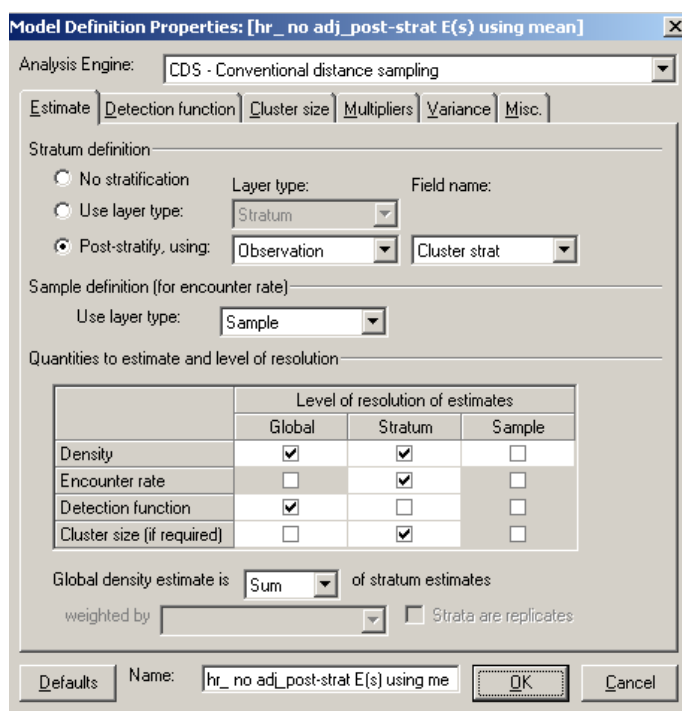
has been set appropriately. Create a new Model Definition where the mean of the observed clusters, rather than the regression method, is used (specified in **Model Definition Properties/Cluster size**). Having run the analysis, look at the cluster size estimation pages.

- a) Why is the “Mean cluster size” on the **Cluster size/Global/Estimates** page different from the mean cluster size in analysis 1 above?
- b) Why is the standard error of “Mean cluster size” in this analysis larger than that of the “Expected cluster size” in analysis 1 above? (Hint: look at the sample sizes.)

Using post-stratification by cluster size

3) Now we come to the “Cluster strat” column in the Observation layer of the data. It was added after the data were entered and is just an indicator column for stratification on the basis of cluster size. All observations with cluster size 1 have been defined to be in cluster stratum 1 and hence have 1 in the “Cluster strat” column. Similarly for cluster size 2. Due to small sample sizes it was not possible to create separate strata for cluster sizes of 3 and above. Therefore, all observations with size 3 or greater have been put in cluster stratum 3 and hence have 3 in the “Cluster strat” column.

- a) Use the “Cluster strat” column as a basis for performing an analysis post-stratified by cluster size. Do this by creating a new analysis with a new Model Definition that uses post-stratification at the Observation level. Fit a detection function pooled across strata, but estimate mean cluster size separately for each stratum (see the picture below for help). There should be no size bias within the strata, so theoretically it should be sufficient to use the mean of the observed cluster sizes for each stratum. Once the analysis is run, note the mean cluster size for the third stratum.



- b) However, when forced to use strata that contain a range of cluster sizes due to small sample sizes (such as stratum 3 in this case), you may suspect that size

bias is still present. It is possible to use the regression method to check this. Create another post stratified analysis which uses the regression method to estimate $E(s)$ in each stratum (again, estimate a pooled detection function and separate cluster size estimates). Compare the regression estimate of $E(s)$ with the mean cluster size (the mean should be identical to the estimate you found in 3(a)). Does it suggest that size bias is present in this third stratum?

- c) Another consideration when using regression with post stratification is the following: is the detection function you are using for the regression the correct one (recall that the explanatory variable in the cluster-size regression is $g(x)$)? In other words, in 3(b) the pooled detection function was used for the regression in the third stratum. However, if you suspect you have size bias in the first place, then you would expect the detection function for larger and smaller cluster sizes to be different - you would expect the detection function for larger cluster sizes to have a wider shoulder (i.e. larger effective strip width and a smaller $f(0)$). Therefore, perform an analysis where you estimate a detection function for each stratum. Look at the results – are the detection functions different between strata? Do they seem plausible? Are you satisfied with the sample sizes used to estimate the detection functions?

Model Definition Properties: [hr_no_adj_post-strat E(s)_using regr_strat f(...)]

Analysis Engine: CDS - Conventional distance sampling

Estimate | Detection function | Cluster size | Multipliers | Variance | Misc.

Stratum definition

No stratification Layer type: Field name:

Use layer type: Stratum

Post-stratify, using: Observation Cluster strat

Sample definition (for encounter rate)

Use layer type: Sample

Quantities to estimate and level of resolution

	Level of resolution of estimates		
	Global	Stratum	Sample
Density	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Encounter rate	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Detection function	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Cluster size (if required)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Global density estimate is Sum of stratum estimates

weighted by Strata are replicates

Defaults Name: hr_no_adj_post-strat E(s)_using re OK Cancel

Overall question: consider all the analyses conducted – which would you use for this dataset?

Introduction to Distance Sampling

Exercise 8: Covariates in the detection function

This exercise consists of four datasets of increasing difficulty. Everybody should work through the first dataset, and the other datasets can be examined later if you wish, or you may work on them when you complete the first analysis. The first analysis will show you the rudiments of conducting an analysis, while the remaining analyses take you deeper into the heart of understanding multiple covariates.

1 A whale of a dataset

Rather than relaxing here in the serenity and tranquility of the Scottish coast, imagine instead that you are a research biologist collecting distance sampling data during December on gray whales as they migrated through the Aleutian chain near Unimak Pass en route to their wintering grounds off Baja California (some luckier, more senior researcher, got the job of data collection on their wintering grounds). These data will now be the focus of your attention for this exercise examining the potential utility of covariates in explaining variation in animal detectability.



Detections were of individuals (not groups), and you chose to record not only distance, but also time of observation (at this latitude at this time of year, the crew was restricted to making observations between 1000 and 1500 during the day). However, because of the low sun angles during much of this time, there was some reason to believe that time of day might play a role in whale detectability. [In what manner might you wish to incorporate this covariate?]

Under extreme weather conditions, observer motion sickness can influence the performance of the observers. An additional covariate, "motion sickness tablet effective dosage at time of observation (MSTDO)" was recorded each time a whale was detected.

The data are available for your inspection in the Distance project **adv_practical_1.dst**. Notice the extreme precision with which the perpendicular distances were measured (how do you suppose this could happen on a rolling ship in the Bering Sea?).

Describe your candidate model set (what models did you construct) and your rationale for the final estimates you provide. You may also comment upon the use of time of observation as a measure of glare from oblique sun angles.

If you have been successful in performing the analysis of this dataset (which can now be revealed to have been simulated), you can continue to sharpen your skills in using covariates in your analysis of distance sampling data by exploring two other data sets, that are considerably more elaborate.

2 Golf tees Data

2.1 The data and distance project

The data come from a survey of clusters of golf tees in grass, conducted by 3rd and 4th year statistics students at the University of St Andrews. It was conducted as a double platform survey; double-platform methods are described later in the workshop and so

we will only use detections recorded for one observer (or platform) for the purposes of this exercise.

Assume that all the data were collected on one 210 metre long transect line, and that this comprises the study area. There were 250 clusters of tees in the study area and 760 individual tees in total.

The population was independently surveyed by two observer teams, of which we will use the data recorded by observer 1. The following data were recorded for each detected group: perpendicular distance, cluster size, observer (team 1 or 2), “sex” (males are yellow=1, females green=0 and golf tees occur in single-sex clusters), and “exposure”. Exposure was a subjective judgment of whether the cluster was substantially obscured by grass (exposure=0) or not (exposure=1). The lengths of grass varied along the transect line, and the grass was also slightly more yellow along one part of the line compared with the rest.

The data are stored in the distance project **GolfteesExercise**. Open the project. Notice that there is already a data filter and several model definitions set up. To avoid overwriting these as they will be used in the double platform exercise, first create a new ‘set’ on the **Analysis** tab and then create a new analysis for that set. Open the model details for the new analysis. First, we need to select only the sightings detected by one observer – we will use observer 1 sightings. Create a new data filter and on the **Data Selection** tab in the data filter, click ‘+’. Choose Layer type ‘Observation’ from the dropdown menu that appears when you click on the cell. Under Selection criteria type ‘observer=1 AND detected=1’. This selects only the detections made by observer 1. The data is already truncated at 4 metres and we will use the same truncation distance.

2.2 CDS analysis of the golf tee data

Now create a new model definition. Start by performing a conventional distance sampling (CDS) analysis using a half-normal key function. To do this, edit the new model definition. Under the Analysis Engine, choose CDS and use the default setting for the detection function. Give it a sensible name and run it.

Look at the results (in the **Analysis** details, **Results** tab). Don’t worry about the warning – this is because there is only one transect and so the encounter rate variance is estimated assuming that the observations are from a Poisson distribution so that $\hat{V}(n) = n$ rather than from inter-transect variation. Make a note of the estimated abundance and associated coefficient of variation (CV). Also have a look at the percentage of variance that was due to the detection function.

2.3 MCDS analysis of the golf tee data

Create a new analysis and a new model definition. This time choose the MCDS analysis engine.

Check that under the **Detection function** tab, the selected key function is half normal and under the Adjustment terms button we have manual selection of zero adjustment terms. MCDS analyses are much harder for the analysis engine to fit than single covariate ones (and a different algorithm is used). In general, it is better to avoid automated selection of adjustment terms and use manual selection instead. Start with zero adjustments terms, and gradually build up 1, 2 etc. checking AIC or one of the other criteria to see if this gives a better fit. It is also a good idea to tick the option in the **Misc.** tab to ‘Report results for each iteration of detection function fitting routine’ (it is ticked by default for the MCDS engine) – this will help you to diagnose any problems that may occur during fitting.

There were 3 additional covariates recorded along with perpendicular distance; cluster size, sex and exposure. Obviously, sex and exposure are factor variables. Sometimes cluster size can be treated as both a factor variable or as a continuous variable: if there are only a few cluster sizes then it can be treated as a factor; however, if cluster size

ranged over a large number of values it would have to be treated as a continuous variable. In this data, cluster sizes ranged from 1 to 8 and it is debatable as to whether you would want to treat it as a factor variable as there are very few large clusters detected. When including cluster size don't forget to tick the cluster size box on the **Covariates** tab – this tells Distance that this covariate is the cluster size covariate. When cluster size is included as a covariate, Distance uses a 'Horvitz-Thompson-like' estimator of abundance (this will have been covered in lectures). In this case, Distance changes a number of options in the **Estimate** and **Cluster size** tabs. In **Estimate**, it changes the 'Sample definition' option and doesn't allow stratification and in **Cluster size** it removes all the options.

Select each of these terms in turn and also in combination on the **Covariates** tab. After running a model, look at the results. The presentation of results is like that in CDS analyses, with a **Log** tab where any warnings or error messages are written, and the **Results** tab which contains details of the analysis. Make a note of the AIC value and look at the detection function plots – notice the difference in the detection function plots when the covariate is specified as a factor variable or a continuous variable.

Once you have decided on the best model, make a note of the estimated abundance, associated CV and percentage of variance accounted for by the detection function. How has this changed?

3 Dolphin Sightings Data

This exercise is optional – so feel free to switch to your own data if you have some. In this example there are several potential covariates and no 'right' answers!

3.1 Reviewing the data

In this example we have a sample of eastern tropical Pacific (ETP) offshore spotted dolphin sightings data, collected by observers placed on board tuna vessels (the data were kindly made available to us by the Inter-American Tropical Tuna Commission – IATTC). In the ETP, schools of yellow fin tuna commonly associate with schools of certain species of dolphins, and so vessels fishing for tuna often search for dolphins in the hopes of also locating tuna. For each school detected by the tuna vessels, the observer records the species, sighting angle and distance (later converted to perpendicular distance and truncated at 5 nautical miles), school size, and a number of covariates associated with each detected school. Many of these covariates potentially affect the detection function, as they reflect how the search was being carried out.

A variety of search methods are used to find the dolphins, but currently the most commonly used are 20x binoculars from the crow's nest, 20x binoculars from another location on the vessel, from a helicopter, or through "bird radar" (high power radars which are able to detect seabirds flying above the dolphin schools). In the example dataset these are coded as 0, 2, 3, and 5, respectively. Some of these methods may have a wider range of search than the others, and so it is possible that the effective strip width varies according to the method being used.

For each sighting the initial cue type is recorded. This may be birds flying above the school, splashes on the water, floating objects such as logs, or some other unspecified cue. In the example they have been coded as 1, 2, 4 and 3, respectively.

Another covariate that potentially affects the detection function is sea state, as measured by Beaufort. In rougher conditions (i.e. higher Beaufort levels), visibility and/or detectability may be reduced. For this example Beaufort levels are grouped into two categories, the first including Beaufort values ranging from 0 to 2 (coded as 1) and the second containing values from 3 to 5 (coded as 2).

The sample data encompasses sightings made over a three month period: June, July and August (months 6, 7 and 8, respectively).

Begin by extracting and opening the project from the archive **Dolphin.zip**. Once it is open, you will see the **Project Browser**, from which you can have a look at the data (**Data** tab).

3.2 Analysis of Dolphin Sightings data

Start by running a set of conventional distance analyses. Are there any problems in the data and if so how might you mitigate them? (Hint – check out the q-q plot, and also try dividing the data into a large number of intervals in the Model Definition | Detection Function | Diagnostics.)

As there are a number of potential covariates to be used in this example, try fitting models with different covariates and combinations of the covariates. All of the covariates in this example are factor covariates except cluster size.

Keep in mind that this is a large dataset (> 1000 observations), and hence estimation may take a while, particularly if you are allowing up to 5 adjustment terms to be fitted. It will be generally more efficient to start fitting models without any adjustment terms, and then adding one at a time if appropriate. Consider also whether to standardize by w or by σ (or try both!).

You will likely end up with quite a few models, so think about how you are going to name and organize them in the Project Browser (for analyses) and Analysis Components window (for model definitions).

Discuss your choice of final model (or models) with your neighbours - did you make the same choices?

4 Passerine data from Marques et al. (2007)

The data from the Auk paper by Marques et al. (2007) is also available on your data stick. It is zipped as the project **ftAMAUK07.zip**. See if you can produce results comparable to those presented in the manuscript (also on your data stick).

Introduction to Distance Sampling

Exercise 9a: Analysis with the use of multipliers

The Problem

The question is how to estimate of the density of sika deer in a number of woodlands in the Scottish Borders. These animals are quite shy and often will be alert to the presence of an observer before the observer detects them, making surveys of the deer challenging. As a consequence, indirect estimation methods have been applied to this problem. In this manner, an estimate of density is produced for some sign generated by deer (faecal pellets) and this estimate is transformed to density of deer by

$$\hat{D}_{deer} = \frac{\frac{\hat{D}_{\text{pellet groups}}}{\text{mean time to decay}}}{\text{dung production rate (per animal)}} = \frac{\text{dung deposited daily}}{\text{dung production rate}}$$

We will produce a pellet group density estimate, then adjust it accordingly to account for the deposition and decay processes operating during the time the data are being acquired. We will also take uncertainty in the production and decay rates into account in our final estimate of deer density.

The Data

Data from 9 woodlands were collected according to the survey design shown below (note differing amounts of effort in different woodlands based on information derived from pilot surveys).



In addition to these data, we also require estimates of the defecation rate. From a consultation with the literature, we learn that sika deer deposit 25 pellet groups daily. The literature source did not provide a measure of variability of this estimate. During the course of our surveys we also followed the fate of some marked pellet groups to estimate the disappearance (decay) rates of a group. A thorough discussion of

methods useful for estimating decay rates and associated measures of precision can be found in Laing et al. (2003) [found on your thumb drive].

There are many factors that might influence both deposition and decay rates, and for purposes of this exercise we will make the simplifying assumption that decay rate is homogeneous across these woodlands; with their mean time to decay of 163 days and a standard error of 13 days. However if you were to conduct a survey such as this, you would want to investigate this assumption more thoroughly.

Pay a visit to http://www.wcsmalaysia.org/analysis/Nest_dung_decay.htm where Mike Meredith of Wildlife Conservation Society in Malaysia thoroughly describes an analysis to estimate decay rates for animal nests or dung.

Analysis Exercises

Use the Distance project **Deer pellets.zip** for the following analyses.

1. Adjust the multipliers in the project (replacing the place-holders in the project, with values provided in the previous section of this exercise).
2. Fit the usual series of models (uniform, half normal, and hazard rate) models to the data.
3. Select the Multipliers button in the Model Definition Properties to specify the layer and the field in the project database for the multipliers you wish to employ (along with their measure of precision).
4. Produce estimates using the woodland as strata, pooling data across strata for fitting the detection function, but using woodland-specific encounter rate to produce woodland-specific estimates of density.
5. Produce an overall estimate of density as mean of woodland-specific densities weighted by the effort allocated within each woodlot.
6. Make special note of the components of variance (contribution of detection function, encounter rate, decay rate, and what happened to defecation rate component?) in each of the strata.

Introduction to Distance Sampling

Exercise 9b: Cue Counting Analysis Exercise

This practical involves analysing an aerial cue counting survey of minke whales in the Atlantic. Minke whales tend to occur singly. An estimate of mean cue rate and its coefficient of variation have been obtained from tagging studies on a number of minke whales in the area.

The sample size is relatively small for a cue counting survey (which require larger sample sizes for reliable estimation of the detection function than line transect surveys), but this is the sample that was generated by the (expensive) survey, so you just need to do the best you can with it.

The data are stored in the distance project **CueCountingExample.zip**. Open the project, and click on the **Data** tab to see how the data are stored. The species code for minke whales is "W" in this project; "bss" is Beaufort sea state code. A simple analysis has been set up but not run in which data filters are used to subset the data so as to use only the data we desire. Have a look at the model definition, in particular, the 'Multipliers' tab.

Question 1: what is $\hat{\eta}$ (see presentation overheads for its meaning) and its coefficient of variation for these data?

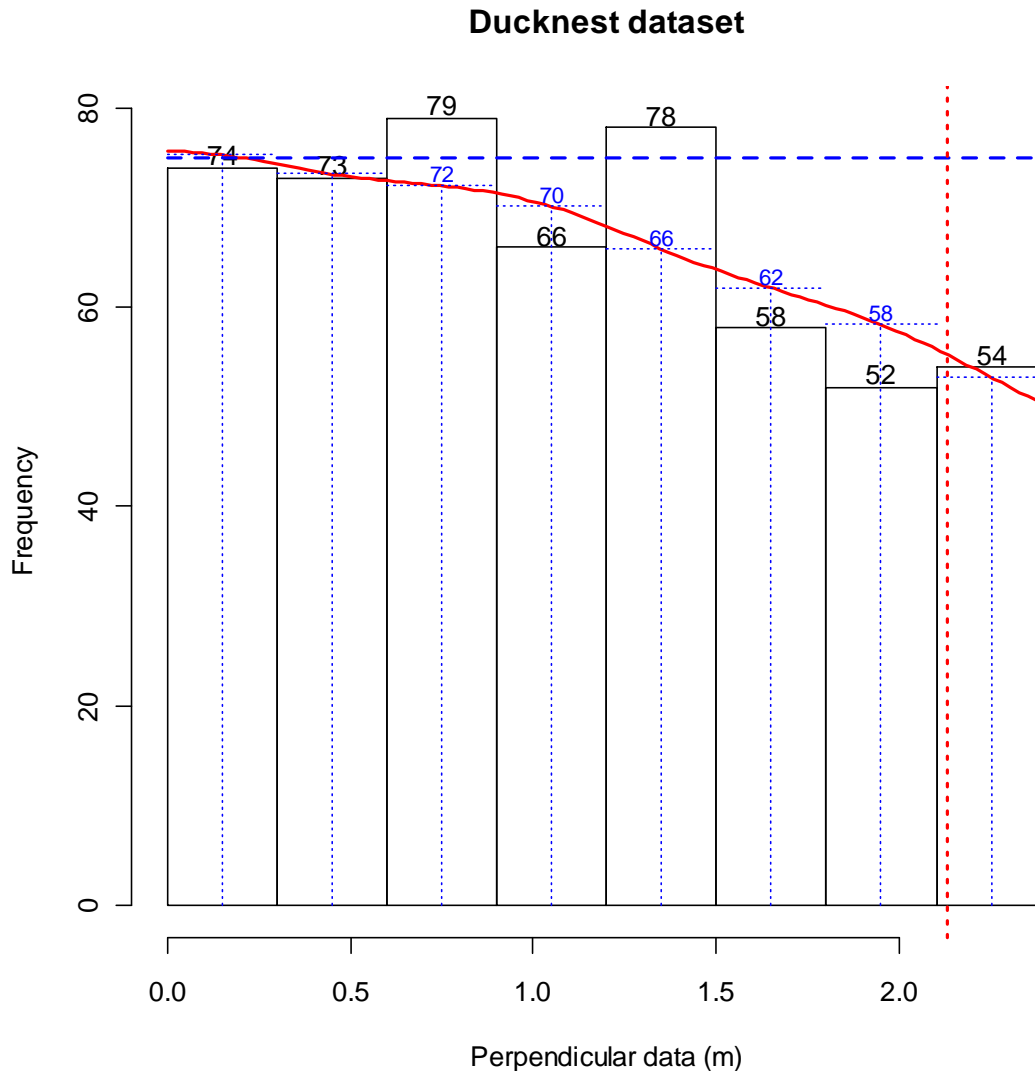
Question 2: what is ϕ (see presentation overheads for its meaning) for this data?

Question 3: Find and fit a suitable detection function model to these data and from this estimate minke whale abundance in the survey region, together with a 95% confidence interval.

We do not describe how you ought to go about selecting a suitable model and assessing its fit (you are becoming experienced using goodness-of-fit statistics and model selection criteria). Note that there was some evidence on the survey of poor-quality distance estimation, so it is worth conducting an analysis on grouped distance data.

Introduction to Distance Sampling

Exercise 1: Line transect Solutions



1) $P_a = \text{area under curve} / \text{area of rectangle}$.

To estimate the area under the curve, I read off the heights of the mid points (in blue) of my fitted curve (red) as follows: 75, 74, 72, 70, 66, 62, 58, 53. So my estimate of area is $(75+74+72+70+66+62+58+53) \times 0.3 = 530 \times 0.3 = 159$. There are lots of other ways to work out the area under a curve – e.g., counting the number of grid squares under the curve on your graph paper or using the trapezoidal rule.

Area of rectangle is height \times width = $75 \times 2.4 = 180$.

So, my estimate of P_a is $159/180 = 0.883$.

How many nests were in the surveyed area? I saw 534 nests, and I estimate the proportion seen is 0.883, so that means I estimate there were $534/0.883=604.7$ nests in the surveyed area. This estimate is for a surveyed area of $2 \times (2.4/1000) \times 2575 = 12.36 \text{ km}^2$. I therefore estimate nest density as $604.7/12.36 = 48.9$ nests per km^2 .

2) The red vertical dashed line shows my estimated effective strip half-width of 2.13m; I estimate that the area below my curve to the right of 2.13 is the same as the area above the curve to the left of 2.13. In this case, the effective area surveyed is estimated as $2\mu L = 2 \times (2.13/1000) \times 2575 = 10.97 \text{ km}^2$, and estimated density is $534/10.97 = 48.7 \text{ nests / km}^2$.

3) For my curve to represent the pdf $f(x)$, I need to rescale such that the area under the curve is 1.0. Since I estimated the area under my curve is 159, I can rescale by dividing all the numbers on the y -axis by 159. The intercept, $f(0)$ is therefore $75/159 = 0.472$. Substituting this into the formula:

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

gives a density estimate of $534 \times (0.472 \times 1000) / (2 \times 2575) = 48.0 \text{ nests per km}^2$ (Note, I had to multiply $f(0)$ by 1000 to convert from m^{-1} to km^{-1} .)

Another way to estimate $f(0)$ is $f(0) = 1/\mu$ – in which case I'd get the same estimate as in part (b).

Distance works by fitting a pdf $f(x)$ to the observed data, and using the estimated $f(0)$ to estimate density. The output also gives μ and P_a , but these are worked out from the estimate of $f(0)$, so Distance would get the same answer whichever formula you used.

Introduction to Distance Sampling

Exercise 2: Line transect analysis of duck nests with Distance

1. You should get very similar estimates of density from different models, provided those models fit the data well. Remember you have
 - the χ^2 goodness-of-fit statistic (why are there 3 of these?)
 - the Kolomogorov-Smirnov and Cramer von Mises tests
 - q-q plot
 - The negative exponential model does not fit

Model	\hat{D} (nests/km ²)	95% c.i. for D
Half-normal (no adjustments)	49.7	(44.2, 55.9)
Fourier series (uniform + cosine)	51.0	(44.9, 58.0)
Hazard-rate (no adjustments)	49.4	(42.3, 57.7)

Compare with 48.6 nests / km² and 48.7 nests / km² from exercise 1.

Introduction to Distance Sampling

Exercise 3: Line transect analysis with Distance

1a) Results of estimating density from simulated data in which true density was 79.8 per km². Findings from some candidate models:

Key function	Adjustments	w (m)	\hat{D}	\hat{D} CV	\hat{D} LCL	\hat{D} UCL
Half normal	0	35.8	87.49	0.16	63	122
Half normal	0	20	84.12	0.17	59	120
Uniform	1	20	86.43	0.17	61	123
Hazard rate	0	20	85.66	0.20	57	129
Neg. exponential	0	20	105.04	0.21	69	159

Not surprisingly for these data (simulated from a half normal detection function with a broad shoulder), the negative exponential model gives a higher estimate than the others, although the confidence interval still includes the true density. The other models provide very similar estimates, though precision is slightly worse for the hazard-rate model (because more parameters fitted). Agreement between the estimate and the known true density is less good if you do not truncate the data, or do not truncate them sufficiently. Take home message: With care, we can get reliable estimates using the wrong model (the data were simulated using a half-normal detection function).

Additional question

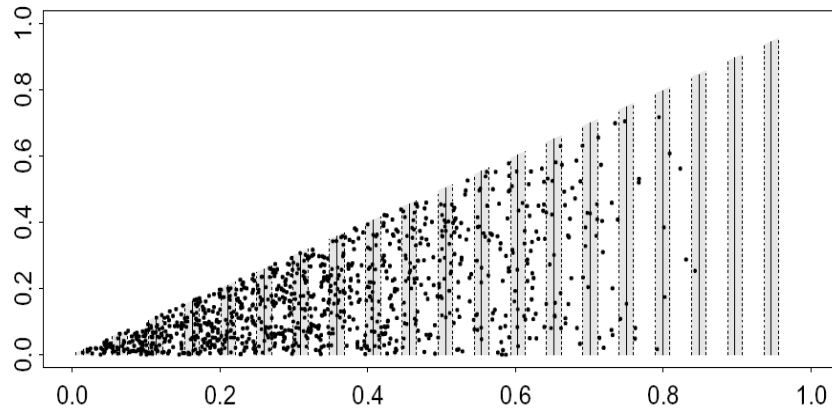
2) These capercaillie data are reasonably well-behaved and different models that fit the data well should give similar results. Note the rounding in the distance data. This means that interval cutpoints for histograms and goodness-of-fit testing, and for the analysis of grouped data if this is required, should be chosen with care (i.e., distance bands ought to be sufficiently broad such that the 'correct' perpendicular distance is in the bands containing the rounded recorded value. e.g. 0, 7.5, 17.5, 27.5, ...

Fitted model	\hat{D}	\hat{D} LCL	\hat{D} UCL	\hat{D} CV
Half normal	4.76	4.01	5.65	0.09
Uniform cosine	4.28	3.22	5.68	0.14
Hazard rate	4.2	3.6	4.9	0.08
Half normal with grouped data	4.06	3.75	4.4	0.09

Introduction to Distance Sampling

Exercise 4: Variance estimation for systematic designs using bootstrap--solutions

Recall the situation in which we have a strong gradient in animal density across our study region, and at the same time we also have a difference in the lengths of our transects; such that short transects are in areas of high animal density, and long transects are in areas of low animal density.



Basic variance estimation, with bootstrapping

8. The precision is very poor: estimated density ranges from about 1000 to 3600: a three-and-a half-fold difference over which we are uncertain. Given that our survey covered 40% of the triangle region, and had a good sample size (254 on 20 transects), this would be a very disappointing result in practice.
9. Bootstrap output [your results may differ slightly as these are created from a random process]:

	Estimate	%CV	#	df	95% Confidence Interval	
Half-normal/Cosine D	2129.2	27.40	999	20.74	1216.2 1164.0	3727.5 3427.2
Half-normal/Cosine N	1064.6	27.40	999	20.74	608.00 582.00	1864.0 1714.0

Note: Confidence interval 1 uses bootstrap SE and log-normal 95% intervals.
Interval 2 is the 2.5% and 97.5% quantiles of the bootstrap estimates.

9. The bootstrap results are very similar to the analytic results, as we would expect. In fact, this did not used to be the case in previous versions of Distance, as the old analytic variance estimator did not perform well when there were extreme trends in both density and line length. You can access the previous default estimator under the Advanced... tab on the Variance page of the Model Definition Properties (it's estimator R3), and more details are given in Fewster et al. (2009) on your thumb drive.
10. The component percentages of variance are as follows:

Component Percentages of Var(D)

Detection probability : 4.3
Encounter rate : 95.7

It should ring an alarm bell to see such a high contribution from Encounter rate. Generally we might expect encounter rate to be in the region of 70% to 80% for line transect surveys.

Post-stratification to improve variance estimation

4. The precision is now greatly improved.

	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	2044.6	8.64	31.41	1715.0	2437.5
N	1022.0	8.64	31.41	858.00	1219.0

and a much smaller and more reasonable (considering the sample size and survey coverage) proportion of the variation comes from estimating encounter rate:

Component Percentages of Var(D)

Detection probability : 44.3
 Encounter rate : 55.7

- The CV is now even smaller, although it could have gone either way since this is an estimator of the same quantity as the last question – just a more precise estimator.

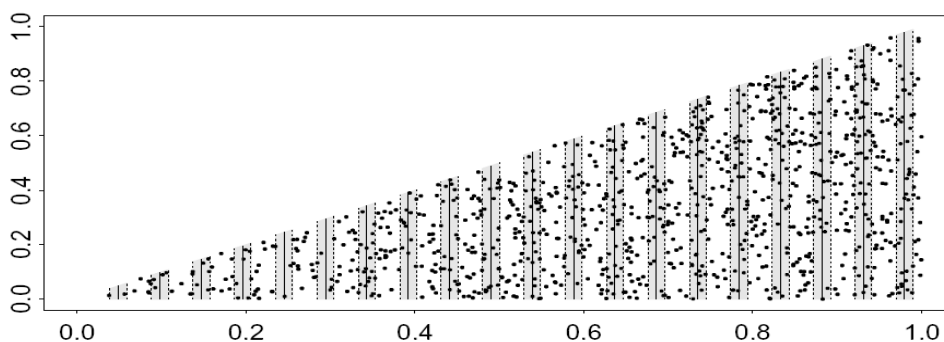
	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	2044.6	7.97	75.87	1745.0	2395.6
N	1022.0	7.97	75.87	872.00	1198.0

The encounter rate degrees of freedom are now 19 (number of lines – 1) rather than 10 (number of post-strata) for the previous question – which is why this is a more precise estimator of the variance.

It must be remembered that we have not made any change to our data by the post-stratification; we are just getting a **better estimate of the variance**. In this case, the increase in precision could make a fundamental difference to the utility of the survey: it might make the difference between being able to make a management decision or not. Usually, trends will not be extreme as they are in this example, and post-stratification will not make a great difference. Such an example is shown below.

Systematic designs where post-stratification is not needed

The following simulated population does not exhibit strong trends across the survey region. Otherwise, the strip dimensions and systematic design are the same as for the previous example.



Without post-stratification: analytic output

	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	1954.0	8.22	50.60	1657.3	2303.9
N	977.00	8.22	50.60	829.00	1152.0

Note: Your bootstrap results will differ slightly, as bootstrapping is a random procedure.

	Estimate	%CV	#	df	95% Confidence Interval	
Half-normal/Cosine						
D	1947.4	10.03	999	50.60	1592.8 1565.0	2380.8 2350.3
Half-normal/Cosine						
N	973.69	10.03	999	50.60	796.00 782.00	1190.0 1175.0

Note: Confidence interval 1 uses bootstrap SE and log-normal 95% intervals.
Interval 2 is the 2.5% and 97.5% quantiles of the bootstrap estimates.

With post-stratification (non-overlapping): analytic output

	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	1954.0	8.38	25.80	1645.4	2320.6
N	977.00	8.38	25.80	823.00	1160.0

Introduction to Distance Sampling

Exercise 5: Notes on point transect exercises

1. Results from selected model options; remember these are simulated data with a half normal detection function and true density 79.6:

Key	Adjustments	# terms	w (m)	\hat{D}	%cv	95% c.i. for D
Half-normal	None	0	34.2	79.6	12.6	(62.1, 102.1)
Half-normal	None	0	20.0	70.8	15.7	(52.0, 96.5)
Uniform	Cosine	1	20.0	75.0	14.4	(56.5, 99.6)
Hazard-rate	None	0	20.0	62.4	18.7	(43.2, 90.0)
Neg. exp.	Simple poly	1	20.0	73.1	29.2	(41.5, 128.6)

We see a fair degree of variability between analyses – reliable analysis of point transect data is more difficult than for line transect data. We see greater loss in precision from truncating data relative to line transect sampling, but if we don't truncate data, different models can give widely differing estimates. For these data, the hazard-rate model appears to have downward bias, and precision is very poor for the negative exponential model.

2. I got the following estimated densities (territories ha⁻¹). I have included estimates for the three other species surveyed (not provided in the projects for the workshop). Method 5 is territory mapping (which does not use distance sampling, and as you note has no measure of precision associated because it is akin to a census method).

Species	Common Chaffinch	Great Tit	European Robin	Winter Wren
Method				
1	1.03 (0.74, 1.43)	0.58 (0.36, 0.94)	0.52 (0.26, 1.06)	1.29 (0.80, 2.11)
2	0.90 (0.62, 1.29)	0.22 (0.13, 0.39)	0.60 (0.38, 0.94)	1.02 (0.80, 1.32)
3	0.71 (0.45, 1.23)	0.26 (0.09, 0.76)	0.82 (0.52, 1.31)	1.21 (0.82, 1.79)
4	0.64 (0.46, 0.90)	0.26 (0.16, 0.42)	0.69 (0.47, 1.00)	1.07 (0.87, 1.31)
5	0.75	0.21	0.84	1.30

To obtain the above estimates, I used a truncation distance of 110m for methods 1 and 2, 92.5m for method 3, and 95m for method 4. For the wren data, I used the uniform key with two cosine adjustments for method 1, the hazard-rate model for methods 2 and 3, and the half-normal key with two Hermite polynomial adjustments for method 4.

Points to note about the wren data: the wren more than any of the other species showed evidence of observer avoidance. This didn't cause too many difficulties, except that the model favoured by AIC for line transect sampling was the hazard-rate model, which had a very flat shoulder out to around 75m. It was implausible that detection was certain out to this distance, so that I selected a model with a slightly inferior AIC value, but with a more plausible fitted detection function. Analyses of the

cue count data are necessarily rather subjective, as the data show substantial over-dispersion (a single bird may give many songbursts all from the same location during a five-minute count). In this circumstance, goodness-of-fit tests are very misleading, and care must be taken not to overfit the data.

3. I obtained good fits to the 1980 savannah sparrow data by truncating at 55m. The half-normal detection function without adjustments fitted well, as did the uniform with cosine adjustments. The hazard-rate model performed less well. There was a marginal preference for fitting the detection function separately in each stratum as judged by AIC, but pooling distance data across strata might offer rather more robust estimation. The estimates of density in the table correspond to a half-normal detection function, fitted separately in each stratum, with a truncation distance of 55m.

For 1981, $w=55m$ was again satisfactory. There was now a clear preference for estimating the detection function separately by stratum, but little to choose between the half-normal model and the uniform model with cosine adjustments. For comparability with 1980, I chose the half-normal model, although AIC showed a very marginal preference for uniform + cosine. (Again, the hazard-rate model provided less good fits overall.)

Estimated densities \hat{D} (birds/ha) of savannah sparrows

Year	Pasture	\hat{D}	95% c.i. for \hat{D}
1980	1	1.43	(0.94, 2.18)
	2	4.12	(3.15, 5.38)
	3	2.35	(1.72, 3.20)
	All	2.63	(2.19, 3.16)
1981	0	1.39	(0.82, 2.36)
	1	0.52	(0.27, 1.03)
	2	1.70	(1.07, 2.71)
	3	1.35	(0.81, 2.26)
	All	1.24	(0.95, 1.62)

Introduction to Distance Sampling

Exercise 6: Automated Survey Design Exercise Solutions

1. Point transect survey of North-eastern Mexico

The completed exercise is archived in the project MexicoUnPrjSolutions.zip

2. Entering geographic data into Distance, and generating Coverage grids

The completed exercise is archived in the project TrapeziumSolutions.zip.

The first 3 designs show results for the equal angle, equal spaced and adjusted angle zigzag designs based on 100 simulations. Even from this small number of runs, it is clear that for the equal angle design coverage probability tends to increase as you move from the right of the survey area, where the trapezium is tall, to the left where the trapezium is shorter. It is easy to see why this is happening by looking at a survey generated using this design (survey 1). For the equal spaced and adjusted angle designs, there doesn't seem to be any pattern in the variation in estimated coverage probability. This variability is largely due to Monte-Carlo error, because we've only done 100 simulations, so before drawing conclusions about these designs, we repeated the exercise with more 10 000 simulations.

These results are shown in designs 4-6. Design 4 is the equal angle zigzag and the pattern of increasing coverage with decreasing trapezium height is now very clear. What about the other two designs? The equal spaced design (Design 5) still looks pretty good, but if you look carefully, there is a hint that coverage is slightly lower on the left side and higher on the right. The coverage probability standard deviation is 0.011. Compare this with the standard deviation for the adjusted angle design (Design 6) – 0.007. Also look at the coverage probability map for the adjusted angle design – there is no evidence of any pattern in coverage probability. We conclude that the equal spaced design has close to even coverage probability, but that only the adjusted angle design has completely even coverage probability.

Note that this result only applies for the adjusted angle design if the study area width is constant perpendicular to the design access. If you try repeating the exercise with a triangular-shaped study area, you will find out that even the adjusted angle design will not have even coverage probability.

3. Systematic parallel line aerial survey of marine mammals in St Andrews bay

I got the following results (yours will be slightly different because the survey locations in each simulation are selected at random). See also the project archived in StAndrewsSolutions.zip

Trackline spacing	On effort trackline length			Total trackline length		
	Min	Max	Mean	Min	Max	Mean
4.5	206.6	228.8	219.6	249.3	275.3	264.7
5.0	184.4	205.6	198.2	220.5	248.8	242.5
5.5	169.7	189.5	178.9	217.1	245.3	224.7
6.0	152.8	176.1	162.1	183.7	220.7	206.1

Based on these, the 5.0km spacing seems to get us closest to our goal of 200km on effort for 250km total trackline length. The maximum total trackline length didn't exceed 250km which is re-assuring if this is an absolute upper limit.

I went ahead and generated one realization of this 5km design, which we will use as the survey plan. It gave me a total trackline of 226.2km, with 184.6km on effort (see StAndrewsSolutions project file). While this is rather less than I wanted, I can't validly throw

this one away and generate another as we could no longer validly claim to have a random start point (I'd effectively only be choosing start points that lead to the amount of trackline length I want) and so would no longer have even coverage probability.

As an aside, it is also interesting to look at the proportion of the total survey time spent on effort – reported in Distance as the proportion on effort/total effort:

Trackline spacing	Mean on effort / total effort
4.5	0.83
5.0	0.82
5.5	0.80
6.0	0.78

Not surprisingly, the greater the spacing between tracklines, the smaller the proportion of time we spend on effort as we have to spend time flying between the transect lines.

Introduction to Distance Sampling

Exercise 7a: Analysis of Stratified Data Outline Solutions

Example analyses, which were used in getting these solutions, and which are referred to below, are in the project file "Stratify solutions.dst".

1. Relevant results are in Analysis "Full geog stratification".
The AICs are 127.90 for the southern stratum and 187.90 for the northern stratum. Detection function model fits are adequate visually and by goodness-of-fit test. Sample sizes are relatively small but not alarmingly so. The southern stratum appears to have a much narrower effective strip width.
2. Relevant results are in Analysis "Pooled $f(0)$ ".
The AIC for the pooled detection function fit is 318.72. The detection function model fit is adequate visually and by goodness-of-fit test. Since $318.72 > (127.9+187.9=315.8)$ estimation of separate detection function in each stratum is preferable.
3. Relevant results are in Analysis "No stratification".
The whale density estimate from the unstratified analysis is around 25% larger than the corresponding estimates from 1. and 2. above. The reason is that the survey design was geographically stratified, with less survey effort in the north stratum, and this is being ignored in the unstratified analysis.

What is **not included in this project** are cluster sizes of the observed minke whale groups (we didn't want to clutter the analysis with that detail). However, there is a bit of a story in geographic variation in cluster sizes. Cluster densities are higher in the southern stratum, but transects from both strata are being treated as if they are representative of the whole survey region. This results in a positively biased cluster density for the region as a whole. In addition, cluster sizes are higher in the South stratum. The estimate of $E(s)$ from the unstratified analysis is a positively biased estimate of $E(s)$ for the North stratum and a negatively biased estimate of $E(s)$ for the South stratum. When it is applied to both strata, it results in a positively biased estimate of whale abundance because the North stratum is much larger and contains roughly twice as many whales as the south stratum.

Moral: Don't perform analyses without taking the survey design into account!

Introduction to Distance Sampling

Exercise 7b: Analysis of Clustered Data Outline Solutions

Example analyses, which were used in getting these solutions, and which are referred to below, are in the project file "Cluster solutions.dst".

1. Relevant results can be found in the analysis "E(s) by ln(s)_g(x)".

- (a) No, the mean observed cluster size is 2.25 (se=0.229) the regression estimate of E(s) is 1.89 (se=0.139). The regression method not only corrects for size bias, but has also given a smaller standard error.
- (b) 7.2% of the variance of the density estimate comes from mean cluster size estimation – so a small amount compared to the variance caused by encounter rate and estimating the detection function.

2. Relevant results can be found in the analysis "truncation E(s)".

The detection function shoulder extends out to about the end of the second distance interval, so all data beyond this were discarded for estimation of cluster size (NB: this truncation does not affect the estimation of the detection function or encounter rate).

- (a) It is different because in this analysis all data beyond the second distance interval have been discarded in an attempt to eliminate any size bias in the data. Compare this result (1.85, se=0.183) with the mean of the observed clusters using the untruncated data (2.25, se=0.229) - note how this result is much closer to the estimate of E(s) using regression (1.89, se=0.139).
 - (b) It is largely because the observed mean cluster size is based on only 41 observations, while the regression estimate in analysis 1 above is based on 88 observations – you pay for discarding data with increased variance.
3. Relevant results can be found in the analyses "Post-stratified E(s) using mean", "Post-stratified E(s)_pooled f(0)_regr" and "Post-stratified E(s)_strat f(0)_regr".
- (a) Mean cluster size is not relevant for strata 1 & 2 as the clusters in each were all the same size. The mean cluster size for the final stratum was 5.81 (se = 0.748).
 - (b) In the analysis "Post-stratified E(s)_pooled f(0)_regr", cluster size strata are pooled for estimation of the detection function and the regression method has been used within cluster size strata. The regression estimate for the third stratum is 4.36 (se = 0.563) which is lower than the mean cluster size, suggesting that size bias is present in this stratum. So, for these data, it would not have been correct to assume that the effect of size bias had been eliminated by post-stratifying (and then using the mean of the observed cluster sizes in each stratum).
 - (c) In the analysis "Post-stratified E(s)_strat f(0)_regr", the detection function has been estimated separately in each cluster size stratum. The detection functions are different from each other - it looks like cluster sizes 3 and above are detected with certainty almost all the way out to 1.2nm. It is questionable whether this is actually possible. In addition, the sample size for the third stratum is very small (only 16 observations) – less than is usually recommended for modelling a detection function.

Overall conclusion: considering all the analyses

- There are questions raised about all the analyses using post stratification. Post stratification and using the mean cluster size per stratum did not eliminate size bias, as we discovered when we checked by using post-stratification with regression. Using a pooled detection function was not ideal, as we suspected that the detection functions would be different for different strata. This was confirmed when detection functions were estimated per stratum. However, the sample size in the third stratum was too small to have enough evidence to believe that the fitted detection function was plausible.
- That leaves the regression and the truncation method to choose between. There were no problems with the regression method, and although the truncation method gave a similar estimate of cluster size, data were thrown away, resulting in a larger standard error.

Overall conclusion: the regression method is the preferred method.

Introduction to Distance Sampling

Exercise 8: Covariates in the detection function

Outline Solutions

1 Simulated whale data

An example of the sort of analysis you might have performed is given in the archived project file `adv_practical_1_solutions.zip`. If you sort by date created or analysis ID, you can see the order I set up the analyses in. I first tried simple half-normal and hazard rate models without covariates, and found that the half-normal model had a lower AIC. I then tried the MSTDO covariate and hour covariates separately (as non-factor covariates). The analysis with MSTDO had a much lower AIC, but the analysis with hour actually had a higher AIC than the analysis without covariates. I tried an analysis with both MSTDO and hour, but this had a higher AIC than MSTDO alone (Table 1). I concluded that the MSTDO covariate was important, but the hour covariate was not.

Although these data did not appear to need any truncation, I briefly confirmed that the same results were obtained with 10% truncation (analyses 7 and 8). Further analyses could look at the effect of adding adjustment terms to the detection function, although since no adjustments were selected with the half-normal without covariates it is likely that none will be required when the MSTDO covariate is used.

Table 1. AIC values for the candidate models.

Model	No truncation	10% truncation
HZ: simple model	125.32	82.46
HN: simple model	123.28	80.80
HN: with MSTDO	111.21	76.06
HN: with HOUR	125.03	82.15
HN: with HOUR + MSTDO	113.14	78.02

2 Analysis of golf tee data

With three covariates there are eight possible detection function models (including perpendicular distance only). The AIC from the CDS model was 311.1 and the lowest AIC I found was 304.3 which included sex as the only additional covariate. You may have found a different model. Table 2 is a summary of the results from the CDS analysis and an MCDS analysis with my best model. The component of variance due to the detection function fitted as a CDS was 64.3% and this reduced to 54.2% when sex was included in the detection function.

As part of an exploratory data analysis it is useful to analyse the data as a CDS but post stratify using the factor variables and fit separate components of the model for each factor level (as long as there are enough observations). The esw's for females (factor level = 0) and males (factor level = 1) are 1.61 metres (%CV=13.0) and 2.65 metres (%CV=10.3), respectively. The esw's for the exposure levels 0 and 1 are 2.41 (13.2) and 2.31 (10.0). The differences between males and females appear to be much larger than the difference between exposure levels indicating that sex would be the more useful covariate to include in the model. Notice how the abundance estimate for the CDS post-stratified by sex and the MCDS including sex are very similar, but the CV is smaller for the latter.

Table 2. Parameter estimates from CDS models and MCDS model which included sex only. CV's are given in parentheses

Parameter	True value	CDS	CDS post stratified by sex		MCDS
			Female (0)	Male (1)	
AIC		311.14	69.7	234.7	304.3
esw (m)		2.34 (7.9)	1.61 (13.0)	2.65 (10.3)	2.24 (6.4)
Ds (clusters per m ²)	0.15	0.13 (7.9)	0.05 (21.3)	0.08 (10.3)	0.13 (11.0)
E[S]	3.04	3.01 (5.9)	2.80 (13.7)	3.13 (6.5)	3.01 (5.9)
D (tees per m ²)	0.45	0.38 (9.9)	0.14 (18.9)	0.25 (12.2)	0.40 (8.8)
N	760	638 (9.9)	243 (18.9)	421 (12.2)	666 (8.8)
			664 (22.5)		

3 Analysis of dolphin sightings data

To obtain an overall impression of the data it is useful to fit a detection function histogram with many intervals (you may have problems fitting to the maximum number of 30, but 25 intervals should be OK). The spikes in the histogram suggest that the data has been rounded to zero and possibly other values. The q-q plot also indicates problems with the model at zero distances. To mitigate these problems, use the diagnostic tab to pool the data into a few intervals – 10 to 15 intervals work OK.

For the MCDS analysis, cluster size was fitted as a continuous variable, whereas, month, Beaufort, cue and search position were fitted as factor variables. Table 3 summarises the results. The number of adjustment terms allowed was limited to a maximum of two. In most cases a half normal function was chosen with either no, or one, adjustment term.

Table 3. Parameter estimates for the different models. Percentage CVs are given in parentheses. Note that CVs for the model containing cluster size are obtained by bootstrapping.

Parameter	CDS	Cluster size	Month	Beaufort	Cue	Search
AIC	3365.9	3359.5	3362.6	3366.9	3368.3	3339.8
esw (nm)	3.00 (4.5)	3.08 (1.9)	3.00 (1.9)	3.00 (1.9)	3.00 (1.9)	2.93 (2.3)
Ds (clusters per nm ²)	181 (4.5)	177	181 (1.9)	181 (1.9)	181 (1.9)	185 (2.3)
E[S]	507 (5.3)	460	529 (5.3)	507 (5.3)	495 (5.3)	589 (5.3)
D (animals per nm ²)	91965 (7.0)	81454	96009 (5.7)	91921 (5.6)	89729 (5.6)	109420 (5.8)

Based on the AIC, it seems as though the model including search method is best, however, there were warning messages about the detection function fitting and cluster size estimation. Before going on and looking at models which include two covariates, it is worth looking at the search model in more detail. The detection functions have very different scale parameters, for example, the detection function for search method 3 (using a helicopter) has a very wide shoulder and so the scale parameter is very large. This suggests that the observers were seeing everything out to 5 nm and so detection does not decrease with distance as it does with the other methods. One assumption of MCDS is that the perpendicular distance distributions of the covariate factor levels have the same shape. It may be worth refitting the model ignoring the observations made by the helicopter. Data can easily be selected/ignored using the Data filter | Data selection tab. The selection criteria will be of the form '[Search method] IN (0,2,5)'

This is a large dataset and so it is worth deciding on your final model before doing any bootstrapping to obtain variances.

4 Hawaiian Passerines

We provide no sample solution to these data, consult the Marques et al. (2007) reprint on your data stick.

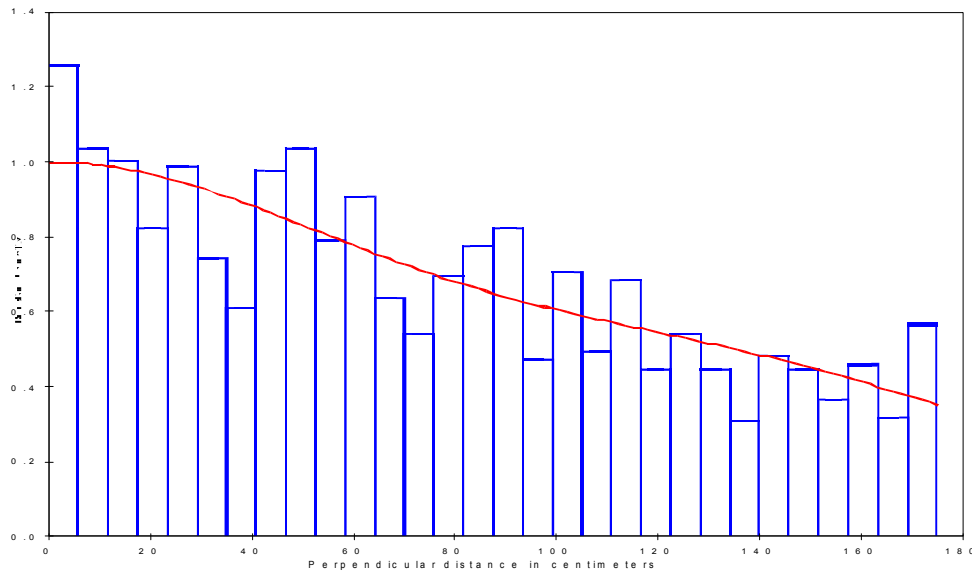
Introduction to Distance Sampling

Exercise 9a: Analysis with the use of multipliers Solution outline

We did not perform a comprehensive examination of fitting a detection function to the pellet groups detected. However, as a general practice, we have truncated the most distant 10% pellet groups. Have a look at “Deer pellets solution.zip”

For management purposes, we would like to produce an estimate of the number of deer inhabiting each woodlot. In scrutinizing the data set, we see there is considerable variability in the number of pellet groups detected within each woodlot, and in some woodlots we detected as few as 4 pellet groups. Hence we cannot reliably estimate woodlot-specific detection functions. Consequently, we will pool data across woodlots to derive a global detection function. To produce woodlot-specific density estimates, we combine woodlot-specific encounter rates with the global detection function.

The global detection function



Encounter rate per kilometer by woodlot

	Encounter rate	CV(n/L)
Block A	715.88	17.25
Block B	360.00	22.99
Block C	37.778	21.51
Block E	35.294	49.26
Block F	145.00	0.00
Block G	80.000	67.70
Block H	15.000	0.00
Block J	70.000	0.00

Note that blocks F, H, and J have but a single transect placed in them. As a consequence, it is not possible to empirically compute a variance for encounter rate in those woodlots.

Results

Produce an overall estimate of density as mean of woodland-specific densities weighted by the effort allocated within each woodlot.--

With considerable effort allocated in woodlot A, where deer density is high, the overall estimate of density is between the estimated density in woodlot A of 74 deer per km⁻² and the lower densities in the remaining woodlots.

Make special note of the components of variance (contribution of detection function, encounter rate, decay rate, and what happened to defecation rate component?) in each of the strata.

Because we now have uncertainty associated not only with the detection function and encounter rate, but also decay rate we are presented with these component of variability for each of the strata for which we requested estimates of density.

In woodlot A, there were 13 transects on which over 1200 pellet groups were detected; uncertainty in the estimated density was 19.0% and the variance components were apportioned as

Component Percentages of Var(D)	

Detection probability	: 4.2
Encounter rate	: 78.1
Decay rate	: 17.7

whereas woodlot E had 5 transects, but only 30 detections overall (resulting in a CV of 48%)

Component Percentages of Var(D)	

Detection probability	: 0.7
Encounter rate	: 96.6
Decay rate	: 2.8

In woodlot F, were only a single transect was placed, the CV of density was 8.9% with the allocation being

Component Percentages of Var(D)	

Detection probability	: 19.1
Decay rate	: 80.9

Do you trust this assessment of uncertainty in the density of deer in this woodlot? We are missing a component of variation because we were negligent in placing only a single transect in this woodlot, and so are left to *assume* there is no variability in encounter rate in this woodlot.

By the same token, we are left to assume there is no variability in defecation rates between deer because we have no measure of uncertainty in this facet of our assessment of deer densities.

Introduction to Distance Sampling

Exercise 9b: Cue Counting Analysis Example Solution

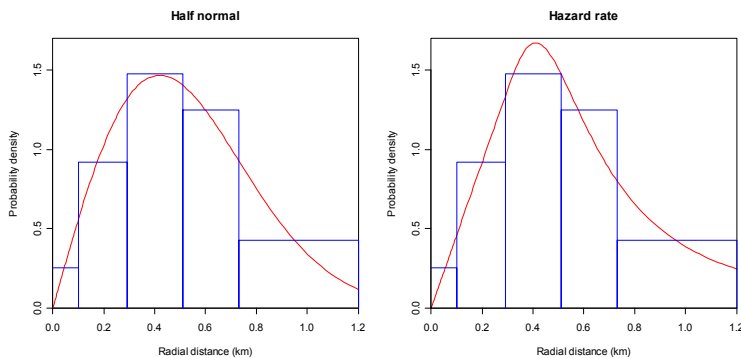
Question 1: $\hat{\eta} = 25$ cues per time unit (per hour in this case). Its standard error is 5, therefore its CV is $5/25=0.2$ (or 20%).

Question 2: Half the circle was searched so the sampling fraction, $\phi/2\pi = 0.5$. Therefore, $\phi = \pi$ (ϕ must be in radians).

Question 3: An example analysis is in the project **D6CueCountingSolution.zip**. A half-normal detection function model with no adjustment parameters was chosen. Minke whale abundance was estimated to be 13,427 whales with 95% confidence interval (5,612; 32,124).

Note the large difference between this and the estimate from the hazard-rate model, which is 10,711 whales, with 95% confidence interval (4,234; 27,097). Although the models produce a warning, this is not in itself a cause for concern, since all it says is that it could not consider many adjustment parameters because the data are in so few intervals – in any event models with no adjustment parameters were chosen (see 2nd page of the analysis output).

Remember that the key parameter in a cue counting analysis is $h(0)$, the slope of the fitted pdf to the observed data at distance zero. The difference between the two estimates is the difference between these slopes for the two models:



Cue-counting estimates of detection probability are more volatile than those from line transect surveys, because on a cue-counting survey you have least data where you need it most to estimate $h(0)$ – namely at distances close to zero. As a consequence, cue-counting surveys require higher cue sample size for reliable estimation than samples of animals for line transect surveys.

Don't worry too much about the apparent lack of fit in the first interval or two in the plots below – remember the sample size is very small in these intervals. Use the plot above and the goodness-of-fit statistics to guide you about the fit of your model.

