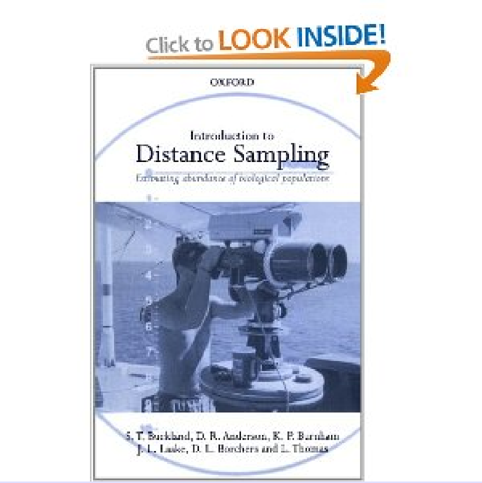
Design and Analysis of Distance Sampling Studies

Carl James Schwarz

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada
cschwarz @ stat.sfu.ca

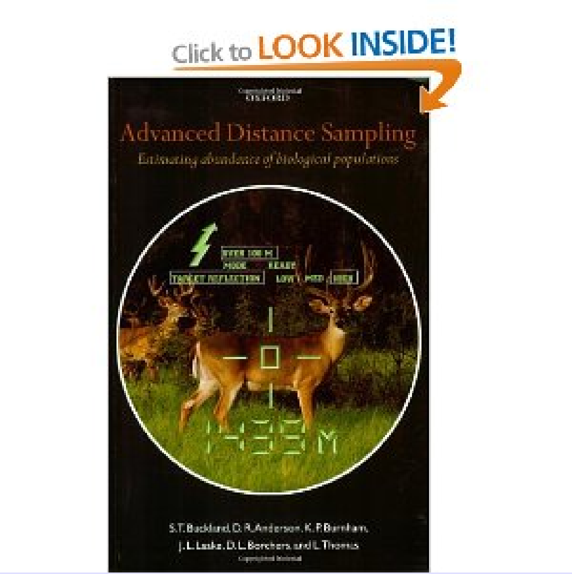
Part 1
Conventional Distance Sampling - Analysis

Books on Distance Sampling



Supersedes old book (available online at
`http://www.colostate.edu/Dept/coopunit/download.html`.

Books on Distance Sampling



DISTANCE Home Page

http://www.ruwpa.st-and.ac.uk/distance/

Program Distance Home Page 13/11/2012 10:32 AM

Distance Home Page

Distance is supported by

Program Distance

- What is Distance?
  - Who uses Distance? - over 26,000 downloads!
- Download Distance
  - Distance 6.0 Release 2
- Distance Support, Updates and Extras
  - Report a Bug in Distance
- Sponsors of Distance
- Program Authors and Citation
  - Paper on program Distance (pdf) - now the default citation

Distance Sampling

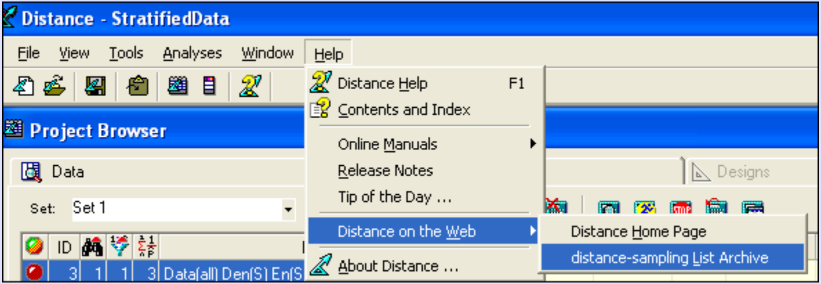
- Distance Sampling Reference Material
  - Online preview of Introduction to Distance Sampling
  - Online preview of Advanced Distance Sampling
  - Non-technical introduction to distance sampling (pdf, short)
  - International Encyclopedia of Statistical Sciences article (pdf, short)
  - Encyclopedia of Environmetrics article (pdf, longer)
  - Pronghorn aerial survey manual
  - Extensive online bibliography
- Distance-sampling Email list
- Workshops and Meetings
  - NEW! Training workshops @ St Andrews, Aug 2012
  - Online abstracts from ISEC 2010 conference
  - Online abstracts from passive acoustic density estimation workshops 2009
  - Online abstracts from ISEC 2008 conference
  - Online abstracts from ICODS 2001 conference

Web site hosted by
Research Unit for Wildlife Population Assessment

Comments and suggestions to
distance@mcs.st-and.ac.uk

Last modified December 8, 2011
Log of changes

DISTANCE - list serve accessible within DISTANCE



What is Distance Sampling

Goal: Estimate **Density** and/or **Abundance**

- Move along a transect and measure objects seen from the transect
  Stay at a point and measure distance of objects seen from the point.
- Detection probability on the transect line/point is certain (probability $= 1$)
- Detection probability declines with distance from the transect/point

Used for:

- Density of whales, carnivores, ungulates, birds, tortoises,
- Transects on ground, in the air
- Visual or auditory detections

Key problem: **Detection probability** < 1**!** Naive estimates lead to underestimates of density and abundance.

When to use Distance Sampling Methods

When to use distance sampling?

- Low density of objects that can be rapidly visually surveyed (e.g. from airplane)
- Only density/abundance is of interest
- Objects are relative immobile relative to observer.
- Detection falls a function of distance.

When to NOT use Distance Sampling Methods

When NOT to use distance sampling:

- Need information on survival rates and other population parameters
- Need to follow individual animals.
- Animals are hidden (multipliers?); animals move very fast
- Animals avoid/attracted to observer.

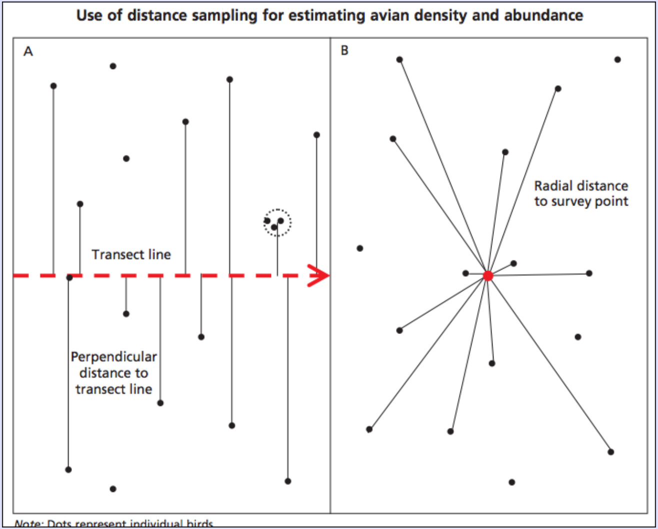
Types of Distance Sampling

Types of Distance Sampling (not a complete list)

- Line transect sampling
- Point sampling
- Cue counting
- Double observer transect sampling
- Spatial density sampling

Types of Distance Sampling

We will focus on Line and Point Distance sampling.

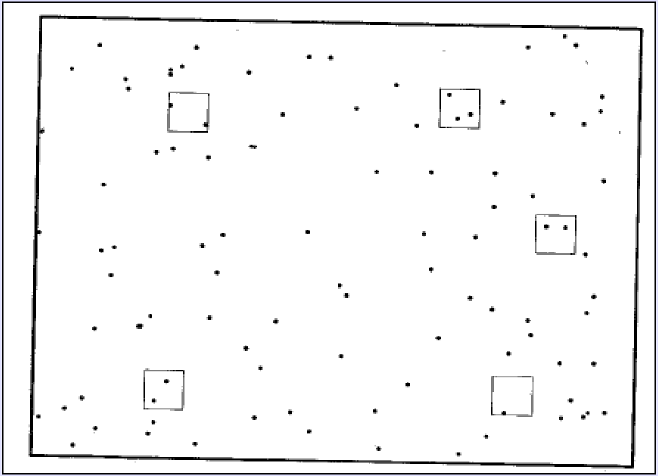


Software for Distance Sampling

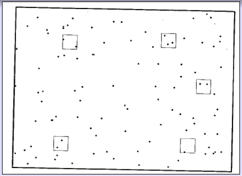
- DISTANCE `http://www.ruwpa.st-and.ac.uk/distance/`
- umarked, Distance, distsamp *R* packages

- Population vs. Sample
- RRRs (Randomization, Replication, Stratification)
- Parameter (D) vs. Statistic (distances measured) vs. Estimate ($\widehat{D}$)
- Standard deviation (s) vs. Standard error (SE) vs. Confidence Interval (CI)

Quadrat sampling with perfect detection:



Five quadrates (1 m^2) sampled in total area of 125 m^2.
Counts are 2, 2, 3, 2, 1. Mean is 2 and standard deviation is 0.707.



Five quadrates (1 m^2) sampled using Simple Random Sampling in total area of 125 m^2.

Counts are 2, 2, 3, 2, 1. Mean is 2 and standard deviation is 0.707.

$$\widehat{D} = \frac{Detection}{Area} = \frac{2+2+3+2+1}{5} = 2$$

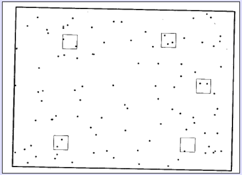
$$se(\widehat{D}) = \frac{s}{\sqrt{n}} = \frac{0.707}{\sqrt{5}} = 0.316$$

$$\widehat{N} = 125 \times \widehat{D} = 125 \times 2 = 250$$

$$se(\widehat{N}) = 125 \times se(\widehat{D}) = 125 \times 0.316 = 39.5$$

What is meaning of standard error (SE)?

How does a standard error differ from a standard deviation?



Approximate 95% confidence interval is found as

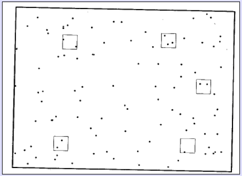
$$\widehat{D} \pm 2SE = 2 \pm 2(.316) = (1.36 \rightarrow 2.64)$$

$$\widehat{N} \pm 2SE = 250 \pm 2(39.5) = (171 \rightarrow 329)$$

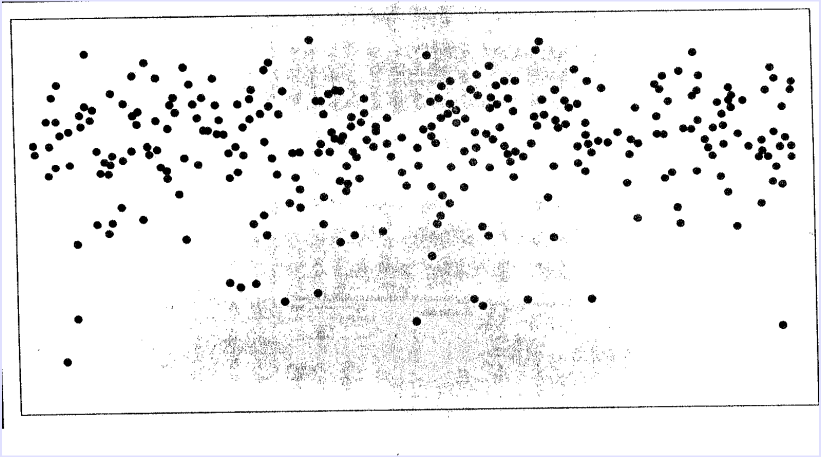
What does these mean?

SE and CI say NOTHING about individual DATA points.

Basic Statistical Review - Quadrat Sampling

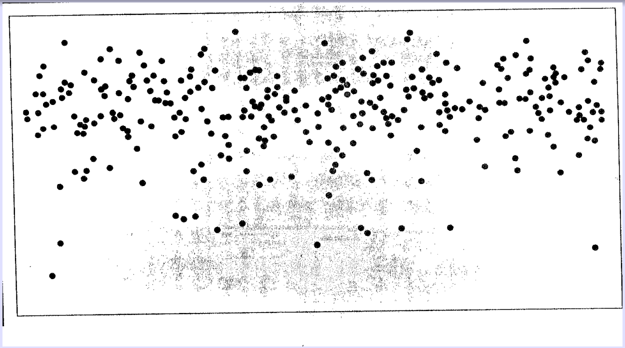


- Precision depends on number of quadrats and size of quadrats
    - If objects are clumped, larger quadrats are better
    - SE typically declines with $\sqrt{effort}$
- Assumes that objects are fixed during survey
- Assumes perfect detectability within each quadrat.
- Best if density is roughly uniform throughout area.



High variability in population leads to poor performance of simple random sampling.

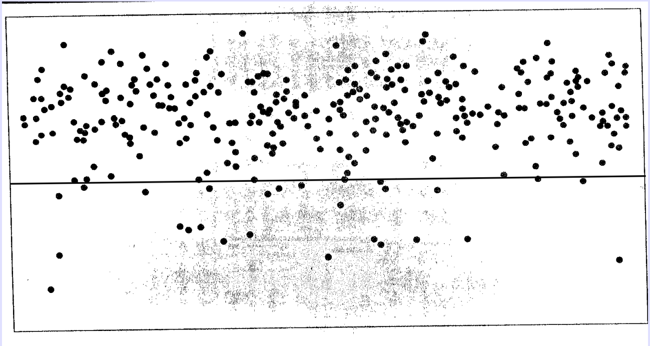
Basic Statistical Review - Quadrat Sampling



Stratified Sampling

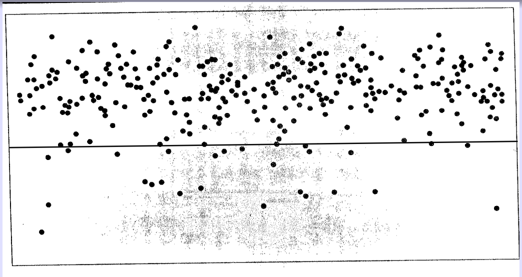
- Divide into more homogeneous strata (3 - 5 strata)
- Allocate effort to strata (equal or proportional)
- Separate survey in each stratum
- Estimate from each stratum
- Rollup

Basic Statistical Review - Quadrat Sampling



Stratify to make population homogeneous in each stratum

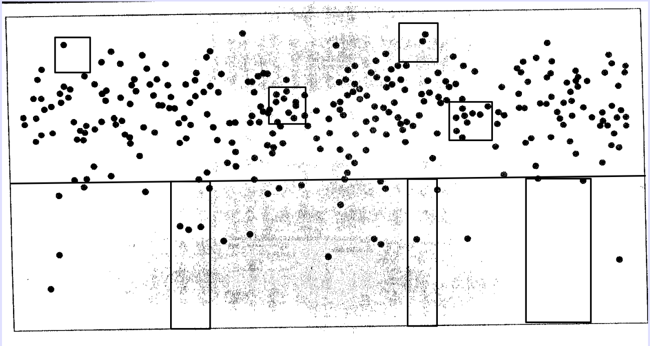
- Strata do NOT have to be contiguous.
- Strata do NOT have to be the same size.
- Pre-stratification is common; post-stratification can be done but you may run into problems in rollup



Allocation of effort:

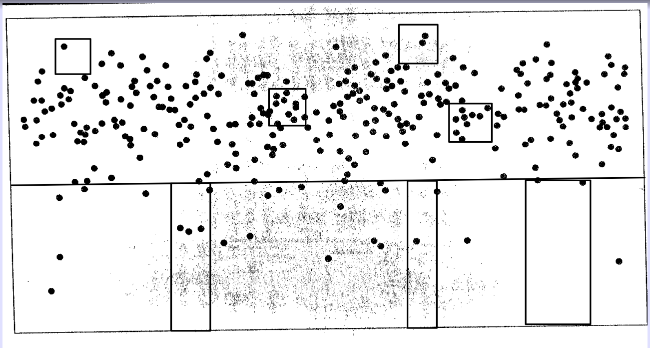
- Equal allocation would put equal effort in each stratum. Is this wise?
- Proportional allocation would put more effort in more important strata
  - Importance $\approx$ size of stratum?
  - Importance $\approx$ density of stratum?
  - Importance $\approx$ value of stratum?
- More complex allocation schemes are also available, but you are unlikely to have relevant information to use them.

Basic Statistical Review - Quadrat Sampling



Separate survey in each stratum

- NOT necessary to use same type of survey in all strata
    - Use the most efficient method for each stratum
    - Of course, use the proper analysis for each stratum!

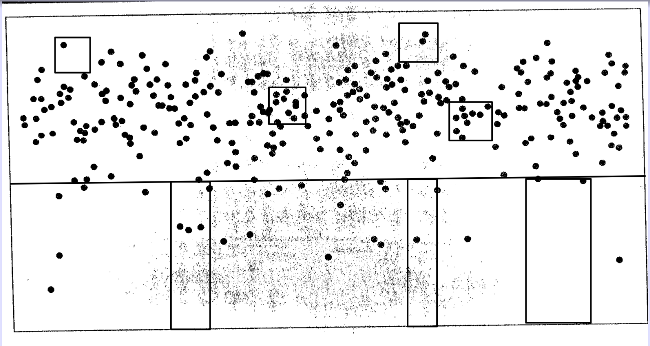


Separate analysis in each stratum

$$\widehat{D}_{high} = \frac{Objects}{Area} = \frac{1+12+2+9}{1.7+1.7+1.9+2.2} = 3.2 \quad se(\widehat{D}_{high}) = 1.40$$

$$\widehat{D}_{low} = \frac{Objects}{Area} = \frac{4+2+2}{8.5+5.5+13.8} = 0.28 \quad se(\widehat{D}_{low}) = 0.10$$

SE is computed differently than previously , i.e. NOT $\frac{s}{\sqrt{n}}$.

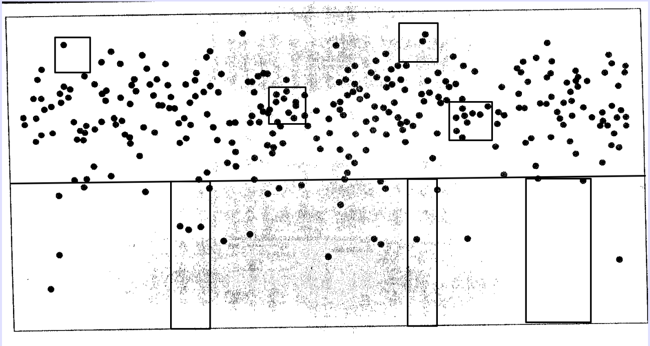


Rollup

Stratum	$\widehat{D}$	$se(\widehat{D})$	Area	$\widehat{N}$	$se(\widehat{N})$
High	3.20	1.40	138	441.6	193.2
Low	0.28	0.10	126	35.3	12.5
Total				476.9	193.6

$$se(\widehat{N}_{Total}) = \sqrt{se(\widehat{N}_{high})^2 + se(\widehat{N}_{low})^2}$$

Basic Statistical Review - Strip Transects



Example of a strip transect is in low density stratum above.

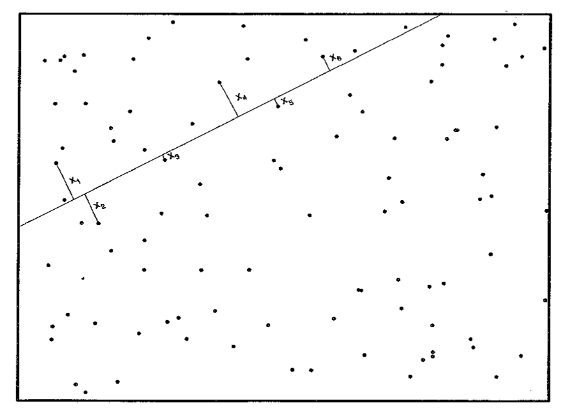
- Quadrats are long and narrow.
- Travel along strip and count objects on either side.
- With strips of half-width w, TOTAL length L, and TOTAL n objects seen
  - $\widehat{D} = \frac{Count}{Area} = \frac{n}{2wL}$

Perils of strip-transects

- Must assume perfect detectability within strip, but no data collected to verify.
- Strips are typically long and narrow with higher travel time.
- Unable to use objects detected outside of strips.
- Sometimes half-width of strips is not well defined and difficult to estimate.

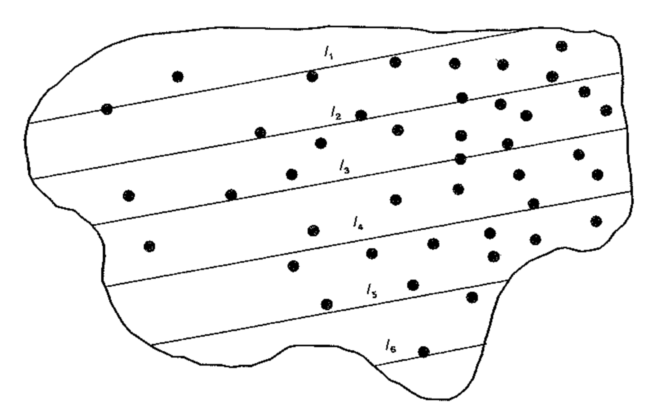
Distance Sampling - Intuitive Basis

Survey Protocol (one transect)



- Randomly place a transect line.
- Travel along line and look for objects.
- Record **perpendicular** distance to each object.
- All objects ON THE TRANSECT are detected.

Survey Protocol (multiple transects)



- Randomly place transects or systematic placement of transect.

Distance Sampling - Intuitive Basis

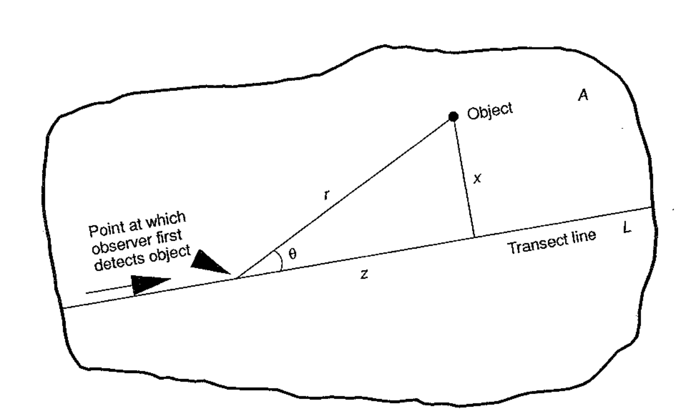
With strips of effective half-width w, TOTAL length L, TOTAL n objects seen, Average DETECTABILITY p

$$\widehat{D} = \frac{Adjusted\ \ Count}{Adjusted\ \ Area} = \frac{n/p}{2wL} = \frac{n}{2wLp}$$

- How is p estimated?
- How is w estimated?

Distance Sampling - Detail

Converting from angle and distance to perpendicular distance:

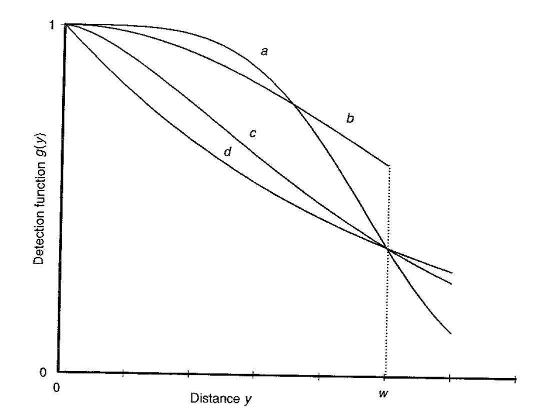


$$x = r\sin(\theta)$$

Distance Sampling - Type of data

- Ungrouped data
  - Actual distances measured to each object.
- Grouped data
  - Distances measured in bins.
  - Dealing with heaping (e.g. distances "tend" to occur at round numbers)
  - Dealing with data as distance intervals and exact distance not available.
- Data truncation, i.e. objects $> w$ are discarded because of unreliability and outliers.

Distance Sampling - Detection function



$g(y)$ = Probability of detection of object at distance y from transect line.

$g(0) = 1$ is critical assumption (but will be relaxed later).

$g(y)$ may also depend on covariates

Use the data to estimate $g(y)$!

Using Program DISTANCE - an introduction

Program DISTANCE

- Data and results stored in PROJECTS
  - Geographic Projects include GIS data and needed for survey design
  - Non-geographic projects can include analysis but not design
- Projects include Project File (*.dst) and Data Folder (.dat)
- Data Folder includes
  - Data file *.mdb + other formats
  - Shapefiles (if geographic project)
  - R folder (e.g. when using MRDS)

Example 1

Create a new folder and copy over the Example1.xls file into the folder.

Program DISTANCE - Example 1

Open the Excel file and notice the data columns and format

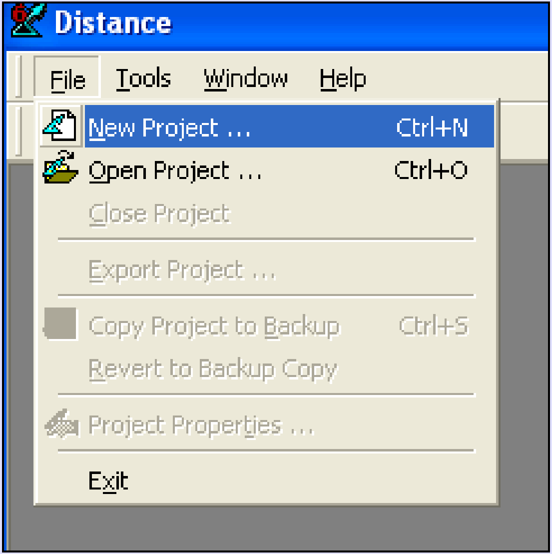
- Stratum label (in this case only 1 stratum)
- Stratum area km^2
- Transect Label (in this case 12 lines)
  - Label each transect separately, even if across strata, e.g. S01L01, S02L01 etc.
- Transect Length (kilometers)
- Distance to object (meters)
  - Notice how a transect is entered that had NO objects (transect 11)
- Other covariates (in this case cluster size)

Notice the data MUST be sorted to group all records appropriately.

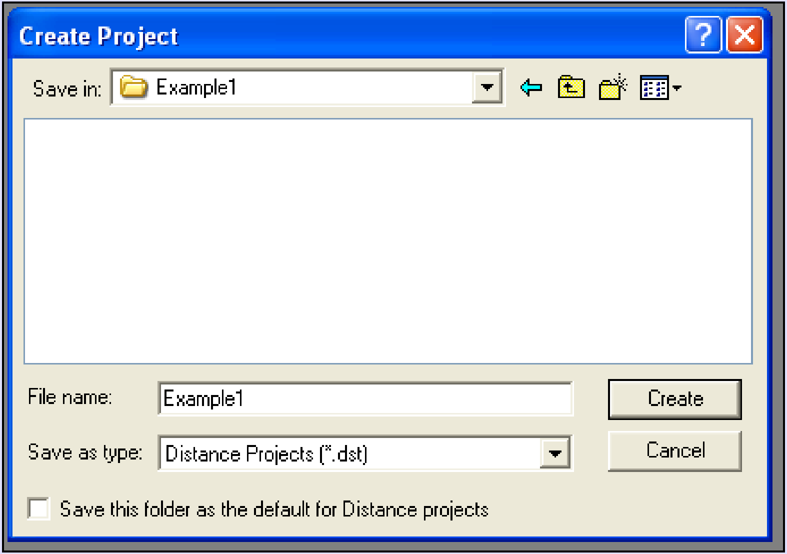
Program DISTANCE - Example 1

Save file as a *.csv file in the same folder.

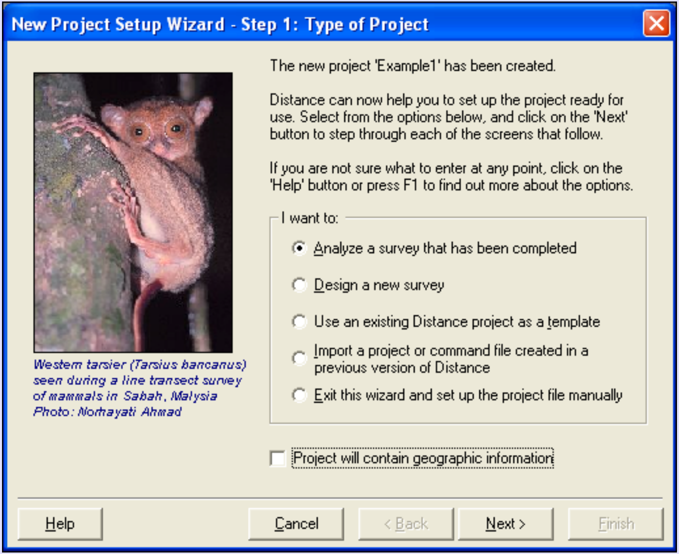
Create a new project with a suitable name . . .



. . . saved in an appropriate location.

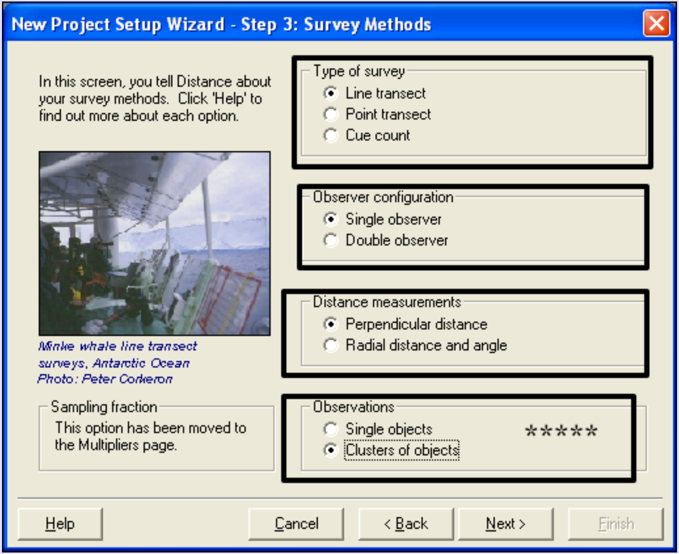


This will be a new project with new data.



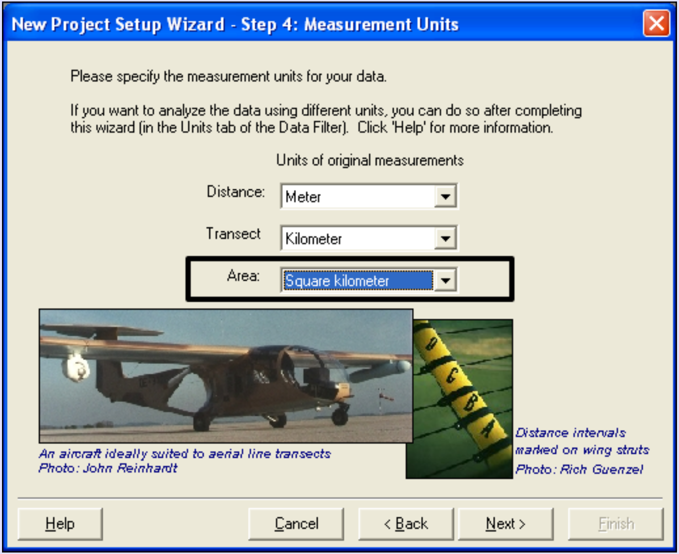
Program DISTANCE - Example 1

. . . collected as specified . . .



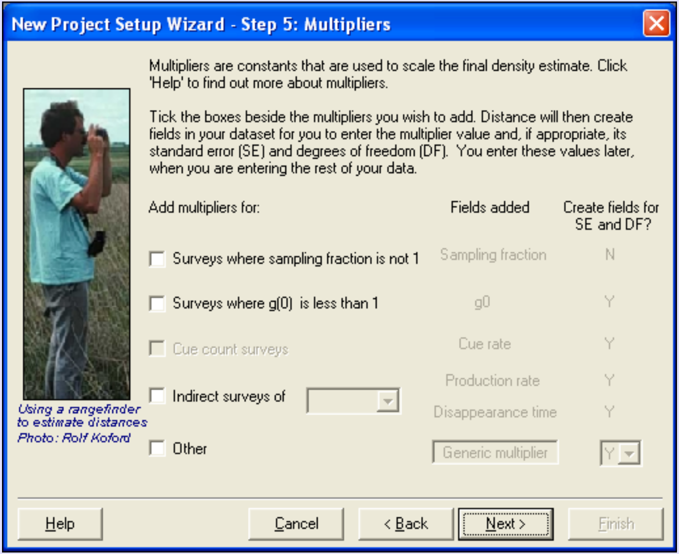
Program DISTANCE - Example 1

. . . With appropriate units . . .



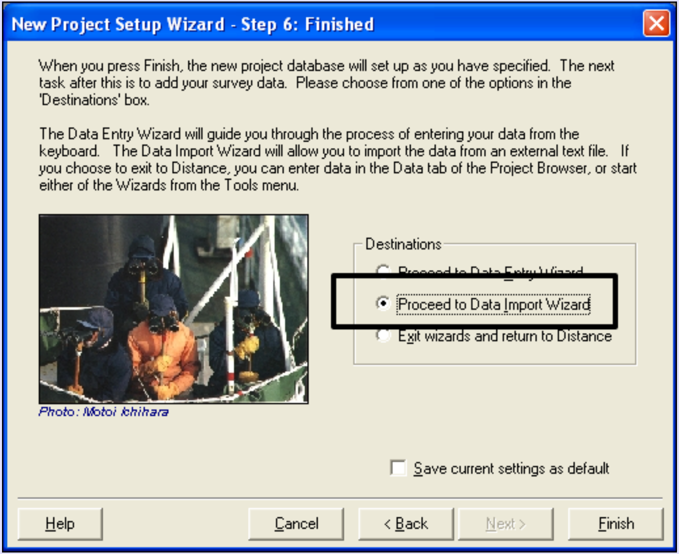
Program DISTANCE - Example 1

… and no unexpected problems …



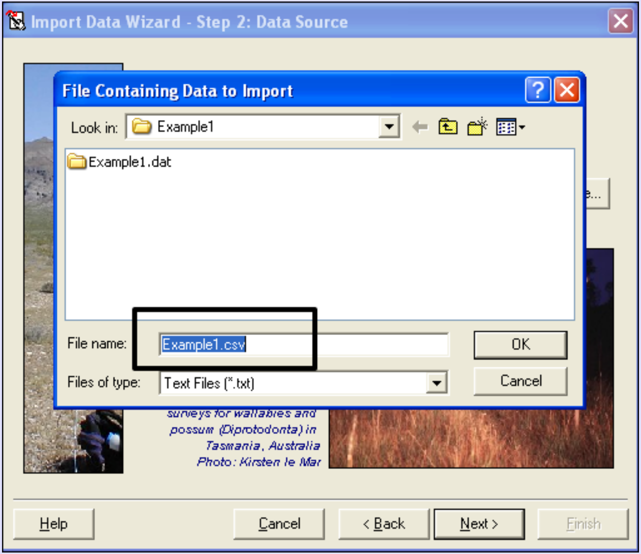
Program DISTANCE - Example 1

Import the data from a pre specified text file:



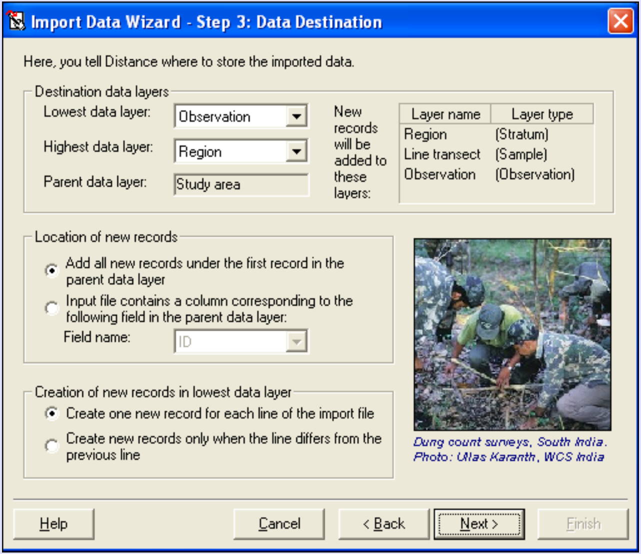
Program DISTANCE - Example 1

. . . specify the file name . . .

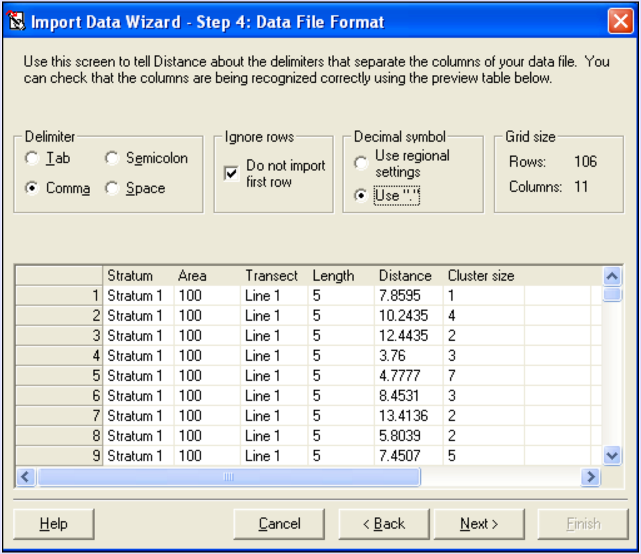


Program DISTANCE - Example 1

. . . use the defaults . . .



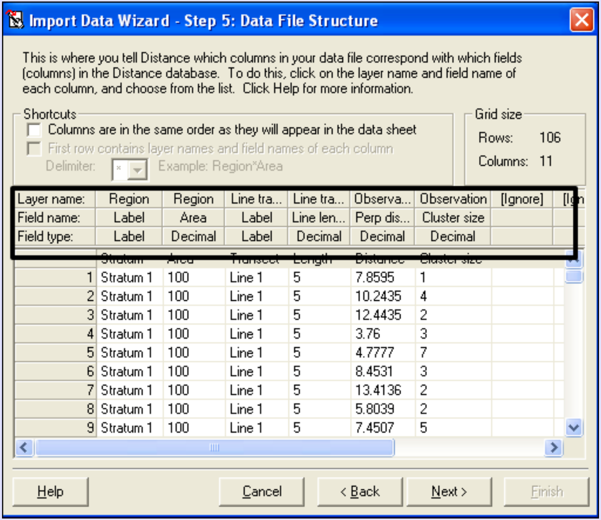
. . . Specify the delimiters, etc. . . .



	Stratum	Area	Transect	Length	Distance	Cluster size
1	Stratum 1	100	Line 1	5	7.8595	1
2	Stratum 1	100	Line 1	5	10.2435	4
3	Stratum 1	100	Line 1	5	12.4435	2
4	Stratum 1	100	Line 1	5	3.76	3
5	Stratum 1	100	Line 1	5	4.7777	7
6	Stratum 1	100	Line 1	5	8.4531	3
7	Stratum 1	100	Line 1	5	13.4136	2
8	Stratum 1	100	Line 1	5	5.8039	2
9	Stratum 1	100	Line 1	5	7.4507	5

Program DISTANCE - Example 1

. . . Ensure that DISTANCE understands all the columns in your data set . . .



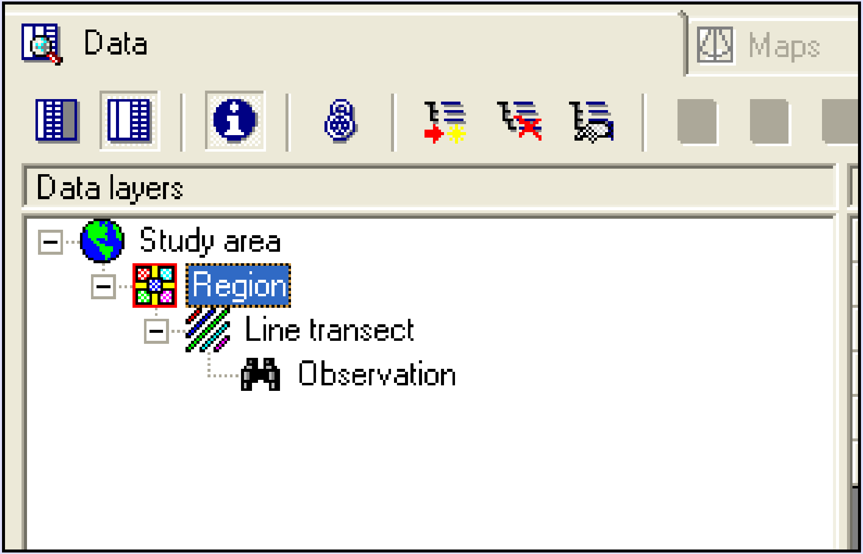
Note REGION = STRATUM

Program DISTANCE - Example 1

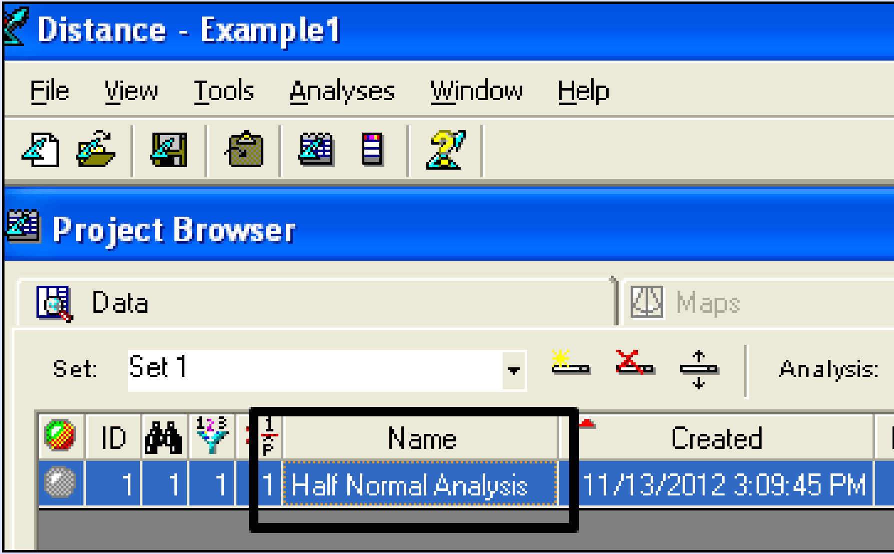
WHEW !

Program DISTANCE - Example 1

Our project has now been created. Notice how you view the data at different levels:

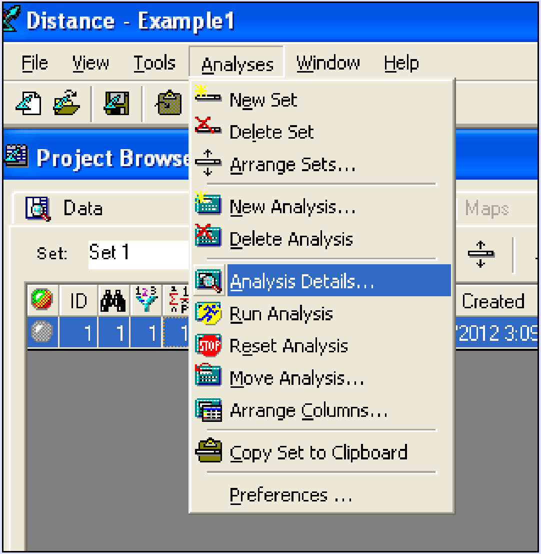


We will now fit several models to this data. Click on the **Analysis Tab**

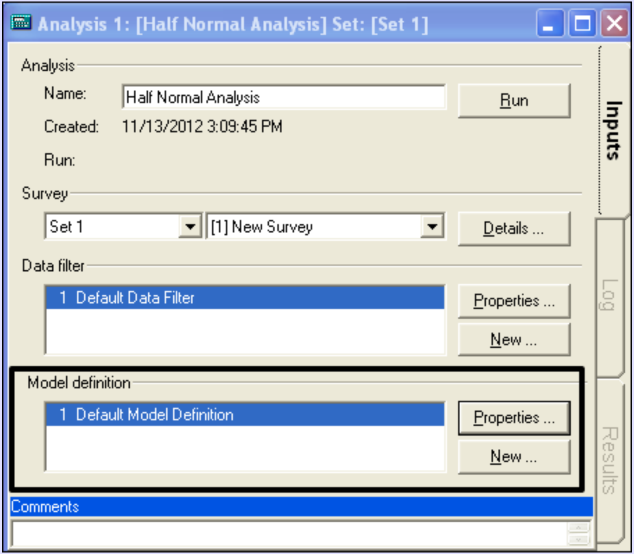


Program DISTANCE - Example 1 - Analysis

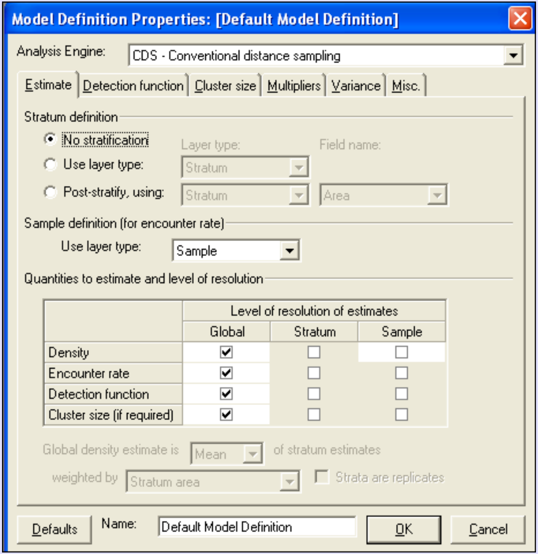
Use the pop-down menu to specify the analysis options for this model.



. . . and click on the button for model properties . . .

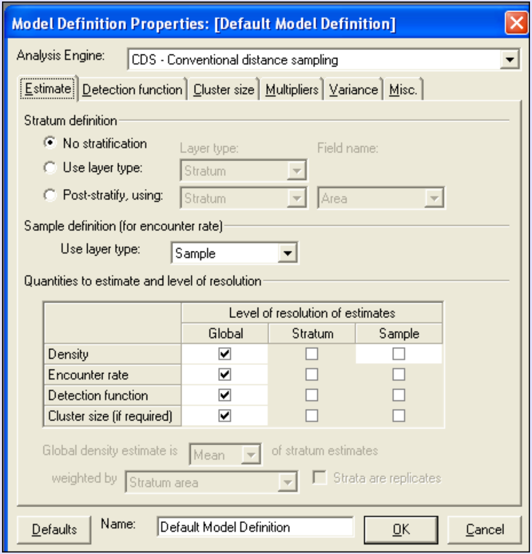


... click through the tabs to set the detection function shape ...



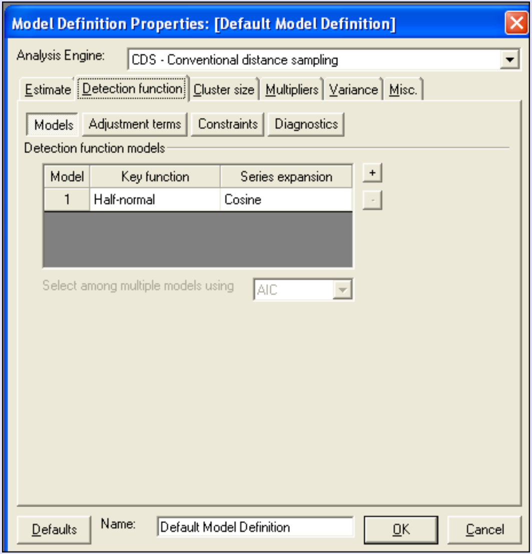
Program DISTANCE - Example 1 - Analysis

. . . click through the tabs to set the various options . . .



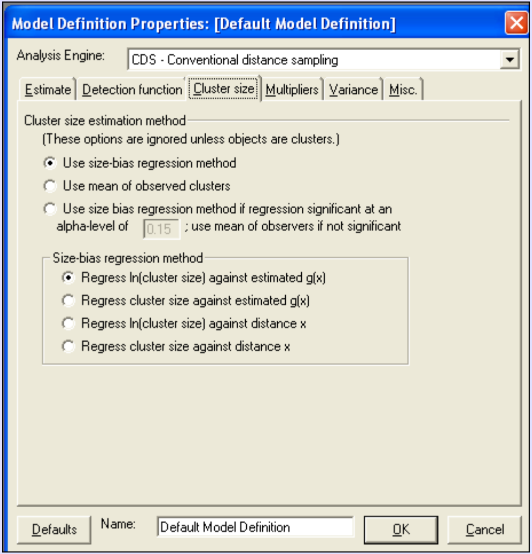
Program DISTANCE - Example 1 - Analysis

... click through the tabs to set the various options ...

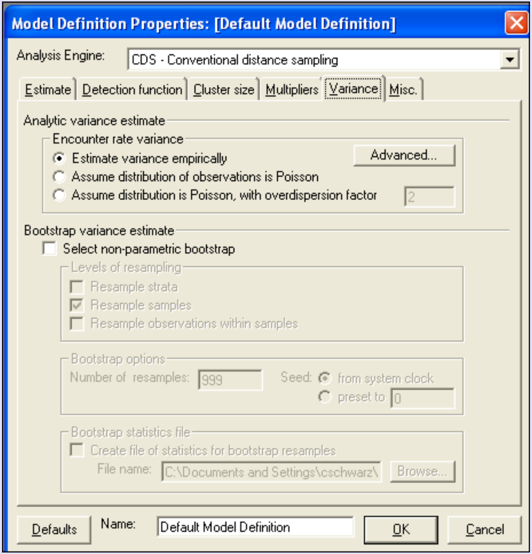


Program DISTANCE - Example 1 - Analysis

. . . click through the tabs to set the various options . . .

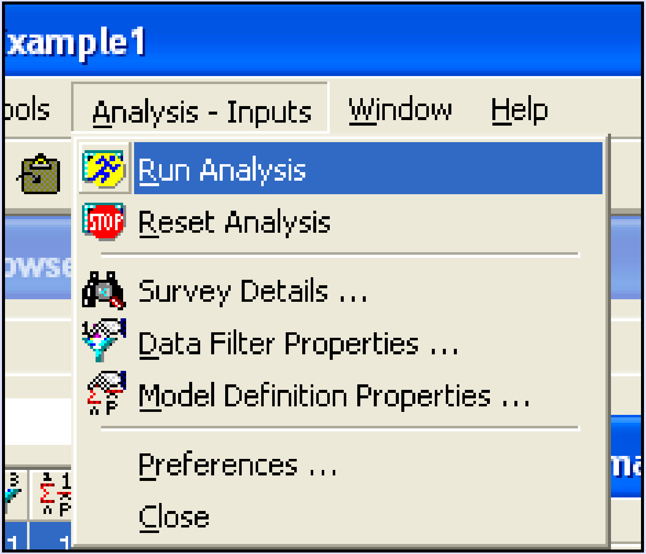


. . . click through the tabs to set the various options . . .



Program DISTANCE - Example 1 - Analysis

... Finally, ready to run the analysis ...



Look at log and results ...

Lots of results - let us take it slow!

Program DISTANCE - Example 1 - Analysis

Estimation options:

- ... *all data combined* implies pooling over all transects and observations to determine the detection function and other features of the model.
- Exact measurements are being used (i.e. not grouped data)
- Fitting a half-normal detection function (with adjustments – see later)

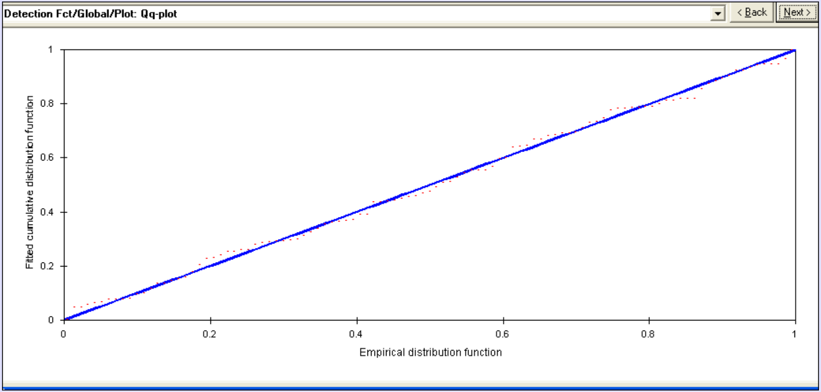
Program DISTANCE - Example 1 - Analysis

Detection function estimates:

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95 Percent Confidence Interval	
A(1)	9.980	0.6201			
f(0)	0.79978E-01	0.49471E-02	6.19	0.70754E-01	0.90404E-01
p	0.34907	0.21592E-01	6.19	0.30881	0.39457
ESW	12.503	0.77341	6.19	11.061	14.134

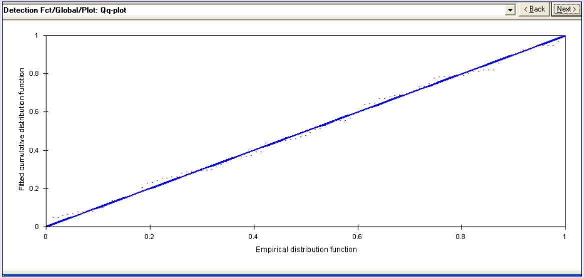
- $A(1)$ parameter describing detection function. For half-normal $=$ standard deviation of normal curve.
- $f(0)$ is the height of the detection function on the line. NOT the detection probability on the line which is assumed to be 1.
- p overall probability of detecting an object over the strip. Varies from 1 (at line) to 0 (beyond largest observation)
- ESW $=$ Effective Strip Width $=$ size of strip transect with equivalent properties where detection $=$ 100%

QQ plot of actual distances vs. theoretical distance is half-normal is correct detection function



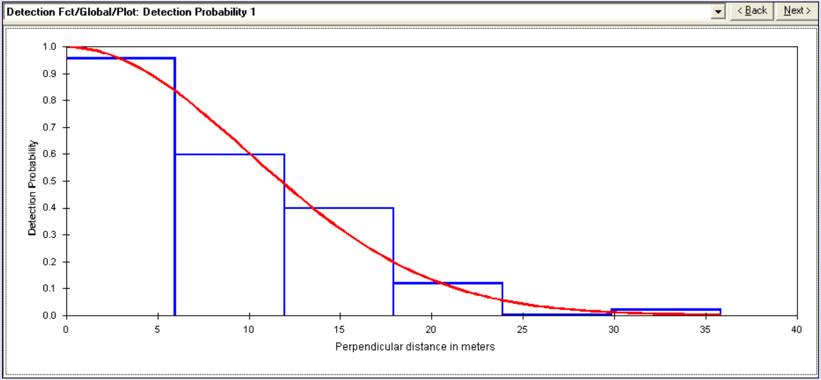
Program DISTANCE - Example 1 - Analysis

QQ plot of actual distances vs. theoretical distance is half-normal is correct detection function



- Points should lie close to the straight line
- Check for "heaping" which shows as horizontal straight lines in QQ-plot
- Format goodness-of-fit tests then follow
    - KS/ CM tests look at distance from points to line. p-values should be large

Plots of observed vs. predicted counts under different number of bins



- Look for large discrepancies between observed and expected counts.
- CAUTION: small counts in a bin can lead to false positive for lack of fit.
- You can change the bin boundaries.

Program DISTANCE - Example 1 - Analysis

Counts of observed vs. predicted counts under different number of bins

Cell i	Cut Points		Observed Values	Expected Values	Chi-square Values
1	0.000	5.97	48	47.30	0.010
2	5.97	11.9	30	33.42	0.350
3	11.9	17.9	20	16.68	0.660
4	17.9	23.9	6	5.88	0.002
5	23.9	29.8	0	1.46	1.464
6	29.8	35.8	1	0.26	2.147

Total Chi-square value = 4.6335 Degrees of Freedom = 4.00

Probability of a greater chi-square value, P = 0.32701

- Look for large discrepancies between observed and expected counts.
- CAUTION: small counts in a bin can lead to false positive for lack of fit (see last line in GOF Test 3 table).
- You can change the bin boundaries.

Program DISTANCE - Example 1 - Analysis

Dealing with clusters rather than individuals

```
Expected cluster size estimated based on regression of: log(s(i)) on g(x(i))


Regression Estimates
--------------------
Slope       =   -0.227763       Std error    =    0.187573
Intercept   =     1.08527       Std error    =    0.142639
Correlation=   -0.1188          Students-t   =   -1.21426
Df          =    103            Pr(T < t)    =    0.113713


Expected cluster size =   2.7116     Standard error     =  0.14952

Mean cluster size     =   2.8571     Standard error     =  0.13763
```

- Larger clusters tend to be more visible at larger distances - check if this is true
- Do NOT replace clusters by individuals as now detections of individual are NOT independent. Estimates will be unbiased, but reported SE will be too small.
- In this case, no evidence that observed cluster sizes varies by distance from transect line. [Look at plot on next window.]

Program DISTANCE - Example 1 - Analysis

Estimates of densities and abundance (Hurrah!)

Parameter	Point Estimate	Standard Error	Percent Coef. of Variation	95% Percent Confidence Interval	
DS	87.476	13.782	15.75	62.691	122.06
E(S)	2.7116	0.14952	5.51	2.4309	3.0247
D	237.20	39.592	16.69	167.71	335.47
N	23720.	3959.3	16.69	16771.	33547.

- DS - density of clusters
- $E(S)$ - estimate of average cluster size
- D - density of individuals
- N - abundance of individuals

Program DISTANCE - Example 1 - Analysis

Partitioning of uncertainty in estimates of D

```
Component Percentages of Var(D)
-------------------------------
Detection probability   :   13.7
Encounter rate          :   75.4
Cluster size            :   10.9
```

- Detection probability - because detection declines with distance
- Encounter rate - because animals are not uniformly distributed on environment
- Cluster size - because animals occur in clusters

WHEW !

Program DISTANCE - Summary of Steps

Get data into DISTANCE

- Get data into Excel and sort to group observations in transects/strata
- Export to text file (e.g. *.csv file)
- Create new project.
- Import data
  - Check that columns (fields) are imported correctly

Program DISTANCE - Summary of Steps - II

Specify the model and run the analysis

- Detection function shape is key - are there theoretical reasons to prefer one over the other?
- Check for model convergence after fitting each model
- Check goodness-of-fit to see that detection function is reasonable.

Program DISTANCE - Exercise

Great Argus pheasants (*Argusianus argus*) sampled studied in Sumatra (Winari 2002).

- An 800-ha study area,
- 144 line transects were placed, each 2,200-m long.
- Birds were detected on 45 of the 144 transects.

Fit a model with half-normal and exponential detection function.

Data from `http://coopunit.forestry.uga.edu/quant_cons_book/companion/chapter9/index.html`

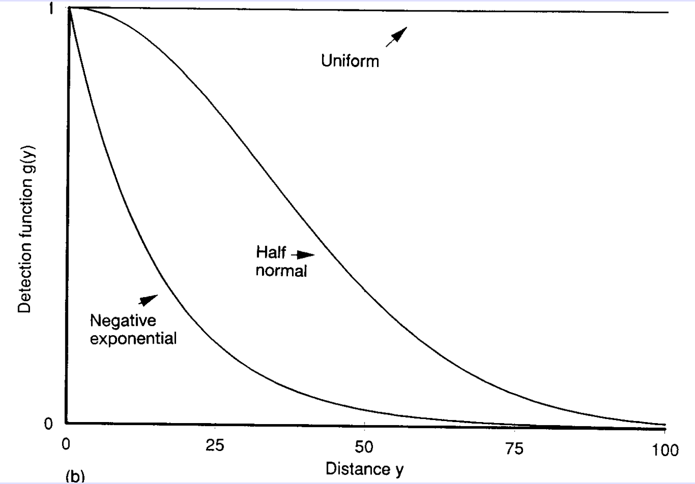
Distance Sampling - Some Theory

Distance sampling - Detection Functions

Theoretical considerations of detection function:

- Should be robust, i.e. flexible and take a wide variety of shapes
- Pooling robust, i.e. pooling data into different groups should not affect fit much.
- Shoulder at zero distance from line, i.e. $g'(0) = 0$ with no spikes
- Should be non-increasing with distance from the transect line.
- Should go to zero as distance increases.

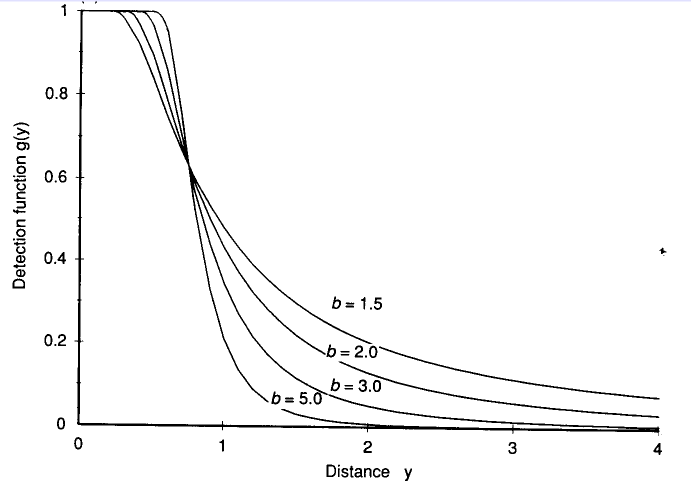
Detection functions has two parts:
"Key Function"



Distance sampling - Detection Functions

Detection functions has two parts:

"Key Function"



Distance sampling - Detection Functions

Detection functions has two parts:
"Series Expansion"

$$g(y) = key(y)[1 + series(y)]$$

where $series(y)$ is suggested by theoretical considerations for each key such as

- cosine series
- simple polynomials
- Hermite polynomials

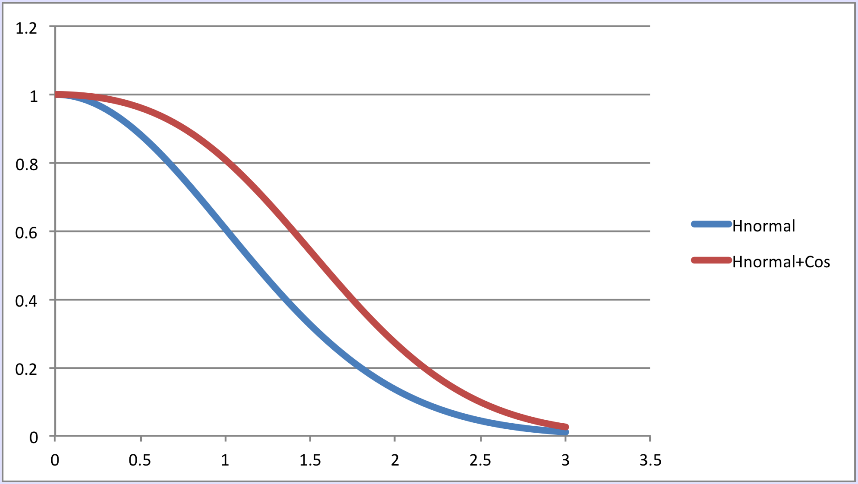
Distance sampling - Detection Functions

Suggest combinations:

Key	Series
Uniform $1/w$	Cosine, $\sum a_j cos(j\pi y/w)$
Uniform	Polynomial, $\sum a_j (y/w)^{2j}$
Half Normal, $exp(-y^2/2\sigma^2)$	Cosine
Half Normal	Hermite
Hazard Rate, $1 - exp(-(y/\sigma)^{-b})$	Cosine
Hazard Rate	Poynomial

- Uniform+Cosine = Fourier series method
- Half-normal has fairly rapid decline
- Hazard rate = marked shoulder, but caution as poor fit gives extreme bias.
- Avoid exponential except in "salvage" mode with poor data and spike at origin.
- DISTANCE can choose appropriate number of series terms.

Example of adding cosine (one term) series to half-normal detection function



Flexibility of key and series terms

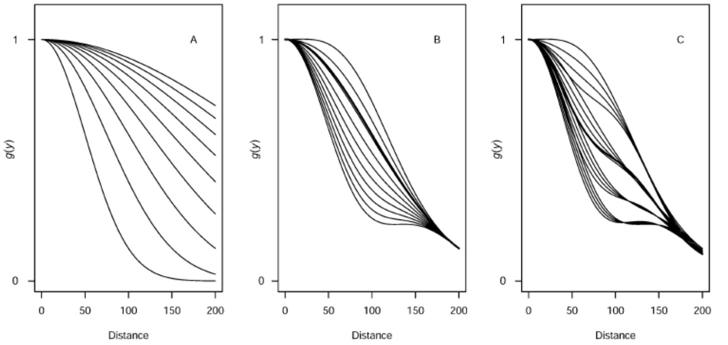
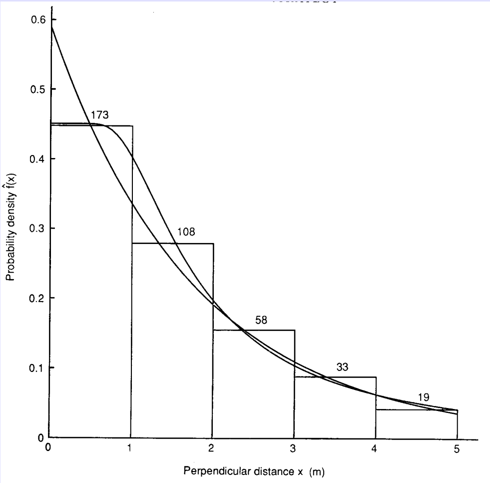


FIG. 1. Illustration of detection-function model formulation and the added flexibility provided by adjustment terms. (A) Half-normal detection function, with different scale parameters ($\sigma = 50, 75, \ldots, 225, 250$). (B) Half-normal with $\sigma = 100$ and several different values of a single cosine adjustment term ($j = 2$ in Table 1). (C) Half-normal with $\sigma = 100$ and several different values of two cosine adjustment terms ($j = 2, 3$ in Table 1).

Distance sampling - Detection Functions

Don't fret too much over choice as several function provide similar fit to data



(Negative exponential vs. hazard-rate model both don't show lack of fit.

Some data truncation at right may be needed!

Distance sampling - Intuitive Estimator

$$E[n] = D \times a \times P$$

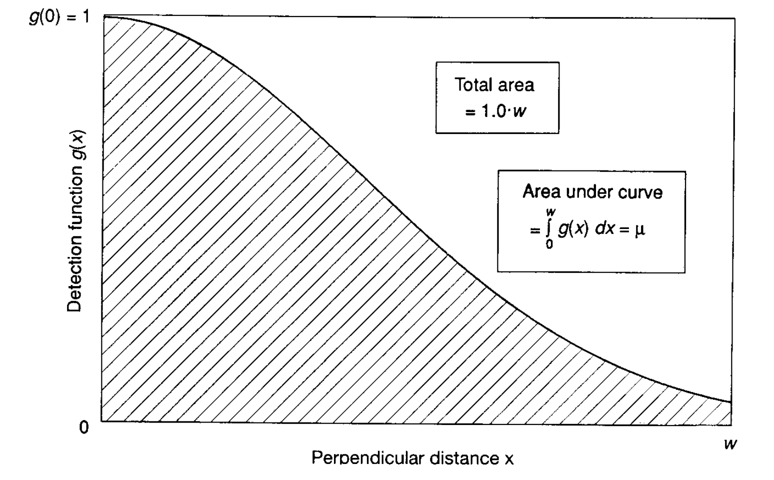
where

- $E[n]$ is expected number of animals seen
- D is the Density
- a is the area searched $= 2wL$ where w is half-width searched and L is length.
- P is the probability of detection an animal within area regardless of distance

This gives

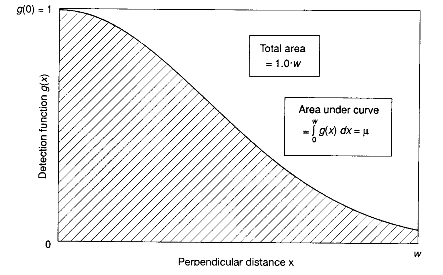
$$D = \frac{E[n]}{a \times P} = \frac{E[n]}{2wL \times P}$$

P determined from detection function:



$$Rectangle\ Area = 1 \times w \qquad P = \frac{Area\ Under\ Curve}{Rectangle\ Area} = \frac{Area\ Under\ Curve}{w}$$

Distance sampling - Intuitive Estimator



Note that because transect line placed randomly, distribution of observed distances $f(y)$ must have same shape but scaled to integrate to 1, i.e. $f(y) = g(y)/(Pw)$ for all y. So

$f(0) = g(0)/(Pw)$ or $P = g(0)/(f(0)w)$. But $g(0) = 1$ giving $P = \frac{1}{f(0)w}$. Finally (!)

$$D = \frac{E[n]}{a \times P} = \frac{E[n]}{2wL \times P} = \frac{E[n]}{2wL \times \frac{1}{f(0)w}} = \frac{E[n]f(0)}{2L}$$

$$\widehat{D} = \frac{n\widehat{f(0)}\widehat{E[S]}}{2L}$$

Implications:

- Not necessary to specify w.
- Likelihood only depends on detection function and distances from line
  - n and L play no part in likelihood to determine detection function.
  - This has implications in stratification model selection (see later slides)

Distance sampling - Obtaining Estimates

Maximum Likelihood Estimation used to estimate DETECTION function only.

- Extracts all information from data
- Produces smallest possible SEs
- In large samples, 95% confidence intervals formed as $Est \pm 2SE$

BUT

- For all but the trivial cases, numerical method (computer) must be used.
- Can be sensitive to outliers.

Distance sampling - Obtaining SEs

$$\widehat{var}(\widehat{D_{Line}}) = \widehat{D}^2 \times \left\{ \frac{\widehat{var}(n)}{n^2} + \frac{\widehat{var}(\widehat{f(0)})}{[\widehat{f(0)}]^2} + \frac{\widehat{var}(\widehat{E[s]})}{[\widehat{E[s]}]^2} \right\}$$

SE of density has multiple components, but largest source is typically from the encounter rate, the number of animals per unit length. SEs of encounter rates are obtained by:

- Empirical variation over transect lines, i.e. variation among n_i/L_i.
- Assuming a distribution of animals (e.g. Poisson process) but animals often show some degree of aggregation or dis-aggregation.
- Bootstrap.

Try the empirical methods with reasonable large sample sizes. Nothing works well with very small sample sizes.

Multiple models can be fit to the data

- Different detection functions (e.g. half-normal vs. hazard-function)
- Is the "series" modification needed for detection function?
- Same or different detection functions across strata?
- With or without covariates affecting detection (e.g. cluster size)

Distance sampling - Model Selection

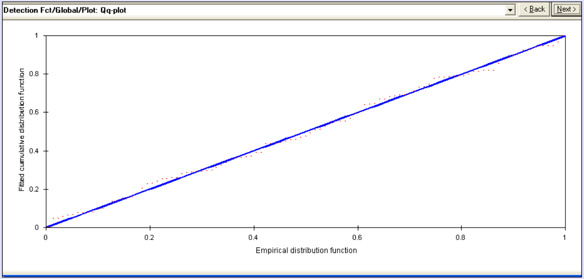
Akaike Information Criteria (AIC) provides a unified way for model selection:

- Fit each model. Get the L=log-likelihood (measure of fit) and k=number of parameters.
- $AIC = -2L + 2k$ measures tradeoff between fit and complexity.
- Smaller AIC is "better" (use arithmetic comparison, e.g $-4 < -3$, but differences of less than 2 or 3 are not "real"
- Model weights can be obtained giving weight of evidence of each model (within the set)
- Model averaging can be used to "average" estimates over different models.

BUT

- AIC does not guarantee a well fitting model – check the fit to the data

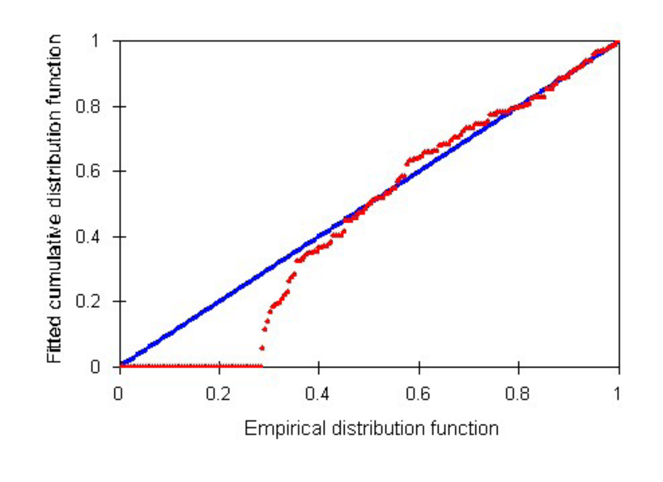
QQ-plots



- Better than observed vs. expected as no grouping required.
- Fit the observed percentile vs. the theoretical percentile
- A good fit lies close to the straight line

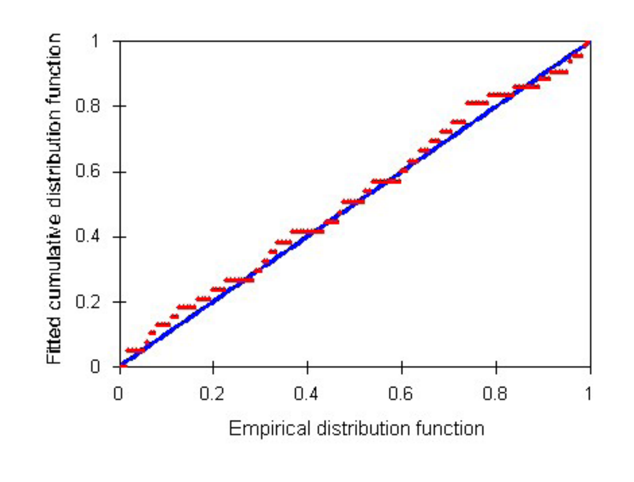
Distance sampling - Goodness-of-fit

QQ-plots



Spike at distance 0 (!)

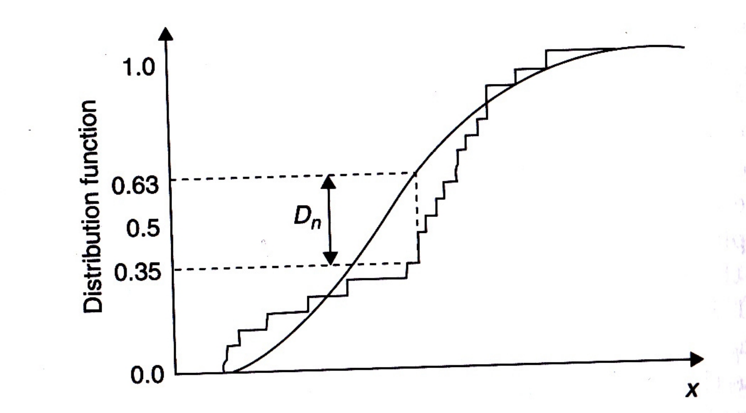
QQ-plots



Reasonable fit, but distances appear to be rounded (several points in steps).

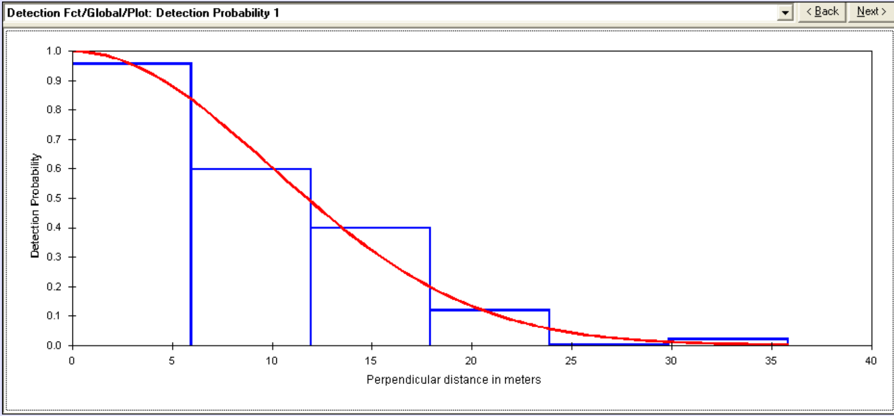
Distance sampling - Goodness-of-fit

Tests for goodness-of-fit (1)



K-S statistic looks at maximum distance between the theoretical and actual cumulative distribution curves.
C-M statistic looks at sum of squared differences between the theoretical and actual cumulative distribution curves.

Tests for goodness-of-fit



Divide the data into bins, count the observed number in each bin and the expected number in each bin (see next slide):

Distance sampling - Goodness-of-fit

Tests for goodness-of-fit

Cell i	Cut Points		Observed Values	Expected Values	Chi-square Values
1	0.000	5.97	48	47.30	0.010
2	5.97	11.9	30	33.42	0.350
3	11.9	17.9	20	16.68	0.660
4	17.9	23.9	6	5.88	0.002
5	23.9	29.8	0	1.46	1.464
6	29.8	35.8	1	0.26	2.147

Total Chi-square value = 4.6335 Degrees of Freedom = 4.00

Probability of a greater chi-square value, P = 0.32701

- Cut-points chosen by DISTANCE to give $\sqrt{n}$, $2/3\sqrt{n}$ and $3/2\sqrt{n}$ intervals.
- Expected value = total probability in interval from distance density function $\times$ total sample size.

Distance sampling - Goodness-of-fit

Tests for goodness-of-fit

Cell i	Cut Points		Observed Values	Expected Values	Chi-square Values
1	0.000	5.97	48	47.30	0.010
2	5.97	11.9	30	33.42	0.350
3	11.9	17.9	20	16.68	0.660
4	17.9	23.9	6	5.88	0.002
5	23.9	29.8	0	1.46	1.464
6	29.8	35.8	1	0.26	2.147

Total Chi-square value = 4.6335 Degrees of Freedom = 4.00

Probability of a greater chi-square value, P = 0.32701

- $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$ and compared to a ch-square distribution with (number of bins -1 -#parms) *df*.
- Evidence of lack-of-fit if the *p*-value is small.
  CAUTION. Be wary of influence of bins with small expected counts (< 2).
  DISTANCE may pool some of the bins with small counts.

Whew!

Review output from Example 1 - Questions ???

Distance sampling theory - Summary - I

- Key action is to find a "good" distribution to the detection distances
  - Shoulder near zero
  - Non-increasing with distance
  - Goes to zero with distance
- Don't fret too much about exact form – once you have a "good-enough" that will suffice.
- DISTANCE will automatically determine number of "series" terms needed for key function.

Distance sampling theory - Summary - II

- Maximum-likelihood estimation (MLE) done numerically.
  - Can be sensitive to outliers (large distances) where single object with small probability of detection gets magnified.
  - Consider truncation at sensible distances.
- Variance estimation (SEs) found using empirical variation among transect lines or bootstrapping.
- Model selection (AIC) to choose among detection functions within models and among models over multiple strata.
  - Data must be the same for all models, i.e. cannot compare models with no truncation with models with truncation.
- LOOK CAREFULLY at goodness-of-fit measures
  - Chi-square measures can be sensitive to small counts in bins.

Multiple Models in DISTANCE

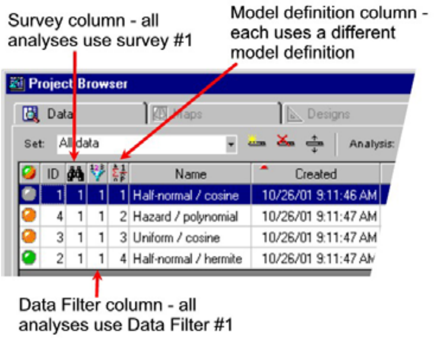
Multiple models can occur for a single data set because

- Different detection functions (e.g. uniform vs. half-normal)
- Different view of data (e.g. all of data vs. truncated data)
- Different strata (e.g. is detection function the same in all strata)

Distance sampling - Analysis components

Each analysis has 3 components

- *Survey* - type of survey that was conducted (see later)
- *Data filter* - which parts of the data are used
- *Model definitions* - which detection function, method of estimating SEs, etc.

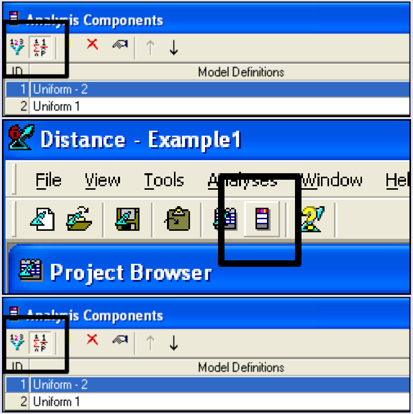


Note: NAME of model is ARBITRARY, so use meaningful notation

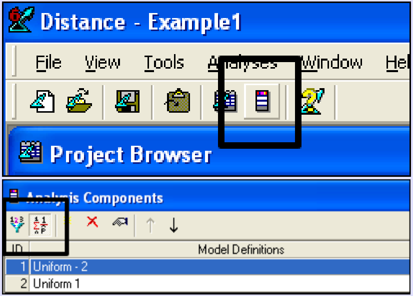
Distance sampling - Analysis components

Two ways to set analysis components

- Within an *Analysis* definition by clicking on the **New...** button
- Using the *Analysis Components* window

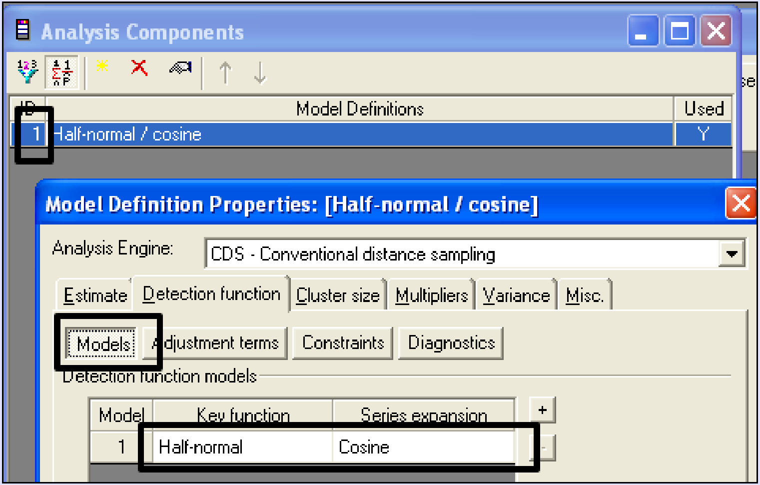


Create two different detection functions for Example 1 – Uniform and Half-normal using the *Analysis Components* window.



Distance sampling - Analysis components

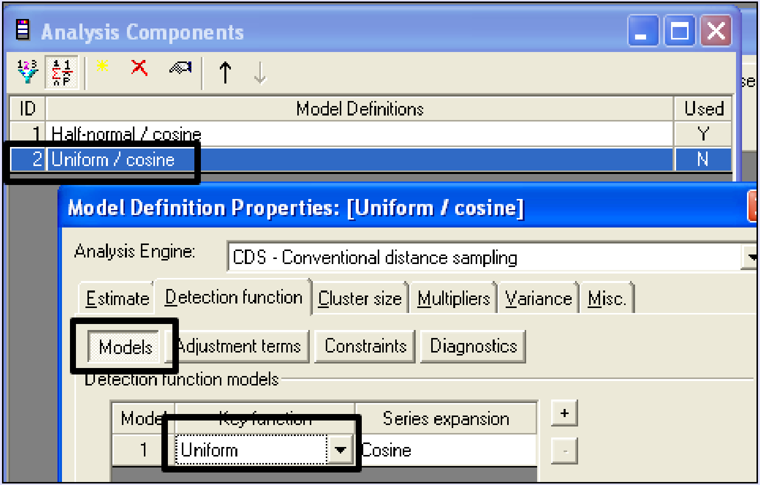
Click on Analysis # to edit, or *New* to create new items.
Specify the **half-normal** model.



Specify a meaningful NAME!

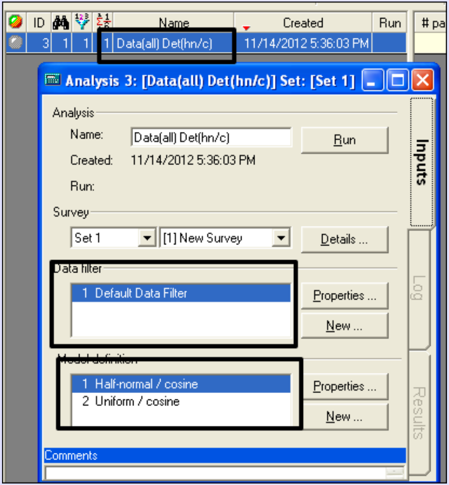
Distance sampling - Analysis components

Click on Analysis # to edit, or *New* to create new items.
Specify the **Uniform** model.



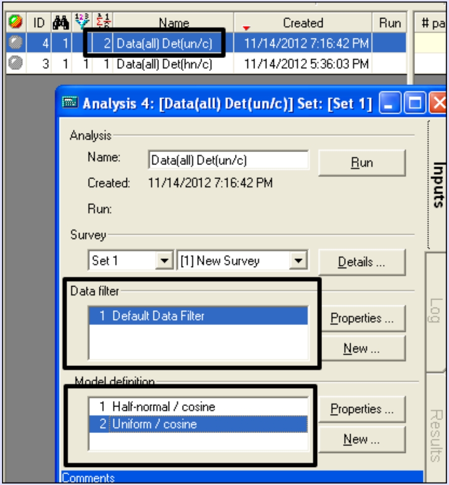
Specify a meaningful NAME!

Create new analysis with same (default) data filter, but different detection functions.



Specify a meaningful NAME for the analysis

Create new analysis with same (default) data filter, but different detection functions.



Specify a meaningful NAME for the analysis

Distance sampling - Run models

Run each model.

	ID				Name	# params	Delta AIC	AIC	ESW/EDR	D	D LCL	D UCL	D CV
	4	1	1	2	Data(all) Det(un/c)	2	0.00	635.66	13.40	222.563	154.522	320.565	0.179
	3	1	1	1	Data(all) Det(hn/c)	1	1.32	636.98	12.50	237.197	167.711	335.472	0.167

- Notice warning light beside one of the models.
- Notice ranking of models by AIC

Explore the two models to check for fit and estimates.

Uniform detection key function with cosines shows flexibility in specification!

Distance sampling - Run models

Add an analysis with **Negative exponential** with cosine series adjustment.

Add an analysis with **Negative exponential** with NO series adjustment.

Hint: Use a manual adjustment with 0 terms.

Distance sampling - Run models

Run all the models.
Sort by AIC.

	ID				Name		# params	Delta AIC	AIC	ESW/EDR	D	D LCL	D UCL	D CV
	4	1	1	2	Data(all) Det(un/c)	11	2	0.00	635.66	13.40	222.563	154.522	320.565	0.179
	5	1	1	3	Data(all) Det(NE/c)	11	3	1.31	636.97	11.97	245.037	156.156	384.506	0.228
	3	1	1	1	Data(all) Det(hn/c)	11	1	1.32	636.98	12.50	237.197	167.711	335.472	0.167
	6	1	1	4	Data(all) Det(NE/none)	11	1	8.61	644.27	8.34	317.555	212.044	475.567	0.202

- Be sure to look at goodness-of-fit statistics etc.
- You can change the information displayed on each half of the output window.
  - RECOMMEND you always display the total detections used.
- No easy way to model average estimates of Density over different models, but see Distance Manual.
- Data is current “locked” to prevent doing silly things, but you can correct data

Distance sampling - Data filters

Data filters are useful to:

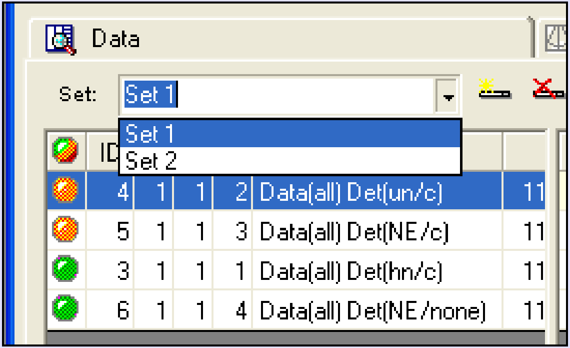
- Deal with outliers, especially long distances away that are "inflated" by detection probability.
  - Empirical evidence that $g(w) = .15$ works well, so discard about 5-10% of data values.
- Deal with subsets of your data (e.g. males vs. females or area A vs. area B)
- Convert from "exact" distances to intervals.

CAUTIONS:

- Put different filters in different sets so you don't compare models across different data sets!

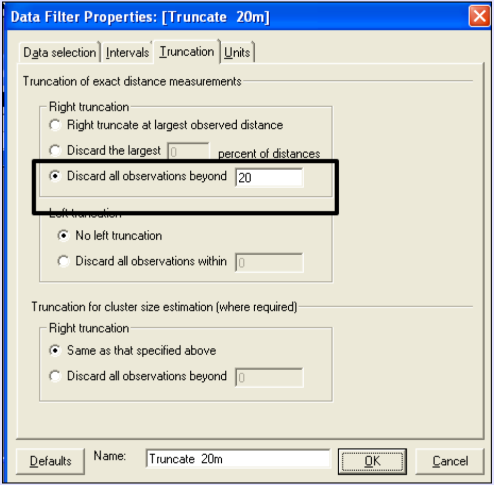
Distance sampling - Data filters

Create a new set



Distance sampling - Data filters

Add a data filter to truncate data at 20 m.



Distance sampling - Data filters

Left truncation?

- Used when detection on line is NOT certain, but higher further away.
- E.g. animals freeze or hide when you walk along the line.
- Detection function is back extrapolated to transect line.
- Alternatively, if detection at distance L' is 100%, subtract L' from observed distances and proceed.
  - Effective survey width is smaller, which causes no bias in estimates.

Distance sampling - Data filters

Do analyses using some of the detection functions created earlier.
Sort by AIC.
Compare results to previous set's results.

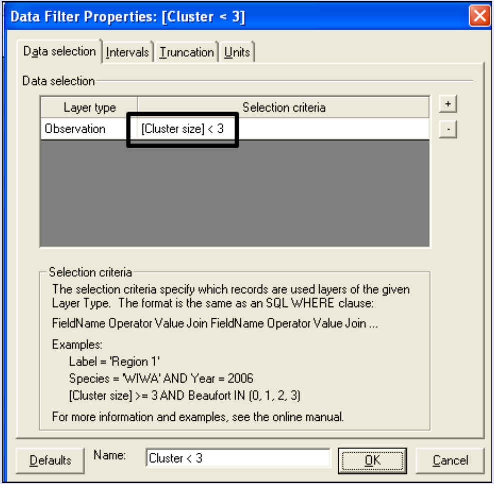
Distance sampling - More complex fIlters

A SQL-like query language can be used to build more complex filters.

- Selection is [Field Name] Operator Value
  - Good practice to put all field names in [ ] esp. if field name contains blanks.
  - Operators are the usual $<$, $>$, $=$, $<>$, $<=$, $>=$, *IN*, etc.
  - Values are numbers or ‘text’
- Join *Selection*s with AND, OR,
- Refer to DISTANCE manual for more details.

Distance sampling - More complex filters

Select data where cluster size is < 3 only.



Look at the output

Assumptions for Distance Sampling

Distance sampling - Assumptions

0. No assumptions made about distribution of animals on landscape!

- D describes the density over the landscape.
- Not necessary that animals distribute themselves independently, i.e. aggregated distributions are allowed.
- RANDOM placement of transects in large study areas implies uniform distribution of perpendicular distance from transect line.
- In small areas, strip width may go off site, leading to potential bias.

Distance sampling - Assumptions

1. Animals on transect line are detected with probability $= 1$.
   - Blind strip along line (aerial surveys, animals hide) dealt with by left truncation.
   - If $g(0) < 1$ then estimates of D are biased downwards.
     - If a correction factor is available, specify a multiplier, (e.g. only males visible) and sex ratio is 1:1.
     - Multiple observer/ platform methods.
   - Caution about "guarding the centerline" can lead to detection functions that are hard to model.
     - High detection rates close to transect line, rapid fall off, then usual decline is hard to model.

2. Animals detected at initial location
   - Movement of animal independent of observer causes little bias if movement is slow
     - CAUTION: Movement could cause an object to be counted twice (herding).
     - CAUTION: Fast moving ($> 1/2$ of observer speed) animals lead to severe bias.
   - Counting of same animal on different transects is ok, as long as movement is random.
     - Animals moving to other transect "balance" animals moving away from other transect.
   - CAUTION if animals are attracted or frightened near the transect line this can lead to serious bias.
     - E.g. animals being herded as observer moves along transect line.
     - Histogram of distances often show hole along transect line.

Distance sampling - Assumptions

3. Measurements of distance from transect line are exact.
    - Even random errors introduce bias because "inflation" by detection probability is larger further from the transect line.
    - Visual estimates of distance are NOT very good, i.e. observers tend to underestimate distances.
    - Beware of heaping, i.e. lots of measurements at 100 m, but fewer at 99 m and 101 m.
    - Try using interval data (binned) (e.g. distances estimated by airplane struts).

Distance sampling - Assumptions

4. Beware of outliers.
   - Objects far away have large influence on estimates because of weighting by 1/prob of detection.
   - Check histograms and truncate.
     - Truncate at $g(w) \approx 0.15$ and truncate a KNOWN half-width w.
     - Truncation introduces no bias.

5. Objects are detected independently of each other.
   - Clusters of animals violate this assumption, but can be dealt with if cluster size is recorded.
   - Pair-bonds of birds where some members of cluster may be hidden?
   - Ill-defined clusters?
   - Clusters that are too large to count?

Distance sampling - Assumptions

6. Detection function is not spiked at 0.
   - Methods work best if there is shoulder to the detection function.
   - Negative exponential (with no series adjustment) is a last resort!

Single Transect Line

- Not possible to use empirical estimate of variance.
- Valid if want to make inference only about area sampled (i.e. along the transect line), but this is doubtful.
- Assume distribution of observations is Poisson with an over dispersion factor of about 3 if sample size is very small and empirical methods cannot be used.
- Refer to DISTANCE manual.

Distance sampling - Common Problems

Missing Data - I

- Missing Distances
  - Nothing entered into DISTANCE.
  - No bias introduced assuming not on line, but loss of precision.
  - If MCAR and on the line, estimate P and use as multiplier
- Line with no observations
  - Do NOT discard the transect.
  - Enter line with length and NO observations.
- Line could not be run due to weather or other factors.
  - Assume missing line is MCAR and do NOT enter in DISTANCE
  - If only part of line can be run, enter the observered portion into DISTANCE.

Distance sampling - Common Problems

Missing Data - II

- Missing Cluster Size
  - By design, i.e. record all clusters, but only count cluster size in every 3rd cluster.
    Enter -1 for cluster size;
    Objects are used to estimate cluster density, but non-zero sizes used to estimate mean cluster size which are then multiplied together.
  - Empty objects, i.e. look for dens and count number of occupants of dens.
    (a) Discard empty dens; use non-zero values; biased estimates of average den size, but fewer dens so ok
    (b) Use empty dens; but cannot adjust for size-biased sampling (not a problem in this example).
- Missing Survey effort (length of transect) – Go home early, nothing can be done.
- Missing Study area size – you can estimate D but not N.

Stratification

Distance sampling - Stratification

Benefits of Stratification

- Account for heterogeneity in density and improve precision and reduce bias.
- Account for heterogeneity in detection and improve precision and reduce bias.
  - Male vs. female
  - Small vs. large clusters
- Stratum specific estimates, e.g. by habitat area.

Costs of stratification

- Reduced sample size in each stratum leading to less precise (larger SEs) in each stratum.
- More parameters to be estimated leading to less precision (larger SEs)
- Model complexity is increased (more things to fiddle with)

Distance sampling - Stratification

Stratify by:

- Geographic region
- Environmental conditions (e.g. habitat)
- Cluster size
- Animal behavior or covariate
- Detection type (e.g. sighting vs. hearing)
- Observer

Distance sampling - Stratification

Types of Stratification

- Pre-stratification (defined in DISTANCE using *Stratum* Layer)
    - Delineate strata in advance.
    - Separate survey in each stratum.
    - Separate estimates in each stratum
        - Some parsimony by have encounter/ detection/ density same across strata
    - Rollup
- Post-stratification
    - Strata are defined after data is collected, e.g. by size of cluster or species detected.
    - Nothing statistically wrong with post-stratification!
        - But... in some cases rollup is impossible, e.g. habitat determination on the fly – what is the total area of each stratum?

Distance sampling - Pre-Stratification

Example of Stratification

- Implemented in DISTANCE using *Stratum* Layer.
  - Look at *Example-Stratification* folder and Excel spreadsheet. Area in km^2; length in km; distances in km.
  - Need area of EACH stratum
  - Don't forget to group data from stratum together; data from transect together, etc
- Save data to *.csv file
- Import into DISTANCE in the usual fashion

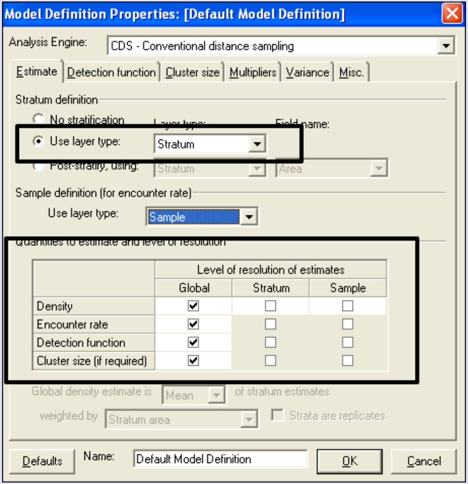
Distance sampling - Pre-Stratification

Data at the transect level.

Study area		Region			Line transect		
ID	Label	ID	Label	Area	ID	Label	Line length
ID	Label	ID	Label	Decimal	ID	Label	Decimal
n/a	n/a	n/a	n/a	ha	n/a	n/a	km
Int	Int	Int	Int	Int	Int	Int	Int
1	StratifiedData	1	Ideal Habitat	85000	1	1	87
					2	2	11
					3	3	16
					4	4	110
					5	5	15
					6	6	10
					7	7	16
					8	8	59
					9	9	10
					10	10	13
					11	11	56
					12	12	1
					13	13	80
		2	Marginal Habitat	600000	14	14	75
					15	15	100
					16	16	110
					17	17	100
					18	18	125
					19	19	100
					20	20	150
					21	21	110
					22	22	115
					23	23	150
					24	24	125
					25	25	110

Set up the analysis in the usual way by selecting the

- Data filter (use the default of all of the data)
- Model definition - specify a stratified estimation procedure



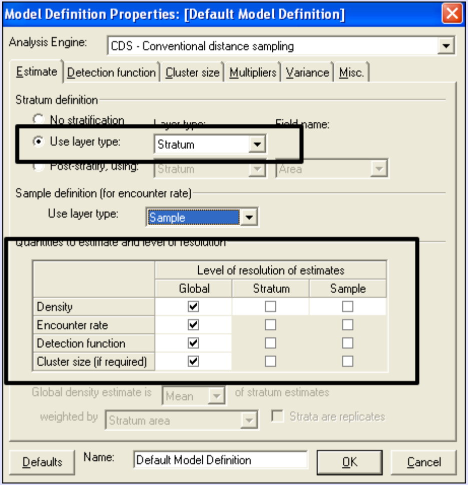
Distance sampling - Pre-Stratification

Specify at which level modeling of features of survey take place

- Global - same feature across all strata, e.g. one detection function for each stratum
- Stratum - different feature across strata, e.g. separate detection function for each stratum
- Sample - different feature for each transect, e.g. separate detection function for each transect

Distance sampling - Pre-Stratification

Start with the simplest model where everything is common across all strata:



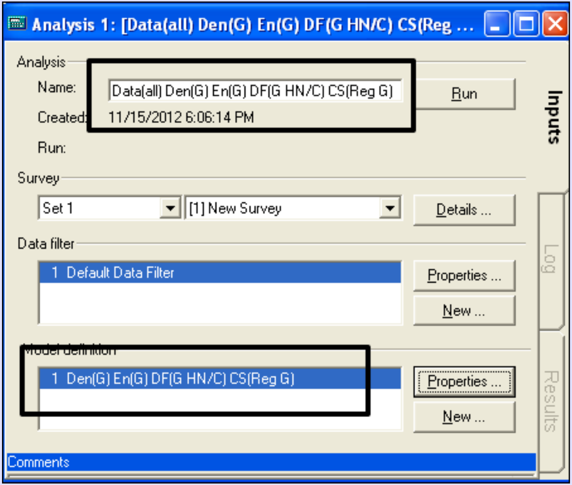
Choose a MEANINGFUL name for this model definition!

Den(G) En(G) DF(G, HN/C), CS(Reg G)

Choose a MEANINGFUL name for this analysis

Data(All) Den(G) En(G) DF(G, HN/C), CS(Reg G)

Run the model



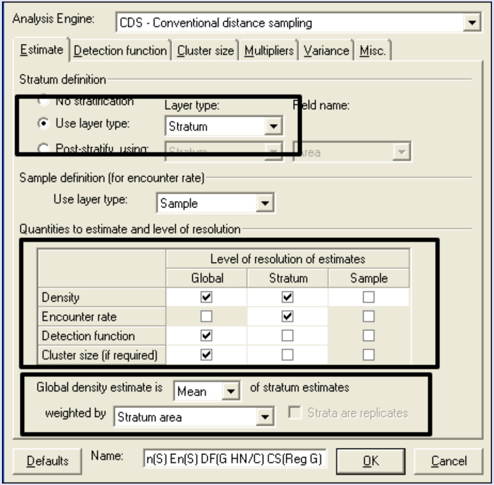
. . . and look at the results . . .

. . . and look at the results . . .

- A single detection function is estimated for all strata
- QQ plot actually doesn't look bad
- Goodness of fit plots don't look bad
- A single regression of cluster size vs. detection prob for all strata
- A single estimate of Density and Abundance over all strata

Distance sampling - Pre-Stratification

Create a model definition where Density varies by stratum, but Detection Function and other attributes are the same across strata.



Notice how estimates of density are combined over strata.

Choose a MEANINGFUL name for this model definition!

Den(S) En(S) DF(G, HN/C), CS(Reg G)

Choose a MEANINGFUL name for this analysis

Data(All) Den(S) En(S) DF(G, HN/C), CS(Reg G)

Distance sampling - Pre-Stratification

... and look at the results ...

- A single detection function is estimated for all strata
- QQ plot actually doesn't look bad
- Goodness of fit plots don't look bad
- A single regression of cluster size vs. detection prob for all strata
- A separate estimate of Density and Abundance for each stratum

. . . BUT . . . the AIC are IDENTICAL!

	ID				Name		# params	Delta AIC	AIC	ESW/EDR	D	D LCL	D UCL	D CV
	1	1	1	1	Data(all) Den(G) En(G) DF(G HN/C) CS(Reg G)		1	0.00	63.88	0.8[illegible]	0.506	0.295	0.866	0.269
	2	1	1	2	Data(all) Den(S) En(S) DF(G HN/C) CS(Reg G) 1		1	0.00	63.88	0.8[illegible]	0.444	0.230	0.857	0.316

Why?

- A single detection function is estimated for all strata which determines the likelihood.
- Different densities give same histogram of distances, but fewer/km of transect
- n/L which should be related to density plays NO part in the model fit!

$$\widehat{D_{pooled}} = \frac{(n_1 + n_2)\widehat{f(0)}}{2(L_1 + L_2)} \neq \frac{A_1 \frac{n_1 \widehat{f(0)}}{2L_1} + A_2 \frac{n_2 \widehat{f(0)}}{2L_2}}{A_1 + A_2} = \widehat{D_{strat}}$$

UNLESS $L_1/A_1` = L_2/A_2$ under proportional allocation.

So . . . there is no simple way to determine which of these two models is better.

Look at the individual estimates of the density to see how different th

Distance sampling - Pre-Stratification

Create a model definition where

Density varies by stratum,

Detection Function varies by stratum,

Cluster relationship with distance same across strata.

Choose a MEANINGFUL name for this model definition!

Den(S) En(S) DF(S, HN/C), CS(Reg G)

Choose a MEANINGFUL name for this analysis

Data(All) Den(S) En(S) DF(S, HN/C), CS(Reg G)

Run this model, but it FAILS!

```
Log messages - 0 warnings, 1 error
** Error: Incompatible resolution levels for estimation  of detection probability and cluster size **
```

Why?

- Different detection functions across strata cannot have same relationship of cluster size and detection!

Not all combinations of analysis options are sensible.

Create a model definition where

Density varies by stratum,

Detection Function varies by stratum,

Cluster relationship varies by stratum.

Choose a MEANINGFUL name for this model definition!

Den(S) En(S) DF(S, HN/C), CS(Reg S)

Choose a MEANINGFUL name for this analysis

Data(All) Den(S) En(S) DF(G, HN/C), CS(Reg S)

Run this model . . .

- Different detection functions (but same key/series) across strata, but each looks ok
- Separate slope of cluster size vs detection probability

AIC suggests that this is a more appropriate model.

4	1	1	4	Data(all) Den(S) En(S) DF(S HN/C) CS(Reg S)	1	2	0.00	57.01		0.415	0.235	0.734	0.278
2	1	1	2	Data(all) Den(S) En(S) DF(G HN/C) CS(Reg G)	1	1	6.87	63.88	0.88	0.444	0.230	0.857	0.316
1	1	1	1	Data(all) Den(G) En(G) DF(G HN/C) CS(Reg G)	1	1	6.87	63.88	0.88	0.506	0.295	0.866	0.269

Some difficulty in the fit with monotonicity of the detection functions so more work is needed . . ..

Distance sampling - Pre-Stratification

Summary

- Specify stratum level in advance in the data
- Look at global and stratum specific models, but not all combinations of effects are possible

Sometimes not possible to specify strata in advance

- DIfferent types of survey effort (observers, years, etc)
- Different types of objects (e.g. species, sex)
- Different sizes of objects (e.g. cluster size) rather than the regression approach.

Need a variable in the data set, either created in Excel and imported, or defined in DISTANCE.

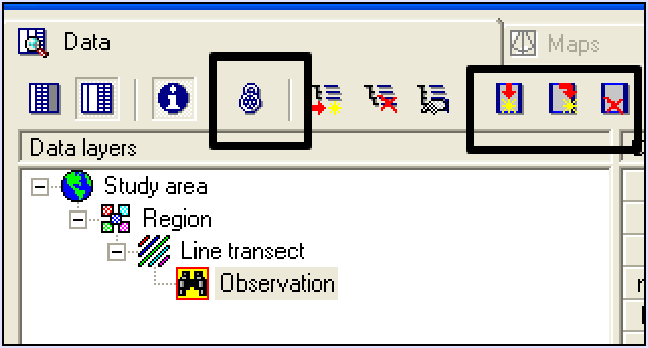
Sometime strata are **Replicates** (like random effects in ANOVA) (specified in DISTANCE as **Strata as replicates** which affects how the variance is computed for the mean over the replicates.

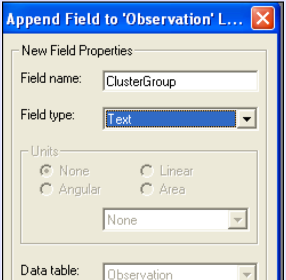
ONLY ONE TYPE OF STRATIFICATION ALLOWED in analysis.

How to get the post-stratification variable into DISTANCE?

- Create in Excel to begin with and import directly as before.
- Create new field (see below) and manually enter.

Create a new field. Go to **Data** layer; unlock; and create field.





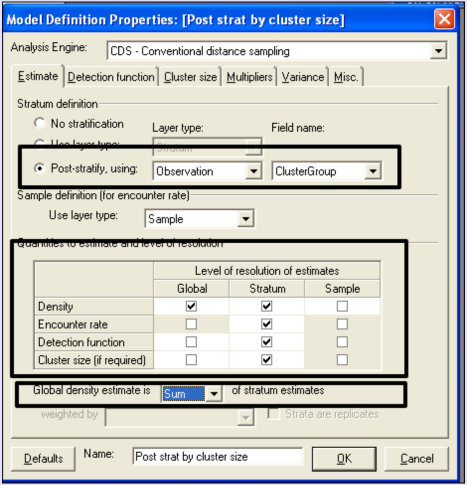
Distance sampling - Post-Stratification

Create Group as 1 or 2+ using appropriate codes. Note how no detections for some transect lines is handled.

Observation			
ID	Perp distance	Cluster size	ClusterGroup
ID	Decimal	Decimal	Text
n/a	m	[None]	n/a
Int	Int	Int	Int
19	1.22	1	1
20	0.68	2	2+
21	0.03	1	1
22	0.26	1	1
23	1.21	2	2+
24	0.52	2	2+
25	0.18	2	2+
26	0.36	2	2+
27	0.61	11	2+
28	0.61	1	1
29	0.6	1	1
30	0.1	2	2+

Distance sampling - Post-Stratification

Create new model definition in usual fashion. Notice specifications



Run the model . . . check the results . . . look at AIC, etc.

Summary

- Create variable outside and import or within DISTANCE
- Define analysis in usual way.
- Think carefully on how to combine the density estimates (either MEAN or SUM)

Distance sampling - Stratification

Super - Summary

- It seldom hurts to stratify (but no more than 2-3 strata)
- Stratification is coarse covariate.
- Different parts of analysis can be stratified
  - Reliable estimation of $f(0)$ requires large sample size, so try global models for detection function.
  - Cluster size stratification and Encounter rate requires smaller sample sizes.
- Pre- and post-stratification are possible, and pre-stratification is preferable to avoid unintentional biases.
- SUM or MEAN of abundances should be chosen as appropriate'
  - Geographic stratification usually wants sums of abundances.
  - Observer stratification usually wants average of abundances.

Line Transect Distance Sampling Analysis Summary

Line Transect Distance Sampling Analysis Summary

Collect data:

- Use good survey design to collect data (see planning later in the course).
- Enter data into spreadsheet
  - Group items together (e.g. observations from the same transect together)
  - Think of stratification variables (pre- and post-) and create
  - Include covariates (see later in this course)
  - Don't just drop missing data
    - Transects with NO observations have blanks for distance from transect line
    - Transects no run (e.g. due to weather) are not entered.
    - Partial transects have only part run entered.
    - Observations with missing distances not entered.
    - Cluster size 0, e.g. detect dens from transect line that is empty; enter 0 for size

Line Transect Distance Sampling Analysis Summary

Preliminary exploration of data:

- Don't plunge into analysis without exploration of data.
- Look for
  - Outliers (large distances have low probability of detection and high inflation).
  - Heaping (distances are 100 m but never 99 m or 101 m).
  - Shape of distance distribution - shoulder, monotonic decline, going toward zero.
  - Hole at zero - animal avoidance etc.
- Consider converting "exact distances" to intervals if problems exist such as heaping.
- Use smaller bins closer to the line; bin width increases with distance from line.

Line Transect Distance Sampling Analysis Summary

Import data into DISTANCE

- Save as *.csv or other text-delimited file.
- Separate folder for each project.
- Keep distance files together.
- Do regular backups.

Line Transect Distance Sampling Analysis Summary

Define *Data Filters*

- For line transect's truncate at about $g(w) = 0.15$
- Truncate at KNOWN half-width w.
- Left truncate if hole at 0?
- Truncate at fixed width value. [IMPORTANT.]
- Use meaningful *Data Filter* name.

Define *Model Definitions*

- Stratification (pre- or post-) gives lots of options.
- Detection function key not crucial as modified by series but avoid negative exponential because no shoulder.
- Deal with cluster size using stratification or regression methods.
- Multipliers for known effects (e.g. only see females and sex ratio is 1:1).
- Empirical variance methods are robust, but with small sample sizes, may be forced to specify Poisson-process.

Line Transect Distance Sampling Analysis Summary

Review the output:

- Check model fit using QQ plots; KS/CM tests; χ^2 tests, but don't rely exclusively on these
    - Horizontal lines in QQ plots are artifacts of heaping and rounding.
    - Small expected counts in χ^2 test can inflate test-statistic and lead to false-poor fit. Look at pooling small cell counts.
- See if estimates are reasonable.
- Rank models of SAME data using AIC. Model averaging must be done outside of DISTANCE.

  Cannot use AIC for model comparison with different data filters.

DON'T PANIC

Orangutan Data

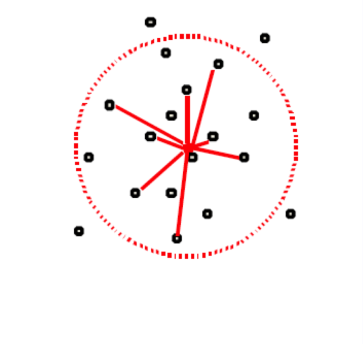
`http://www.wcsmalaysia.org/analysis/O-u_nests_DISTANCE.htm`

Read the *.pdf file in the directory

Distance Sampling - Point Transects

The same, but different.

Distance sampling - Point Transects



Point transects

- Locate points in landscape using good survey design
    - Randomization, replication, stratification.
- Measure distance to objects from point.
- Sighting radius may be fixed or variable.

Distance sampling - Point Transects

Why point transects?

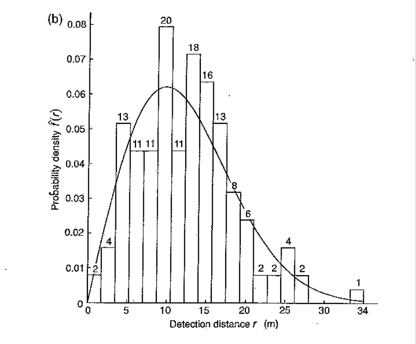
- Concentrate on detection without moving.
- Easy to travel to point where line transect must stay on transect line.
- Patchy habitats sampled more easily.
- Easier to model density as a function of habitat compared to line transects because each point is all one habitat while transects may traverse several habitats.
- Easier to estimate distance (flags on ground); not necessary to convert to perpendicular distances.

Why NOT point transects?

- Disturb (flush) animals away from point. Which would have been detected from the point?
- Detections while traveling between points are ignored.
- Inefficient for low density populations.
- Analysis is less robust to animal movement.
- Objects enter/leave area around point during survey.

Distance sampling - Point Transects

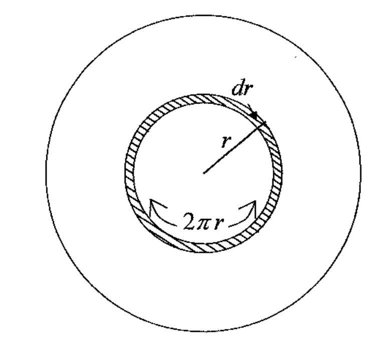
What is key difference in data and analysis?



Distribution of distance not monotonic declining like in line transects. Why?

Distance sampling - Point Transects

What is key difference in data and analysis?



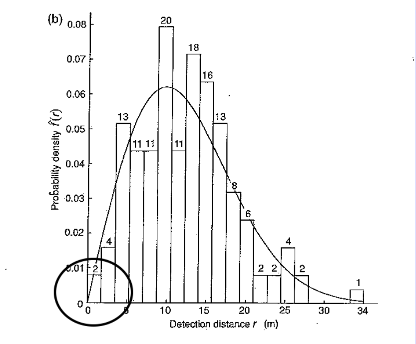
Detection declines with distance, but more area searched at distance r from point.
Observed distances is a convolution of detection and area swept.

$$f(r) \propto 2\pi r \Delta r \times g(r)$$

$$\widehat{D} = \frac{n\widehat{f'(0)}\widehat{E[s]}}{2\pi k}$$

where

- k is the number of points.
- n is the number of objects detected.
- $E[S]$ is the expected cluster size.
- $f'(0)$ is the DERIVATIVE (slope) of distribution of distances.



Distance sampling - Point Transects - Theory

$$\widehat{D} = \frac{n\widehat{f'(0)}\widehat{E[s]}}{2\pi k}$$

Implications:

- $f(y)$ and $g(y)$ no longer have same shape
- $f'(0)$ MUCH harder to estimate than $f(0)$
- Only distances matter in likelihood, so be careful in stratification where the detection function is common across strata as you cannot distinguish between the fits.
- In both line and point transect sampling, behavior at 0 is CRITICAL for estimation

Distance sampling - Point Transects - Theory

$$\widehat{var}(\widehat{D_{Point}}) = \hat{D}^2 \times \left\{ \frac{\widehat{var}(n)}{n^2} + \frac{\widehat{var}(\widehat{f'(0)})}{[\widehat{f'(0)}]^2} + \frac{\widehat{var}(\widehat{E[s]})}{[\widehat{E[s]}]^2} \right\}$$

Implications:

- Not all components of variation under control of experimenter – stratify??
- Empirical estimates of variance components are robust.
- MLE theory only provides estimate of middle term.

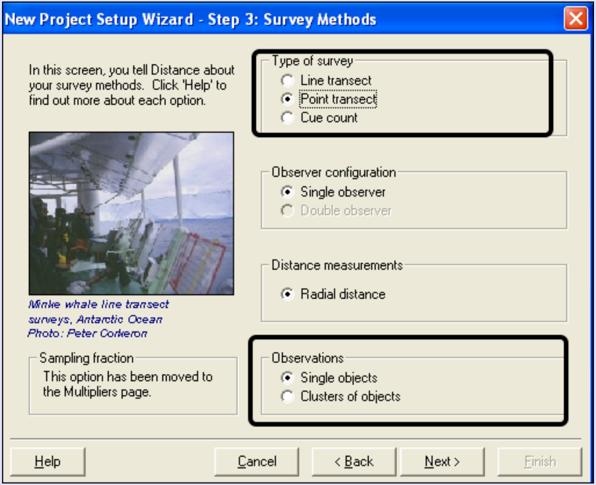
Model fitting similar to that of line transect sampling:

- Choose form of detection function (key function) + adjustment (series)
  - Does not depend on ANGLE of observation
- Truncate data to avoid outliers ($g(r) \approx 0.10$) as more observations appear in tails than in line transect sampling.
- Truncate at fixed radius value. [IMPORTANT.]
- MLE
- Model assessment etc.

Example of point transect analysis

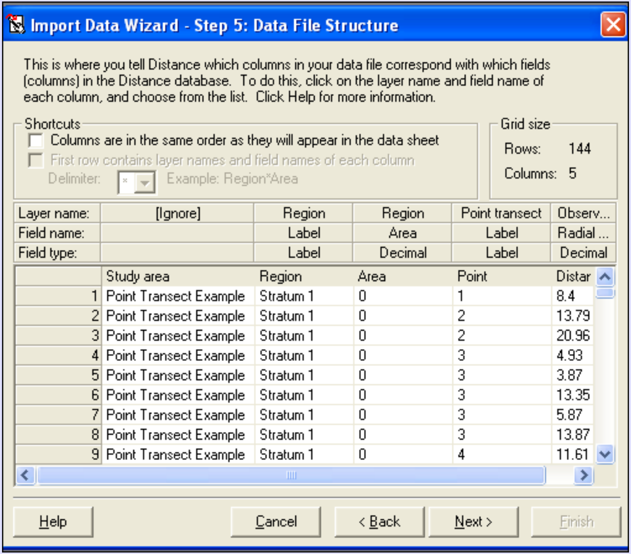
Distance sampling - Point Transects - Example

Open the Example-point directory in the Sample Data folder. Open the Example-point.xls file and import the data in the usual way. Start a project – be sure to specify a point transect.



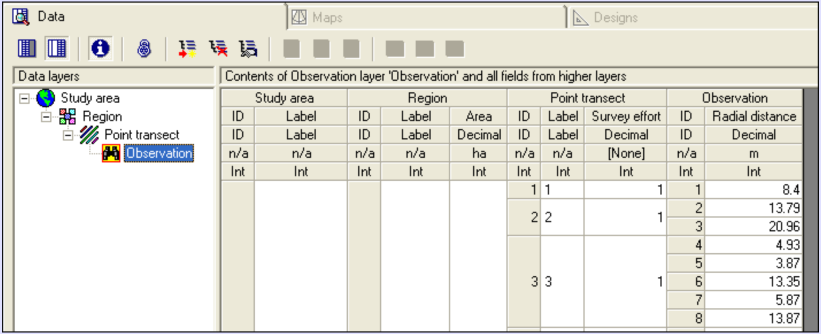
Distance sampling - Point Transects - Example

Import the data. [No clustering in this example.]



Distance sampling - Point Transects - Example

Data is ready for analysis.



Study area		Region			Point transect			Observation	
ID	Label	ID	Label	Area	ID	Label	Survey effort	ID	Radial distance
ID	Label	ID	Label	Decimal	ID	Label	Decimal	ID	Decimal
n/a	n/a	n/a	n/a	ha	n/a	n/a	[None]	n/a	m
Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
					1	1	1	1	8.4
					2	2	1	2	13.79
								3	20.96
					3	3	1	4	4.93
								5	3.87
								6	13.35
								7	5.87
								8	13.87

Do analysis using all of data; Half-normal/cosine detection function.

Don't forget to use meaningful analysis, data filter, and detection function names.

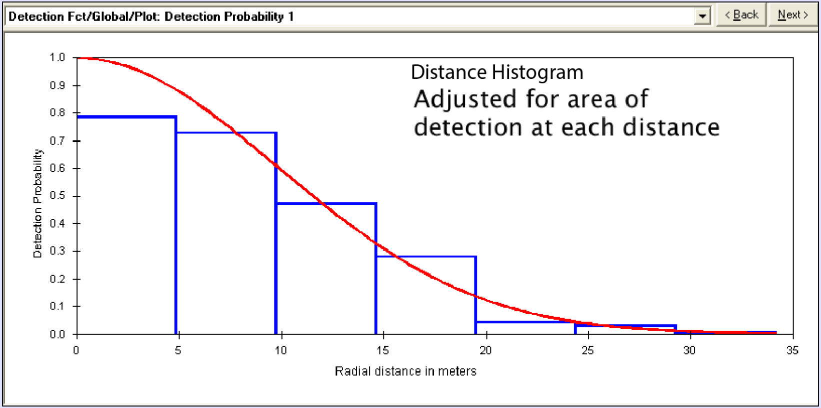
Run the analysis and look at output.

Data(all) HN/C

- Estimates of $A(1)$, $h(0) = f'(0)$, EDR, p
- QQ plot looks ok; KS/CM look ok.

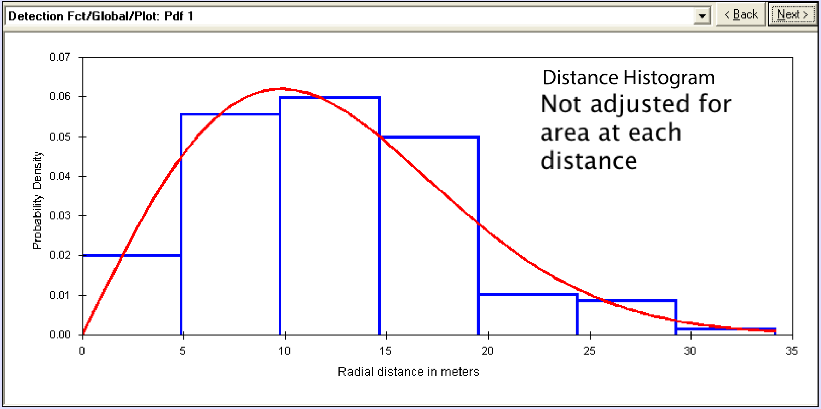
Distance sampling - Point Transects - Example

Data(all) HN/C: Two-different histograms of distance and detection function.



Red shows actual detection function; blue is ADJUSTED distance measurements.

Data(all) HN/C: Two-different histograms of distance and detection function.



Red is combination of detection function x area searched; blue is ACTUAL distance measurements.

Distance sampling - Point Transects - Example

Data(all) HN/C: Estimates of density.

Density Estimates/Global

```
Effort        :    30.00000
# samples     :    30
Width         :    34.16000
# observations:   144

Model  1
   Half-normal key, k(y) = Exp(-y**2/(2*A(1)**2))


              Point         Standard     Percent Coef.        95% Percent
 Parameter    Estimate        Error      of Variation     Confidence Interval
 ---------   -----------   -----------   --------------   ----------------------
   D          79.621        7.7053          9.68          65.806        96.338
 ---------   -----------   -----------   --------------   ----------------------

Measurement Units
---------------------------------
Density: Numbers/hectares
    EDR: meters

Component Percentages of Var(D)
-------------------------------
Detection probability   :  84.9
Encounter rate          :  15.1
```

Try a different detection function.

Try truncation (recall $g(w) \approx 0.1$ in point transect - where is this?

Do NOT compare models using AIC that have different data filters!

	ID				Name	# params	Delta AIC	AIC	ESW/EDR	D	D LCL	D UCL	D CV
	1	1	1	1	Data(all) HN/C	1	0.00	919.16	13.85	79.621	65.806	96.338	0.097
	2	1	2	1	Data(<25m) HN/...	1	0.00	864.73	13.85	77.469	62.360	96.239	0.110

Distance sampling - Point Transects - Assumptions

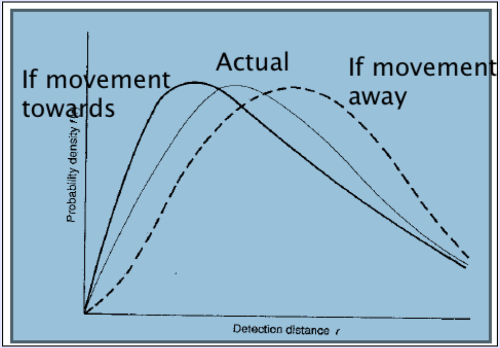
0. No assumptions made about distribution of animals on landscape!

- D describes the density over the landscape.
- Not necessary that animals distribute themselves independently, i.e. aggregated distributions are allowed.
- RANDOM placement of points in large study areas implies uniform distribution of distance from point.
- If D varies over short distances, then reliable estimates still possible – be caution of GOF tests.

Distance sampling - Point Transects - Assumptions

1. Animals at point are detected with probability $= 1$, i.e. $g(0)=1$
   - Be cautious of animals hiding at distance 0.

2. Animals detected at initial location - NO movement.
   - Movement of animal causes bias in point transect, unlike in line transects because detection when moving toward observer is higher than movement away from observer.
   - CAUTION if animals are attracted or frightened near the point this can lead to serious bias. Notice how slope at zero changes below.



Distance sampling - Point Transects - Assumptions

3. Measurements of distance from point are exact.
    - Even random errors introduce bias because "inflation" by detection probability is larger further from the transect line.
    - Visual estimates of distance are NOT very good, i.e. observers tend to underestimate distances.
    - Beware of heaping, i.e. lots of measurements at 100 m, but fewer at 99 m and 101 m.
    - Try using interval data (binned) (e.g. distances estimated by airplane struts).

4. Beware of outliers.
   - Objects far away have large influence on estimates because of weighting by 1/prob of detection.
   - Check histograms and truncate at about $g(w) = 0.1$
     - Truncation introduces no bias.

Distance sampling - Point Transects - Assumptions

5. Objects are detected independently of each other.
   - Clusters of animals violate this assumption, but can be dealt with if cluster size is recorded.
   - Pair-bonds of birds where some members of cluster may be hidden?
   - Ill-defined clusters?
   - Clusters that are too large to count?

Distance sampling - Point Transects - Assumptions

6. Detection function is not spiked at 0.
    - Methods work best if there is shoulder to the detection function.
    - Because SLOPE at 0 is to be estimated, this is difficult with so few animals detected close to 0.
    - Negative exponential (with no series adjustment) is a last resort!

Blue grosbeaks (*Guiraca caerulea*) in Georgia, USA.
Points (1-2) were randomly placed in 39 agricultural fields that were being included in a longleaf pine (*Pinus palustris*) restoration program.
`http://coopunit.forestry.uga.edu/quant_cons_book/companion/chapter9/index.html`

Fit a simple model with no stratification. Treat the 3 time periods as "independent".

Distance sampling - Point Transects - Exercise

HouseWren

- 10 study blocks of 16 ha along 30 km of South Platte River
- 14-16 points in each block
- 4 observers visited each point.

Post-stratify by observer and fit appropriate models to estimate D.

Distance sampling - Point Transects - Summary

Similar analysis to Line transect analysis

- Distribution of distances will not be monotonic but this is an artifact of convolution of detection function and area searched increasing with distance.
- Truncate distances at $g(w) = 0.1$.
- Truncate at fixed radius value.

Indirect Methods - Using Multipliers

Distance sampling - Indirect methods

Sometime animals are too hard to count, but animals signs can be counted

- dung and nests
- only part of population can be detected (e.g. females are hidden)
- animals not continuously available
- fast moving objects

Need a conversion factor (multiplier) to go from cues to animals.

Distance sampling - Indirect Methods - Example

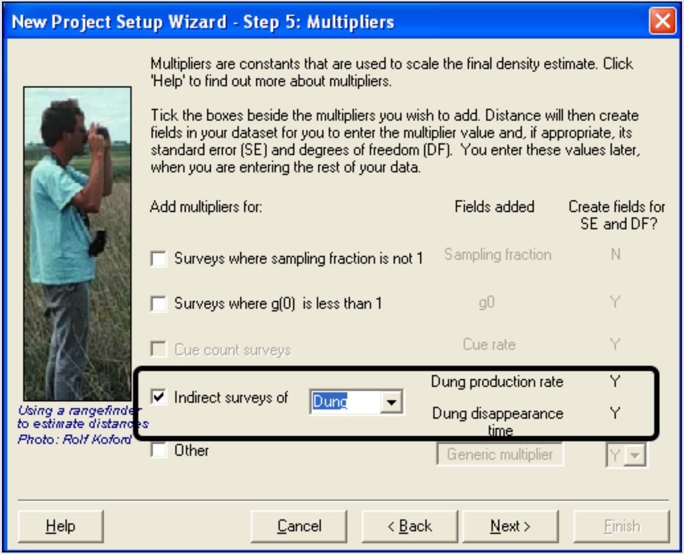
Elephant Dung example.

Import the data. Units are:

- Distances from the transect are in Centimetres,
- Transect length is in Metres
- Results for dung and elephant density per Square kilometre.

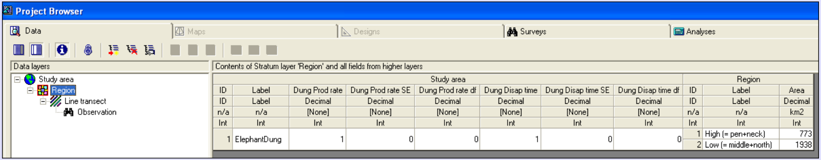
Leave room for multipliers (see next slide)

Distance sampling - Indirect Methods - Example



We will fill in multiplier later . . .

... We eventually get the data imported. Notice slots for multipliers.



Study area								Region		
ID	Label	Dung Prod rate	Dung Prod rate SE	Dung Prod rate df	Dung Disap time	Dung Disap time SE	Dung Disap time df	ID	Label	Area
ID	Label	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	ID	Label	Decimal
n/a	n/a	[None]	[None]	[None]	[None]	[None]	[None]	n/a	n/a	km2
Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
1	ElephantDung	1	0	0	1	0	0	1	High (= pen+neck)	773
								2	Low (= middle+north)	1938

Fit a model with

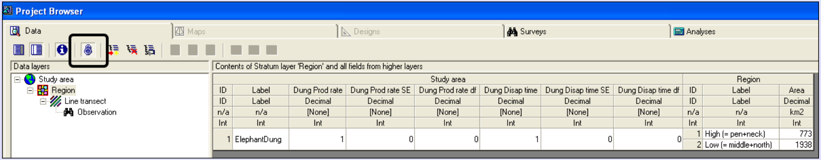
- Common detection function in both stratum (one stratum has too few observations)
- Different encounter rates in strata

Distance sampling - Indirect Methods - Example

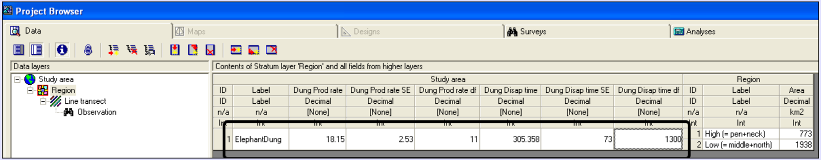
Put in the multipliers

- Dung production of 18.15 (SE 2.53, df 11) piles/elephant per day.
- Average time to disappearance is 305.356 (SE 7.3, df 1300) days.

First unfreeze the data:



Enter the multipliers



Study area								Region		
ID	Label	Dung Prod rate	Dung Prod rate SE	Dung Prod rate df	Dung Disap time	Dung Disap time SE	Dung Disap time df	ID	Label	Area
ID	Label	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	ID	Label	Decimal
n/a	n/a	[None]	[None]	[None]	[None]	[None]	[None]	n/a	n/a	km2
Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
1	ElephantDung	18.15	2.53	11	305.358	73	1300	1	High (= pen+neck)	773
								2	Low (= middle+north)	1938

. . . and rerun the analysis . . . .

Distance sampling - Indirect Methods - Example

All se's etc are updated automatically, etc and dung converted to elephants

		Estimate	%CV	df	95% Confidence Interval	
Stratum: 1. High (= pen+neck)						
Half-normal/Cosine						
	D	0.53053	31.26	200.30	0.29053	0.96876
	N	410.00	31.26	200.30	225.00	749.00
Stratum: 2. Low (= middle+north)						
Half-normal/Cosine						
	D	0.26965E-01	48.59	36.61	0.10606E-01	0.68559E-01
	N	52.000	48.59	36.61	21.000	133.00

How to derive the conversion factors? Not part of this course, but send me an email.

Distance sampling - Indirect Methods - Summary

Use DISTANCE to get estimate of animal signs and then convert to animals.

DISTANCE will automatically propagate uncertainty into final estimate.

Estimating Changes in Density

Given $\widehat{D}_1$ and $\widehat{D}_2$, estimate (with SE) $D_1 - D_2$.

Not possible to do directly in DISTANCE because likelihood only models the detection function.

Two cases

- $\widehat{D}_1$ and $\widehat{D}_2$ are from different surveys, e.g. Year 1 vs. Year 2, analyzed independently.
- $\widehat{D}_1$ and $\widehat{D}_2$ have a common detection function estimated from pooled data. E.g. test if density is same in two strata in same survey.

Distance sampling - Estimating Changes in Density

Independent sample case (easy):

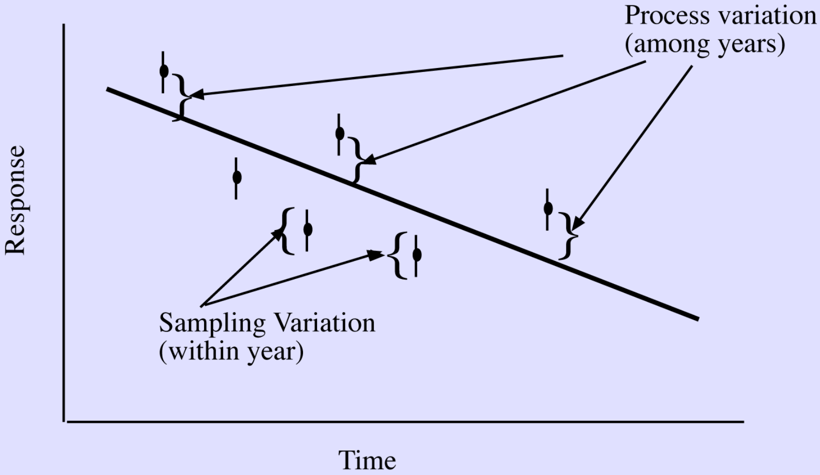
$$\widehat{D_1 - D_2} = \widehat{D}_1 - \widehat{D}_2$$

$$se(\widehat{D_1 - D_2}) = \sqrt{se^2(\widehat{D}_1) + se^2(\widehat{D}_2)}$$

To test for a change, look to see if 95% c.i. includes the value 0, or create Z test.

Some adjustment can be made for small sample sizes and using a t distribution. See Distance sampling book.

Independent sample case (easy):
Use $\widehat{D}_i$ directly in trend analysis; se not needed; beware of process error.



Distance sampling - Estimating Changes in Density

Non-independent sample case (hard):

$$\widehat{D}_1 = \frac{n_1 \widehat{f(0)} \widehat{E_1[S]}}{2L_1} \qquad \widehat{D}_2 = \frac{n_2 \widehat{f(0)} \widehat{E_2[S]}}{2L_2}$$

Then

$$\widehat{D}_1 - \widehat{D}_2 = \left\{ \frac{n_1 \widehat{E_1[S]}}{2L_1} - \frac{n_2 \widehat{E_2[S]}}{2L_2} \right\} \times \widehat{f(0)}$$

Use Delta method to find variance of this – see Distance book, Section 3.6.5.
No easy way to look at trend because ALL estimates are correlated because same detection function and/or transects/points used over time.

Distance Sampling Analysis Super Summary

Distance Sampling Analysis - Super Summary

Preliminary exploration of data:

- Don't plunge into analysis without exploration of data.
- Look for
  - Outliers (large distances have low probability of detection and high weights.
  - Heaping (distances are 100 m but never 99 m or 101 m.
  - Shape of distance distribution
    - Line transect distance distribution should have shoulder, decline as distance increases. Hole at zero is problematic.
    - Point transect distance distribution will have uni-modal (artifact of convolution of detection and area searched). Hole at zero is expected.
- Consider converting "exact distances" to intervals if problems exist such as heaping.

Distance Sampling Analysis - Super Summary

Import data into DISTANCE

- Include stratification and covariates even if not using.
- Save as *.csv or other text-delimited file.
- Separate folder for each project.
- Keep distance files together.
- Do regular backups.

Distance Sampling Analysis - Super Summary

Define *Model Definitions*

- Stratification (pre- or post-) gives lots of options.
- Detection function key not crucial as modified by series but avoid negative exponential because no shoulder.
- Deal with cluster size via stratification, regression, or covariate methods,
- Multipliers for known effects (e.g. only see females and sex ratio is 1:1).

Distance Sampling Analysis - Super Summary

$$\widehat{D} = \frac{n\widehat{f(0)}\widehat{E[S]}}{2L} \quad \widehat{D} = \frac{n\widehat{f'(0)}\widehat{E[s]}}{2\pi k}$$

- Right truncate at $g_{Line}(w) \approx 0.15 \quad g_{Point}(w) \approx 0.10$
- Truncate at fixed width. [IMPORTANT.]
- Detection function
  - Only contribution to the likelihood which has implications later
  - Shoulder at 0; no spike at 0; left truncation if needed.
  - Monotonic decline; group distances if measurement error.
  - Decline to 0 as distance increases.
  - Many key function + series give similar fit but avoid neg. exp.
- Deal with cluster size via modeling, stratification, or covariate.
- Multipliers for known effects (e.g. seeing fraction of population, dung counts, etc.)
- Stratification/covariates can improve precision but don't go overboard.
- AIC for model selection but cannot change data filter among models being compared.
  - No simple way to model average, but see DISTANCE manual.

Distance Sampling Analysis - Super Summary

$$\widehat{var}(\widehat{D_{Line}}) = \hat{D}^2 \times \left\{ \frac{\widehat{var}(n)}{n^2} + \frac{\widehat{var}(\widehat{f(0)})}{[\widehat{f(0)}]^2} + \frac{\widehat{var}(\widehat{E[s]})}{[\widehat{E[s]}]^2} \right\}$$

$$\widehat{var}(\widehat{D_{Point}}) = \hat{D}^2 \times \left\{ \frac{\widehat{var}(n)}{n^2} + \frac{\widehat{var}(\widehat{f'(0)})}{[\widehat{f'(0)}]^2} + \frac{\widehat{var}(\widehat{E[s]})}{[\widehat{E[s]}]^2} \right\}$$

- Some components of variation you have little control over and are the same regardless of effort.
- Variance components are estimated by:
  - Empirical variance methods over replicate lines/points are robust for $var(n)$, but with small sample sizes, may be forced to specify Poisson-process.
  - MLE theory gives $\widehat{var}(\widehat{f(0)})$ and $\widehat{var}(\widehat{f'(0)})$
  - Empirical variance methods for $\widehat{var}(\widehat{E[s]})$

Line Transect Distance Sampling Analysis Summary

Review the output:

- Check model fit using QQ plots; KS/CM tests; χ^2 tests, but don't rely exclusively on these
    - Horizontal lines in QQ plots are artifacts of heaping and rounding.
    - Small expected counts in χ^2 test can inflate test-statistic and lead to false-poor fit. Look at pooling small cell counts.
- See if estimates are reasonable.
- Rank models of SAME data using AIC. Model averaging must be done outside of DISTANCE but refer to DISTANCE manual.

  Cannot use AIC for model comparison with different data filters.

DONT PANIC