Introduction to Distance Sampling

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.

<u>Using regression</u>

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?

<u>Using truncation</u>

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.

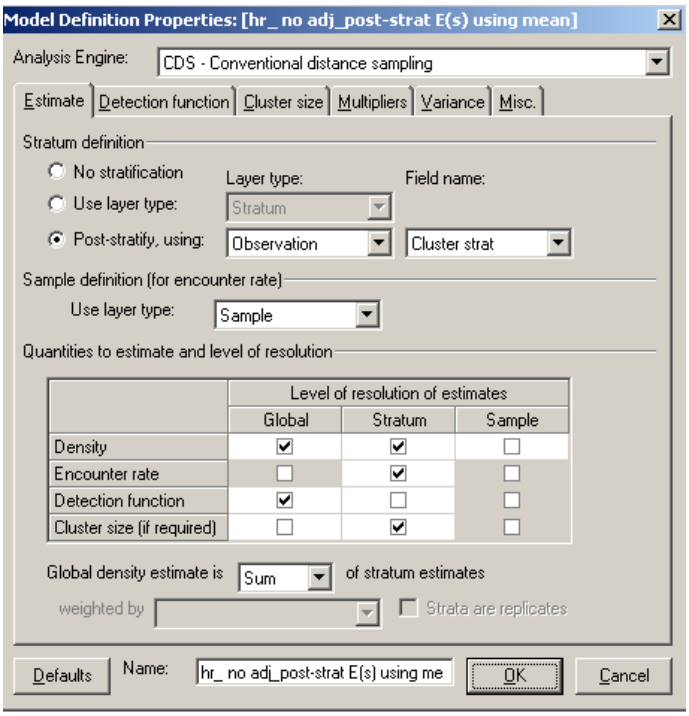
a) Why is the "Mean cluster size" on the **Cluster size/Global/Estimates** page different from the mean cluster size in analysis 1 above?

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.)

<u>Using post-stratification by cluster size</u>

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.

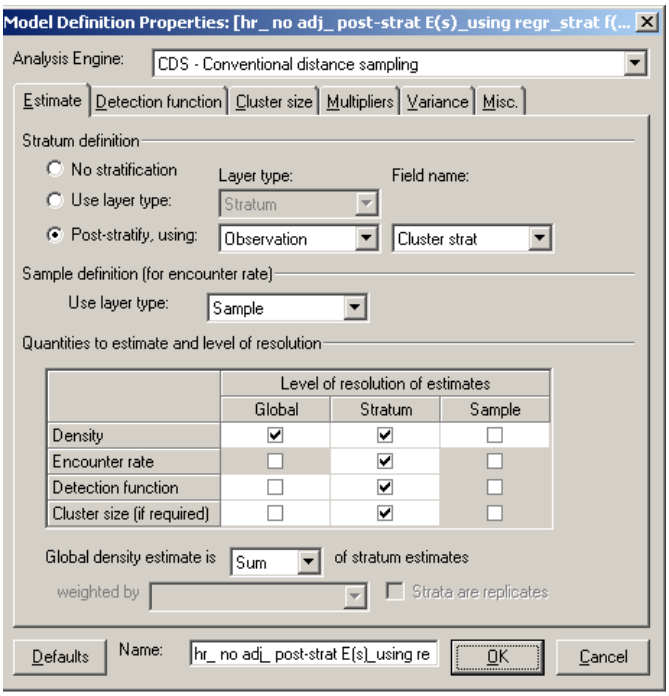
a) Use the "Cluster strat" column as a basis for performing an analysis post-stratified by cluster size. Do this by creating a new analysis with a new Model Definition that uses post-stratification at the Observation level. Fit a detection function pooled across strata, but estimate mean cluster size 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?



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