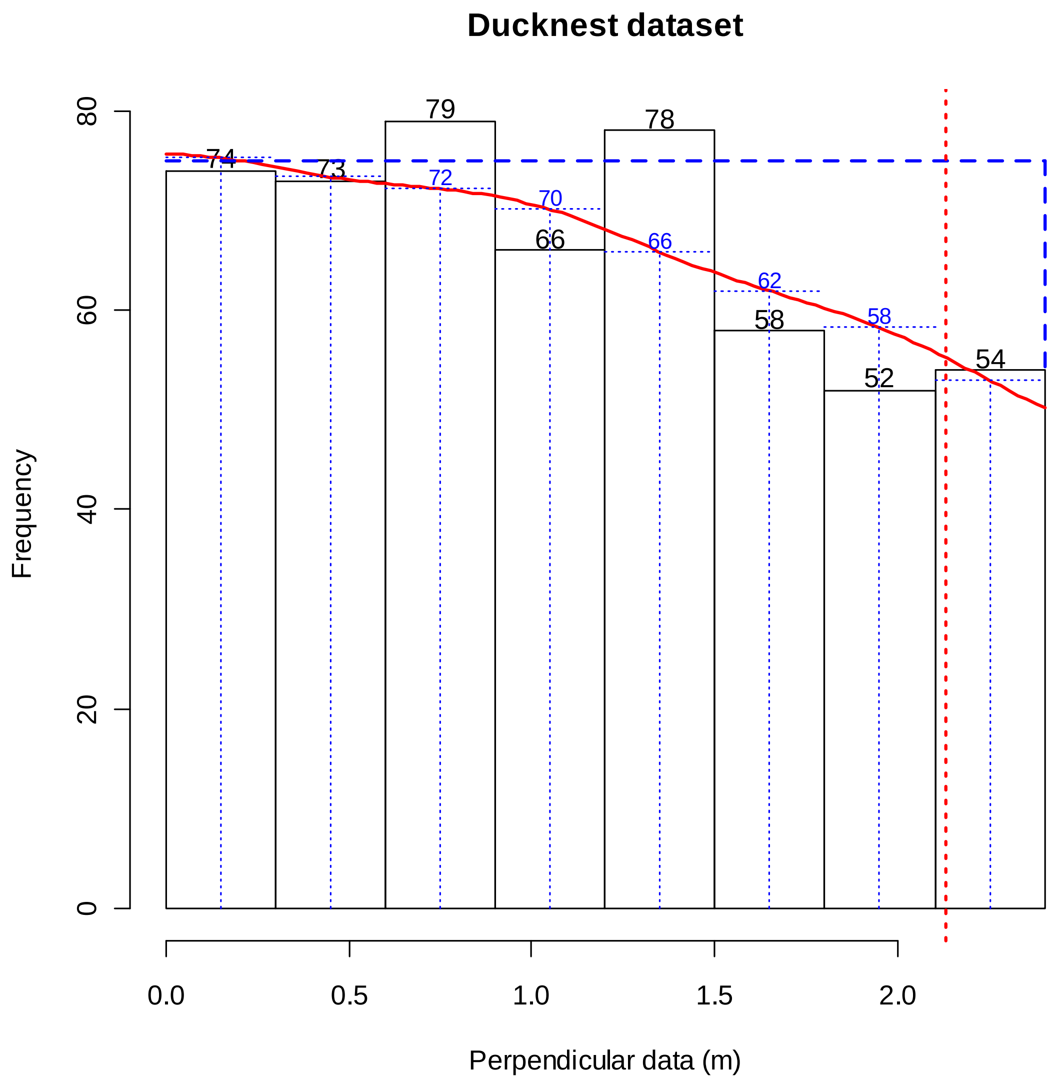
Introduction to Distance Sampling

Exercise 1: Line transect Solutions



1) P_a = area under curve / 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) ×0.3 = 530×0.3 = 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 × width = 75×2.4 = 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 $2wL$ = 2×(2.4/1000) ×2575 = 12.36km^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 $2\mu L = 2\times(2.13/1000)\times 2575 = 10.97$ km^2, and estimated density is 534/10.97 = 48.7 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 $534\times(0.472\times1000)/2\times2575 = 48.0$ 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 $\chi 2$ 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 / km^2 and 48.7 nests / km^2 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 km^2. 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

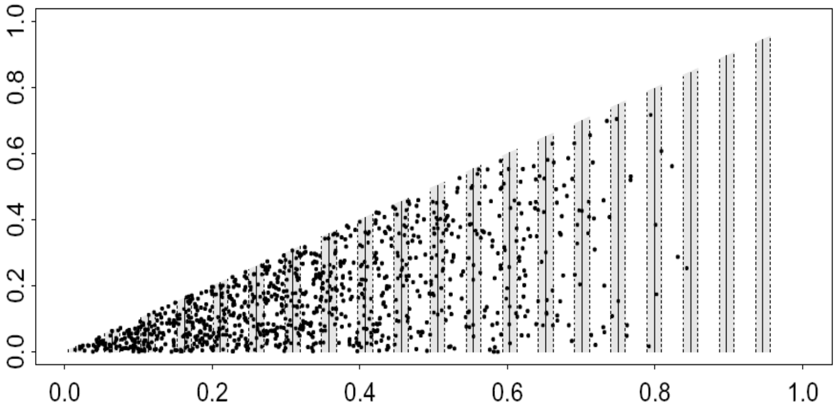
2) These capercaillie data are reasonably well-behaved and different models that fit the data well should give similar results. Note the rounding in the distance data. This means that interval cutpoints for histograms and goodness-of-fit testing, and for the analysis of grouped data if this is required, 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        3727.5
                                                         1164.0        3427.2
Half-normal/Cosine
                N        1064.6     27.40  999    20.74  608.00        1864.0
                                                         582.00        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
```

5. 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.

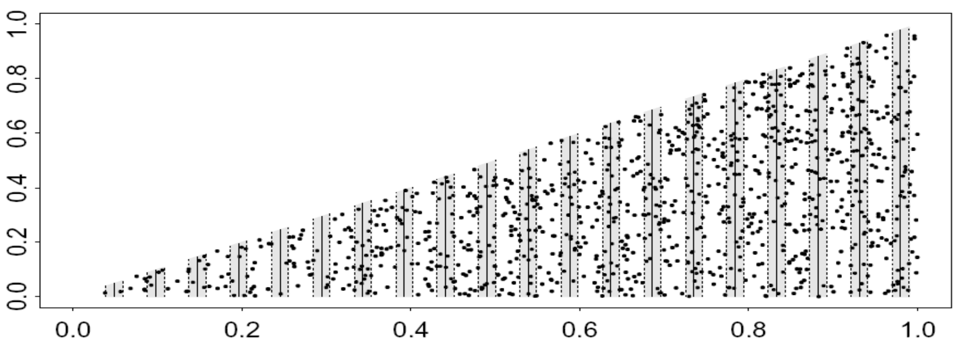
	Estimate	%CV	df	95% Confidence Interval	
Half-normal/Cosine					
D	2044.6	7.97	75.87	1745.0	2395.6
N	1022.0	7.97	75.87	872.00	1198.0

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.



<u>Without post-stratification: analytic output</u>

	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	2380.8
					1565.0	2350.3
Half-normal/Cosine						
N	973.69	10.03	999	50.60	796.00	1190.0
					782.00	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	0.58	0.52	1.29
	(0.74,1.43)	(0.36,0.94)	(0.26,1.06)	(0.80,2.11)
2	0.90	0.22	0.60	1.02
	(0.62,1.29)	(0.13,0.39)	(0.38,0.94)	(0.80,1.32)
3	0.71	0.26	0.82	1.21
	(0.45,1.23)	(0.09,0.76)	(0.52,1.31)	(0.82,1.79)
4	0.64	0.26	0.69	1.07
	(0.46,0.90)	(0.16,0.42)	(0.47,1.00)	(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)
			243 (18.9)	421 (12.2)	
N	760	638 (9.9)	664 (22.5)		666 (8.8)

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.

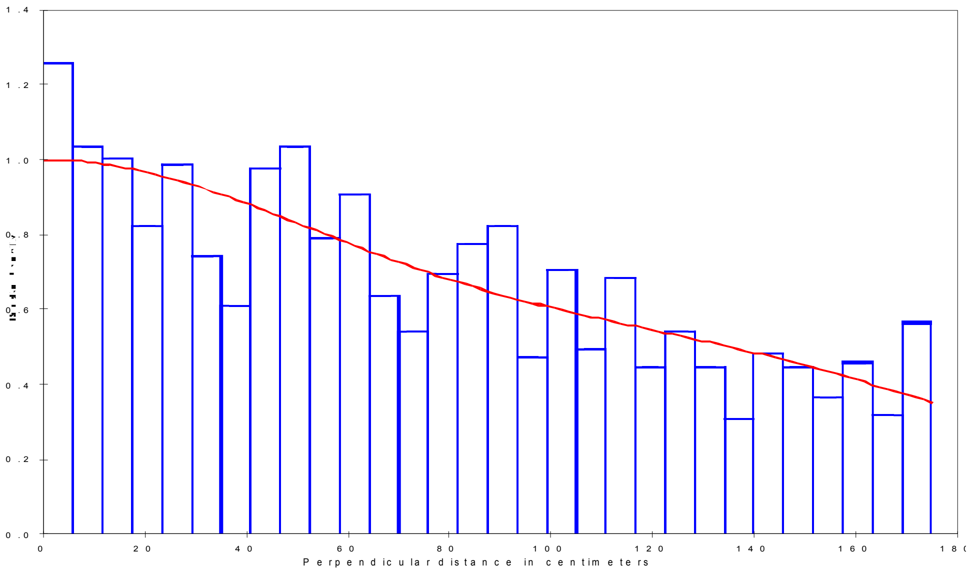
Introduction to Distance Sampling

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

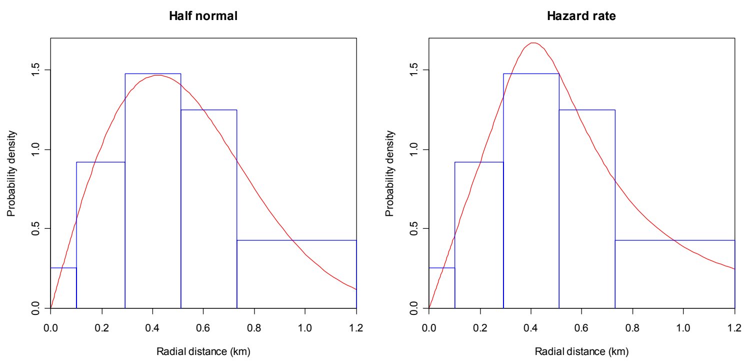
Question 1: $\hat{\eta} = 25$ cues per time unit (per hour in this case). Its standard error is 5, therefore its CV is 5/25=0.2 (or 20%).

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.