Measures of Precision

Overview

- How to quantify uncertainty
- Why variance is important
- Components of variation in distance sampling
- Controlling variance
- Estimating variance
  - Analytic
  - Bootstrap
- Confidence Intervals

How do estimates behave?

Consider an artificial population

D = 500 per unit^2 (no density gradient)

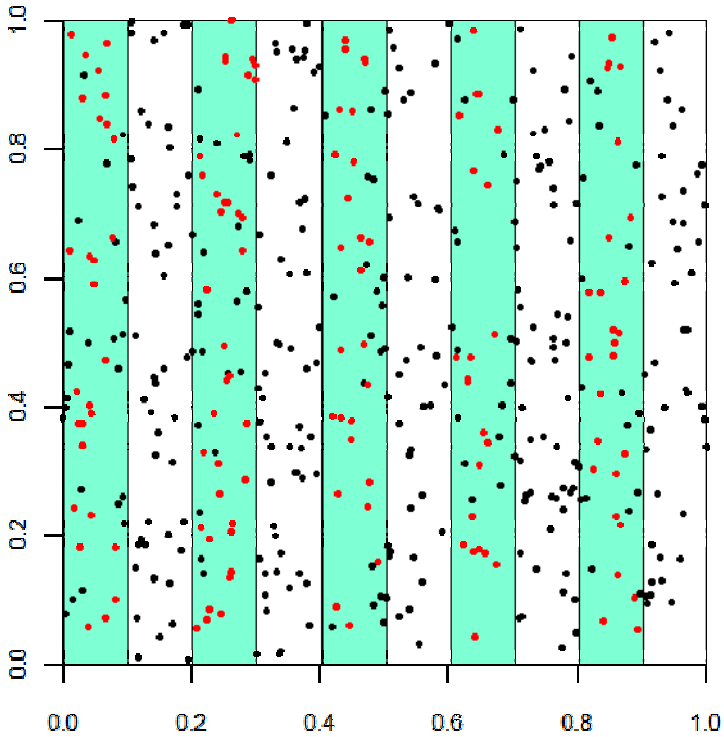
Design: 5 transects equally-spaced (w=0.05)

Results:

n = 140

$\hat{f}(0)$ = 34.6

$\hat{D}$ = 484.4



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

How do estimates behave?

Consider a duplicate survey

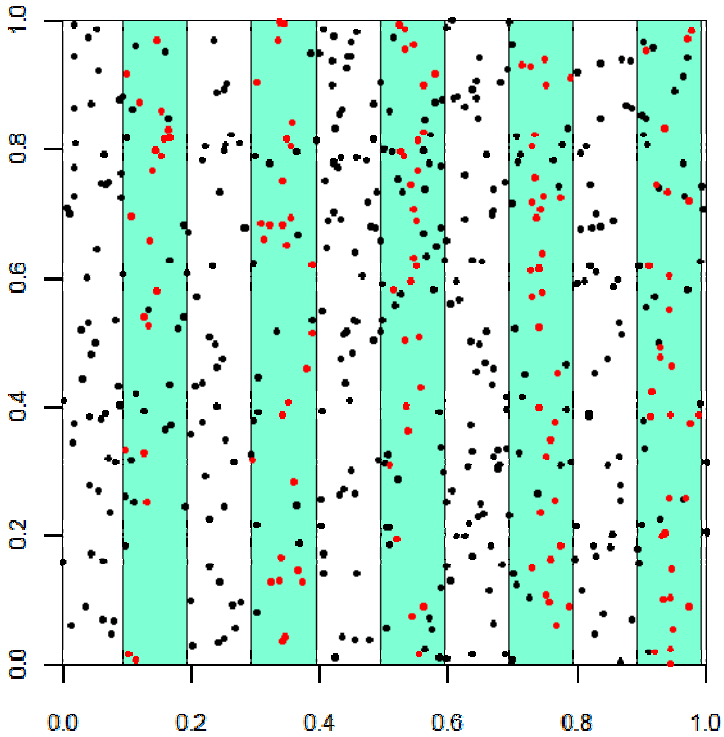
- Same population model
- Same survey design (with a new random start point)

Results:

n = 139

$\hat{f}(0)$ = 37.6

$\hat{D}$ = 522.1



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

How do estimates behave?

Imagine repeating this process over and over, using the same survey design and a population drawn from the same density model

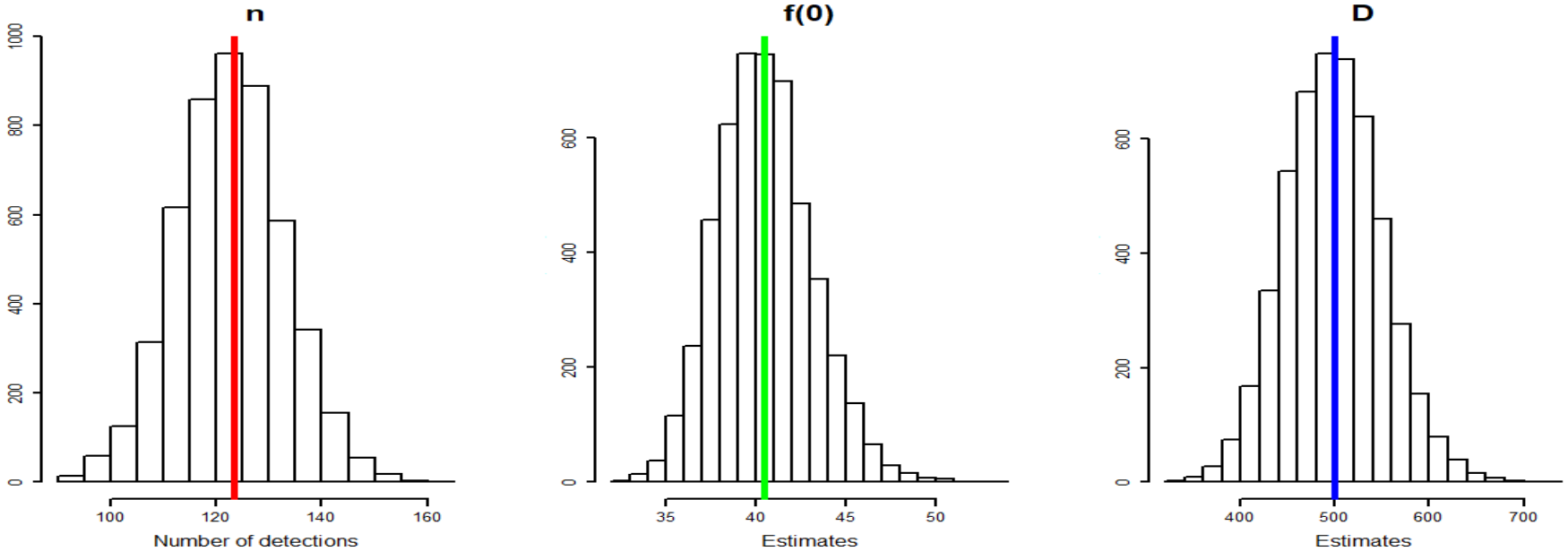
Each survey will yield:

- A different value for n
- A different value for $\hat{f}(0)$
- A different value for $\hat{D}$

How do estimates behave?

What happens if we repeat this simulated survey 10,000 times?

We end up with distributions for n, $\hat{f}(0)$ and $\hat{D}$



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

How do estimates behave?

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

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

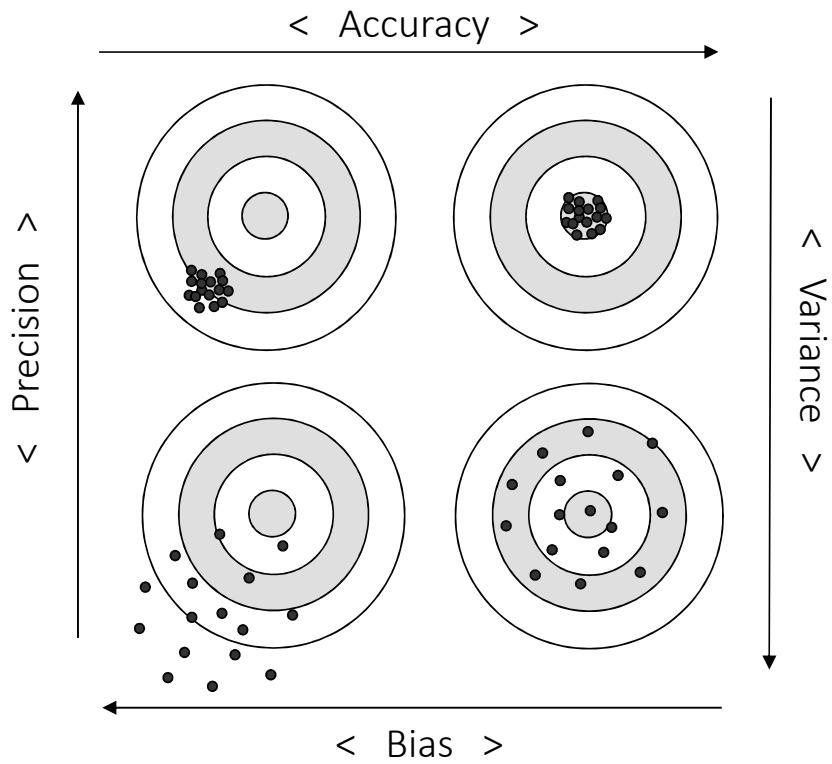
How variable are the estimates?

E.g. what is the variance of the distribution for $\hat{D}$?

What is the average value of the estimates?

E.g. is the distribution for $\hat{D}$ centred on the truth?

Bias vs. Variance



Low precision = high variance = high uncertainty

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Quantifying uncertainty

Different ways of measuring uncertainty:

1. Variance = the average squared difference from the mean (the inverse of precision)

   If the estimator for *D* is unbiased, then

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

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

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

Quantifying uncertainty

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

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

Useful for comparing variances when the scale and/or the units of measurement differ

E.g. consider two variables: X has mean = 100 and variance = 400,
Y has mean = 1 and variance = 0.04

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

Quantifying uncertainty

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

   Calculations are based on variance

   Different ways to calculate CIs, depending on the data, e.g.

   Normal

   Lognormal (available in Distance)

   Bootstrap (available in Distance)

   More about CIs later…

Why is variance important?

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

Variance by components

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

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

Constant (no variance)

Encounter rate

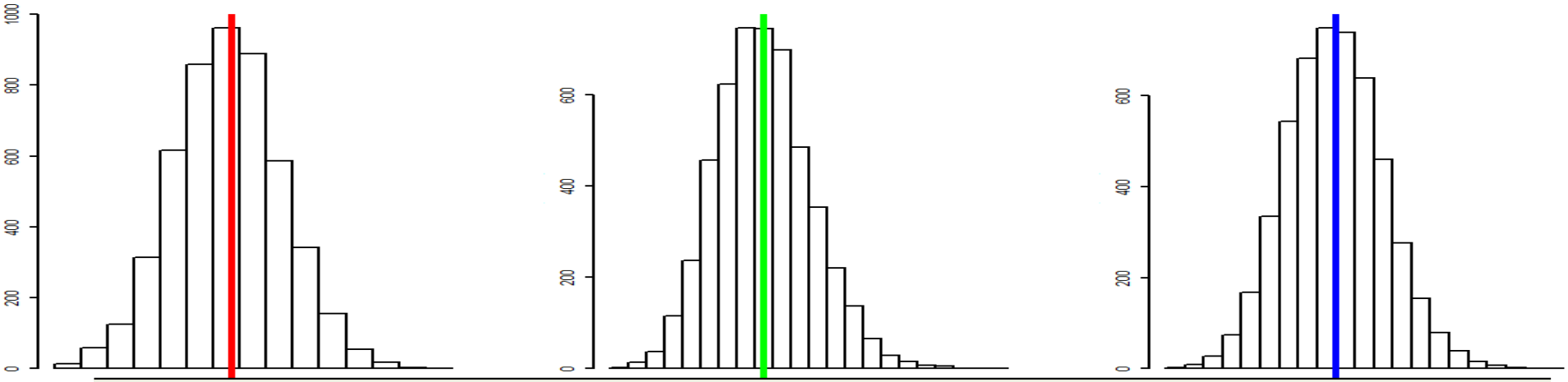
Detection function

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Variance by components

We can calculate variance measures separately for each component



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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Variance by components

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

Variance by components

Distance provides several variance measures for each component

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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Encounter rate variance

The **encounter rate** = n/L = the number of detections per unit of distance

The variance of n/L is related to the variance of n, and therefore to the variances of counts for individual transects

Therefore, if counts from individual transects are highly variable the variance of n/L will also be high $Var[n] = Var[n_1] + \ldots + Var[n_k]$ ← **assumes independence**

Controlling variance

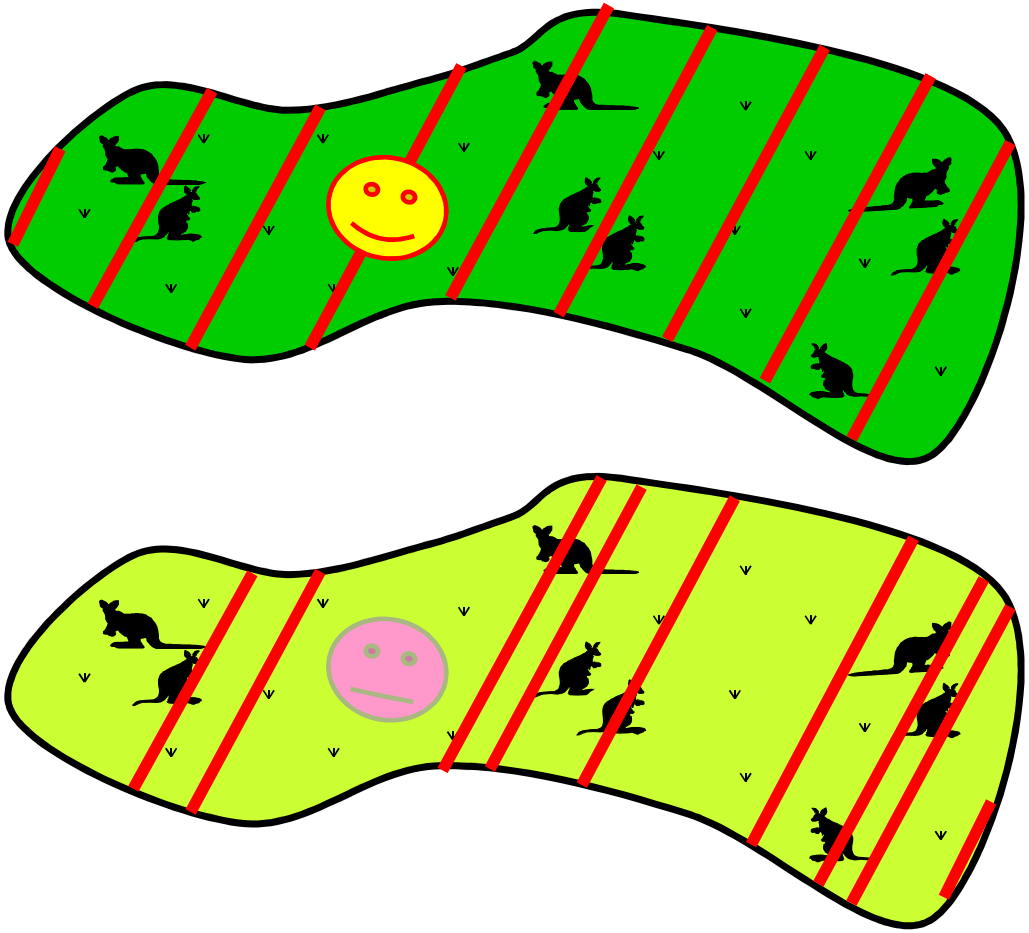
- We can use this knowledge of encounter rate variance to help design good surveys
- Three main ways we can reduce encounter rate variance:
  - Use systematic survey designs
  - Run transects parallel to density gradients
  - Use designs with several transects

Controlling variance

1. **Use systematic survey designs**

   These give lower variance than completely random designs

   More likely to give even coverage of the survey region

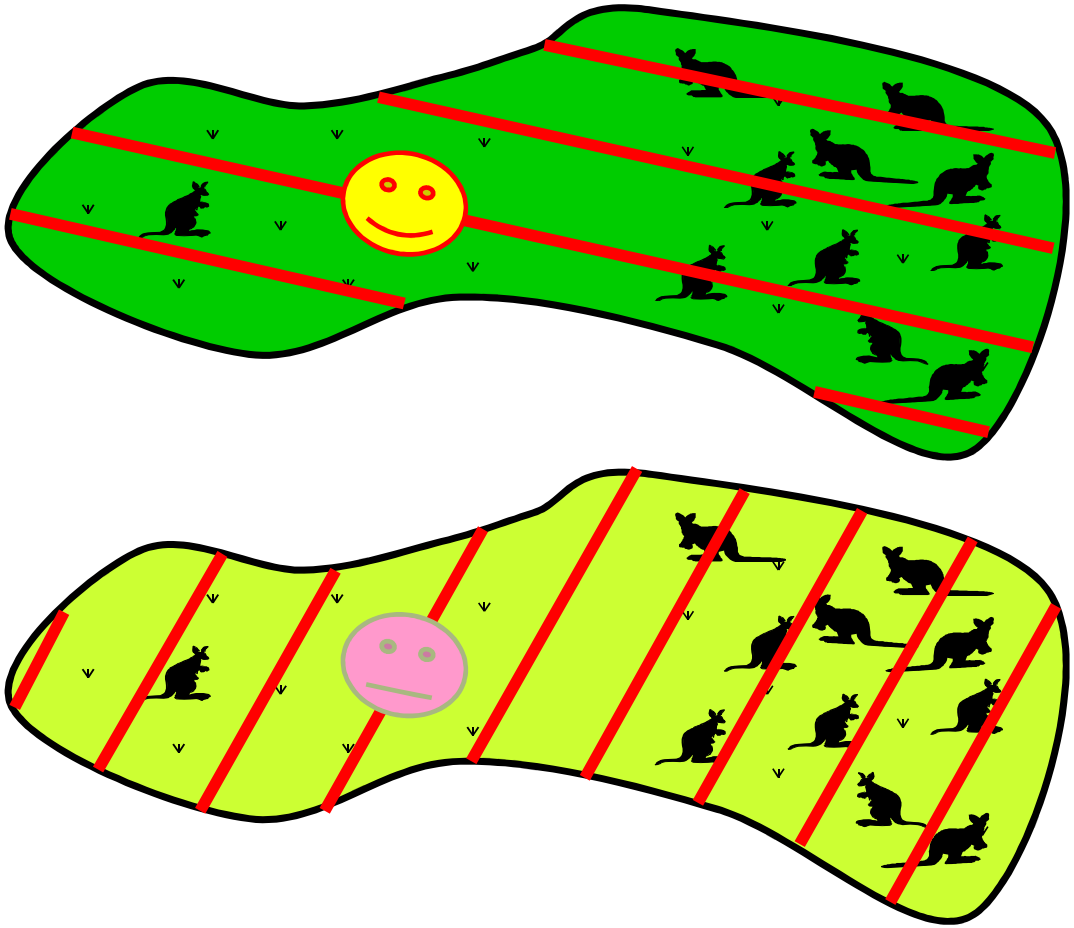


Controlling variance

2. **Run transects parallel to known density gradients**

   i.e. perpendicular to gradient contours

   Lines are more likely to have similar encounter rates since each line will cover the full range in density



CREEM
Centre for Research into Ecological and Environmental Modelling

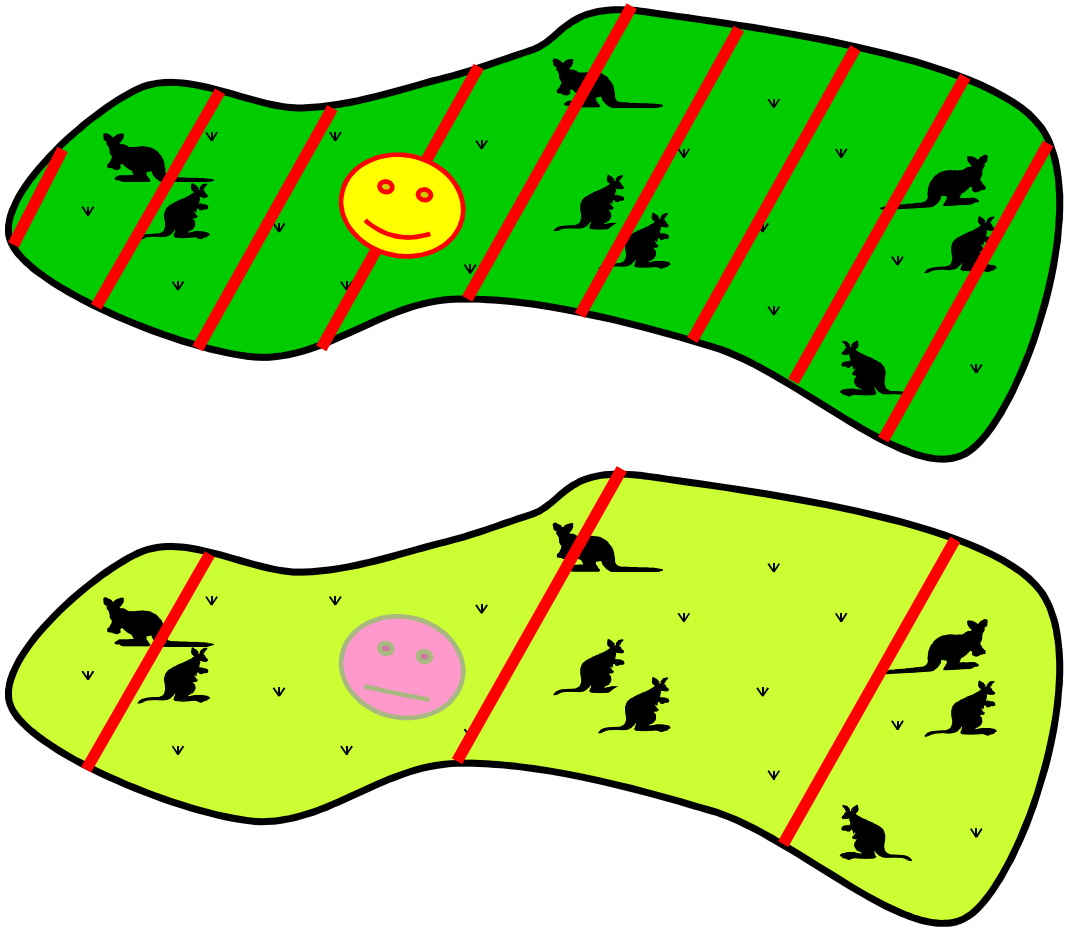
University of St Andrews

Controlling variance

3. Use many lines (or points)

This will give good spatial coverage

Using longer lines will also help to reduce variation between lines (because each transect will have a larger sample size)



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Recap

- Density estimates will vary between (hypothetical) duplicate surveys
- Density estimators with low bias and low variance are more reliable, since individual estimates are more likely to be close to the truth
- High variance in the encounter rate and the fitted detection function will lead to high variance in $\hat{D}$
- Good survey design can help to lower the variance of $\hat{D}$ by reducing encounter rate variance

Estimating variance

- So how do we measure the variance?
- In reality we rarely know the true variance of our estimator, and we can't carry out 1000s of real surveys…
- What we need to do is estimate the variance using the data from a single survey
- Two methods:
  - Analytic
  - Bootstrap

Estimating variance – Analytic

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

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

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

Estimating variance – Analytic

We can check this approximation works using the results of our simulation,

$$\left\{cv(\hat{D})\right\}^2 = 0.1125^2 = 0.01266$$

$$\left\{\mathrm{cv}\left(\frac{\mathrm{n}}{\mathrm{L}}\right)\right\}^2 + \left\{\mathrm{cv}(\hat{\mathrm{f}}(0))\right\}^2 = 0.0869^2 + 0.0704^2 = 0.01251$$

We can rearrange the squared CV to get an estimate of the variance

$$\mathrm{var}(\hat{D}) \approx \left\{\hat{D} \times \sqrt{\left\{cv(\hat{D})\right\}^2}\right\}^2$$

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Estimating variance – Analytic

- To estimate the variance of $\hat{D}$ we need to estimate the squared CVs of the components
- We therefore need estimates for the variances of n/L and $\hat{f}(0)$
- The variance of $\hat{f}(0)$ can be estimated using standard methods (because its parameters are obtained via maximum likelihood)
- But n/L is part of the data, not an estimate, and estimating its variance is more difficult...

Estimating variance – Analytic

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

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

k = number of lines

l_i = effort for line i

n_i = count for line i

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

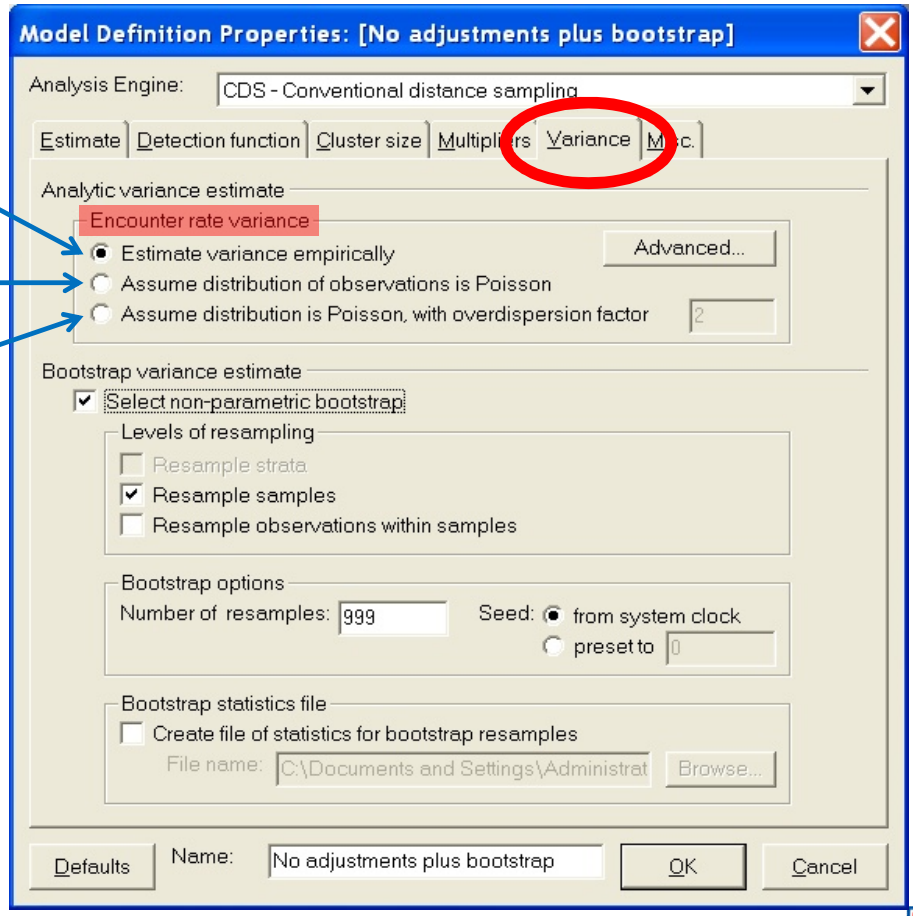
Estimating variance – Analytic

- Note that the formula used in Distance assumes independently arranged lines
- It therefore tends to overestimate encounter rate variance for systematic designs (for which there are no analytic variance estimators)
- Stratification-based methods can help to remove this bias (see later lecture)
- Highlights the difference between the true (and unknown) variance and our ability to estimate it

Estimating variance – Analytic

Three options in Distance:

1. Use the formula we have just seen
2. Assume *var(n)=n* (last resort)
3. Assume *var(n)=ϑn*
   (slightly better last resort)



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Estimating variance – Analytic

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

Measurement Units

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

Component Percentages of Var(D)

Detection probability : 11.5
Encounter rate : 80.7
Cluster size : 7.9

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

Similarly, N and D always have the same CV

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Estimating variance – Analytic

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

E.g.

$$100\% \times \frac{\left\{cv(\hat{f}(0))\right\}^2}{\left\{cv(\hat{D})\right\}^2} =$$

The percentage relative contribution made by f(0)

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

Estimating variance – Analytic

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

Measurement Units

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

Component Percentages of Var(D)

Detection probability : 11.5
Encounter rate : 80.7
Cluster size : 7.9

$12.17^2 / 35.92^2 \times 100\%$

$32.25^2 / 35.92^2 \times 100\%$

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Estimating variance – Bootstrap

The bootstrap method works as follows:

1. Generate a new sample by repeatedly sampling from the original sample randomly and with replacement
   - *some units from the original sample may appear more than once in the new sample (others might not appear at all)*
   - *each new sample must be the same size as the original sample*
2. Calculate a density estimate using the new sample
3. Repeat steps 1 & 2 a large number of times (e.g. 999) to obtain multiple density estimates

Estimating variance – Bootstrap

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

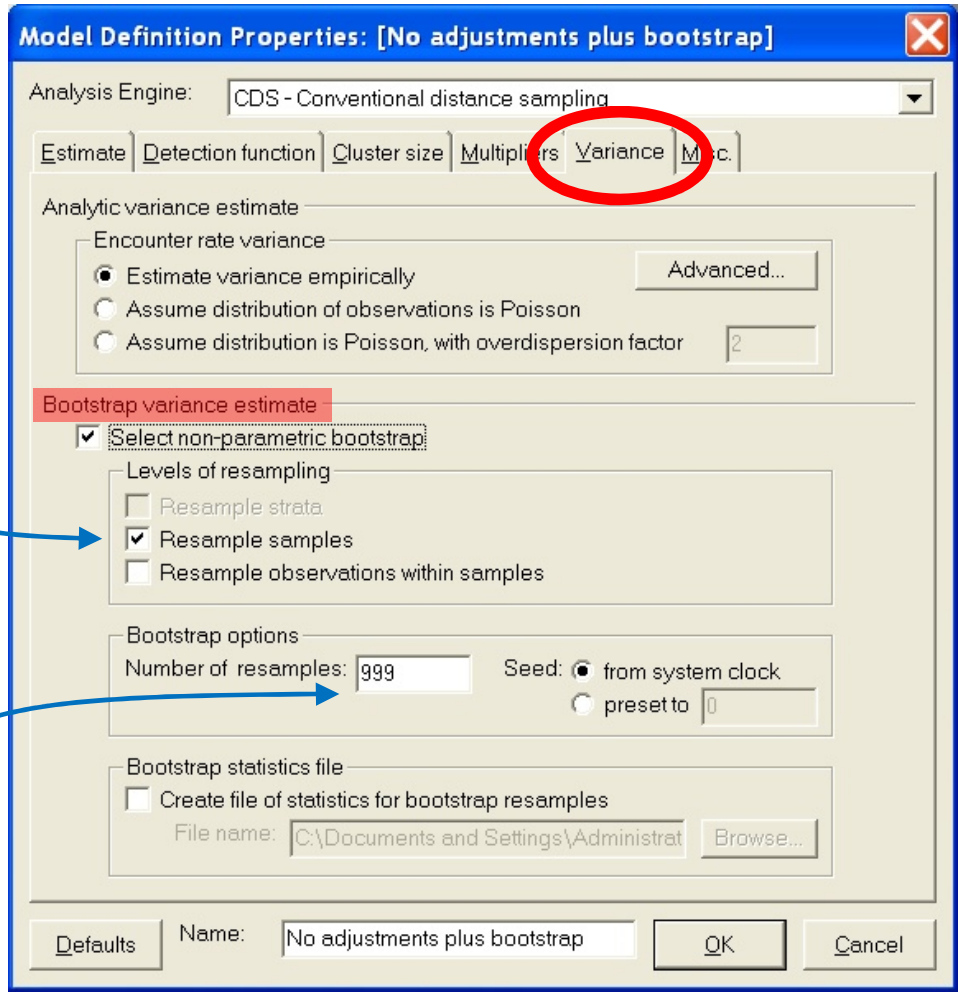
Estimating variance – Bootstrap

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

Estimating variance – Bootstrap

The usual option (samples = transects)

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



CREEM
Centre for Research into Ecological and Environmental Modelling

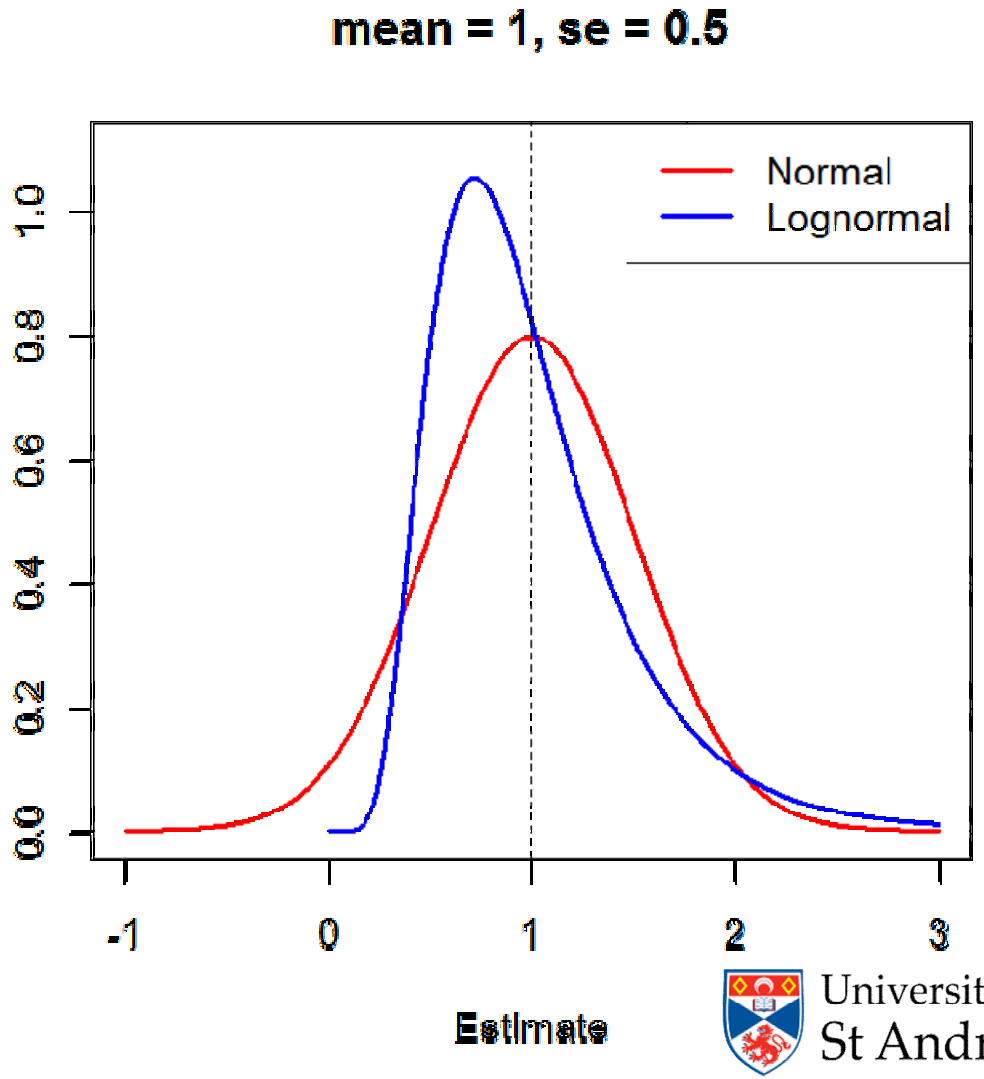
University of St Andrews

Confidence Intervals

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

Confidence Intervals - Analytic

- Two choices:
  - Normal
    - *symmetrical*
    - *easy to use*
    - *allows negative values*
  - Lognormal
    - *asymmetric (skewed)*
    - *trickier to use*
    - *typically higher interval limits*
    - *does not allows negative values*



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Confidence Intervals - Analytic

Distance uses 95% lognormal CIs

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

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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Confidence Intervals – Bootstrap

We can use the bootstrap estimates to construct CIs for the true density in two ways:

Parametric

Use the lognormal CI method with the bootstrap estimate of variance instead of the analytic estimate

Non-parametric

Place the bootstrap estimates in order of increasing size and use percentiles as the CI limits (e.g. for a 95% CI using 999 bootstrap estimates, take the 25th estimate as the lower limit and the 975th estimate as the upper limit)

Confidence Intervals – Bootstrap

Both options are provided in Distance

Pooled Estimates:

	Estimate	%CV	#	df	95% Confidence Interval	
DS	0.20906E-01	38.14	999	19.43	0.96777E-02	0.45162E-01
					0.12166E-01	0.41557E-01
D	0.39125E-01	42.24	999	23.93	0.16953E-01	0.90294E-01
					0.22020E-01	0.81797E-01
N	27987.	42.24	999	23.93	12127.	64589.
					15752.	58511.

Option 1 (parametric)

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

Option 2 (non-parametric)

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Other advantages of the bootstrap

Ambivalent model fit

E.g. different detection functions can be fitted to each bootstrap resample if it is difficult to choose between competing models (model uncertainty is therefore incorporated into the bootstrap density estimates)

Checking independence of components

If the bootstrap and analytical CIs are very different, then the ‘squared CVs add’ rule may have been violated (i.e. due to non-independence of the components of the density estimator – a necessary assumption for this approximation to work well)

Further reading

- Section 3.6 of Buckland et al. (2001) Introduction to Distance Sampling
- Sections 6.3.1.2 (lines) and 6.3.2.2 (points) of Buckland et al. (2015) Distance Sampling: Methods and Applications
- Fewster et al. (2009) Estimating the encounter rate variance in distance sampling. Biometrics 65: 225-236.