Measures of Precision

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

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)

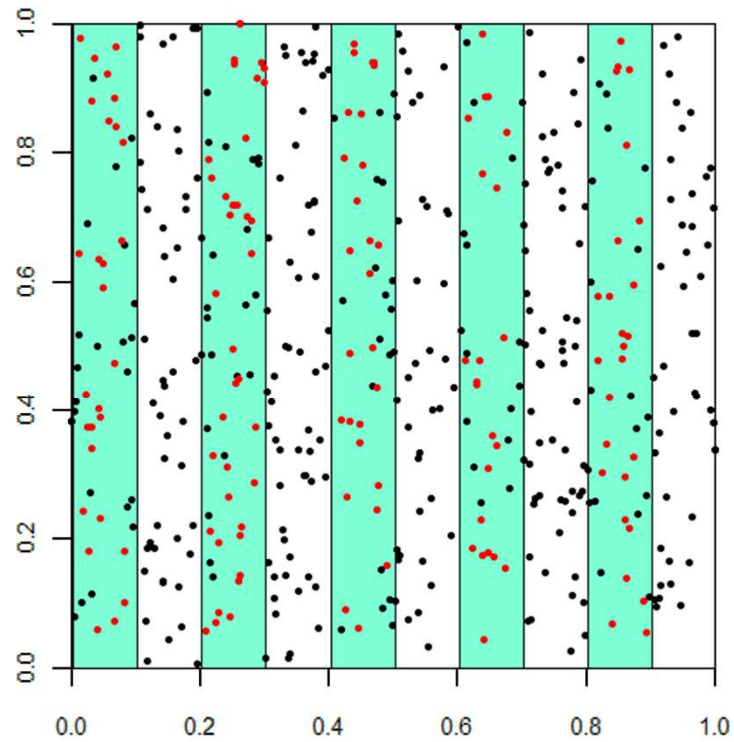
Design: 5 transects equally-spaced (w=0.05)

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

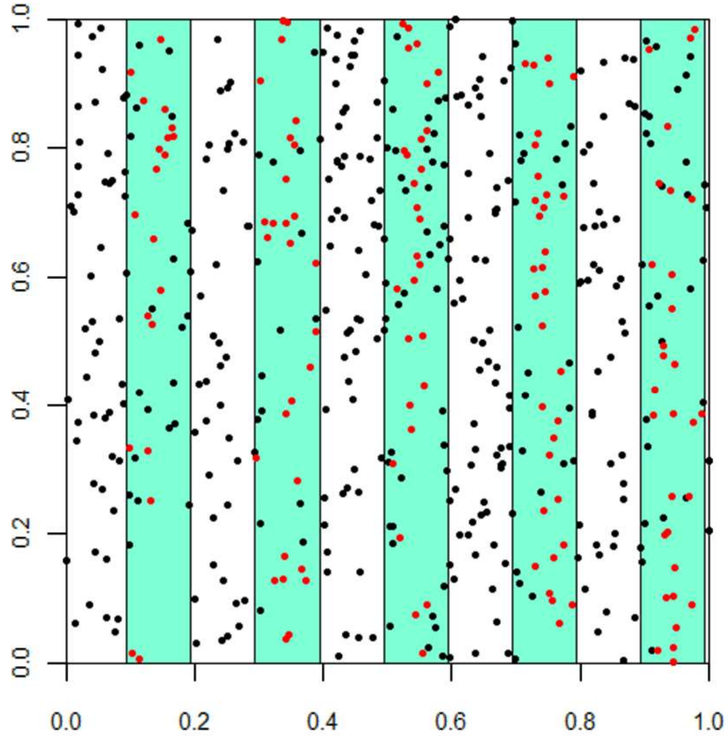
- Same population model
- Same survey design (with a new random start point)

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

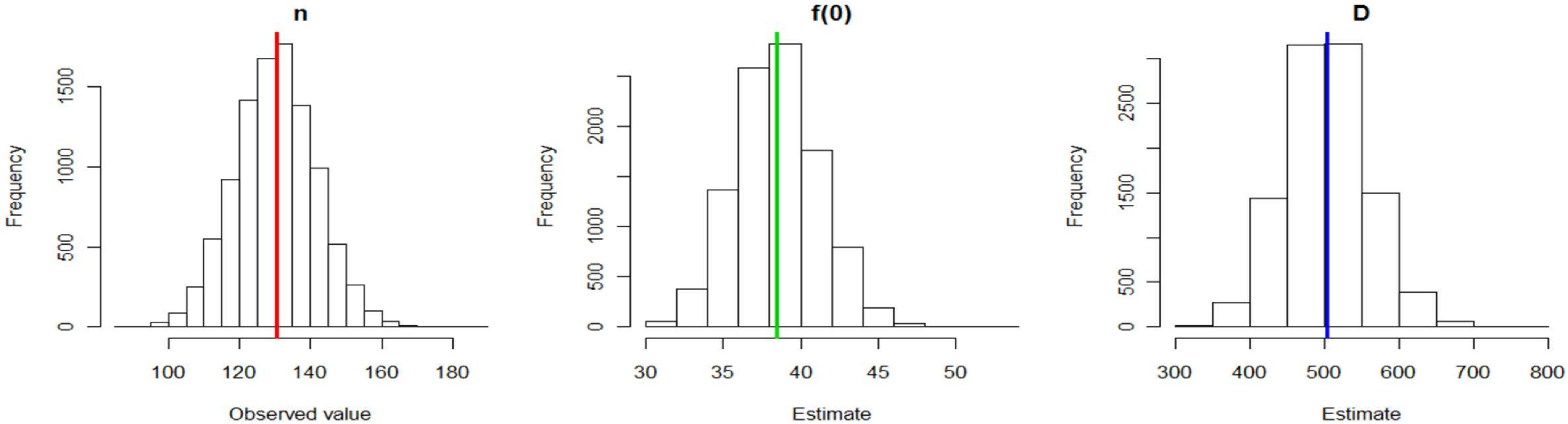
Each survey will yield:

- 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 $\widehat{D}$



Note, $\hat{f}(0) = {}^{1}/_{w\hat{P}_a}$

How do estimates behave?

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

$$\widehat{D} = \frac{n}{2wL\,\hat{}_{a}}$$

How variable are the estimates?

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

What is the average value of the estimates?

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

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)

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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Quantifying uncertainty

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

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

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

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

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

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!)

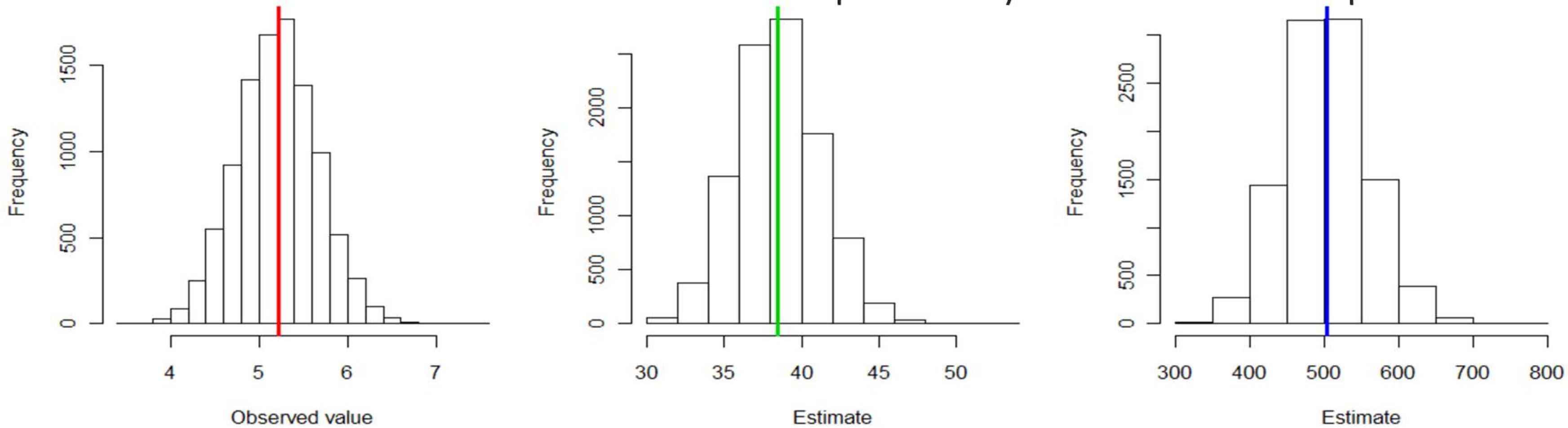
Variance by components

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

$$\hat{D} = \frac{n}{2wL\hat{P}_a} = \underbrace{\frac{1}{2w}}_{\text{Constant (no variance)}} \times \underbrace{\frac{n}{L}}_{\text{Encounter rate}} \times \underbrace{\frac{1}{\hat{P}_a}}_{\text{Detection function}}$$

Variance by components

We can calculate variance measures separately for each component



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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Variance by components

- The variance of $\widehat{D}$ is affected by the variance of its components
- If the variance of n is high,
- *then the variance of* $\frac{n}{L}$ *will be high and*
- *the variance of* $\widehat{\mathrm{D}}$ *will be high*
- If the variance of $\widehat{P}_a$ is high
- *then the variance of* $\widehat{\mathrm{D}}$ *will be high*
- For reliable estimates,
- we want $Var\left[\frac{n}{L}\right]$ and $Var[\widehat{P}_a]$ to be low

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Variance by components

Distance provides several variance measures for each component

```
                           Estimate          SE         CV
Average p                 0.3491863  0.02160949 0.06188528
N in covered region 300.6991117 30.11200030 0.10013997

Summary statistics:
   Region Area CoveredArea Effort   n  k     ER      se.ER      cv.ER
1 Default    1      3436.8     48 105 12 2.1875 0.3169604 0.1448962

Abundance:
  Label   Estimate        se        cv       lcl       ucl       df
1 Total   8.749392  1.378541 0.1575585  6.270328  12.20859 15.32522

Density:
  Label   Estimate        se        cv       lcl       ucl       df
1 Total 0.08749392 0.01378541 0.1575585 0.06270328 0.1220859 15.32522
```

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 multiple transects

Estimating variance – Analytic

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

$$\left\{cv\left(\widehat{D}\right)\right\}^2 = \left\{cv\left(\frac{n}{L}\right)\right\}^2 + \left\{cv\left(\widehat{P}_a\right)\right\}^2$$

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

Estimating variance – Analytic

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

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

$$\{cv(\hat{D})\}^2 = 0.1126^2 = 0.01266$$

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

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

$$var(\hat{D}) \approx \hat{D}^2 \times \{cv(\hat{D})\}^2$$

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Estimating variance – Analytic

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

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

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

```
                          Estimate          SE         CV
Average p                0.3491863  0.02160949 0.06188528
N in covered region 300.6991117 30.11200030 0.10013997

Summary statistics:
   Region Area CoveredArea Effort   n  k     ER     se.ER     cv.ER
1 Default    1      3436.8     48 105 12 2.1875 0.3169604 0.1448962

Abundance:
  Label    Estimate        se        cv        lcl       ucl       df
1 Total    8.749392   1.378541 0.1575585   6.270328  12.20859 15.32522

Density:
  Label    Estimate        se        cv        lcl       ucl       df
1 Total 0.08749392 0.01378541 0.1575585 0.06270328 0.1220859 15.32522
```

```
Component percentages of variance:
 .Label Detection     ER
  Total     15.43  84.57
```

Abundance and Density always have the same CV

Estimating variance – Analytic

```
                           Estimate          SE          CV
Average p                 0.3491863  0.02160949 0.06188528
N in covered region 300.6991117 30.11200030 0.10013997

Summary statistics:
   Region Area CoveredArea Effort   n  k     ER     se.ER     cv.ER
1 Default    1      3436.8     48 105 12 2.1875 0.3169604 0.1448962

Abundance:
  Label   Estimate        se        cv       lcl       ucl       df
1 Total   8.749392  1.378541 0.1575585  6.270328  12.20859 15.32522

Density:
  Label   Estimate        se        cv       lcl       ucl       df
1 Total 0.08749392 0.01378541 0.1575585 0.06270328 0.1220859 15.32522
```

```
Component percentages of variance:
 .Label Detection    ER
  Total     15.43 84.57
```

$$= \frac{0.1449^2}{0.1576^2} \times 100$$

$$= \frac{0.0619^2}{0.1576^2} \times 100$$

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. $$100\% \times \frac{\{cv(\hat{P}_a)\}^2}{\{cv(\hat{D})\}^2} =$$ The percentage relative contribution made by $\hat{P}_a$

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

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

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 $var(\hat{D})$

Estimating variance – Bootstrap

Basic function to generate a bootstrap:

```
bootdht(model, flatfile, nboot, summary_fun)
```

`model` – detection function model

`flatfile` – data object used to fit model

`summary_fun` – function to harvest required statistic from each bootstrap sample

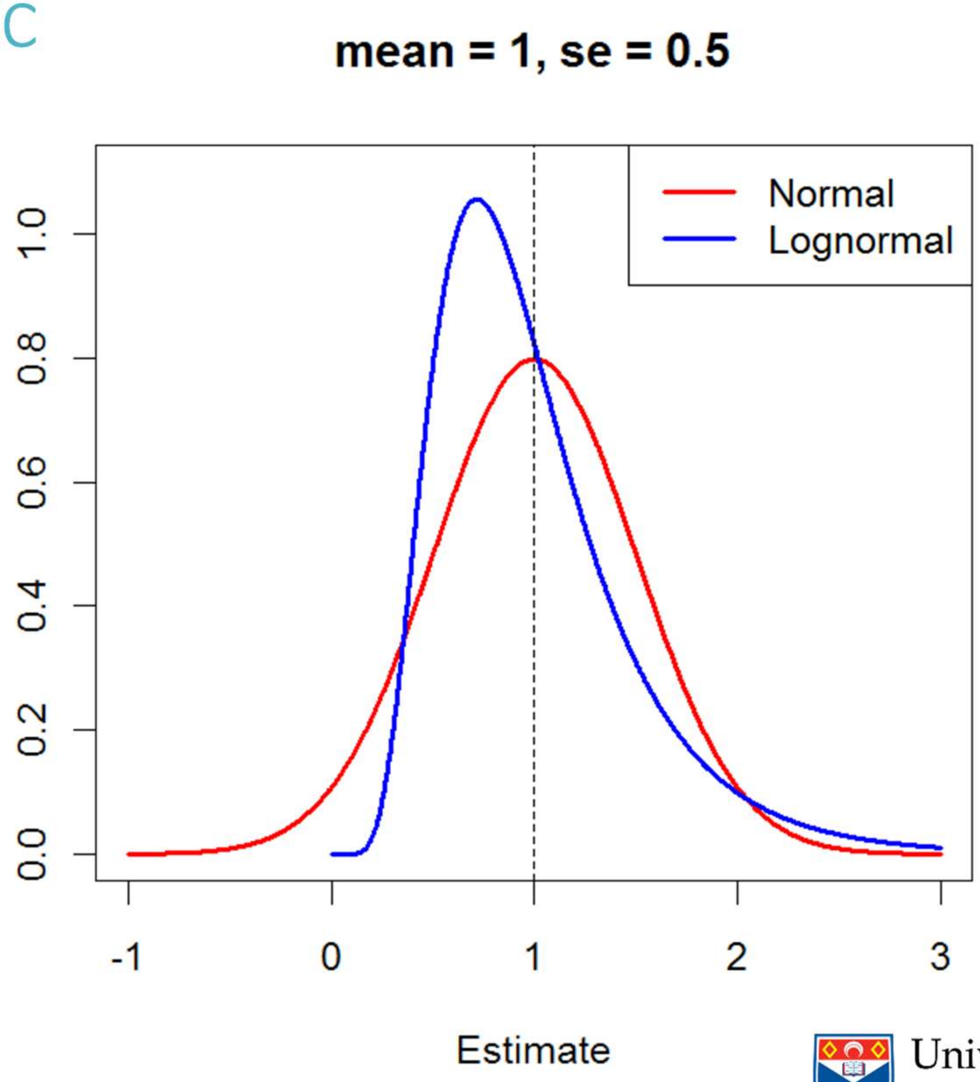
`nboot` – the number of bootstrap samples to use

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, it would be beneficial 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



Confidence Intervals - Analytic

Distance uses 95% lognormal CIs

```
Abundance:
  Label Estimate        se        cv      lcl      ucl       df
1 Total 8.749392    1.378541 0.1575585 6.270328   12.20859 15.32522

Density:
  Label    Estimate        se        cv       lcl       ucl       df
1 Total 0.08749392 0.01378541 0.1575585 0.06270328 0.1220859 15.32522
```

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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Confidence Intervals – Bootstrap

The nonparametric option is provided in Distance

```
Bootstrap results

Boostraps                : 999
Successes                : 999
Failures                 : 0

  Estimate   se   ucl  lcl   cv
N      8.58 1.44 11.67 5.94 0.17
D      0.09 0.01  0.12 0.06 0.17
```

Standard error divided by the mean

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Further reading about precision

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

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

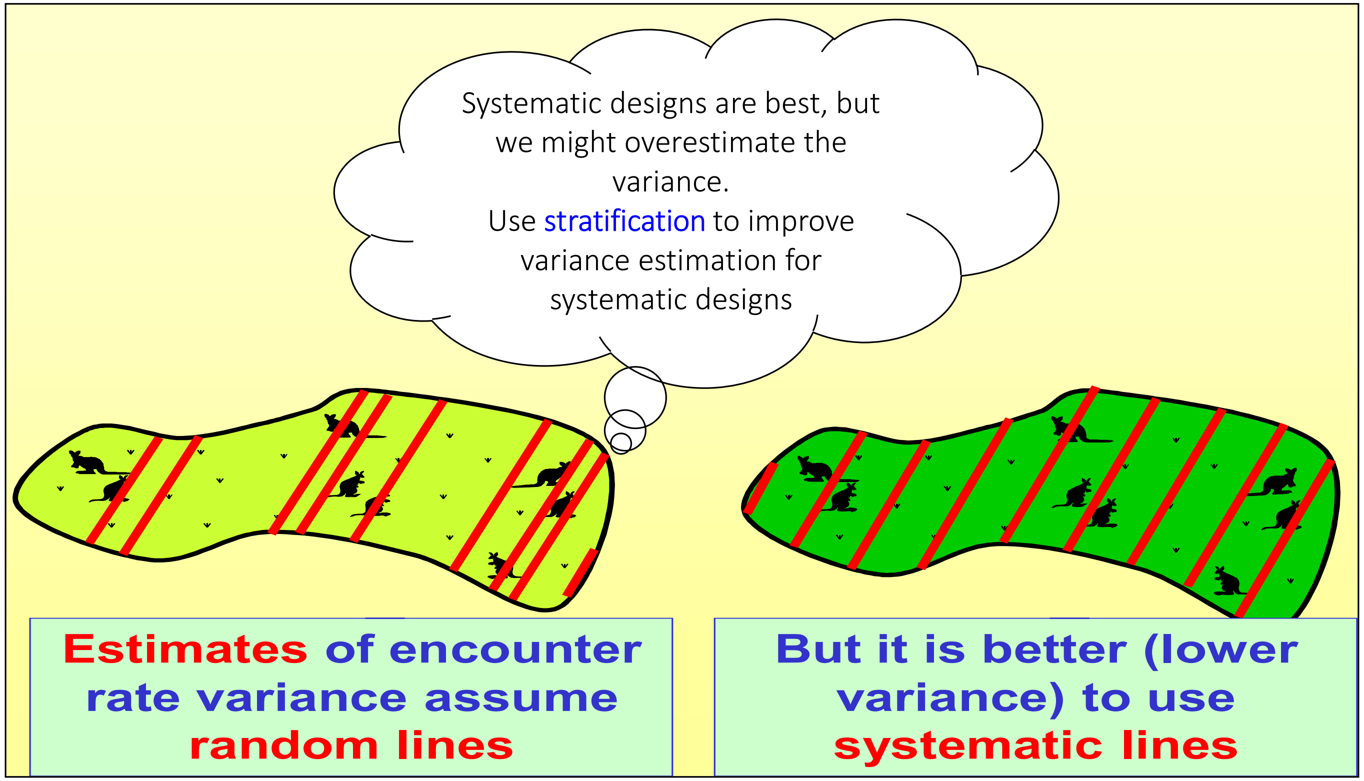
Producing a better estimate of variance when systematic samplers are used

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

Systematic samples

Problem:

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

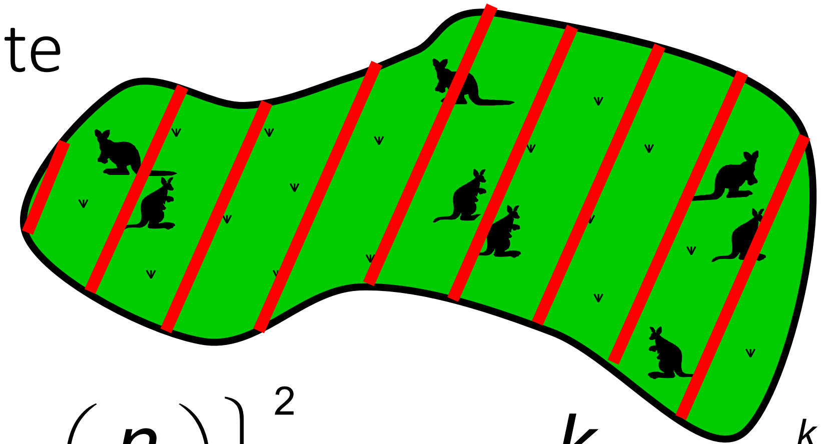


No unbiased estimator exists for estimating variance from a single systematic sample

Systematic samples advice

Usually, do nothing!

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

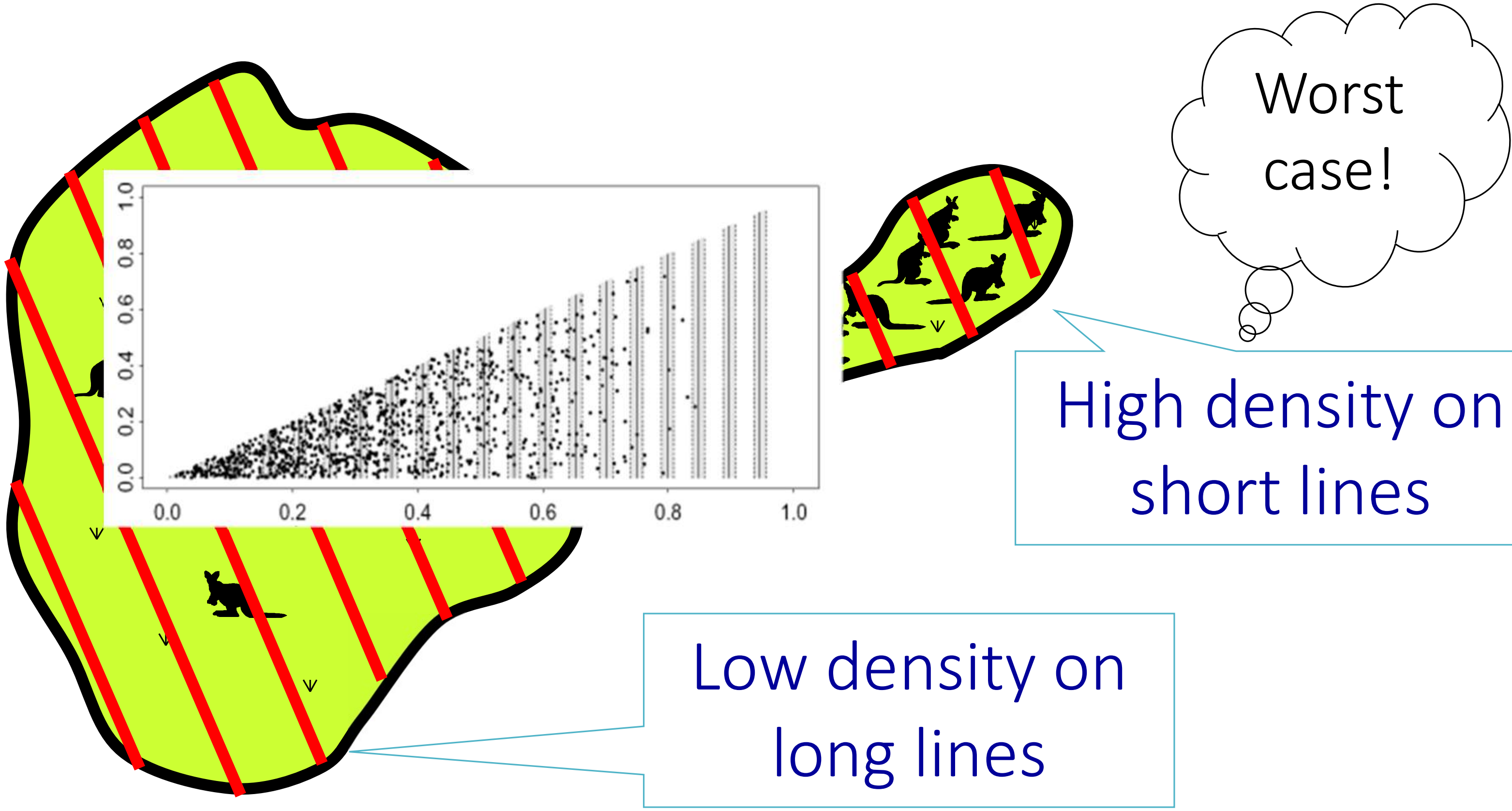


$$\left\{ cv\left(\frac{n}{L}\right)\right\}^2 = \frac{k}{n^2(k-1)}\sum_{i=1}^{k} \ell_i^2 \left(\frac{n_i}{\ell_i} - \frac{n}{L}\right)^2$$

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

If there are strong trends, variance might be significantly overestimated



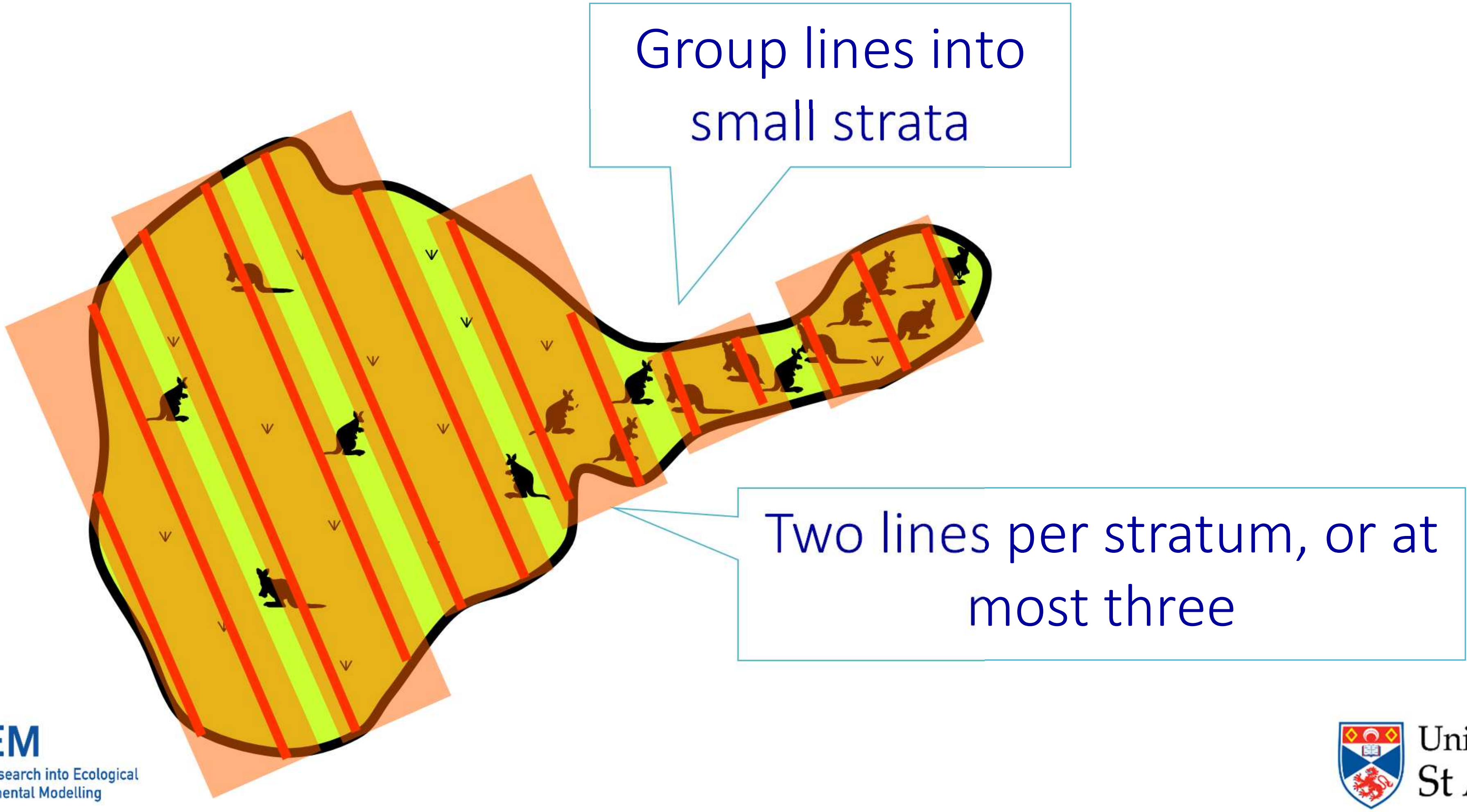
High density on short lines

Low density on long lines

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Post-stratification can give much better variance estimates



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

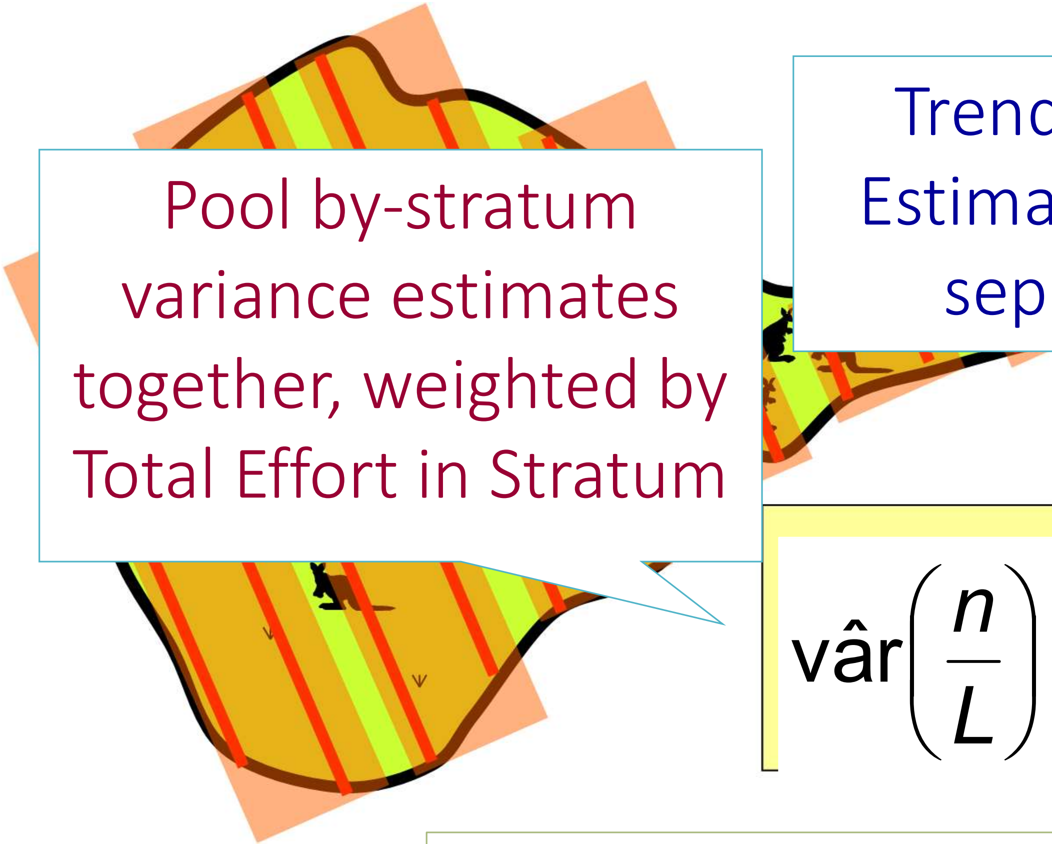
In Distance:

The encounter rate variance can be specified in the `dht2` function with the `er_est` argument

`dht2(model, flatfile, er_est)`

- The options follow the notation used in Fewster *et al*. (2009)
- The default is `er_est="R2"` – random line placement with unequal line length
- For systematic estimators, successive pairs of lines will be grouped together, according to the Sample.Label and so labels should be numeric (e.g. lines 1 and 2 grouped)
- If there are an odd number of lines, the last 3 will be grouped

Post-stratification can give much better estimates of variance



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

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

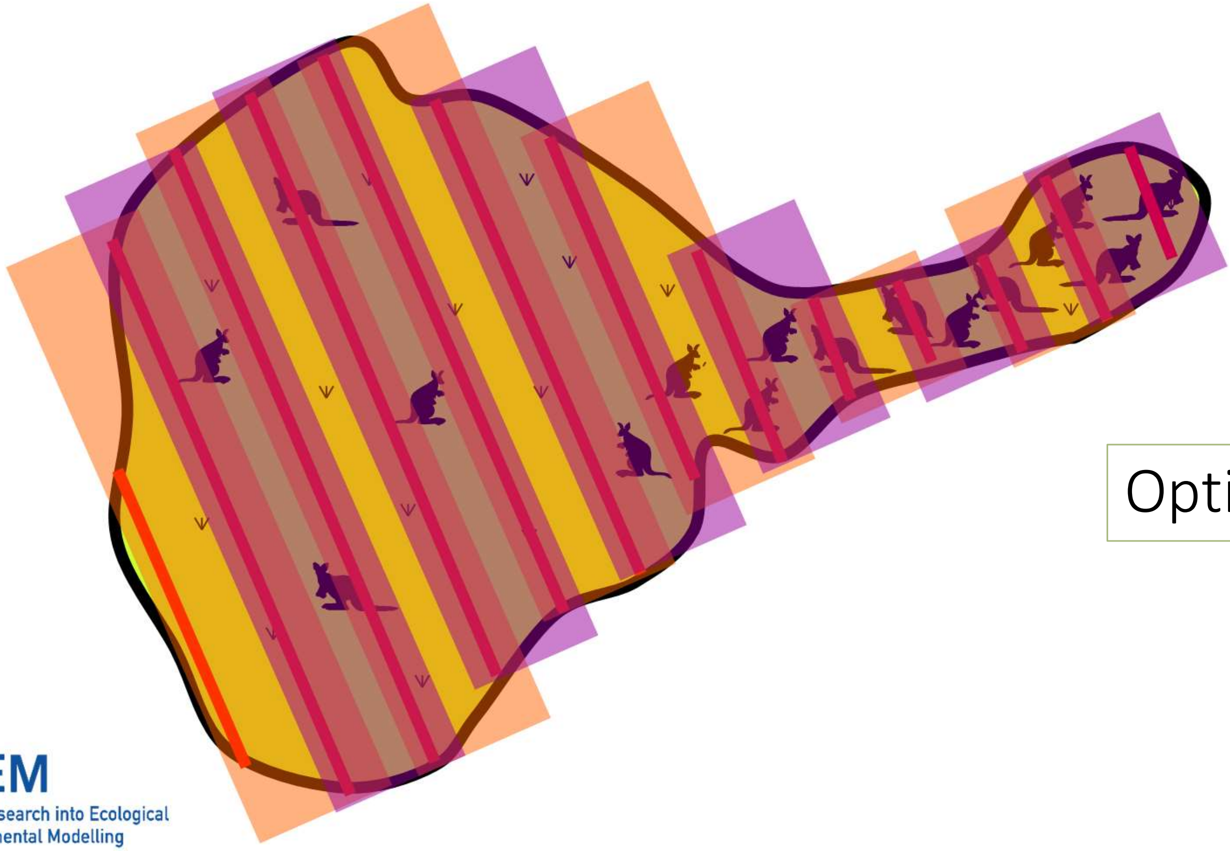
$$\hat{\mathrm{var}}\left(\frac{n}{L}\right) = \frac{1}{L^2}\sum_{h=1}^{H} L_h^2 \, \hat{\mathrm{var}}_h\left(\frac{n_h}{L_h}\right)$$

Option is `er_est="S2"`

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

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



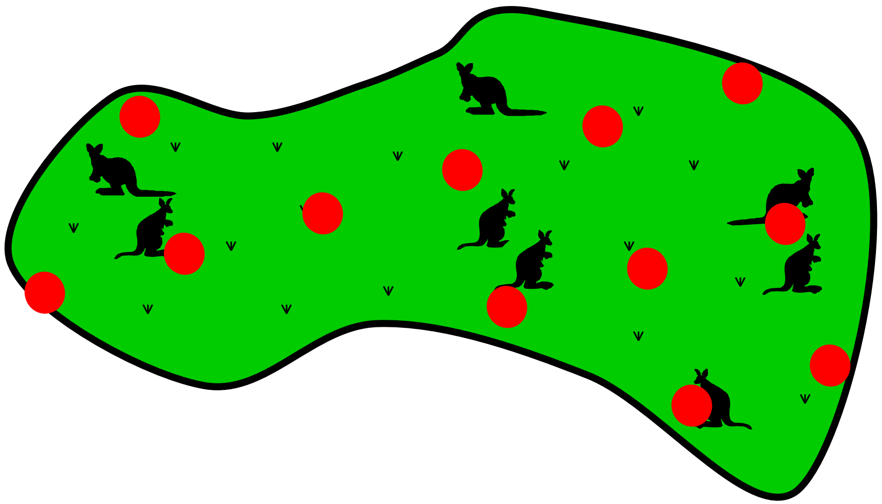
Option is `er_est="O2"`

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Point transect surveys

Default (and only) option is `er_est="P2"`



CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews