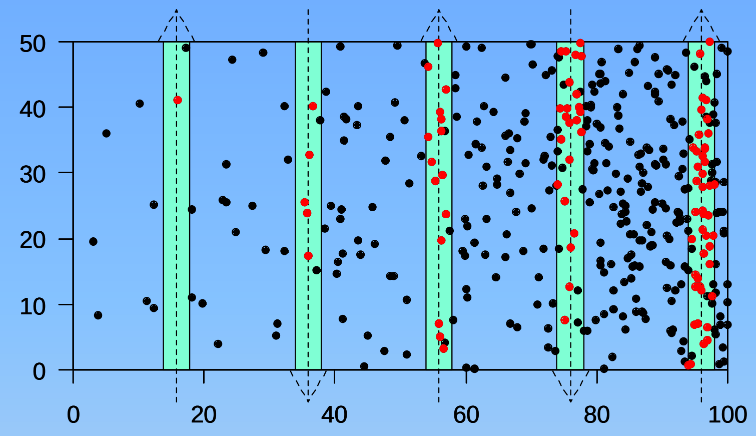
Model-based methods for distance sampling

CS Oedekoven and ST Buckland

National Centre for Statistical Ecology

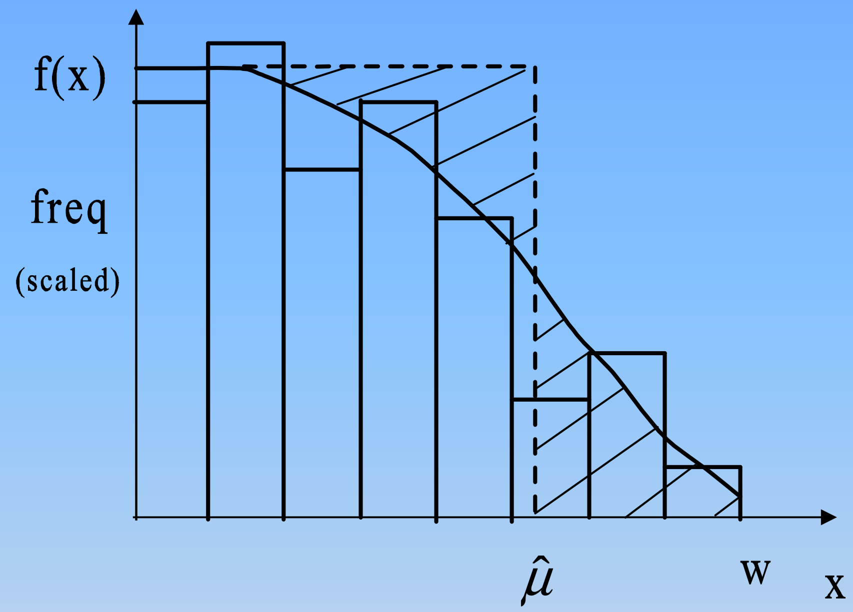
CREEM
Centre for Research into Ecological and Environmental Modelling

Conventional distance sampling



- A form of plot sampling, where the plots are circles (point transect sampling) or strips (line transect sampling)
- Not every animal on the sampled plots is detected, but we assume all animals on the line or point are detected

Line transect sampling



We use the fall-off in detections with distance to estimate the effective width of the strip surveyed and hence

$$\hat{P}_A = \hat{\mu}/w$$

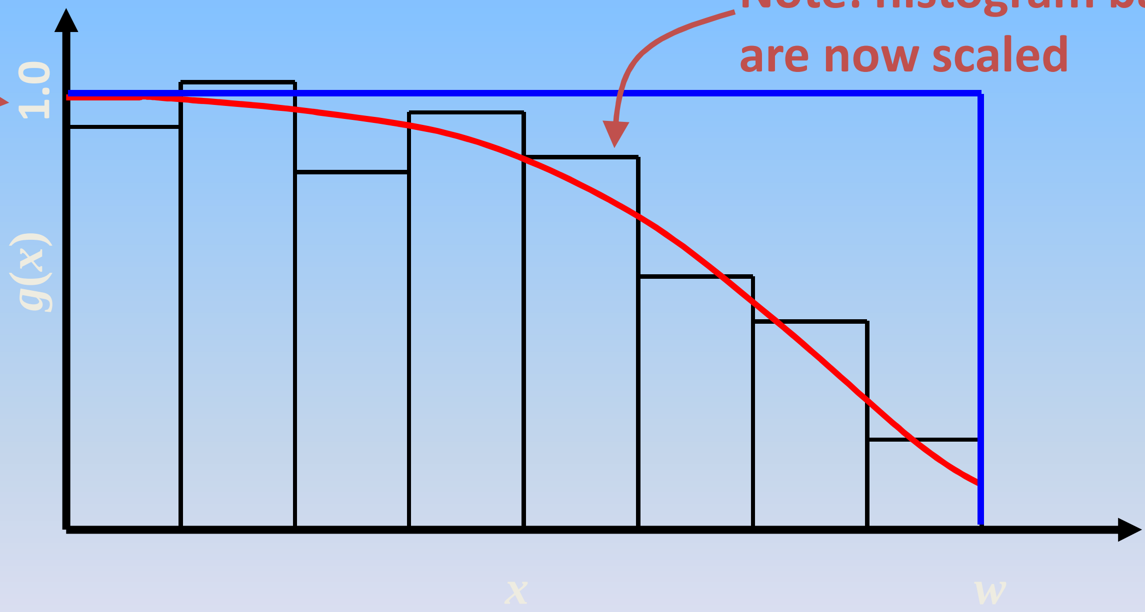
This assumes that animals are uniformly distributed wrt distance from the line

The detection function, *g*(*x*)

- *g*(*x*) = probability of detecting an animal, given that it is at distance *x* from the line

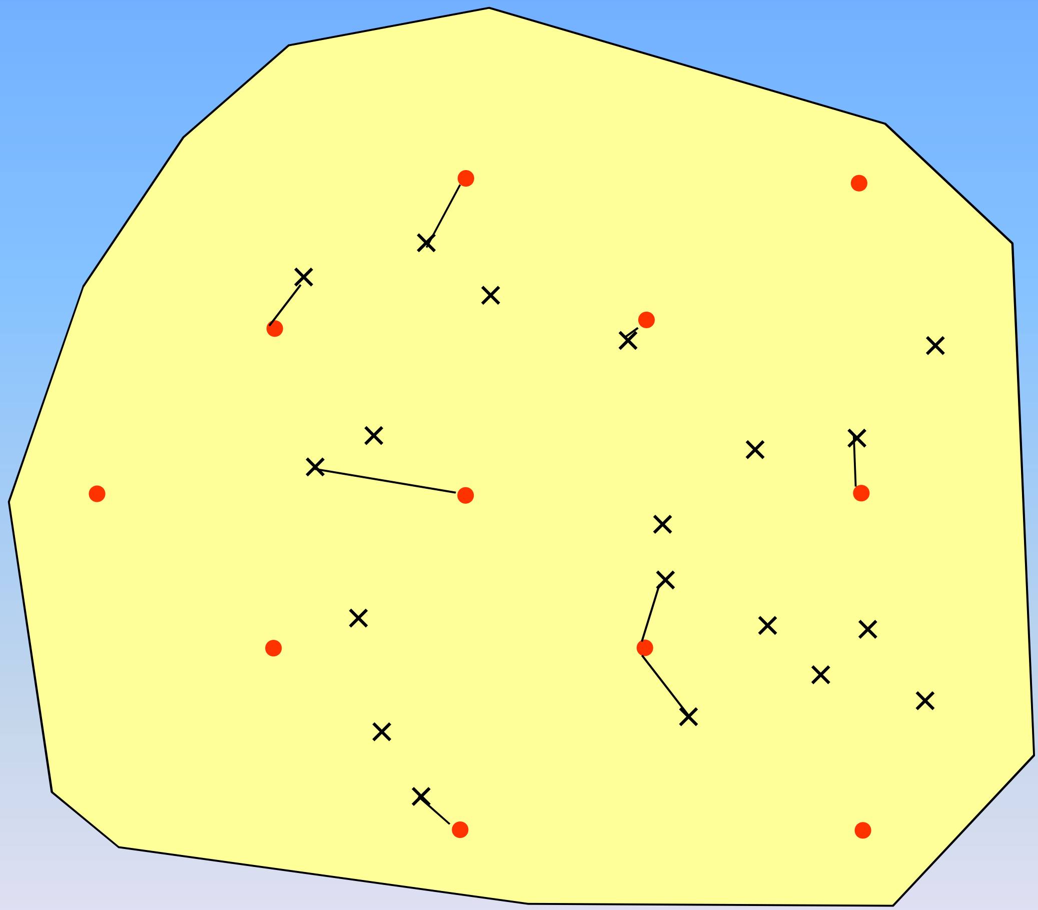
We assume $g(0) = 1$

Note: histogram bars are now scaled



$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{\int_0^w \hat{g}(x)dx}{w}$$

Point transect sampling



Random points or systematic grid of points randomly placed; observer records distance to any detected animals

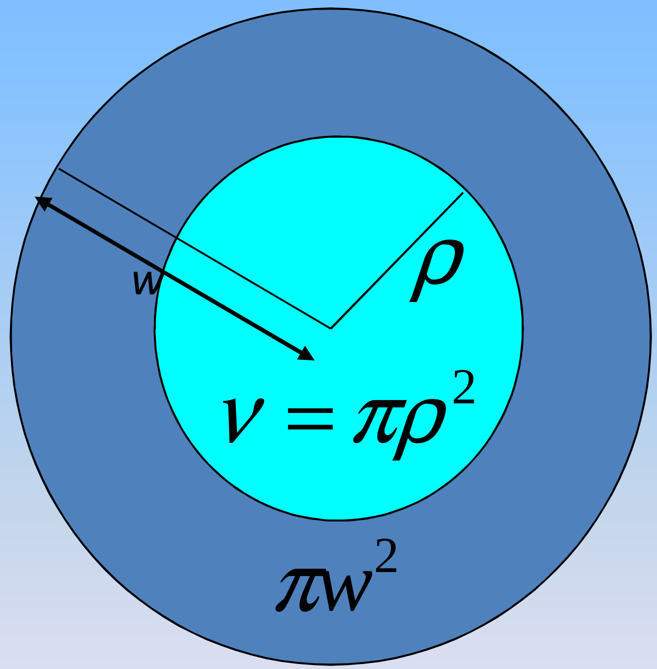
Point transect sampling

For *k* point counts with certain detection to distance *w*:

$$\hat{D} = \frac{n}{k\pi w^2}$$

How does this change if detection is uncertain?

Effective radius and effective area



ρ = effective radius

ν = effective area

Covered area:

$$a = k\pi w^2$$

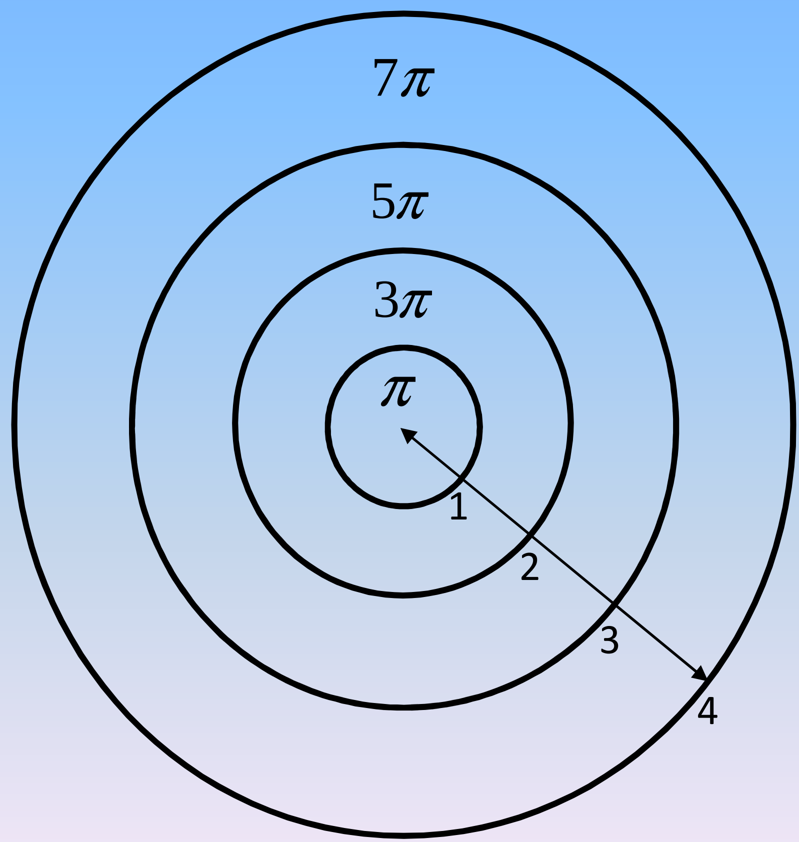
Proportion detected:

$$P_a = \frac{k\pi\rho^2}{k\pi w^2} = \frac{\rho^2}{w^2}$$

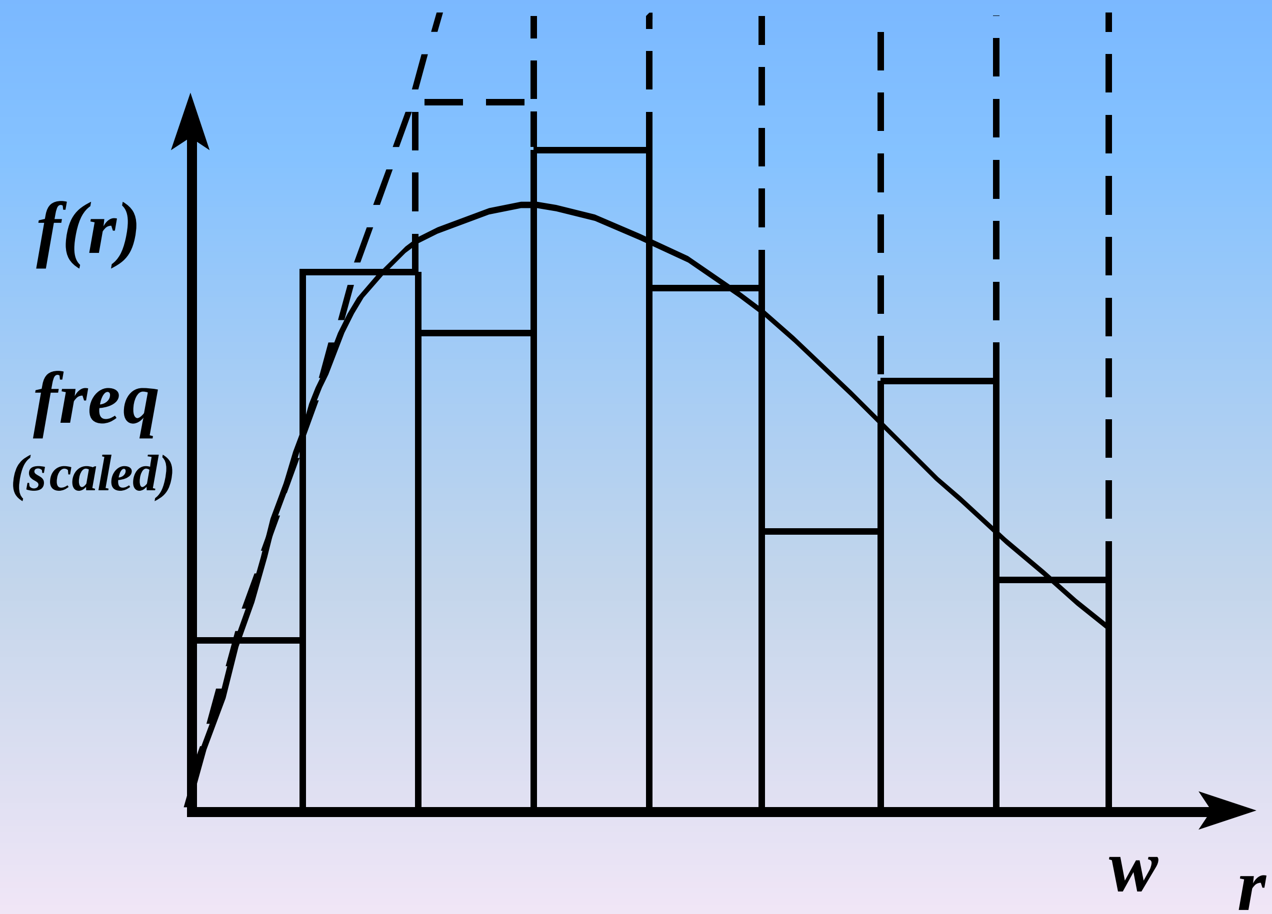
Estimated density:

$$\hat{D} = \frac{n}{a\hat{P}_a} = \frac{n}{k\pi w^2 \times \hat{\rho}^2 / w^2} = \frac{n}{k\pi\hat{\rho}^2}$$

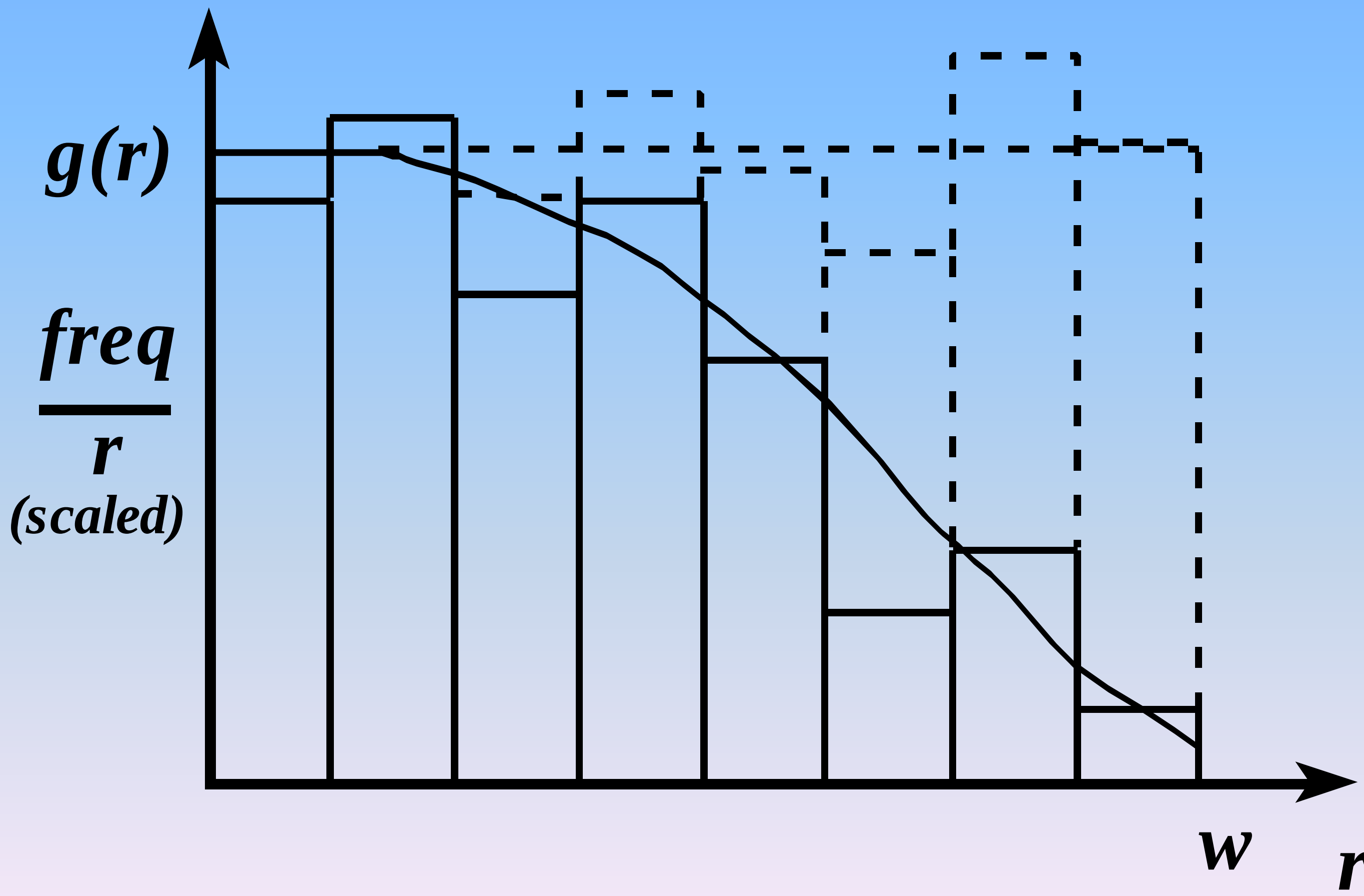
Area and hence number of birds increase linearly with distance:

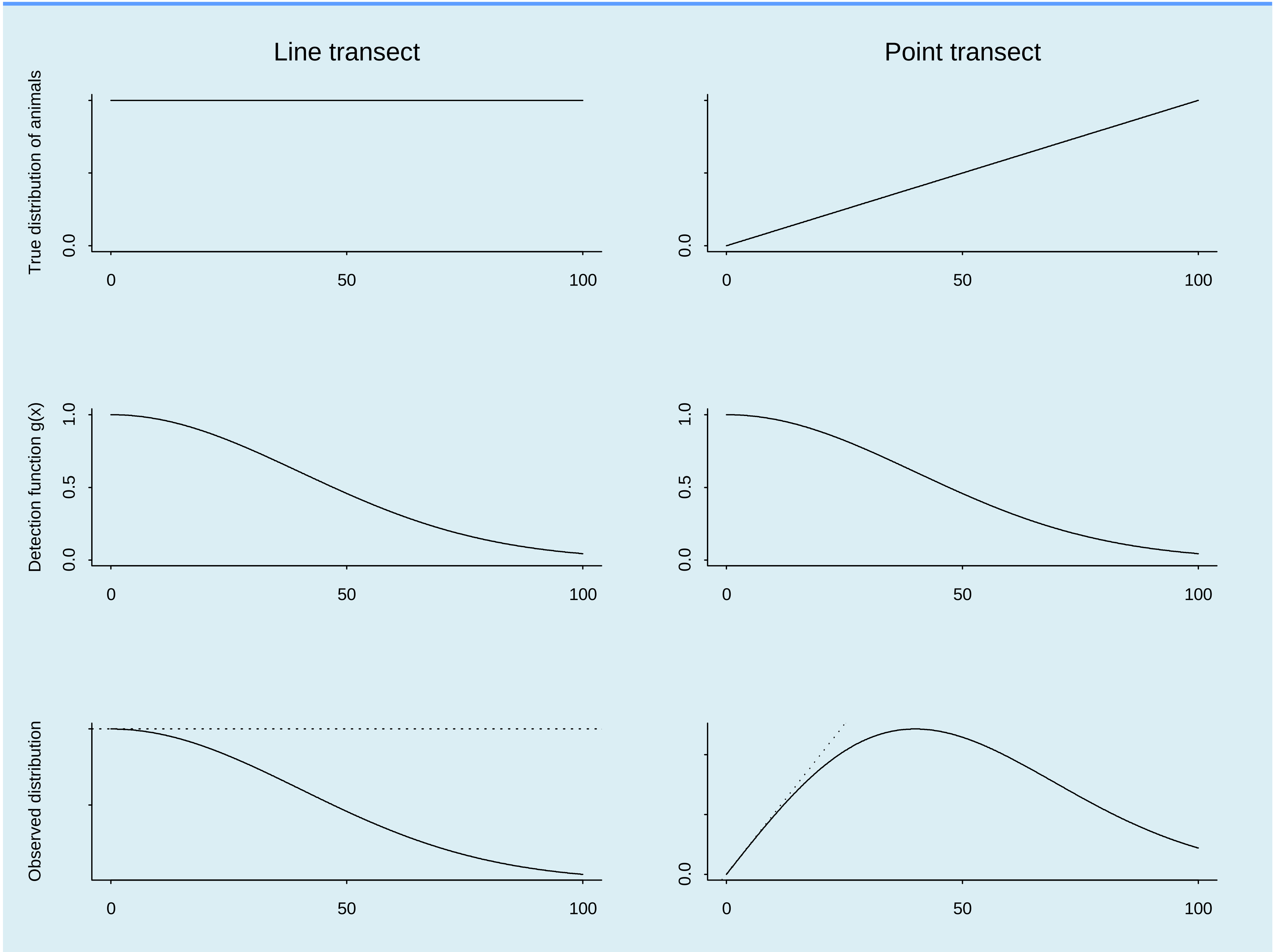


Probability density function



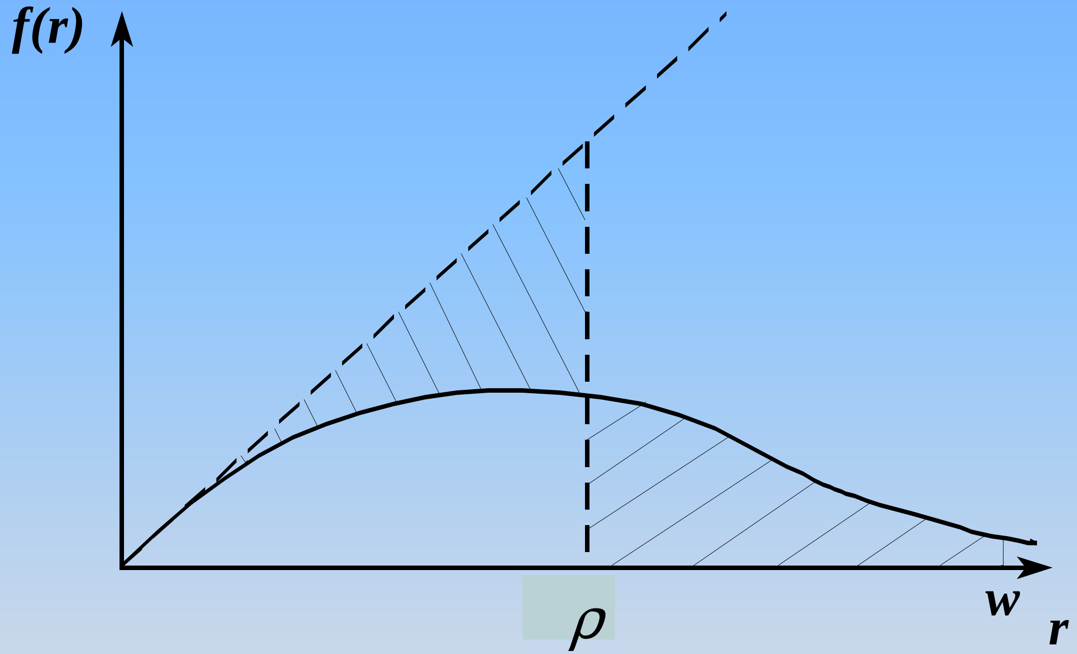
Detection function





Line transect
Point transect
True distribution of animals
Detection function g(x)
Observed distribution
0.0
0.5
1.0
0
50
100

The effective radius ρ …

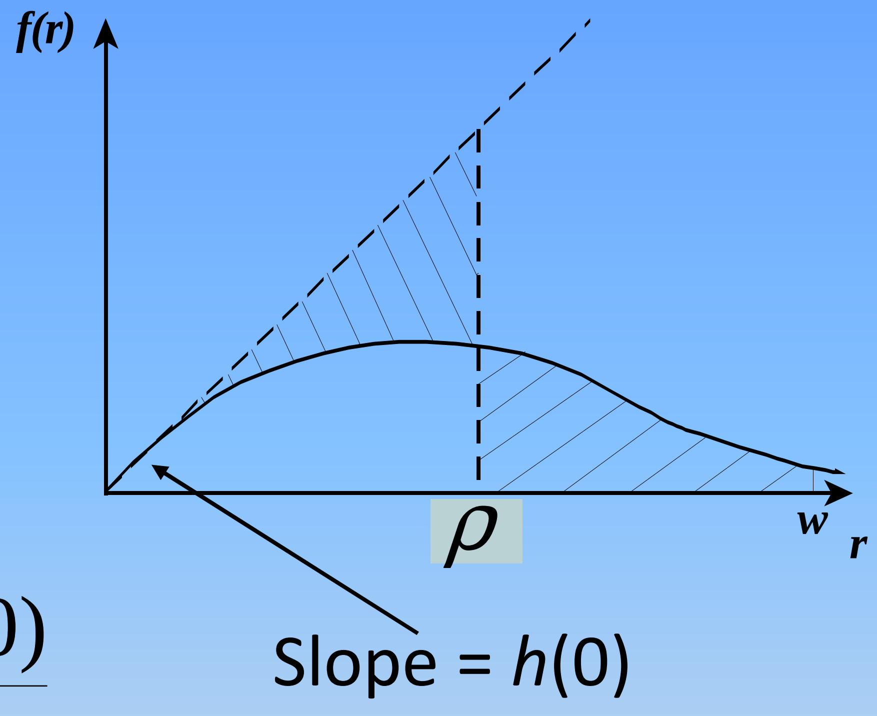


… is the distance such that as many birds beyond ρ are detected as are missed within ρ of the point.

Area under curve:

$$\int_0^w f(r)dr = 1$$

Area of triangle:



$$\frac{\rho \times \rho f'(0)}{2} = \frac{\rho^2 h(0)}{2}$$

Hence $\hat{\rho}^2 = \frac{2}{\hat{h}(0)}$ and $\hat{v} = \frac{2\pi}{\hat{h}(0)}$

so that

$$\hat{D} = \frac{n\hat{h}(0)}{2\pi k}$$

Conventional distance sampling

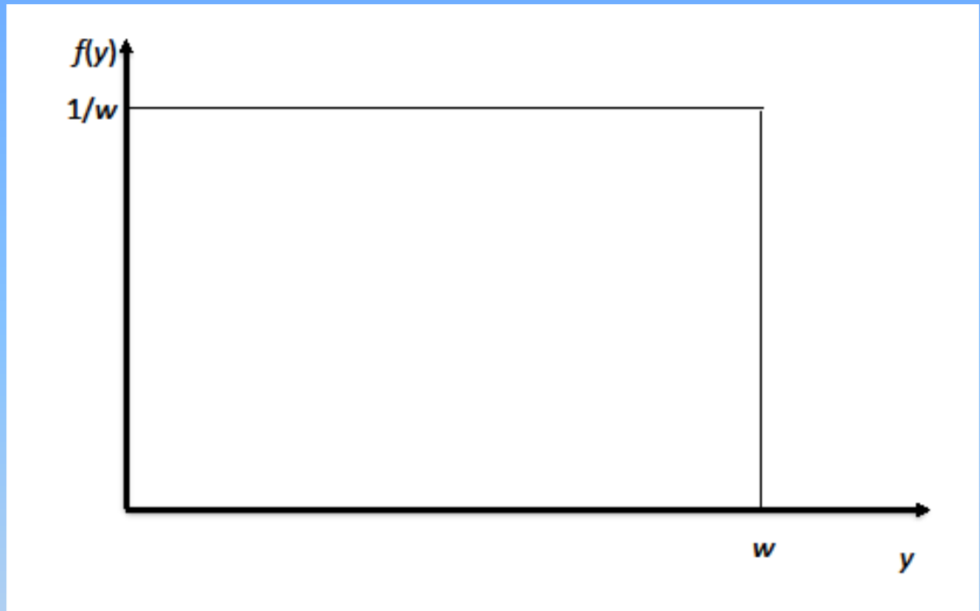
- Mixture of model-based and design-based
- Model-based: estimating the probability that an animal on a plot is detected
- Design-based: extrapolation of density on the sample plots to the whole study area …
  … and to ensure animals are uniformly distributed wrt distance from the line (line transects) or on plots (point transects)

Conventional distance sampling (CDS):

$$L_y = \prod_{i=1}^{n} f_y(y_i) = \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a}$$

where y_i is distance from the line or point
and $\pi(y_i)$ is uniform for line transect sampling,
and triangular for point transect sampling

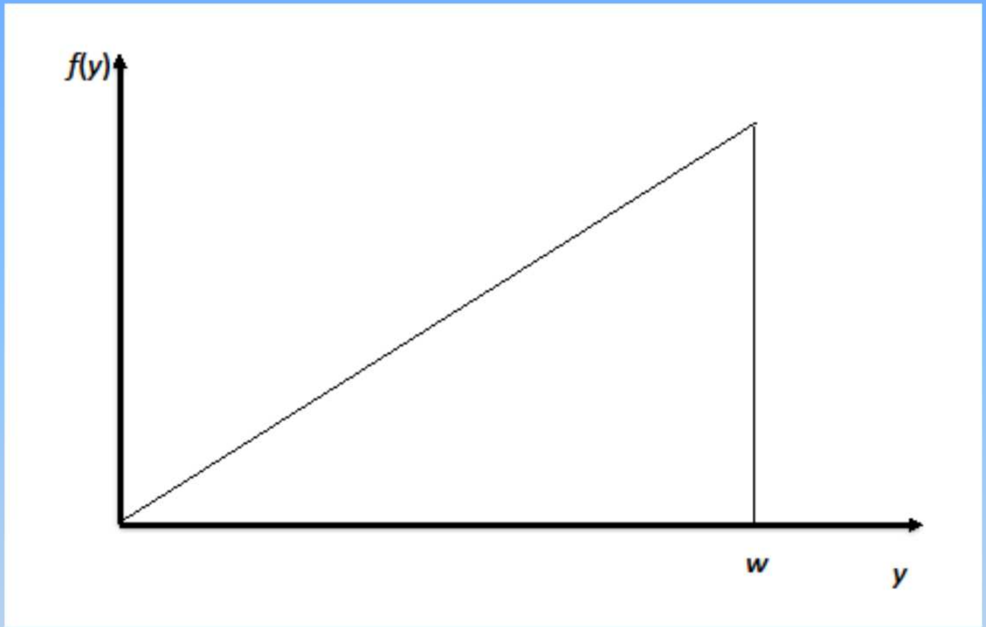
CDS: line transect sampling, $\pi(y) = \frac{1}{w}$



$$L_y = \prod_{i=1}^{n} f_y(y_i) = \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a} = \frac{\prod_{i=1}^{n} g(y_i)}{(wP_a)^n}$$

where $P_a = \frac{1}{w}\int_0^w g(y)dy$

CDS: point transect sampling, $\pi(y) = \frac{2y}{w^2}$



$$L_y = \prod_{i=1}^{n} f_y(y_i) = \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a} = \frac{2^n \prod_{i=1}^{n} y_i g(y_i)}{(w^2 P_a)^n}$$

where $P_a = \frac{2}{w^2} \int_0^w y g(y) dy$

Full likelihood:

$$L_{n,y} = L_n \times L_y$$

where L_n might be a binomial or Poisson likelihood

We can use either maximum likelihood
or Bayesian methods for inference

Multiple covariate distance sampling (MCDS):

$$L_{y|z} = \prod_{i=1}^{n} f_{y|z}(y_i|z_i)$$ where z_i is (vector of) covariate value(s)

Estimate probability of detection then use Horvitz-Thompson-like estimator:

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{\hat{p}_i}$$ or for clustered populations $$\hat{N} = \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i}$$

Full likelihood:

$$L_{n,z,y} = L_n \times L_z \times L_{y|z}$$

Intermediate option:

$L_n \times L_{y|z}$ and use Horvitz-Thompson-like estimator

Mark-recapture distance sampling (MRDS):

$L_{\omega} L_{y|z}$ where ω represents possible capture histories

Estimate probability of detection then use Horvitz-Thompson-like estimator

Full likelihood:

$$L_{n,\omega,z,y} = L_n \times L_{\omega} \times L_z \times L_{y|z}$$

Intermediate option:

$L_n \times L_{\omega} \times L_{y|z}$ and use Horvitz-Thompson-like estimator

See Borchers and Burnham (2004)

Model-based CDS

$$L_n = \binom{N}{n} (\pi_c P_a)^n (1 - \pi_c P_a)^{N-n}$$

$$L_y = \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a}$$

$$L_{n,y} = L_n \times L_y = \binom{N}{n} (\pi_c)^n (1 - \pi_c P_a)^{N-n} \prod_{i=1}^{n} g(y_i)\pi(y_i)$$

Borchers et al 2002; Royle and Dorazio 2008

Model-based CDS

For grouped distance data, we simply replace L_y by a multinomial likelihood

Model-based MCDS

$$L_n = \binom{N}{n} (\pi_c P_a)^n (1 - \pi_c P_a)^{N-n}$$

$$L_z = \prod_{i=1}^{n} \frac{P_a(\mathbf{z}_i)\pi_z(\mathbf{z}_i)}{P_a}$$

$$L_{y|z} = \prod_{i=1}^{n} \frac{g(y_i, \mathbf{z}_i)\pi(y_i)}{P_a(\mathbf{z}_i)}$$

$$P_a(\mathbf{z}_i) = \int_0^w g(y_i, \mathbf{z}_i)\pi(y_i)dy$$

$$P_a = \int_{\mathbf{z}} P_a(\mathbf{z})\pi_{\mathbf{z}}(\mathbf{z})d\mathbf{z}$$

$$L_{n,z,y} = L_n \times L_z \times L_{y|z} = \binom{N}{n} (\pi_c)^n (1 - \pi_c P_a)^{N-n} \prod_{i=1}^{n} \pi_z(\mathbf{z}_i) g(y_i, \mathbf{z}_i)\pi(y_i)$$

Similarly for model-based MRDS

Plot count models

We need to consider two cases:

1. We wish to estimate abundance N in the wider survey region

2. We wish to model plot abundance/density, e.g. in a designed distance sampling experiment

Plot count models

We could take our previous approach, and apply it to plots to replace L_n by:

$$L_{\{n_k\}} = \frac{N!}{\prod_{k=1}^{K} n_k!\,(N-n)!} \prod_{k=1}^{K} (\pi_k P_k)^{n_k} \left[1 - \sum_{k=1}^{K} \pi_k P_k\right]^{N-n}$$

Or if inference is restricted to plots:

$$L_{\{n_k\}} = \frac{N_c!}{\prod_{k=1}^{K} n_k!\,(N_c-n)!} \prod_{k=1}^{K} (\pi_k P_k)^{n_k} \left[1 - \sum_{k=1}^{K} \pi_k P_k\right]^{N_c-n}$$

Plot count models

When inference is restricted to plots, Poisson models are simpler:

$$L_{\{n_k\},y} = L_{\{n_k\}} \times L_y$$
$$= \prod_{k=1}^{K} \frac{\lambda_k^{n_k} e^{-\lambda_k}}{n_k!} \times \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a}$$

where count n_k has expectation

$$E(n_k) = \lambda_k = \exp\left(\sum_{q=1}^{Q} x_{qk}\beta_q + \log_e(a_k P_k)\right)$$

with $P_k = \frac{n_k}{N_k}$ which must be estimated

Plot count models

Model-based CDS – designed DS expts, grouped data:

$$L = \prod_{k=1}^{K} \prod_{j=1}^{u} \frac{(f_j \lambda_k)^{m_{jk}} e^{-f_j \lambda_k}}{m_{jk}!}$$

where $$f_j = \int_{c_{j-1}}^{c_j} f(y)dy = \int_{c_{j-1}}^{c_j} \frac{g(y)\pi(y)}{P_a} dy$$

and the c_j are cutpoints

Plot count models

Model-based CDS – designed DS expts, grouped data:

$$L = \prod_{k=1}^{K} \prod_{j=1}^{u} \frac{(f_j \lambda_k)^{m_{jk}} e^{-f_j \lambda_k}}{m_{jk}!}$$

The above can be shown to be equivalent to the plot abundance likelihood of Royle et al. (2004) and is a special case of the likelihood of Oedekoven et al., 2013 (indigo buntings).
We can extend this to MCDS provided covariates apply to plots or higher.

Plot count models – extensions

Random effects either in the model for λ_k (Oedekoven et al., 2013, 2014) or in the model for the scale parameter in the detection function (Oedekoven et al., 2015).

Spatial covariates in the model for λ_k allow density to be estimated throughout the study area, and hence abundance for any region of interest to be estimated.

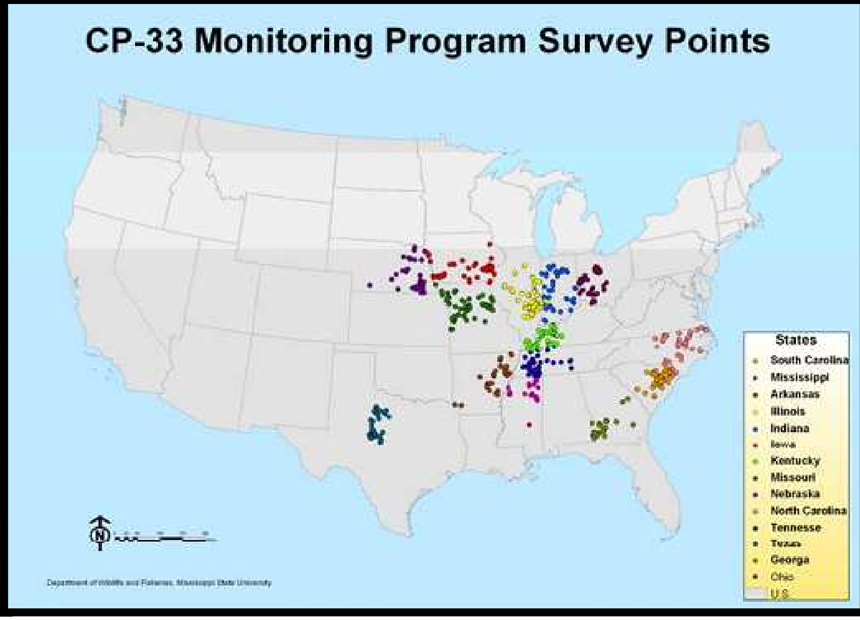
Plot count models – extensions

Point process models: Hedley and Buckland (2004); Johnson et al. (2010); Yuan et al. (in press).

$$\lambda_k = \int_k D(l) g\big(y(l)\big) dl$$

where the integral is over plot k, $D(l)$ is animal density at location l, and $y(l)$ the distance of location l from the line or point.

Case studies: point transects



>400 sites: pairs of 1 Control and 1 Treated Point

Repeated surveys

2 years

Indigo buntings

Oedekoven et al. (2013)

3 years

Northern bobwhite coveys

Oedekoven et al. (2014)

Three possible strategies

1. 2-stage. Model the detection function, then conditional on that fit, model the plot counts. Propagate uncertainty from first stage to second using the bootstrap.
2. Maximize the integrated likelihood, with the (unknown) probability of detection in the offset of the count model.
3. Define the likelihood as in 2, but use Bayesian methods to draw inference.

Bobwhites – exact distances

For site j, point p, visit r:

$$D_{jpr} = E[n_{jpr}] / \upsilon_{jpr}$$

$$E[n_{jpr}] = D_{jpr} \times \upsilon_{jpr}$$

$$n_{jpr} \sim Poisson(\lambda_{jpr})$$

$$\lambda_{jpr} = \exp(\beta_0 + b_j + \sum_{k=1}^{K} x_{kjpr}\beta_k + \log(\upsilon_{jpr}))$$

$$b_j \sim Normal\left(0, \sigma_b^2\right)$$

Oedekoven et al. (2014)

Integrated likelihood

Count: $$L_n(\boldsymbol{\beta}, \sigma_b \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} \prod_{p=1}^{P_j} \prod_{r=1}^{R_j} \frac{(\lambda_{jpr})^{n_{jpr}} e^{-\lambda_{jpr}}}{n_{jpr}!} \times \frac{1}{\sqrt{2\pi\sigma_b^2}} e^{-\frac{b_j^2}{2\sigma_b^2}} db_j$$

Det Fct: $$L_y(\Theta) = \prod_{i=1}^{n} \frac{g(y_i)\pi(y_i)}{P_a}$$

Integrated: $$L_{y,n}(\boldsymbol{\beta}, \sigma_b, \boldsymbol{\theta}) = L_y(\boldsymbol{\theta}) \times L_n(\boldsymbol{\beta}, \sigma_b \mid \boldsymbol{\theta})$$

Oedekoven et al. (2014)

N. bobwhite coveys: Model probabilities

Model	RJMCMC	Two-stage
Detection Function		
MCDS: Type	< 0.001	--
MCDS: State	--	0.01
MCDS: Year + State	--	0.16
MCDS: Type + State	< 0.001	0.02
MCDS: Year + Type + State	1.00	0.81
Count		
Type + State	--	0.003
Year + Type + State	--	0.01
Type + JD + State	0.89	0.1
Year + Type + JD + State	0.11	0.89

Oedekoven et al. (2014)

N. bobwhite coveys: Summaries for parameters

	RJMCMC		Two-stage	
	Mean	SD	MLE	BSE
Count model: random effects				
Standard deviation	0.82	0.05	0.78	0.04
Count model: fixed effects				
Intercept Density	-13.10	0.18	-13.23	0.33
Year 2007	-	-	0.17	0.13
Year 2008	-	-	0.17	0.11
Type Treatment	0.62	0.07	0.63	0.12
Julian Day	-0.01	0.002	-0.01	0.003
State IA	[illegible]	[illegible]	[illegible]	[illegible]
State IL	[illegible]	[illegible]	[illegible]	[illegible]
State IN	[illegible]	[illegible]	[illegible]	[illegible]
State KY	-0.47	0.25	-0.44	0.34
State MO	0.01	0.22	0.05	0.34
State MS	-0.43	0.25	-0.37	0.34
State NC	-1.39	0.26	-1.31	0.36
State SC	0.01	0.27	0.08	0.42
State TN	-1.10	0.28	-1.03	0.38
State TX	1.74	0.18	1.46	0.29

exp(0.62) = 1.86
95% CRI: (1.62,2.12)

exp(0.63)=1.88
95% CI: (1.43,2.03)

Oedekoven et al. (2014)

Indigo buntings – grouped distances

$$E[n_{jpri}] = N_{jpr} \times f_{ui}$$

$$n_{jpri} \sim Poisson(\lambda_{jpr} f_{ui})$$

$$\lambda_{jpr} = \exp(\beta_0 + b_j + \sum_{k=1}^{K} x_{kjpr} \beta_k)$$

$$b_j \sim Normal\left(0, \sigma_b^2\right)$$

$$f_{ui} = \int_{c_i}^{c_{i+1}} \pi(y) g(y) dy$$

Oedekoven et al. (2013)

Integrated likelihood

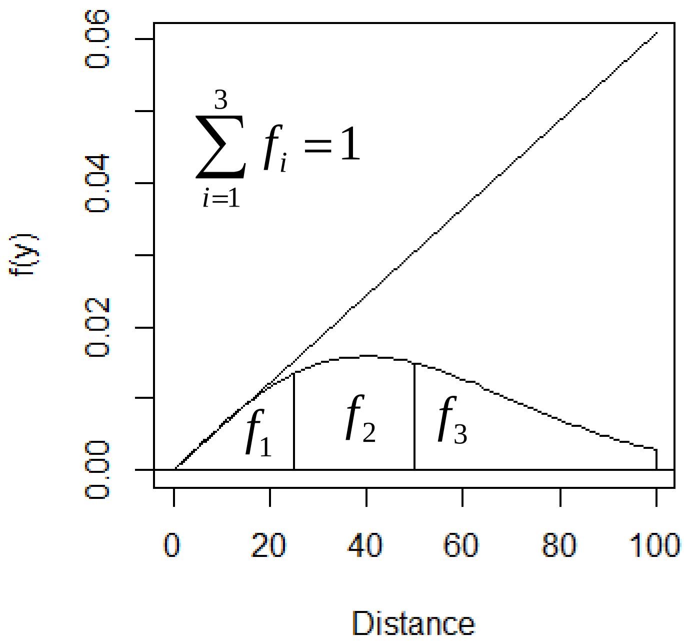
$$L_{yG,n}(\boldsymbol{\beta}, \sigma_b, \boldsymbol{\theta}) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} \prod_{p=1}^{P_j} \prod_{r=1}^{R_j} \prod_{i}^{I} \frac{(\lambda_{jpr} f_{ui})^{n_{jpri}} e^{-(\lambda_{jpr} f_{ui})}}{n_{jpri}!} \times \frac{1}{\sqrt{2\pi\sigma_b^2}} e^{-\frac{b_j^2}{2\sigma_b^2}} db_j$$

Oedekoven et al. (2013)

Conditional vs. unconditional likelihood of observed distances

$$f(y) = \frac{\pi(y)g(y)}{\int_0^w \pi(y)g(y)dy} \qquad f_i = \frac{\int_{c_i}^{c_{i+1}} \pi(y)g(y)dy}{\int_0^w \pi(y)g(y)dy}$$

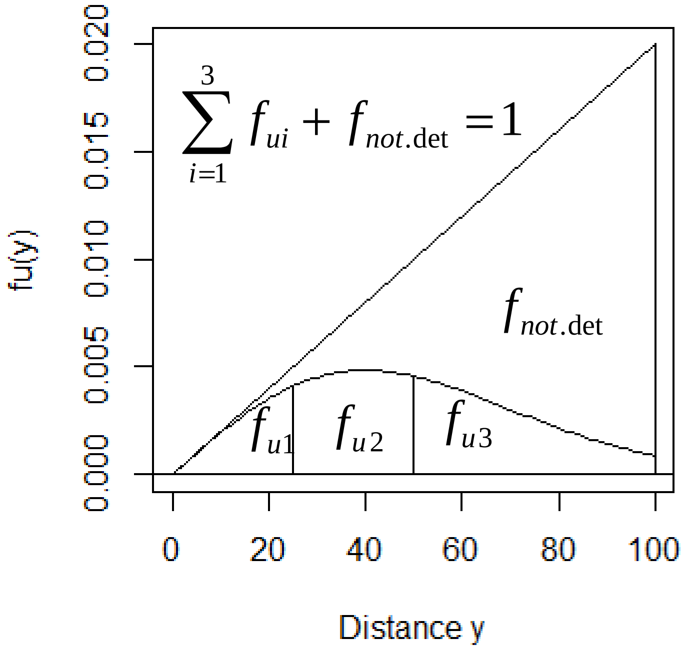
Buckland et al. 2001



Oedekoven et al. (2014)

$$f_u(y) = \pi(y)g(y) \qquad f_{ui} = \int_{c_i}^{c_{i+1}} \pi(y)g(y)dy$$

Royle et al. 2004



Oedekoven et al. (2013)

But the two approaches are equivalent!

Indigo buntings: Summaries for parameters

	Integrated		Two-stage	
	Mean	SD	MLE	BSE
Count model: random effects				
Standard deviation	0.50	0.02	0.49	0.04
Count model: fixed effects				
Intercept Abundance	-0.99	0.28	-	-
Intercept Density	-	-	-10.91	0.43
Type Treatment	0.30	0.02	0.30	0.04
Julian Day	[illegible]	[illegible]	[illegible]	[illegible]
State IL	[illegible]	[illegible]	[illegible]	[illegible]
State IN	[illegible]	[illegible]	[illegible]	[illegible]
State KY	[illegible]	[illegible]	[illegible]	[illegible]
State MO	0.53	0.29	0.32	0.36
State MS	1.25	0.31	0.91	0.37
State OH	1.11	0.30	0.92	0.37
State SC	0.59	0.31	0.28	0.39

exp(0.30) = 1.35
95% CI: (1.30,1.40)

exp(0.30)=1.35
95% CI: (1.25,1.46)

Oedekoven et al. (2013)

References

Borchers, D.L., Buckland, S.T. and Zucchini, W. 2002. *Estimating Animal Abundance: Closed Populations*. Springer Verlag, London. 314pp.

Borchers, D.L. and Burnham, K.P. 2004. General formulation for distance sampling. Pp. 6-30 in *Advanced Distance Sampling* (eds Buckland et al.). OUP.

Buckland, S.T., Russell, R.E., Dickson, B.G., Saab, V.A., Gorman, D.N. and Block, W.M. 2009. Analysing designed experiments in distance sampling. *JABES* **14**, 432-442.

Hedley, S.L. and Buckland, S.T. 2004. Spatial models for line transect sampling. *JABES* **9**, 181-199.

Johnson, D.S., Laake, J.L. and Ver Hoef, J.M. 2010. A model-based approach for making ecological inference from distance sampling data. Biometrics 66, 310—318.

Oedekoven, C.S., Buckland, S.T., Mackenzie, M.L., Evans, K.O. and Burger, L.W. 2013. Improving distance sampling: accounting for covariates and non-independency between sampled sites. *J. App. Ecol.* **50**, 786-793.

Oedekoven, C.S., Buckland, S.T., Mackenzie, M.L., King, R., Evans, K.O. and Burger, L.W. 2014. Bayesian methods for hierarchical distance sampling models. *JABES* **19**, 219-239.

Oedekoven, C.S. , Laake, J.L. and Skaug, H.J. 2015. Distance sampling with a random scale detection function. Environmental and Ecological Statistics.

Royle, J.A., Dawson, D.K. and Bates, S. 2004. Modeling abundance effects in distance sampling. Ecology 85, 1591-1597.

Royle, J.A. and Dorazio, R.M. 2008. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press.

Yuan, Y., Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., Buckland, S.T., Rue, H. and Gerrodette, T. in press. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. Annals of Applied Statistics.