Choosing a Detection function

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Overview

Formal definition

Criteria for a good detection function model

Key functions and adjustment terms

Fitting models using `ds()`

Choosing the number of parameters

Comments about truncation

Formal definition

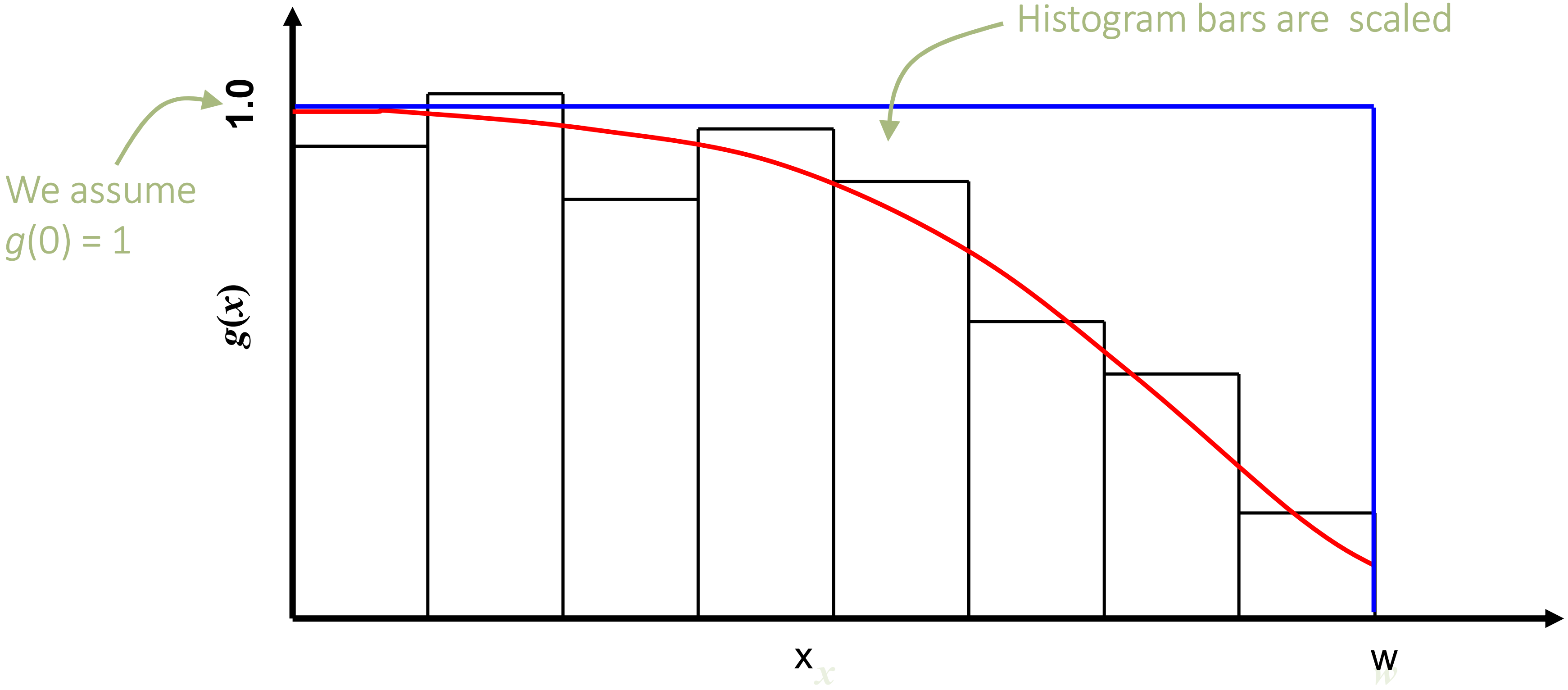
The detection function describes the relationship between distance and the probability of detection

Formally denoted by $g(x)$ (usually referred to as 'g of x')

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

Key to the concept of distance sampling

The detection function, g(x)



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

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Modelling g(x)

g(x) represents the underlying relationship between detection probability and distance

However, the true form of g(x) is unknown to us

We need to estimate g(x) by fitting a model to our data

i.e., we need to find a curve that will approximate the underlying relationship

Criteria for robust estimation

Four main criteria for a good model:

1. Model robustness – use a model that will fit a wide variety of plausible shapes for g(x)
2. Shape criterion – use a model with a 'shoulder' – i.e. $g'(0)=0$
3. Pooling robustness – use a model for the average detection function, even when many factors affect detectability
4. Estimator efficiency – use a model that will lead to a precise estimator of density

Key functions

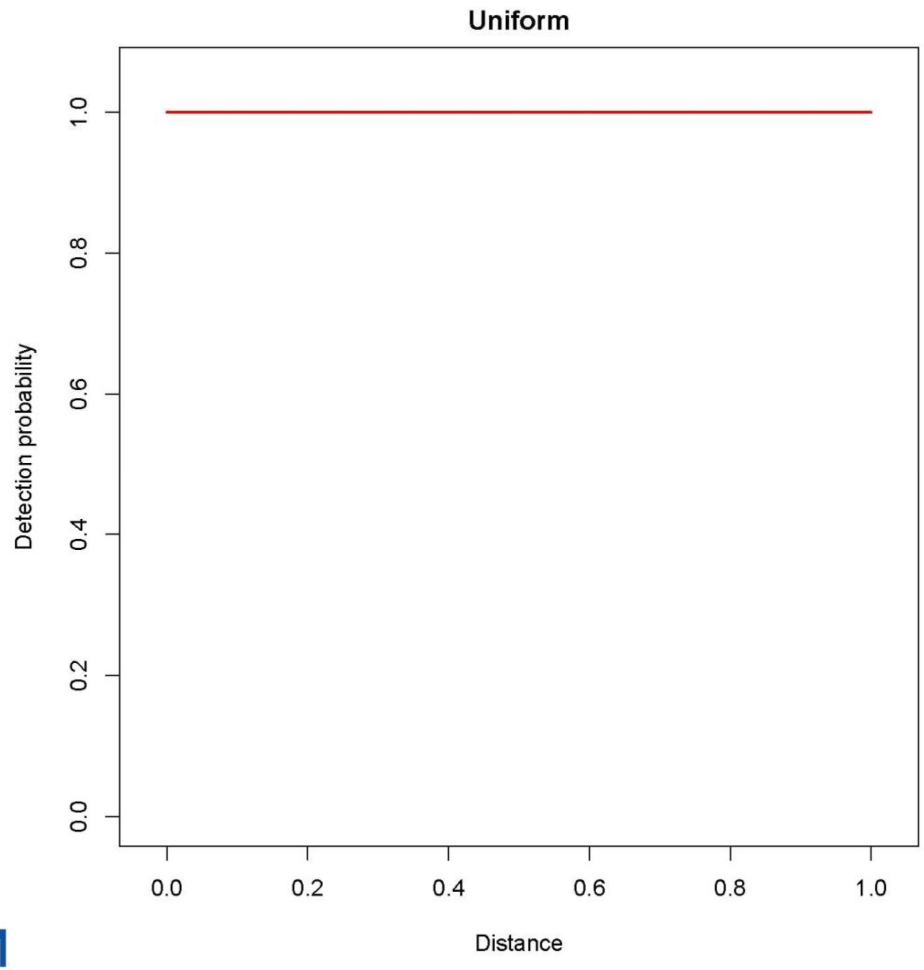
The first step in constructing a model for g(x) is to choose a key function

This determines the basic model shape

Three key functions available in `ds()`:

1. Uniform
2. Half normal
3. Hazard rate

Key functions (cont.)



- Model formula:

$$g(x) = 1,\ x \leq w$$

- Parameters = 0
- Shape criterion?

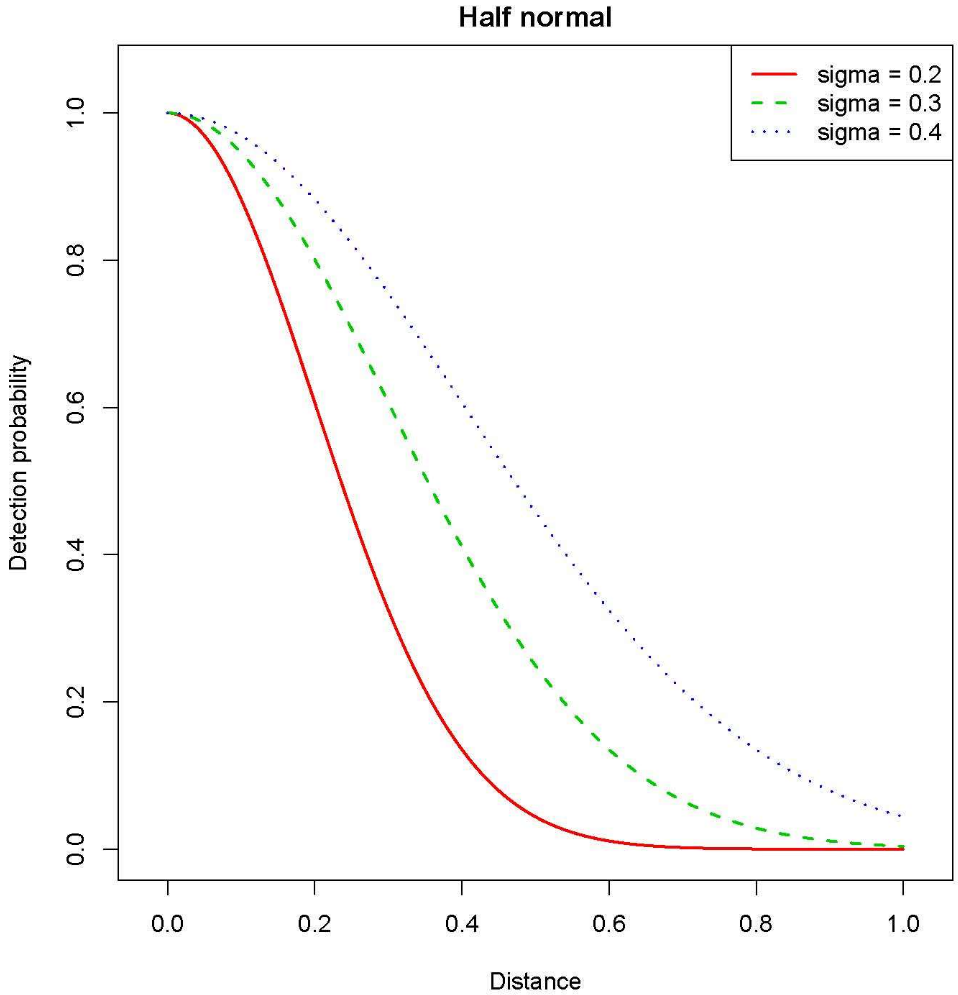
  Yes
- Model robust?

  No

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Key functions (cont.)



- Model formula:

$$g(x) = \exp\left(\frac{-x^2}{2\sigma^2}\right),\ x \leq w$$

- Parameters = 1
- Shape criterion?

  Yes

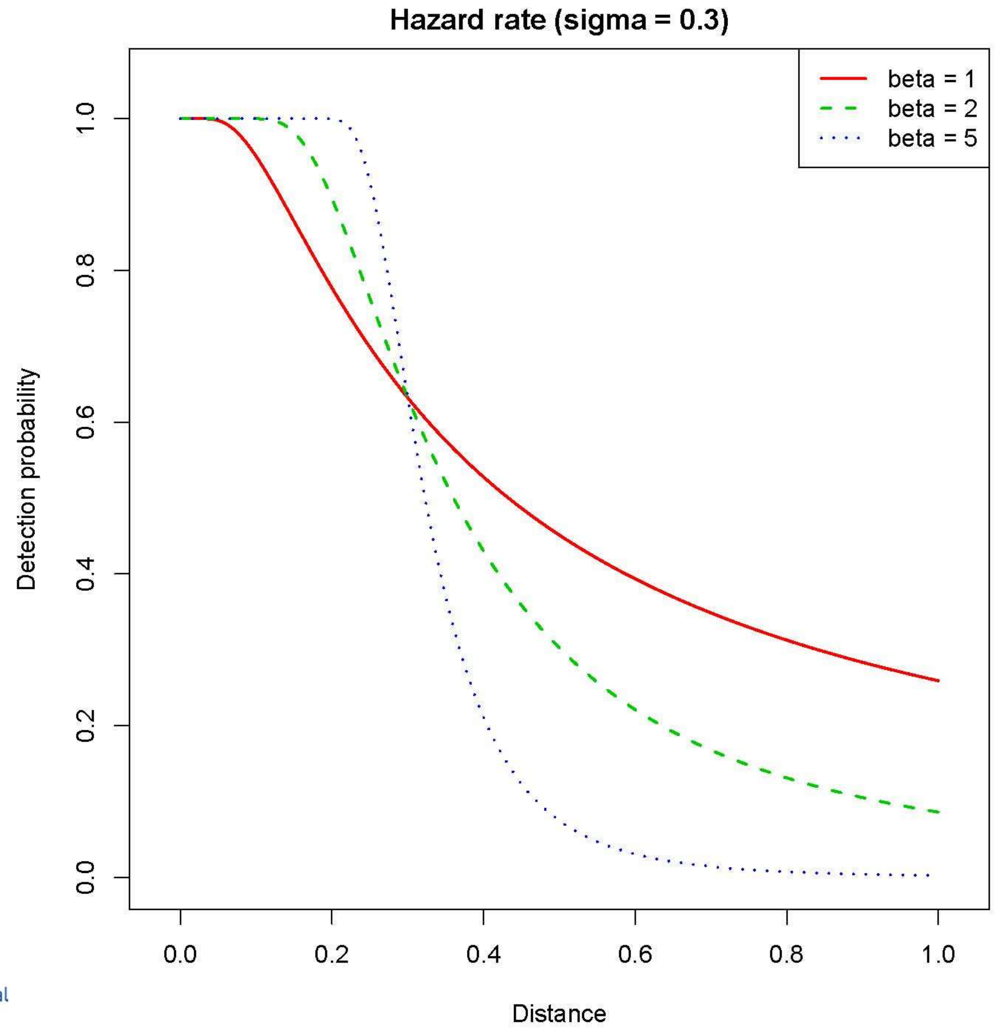
- Model robust?

  Somewhat

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Key functions (cont.)



- Model formula:

$$g(x) = 1 - \exp\left[-\left(\frac{x}{\sigma} \right)^{-\beta} \right], x \leq w$$

- Parameters = 2
- Shape criterion?

  Yes

- Model robust?

  Yes

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Key functions in Distance

Load package (at start of R session)

```
library(Distance)
```

Fit detection function

```
ds(data, key)
```

Contains column called distance

Options are “hn”, “hr” and “unif”
E.g. `ds(data, key=“hn”)`

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Adjustment terms

Models can be made more robust by adding a series of adjustment terms (also called series expansion or series adjustment) to the key function

Key function × (1 + Series)

Series = $\alpha_1 \times \text{term}_1 + \alpha_2 \times \text{term}_2 + \ldots..$ etc.

The α_i parameters must be estimated

Resulting curve model is scaled so that g(0)=1

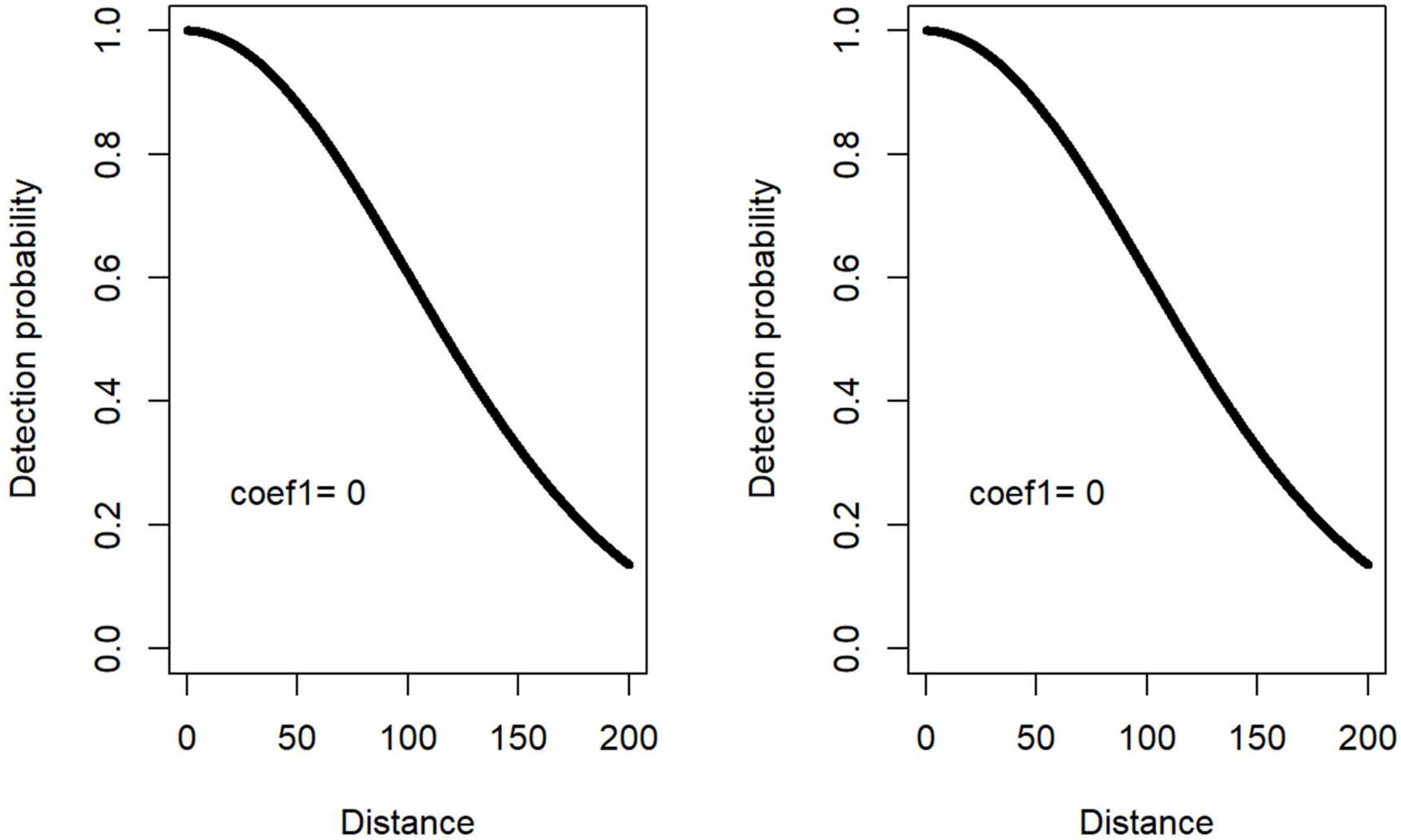
The number of adjustment terms needs to be chosen

Adjustment terms

Distance allows the selection of three types of series (one type per model)

Key function	Series adjustment
Uniform*	Cosine*
Half normal†	Hermite polynomial†
Hazard rate	Simple polynomial

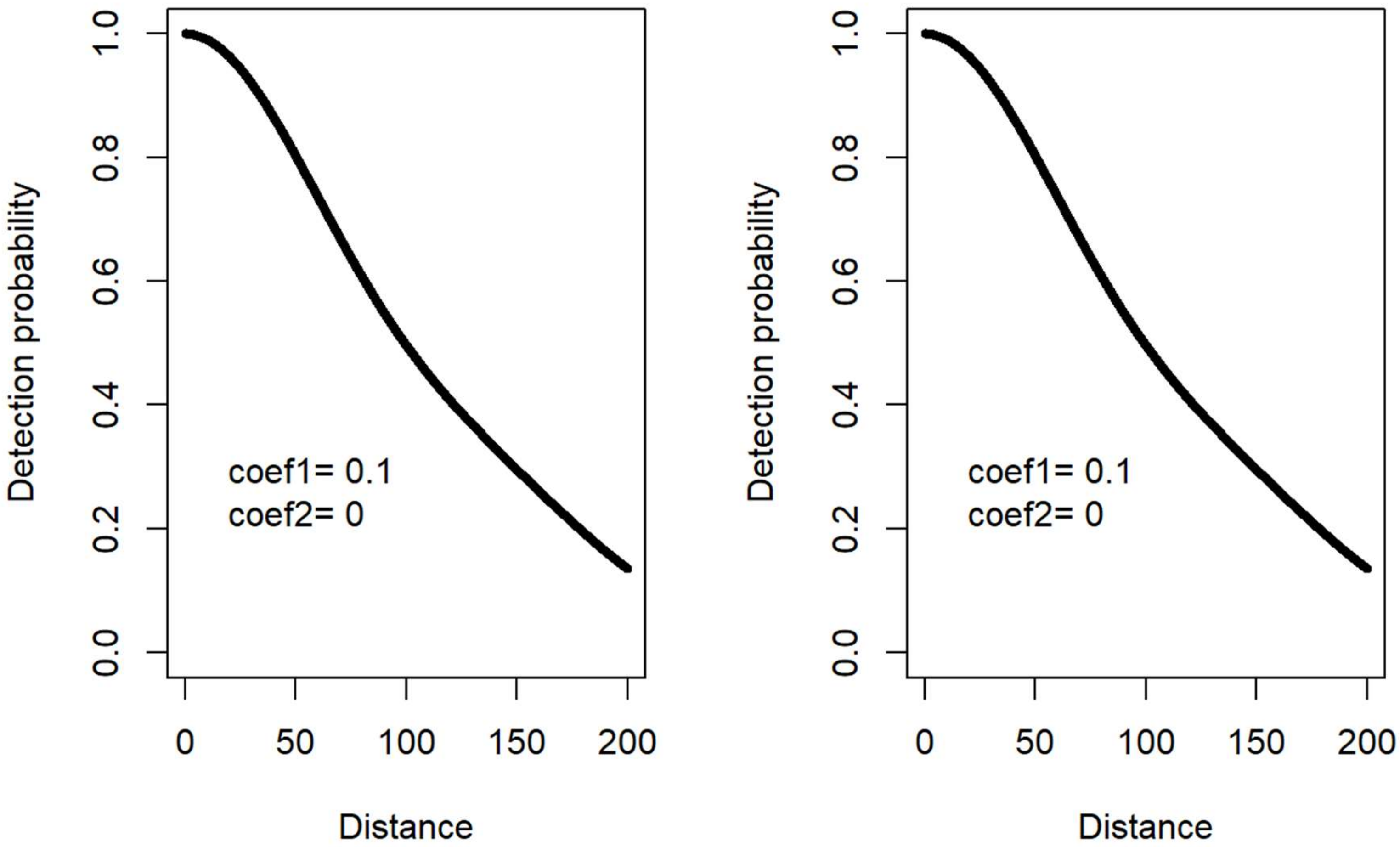
Half normal key, single cosine adjustment term



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Half normal key, two cosine adjustment terms



Adjustments in Distance

Fit a half normal detection function with cosine adjustments

```
ds(data, key="hn", adjustment="cos")
```

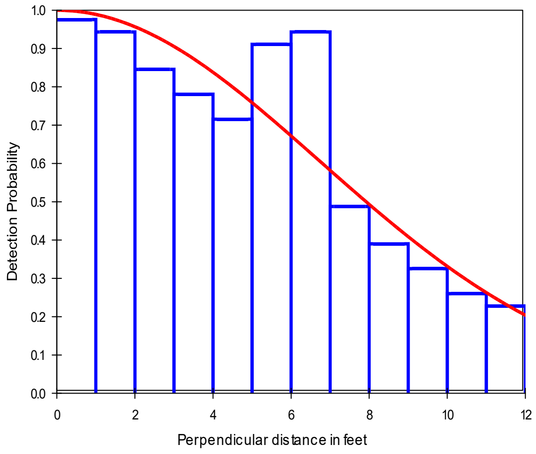
Options are

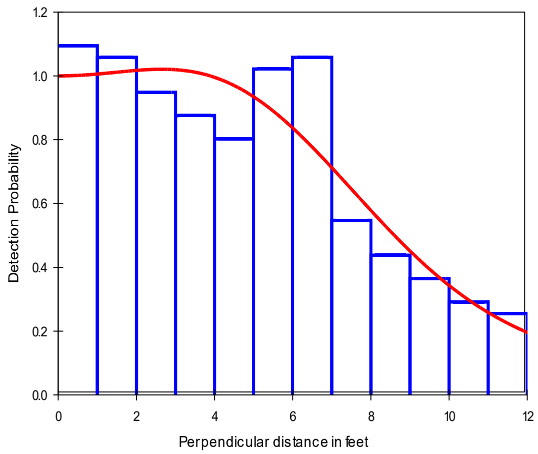
- "cos" - cosine
- "herm" – Hermite polynomial
- "poly" – simple polynomial
- NULL – no adjustments will be fitted

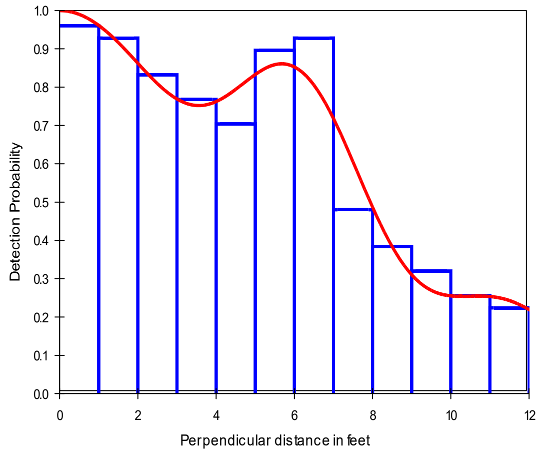
CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Adjustment terms – how many?







Half normal	Half normal	Half normal
0 adjustment terms	1 adjustment term	5 adjustment terms
1 parameter	2 parameters	6 parameters
$\hat{P}_a = 0.65$	$\hat{P}_a = 0.72$	$\hat{P}_a = 0.63$
$CV(\hat{P}_a) = 5.8\%$	$CV(\hat{P}_a) = 11.6\%$	$CV(\hat{P}_a) = 19.9\%$

Note: There is a monotonicity constraint in Distance that is switched on by default to prevent detection functions from increasing. The constraint had to be turned off to produce the third plot. The third plot is for demonstration only – it would not be a good detection function to choose (unless there was a biological reason why detection probability would increase at those distances).

How many parameters?

- Models with too few parameters will not be flexible enough to describe the underlying relationship

- Adding parameters will improve the fit

- But models with too many parameters will be too flexible and will also describe the random noise in the data

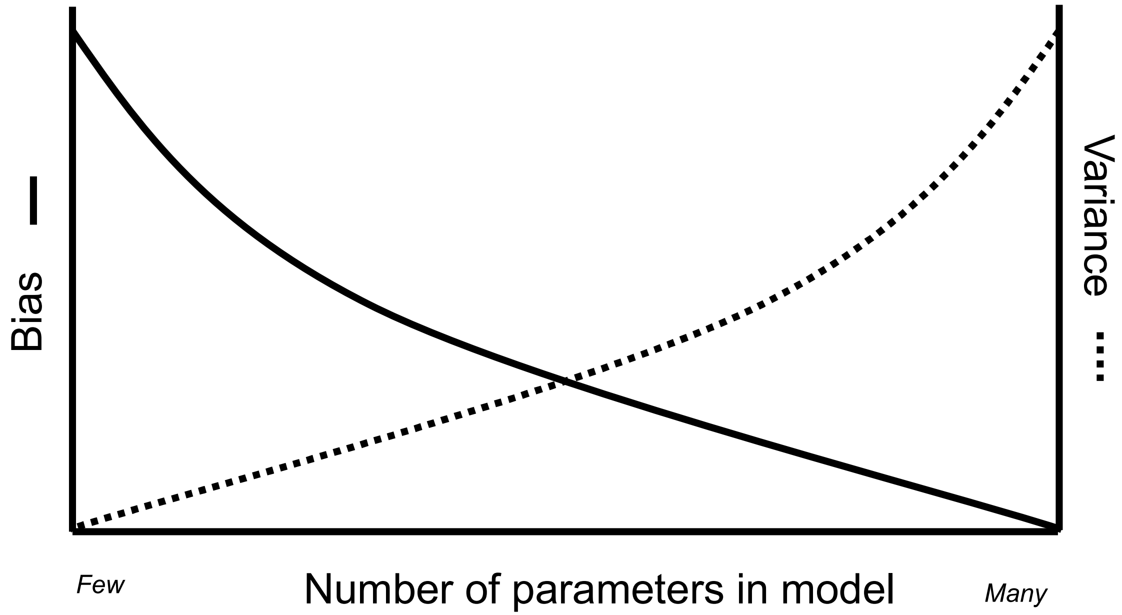
- We generally seek models with an intermediate number of parameters

How many parameters?

This problem can also be expressed as a trade-off between bias and variance

Models with too few parameters tend to produce estimates with low variance and high bias

Models with too many parameters tend to produce estimates with low bias and high variance (note the increasing CV for the estimate of P_a on the earlier slide)



CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Truncation

$$\hat{N} = \frac{nA}{2wL\hat{P}_a}$$

Need to choose the value of *w* (right truncation)

Detections at large distances contribute little to estimating the shape of g(x) at small distances (i.e. the shoulder) and may lead to poor fit and high variance

Typically, we might truncate around 5% of observation for line transects (perhaps nearer 10% for point transects)

Can also use estimated values of g(x) from fitted model as truncation criterion; truncate at *w* when *g(w)=0.15*

See supplement to Practical 2 regarding truncation

Three ways to think about detectability in distance sampling

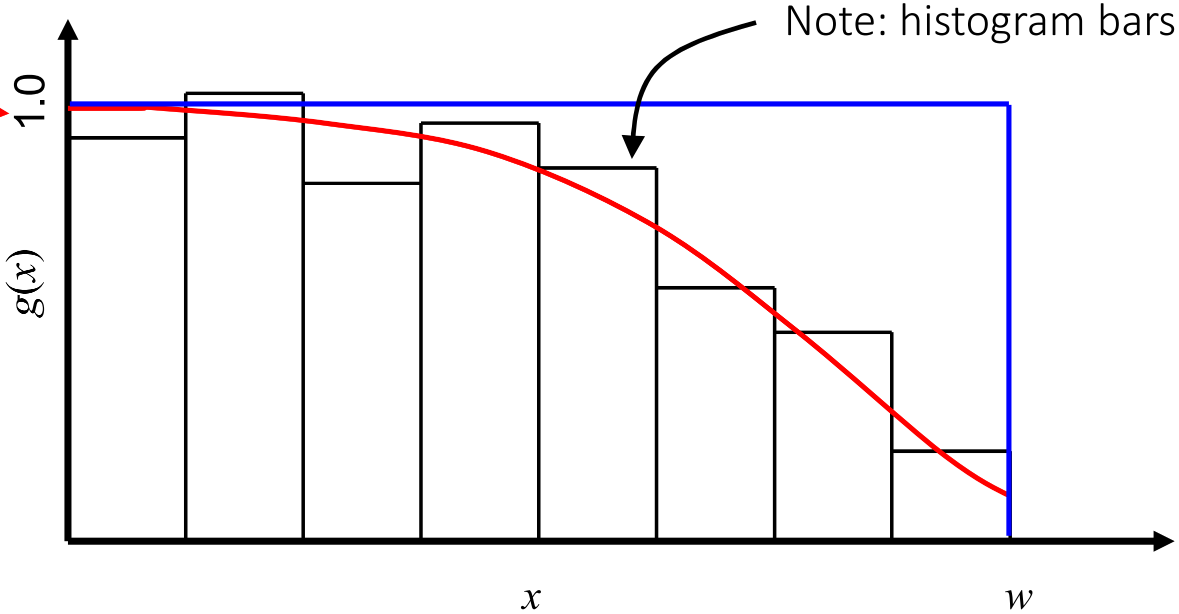
CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

1. 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 scaled

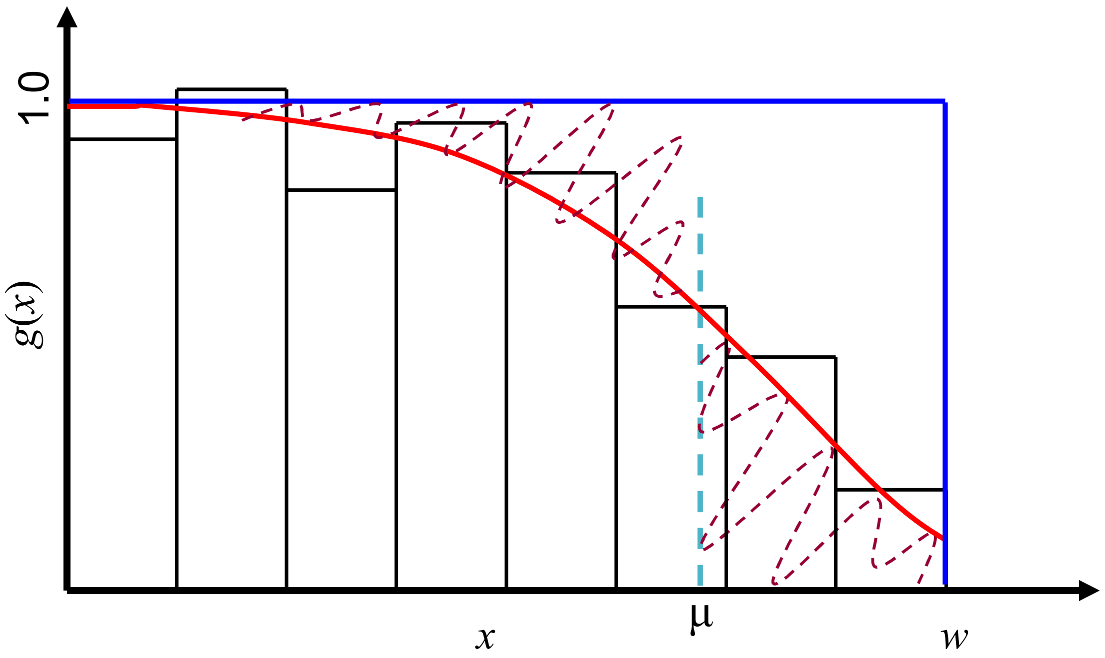
$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{\int_0^w \hat{g}(x)dx}{1 \times w}$$

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

2. Effective strip (half) width, μ

- Instead of a line transect out to w, where proportion P_a objects are seen, think of a strip transect out to some distance μ.



The ESW, μ, is the distance at which as many objects are detected beyond μ as are missed within μ

Line transect out to w

$$\hat{N} = \frac{nA}{\underbrace{2wL\hat{P}_a}_{\text{Area covered}}}$$

Strip transect out to μ

$$\hat{N} = \frac{nA}{\underbrace{2\hat{\mu}L}_{\text{Area effectively covered}}}$$

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

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

3. The probability density function, $f(x)$

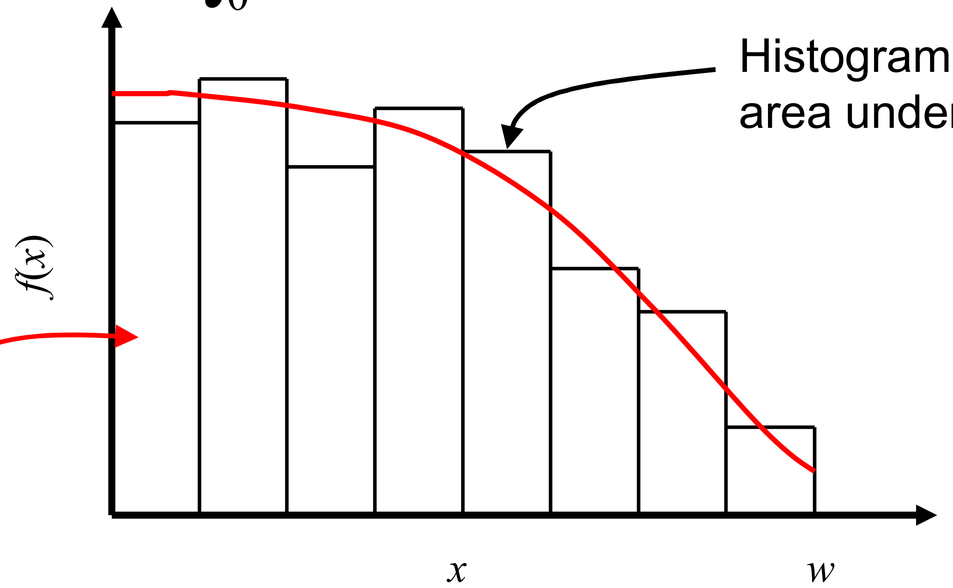
$f(x)dx$ = probability of observing an animal between distance x and $x+dx$, given it was observed somewhere in $(0,w)$

$f(x)$ is called the probability density function (pdf) of the observed distances

Because observations are between 0 and w, the area under $f(x)$ is 1.0

$$\int_0^w f(x)dx = 1$$

Area under $f(x)$ is 1



Histogram bars are scaled so that area under histogram is 1.

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Why is $f(x)$ useful?

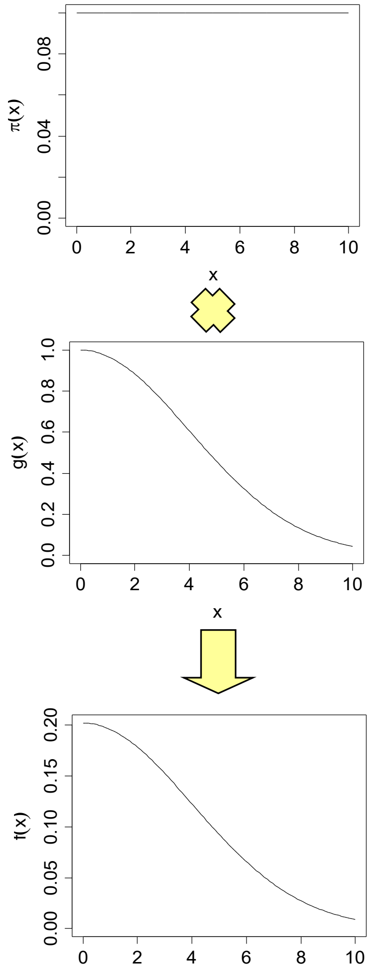
1. Useful for point transects, as it gives the expected distribution of detection distances

True distribution of animals

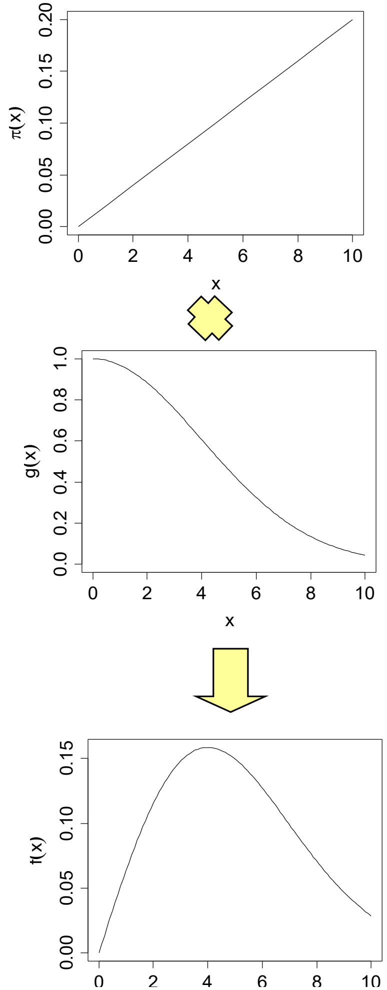
Detection function, $g(x)$

Observed distribution, $f(x)$

Line transect



Point transect



see lecture on point transects

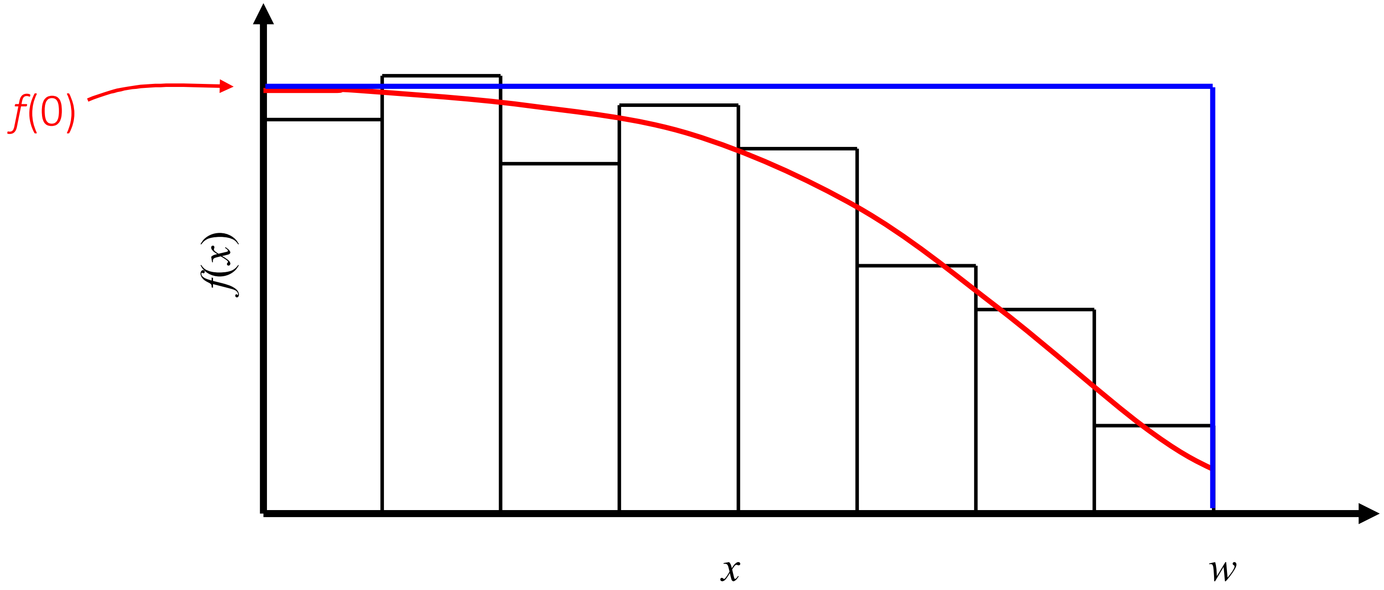
CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Why is $f(x)$ useful?

2. Gives another way to estimate P_a

   Lots of statistical machinery to fit pdfs, so this is the way `ds()` does it.



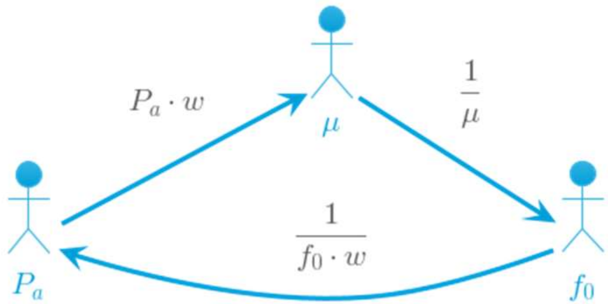
$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{1}{\hat{f}(0)w}$$

$$\hat{N} = \frac{nA}{2wL\hat{P}_a} = \frac{nA}{2wL\left(\frac{1}{\hat{f}(0)w}\right)} = \frac{nA\hat{f}(0)}{2L}$$

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Formulae – line transects



Three ways to think about line transects

1. Proportion seen or average probability of detection in covered region, P_a

$$\hat{N} = \frac{nA}{2wL\hat{P}_a} \qquad \hat{D} = \frac{n}{2wL\hat{P}_a}$$

2. Effective strip (half-)width, ESW, μ.

$$P_a = \mu / w$$

$$\hat{N} = \frac{nA}{2\hat{\mu}L} \qquad \hat{D} = \frac{n}{2\hat{\mu}L}$$

3. Pdf of observed distances, $f(x)$, evaluated at 0 distance

$$f(0) = 1/\mu$$

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

CREEM Centre for Research into Ecological and Environmental Modelling

Note where the "hats" are found on the right hand side of equations

University of St Andrews

Notation – line transects

Known constants and data:

k = number of lines

l_j = length of j^{th} line, j=1,...,k

$L = \Sigma l_j$ = total line length

n = number of animals or clusters detected

x_i = distance of i^{th} detected animal or cluster from the line, i=1,...,n

w = truncation distance for x

A = size of region of interest

a = area of "covered" region = $2wL$

s_i = size of i^{th} detected cluster, i=1,...,n

CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

Notation – line transects

Parameters and functions:

N = population size / abundance of animals

N_s = abundance of clusters

D = density = animals per unit area = N/A

D_s = density of clusters

$g(x)$ = detection function

$f(x)$ = probability density function (pdf) of observed distances

$f(0)$ = $f(x)$ evaluated at 0 distance

μ = effective strip (half-)width

P_a = probability of detecting an animal or cluster given it is in the covered area a

$E(s)$ = mean size of clusters in the population

CREEM Centre for Research into Ecological and Environmental Modelling

University of St Andrews