

# Choosing a Detection function

# Overview of detection function modelling

Formal definition

Criteria for a good detection function model

Key functions and adjustment terms

Choosing the number of parameters

Comments about truncation

## Formal definition

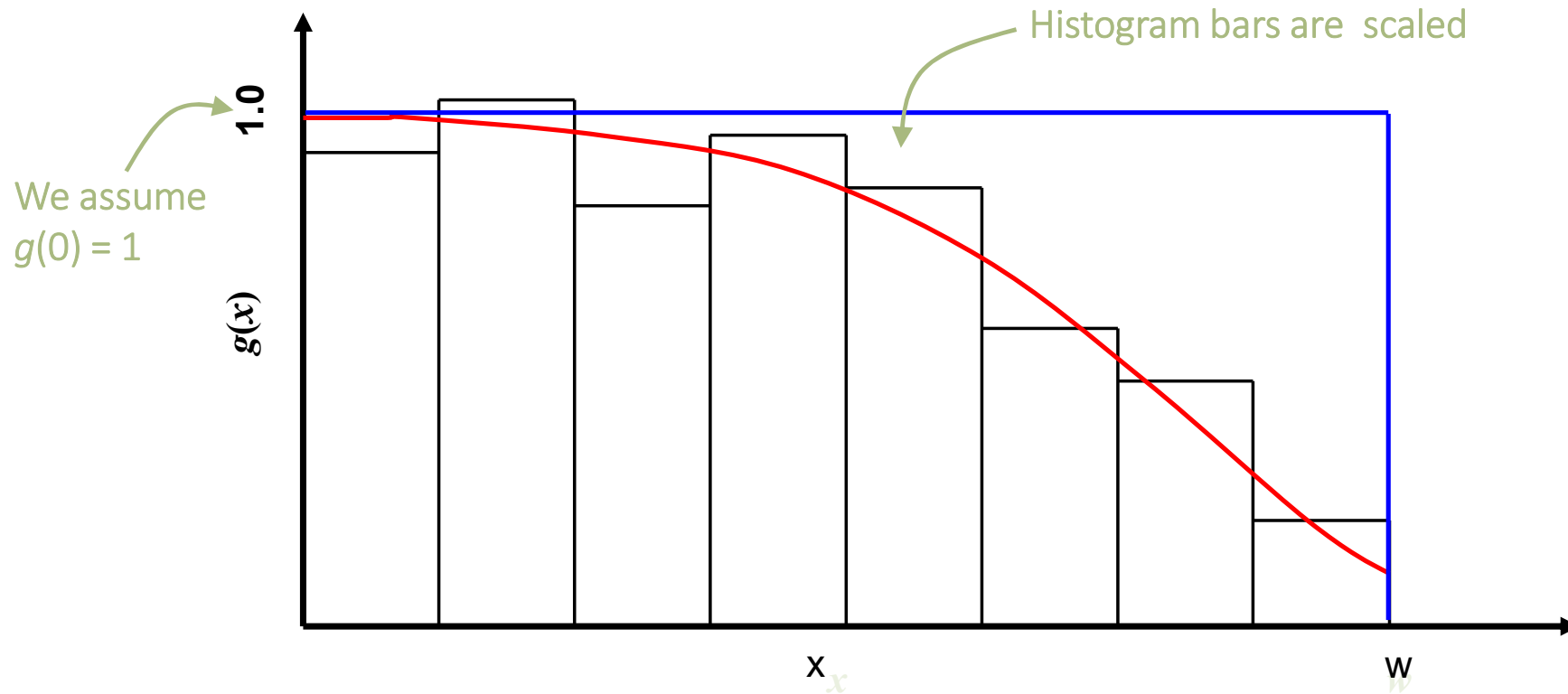
The **detection function** describes the relationship between distance and the probability of detection

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, $g(x)$



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

## 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

# Key functions

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

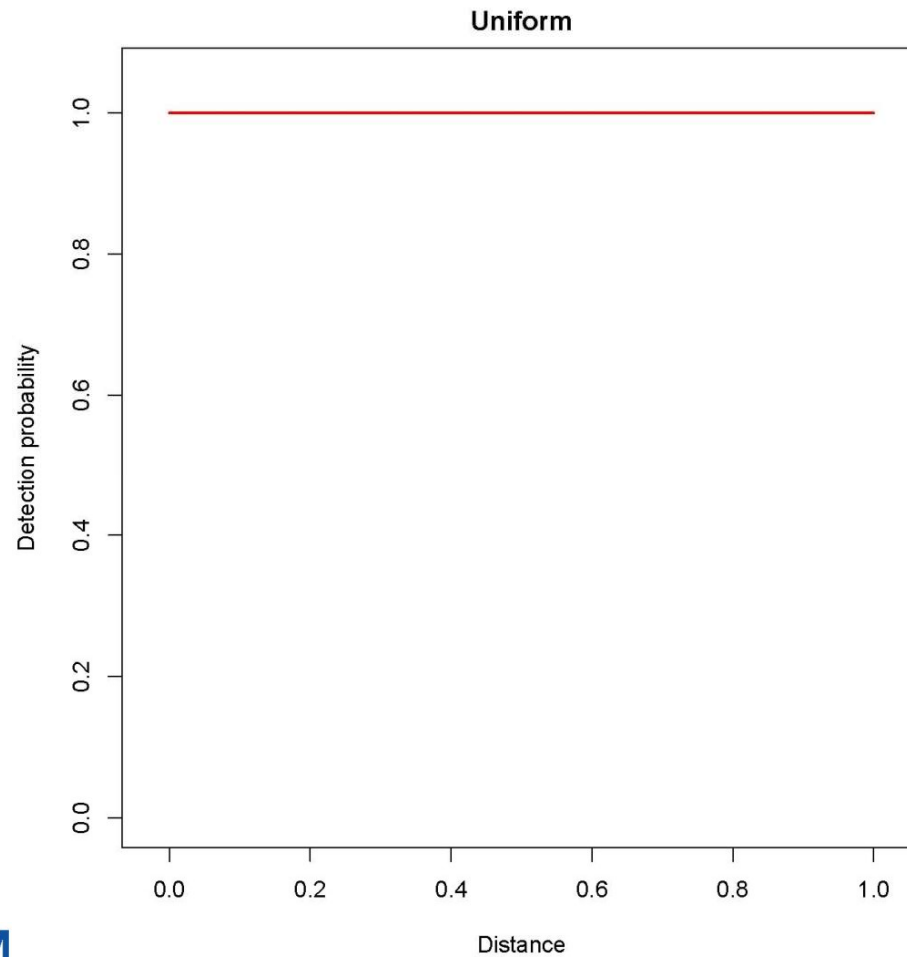
This determines the basic model shape

Three key functions available in distance sampling software:

- Uniform
- Half normal
- Hazard rate



# Key functions (cont.)



- Model formula:

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

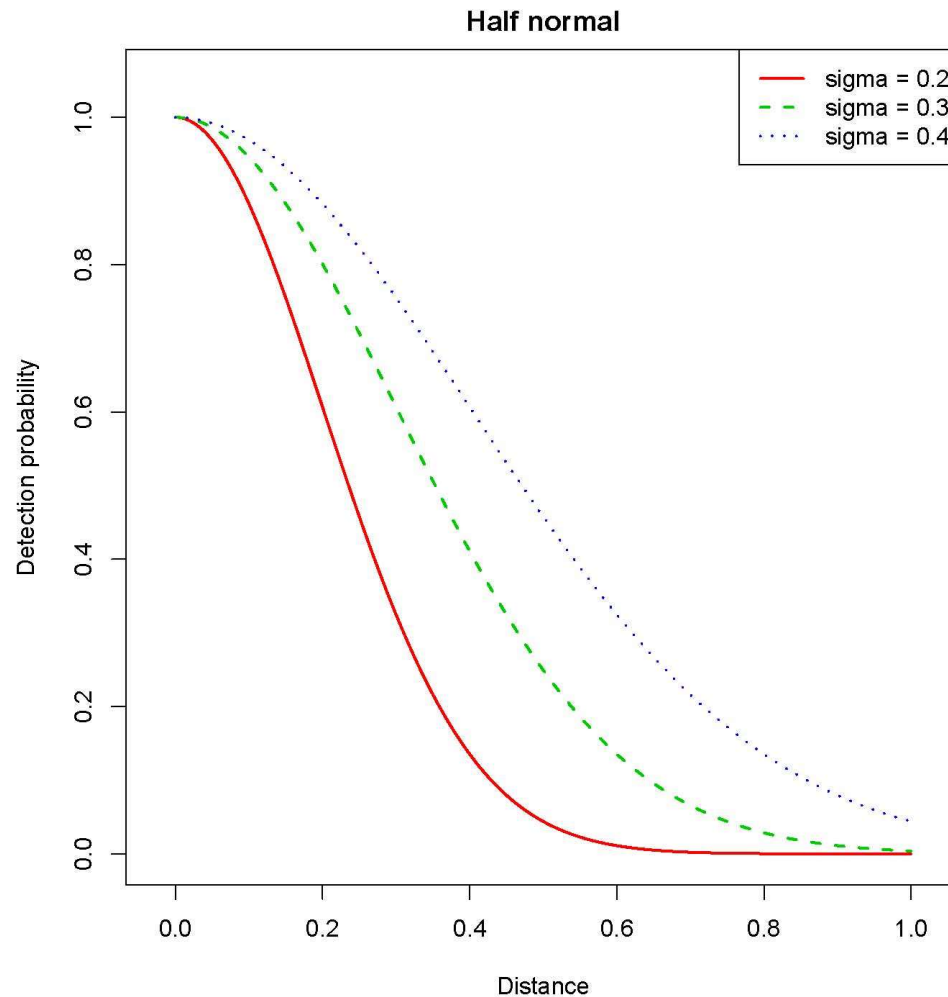
- Parameters = 0
- Shape criterion?

Yes

- Model robust?

No

## Key functions (cont.)



- Model formula:

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

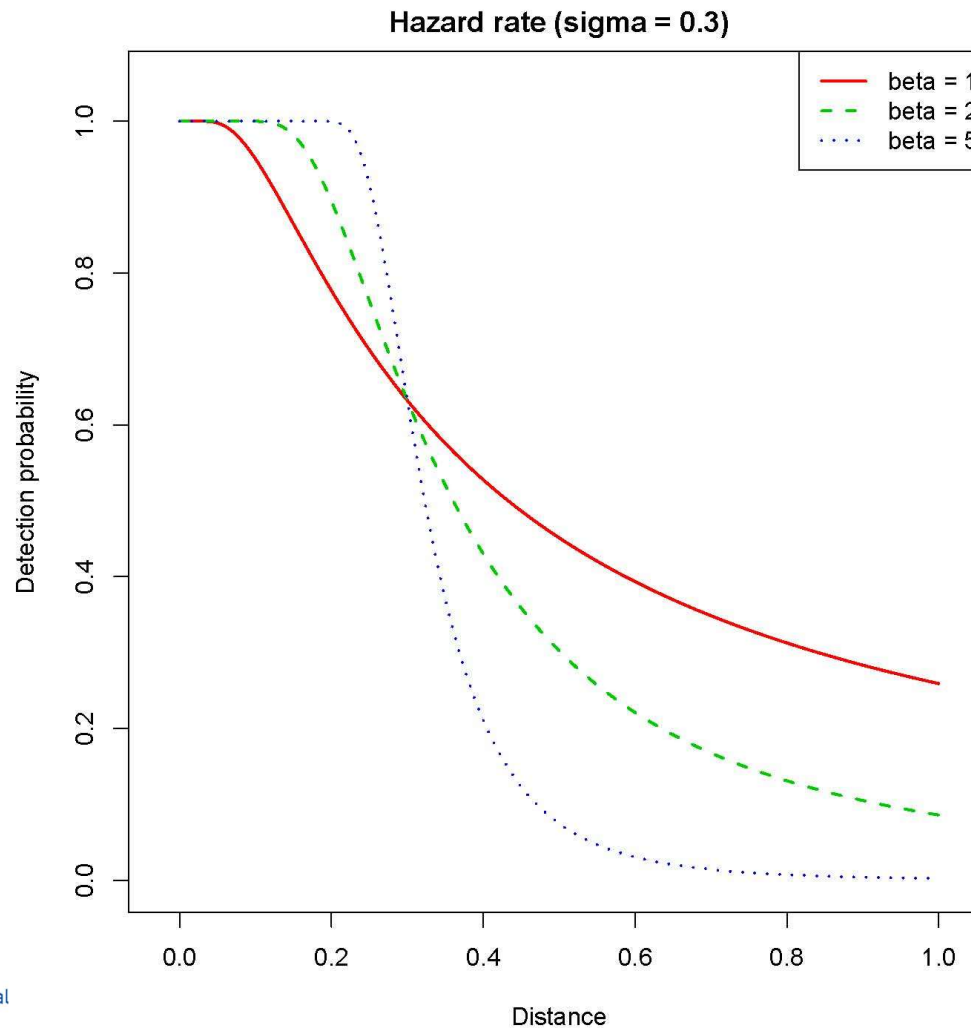
- Parameters = 1
- Shape criterion?

Yes

- Model robust?

Somewhat

## 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

# Adjustment terms

# 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  $\times$  (1 + Series)

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

The  $\alpha_i$  parameters must be estimated

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

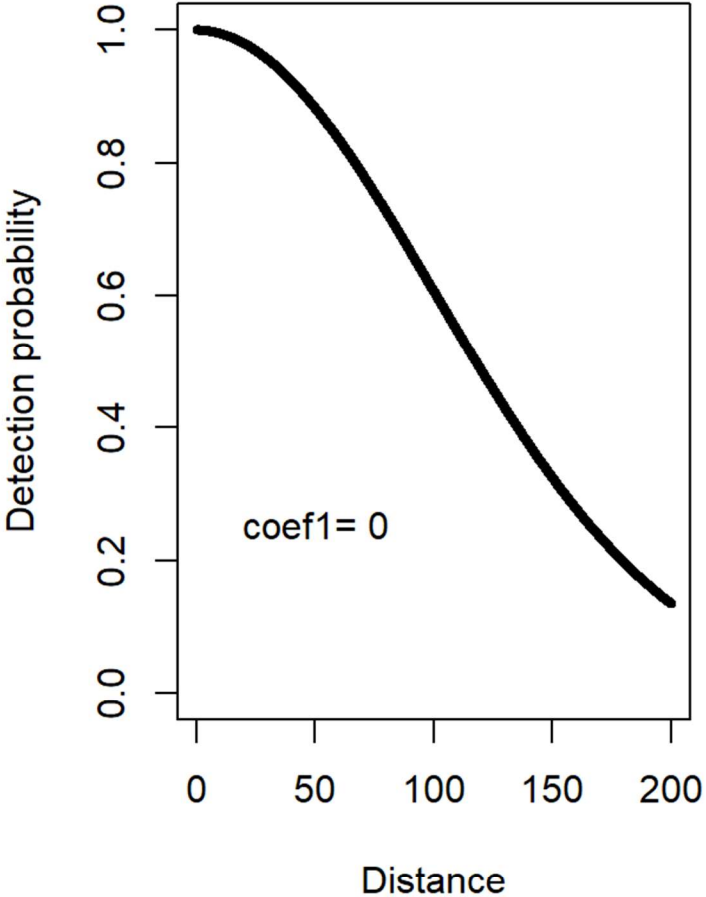
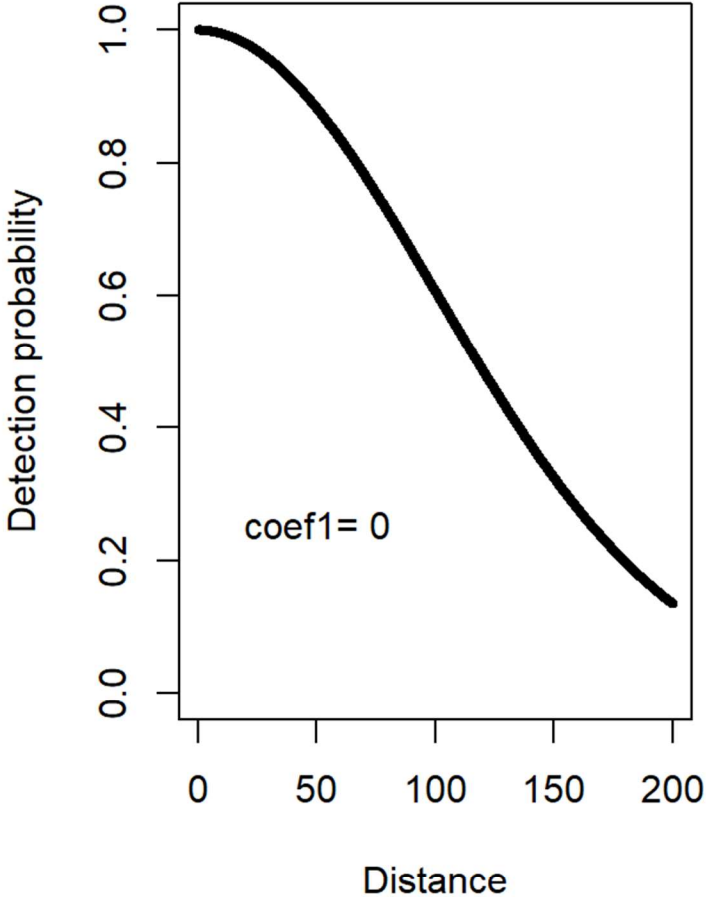
The number of adjustment terms needs to be chosen

# Adjustment terms

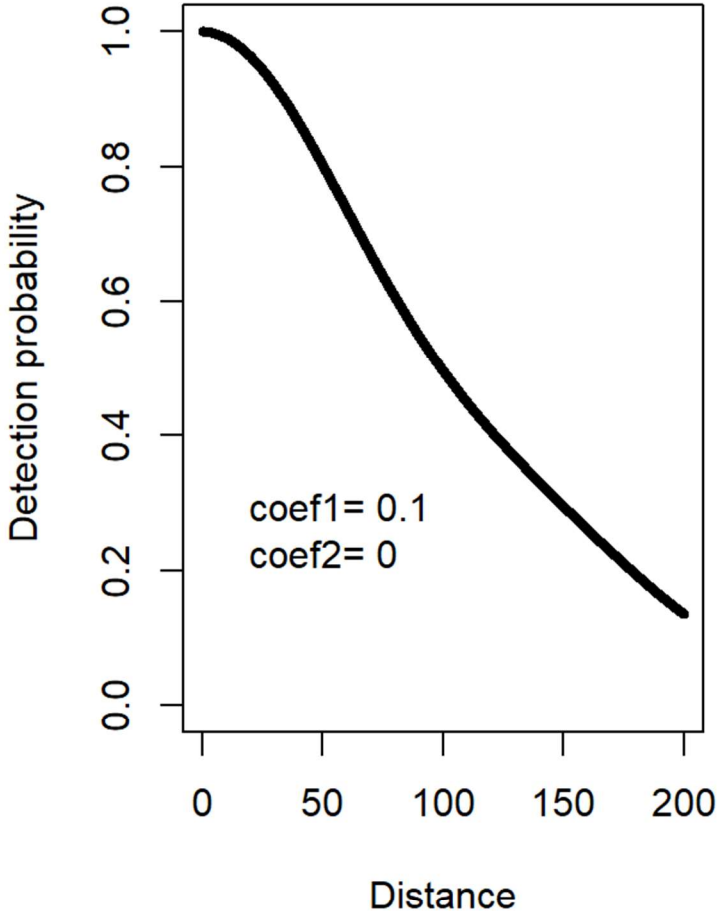
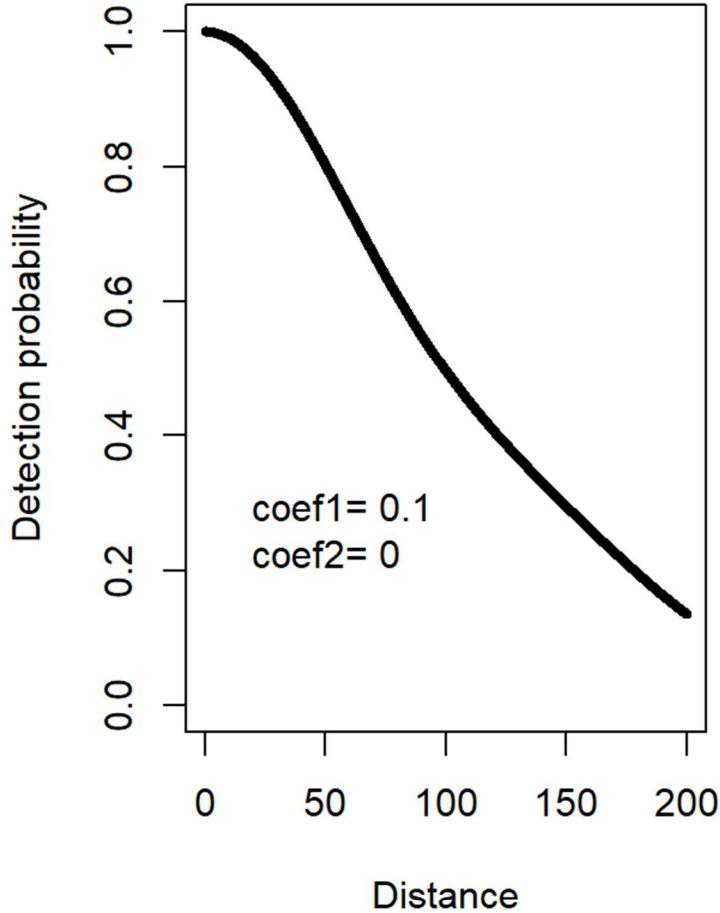
Distance allows the selection of three types of series (one type per model)

Key function	Series adjustment
Uniform*	Cosine*
Half normal <sup>†</sup>	Hermite polynomial <sup>†</sup>
Hazard rate	Simple polynomial

# Half normal key, single cosine adjustment term

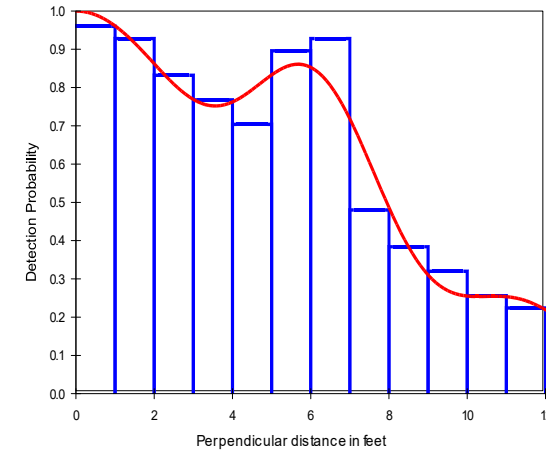
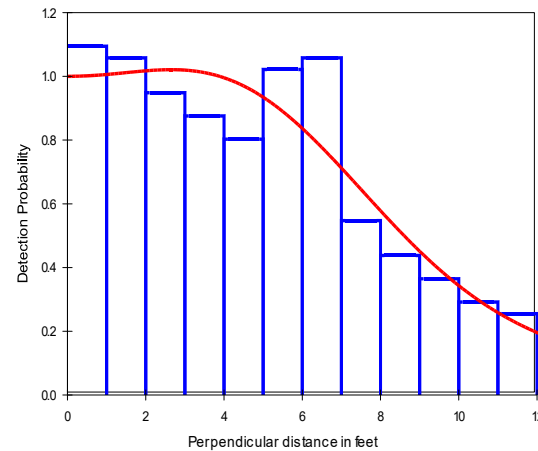
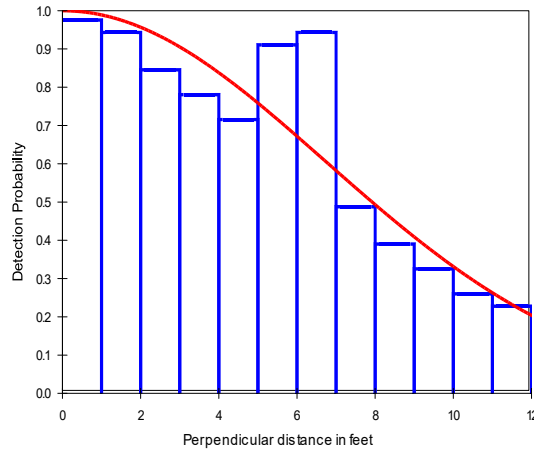


# Half normal key, two cosine adjustment terms





# 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).

# Bias vs variance tradeoff

# How many parameters?

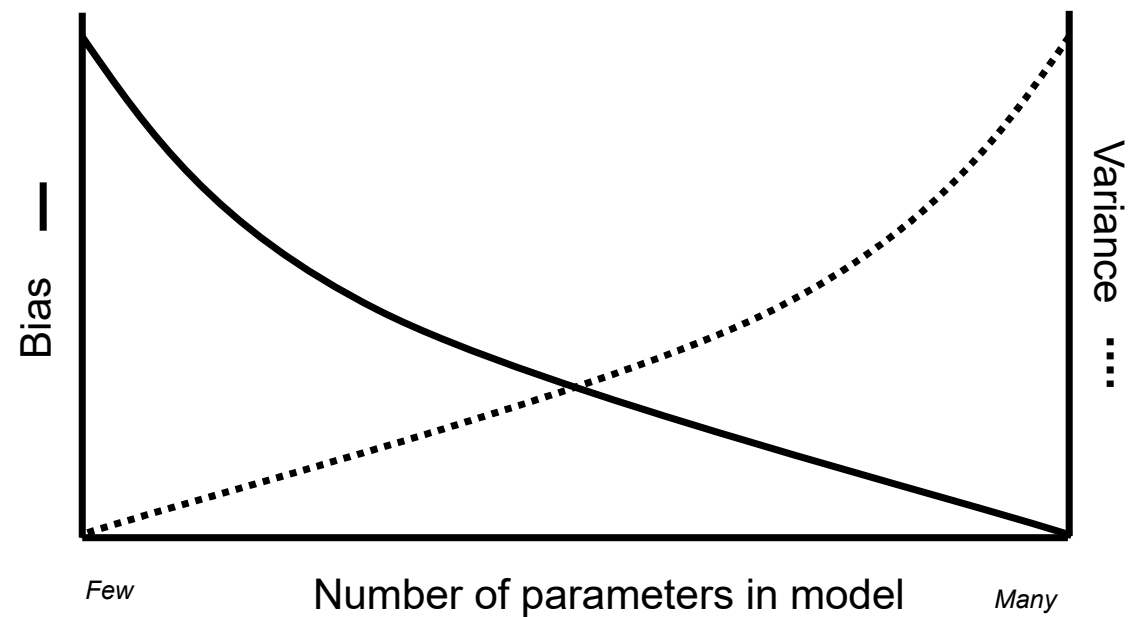
- Models with too few parameters will not be flexible enough to describe the underlying relationship
- Adding parameters will improve the fit
- But models with too many parameters will be too flexible and will also describe the **random noise** in the data 
- We generally seek models with an intermediate 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)



# Data truncation prior to model fitting

# 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

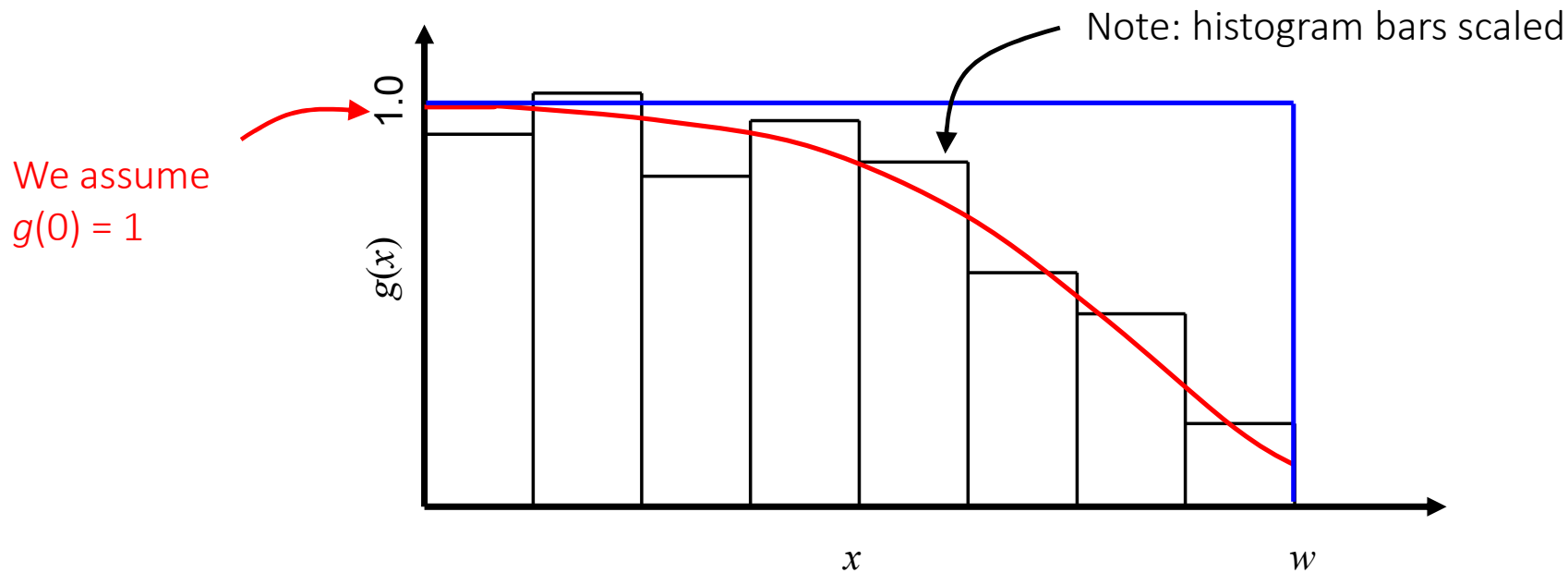
Typically, we might truncate around 5% of observation for line transects (perhaps nearer 10% for point transects)

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

# Alternative derivations for understanding detectability

# 1. The detection function, $g(x)$

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

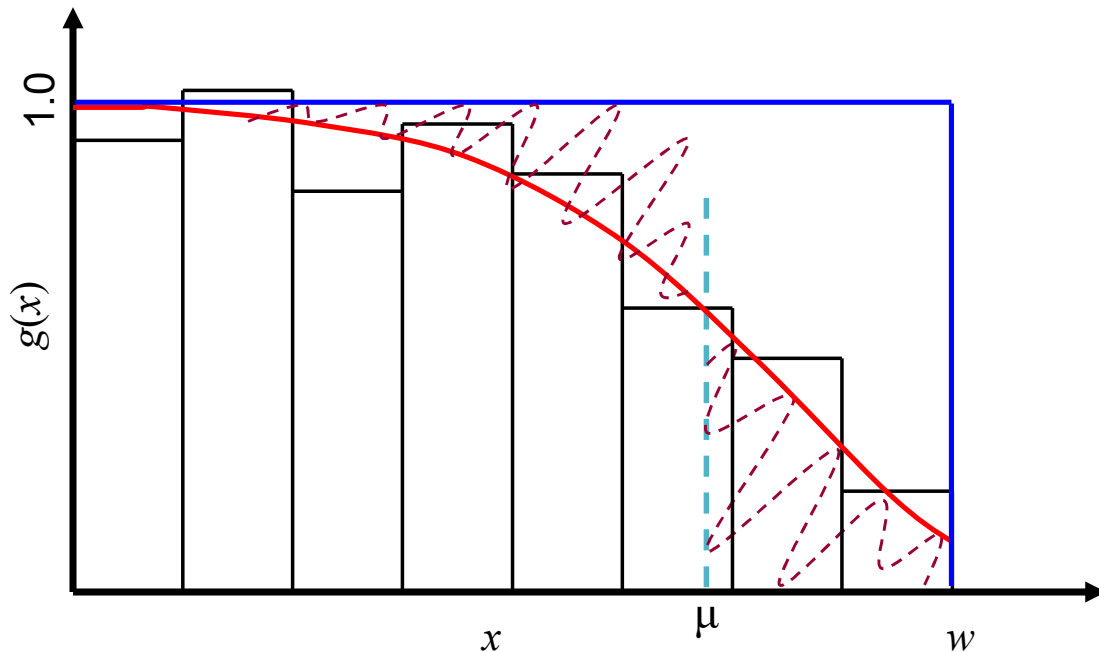


$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{\int_0^w \hat{g}(x) dx}{1 \times w}$$



## 2. Effective strip (half) width, $\mu$

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



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

Line transect out to  $w$

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

Strip transect out to  $\mu$

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

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

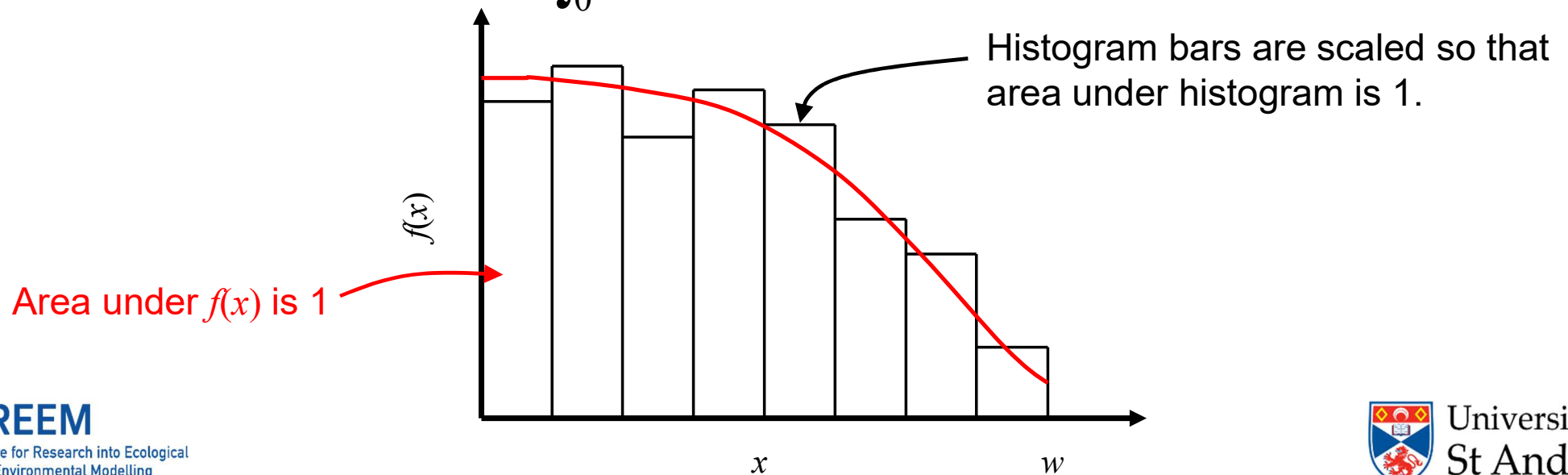
### 3. The probability density function, $f(x)$

$f(x)dx$  = 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

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

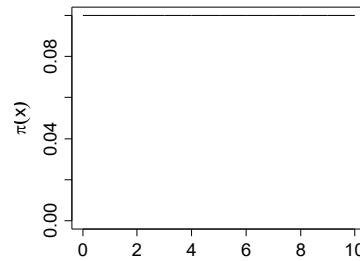


# Why is $f(x)$ useful?

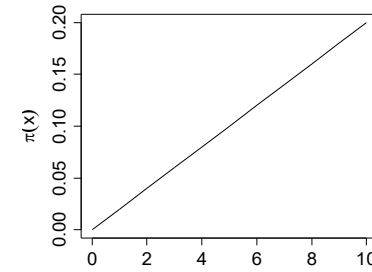
1. Useful for point transects, as it gives the expected distribution of detection distances

True distribution of animals

Line transect

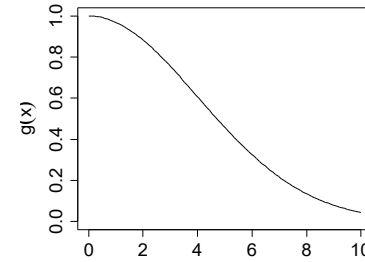
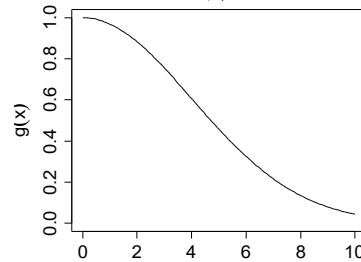


Point transect

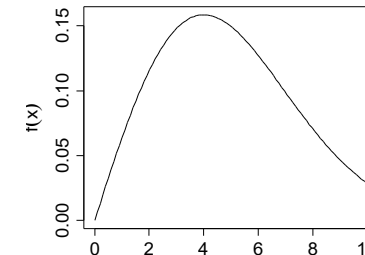
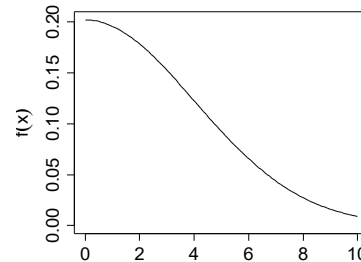


see lecture on point transects

Detection function,  $g(x)$

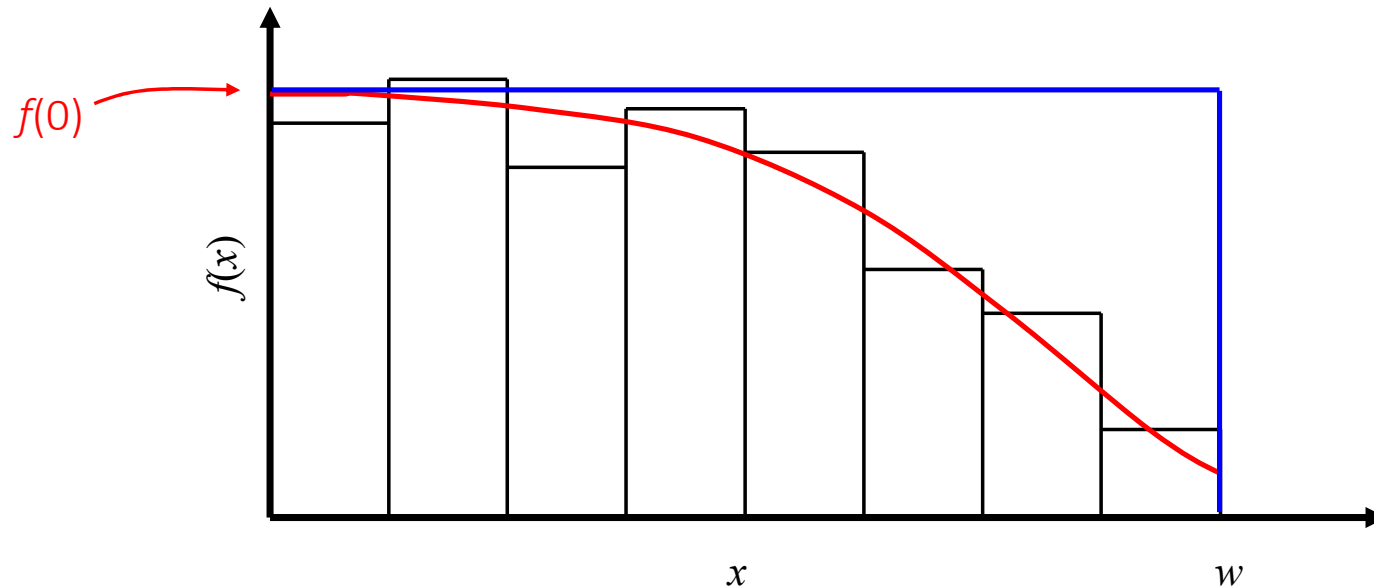


Observed distribution,  $f(x)$



# Why is $f(x)$ useful?

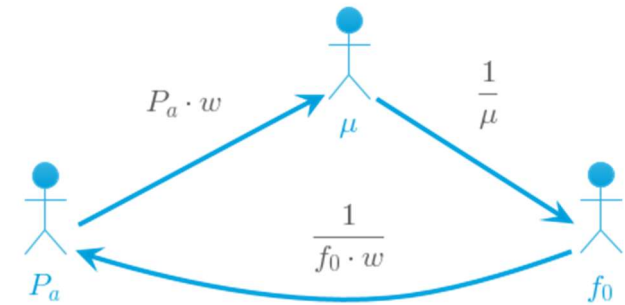
2. Gives another way to estimate  $P_a$   
 Lots of statistical machinery to fit pdfs



$$\hat{P}_a = \frac{\text{area under curve}}{\text{area under rectangle}} = \frac{1}{\hat{f}(0)w} \quad \hat{N} = \frac{nA}{2wL\hat{P}_a} = \frac{nA}{2wL\left(\frac{1}{\hat{f}(0)w}\right)} = \frac{nA\hat{f}(0)}{2L}$$

## Formulae – line transects

Three ways to think about line transects



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

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

2. Effective strip (half-)width, ESW,  $\mu$ .

$$P_a = \frac{\mu}{w}$$

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

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

$$f(0) = \frac{1}{\mu}$$

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

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

# Notation (knowns and unknowns)

## Notation – line transects

Known constants and data:

$k$  = number of lines

$l_j$  = length of  $j^{\text{th}}$  line,  $j=1, \dots, k$

$L = \sum l_j$  = total line length

$n$  = number of animals or clusters detected

$x_i$  = distance of  $i^{\text{th}}$  detected animal or cluster from the line,  $i=1, \dots, n$

$w$  = truncation distance for  $x$

$A$  = size of region of interest

$a$  = area of “covered” region =  $2wL$

$s_i$  = size of  $i^{\text{th}}$  detected cluster,  $i=1, \dots, n$

## 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

$\mu$  = 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