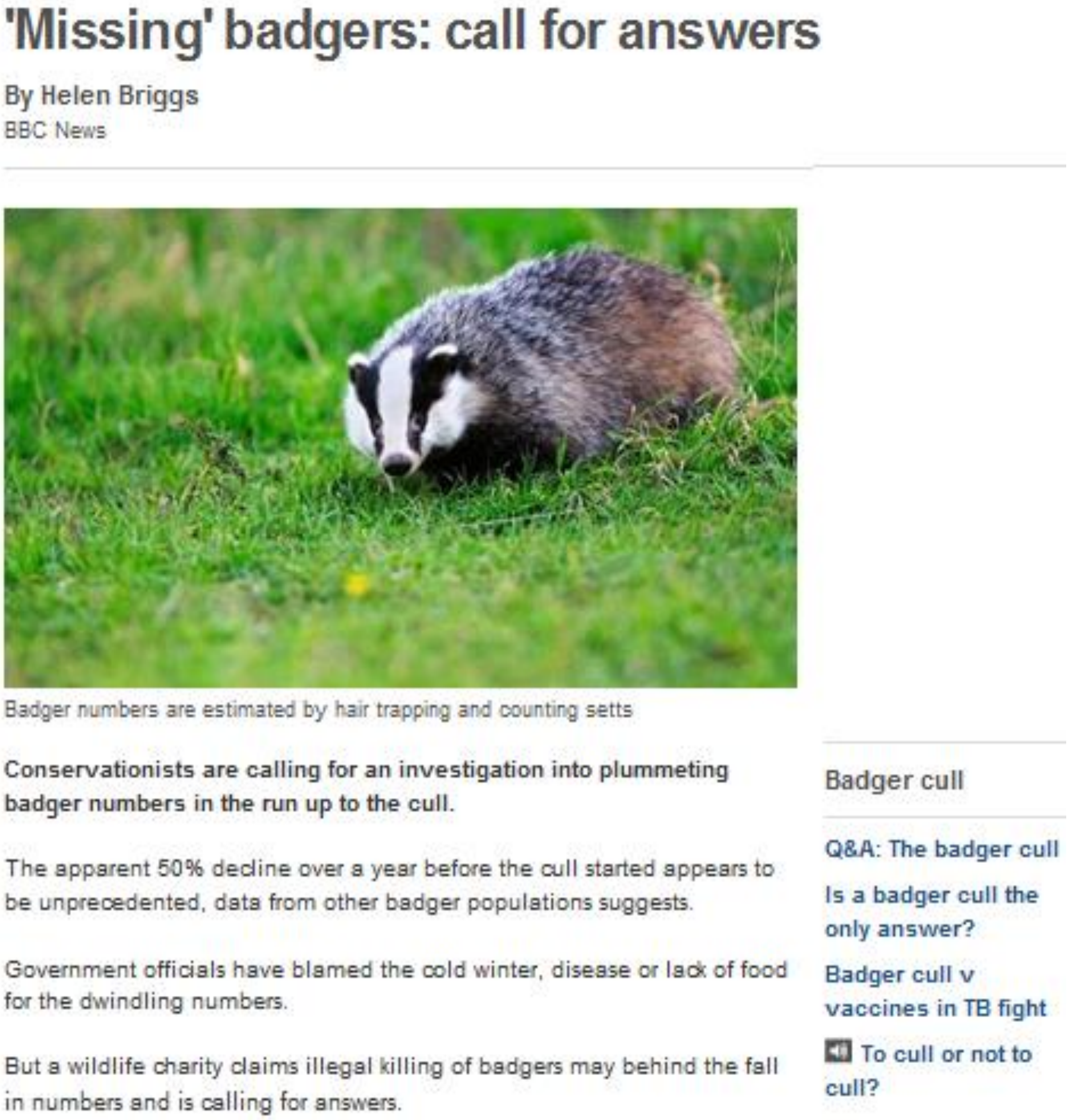
Mark-recapture with identification errors

Richard Vale

16/07/14



'Missing' badgers: call for answers

By Helen Briggs
BBC News

Badger numbers are estimated by hair trapping and counting setts

Conservationists are calling for an investigation into plummeting badger numbers in the run up to the cull.

The apparent 50% decline over a year before the cull started appears to be unprecedented, data from other badger populations suggests.

Government officials have blamed the cold winter, disease or lack of food for the dwindling numbers.

But a wildlife charity claims illegal killing of badgers may behind the fall in numbers and is calling for answers.

Badger cull

Q&A: The badger cull

Is a badger cull the only answer?

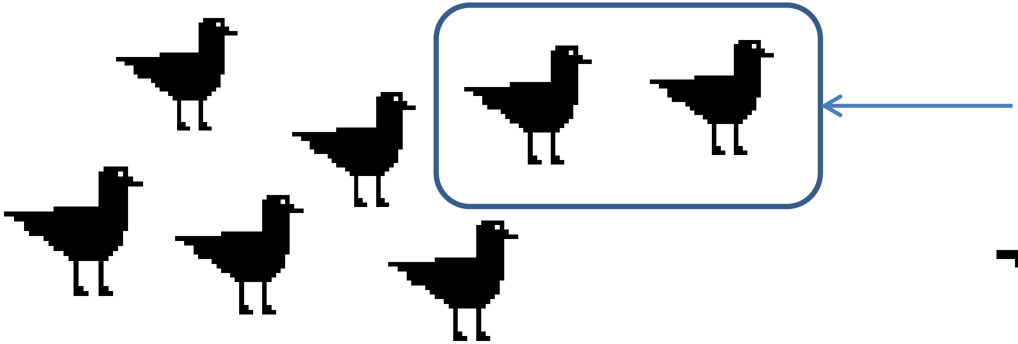
Badger cull v vaccines in TB fight

To cull or not to cull?

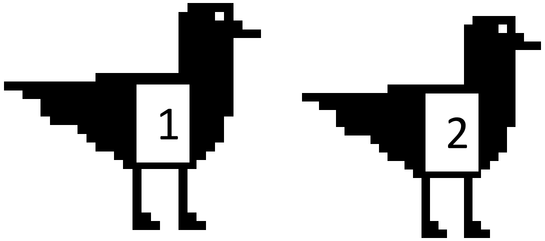
Hard to estimate the size of an animal population.

One popular method: mark-recapture sampling

A population

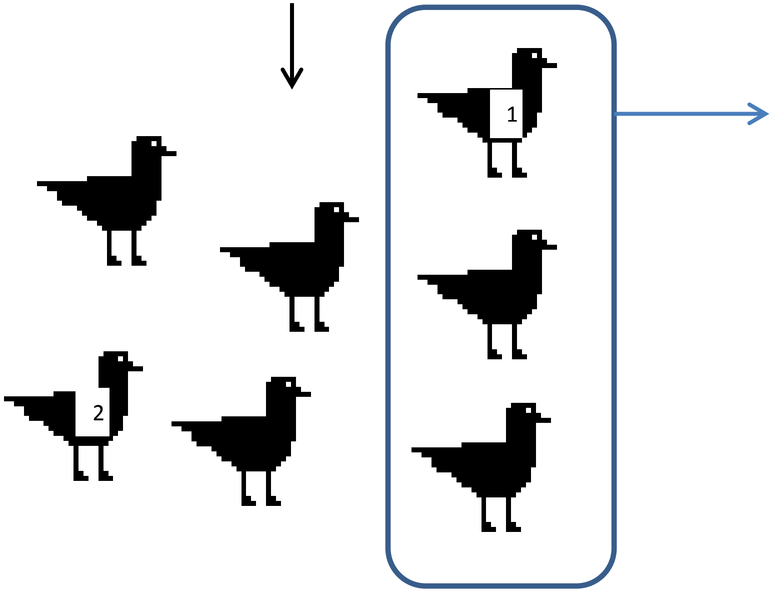


A sample, taken by a biologist



Sampled animals are identified with a mark.

Sampled animals are returned to the population and mingle with the others.



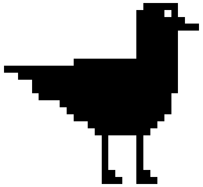
Later, another sample is taken. Some of the marked animals may be recaptured, while others may not be.

The process of capturing animals, marking them and releasing them is repeated several times …

Here is how the data might look. Notice that some individuals were never captured.

Aim: estimate the unknown population size N.

If captures are assumed to be independent Bernoulli trials with constant capture probability p, then we can write down the likelihood function and maximise it.

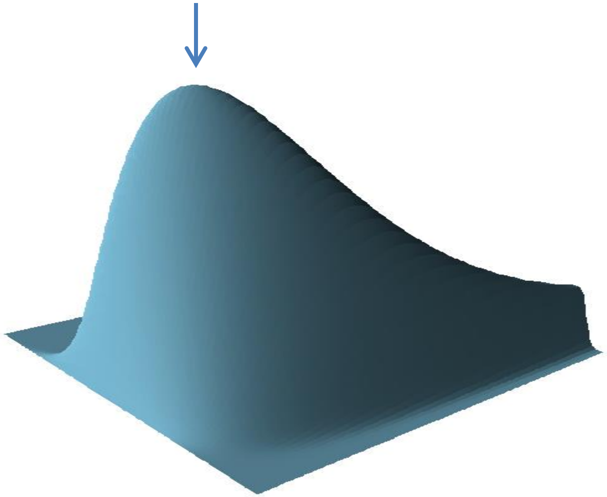


1	1	1	0
2	1	0	1
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1

Individual (mark)

Indicators: 1 if the individual was captured on that particular occasion and 0 otherwise.

Maximum at $N = 8.016, \;\; p = 0.33$



Usually capture probabilities are assumed to depend on capture occasion *t*. Then the model is called M_t.

Consider animal *i*. Define:

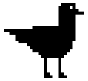
$$\omega_{it} = \begin{cases} 1 & \text{if captured at time } t \\ 0 & \text{otherwise} \end{cases}$$

Then the probability of a particular capture history is:

$$P(\omega_{i1} \ldots \omega_{it}) = \prod_{t=1}^{T} p_t^{\omega_{it}} (1 - p_t)^{1-\omega_{it}}$$

For example, in the example from the previous page:

$$P(110) = p_1 p_2 (1 - p_3)$$



1	1 1 0
2	1 0 1
3	0 1 0
4	0 1 0
5	0 0 1
6	0 0 1

Summing over all ways of assigning captures to animals, the likelihood for model M_t:

$$\propto \frac{N!}{(N - N_{obs})!} \prod_{i=1}^{N} \prod_{t=1}^{T} p_t^{\omega_{it}} (1 - p_t)^{1-\omega_{it}}$$

$$= \frac{N!}{(N - N_{obs})!} \prod_{t=1}^{T} p_t^{n_t} (1 - p_t)^{N-n_t}$$

where N_{obs} is the number observed and n_t is the number captured at time *t. (*The important point here is that the likelihood only depends on these quantities.)

Remarks:

1. Is this a sensible way to treat the p_t? Are the capture occasions really independent?
2. Can show that likelihood has no global maximum in case there are no repeated captures.

A complication: it is difficult to label some kinds of animals.

Alternatives:

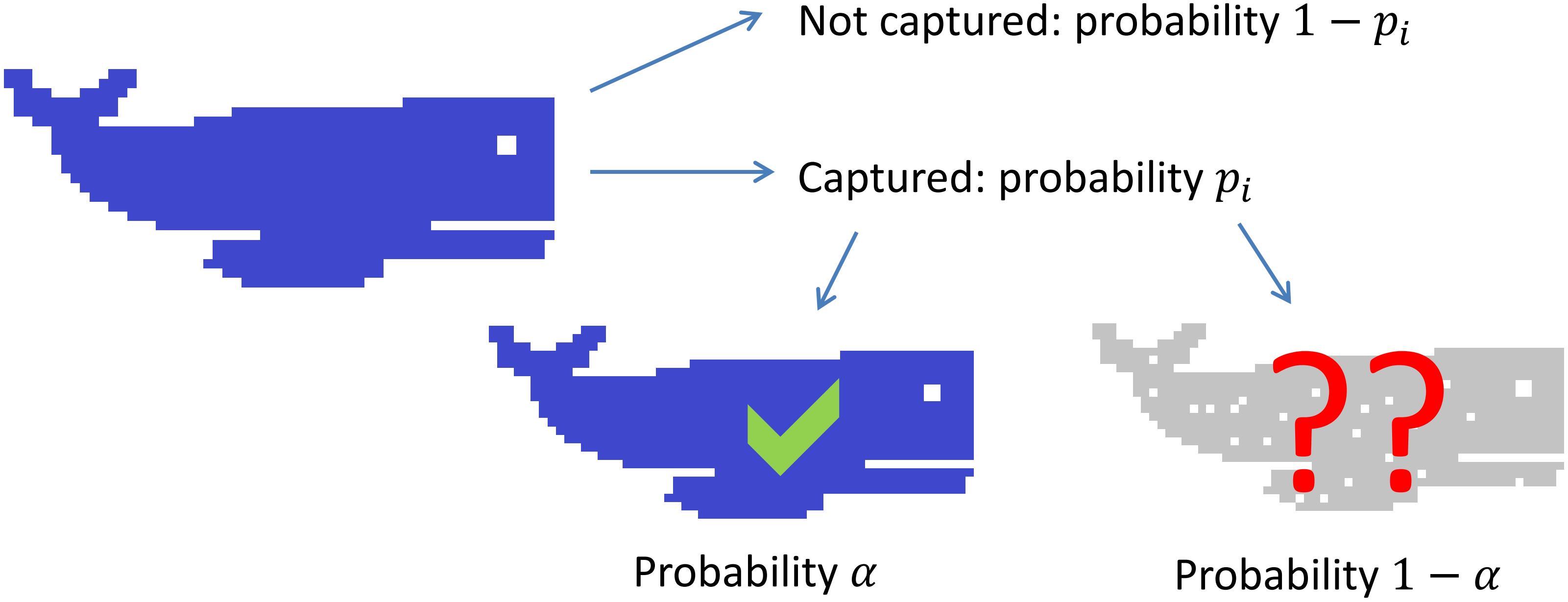
- Genetic samples (from hair, faeces etc.)
- Human observers
- Photographs

These methods might lead to misidentification errors.

Model $M_{t,\alpha}$

This version of the mark-recapture model was invented by Lukacs/Burnham (2005) and Yoshizaki et al. (2011).

- probability p_i of an animal being captured at time i.
- Captured animals are correctly identified with a fixed probability α.
- A misidentified animal produces a ghost record which is seen only once.

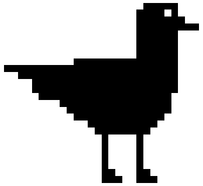


Not captured: probability $1 - p_i$

Captured: probability p_i

Probability α

Probability $1 - \alpha$



1	1	1	0
2	1	0	1
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1

Possible ghost records

Under model $M_{t,\alpha}$, there could be as few as three individuals in the population. ($\widehat{N} = 3$, $\hat{\alpha} = 0.6$)

$$\omega_{it} = \begin{cases} 2 & \text{if seen at time } t \text{ and misidentified} \\ 1 & \text{if seen at time } t \text{ and correctly identified} \\ 0 & \text{otherwise} \end{cases}$$

(Of course, we don't observe ω_{it})

$$P(\omega_{i1} \ldots \omega_{it}) = \prod_{t=1}^{T} (\alpha p_t)^{\omega_{it}=1} (1-p_t)^{\omega_{it}=0} (1-\alpha)^{\omega_{it}=2}$$

The likelihood is the sum over all possible assignments of capture histories to the animals which give the observed data:

$$= \sum \prod_{i=1}^{N} \prod_{t=1}^{T} p_t^{\omega_{it}=1} (1-p_t)^{\omega_{it}=0} \alpha^{\omega_{it}=1} (1-\alpha)^{\omega_{it}=2}$$

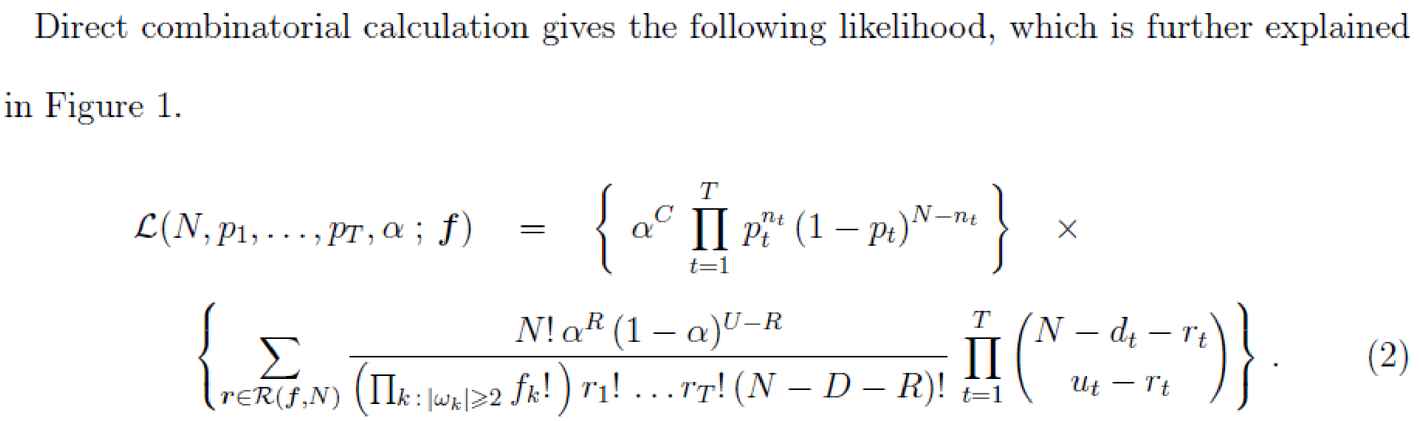
$$= \sum \prod_{t=1}^{T} p_t^{n_t} (1-p_t)^{N-n_t} \alpha^{C-G} (1-\alpha)^{G}$$

C = total number of captures,

G = number of ghosts

To compute the likelihood, you just need to **count** how many ways you can get the observed data with exactly *G* ghosts, for any given *G.*

This is a purely combinatorial problem; there is no probability involved!



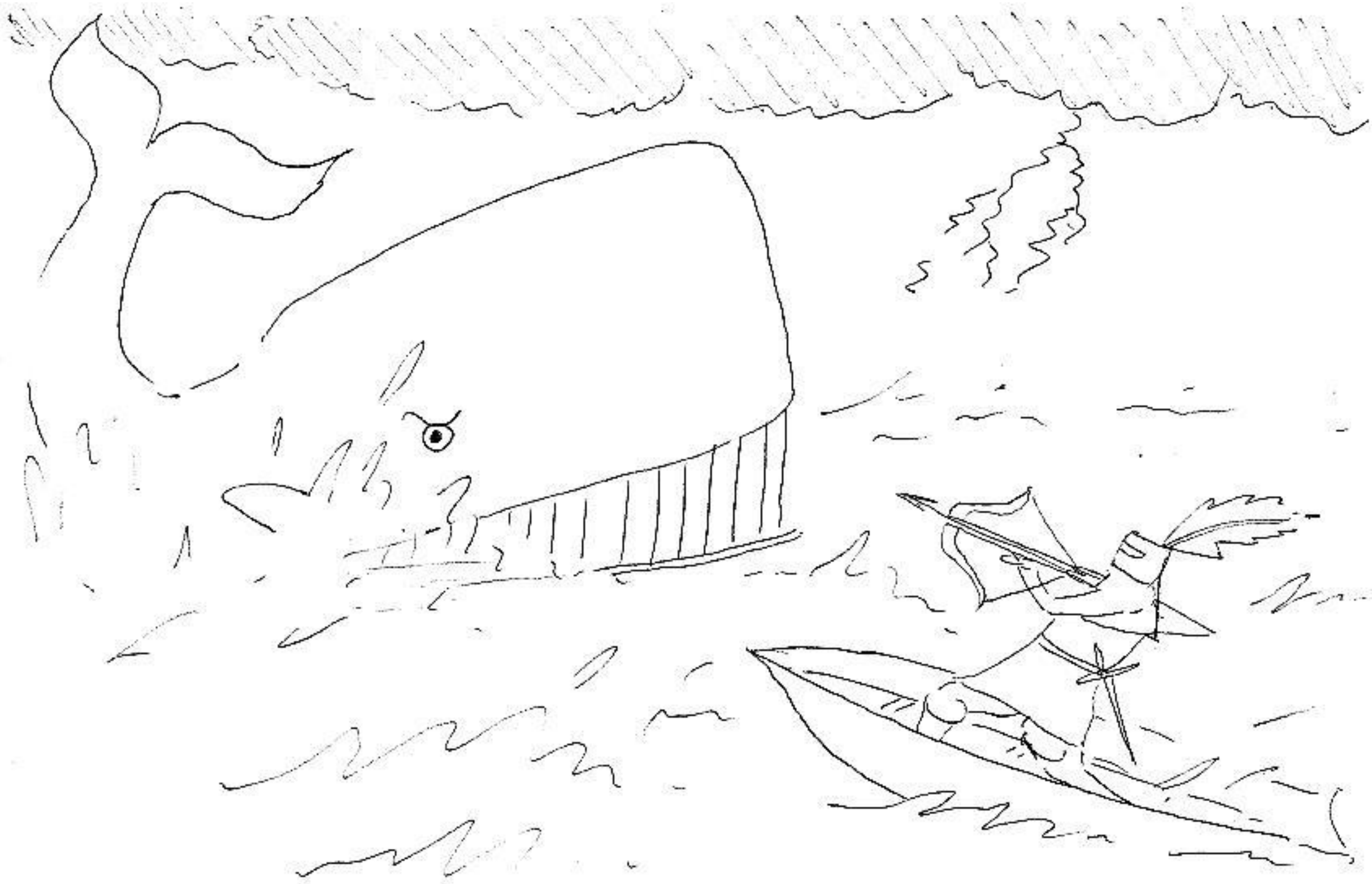
Direct combinatorial calculation gives the following likelihood, which is further explained in Figure 1.

$$\mathcal{L}(N, p_1, \ldots, p_T, \alpha\,;\, \boldsymbol{f}) \quad = \quad \left\{ \alpha^C \prod_{t=1}^{T} p_t^{n_t} (1-p_t)^{N-n_t} \right\} \quad \times$$

$$\left\{ \sum_{\boldsymbol{r}\in\mathcal{R}(\boldsymbol{f},N)} \frac{N!\,\alpha^R (1-\alpha)^{U-R}}{\left(\prod_{k\,:\,|\omega_k|\geqslant 2} f_k!\right) r_1! \ldots r_T!\,(N-D-R)!} \prod_{t=1}^{T} \binom{N-d_t-r_t}{u_t-r_t} \right\} . \tag{2}$$

The model can already be fitted by Bayesian methods (Link et al. 2010), so what is the advantage of having a new expression for the likelihood? The main advantage is that the model can be fitted much more quickly, and simulation studies can be carried out on a large scale.

“We apply method $M_{t,\alpha}$ to genetic and photographic surveys of the endangered New Zealand population of the southern right whale (Eubaleana australis). Genetic samples were collected in the T=4 austral winters of 1995-1998 using small biopsy darts deployed from a crossbow.”



- Used R, ADMB to calculate the likelihood function and maximise it. (Note: actually implementing the formula is not trivial as a naïve approach runs out of memory.)

- Simulation studies verify that the likelihood is correct (can also do explicit calculations in small cases.)

- The whales were identified using genetic markers. It is assumed that some errors have occurred. These can either be corrected by pre-processing the data or by applying model $M_{t,\alpha}$. Note that there are almost no recaptures if two whales which had any difference at all are recorded as different captures (recall that this causes model M_t to break down.)

- Results:

95% confidence interval for $\widehat{N}$: $(49, 419)$. Conclusion: there are some whales.

Our simulation studies show that parameter estimates tend to be biased unless the sample sizes and capture probabilities are unrealistically large. Other authors have also had trouble applying the model to real data (or neglected real data altogether). It seems that without strongly informative priors, $M_{t,\alpha}$ gives very large error estimates.

Research directions:

- Pin down why the model fails by analysing simpler models in more detail
- Develop better approaches for photographic studies

References:

For this talk:

- Vale, Fewster, Carroll, Patenaude, *Biometrics*, 2014

For the model:

- Lukacs, Burnham*, J.Wildlife Management*, 2005.
- Link, Yoshizaki, Bailey, Pollock, *Biometrics*, 2010
- Link, Barker (book) 2009
- Yoshizaki, Brownie, Pollock, Link, *Environmental and Ecological Statistics*, 2011.

For the data:

- Carroll et al, *Marine Biology*, 2011
- Carroll et al, *Ecological Applications* 2013.