Modeling data on ecological communities

Bert van der Veen

Department of Mathematical Sciences, NTNU

The goal of this presentation

Instill basic thinking about study design and data properties

1) Study design and sampling matter a lot
2) Adjust the model, not the data
3) Consider the "true" model: what is your ecological process?
4) Garbage in, garbage out

Leading example: picking orchids

Sampling data

Figure 1: www.ugent.be

A nice field with orchids.

How do we find the proportion of orchids?

What is the proportion of orchids?

We decide to walk through the field and at 10 places record when we find an orchid (1) or not (0)

1. First time: 5 orchids from 10 picks ($5/10 = 0.5$)

What is the proportion of orchids?

We decide to walk through the field and at 10 places record when we find an orchid (1) or not (0)

1. First time: 5 orchids from 10 picks ($5/10 = 0.5$)
2. Second time: 2 orchids from 10 picks ($2/10 = 0.2$)
3. Third time: 8 orchids from 10 picks ($8/10 = 0.8$)

What is the proportion of orchids?

We conclude, half of the flowers are orchids ($15/30 = 0.5$). But encounter this guy:

He tells us that the true proportion of orchids is 0.4.

- What caused our estimate of the proportion of orchids to be inaccurate?
- And why did we not get the same proportion of orchids every time?

Aspects of sampling

There are loads of things that affect our sampling

- Where we look
- When we look
- How often we look
- The resources we have
- Who looks
- Things that walk away
- Things that get eaten

Can mess that up (and often do), consequence: we need to adjust our analysis

Preferential sampling

"I want to survey community A"

or

"I sample on an elevation gradient"

a) You have predefined your community; the predefinition affects your results
b) You have predefined your environment; the predefinition affects your results

Preferential sampling

"I want to survey community A"

or

"I sample on an elevation gradient"

a) You have predefined your community; the predefinition affects your results
b) You have predefined your environment; the predefinition affects your results

You self-limited the scope of your study, self-selected results for diversity, composition, environment, and so on.

Preferential sampling

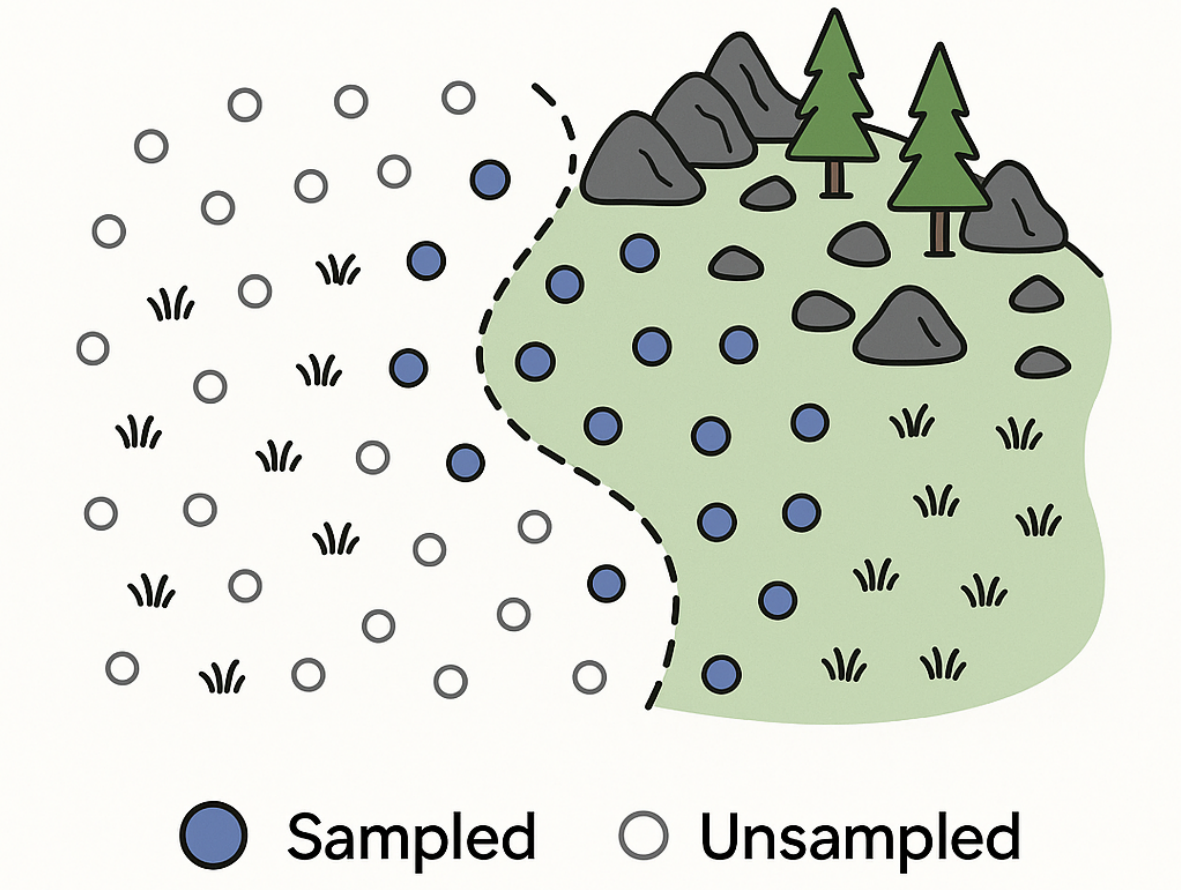
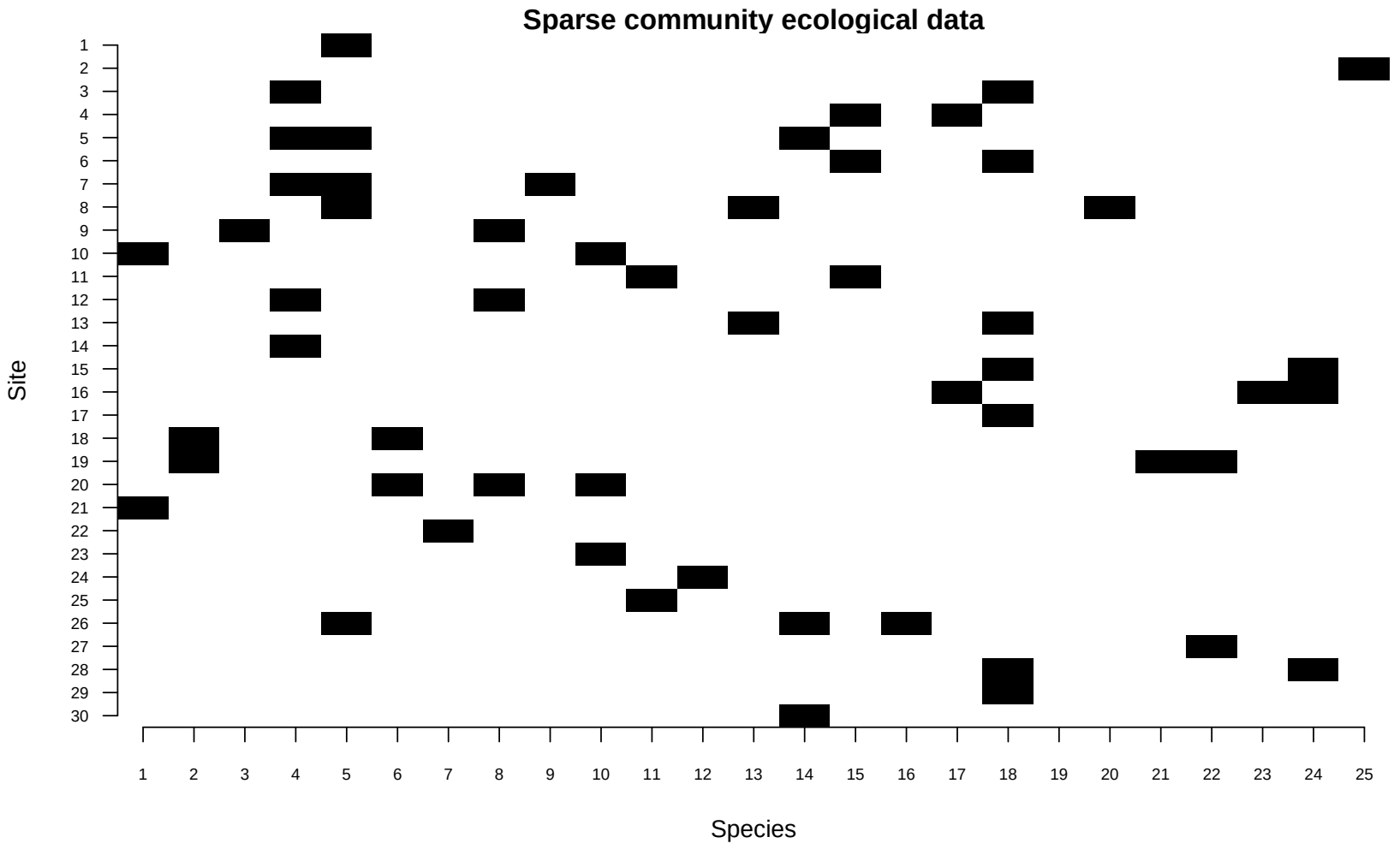
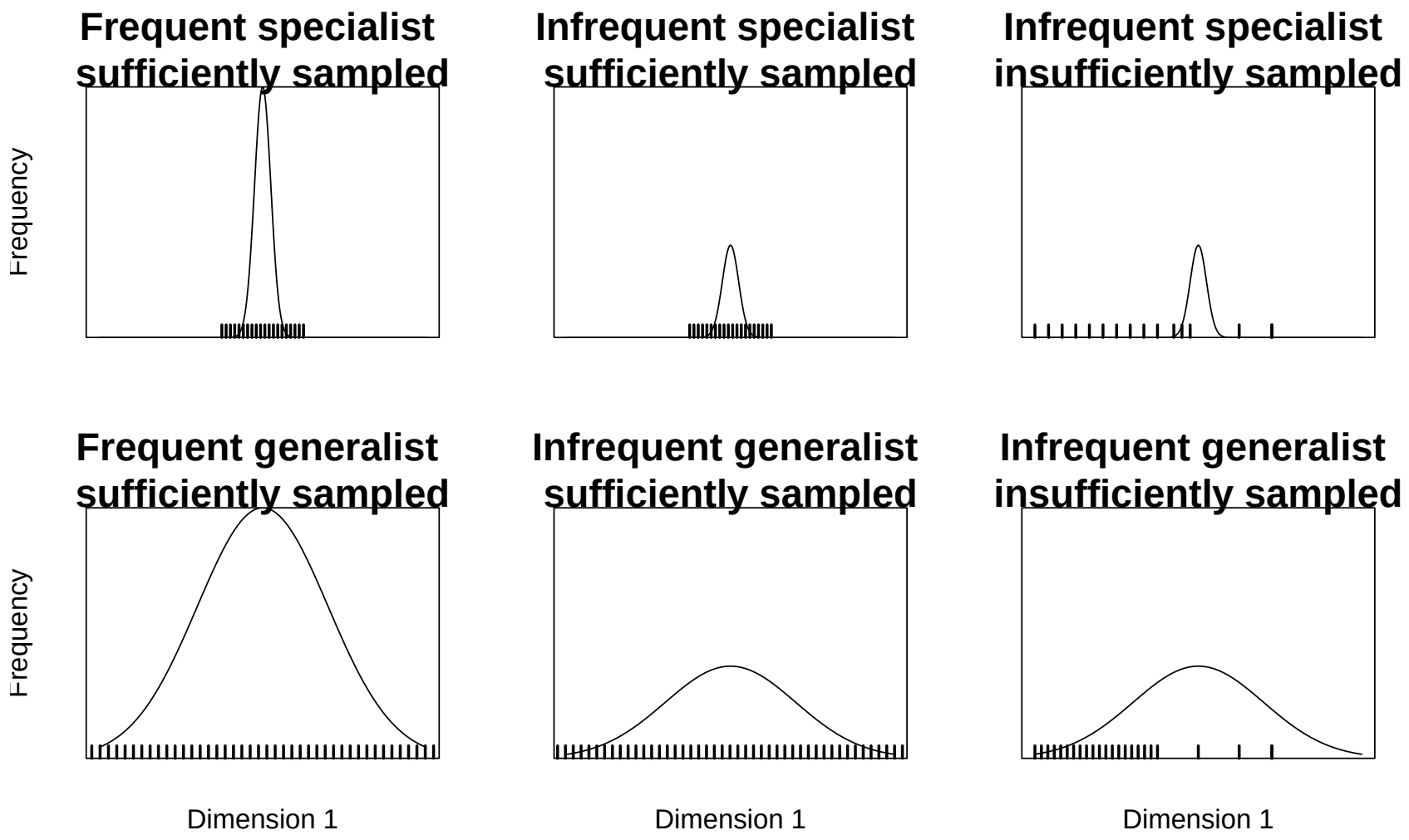


Figure 2: Thanks chatGPT

Consequences: few observations



"Rare" species



Sample size

Field work is hard, takes time, costs money.

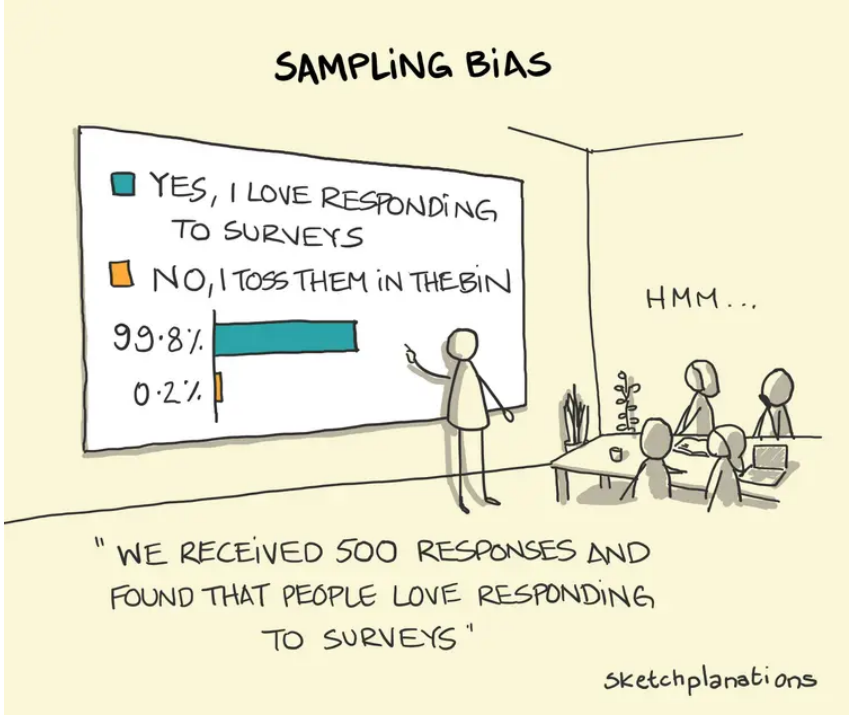
- Community ecological studies often have low samples
- And are noisy
- Combined with strong mean-variance relations this causes issues
- Studies are underpowered and lack information
- Many species have few observations
- Drawing conclusions is sometimes not possible
- Can largely be avoided with power analysis

Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in community ecology

- Opportunistic (eek)
- Random sampling
- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling

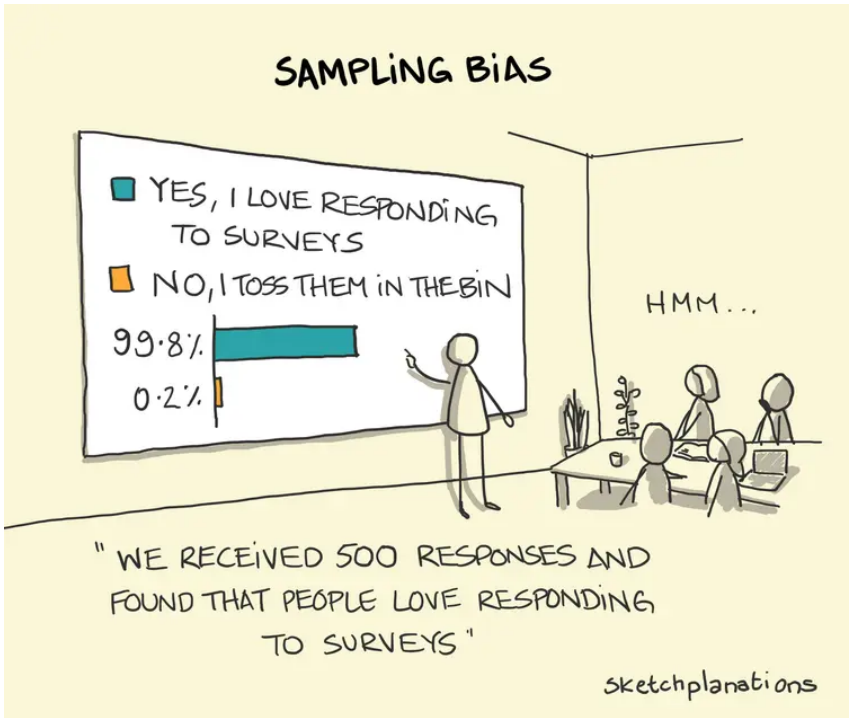


Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in community ecology

- Opportunistic (eek)
- Random sampling
- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling



Sampling design affects our sample size, and the ecological results. It needs to be taken into account during analysis.

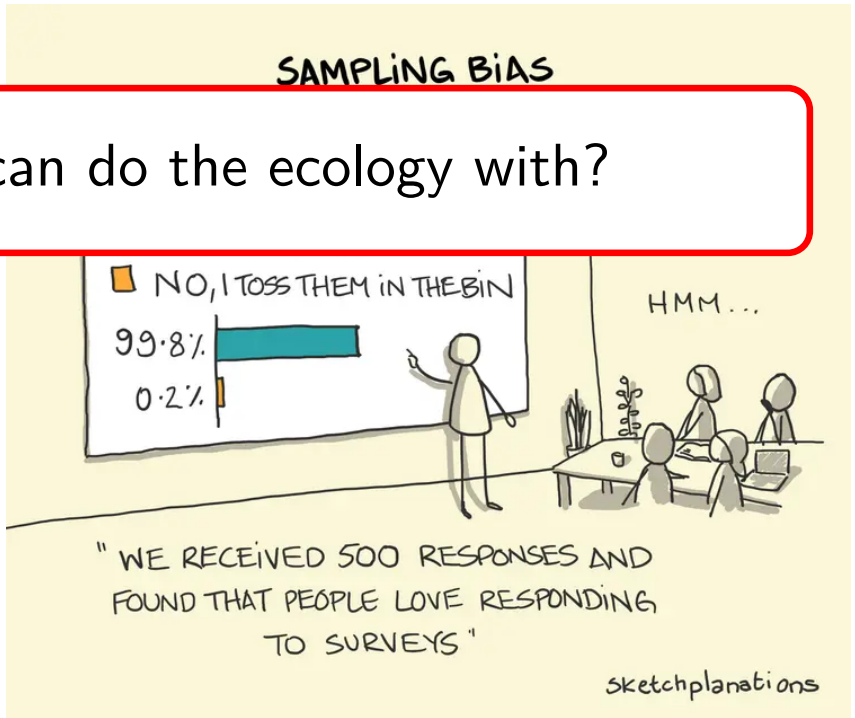
Minimizing impact of the sampling process

We can minimize the effects of sampling by considering its effects *a-priori*

There are many sampling designs in

Does it give data that you can do the ecology with?

- Random sampling
- Systematic sampling
- Stratified sampling
- Stratified-random sampling
- Adaptive sampling
- Cluster sampling
- Paired sampling



Sampling design affects our sample size, and the ecological results. It needs to be taken into account during analysis.

Detection bias

This one is not often covered, but certain species are harder to sample (identify or find) than others.

- Not considering it: you assume perfect detection
- Plants are easier than moving things
- Plants or flower are seasonal
- Pollinators fly at particular conditions
- Insects have different life stages (some easier to detect)
- Some people are better at finding things
- We should consider where species **can occur**

Classification error

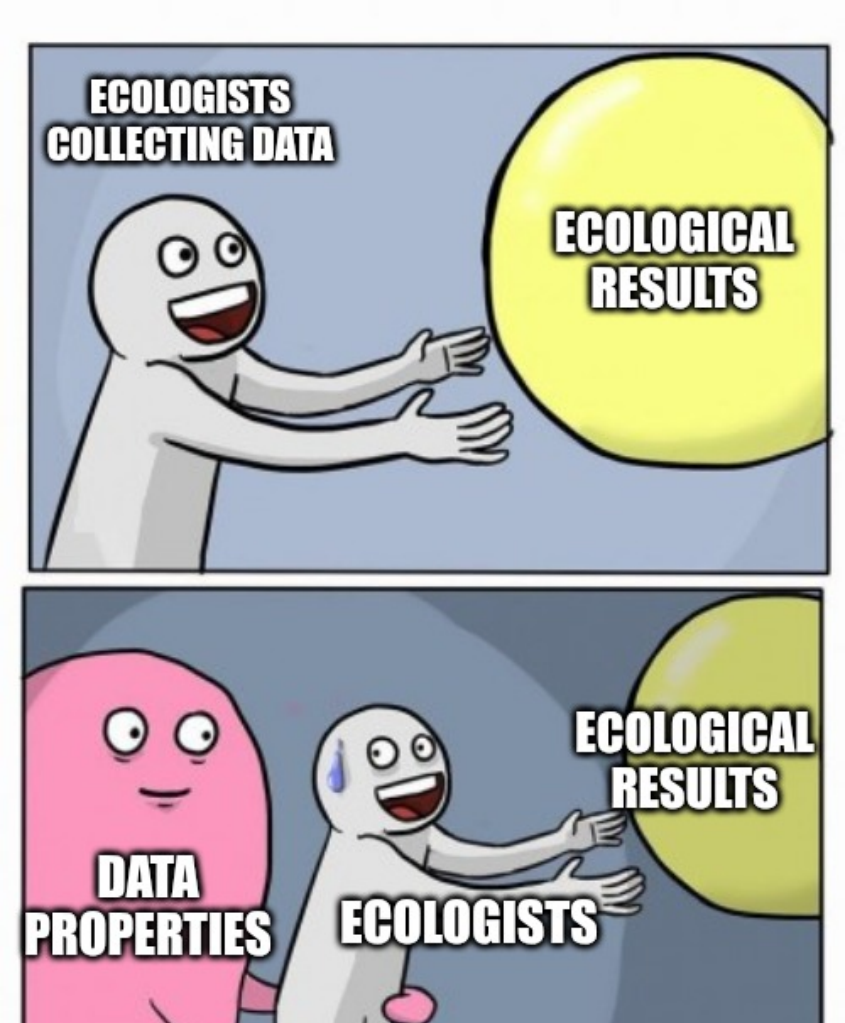
Classification mistakes introduce error: we confuse a species with another.

Exacerbated if you have multiple observers.

Getting results

You have got your data, and are ready to do some ecology!

Data of ecological communities has various common properties that tend to get in the way.



The data

The properties of data depend on the type. We characterize these by a distribution.

For (binary) orchid data: $y_i \sim Binom(p, n_{picks})$, with $p(orchid) = p$

The data

The properties of data depend on the type. We characterize these by a distribution.

For (binary) orchid data: $y_i \sim Binom(p, n_{picks})$, with $p(orchid) = p$

The distribution informs us what the probability is to observe a data point as a function of some model
This type of probabilistic framework facilitates us in getting an estimate for $p(orchid)$

The binomial distribution

$$f(y_i; n_{picks}, p) = constant \times p^{y_i}(1-p)^{n_{picks}-y_i} \quad (1)$$

Moments

- mean: $\mathbb{E}(y_i) = n_{picks} \times p(orchid)$
- variance: $\text{var}(y_i) = n_{picks} \times p(orchid)(1 - p(orchid))$

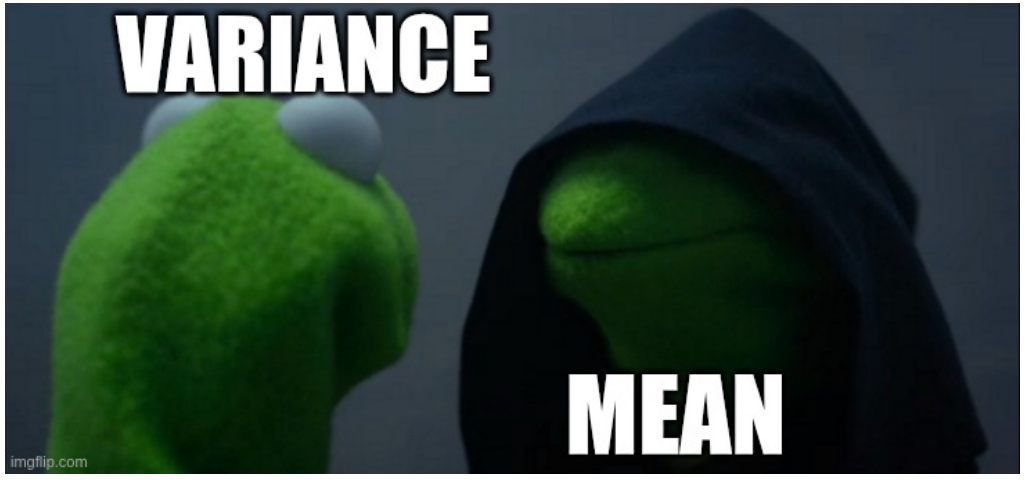
R-functions

- Density: `dbinom`
- Number generator: `rbinom`

Mean-variance relationships

Unless your data come from a normal distribution, the variance depends on the mean

- Ecological data often have strong mean-variance relationships
- This will muck up your results if not accommodated



Simulation: counting orchids once

```
p.orchid = 0.4 # The true proportion of orchids
n.picks = 10 # The number of picks in the field
n.times = 1 # number of fields
# Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
y/n.picks # Proportion of orchids
```

```
## [1] 0.5
```

Simulation: counting orchids once

What if we sample the whole field once?

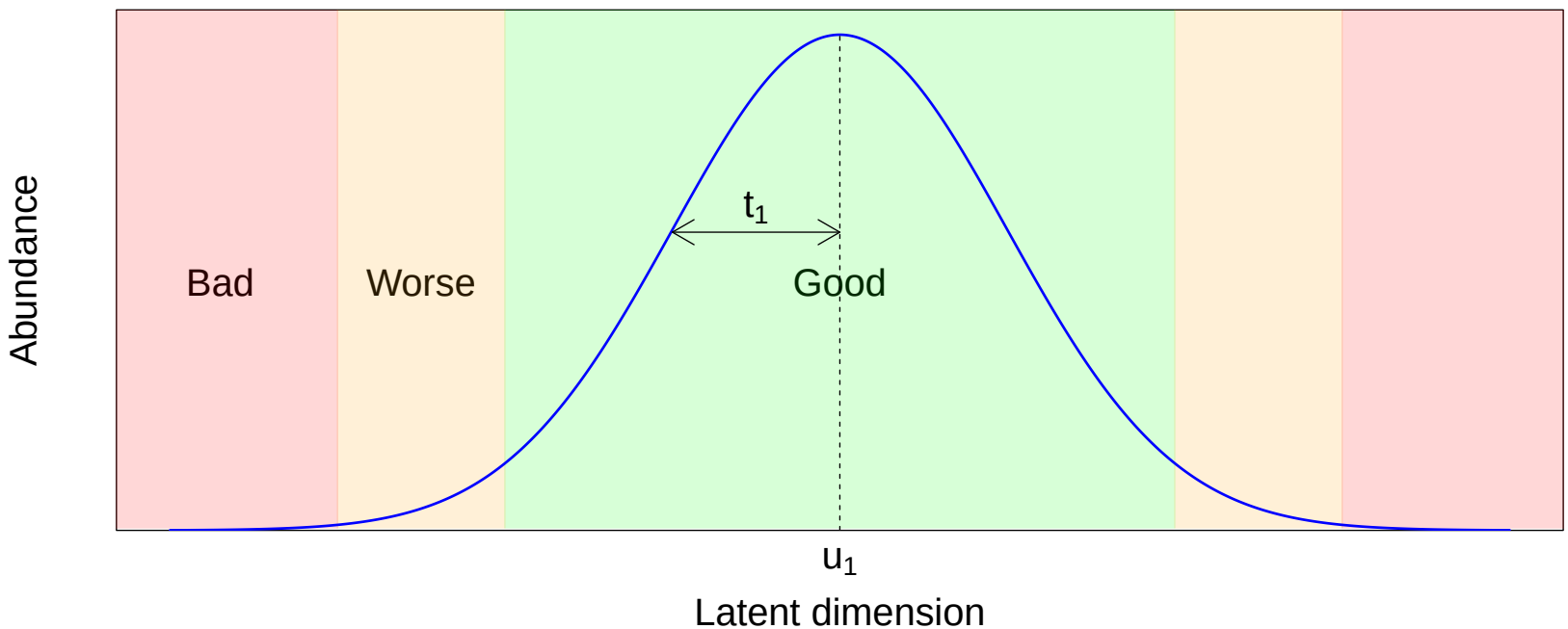
```
n.times = 1e5 # The number of picks in the field
n.picks <- 1 # number of fields
 # Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
mean(y/n.picks) # Proportion of orchids
```

```
## [1] 0.40021
```

Rare species

Shelford's law of tolerance (1931) tells us:

- There are specialist and generalist species
- Many species naturally occur only at a few places



Dimensionality

There are often many species in the data; sieving through results is difficult, and analysis can be computationally intensive.

At the same time, data are sparse.

Other things

- Non-linearity
- Compositionality
- Ordering

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*] Jean M. Macklaim[1] Vera Pawlowsky-Glahn[2] Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

Models

Traditional methods of analysis in community ecology are not good at dealing with data properties.

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

Forum | Open Access

The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)

David I. Warton, Francis K. C. Hui

First published: 26 July 2017 | https://doi.org/10.1111/2041-210X.12843 | Citations: 28

Models to the rescue

Plant Ecol (2015) 216:669–682
DOI 10.1007/s11258-014-0366-3

Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

General attitude

Repeat after me:

We adjust the model, not the data adjusting the data is bad

Process-based thinking

1. There is a sampling process
2. There is an ecological process

Our data is the result of both, our primary interest is the latter.

Process-based thinking

1. There is a sampling process
2. There is an ecological process

Our data is the result of both, our primary interest is the latter.

Our goal is to disentangle 1. from 2.

Orchids

In case of the orchids, what affects where we see orchids?

1. Where we look (sampling/observation)
2. Where they are (ecology)

In both cases, if we look in the wrong place, at the wrong time, or in the wrong way, we may not see orchids (even if they are there).

- It is not interesting if you observed more orchids because you are better at finding them
- It is interesting if you observed more orchids because the places you went provided more suitable growing conditions

To do ecology

We need to carefully consider properties of our data, how it is sampled, and thus what our analysis needs to accommodate.

- Strong mean-variance
- Repeated designs (e.g., multiple observers)
- What am I trying to answer ecologically?

Garbage in, garbage out (GIGO)

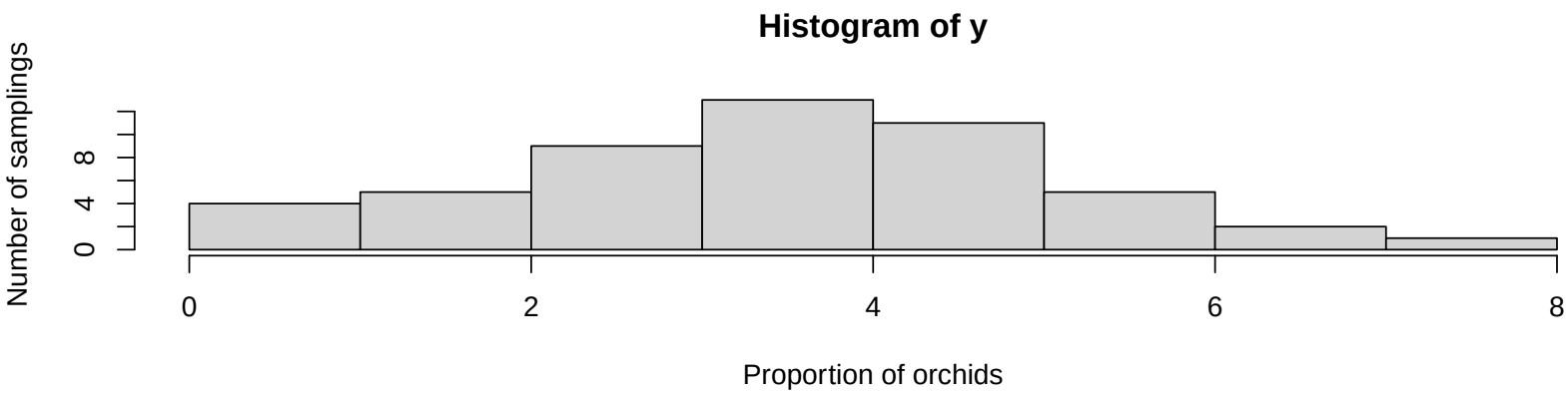
In absence of a good study design, models can help. However, a fancy hammer is not a panacea.

Estimating parameters and quatifying uncertainty

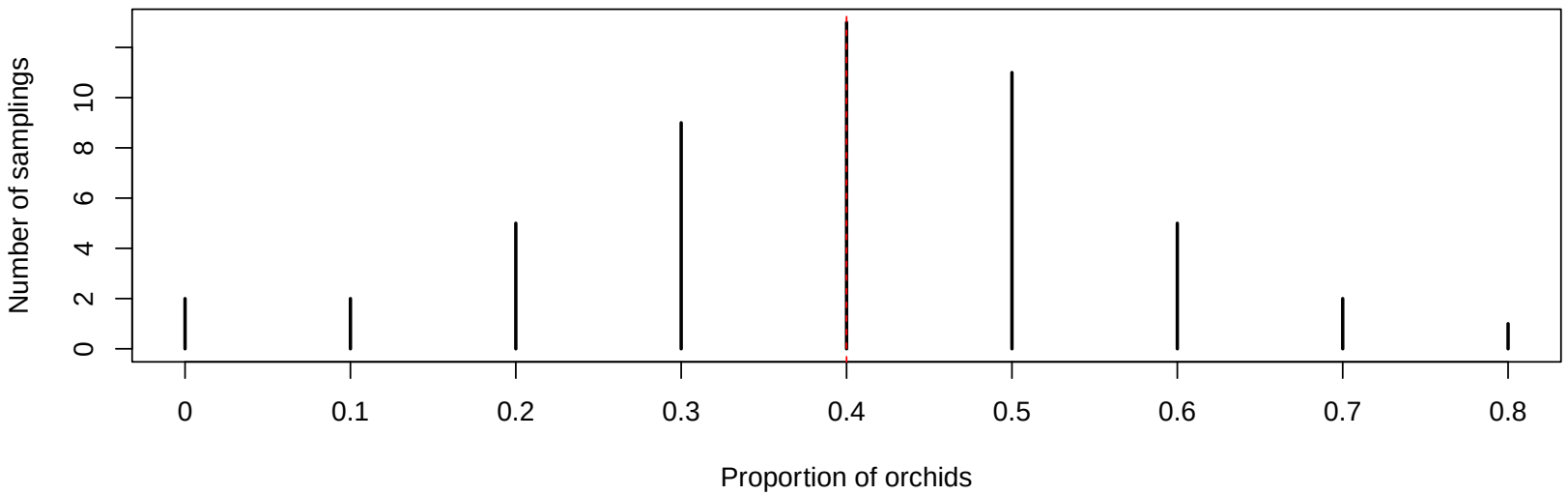
- We do not usually have infinite amounts of data
- So how can we quantify variability?
- The model allow us to do that!

Simulation: counting orchids 50x10 times

```
n.times <- 50 # The number of picks in the field
n.picks = 10 # number of fields
# Collect data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
hist(y, xlab = "Proportion of orchids",
     ylab = "Number of samplings")
```



Simulation: counting orchids 50x10 times



As you see, we have variability in our estimate of the proportion of orchids.

- Can we summarize this variation?
- Preferably without collecting data many times

Q: Are more than half of the flowers in this field orchids?

- Exclude rare events, and decide on an acceptable margin of error (5%)
- What is the range our parameter is estimated to be 95% of the time?

```
quantile(y/n.picks,c(0.025,.975))
```

```
##   2.5%  97.5%
## 0.0225 0.7000
```

So 50 times 10 picks tell us little.

Q: Are more than half of the flowers in this field orchids?

```
set.seed(12345)
n.picks = 100
n.times <- 50
# Collecting data
y <- rbinom(n.times, size = n.picks, prob = p.orchid)
quantile(y/n.picks,c(0.025,.975))
```

```
##    2.5%   97.5%
## 0.31225 0.48000
```

50 times 100 picks reduces the variability in the proportion of observed orchids. Most of the time we will not find orchids on all our picks.

The likelihood: single data point

$$\mathcal{L}(y_i; \Theta) = f(y_i; \Theta) \tag{2}$$

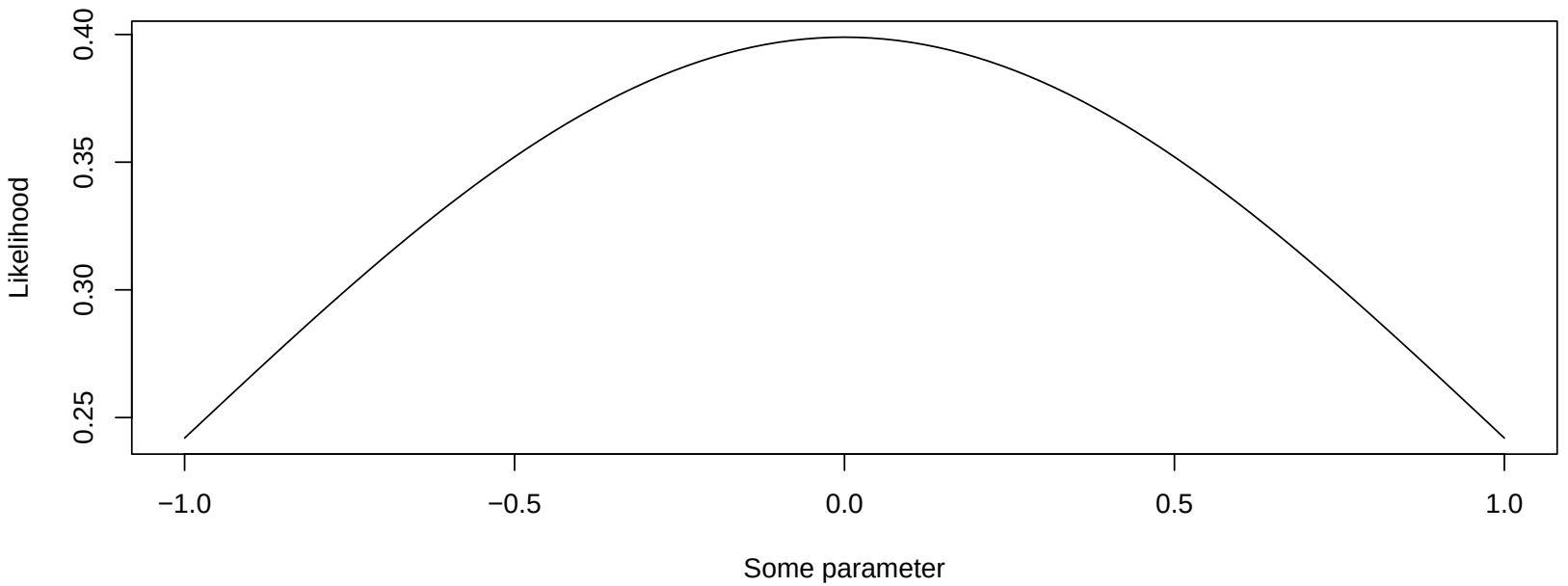
The probability of obtaining our data assuming Θ is the true parameter(s).

The likelihood: multiple data points (2)

$$\mathcal{L}(\mathbf{y};\Theta) = \prod_{i}^{n} f(y_i;\Theta) \tag{3}$$

We just multiply! (assumes independence)

The likelihood (3)

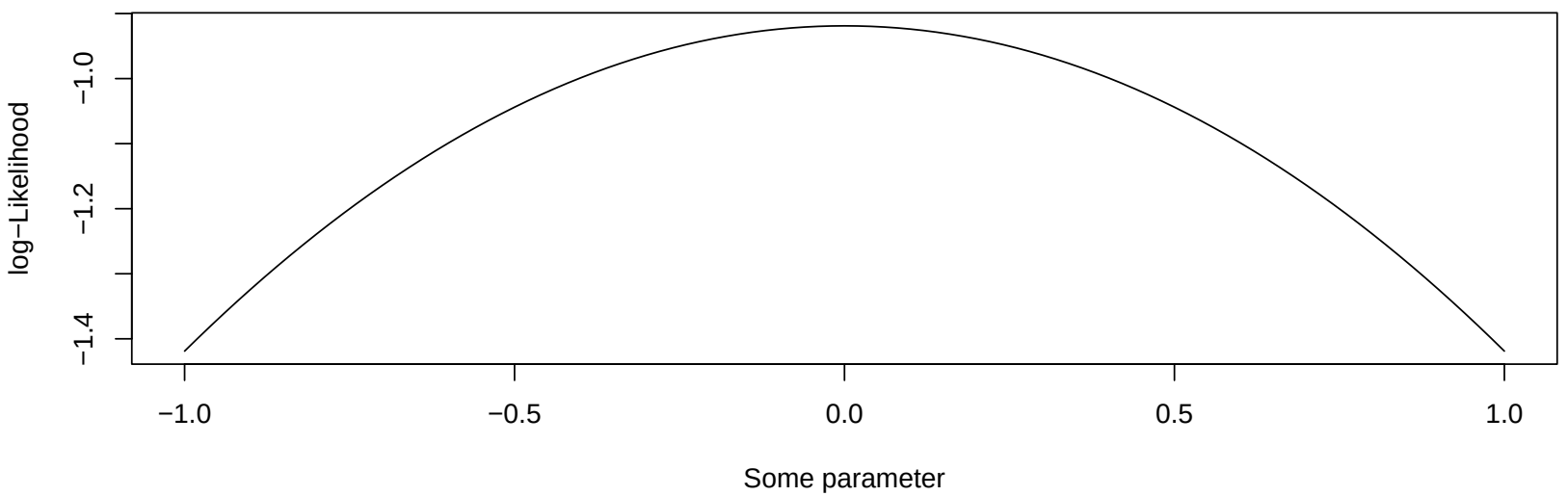


Likelihood tells us about:

- The (set of) parameter estimates that most likely generated the data
- The information contained in our data

The log-likelihood

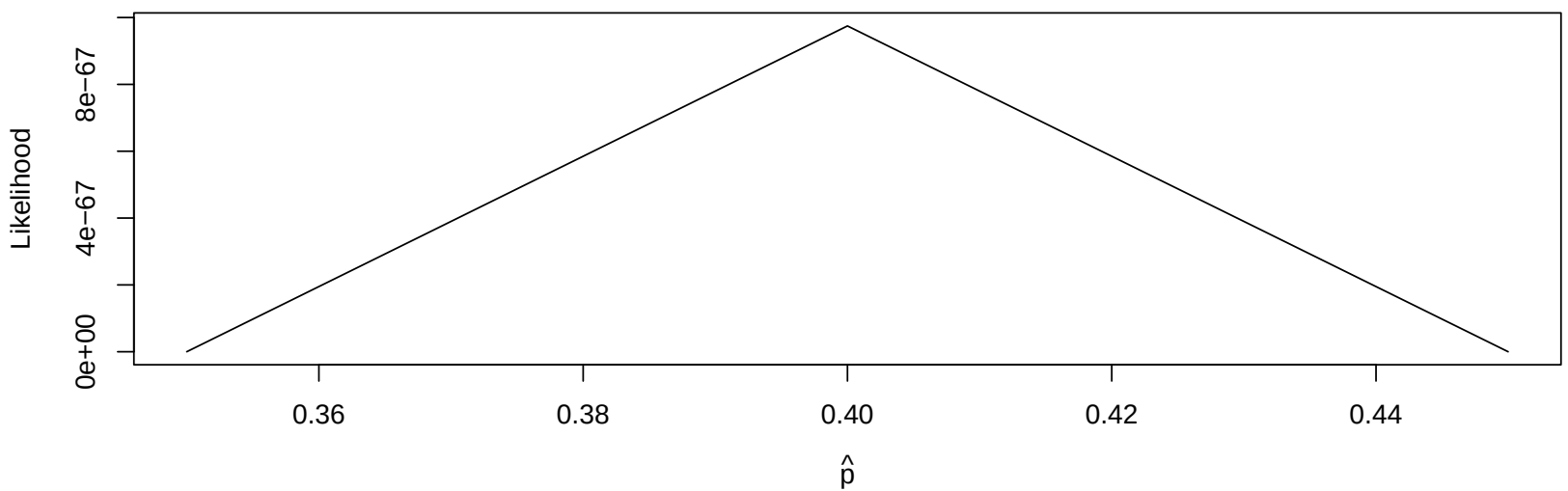
$$\log\{\mathcal{L}(\mathbf{y};\Theta)\} = \sum_{i}^{n} \log\{f(y_i;\Theta)\} \tag{4}$$



Usually, we work with the log-likelihood. The maximum is the same and it is easier. So we just add things together.

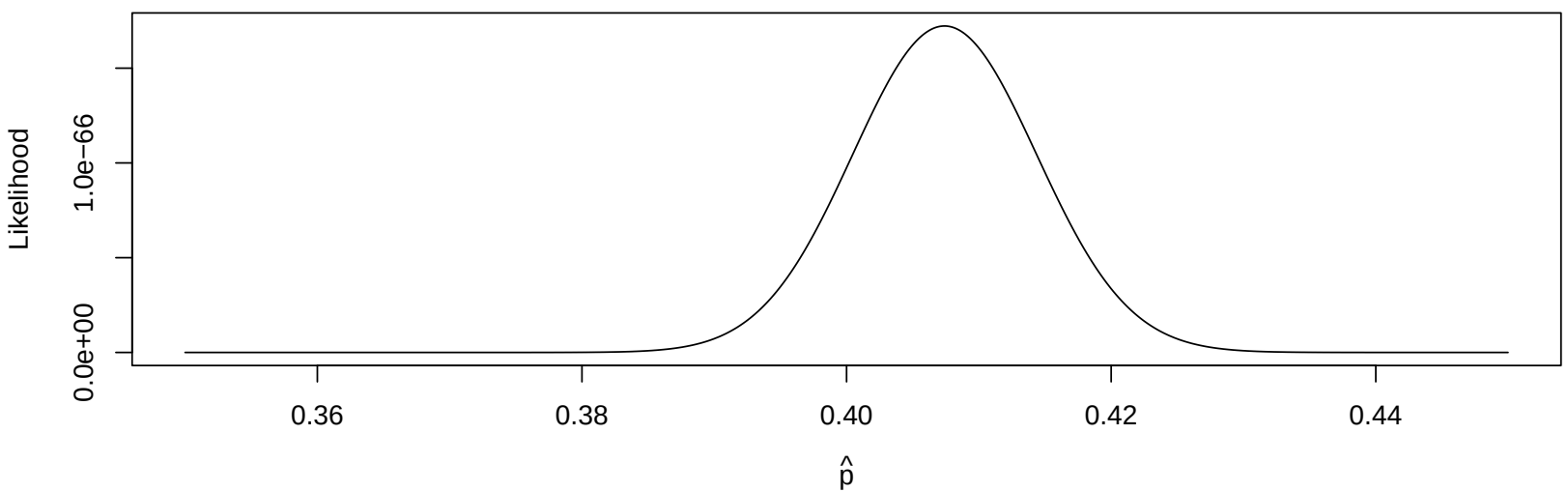
Finding the proportion of orchids

```
ll <- function(p, n.picks, y)prod(dbinom(y, n.picks,p))
phat <- seq(0.35,0.45,length.out=3)
plot(sapply(phat, ll, n.picks = n.picks, y = y),
     x = phat, type = "l", xlab=expression(hat(p)), ylab="Lik
```



Finding the proportion of orchids (2)

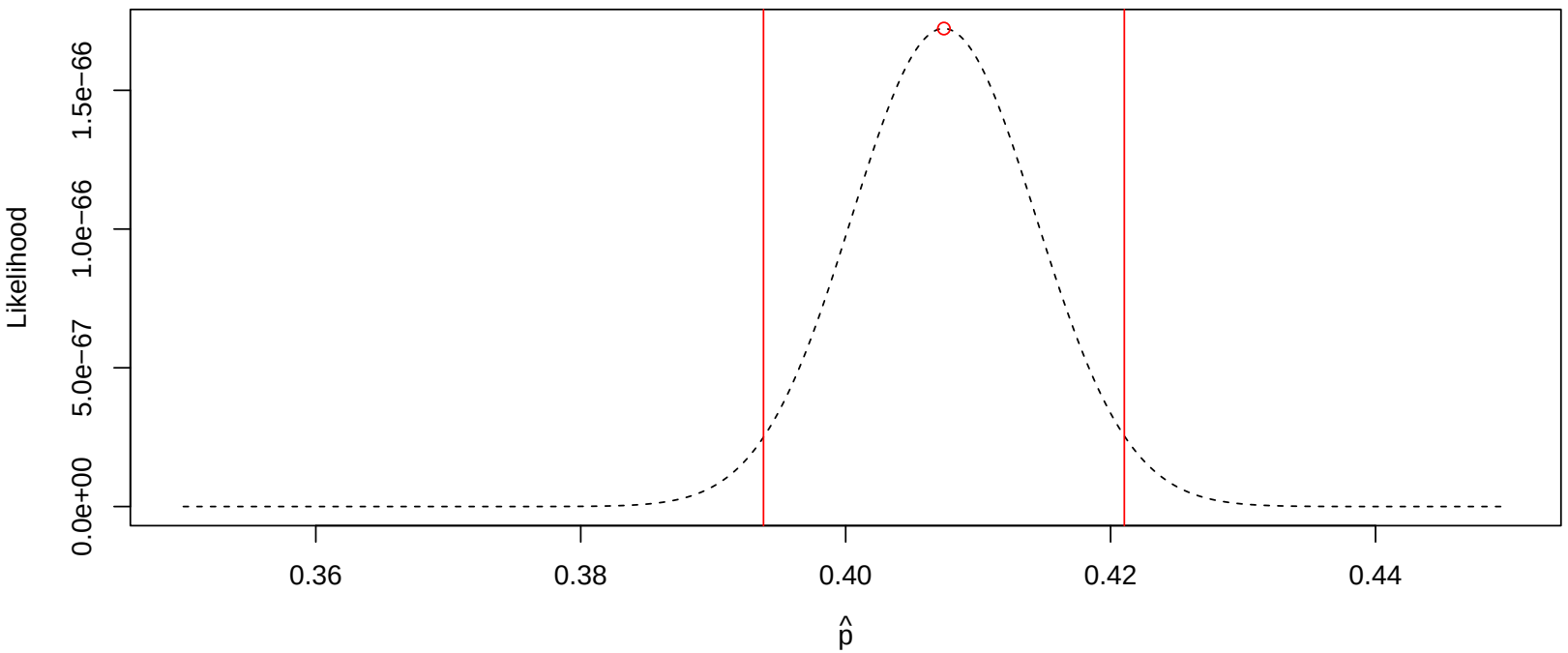
```
ll <- function(p, n.picks, y)prod(dbinom(y, n.picks,p))
phat <- seq(0.35,0.45,length.out=1000)
plot(sapply(phat, ll, n.picks = n.picks, y = y),
     x = phat, type = "l", xlab=expression(hat(p)), ylab="Lik
```



Uncertainty

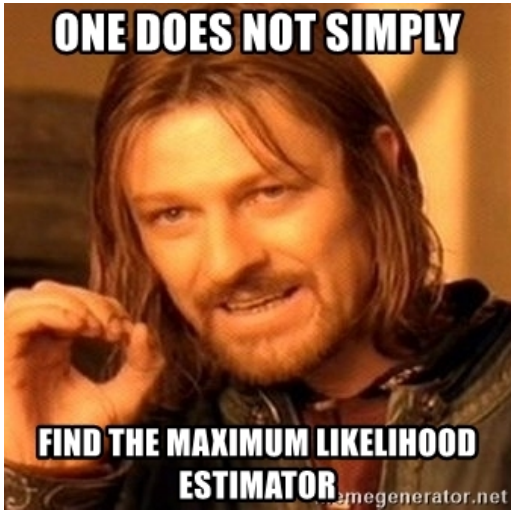
(an estimate of) Width of the likelihood:

$$\frac{\partial^2 \log\{\mathcal{L}(\mathbf{y}; n_{picks})\}}{\partial p^2} = -\sum_{i=1}^{n_{times}} \frac{y_i}{p^2} + \frac{n_{picks} - y_i}{(1-p)^2} \tag{5}$$



Putting it all together

- We collect data
- We estimate a parameter of interest
- If we collected data again, we get many different estimates
  - This forms a *sampling* distribution
- We summarize this variability
- The width of this sampling distribution tells us the variability
- Instead of collecting data many times, we estimate parameters with MLE
  - This also allows us to quantify the variability



Why is uncertainty so important

- We are not interested in an answer for **this** dataset
- But for an answer for **all** datasets
- If we have new data, our answer might change a little
- Uncertainty tells us if our answer is robust to sampling new data
- I.e., not so important for the *dataset* but important for *multiple datasets*

Afterall, we are looking for a robust recommendation.

Confidence intervals

An interval that contains the true value in 95% of repeated samples (in large samples)

Be careful with interpretation, and with assumptions.

- Any computed interval either contains the truth, or it does not
- Not the range that the true parameter falls in with 95
- Other misinterpretations

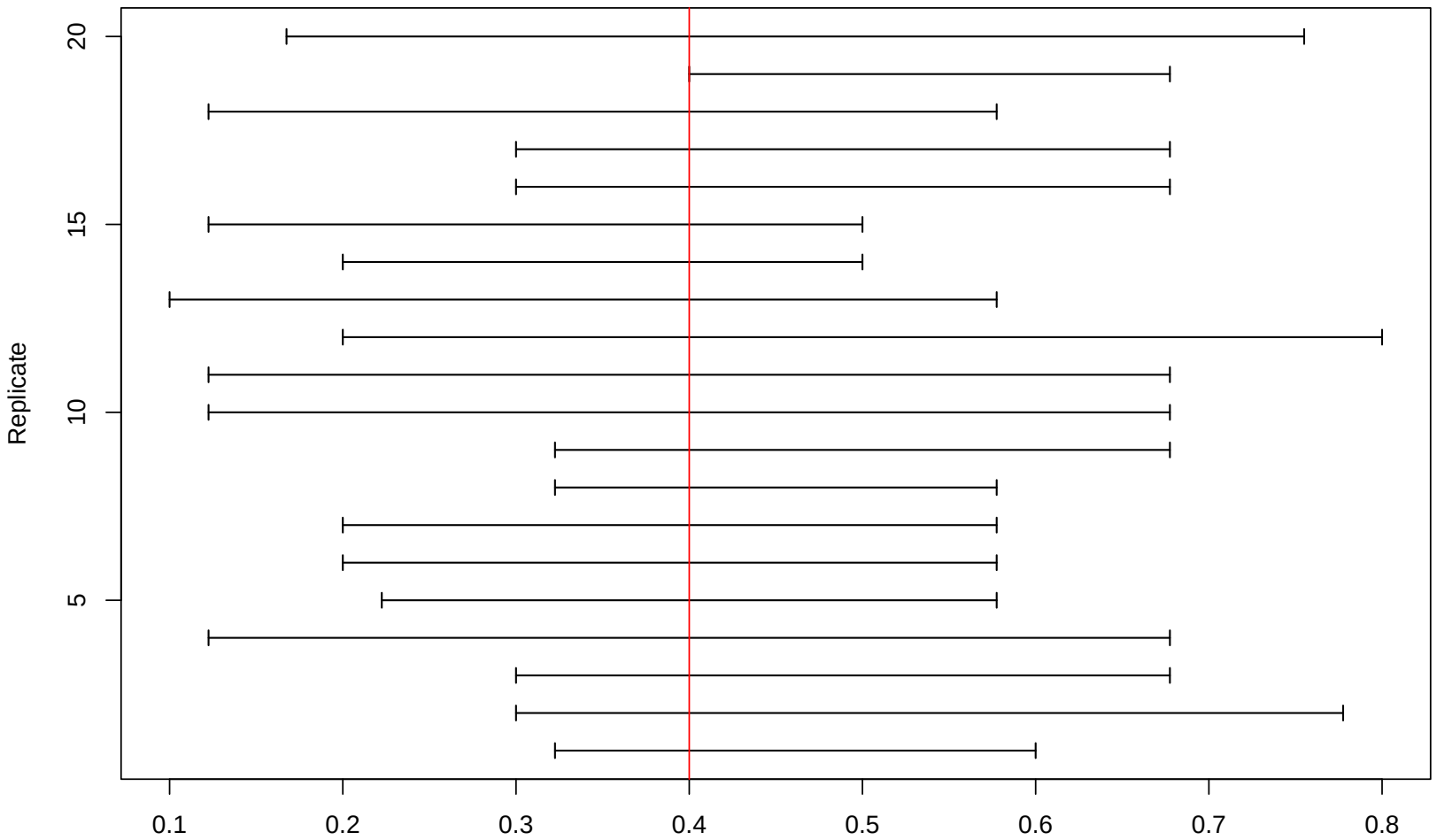
Assumes:

- Asymptotic normality
- inverse Hessian gives covariance of estimators
- Can be interpreted as a kind of statistical test
- Or generally as "evidence"

Gets smaller with:

- More information
- Less variability
- The confidence level

Repetition



Summary

- Most data properties can be accommodated with models
- It requires consideration of sampling and ecological processes
- Choose the appropriate model, not the software you like
- Some properties are more difficult to accommodate
- Sparsity, sample size issues, and misclassification are tough
- Many issues do not show in traditional methods
- Models will be more honest to you