Lecture 1

Introduction to statistical inference in ecology

WILD6900 (Spring 2021)

Reading

> Hobbs & Hooten 3-16

Ecology [1]

> the comprehensive science of the relationship of the organism to the environment (Haeckel 1866)

> the study of the natural environment, particularly the interrelationships between organisms and their surroundings (Ricklefs 1973)

> the scientific study of the distribution and abundance of organisms (Andrewartha 1961)

> where organisms are found, how many occur there, and why (Krebs 1972)

Ecological state variables

State variables are the ecological quantities of interest in our model that change over space or time [2]

Abundance

> the number of individual organisms in a population at a particular point in time

Occurence

> the spatial distribution of organisms with a particular region at a particular point in time

Richness

> the number of co-occuring species at a given location and a particular point in time

Ecological parameters

Parameters determine how the state variables change over space and time

- Survival
- Reproduction
- Movement
- Population growth rate
- Carrying capacity
- Colonization/extinction rate

Models of populations

Inference in ecology **requires** models

Models link **observations** to **processes**

Models are tools that allow us understand processes that we **cannot directly observe** based on quantities that we **can** observe

By necessity, models are simplifications of reality

Types of expertise

1) Domain expertise

> knowledge based on experience and understanding of the *ecological* system of interest

2) Statistical expertise

> knowledge of probabilistic modeling and computation

Useful models should be consistent with *both* domain and statistical expertise!

Notation

Parameter(s)

$$\theta$$

Observation(s)

$$y$$

Predictor(s)

$$x$$

Lightface = scalar

$$(y, \theta, x)$$

Boldface = vector

$$(\mathbf{y}, \theta, \mathbf{x})$$

Probability distribution

$$[a|b, c]$$

Deterministic function

$$g()$$

A line of inference in ecology

Process models

Process models

$$g(\theta_p, x)$$

- Mathematical description of our hypothesis about how the *state variables* we are interested in change over space and time
- Represent the **true** value of our state variables at any given point in space or time
- Deterministic
- Abstraction

Process models

- Abstraction = uncertainty [3]
- Unmodeled sources of variation = σ_p^2
- State variable (z) modeled as a *probability distribution*

$$[z|g(\theta_p, x), \sigma_p^2]$$

Process models

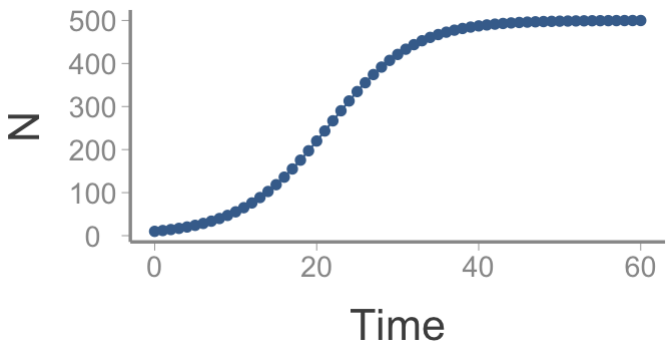
Example

We are interested in predicting the population growth of species a as a function of abundance in the previous year [4]

We hypothesize that population growth rate will be highest at low densities and lowest (maybe even negative) at high density

This leads us to believe that the *discrete logistic equation* is a good descriptor of our system:

$$N_{t+1} = g(\theta_p, x) = N_t e^{r_0[1-(N_t/K)]}$$



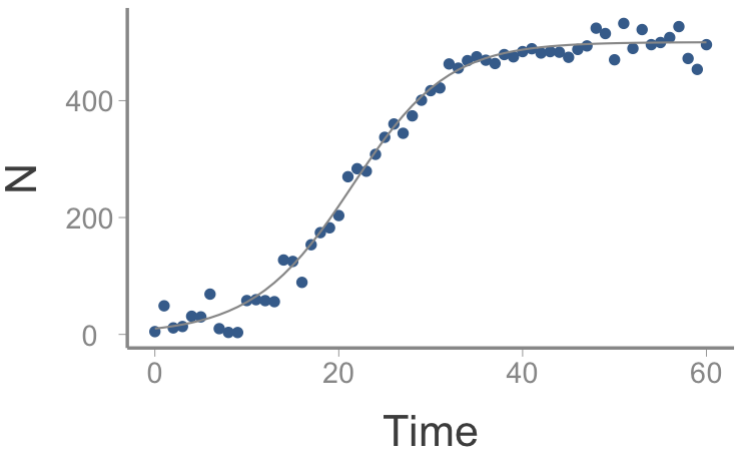
Process models

Example

In reality, N will not fall exactly on the line predicted by the Ricker model

There will obviously be processes other than density-dependence that are influencing population size in each year that are not accounted for by our model

- In other words, there is *process uncertainty* $(\sigma_p^2 > 0)$:



Process models

Interpreting σ_p^2

- The process model represents the **true** value of N, *not* our observation of it.
- σ_p^2 is as a measure of how well our process model fits reality
- To minimize process uncertainty, we need *a better model*. No amount of additional data will lower σ_p^2.

In our example, maybe we have environmental covariates (rainfall, temperature, etc.) that we also think are important. To reduce process uncertainty, we would need to modify our process model to include these effects.

Sampling models

Sampling models

- Obtaining probability distributions about our state-variables and parameters requires **data**
- Data are samples of the true population
    - Our sample will not perfectly represent the true state of the system
- As for the process model, we can represent sampling uncertainty σ_s stochastically using a probability model [5]:

$$\left[u_i \middle| z, \sigma_s^2\right]$$

Sampling models

Example

In our population size example, suppose we conduct $i = 1, 2, 3, \ldots, K$ transect or point counts to estimate abundance. The area of our counts (we'll call it a) is not the entire area of our population $(a < A)$. If we want to estimate N_t, we need a model linking our counts (call them $n_{t,i}$) to the true abundance. If we assume individuals are uniformly distributed across our study area, then perhaps we could use:

$$\left[\sum_{i=1}^{K} n_{t,i} \middle| \frac{N_t}{a}, \sigma_s^2 \right]$$

In this case, our counts n_t will be different if we had chosen different transect routes or points. This is what σ_s^2 represents.

Separating σ_s^2 from σ_p^2 is important because we *can* lower σ_s^2 by collecting larger sample sizes or increasing replication

Observation models

Observation models

- σ_s^2 only captures uncertainty that is due to the randomness of our sampling process
- Even when sampling, we rarely observe the true state perfectly
  - Animals are elusive and may hide from observers
  - Individuals may not be "available" during our sample
  - Even plants may be cryptic and hard to find
- This *observation uncertainty* (σ_o^2) can lead to biased estimates of model parameters, so generally requires its own model [6]

$$[y_i | d(\Theta_o, u_i), \sigma_o^2]$$

Observation models

Example

If we used the $n_{t,i}$ as our observations, we would have to make the assumption that we counted every individual in our sampling area. This is almost never a good assumption in studies of animals (and even plants!)
Another way to say this is that $n_{t,i}$ is the *true* number of individuals in our sampling area but usually cannot count every individual [7]

In our count model, we might define a parameter ψ that is the probability that individual that is present in our sample is counted by the observer (we could further use a generalized linear model to account for the effects of, e.g., weather or observer skill, on ψ):

$$[y_{t,i}|\psi n_{t,i}, \sigma_o^2]$$

where σ_o^2 is uncertainty about the value of ψ.

Parameter models

Parameter models

- In Bayesian inference, parameter models express what we know about our parameters *prior to* collecting data
- Parameter models are more commonly referred to as *priors* [8]
- Every parameter in our model requires a probability distribution describing the prior probability we place of different values the parameter could take

$$[\theta_p][\theta_o][\sigma_p^2][\sigma_s^2][\sigma_o^2]$$

- These distributions can provide a lot or a little information about the potential value of each parameter

The full model

The full model

With each of our models created, we are prepared to right out the full model:

$$\Big[\underbrace{z, \theta_p, \theta_o, \sigma_p^2, \sigma_s^2, \sigma_o^2, u_i}_{unobserved} \Big| \underbrace{y_i}_{observed}\Big] \propto \underbrace{[y_i|d(\Theta_o, u_i), \sigma_o^2]}_{Observation\ model} \underbrace{[u_i|z, \sigma_s^2]}_{Sampling\ model} \underbrace{[z|g(\theta_p, x), \sigma_p^2]}_{Process\ model} \underbrace{[\theta_p][\theta_o][\sigma_p^2}_{Parameter}$$

Notice that:

- all of the quantities that we do not directly observe (parameters and state-variables) are on the left side of the conditioning symbol "|"
- This means that they are treated as *random variables* that are governed by probability distributions
- treating all unobserved quantities as random variables and specifying probability distributions for each quantity is what makes this model *Bayesian*.