Ecological Statisitics and Data

Eric Scott

2019-01-16

Welcome to Ecological Statistics and Data!

"Modeling" vs. "statistics"

- 20th century: design experiments to fit statistical analysis assumptions
- 21st century: option to design statistical models to be more like data
  - But, with great power comes great responsibility

"Ecological"

- Ecologists often count things and watch when they reproduce and die
- These types of data have special properties not typically covered in as much depth in more general or introductory courses
- Focus on linear models and their extensions

Introductions

- Me: your instructor, a chemical ecologist studying tea chemistry and climate change
- Avalon: your TA, an ecologist studying fireflies and light pollution
- You: Undergraduates and graduate students from several departments with varied backgrounds in math, statistics, computer programming, and ecology.

I want you to take advantage of that variation and learn from eachother. Also, don't be afraid to ask questions! I'm likely to incorrectly assume you know something that you don't!

Today's Outline

1. Logistics
2. Probability
3. Homework

Logistics

- General format (lectures mixed with computer exercises – find a partner or 2)
- Books: Bolker, R4DS
- Syllabus
- Access to laptops
- Lab

Outline for next few weeks

1. Probability
2. Liklihood
3. Bayes' Theorum
4. Linear models

Probability: textbook explanation #1

Vocabulary:

- "event" something that does or does not occur (person has a cold or not; stalk of a plant has a flower or not; you heard a woodpecker or not)
- "trial" = single observation or data point or "experiment" in which the event can occur or not (single person; single stalk; single walk in the woods when you listened for woodpeckers)

Notation: - $P(A)$ = probability that an event (A) will occur in a single trial - N = total number of trials - N_A = number of times A occurs

technical definition of probability:

> If an observation is made N times and event A occurs N_A times, then with a high degree of certainty, the relative frequency of N_A/N is close to $P(A)$, the probability of A in a single trial, $P(A) \approx N_A/N$, provided N is sufficiently large.

- Basically describing sampling randomly to get at a *true* probability

Mathematical restatement:

$$P(A) = \lim_{N \to \infty} \left(\frac{N_A}{N} \right)$$

- How certain?
- How close?
- How large?
- ... this is why we need statistics.

Example application

Common ecological application: survival and reproduction of perennial wildflowers
Useful to know to understand and predict what will happen to populations over time

- Mark individual plants and watch their performance over time.
- What is the probability that a plant will survive from one year to the next?
- What is the probability that a plant that survives will flower?

FOR A PARTICULAR DATA SET, we can use the definition of probability to estimate these probabilities ("provided N is **sufficiently large**")

Say we have 5 plants, 4 live, and 1 of these flower

Plant #	fate
1	Flowers
2	Vegetative
3	Vegetative
4	Vegetative
5	Dead

N = total # trials N_A = number of times A occurs

Survival:
- What is N? [5]
- What is N_A?[4] - What is $P(A)$?

$P(A) \approx 0.8$

Flowering, for plants that survive:
- What is N? [4] - What is N_A? [1]

$P(A) \approx 0.25$

FOR THOUGHT: Based on these data, what is the probability that a plant survives AND flowers?

Sampling and Scope of Inference

Definition of probability assumes trials are representative of a "population"

"population" (statistics definition) = larger group for which you estimating the probability

Trials should be **representative** and **independent**.
- Sampling types: + Random + Stratified (e.g. randomly selected transects) + Systematic (e.g. every 10 paces along transect) + Haphazard - Scope of inference = "population" over which your data are a **representative** and **independent** observations

Homework

Measure a probability of some event by sampling a population.