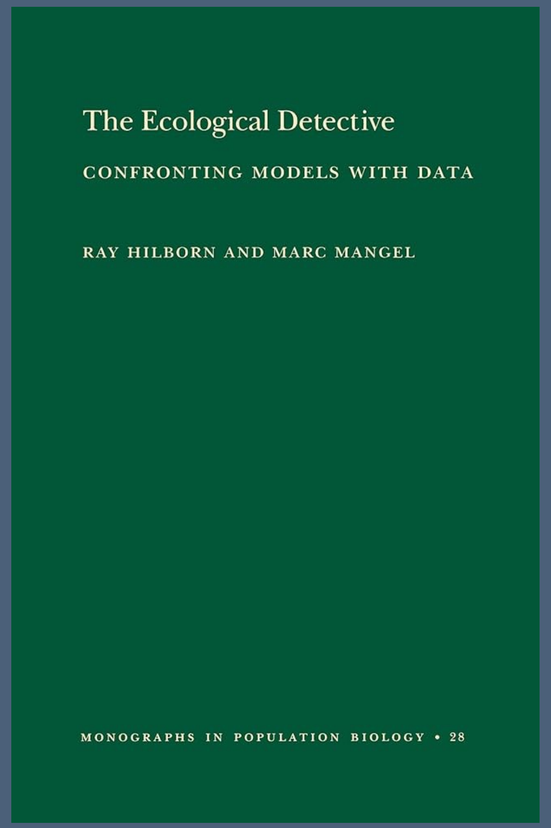
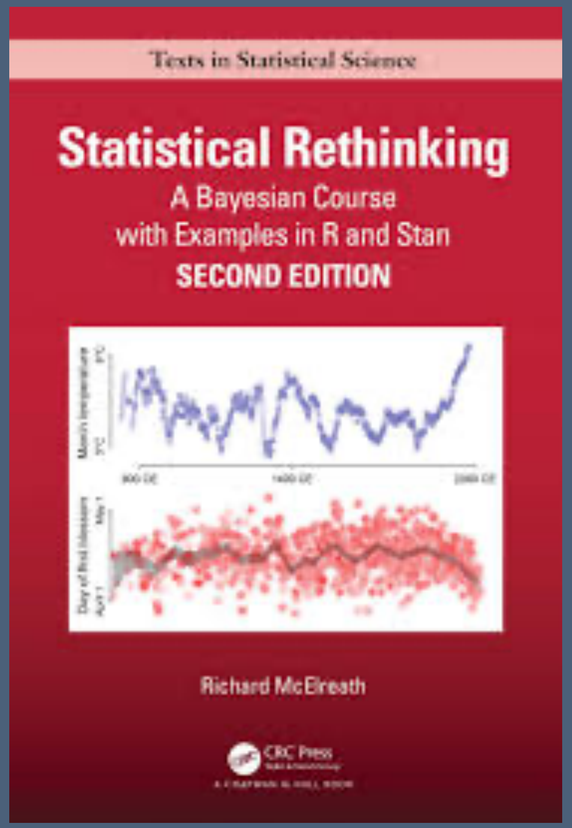
Connecting theory and data

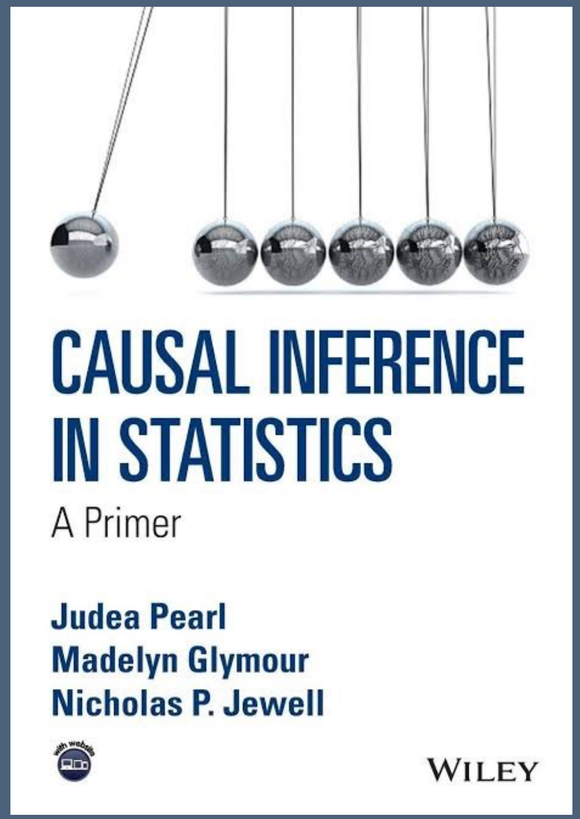
Overview of the 2025 course

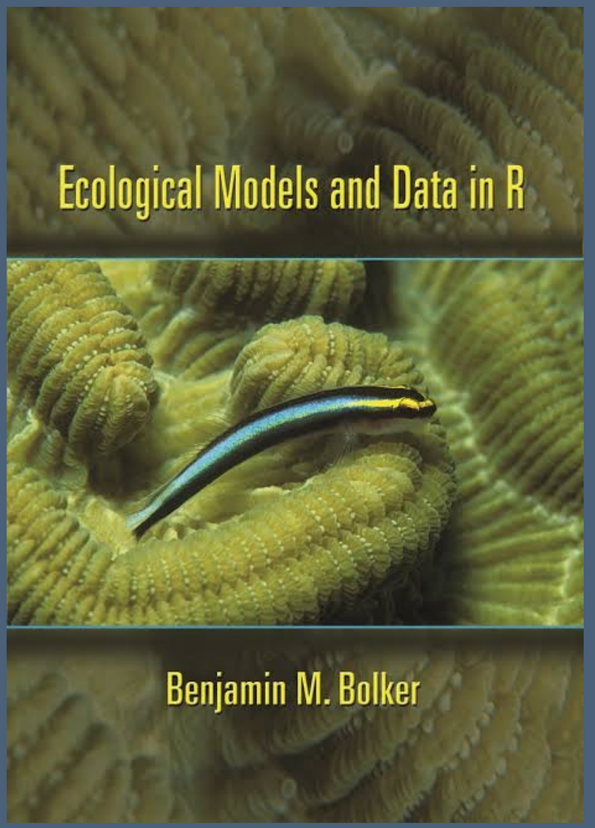
Sara Mortara & Diogo Melo

Our main references









Linking data and theory

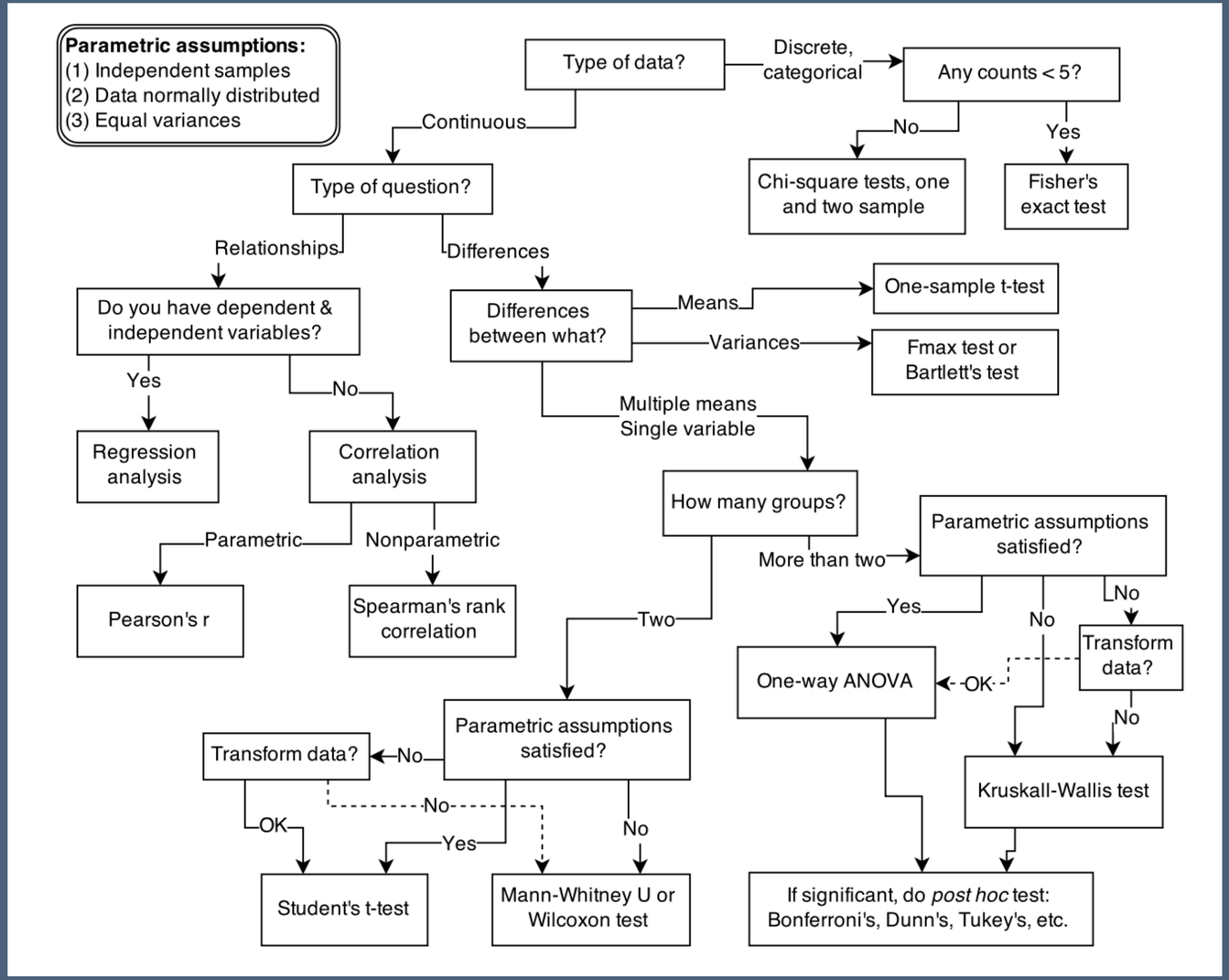
How do we address ecological problems characterized by complexity and uncertainty?

Ibã Huni Kuin

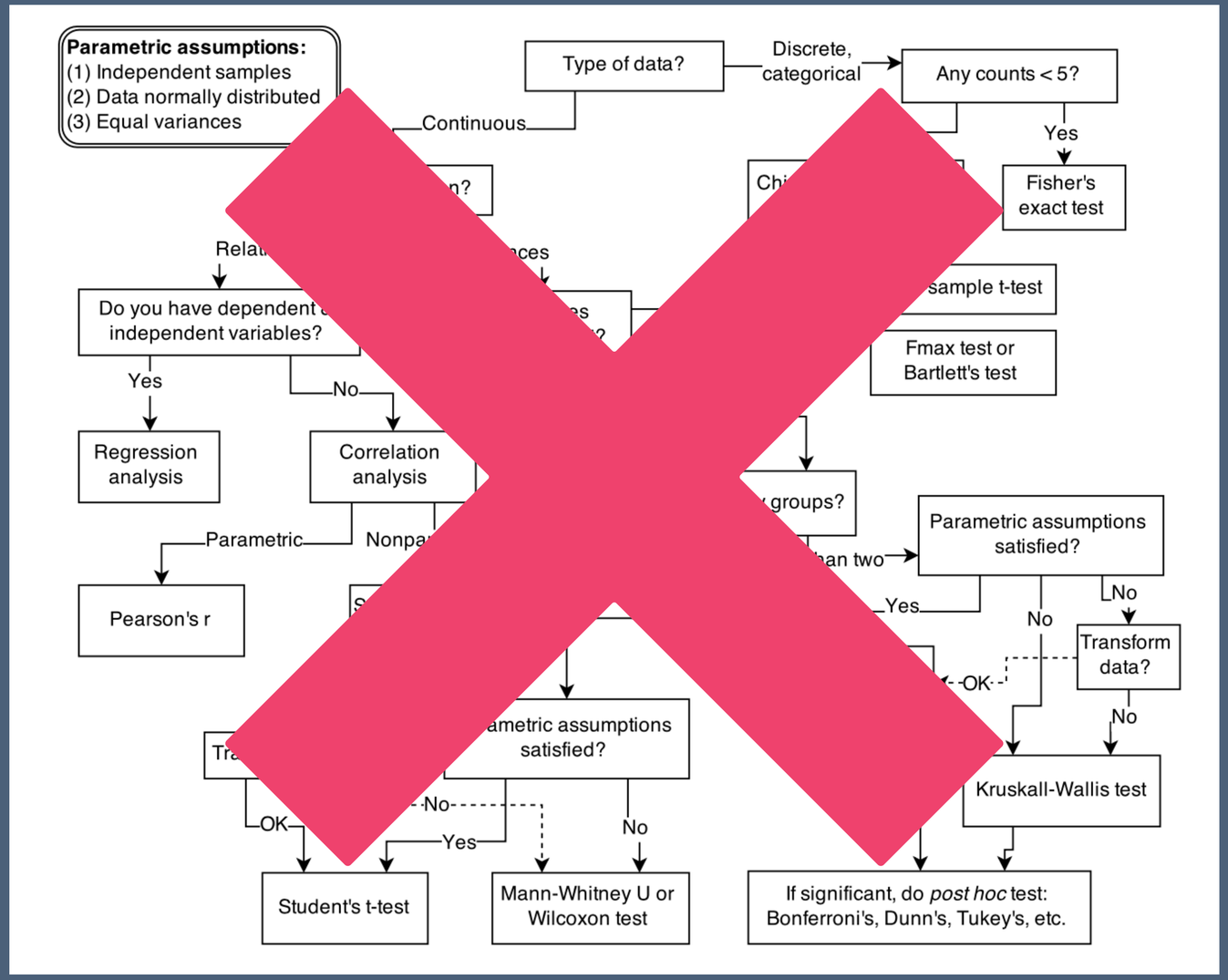
- Use a **model** to simplify
- **Theories** to summarize the understanding on how
- **Data**: measurable variables carrying uncertainty
- **Statistical models**: confront theory with data

Nossa pintura é miração, é quando você viaja no pensamento, viaja no mundo inteiro

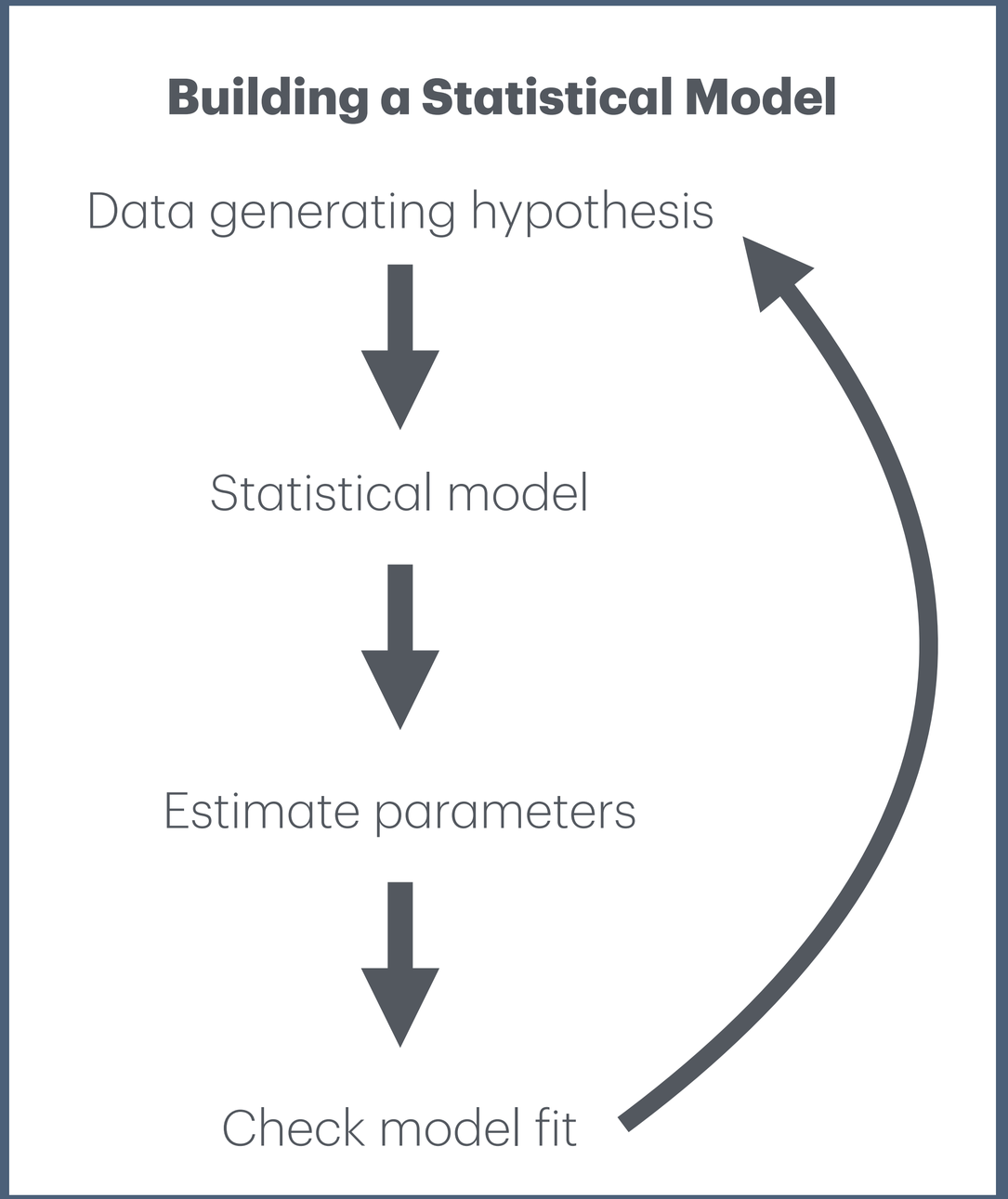
Let's move beyond recipes and think scientifically



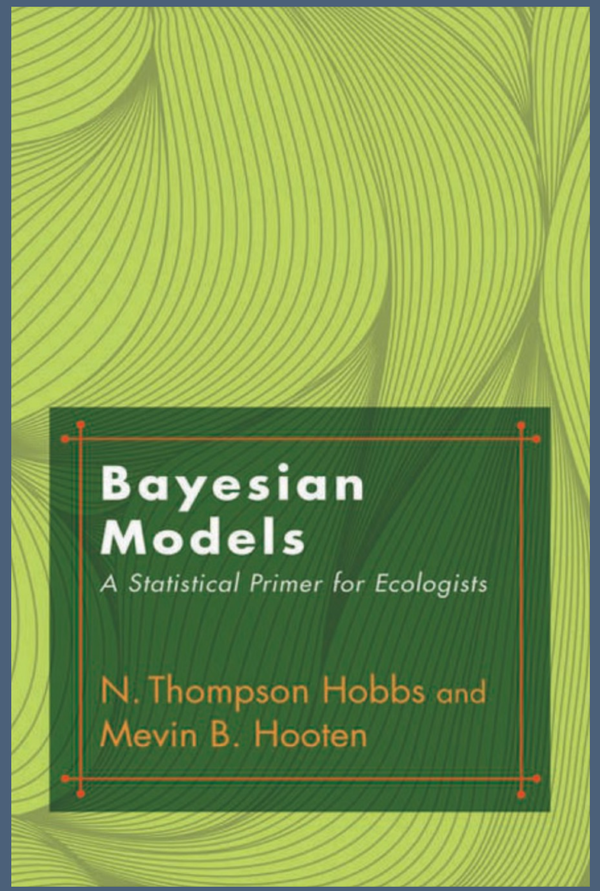
Let's move beyond recipes and think scientifically

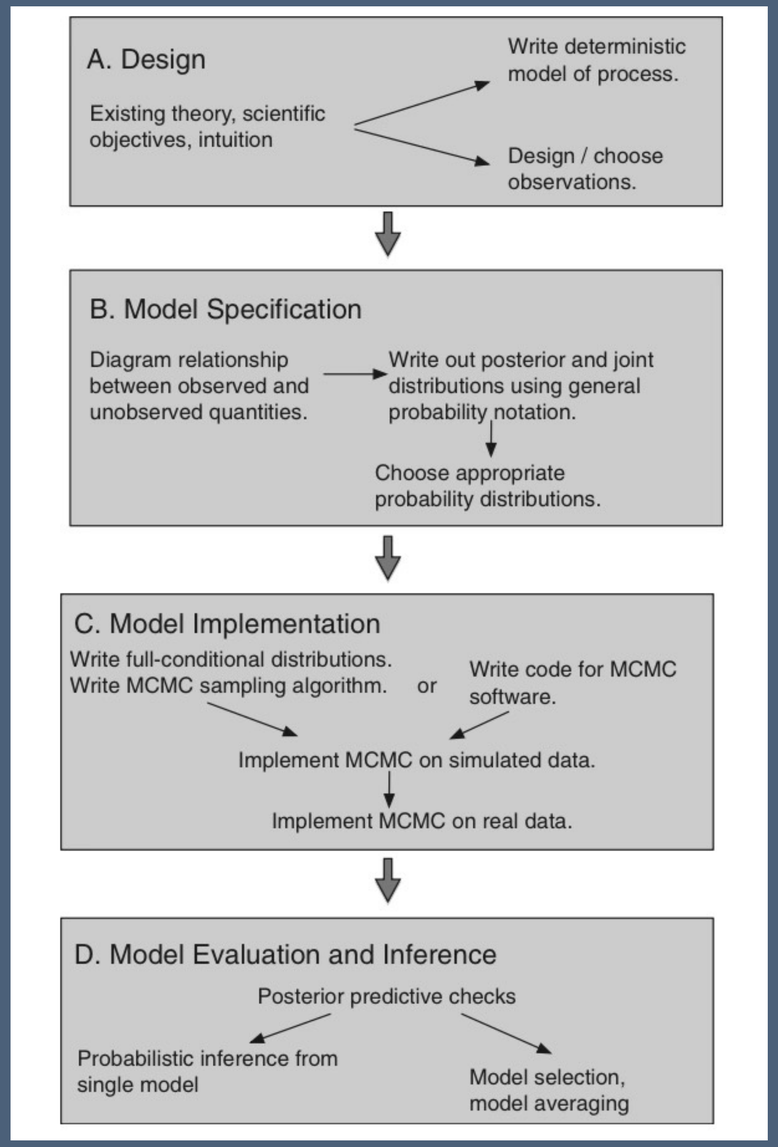


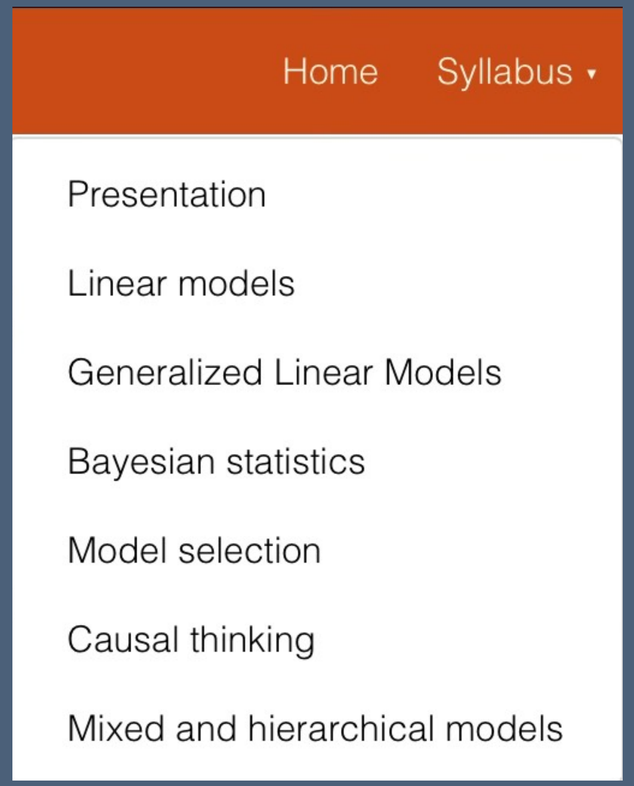
The basic model



Step by step of the modeling process







The Bayesian view

The likelihood

The probability of each value of y

- What does this mean?

$$y_i \sim N(\mu, \sigma)$$

- We can also write this as:

$$P(y \mid \mu, \sigma)$$

The likelihood of y

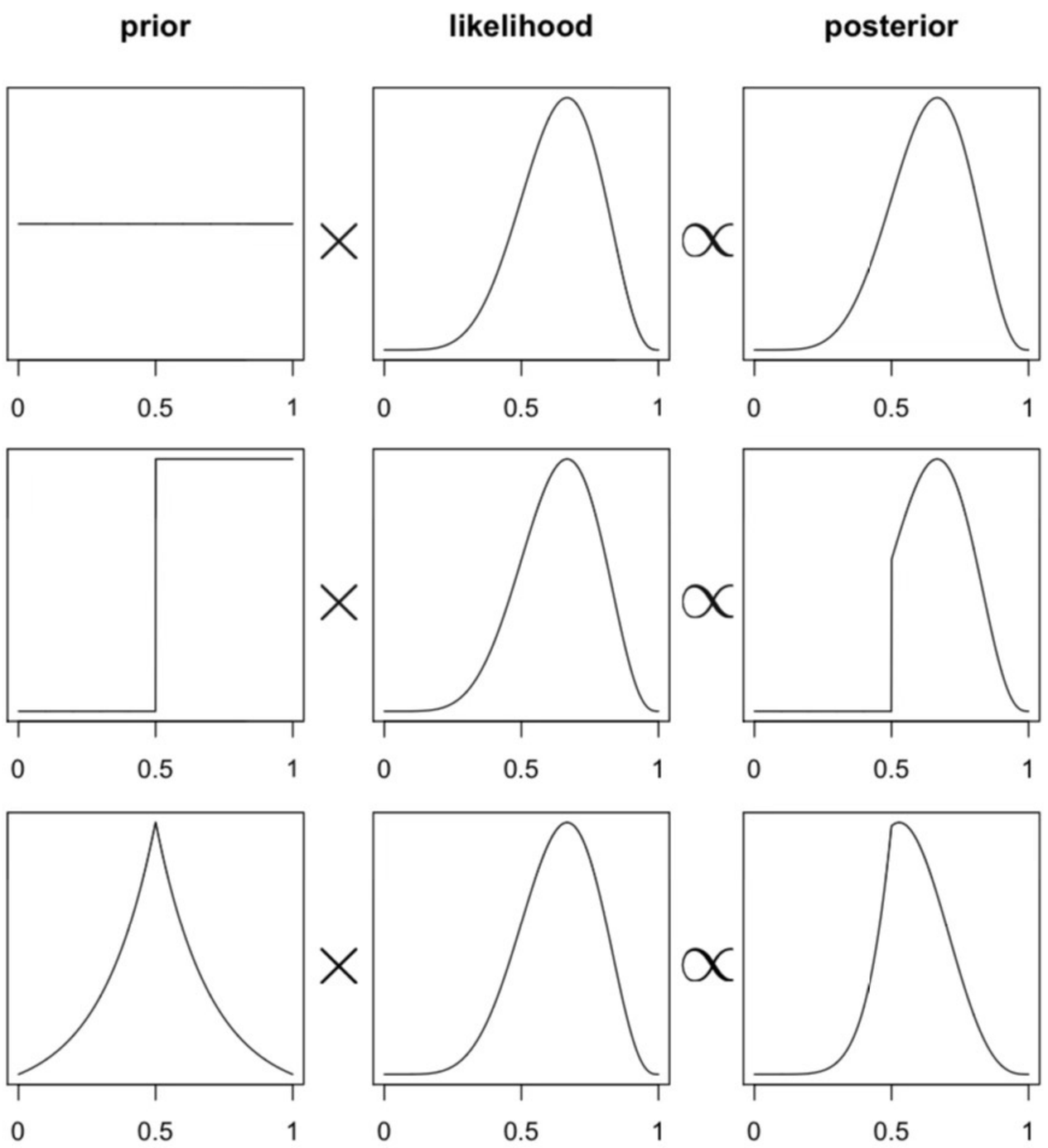
- By the product rule:

$$P(\mu, \sigma \mid y) = \frac{P(y \mid \mu, \sigma)P(\mu, \sigma)}{P(y)}$$

- $P(\mu, \sigma) = P(\mu)P(\sigma)$: the prior distribution
- $P(\mu, \sigma \mid y)$: The posterior distribution
- $P(y)$: A constant, the “evidence"

Bayesian definition of model

The posterior distribution as a product of the prior distribution and likelihood



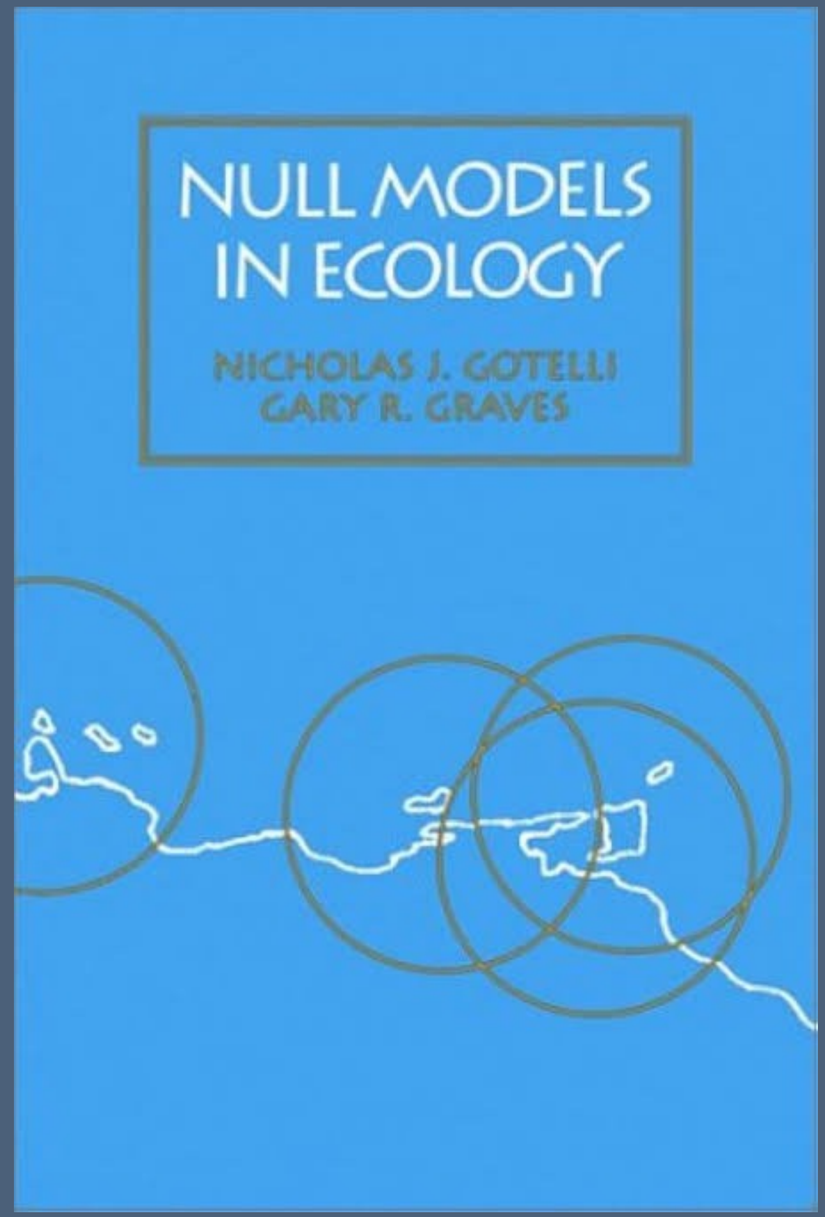
Bayesian Vs Frequentist inference in Ecology

(...) The controversy has also seemingly distracted attention from, or completely overlooked, more important issues such as the misinterpretation and misreporting of statistical methods, regardless of whether Bayesian or frequentist methods are used

Michael McCarthy Chap 1 in *Ecological Statistics*

Null models in Ecology

- Using null models to create robust inference
- Maintains some constraint from observed data (species richness, abundance distributions, ...)
- Custom-build for specific ecological questions
- Goes beyond the basic null hypothesis testing by creating a baseline expectation that is data-driven



Null and neutral are different

FORUM is a lighter channel of communication between readers and contributors; it aims to stimulate discussion and debate, particularly by presenting new ideas and by suggesting alternative interpretations to the more formal research papers published in ECOGRAPHY and elsewhere. A lighter prose is encouraged and no summary is required. Contributions should be concise and to the point, with a relatively short bibliography. Formal research papers, however short, will not be considered.

Null versus neutral models: what's the difference?

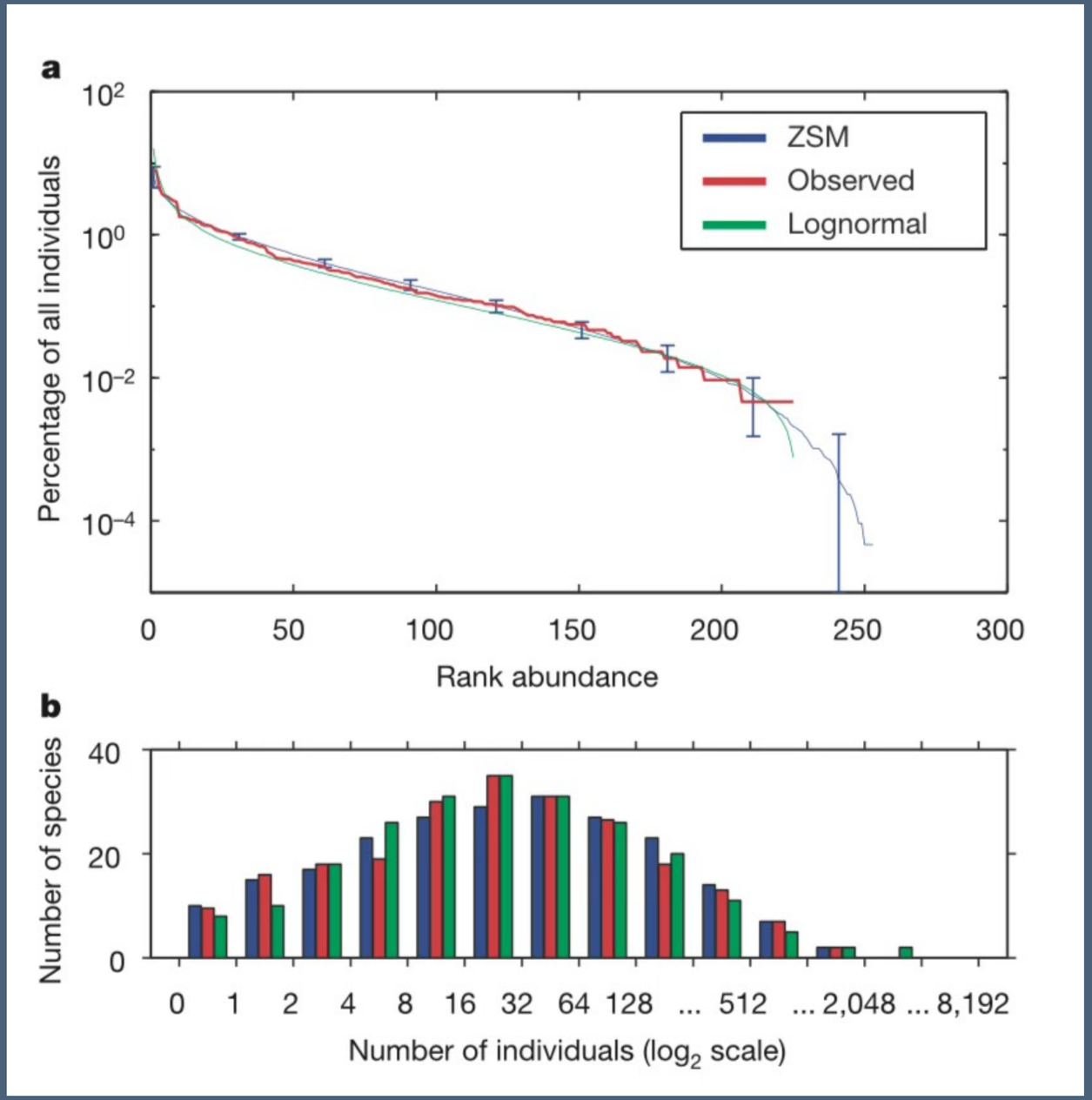
N. J. Gotelli, (ngotelli@uvm.edu), Dept of Biology, Univ. of Vermont, Burlington, VT 05405, USA. – Brian J. McGill, Dept of Biology, McGill Univ., Stewart Biology Bldg, 1205 Dr. Penfield Ave., Montreal, QC H3A 1B1, Canada

The neutral model posits that random variation in extinction and speciation events, coupled with limited dispersal, can account for many community properties, including the relative abundance distribution. There are important analogies between this model in ecology and a three-tiered hierarchy of models in evolution (Hardy Weinburg, drift, drift and selection). Because it invokes random processes and is used in statistical tests of empirical data, the neutral model can be interpreted as a specialized form of a null model. However, the application and interpretation of neutral models differs from that of standard null models in three important ways: 1) whereas most null models incorporate species-level constraints that are often associated with niche differences, the neutral model assumes that all species are functionally equivalent. 2) Null models are usually fit with constraints that are measured directly from the data set itself. In contrast, the neutral model requires parameters for speciation, extinction, and migration rates that are almost never measured directly, so their [illegible]. 3) Most neutral models are stochastic models that might be viewed as a mechanism that contributes to pattern along with other processes. Alternatively, the fit of data to the neutral model can be compared to the fit to other process-based models that are not based on neutrality assumptions. Finally, the neutral model can also be tested directly if its parameters can be estimated independently of the test data. However, these approaches may require more data than are often available. For these reasons, simple null model tests will continue to be important in the evaluation of the neutral model.

The neutral model (Bell 2000, Hubbell 2001) has generated great interest and controversy among ecologists. Some of these debates echo earlier controversies in the 1980s over null model analysis (Gotelli and Graves 1996). Indeed, Enquist et al. (2002) have claimed that the

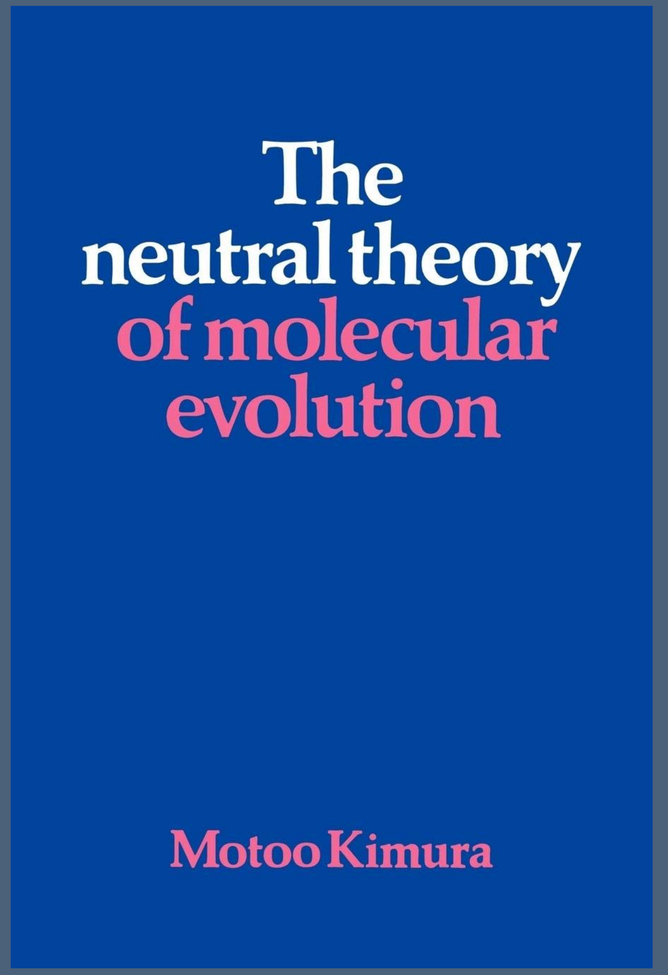
A test of the unified neutral theory of biodiversity

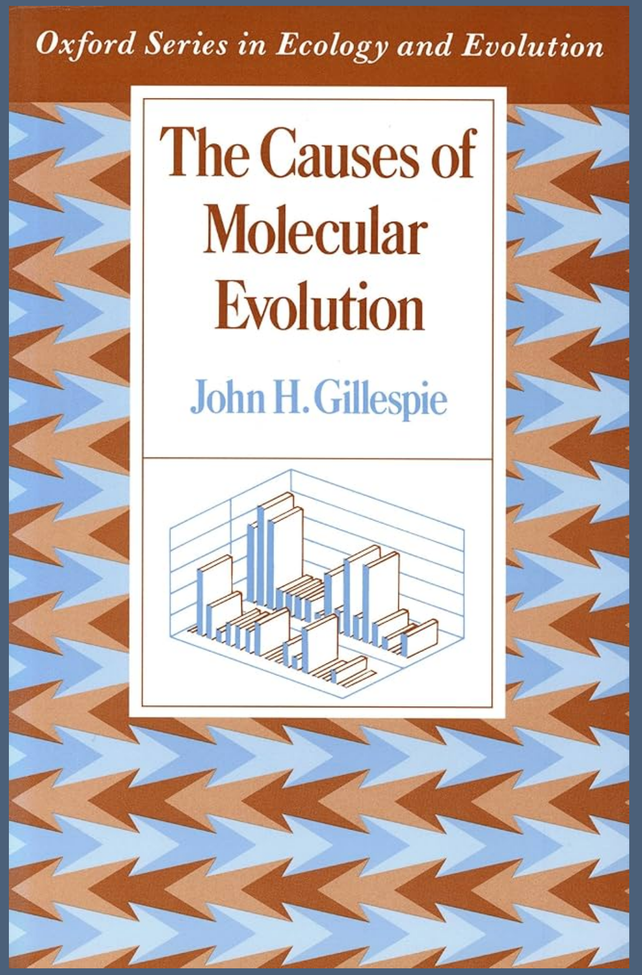
- Assumption that Zero-sum multinomial (ZSM) distribution fits better the data
- Comparison of neutral (ZSM) with a null assumption (Lognormal)
- ZSM fail to fit empirical data better 95% of the time



Null hypothesis and scientific models

- Null models are not unique:
  - What is a null pedigree?
  - Or a null population dynamics?
  - Null interaction network?
- Different hypothesis can lead to the same statistical models
  - Kimura and Gillespie debates:
    - Kimura showed that data fit the neutral model, Gillespie came up with a selective model that showed the same prediction





Permutations frequently don't work

Nulls are not unique, some structure can't be permuted away

Behavioral Ecology and Sociobiology (2022) 76:151
https://doi.org/10.1007/s00265-022-03254-x

METHODS

Common permutation methods in animal social network analysis do not control for non-independence

Jordan D. A. Hart[1] · Michael N. Weiss[1,2] · Lauren J. N. Brent[1] · Daniel W. Franks[3]

Received: 23 September 2022 / Revised: 4 October 2022 / Accepted: 10 October 2022 / Published online: 29 October 2022

Abstract
The non-independence of social network data is a cause for concern among behavioural ecologists conducting social network analysis. This has led to the adoption of several permutation-based methods for testing common hypotheses. One of the most common types of analysis is nodal regression, where the relationships between node-level network metrics and nodal covariates are analysed using a permutation technique known as node-label permutations. We show that, contrary to accepted wisdom, node-label permutations do not automatically account for the non-independences assumed to exist in network data, because regression-based permutation tests still assume exchangeability of residuals. The same assumption also applies to the quadratic assignment procedure (QAP), a permutation-based method often used for conducting dyadic regression. We highlight that node-label permutations produce the same *p*-values as equivalent parametric regression models, but that in the presence of non-independence, parametric regression models can also produce accurate effect size estimates. We also note that QAP only controls for a specific type of non-independence between edges that are connected to the same nodes, and that appropriate parametric regression models are also able to account for this type of non-independence. Based on this, we suggest that standard parametric models could be used in the place of permutation-based methods. Moving away from permutation-based methods could have several benefits, including reducing over-reliance on *p*-values, generating more reliable effect size estimates, and facilitating the adoption of causal inference methods and alternative types of statistical analysis.

Keywords Animal social network analysis · Mixed models · Node-label permutations · Permutation tests

BRIEF COMMUNICATION

doi:10.1111/j.1558-5646.2010.00973.x

POOR STATISTICAL PERFORMANCE OF THE MANTEL TEST IN PHYLOGENETIC COMPARATIVE ANALYSES

Luke J. Harmon[1,2] and Richard E. Glor[3]

[1]*Department of Biological Sciences, University of Idaho, Moscow, Idaho, 83844*

[2]*E-mail: lukeh@uidaho.edu*

[3]*Department of Biology, RC Box 270211, University of Rochester, Rochester, New York, 14627*

Received May 21, 2008
Accepted January 21, 2010

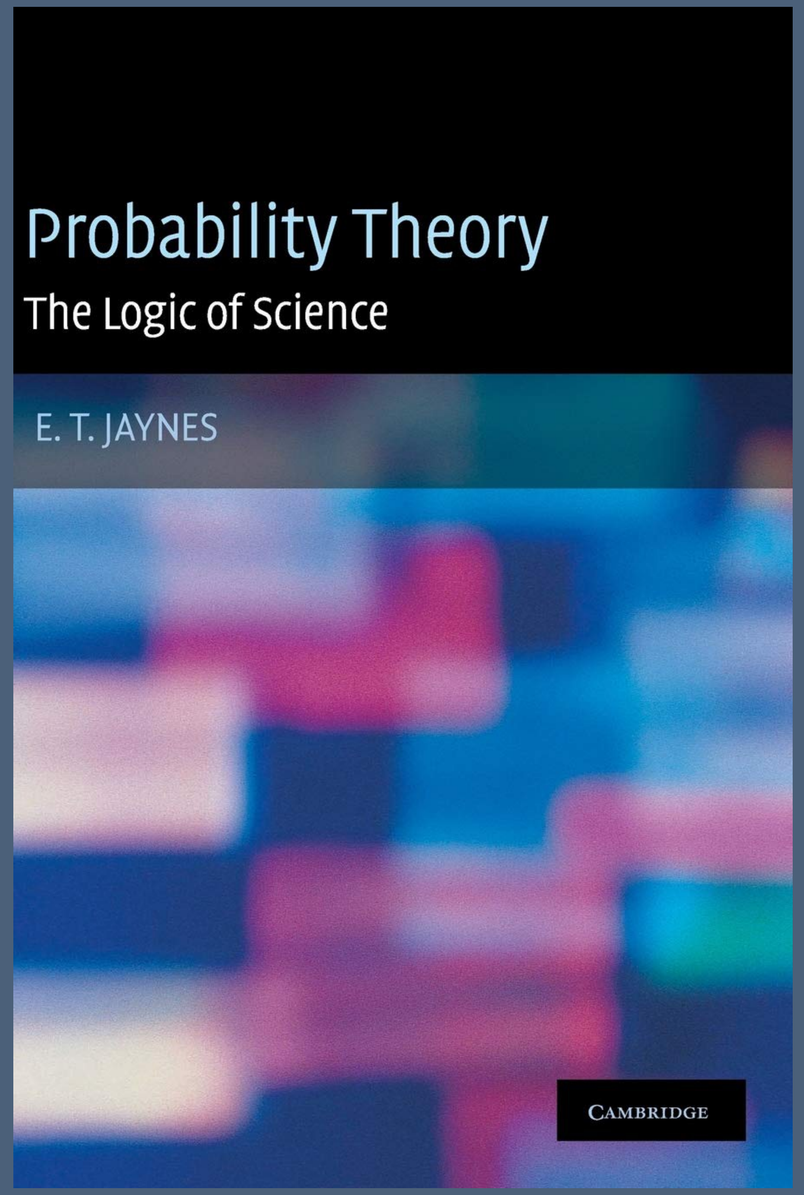
The Mantel test, based on comparisons of distance matrices, is commonly employed in comparative biology, but its statistical properties in this context are unknown. Here, we evaluate the performance of the Mantel test for two applications in comparative biology: testing for phylogenetic signal, and testing for an evolutionary correlation between two characters. We find that the Mantel test has poor performance compared to alternative methods, including low power and, under some circumstances, inflated type-I error. We identify a remedy for the inflated type-I error of three-way Mantel tests using phylogenetic permutations; however, this test still has considerably lower power than independent contrasts. We recommend that use of the Mantel test should be restricted to cases in which data can only be expressed as pairwise distances among taxa.

KEYWORDS: Comparative methods, independent contrasts, phylogenetic signal, statistical power, type-I error.

Bayes is a practical choice

Also, it's just probability theory...

- Using Bayesian methods is simply easier in real life:
  - Simple models are the same, so why bother?
  - Because complicated models are possible:
    - Real data is messy: missing data, replicates, correlated observations, mark-recapture...
- Models are **generative**, easy to simulate from and easy(er?) to build using scientific knowledge



Inference and prediction

We are most of the time doing both at the same time!

- Use the tools we have:
  - Build DAGs to understand the causal implication of the assumptions we are making about our system
  - Use cross-validation and WAIC to check model fit and predictive accuracy
  - Plot model predictions and parameter estimates
- Simulate, simulate, **simulate**!
- Start with simple models, gradually make them complex