Five points on the Frequentist vs. Bayesian debate from an evolutionary ecologist

Holger Schielzeth
Overview

1. Pressing problems for the field (and beyond)
2. $P$ values and null hypothesis testing
3. Effect sizes estimation and quantification of uncertainty
4. Specific practical issues with Bayesian analyses
5. Education and capacity building

Conclusions
Statistical issues in Ecology and Evolution
Special challenges: Diversity of species

- Mostly quasi-replication in different species
- Poor knowledge about the ecology of a species
- Widespread use of (more or less good) proxies
- Poor mechanistic understanding (far from causality)
- Poorly developed experimental setup
- Limits to sample sizes in natural populations
Special challenges: Diversity of species

<table>
<thead>
<tr>
<th>Journal</th>
<th>True replication</th>
<th>Quasireplication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Partial and Conceptual</td>
</tr>
<tr>
<td>Behavioral Ecology and Sociobiology</td>
<td>0</td>
<td>24 (33.8)</td>
</tr>
<tr>
<td>Animal Behaviour</td>
<td>0</td>
<td>28 (24.8)</td>
</tr>
<tr>
<td>Behavioral Ecology</td>
<td>0</td>
<td>31 (26.3)</td>
</tr>
</tbody>
</table>

Example I: Experimental setup

Example II: Experiments

LQ = Standard diet

HQ = Supplemented diet

Special challenges: Exploration

• Many of the analysis are exploratory
• Often merely descriptive data collection
• Sometimes motivated by sound theory
Example III: Selection analysis

Griffin et al. in prep.

Example IV: Variance-covariance

\[ \Delta z = G \beta \]

\[
G = \begin{bmatrix}
47 & .47 & .46 \\
.65 & .69 & .74 \\
\end{bmatrix}
\]


Pressing problems
Reproducibility crisis

Winner’s curse

Reproducibility crisis

Meta-analysis of temporal trends in meta-analyses

Reasons

General

• Data collection stage
  – Conditional stopping rules
  – Many small instead of few large experiments
• Data analysis stage
  – Flexible choice of response variable
  – Flexible choice of covariates
  – Removal of putative outliers
  – Flexible use of transformations
• Publication stage
  – Hypothesizing after results are known

E&E specific

• Censored data
• Transformations
• Interpretation of non-Gaussian models
• Use of proxies
• Standardizations for comparability
• Multicollinearity
Point 1

There are several influential issues that are detrimental to the field.

These are more pervasive than the difference between frequentist vs. Bayesian philosophies.
$P$ values and
null hypotheses testing
Deviation from randomness

• For (not so) far-fetched hypotheses
e.g. Strategic sex ratio adjustments
(Trivers-Willard hypothesis)

Deviation from randomness

• For experimental treatments

Example: gene mapping

Hypothesis testing plot (location across the genome)

Effect size plot (proportion variance explained)

When $P$ values are obsolete

- For variance components
  - (Autosomal) genetic component to some phenotype
- Any post-hoc hypothesis
  - $P$ values invalid
- Exploratory (backward) model selection
  - $P$ values are misused for this purpose
False positives

Main effects only

Including 2-way interaction

Winner’s curse

Effect sizes

'Significant' predictors

'Insignificant' predictors

Evidence and interpretation of $P$

**Evidence for $H_A$**

- Evidence against $H_0$ does not provide evidence for a specific theory that motivates $H_A$

**Evidence against $H_0$**

- Absence of evidence against $H_0$ does not provide evidence for $H_0$ being true

- Frequent misunderstanding
- Better training is needed!

=> Bayesian inference (Bayes factors, credibility intervals) also include an element of dichotomous decisions.
Null hypothesis tests are useful in some specific situations.

But the overall $P$ values promote binary thinking, invite incorrect conclusions and sometimes test irrelevant (straw-man) null hypotheses.

Not good for comparing multiple distinct hypotheses.
Unbiased effect size estimation
Point 3

If we are interested in an effect and we even bother to collect data on it, I think we want to know its effect size, not its presence/absence.

Effect size estimates make a study valuable for future research (including meta-analyses).
We want to know $P(\theta | D)$

- ML estimates, posterior modes
- Standard errors, posterior standard deviations
- Credibility and confidence intervals

=> Summaries of distributions
Uncertainty estimates

Uncertainty (CI)
- Parametric bootstrapping
- Profile likelihood

=> Conditional on model (estimating procedure)

Bayesian (CrI)
- Conditional on model
- Conditional on priors

- In practice so similar in most real applications
- Model structure (covariates, transformations, missing traits, hidden structure) have a far greater influence on uncertainty estimates than frequentist vs. Bayesian.
Width of uncertainty interval

- Criterion for quality of the data (at least for exploratory studies).
- Better than significant/non-significant bias in current publishing decisions
Bayesian method specifics
Choice of priors

- Little training in the community
- Weakly developed thinking in distributions
- Lack of information on what represents an uninformative prior
- Often weak (biological) theory
- Informativeness depends on interaction between prior and data
A bit of Drosophila genetics

Female

Male

X chromosomes

Y chromosome
Priors for VCV matrices:
$V_A$ – the additive genetic variance

Different scale matrices

Parameter expanded

ASReml

X lines

A lines
Priors for VCV matrices:
\( r_{MF} \) – the cross-sex genetic correlation

Different scale matrices
Parameter expanded

ASReml

X lines

A lines
Bugs in coding

• Tend to require more manual coding
• Coding is susceptible to errors by beginners and by experts
Arbitrary decisions with respect to BF

Probability distribution for competing hypotheses

\[ P(\theta|H) \]

Preference regarding unfamiliar brother

- **H₁**: Females recognize and avoid him.
- **H₂**: Females recognize and prefer him.
- **H₃**: Females don’t recognize him.
- **H₄**: Females recognize him, but don’t care.
Randomisation test: $P = 0.011$

Time spent with unfamiliar brother [%]

Female identity (ranked)

n = 64 ♀♀
Point 4

There are specific issues with Bayesian estimation procedures that are potentially dangerous, like attempts to choose uninformative priors.

Bayes factor depends on choice of $P(\theta | H)$. 
Education and capacity building
Current mistakes

• Erroneous focus on P values
  => difficult, deterring and less important

• Erroneous focus on standard tests (ANOVA, paired and unpaired t-test, non-parametric tests)
  => messy, fragmented and incoherent

• Erroneous focus on dichotomous outcomes
  => hampers accumulation and integration of evidence
Useful solutions

- Deeper discussion of (probability) distributions
- Focus on modular structure of linear models
- Stronger focus on sampling design
- Resampling strategies (as the basis for permutations and parametric bootstrapping)
- Better introduction to likelihood principle
- $t(s)$ for every $F$
- Quantification of prior knowledge
Point 6

Current stats education is quite poor (at least in E&E), both with respect to frequentist and Bayesian methods.

A unified curriculum including a stronger focus on distributions and uncertainty would be useful.
5-point summary
My five points

Several influences other than frequentist vs. Bayesian are more influential on inference.

NHT are useful in some specific situations, but the overall detrimental influence of a focus on $P$ values is undisputed.

Unbiased parameter estimation and appropriate quantification of uncertainty should be our main concern.

Specific issues with Bayesian analysis that are potentially dangerous (in particular the choice of priors).

Shift of focus in education on uncertainty and distributions.