Statistics Writing

How to write more gooder

Samuel Robinson, Ph.D.

Nov 3, 2023

Outline

- Types of scientific writing
  - IMRaD manuscripts
  - Figures and tables
- Writing about statistics
  - Translating model results
- How peer review works

Part 1: Types of scientific writing

Where do I start?

- You've finished fitting your models, and the results make sense to you, but...
- How do I translate all these numbers into "real" English?
- Where do I put all these numbers in the paper?
- Do I need figures and tables?

Answer: “It depends”

What is your story? Who is your audience?

- How do these numbers serve the questions I’m asking?
- Do these numbers help my audience to understand what I found?
- Would figures or tables help to prove my point more concisely or easily?
- How do these numbers relate to the rest of the literature?

A bit of history

- (European) Universities are largely offshoots of the Christian monastic tradition
- What we now call science started in about the 1600s, largely as offshoots of astrology and alchemy
  - Biology began slightly later (1700s-1800s), as offshoots of medicine and natural history
- "Natural philosophers" (scientists) would write letters to each other about what they were up to
- Eventually, organizations of scientists began publishing research results publicly (e.g. *Philosophical Transactions of the Royal Society*, 1665)
- Peer review was sparse, and was usually done by the editor or a board. External peer review wasn't widespread until 1950-1970
- Early science writing is *extremely* varied, and is much different from modern science writing

What is science writing for?

- "Recording secret knowledge" (Newton)
- "Describing *exactly* how an experiment proceeded" (Bacon)
- Modern science writing does mostly the latter

Common types of scientific writing

1. IMRaD papers: "standard" scientific papers
   - Introduction, Methods, Results, and *Discussion*
2. Meta-analyses
3. Review papers

How does this relate to statistics?

- Early use of statistics in science was fairly "vibes-based", at least until computers became more readily available (1950s onward)
  - Not necessarily a bad thing!
- More complex and extensive data collection requires more complex modeling approaches
  - Trade-off between realism and "explainability"
- Pushback from some quarters: One aspect of the ongoing replication crisis
  - *Statistics are political*

"I have heard from graduate students opting out of academia, assistant professors afraid to come up for tenure, mid-career people wondering how to protect their labs, and senior faculty retiring early, all because of methodological terrorism" - Susan Fiske, APS Past President

"[Fiske is] seeing her professional world collapsing... her work and the work of her friends and colleagues is being questioned in a way that no one could've imagined ten years ago. It's scary, and it's gotta be a lot easier for her to blame some unnamed"terrorists" than to confront the gaps in her own understanding of research

Part 2: Writing about statistics

Models as evidence for arguments

- Scientific discourse can be thought of as a series of logical arguments
- When making an argument, you bring evidence to support your claims
- We use experiments/observations, mathematics, and previous literature to support our claims
  - None of these are assumption-free: The reader must be convinced that these are appropriate!
- Models also act as a *piece of evidence*, translating raw data into "ammunition" for your claim
  - Model structure and performance checks (residual plots, etc.) should *also* convince the reader that this is believable

Show the *bare minimum* number of statistics needed to convince people. If it's not relevant to your story, move it somewhere else.

Example arguments:

- Premise 1: Bats eat bugs
- Premise 2: Forests have lots of bugs
- Claim: Therefore, bats should prefer forests [1]

Example 1:

- Evidence: The model of my data **supports this claim**
- Conclusion: This means that our understanding of bugs, bats, and forests is pretty good

Example 2:

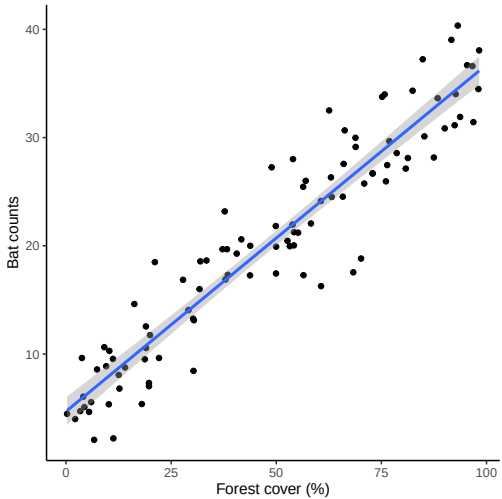
- Evidence: The model of my data **does not support this claim**
- Conclusion: One of these premises is wrong, or we left out an important premise

[1] Inductive reasoning

Models as reflections of reality

- Models are meant to reflect an *underlying biological process*
- Things like effect size (mean/SE) reflect the relative strength of the factors involved
- Things like R^2 reflect how well the model fits the data *overall*
- Causality is implied, but has to be justified

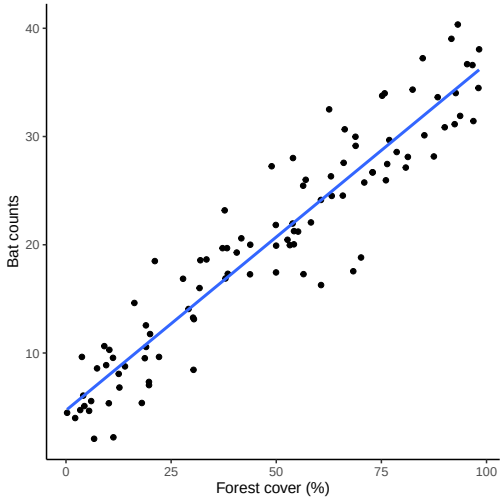
"Keep your eye on the biology!"



What might the underlying physical process be here?

Evidence type 1: coefficients

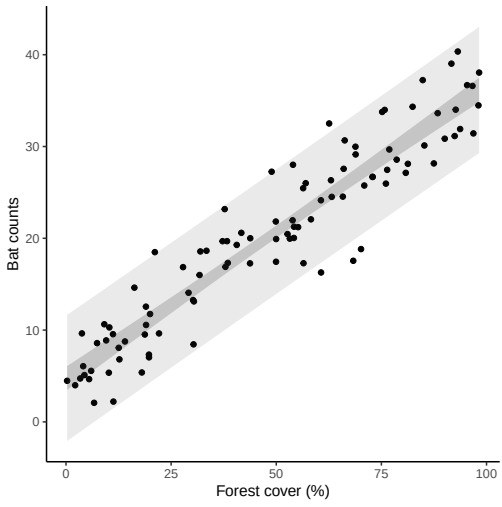
- Slopes and intercepts have physical interpretations
  - Intercept: How many bats at 0 % forest?
  - Slope[2]: $+$ 1 % forest $=+$ 1 bat
- Interpretation can be:
  - Yes/no: "Is there any relationship?"
  - Directional: "Is the relationship positive?"
  - Magnitude: "How big is the slope?"



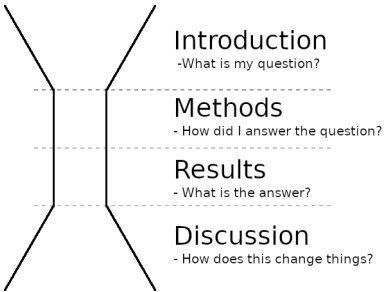
[2]For GLMs, slopes are in log or log-odds (logit) units

Evidence type 2: variance

- Variance has a physical interpretation
  - What is the variation in bat counts at a given level of forest?
- R^2 relates actual to modeled variance: what % of variance does your model explain?
- GLMs: different distributions model variance differently
- Hierarchical models deal with many levels of variance
  - Tells you where the variance in your system is coming from



IMRaD Paper Structure



- Most scientific papers follow the IMRaD canon
- Allows the reader to quickly assess whether this paper is useful and skip to important sections only
- Generally, statistics are discussed in the *Methods* and *Results* sections only

Introduction

I find this section the trickiest to write, so I usually write it last.

- Set up your research question, using the literature
  - Moves from general ("Animals need food") to specific premises ("Bats need bugs")
- Establish the *knowledge gap* that your research will address
  - "Forest have lots of bugs, but nobody has checked whether there are bats there too!"
- Last paragraph: strong statement that sums up what you're expecting to see
  - "Bats eat bugs, and forests have lots of bugs. Therefore, bats should prefer forests."

Methods

I usually write this section first, as it gets me "warmed up" for the rest of it.[3]

- Establish how you collected the data, and how you analyzed it
  - This defends against criticism of your model or your data, and makes your results more believable
- The detail you use depends how "unusual" your model is, which depends on your audience
- Clarify what the dependent, independent variables, and random effects in your models are
- Sometimes you can just use the actual R model formula:
  - "I fit the model using `lme4` in R using the following model structure for bat counts (while accounting for frogs and unicorns):"

```
lmer(batCounts ~ forest + frogs + unicorns + (1|site))
```

[3] You can even write it before you collect your data!

Results

I write this section after I write the Methods section

- Brief summary of what you collected[4]
  - "I caught 420 bats at my 69 sampling sites."
- Present your results as an answer to the questions that you posed in the Introduction.
  - "Forest cover caused an increase of 3 bats for each 10% of forest ($p<0.001$), while frogs had no effect ($p=0.7$)"
  - Try to keep the language as normal and direct as possible
  - Having tons of p-values and other numbers can make the text hard to read
- If something weird happened, just say it and move on. Speculate on *why* in the Discussion.
  - "Surprisingly, unicorns had a negative effect on bat counts."

[4] Can sometimes go at the end of the Methods

Discussion

I write this after my model Results. Here you can name-drop all the relevant papers you've read.

- Relate your results to your research question. Did your results match your expectations?
- Move from specific ("Bats need bugs") to general ("Animals need food"); opposite of the Introduction
- Put the Results you found into the context of the rest of the literature. If your results contradict other studies, why do you think that occurred?
  - "Barclay et al. (2017) showed that bats don't like forests, but our results may differ because..."
- **So what?** What new things have we learned? How might this affect theory or practice? Should non-bat people pay attention to this paper?

Figures and Tables

- Figures can be excellent tools for telling your story, but...
  - Figures take up lots of room, cost $ in publications, and can overwhelm the reader if there are too many
  - Many resources for good figure design
- Tables are kind of boring, but are great for conveying lots of numbers at once
  - Useful for showing information on large numbers of coefficients
  - If you have lots of models, `library(broom)` provides summaries of all of them at once
- Tables and figures (+ captions) should be readable without knowing the rest of the text

Suggestions:

1. Choose 2 or 3 figures and tables to be the **Main Characters** in your Results section.
2. Use them to illustrate what your models show.
3. Move the rest into a supplemental or appendix.

Final remarks

- Good writing is re-writing
  - What is obvious to you may not be obvious to your readers. Revision is annoying and painful, but it **will** help!
  - "[Good writing is:] Telepathy, of course" (Stephen King)
  - I find writing in point form useful, as it lets me hash out the general paragraph and section structure
- Use the literature
  - There are tons of poorly-written papers out there, but...
  - Was there a paper that you found easy to understand? Re-read it, and figure out why!
  - Check out how other scientists display their statistics, and imitate/avoid their style
- Use your supervisor and committee members
  - They have a much wider picture of the field, and have lots of writing and editing experience
  - This means that they can be a good stand-in for your audience

Remember: GOOD WRITING IS GOOD STORYTELLING

Practice

Pull up a blank text document. Using a model that you have fit to some data, write:

- Description of your research question for the Introduction. What do you expect to see, and why?
- Short description of the model for the Methods
- Short summary of what the model is telling you for the Results

Point form is fine, but try to find important points that you could convert into topic sentences in an actual paragraph