A few suggestions for writing a thesis or research article in statistics

Jonas Peters and Niklas Pfister

January 20, 2023

The ability to clearly communicate research in written form is necessary in order to actively contribute to science. This article contains a collection of rules, suggestions and comments that can be helpful when starting to write a thesis or research article. We believe the key to good scientific writing is to (1) put yourself into the shoes of the intended readers and to (2) be concise and consistent. Some parts of this article are based on our personal preferences. It is fine to disagree with them – but if you deviate from these (or other) suggestions, it should always be a conscious decision.

1 Writing guide

1.1 General rules

- Throughout the writing process, you should always think about the reader and ask yourself: What can I do to help her/him understand my work better?
- The most important thing is to explain the problem you are trying to solve. If the reader does not understand the problem she will stop reading.
- Do not write a description or report of what you did (and when you did it). This is usually boring to read and didactically suboptimal. Rather focus on making the reader understand the content of the paper (the problem, the ideas, the results etc).
- If the paper makes use of code, this should always be publicly available and published under a license (e.g., github allows you to choose "GNU Affero General Public License v3.0").

1.2 Structuring

- The structure (sections, subsections etc) is important. Invest some time in coming up with a good structure and discuss the structure with others.

- The following three part structure is a good starting point:
  - *Introduction:* This is generally the first section and should motivate your problem and relate it to existing work. It is common to write this section last and collect (e.g., as bullet points) topics that you want to mention while writing the rest.
  - *Main part:* This can consist of multiple sections and the structure should help the reader. Common components of this part are required background material, model/method description, theoretical results, numerical simulations.
  - *Conclusion/Discussion:* This is generally the last section and should summarize the main contributions/findings, discuss potential shortcomings/weaknesses and point to future directions of research.

1.3 Mathematical formalism

Mathematical formalism is a way of expressing abstract concepts in an unambiguous way. When writing mathematical expressions it is therefore particularly important to be as precise as possible.

Rules on predicate logic A predicate is the formalization of the concept of a mathematical statement. A statement is understood as an assertion that, depending on the value of the contained variables, evaluates to being either true or false. Predicates consists of a concatenation of formulas that are connected via logical connectives such as negation (text: "not"; symbol: "$\neg$"), logical conjunction (text: "and"; symbol: $\wedge$), logical disjunction (text: "or"; symbol: $\vee$), existential quantification (text: "there exist(s)"; symbol: $\exists$) and universal quantification (text: "for all"; symbol: $\forall$).

- Always use the quantors "for all" and "there exists" in front of a predicate. Generally, we use the written form of these quantors and only use the symbolic forms "$\forall$" and "$\exists$" in display expressions if it increases clarity.
- Do not use expressions such as "for", "for some", "for each" or "for any" – they can be ambiguous.
- Try to form full sentences, even if you use a formula in an `equation` or `align` environment. Ergo, sometimes one need to put "." or "," after an equation.
- Avoid expressions like "for all $m \geq 0$" as it is not entirely clear whether it means "for all $m \in [0, \infty)$" or "for all $m \in \{0, 1, 2, \ldots\}$". Instead use the more precise version.
- Always use the quantors in front of the predicate. For example, write:

  For all $x \in \mathbb{R}$ there exists $y \in \mathbb{R}$ such that for all $t \in [0, \infty)$ it holds that

  $$e^x \leq y\, e^t.$$

This avoids ambiguity as illustrated by the following example:

For all $x \in \mathbb{R}$ there exists $y \in \mathbb{R}$ such that

$$e^x \leq y\, e^t$$

for all $t \in [0, \infty)$.

Problem: It is not entirely clear whether this means "$\forall\, x \in \mathbb{R}\colon \exists\, y \in \mathbb{R}\colon \big(e^x \leq y\, e^t \,\forall\, t \in [0,\infty)\big)$" or whether it means "$\forall\, x \in \mathbb{R}\colon \big(\big(\exists\, y \in \mathbb{R}\colon e^x \leq y\, e^t\big) \,\forall\, t \in [0,\infty)\big)$".

General mathematics rules

- Try to make the paper as self-contained as possible. In particular, define non-standard concepts (i.e., anything the reader cannot be expected to know without looking it up in a specific paper) directly in the paper.
- Formal results (such as theorems, lemmas, propositions and corollaries) should be self-contained, too. For example, it should be clear what every term means or where it can be found. Often, it helps if you write such results in a two part structure. First, list all assumptions and secondly state the result.
- Make sure that every formal result has a proof that is easy to find for the reader. In statistics, it is common to relegate proofs to the appendix to avoid breaking the flow of the paper.
- Proofs should have a clear uniform structure and every step (even the ones you find obvious) should be clearly argued. It is good practice to end each proof we a sentence like: "This completes the proof of Proposition 2."
- Equations should only be numbered if you are referring to them later.

1.4 LateX and BibTex

Everyone has there one preferences when it comes to LaTeX. Here, we list some basic style rules.

- Avoid using '\\' and '`\newline`' – leave a free line instead.
- Avoid free lines before and after figures, propositions, theorems.
- Avoid free lines after section.
- LaTeX adds extra space after '.'. Thus use '\' after a '.' that is not a full stop, e.g., 'choose xx s.t.\ something holds'.
- If you refer to Section 3, for example, use '`Section~\ref{sec:starting}`'. In particular, use capitals. The '~' prevents linebreaks.

- Clean up warnings and errors already from the start.
- Be consistent with formatting choices: e.g., what do you put in italic?
- Figures:
  - Use vector graphics (e.g., PDF or SVG) for plots instead of more classical image formats such as JPEG or PNG.
  - Try to adapt the formatting of plots (caption, title, labels, etc) as much as possible to the format you use in the text.
  - Always add captions and make these as self-contained as possible. In particular, include interpretations, e.g., "In all experiments, the empirical type I error stays under 5%." Do not worry about having a bit of redundancy in the figure captions. Many people first look at the figures to decide whether they read the document.
  - Make sure that your figures are well-positioned in the final document. Do not do this until you have finished everything else though.

References are managed by BibTeX within LaTeX. We suggest using the following setup (but other options are also possible):

- Use the natbib package ('`\usepackage{natbib}`'), this defines the commands '`\citep`' for bracketed citations (e.g., "It has been shown that 4 is a great number [Hansen et al., 2012].") and '`\citet`' for in-text citations (e.g., "Hansen et al. [2012] show that 4 is a great number."). If possible – in particular for single citations – it is preferable to use in-text citations.
- You can use square brackets in citations to add more details. '`\citet`' allows for one argument (i.e., '`\citet[Theorem~12]{hansen12}`' leads to Hansen et al. [2012, Theorem 12]), while '`\citep`' allows for two arguments, one before and one after the citation (i.e., '`\citep[e.g.,][Theorem~12]{hansen12}`' leads to [e.g., Hansen et al., 2012, Theorem 12]).
- Make sure the references are complete and consistent. In particular, check capitalization, abbreviations of first names, etc).

1.5 Language

Written English can be difficult in particular for non-native speakers. It is therefore advisable to spend some time revising a text to improve the English. Below we list a some guidelines that might help in the process. We also recommend that you think about the following two points for each paragraph:

- Does this paragraph make logical sense and does it help convey the intended message?
- Can the paragraph (or parts of it) be changed to make it shorter an clearer?

Neither of us are experts in written English and the points we have collected here are a summary of suggestions we have gotten.

Less is more

- Shorten all structure words:
  - due to the fact that → because
  - in view of the above → therefore
  - in the course of → during
  - the fact that → that
  - in relation to → about
  - for the purpose of doing → to (do)/ for (doing)
  - until such time as → until
- Eliminate all redundancies:
  - various kinds of theories → various theories
  - in the month of August → in August
  - to a large extent → largely
  - in a professional manner → professionally
  - in the amount of 500 Euro → for 500 Euro
  - fear and trembling → dread
  - each and every one → each one/ every one
  - aggregate together → aggregate

Some general rules

- Usually, there is no comma before if.
- Use present tense when using citations "Müller et al (2019) show how to...".
- Expressions like "note that" or "observe that" can almost always be left out; instead: improve the (logical) connections between sentences.
- Avoid subjective language; avoid 'very'.
- Avoid "it is" and "there is/are". For example, instead of "It is shown by Theorem 4 that ...", write "Theorem 4 demonstrates that...".
- Avoid "It is argued that..." and "It is obvious that...".
- Do not use "this" (or "these") by itself. Add the appropriate noun. For example: "This discrepancy led to...".

- "which" versus "that":
  - Use "which" (with comma) for nonessential material, i.e., material that could be left out without breaking the meaning of the sentence.
  - Use "that" (without comma) for essential material, i.e., material that could be left out without breaking the meaning of the sentence.
- Prefer active over passive voice. For example, use "The findings appear in Figure 1" instead of "The findings are presented in Figure 1".
- Avoid the future tense when describing the content appearing later in the paper, e.g., "we show below" instead of "we will show below".

Acknowledgements

Many of the above comments are not due to us. We thank everyone who has contributed ideas, including Arnulf Jentzen, who recommended some of the suggestions on how to present mathematical statements to the authors, and Nathalie Reid, whose writing course we attended.

References

James Hansen, Makiko Sato, and Reto Ruedy. Perception of climate change. *Proceedings of the National Academy of Sciences*, 109(37):E2415–E2423, 2012.

W. Rautenberg. *A concise introduction to mathematical logic*. Springer, 2010.