Advanced Research Writing in Statistics and Data Science

Kun Chen

2025-08-21

Contents

Preface

This book aims to train students in statistics and data science on academic writing with professional tools such as LaTeX, BibTeX, R, and Git.

This book is currently under-development, and it is meant to be an extended version of the book "Statistical Writing" https://github.com/statds/stat-writing written by Dr. Elizabeth Schifano and Dr. Jun Yan.

The notes are prepared with the **bookdown** R package [**?**], which can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 1

Introduction

What does a statistical or data science paper look like? As with all scientific papers, it should have some commonly expected structures: title, abstract, keywords, introduction, methods, results, discussion, acknowledgements, references, appendix, and supplement. Due to the specificity of the statistical discipline, machine learning practices, and application domains, however, research papers can look very different from one another.

1.1 Types of Statistical Papers

1.1.1 Theory papers

A **theory paper** in statistics and probability is closest in form to a mathematical paper. It typically includes the statement of the problem, formulation of assumptions, and the presentation of theorems, lemmas, and proofs. The purpose is often to establish fundamental properties of certain statistical or probabilistic tools/approaches, including but well beyond consistency, efficiency, optimality, convergence, asymptotic distributions, and non-asymptotic error bounds.

While theory papers may not always feature data, simulations, or applications, they form the **mathematical and probabilistic foundation** upon which methodology and applied work are built.

Most of the articles in journals such as *Annals of Statistics*, *Annals of Probability*, *Bernoulli*, *Probability Theory and Related Fields*, among others, are considered as theory papers. In other words, these journals should be the primary outlets for a theory paper.

Examples include **?**.

1.1.2 Method papers

A **method paper** focuses on proposing a novel methodological contribution that can be applied to a general class of real-world problems. A commonly seen structure is:

- Introduction
- Methods (e.g., estimation, hypothesis tests, diagnostic procedures)
- Theoretical properties
- Simulation studies
- Applications/illustrations
- Discussion/Conclusions

Such papers often include a blend of theory (e.g., asymptotic guarantees) and empirical validation. See, for example, **?**; **?**.

1.1.3 Applied papers

An **applied paper** focuses on addressing a concrete scientific question in a particular domain using statistical or data science methods. Its structure often includes:

- Introduction
- Data description
- Methods (applied or adapted)
- Results
- Discussion

Sometimes simulation studies are added to assess sensitivity. Applied papers may involve applying existing methods or developing new ones motivated by the application. Examples include **?**; **?**; **?**.

1.2 Data Science and Machine Learning Papers

Beyond traditional statistics journals, researchers often publish in **data science and machine learning outlets** such as *NeurIPS*, *ICML*, *KDD*, and *AAAI*. These papers emphasize:

- Conciseness (strict page limits)
- Algorithmic novelty
- Benchmark comparisons on standard datasets
- Open-source reproducibility
- Clarity for an interdisciplinary readership

Compared to *JASA* or *Annals of Statistics*, these outlets prioritize **empirical performance and novelty** over lengthy theoretical development, though many still include core mathematical analysis.

1.3 Audience and Purpose

An author should always keep the **target audience** in mind. Statistical journals span a wide spectrum from applied to theoretical. Machine learning venues differ again in expectations. Even technical reports, white papers, or grant proposals have unique readerships. Regardless of outlet, any scientific writing should be convincing and concise. Authors must show that their work is important, valid, and useful — and avoid wasting the reader's time.

1.4 Scientific Writing Resources

Many resources on scientific writing are available. For example, **?** was selected by its publisher, *American Scientist*, as one of its 36 "Classic Articles" from the first 100 years of publishing history. Popular books are **?**, **?**, **?**, and **?**.

1.5 About This Book

This book, *Advanced Research Writing in Statistics and Data Science*, extends earlier efforts such as *stat-writing* by broadening the scope to include not only statistical research writing but also writing for **data science and machine learning communities**.

The book emphasizes three guiding principles:

1. **Clarity and Persuasion** – communicating complex technical ideas to both expert and interdisciplinary audiences.

2. **Genre Awareness** – tailoring style and structure to journals, conferences, and other outlets.

3. **Learning by Doing** – the book includes examples, annotated excerpts, and practical exercises.

1.6 Outline of the Book

The planned content includes:

- **Introduction** (this chapter): Types of research papers and outlets.
- **Clarity and Conciseness**: Avoiding jargon, redundancy, and ambiguity.
- **Titles, Abstracts, and Introductions**: Crafting compelling entry points.
- **Methods and Results**: Presenting technical material effectively.
- **Figures, Tables, and Captions**: Communicating visually.
- **Discussion and Conclusions**: Framing contributions and limitations.
- **Peer Review and Revision**: Writing reviews, responding to reviewers, revising.
- **Grant Proposals and Research Statements**: Writing beyond papers.
- **Ethics and Integrity**: Authorship, plagiarism, and responsible use of AI.
- **Exercises and Practice**: Rewriting, reviewing, and polishing tasks.

Each chapter will include **examples drawn from real papers across statistics, biostatistics, and machine learning** and will feature **exercises** that mimic the real writing and review process.

Chapter 2

Introduction

What does a statistical paper look like? As with all scientific papers, it should have some commonly expected structures which include components such as title, abstract, keywords, introduction, methods, results, discussion, acknowledgements, references, appendix, and supplement. Due to the specificity of the statistical discipline and application areas, however, statistical papers could look quite different one from another.

There are different types of statistical papers. A theory paper would look similar to a paper in mathematics, with statement of the problem, presentation of some theorems, and technical proofs. Such papers are not covered here. We focus on two types of statistical papers: application papers and method papers. Application papers focus on a specific application problem in a certain domain where the research discoveries depend on applications of existing or novel statistical methods. Method papers, on the other hand, aim to provide a general methodological solution to a class of applied problems. Often, methods paper have a theoretical component, for example, the establishment of the asymptotic properties of a new estimator. An applied paper in statistics could be a method paper in the domain of the problem it solves.

An author should always keep the target audience in mind when writing. There are many statistical journals on the wide spectrum from applied to theoretical papers. Each one has its own aims and scope, with different target readerships. Writings such as customary statistical reports that are not intended for journal publications also have target readerships. Regardless of the audience, any scientific paper should be convincing and concise. You need to show the readers that your work is important, valid, and useful. You don't want to waste the time of any readers.

2.1 Applied papers

An applied paper has a widely accepted structure:

- Introduction
- Data description
- Methods
- Results
- Discussion

An applied paper can be applying existing statistical methods to solve an applied problem. See, for example, **?**; **?**.

When sensitivity analysis is desired for the applications, one can have a section on simulation studies. See, for example, **?**; **?**.

Some applied papers can involve novel methodology development that is motivated by an applied problem. In such cases, simulation studies become necessary, where you validate your methods with simulated data so you can check your estimator with the truth. Such check is not feasible when analyzing real data. See, for example, **?**; **?**.

2.2 Methods papers

A methods paper focuses on a novel method that is applicable to a general class of problems arising in different domains. A commonly seen structure is:

- Introduction
- Methods (e.g., estimation, hypothesis tests, diagnosis)
- Properties
- Simulations
- Illustrations (with real applications)
- Discussion/Conclusions.

The simulations section is often needed for methods papers. Any method has assumptions. Any reasonably good method should work as expected when the assumptions hold. It would be even better if it remains working when some of the assumptions are violated. Simulation studies can be designed to check whether the proposed estimators are unbiased and more efficient than competing estimators; whether the proposed tests retains their sizes and are more powerful than competing tests.

Here are some examples: **?**; **?**.

2.3 Scientific Writing

Many resources on scientific writing are available. For example, **?** was selected by its publisher, *American Scientist*, as one of its 36 "Classic Articles" from the

first 100 years of its publishing history. Popular books are **?**, **?**, **?**, and **?**.

Chapter 3

Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

3.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
    - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world` = `# Hello world {#hello-world}`.
    - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter 3.
    - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

3.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 3.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Don't miss Table 3.1.

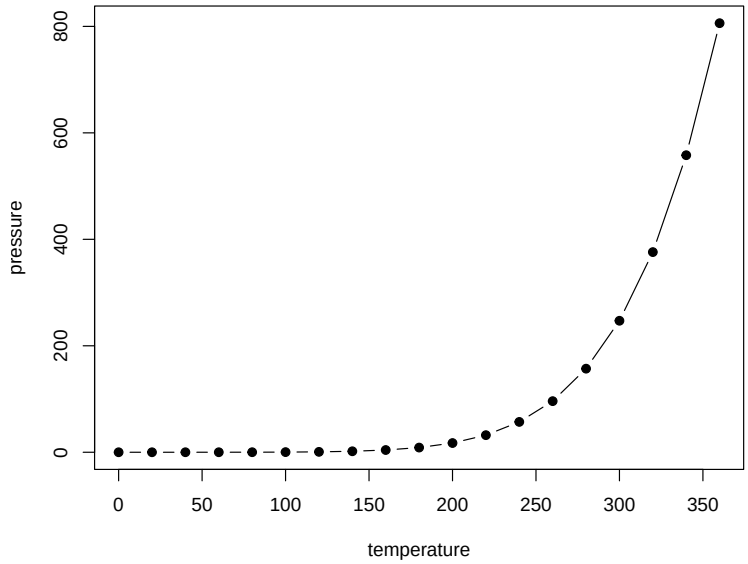

Figure 3.1: Here is a nice figure!

```
knitr::kable(
  head(pressure, 10), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

Table 3.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 4

Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART\*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 5

Footnotes and citations

5.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

5.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package [**?**] (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** [**?**] (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 6

Blocks

6.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (6.1).

6.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 6.1.

Theorem 6.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

6.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 7

Sharing your book

7.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

7.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

7.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book- all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

```
?bookdown::gitbook
```