Advanced Research Writing in Statistics and Data Science

Kun Chen, Elizabeth Schifano, Jun Yan

2025-08-22

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 [Xie, 2016], 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.

What does a statistical or data science paper look like, and why should we study it?
For many graduate students and researchers, writing a paper is one of the most challenging yet most important parts of their training. A paper is more than just record of results; correctness is only the bottom line. A paper is the primary way for researchers to communicate ideas, establish credibility, and contribute to scientific literature. Good writing makes research visible, while poor writing may jeopardize delay the acknowledgement of even the most important discoveries.

As with all scientific papers, statistical and data science articles generally follow a set of commonly expected structures: title, abstract, keywords, introduction, methods, results, discussion, acknowledgements, references, appendix, and supplement. Yet, due to the specificity of the statistical discipline, the practices of machine learning, and the variety of application domains, research papers can look very different from one another.

This book (and its companion course) is motivated by the need to make these conventions explicit, to demystify the writing process, and to provide practical guidance through many examples and exercises. It is designed for graduate students and researchers who wish to sharpen their writing, whether they are preparing a dissertation chapter, a methodological paper, or an applied paper. Our goal is to teach students *how* statistical and data science papers are structured, *why* they are written in such ways, and *what* we shall do to adapt writing for different audiences and outlets.

1.1 Types of Papers in Statistics and Data Science

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 Barber et al. [2021].

1.1.2 Methodological papers

A **methodological 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
- Literature review
- Methods (e.g., estimation, hypothesis tests, diagnostic procedures)
- Theoretical properties
- Simulation studies
- Applications/Illustrations
- Discussion/Conclusions

Such papers emphasis on methodological development, with a blend of theory (e.g., asymptotic or non-asymptotic guarantees) and empirical validations.

Such papers emphasize methodological development, often with a blend of theory (e.g., asymptotic or non-asymptotic guarantees) and empirical validations.

Astrong methodological paper should not be just a mechanical combination or minor extension of existing methods. Instead, it should be driven by a clear motivation from real-world applications.

In other words, the most impactful methodological contributions are those that connect with practical relevance. They are inspired by genuine applied needs, but provide solutions that are general enough to influence future work in an important domain or even across different domains.

Example include Li et al. [2023] and Lau and Yan [2022].

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 Price and Yan [2022]; Caplan et al. [2019]; Jiao et al. [2022].

It is important to note that what we call "applied papers" here includes a large portion of scientific papers that rely on data analytics, statistics, and machine learning methods. In many scientific domains, these papers may in fact be considered theoretical or methodological contributions within that field. For instance, an applied paper with genomical data analysis could be regarded as a methodological paper in genetics.

Such works are most often interdisciplinary, typically resulting from close collaborations between statisticians, data scientists, and domain experts. They showcase how quantitative methods advance science in other fields, while also motivating the development of new techniques within statistics and machine learning.

1.1.4 Data Science and Machine Learning Papers

Beyond traditional statistics journals, researchers often publish in data science, machine learning, and data mining journals and conferences, including outlets such as *NeurIPS*, *ICML*, *KDD*, and *AAAI*. These conference 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 often prioritize empirical performance and novelty over extensive theoretical justifications or methodological development, though many still include core mathematical or probabilistical analysis. This emphasis partly reflects the nature of the research topics: for many cutting-edge machine learning and AI methods, their theoretical understanding is still evolving and often lags behind practice. As a result, papers are judged primarily on their ability to demonstrate empirical advances on benchmark datasets or practical applications.

1.2 Scientific Writing Resources

Many resources on scientific writing are available. For example, Gopen and Swan [1990] 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 Oshima and Hogue [2000], Gopen [2004], Hairston and Keene [2003], and Lebrun and Lebrun [2021].

1.3 About This Book

This book, *Advanced Research Writing in Statistics and Data Science*, extends earlier efforts from *Statistical Writing*, by broadening its scope and including hands-on examples and exercises.

The book emphasizes three guiding principles:

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

2. **Adaptivity**: tailoring style and structure to journals, conferences, and other outlets.

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

1.4 Outline of the Book

The planned content includes:

- **Introduction** (this chapter): Types of research papers — theory, methods, and applied — and the conventions of scientific writing across statistics, data science, and machine learning.
- **Tools and Workflows for Writing**: Version control (Git/GitHub), LaTeX, R Markdown/Bookdown, and reproducibility practices.
- **Getting Started with Writing**: The research and writing lifecycle, writing proposals, and planning strategies.
- **Writing Specific Sections**: Titles, abstracts, introductions, methods, results, discussion, figures, tables, captions, and style.
- **Writing for Different Outlets**: How to adapt writing for different audiences and venues — e.g., writing statistical analysis sections in interdisciplinary papers, preparing concise ML/AI conference papers, and meeting expectations in statistics journals.
- **Peer Review and Revision**: How to review manuscripts constructively, write referee reports, and respond to reviewers effectively.
- **Grant Proposals and Research Statements**: Writing for funding agencies, fellowships, and academic job applications.
- **Ethics and Integrity**: Authorship, plagiarism, collaboration, and responsible use of AI tools.
- **Exercises and Practice**: Rewriting, reviewing, and polishing tasks, designed to simulate authentic writing and reviewing experiences.

Each chapter will include **examples drawn from real papers across statistics, biostatistics, and machine learning** and will feature **exercises** that encourage hands-on practice and reflection.

1.5 Before We Start

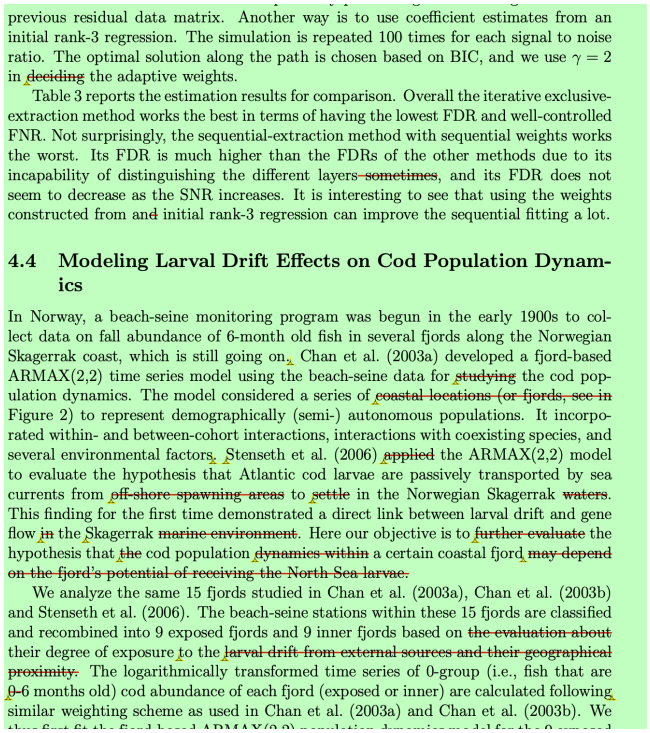
An author should always keep the target audience in mind. Statistical journals span a wide spectrum from applied to theoretical. Machine learning venues differ again in expectations. Even technical reports, white papers, or grant proposals have unique readerships.

Regardless of outlet, any scientific writing should be convincing and concise. Authors need to show clearly that their work is important, valid, and useful.

Ultimately, strong writing can never substitute for strong research. In short, keep in mind that **a good paper must rest on good work**. Just as important, a good paper is made through revision.

Or, as the Chinese phrase puts it: .

You can view an example of a marked-up draft with comments from my PhD advisor:
Download advisor's comments (PDF)



previous residual data matrix. Another way is to use coefficient estimates from an initial rank-3 regression. The simulation is repeated 100 times for each signal to noise ratio. The optimal solution along the path is chosen based on BIC, and we use $\gamma = 2$ in deciding the adaptive weights.

Table 3 reports the estimation results for comparison. Overall the iterative exclusive-extraction method works the best in terms of having the lowest FDR and well-controlled FNR. Not surprisingly, the sequential-extraction method with sequential weights works the worst. Its FDR is much higher than the FDRs of the other methods due to its incapability of distinguishing the different layers sometimes, and its FDR does not seem to decrease as the SNR increases. It is interesting to see that using the weights constructed from and initial rank-3 regression can improve the sequential fitting a lot.

4.4 Modeling Larval Drift Effects on Cod Population Dynamics

In Norway, a beach-seine monitoring program was begun in the early 1900s to collect data on fall abundance of 6-month old fish in several fjords along the Norwegian Skagerrak coast, which is still going on. Chan et al. (2003a) developed a fjord-based ARMAX(2,2) time series model using the beach-seine data for studying the cod population dynamics. The model considered a series of coastal locations (or fjords, see in Figure 2) to represent demographically (semi-) autonomous populations. It incorporated within- and between-cohort interactions, interactions with coexisting species, and several environmental factors. Stenseth et al. (2006) applied the ARMAX(2,2) model to evaluate the hypothesis that Atlantic cod larvae are passively transported by sea currents from off-shore spawning areas to settle in the Norwegian Skagerrak waters. This finding for the first time demonstrated a direct link between larval drift and gene flow in the Skagerrak marine environment. Here our objective is to further evaluate the hypothesis that the cod population dynamics within a certain coastal fjord may depend on the fjord's potential of receiving the North Sea larvae.

We analyze the same 15 fjords studied in Chan et al. (2003a), Chan et al. (2003b) and Stenseth et al. (2006). The beach-seine stations within these 15 fjords are classified and recombined into 9 exposed fjords and 9 inner fjords based on the evaluation about their degree of exposure to the larval drift from external sources and their geographical proximity. The logarithmically transformed time series of 0-group (i.e., fish that are 0-6 months old) cod abundance of each fjord (exposed or inner) are calculated following similar weighting scheme as used in Chan et al. (2003a) and Chan et al. (2003b). We

Chapter 2

Tools and Workflows for Writing

Many people use MS Word when it comes to writing. Not withholding the importance of the invention of MS Office, it is not the right tool to write methodological or theoretical papers in statistics. To many, writing a statistical paper using MS Word would be as interesting as running a statistical data analysis using MS Excel. Simply put, MS Office is great for office staff to do routine office documentary tasks.

For professional writing, one need to be aware of the professional tools and invest time to master them.

2.1 Word vs. LaTeX

The right, high - quality, professional typesetting system is LaTeX. LaTeX is a typesetting language that makes it easier and cleaner to write documents involving extensive mathematical content. It is the standard in Statistics, Mathematics, Physics, Chemistry and other disciplines that require many mathematical formulas.

LaTeX separates the appearance of a document from its content. This allows authors to be able to focus on writing the content without having to worry about its appearance until the end. There are many different professionally looking appearances one can choose or design, allowing for easy adaptation to different formats and styles.

A LaTeX document has `.tex` extension, and can be edited by your favorite text editor. The final output of the document can have different formats, the most

popular of which is `pdf`, which stands for *portable document format* . It can be opened on any platform (computer operating system).

The source `.tex` file is a plain text file. Just like source code of any programming language, a plain text file allows version control, which makes tracking and managing the source easy and professional. The most popular version control tool today is `git`.

Example 2.1. LaTeX vs. Word

Suppose you want to write down a likelihood function for a Gaussian model:

$$L(\mu, \sigma^2 \mid x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

- **In LaTeX** , the source code is simple and transparent:

```
L(\mu, \sigma^2 \mid x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n / 2}\exp\left\{
-\frac{1}{2\sigma^2}\sum_{i = 1}^n(x_i - \mu)^2\right\}.
```

This produces beautifully typeset mathematics, is reproducible, and can be revised easily. In MS Word, one must insert each symbol manually through the Equation Editor:Greek letters via dropdowns, summations via menus, exponents via special boxes. Formatting quickly becomes clumsy, and editing dozens of formulas is tedious. Copy-pasting often breaks structure, and version control is nearly impossible.

For one or two equations Word may be acceptable. But for an entire paper with dozens of equations, cross-references, and theorems, LaTeX is the only tool that is professional, efficient, and sustainable.

Exercise 2.1. Compare Word and LaTeX

1. Typeset the following equation in **MS Word** using its built - in Equation Editor:
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \mid y - X\beta \mid_2^2 + \lambda \mid \beta \mid_1 \right\}.$$

2. Now, paste this LaTeX code into Overleaf (or another LaTeX editor) and compile:

```
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p}
\left\{\frac{1}{2n}\ | y - X\beta\ | _2^2 + \lambda \ | \beta\ | _1 \right\}
```

- Which approach is faster ?
- Which result looks more professional ?
- Which is easier to revise and share ?

While LaTeX is the professional standard for statistical and mathematical writing, statisticians must also remain flexible when collaborating with scientists

from other fields. In many interdisciplinary projects, collaborators still prefer Word as the common platform for writing and revision. This can also often be the case when developing grant applications to National Insititutes of Health (NIH). This is understandable, since for papers in fileds such as biology, medicine, or social sciences, statistical formulas are often minimal or should even be avoided altogether. In such cases, it is often best to accommodate collaborators by using Word for the main text, while reserving LaTeX for supplementary technical sections, appendices, or internal drafts where precise mathematical notation is required.

2.2 Git for Version Control

Many tutorials are available in different formats. Here is a YouTube video "Git and GitHub for Beginners—Crash Course' '.

The video also covers GitHub, a cloud service for Git. Other similar services are, for example, bitbucket and GitLab. A cloud service gives you a cloud back up of your work and makes collaboration with co - workers easy.

There are tools and tutorials that make learning Git easy.

- Here is a collection of online Git exersices
- Here is a game called `Oh My Git`, an open source game about learning Git!

2.2.1 Set Up

- Download Git here.
- Make a GitHub Account here if you don't have one yet.
- Get started with your GitHub account by following the help page.
  - One important step is the set-up.
  - The connection between your local and GitHub repositories needs to be set up onlyonce.
  - One easy way is with a personal access token, as illustrated in a YouTube video.

2.2.2 Most Frequently Used Git Commands

- `git clone`:
  - Clones a remote repository to a local folder.
  - Requires either HTTPS link or SSH key to authenticate.
- `git pull`:

  - Downloads any updates made to the remote repository and automatically updates the local repository.
- `git status`:
  - Returns the state of the working directory.
  - Lists the files that have been modified, and are yet to be or have been staged and/or committed.
  - Shows if the local repository is begind or ahead a remote branch.
- `git add`:
  - Adds new or modified files to the Git staging area.
  - Gives the option to select which files are to be sent to the remote repository.
- `git rm`:
  - Used to remove files from the staging index or the local repository.
- `git commit`:
  - Commits changes made to the local repository and saves it like a snapshot.
  - A message is recommended with every commit to keep track of changes made.
- `git push`:
  - Pushes commits made on local repository to the remote repository.

Example 2.2. A clean daily Git workflow

1. Pull the latest changes

```
git pull
```

2. Write/edit files (e.g., intro.Rmd, style.css).

3. Check what has changed

```
git status
git diff
```

4. Stage only what you intend to commit

```
git add intro.Rmd style.css
```

5. Commit with a focused message

```
git commit -m "intro: refine motivation; style: add callout color"
```

6. Push to remote

```
git push
```

2.2.3 Tips on using Git:

- Use the command line interface instead of the web interface (e.g., upload on GitHub)
- Make frequent small commits instead of rare large commits.

- Make commit messages informative and meaningful.
- Name your files/folders by some reasonable convention.
- Lower cases are better than upper cases.
- No blanks in file/folder names.
- Keep the repo clean by not tracking generated files.
- Creat a `.gitignore` file for better output from `git status`.
- Keep the linewidth of sources to under 80 for better `git diff` view.

2.2.4 Pull Request

To contribute to an open source project (e.g., our classnotes), use pull requests. Pull requests "let you tell others about changes you've pushed to a branch in a repository on GitHub. Once a pull request is opened, you can discuss and review the potential changes with collaborators and add follow-up commits before your changes are merged into the base branch."

Watch this YouTube video: GitHub pull requests in 100 seconds.

Example 2.3. A useful `.gitignore` for LaTeX/Bookdown

Create a file named `.gitignore` in the project root:

```
# RStudio / R files
.Rproj.user/
.Rhistory
.RData
.Ruserdata

# Bookdown build output
_book/
_bookdown_files/
*.utf8.md
*.knit.md

# LaTeX build files
*.aux
*.bbl
*.blg
*.brf
*.fls
*.fdb_latexmk
*.idx
*.ilg
*.ind
*.lof
*.log
*.lot
*.maf
```

```
*.mtc*
*.out
*.pdf        # <-- if you want to keep generated PDFs out of git
*.synctex.gz
*.toc
*.ttt
*.xdv

# Common OS junk
.DS_Store
Thumbs.db

# Editor files
*.swp
*.swo
```

Now `git status` will show only meaningful source changes.

2.3 LaTeX

2.3.1 LaTeX Templates

A LaTeX source file has extension name `.tex`. It is a plain text file that can be edited by any text editor. It can be tracked easily for differences between any two versions. Different document classes are predefined such as `letter`, `article`, `report`, `beamer` (for presentations), and `book`. Customized document classes can be defined once you know more about LaTeX.

- The instructions in this section are practiced in a demo repo from Professor Jun Yan.
- Anthony Zeimekakis was an undergraduate student who worked with us on a thesis. The tex sources, data, and code are in a GitHub repo, which can be used as a template too. This paper was published in American Statistician, two years after Anthony graduated. Interested students beaware that this is a serious commitment.
- You may also directly download LaTeX templates (zip) here:
  - Download LaTeX template for writing a paper
  - Download LaTeX template for making a presentation

2.3.2 Editing and Compiling LaTeX

Working efficiently with LaTeX largely depends on the editing environment. Although a plain text editor is sufficient for editing, certain tools can make the

writing, compiling, and debugging process much smoother.

2.3.2.1 Local Editors

- *Emacs with AUCTeX*: a very powerful, customizable editor popular with many statisticians. AUCTeX provides syntax highlighting, automatic indentation, compilation shortcuts, and forward/backward search between the source and the PDF.
  - Here is a short tutorial.
- *RStudio*: although primarily for R, it supports LaTeX via R Markdown and Bookdown projects, making it very convenient for statistical writing.
  - Here is a template from the UConn Data Science Lab.
- *TeXstudio* and *TeXmaker*: cross-platform editors that offer integrated compiling, autocompletion, spell checking, and bibliography management.

2.3.2.2 Online Editors

- *Overleaf*: a widely used online LaTeX editor that runs in the browser. It allows real-time collaboration, version history, and integrates easily with GitHub. Overleaf is especially helpful when working with co-authors who are less comfortable setting up LaTeX locally.

2.3.2.3 Compiling LaTeX

For compilation, the traditional workflow involves running `pdflatex` and `bibtex` manually. Tools like `latexmk` can automate this process by detecting which passes are needed and executing them automatically. Most editors (e.g., TeXstudio, VS Code, Emacs+AUCTeX) integrate `latexmk` so you only need to press one shortcut key to recompile the document.

Exercise 2.2. Compile a LaTeX template

Let us start from a basic template in the demo repo. Clone it to an appropriate location on your own computer. Go to the `manuscript` folder and compile the pdf product with the following:

```
pdflatex statspaper
bibtex statspaper
pdflatex statspaper
pdflatex statspaper
```

It is the `bibtex` step that incorporates the references from the bib files. Two rounds of `pdflatex` are necessary for LaTeX to get all the cross-referencing settled.

The whole process could be automated by:

```
latexmk -pdf statspaper
```

Advanded users may take a look at the `Makefile`, in which different targets can be set up and the needed opertations for each target is automated.

2.3.2.4 Choosing the Right Tool

- If you prefer maximum control and flexibility, *Emacs with AUCTeX* or *VS Code with LaTeX Workshop* are excellent.
- If you want a quick, friendly environment, *RStudio* work well.
- For collaborations, especially interdisciplinary projects or student mentoring, *Overleaf* is often the most practical choice, since it removes the need to install LaTeX locally.

Exercise 2.3. Trying Different Editors

1. If you already have LaTeX installed, open the same `.tex` file in two different editors (e.g., Emacs and TeXstudio, or VS Code and RStudio). Alternatively, you may use Overleaf.
2. Try compiling the document and correcting a small error (e.g., a missing bracket).
3. Discuss:
   - Which editor felt more comfortable?
   - Which features (syntax highlighting, autocompletion, error display) were most helpful?
   - If you were collaborating with someone outside statistics, which tool would you recommend?

2.3.2.5 Tips on Getting Started

- Read the compiling log and fix the errors/warnings.
- Googling the error/warning messages usually helps.
- Limit the preamble to include only what is necessary.
- Set up document margins with the `geometry` package.
- No manually controlling spaces.
- Familiarize yourself with LaTeX symbol tables.
- Keep line widths under 80 characters in source files.
- Separate paragraphs in source files by double blank lines.
- Define acronyms at their first occurrences and only once.
- Use GPT or other LLM tools. While these tools are very convenient and getting better and better, it is still helpful if you know the basic about LaTeX.

Perfect — that will be very engaging for students. You can simulate **realistic LaTeX error messages**, show them in an example block, and then ask students to diagnose what they mean. Here's a ready-to-drop Rmd block:

Example 2.4. Common LaTeX Error Messages

Here are some typical error messages you might encounter.

1. **Missing end of document**

```
! LaTeX Error: \begin{document} ended by \end{article}.
```

2. **Undefined control sequence**

```
! Undefined control sequence.
l.6 This paper is written in \Latex
```

3. **Missing $ inserted**

```
! Missing \$ inserted.
l.12 The regression coefficient is beta\_1
```

4. **Citation undefined**

```
LaTeX Warning: Citation 'smith2020' on page 3 undefined on input line 45.
```

- What do each of these errors/warnings mean?
- How would you go about fixing them?

2.3.3 Math Equations

For serious math typesetting, use packages `amsmath`, `amsthm`, and others.

Tips on using math:

- Punctuate equations as they always are part of sentences.
- Add spaces between symbols for better readability in sources.
- Do not start a sentence with a math symbol; rephrase to avoid it.
- No fractions (`\frac`) in inline math expressions.
- No breaking inline math expressions into different lines in tex sources.
- No labeling equations that are not referenced.
- Reference labeled equations with `\eqref` instead of `(\ref)`.
- Keep fonts consistent for the same notations (e.g., n not n; AIC not AIC).
- Use appropriate sizes for parentheses.
- When multiple parentheses are needed in mathematical expressions, use the following ordering unless the journal specifies otherwise [{(math here)}].
- Use predefined math functions (e.g, exp not exp; Pr not P).
- Use `\allowdisplaybreak` to allow page breaks in aligned equations.
- Use `\dd` for differentiation operator (available from package `physics`).
- No breaking long equations arbitrarily in tex source; break them into short lines at appropriate places and add sufficient spaces to make the sources more readable.
- Align at appropriate places in multiline equations.

Example 2.5. Cleaning Up Math Expressions in LaTeX

Below are some math expressions written poorly in LaTeX.
Your task is to (1) identify what's wrong with each one, and (2) rewrite it following the best practices listed above.

1. Inline math with fraction:

```
$ \frac{a}{b} + c $
```

2. Missing equation label:

```
\begin{equation}
y = \beta_0 + \beta_1 x + \epsilon
\end{equation}
```

3. Wrong font for notations:

```
$ AIC = -2logL + 2k $
```

4. Misplaced function and parentheses:

```
$ P(X>c) = exp(-\lambda c) $
```

Exercise 2.4. A Statistically Correct Paragraph with Bad LaTeX

The following paragraph deliberately violates many LaTeX best practices. Your task is to identify and fix all the issues.

```
Let $\mathbf{m}=(m_1,\ldots,m_p)^T \in N^{p}$ be a vector of taxon counts
and $M=\sum_{j=1}^{p} m_j$ be the total count. Let $S \in \{1, ..., K\}$
be the unobservable hidden state variable indicating the cluster membership.
Assume $\mathbf{m}$, in any given state S, follow a Dirichlet-Multinomial (DM)
distribution. More specifically
\begin{equation}\label{eq:dm}
  \mathbf{m} | (S=k) \sim DM(M,\theta_k,\alpha^{[k]})
\end{equation}
or, equivalently, the hierarchical structure
\begin{equation*}
  \begin{aligned}
    \mathbf{m} | \mathbf{p} &\sim Multinomial(M,\mathbf{p}) \\
    \mathbf{p} | (S=k) &\sim Dir(\theta_k,\alpha^{[k]})
  \end{aligned}
\end{equation*}
where $\alpha^{[k]}=(\alpha_{1}^{[k]},...,\alpha_{p}^{[k]})^T \in \Delta^{p-1},
\theta_k>0$
is an over-dispersion parameter and $\mathbf{p}$ follows a Dirichlet
distribution under
each state k in the hierarchical formulation The conditional mean of
taxon j's count is
$E(m_{j} | S = k) = M\alpha_{j}^{[k]}$ and the conditional variance
```

```
is
$Var(m_j | S=k) =
\frac{M\alpha_{j}^{[k]}(1-\alpha_{j}^{[k]})(M \theta_k + 1)}{\theta_k + 1}$
For completeness, note also that $P(S=k|\mathbf{m})>0$ for
all k $[\{(m_1+m_2]\}]$ (see (\ref{eq:dm})).
```

Let $\mathbf{m} = (m_1, \dots, m_p)^T \in N^p$ be a vector of taxon counts and $M = \sum_{j=1}^{p} m_j$ be the total count. Let $S \in \{1, ..., K\}$ be the unobservable hidden state variable indicating the cluster membership. Assume $\mathbf{m}$, in any given state S, follow a Dirichlet-Multinomial (DM) distribution. More specifically

$$\mathbf{m}|(S = k) \sim DM(M, \theta_k, \alpha^{[k]}) \tag{2.1}$$

or, equivalently, the hierarchical structure

$$\begin{aligned} \mathbf{m}|\mathbf{p} &\sim Multinomial(M, \mathbf{p}) \\ \mathbf{p}|(S = k) &\sim Dir(\theta_k, \alpha^{[k]}) \end{aligned}$$

where $\alpha^{[k]} = (\alpha_1^{[k]}, ..., \alpha_p^{[k]})' \in \Delta^{p-1}, \theta_k > 0$ is an over-dispersion parameter and $\mathbf{p}$ follows a Dirichlet distribution under each state k in the hierarchical formulation The conditional mean of taxon j's count is $E(m_j|S = k) = M\alpha_j^{[k]}$ and the conditional variance is $Var(m_j|S = k) = \frac{M\alpha_j^{[k]}(1-\alpha_j^{[k]})(M\theta_k+1)}{\theta_k+1}$. For completeness, note also that $P(S = k|\mathbf{m}) > 0$ for all k [{$(m_1 + m_2]$}] (see (2.1)).

- Display equations missing terminal punctuation.
- Conditional bars written as `|` instead of `\mid`.
- Distribution names (`DM`, `Multinomial`, `Dir`) as bare text in math (should use `\text{}`/`\mbox{}` or `\mathrm{}`).
- Wrong/undefined sets and symbols.
- Inconsistent notation/fonts (transpose).
- Inline fraction `\frac{…}{…}` in inline math.
- Mismatched/incorrect delimiter ordering: `[\{( … ]\}]`.
- Unnumbered vs numbered environments used inconsistently; labeled equation `\label{eq:dm}` referenced as `(ref{…})` instead of `\eqref{…}`; or labeled without a proper reference.
- Missing commas and periods in sentences around math; weak spacing/readability in sources.

2.3.4 Tables

If you are manually typing a LaTeX table source, think if you can generate the source. There are multiple R packages that can generate the tex source from a given dataset. See package `xtable` for example.

Tips on professional LaTeX tables:

- Use `tbp` for floating locations; avoid `h`.
- Make it self-contained with an informative caption.
- Captions should be located above the table unless the journal specifies otherwise.
- Avoid vertical lines.
- Put negative signs in math mode.
- Use better top, middle, and bottom rules from package `booktabs`.
- Allow hierarchy by `cmidrule()`.
- Do not change font size for tables. Change table layout to fit instead of re-sizing it.
- Right adjust numbers with decimal places.
- Use consistent number of decimal places within a column or row of same types of measurements.
- Avoid having many leading 0's in decimal entries.

See Small Guide to Making Nice Tables by Markus Puschel

Perfect — this is a great place to design a **"bad table" exercise** so students can practice spotting and fixing formatting issues. Here's a deliberately messy version of your table, full of bad practices (tiny font, vertical lines, inconsistent decimals, misaligned numbers, redundant symbols, etc.).

You can drop this directly into your bookdown/LaTeX as an **exercise block**.

Exercise 2.5. Fix the Table

Below is the same descriptive table as before, but typeset with many bad practices.
Your task is to identify the issues (at least 8) and then rewrite the table according to professional LaTeX table guidelines.

```
\begin{table}[h]
\caption{Table showing descriptive characteristics of the case control data.}
\centering
\tiny
\begin{tabular}{|l|c|c|c|c|}
\hline
 & Multi-record cases & Multi-record controls & Single-record cases &
 Single-record controls \\
\hline
Total & 487.0 & 2435 & 1408.00 & 7040 \\
\hline
Sex &  &  &  &  \\
Female & 310(63.7%) & 1505 (61.8) & 952(67.61 %) & 4760 \\
Male & 177 (36.34) & 885 & 456 (32.4 %) & 2280.000 \\
\hline
Race &  &  &  &  \\
Asian & 12 & 60.00 & 27(1.9\%) & 135.000 \\
```

```
Black & 36 (7.39 \%) & 180 & 106 (7.5%) & 530 \\
Hispanic & 32 (6.57%) & 160 & 116 (8.2) & 580 \\
Other & 24 (4.93 %) & 120 & 128(9.09%) & 640 \\
White & 383 (78.64 \% ) & 1915 & 1031 (73.2\%) & 5155 \\
\hline
Age &  &  &  &  \\
10-14 & 46 (9.45) & 237 (9.7%) & 176(12.5) & 830 (11.8%) \\
15-19 & 247(50.7 \%) & 1161 & 711 (50.50) & 3449 (49) \\
20-24 & 194 (39.84\%) & 1037 & 521(37.0\%) & 2761 (39.22\%) \\
\hline
\end{tabular}
\end{table}
```

Table 2.1: Descriptive characteristics of the case-control data. The percentages are not shown for sex and race in the controls, as they are matched exactly with those of the cases.

	Multi-record cases	Multi-record controls	Single-record cases	Single-record controls
Total	487	2435	1408	7040
Sex				
Female	310 (63.66%)	1505	952 (67.61%)	4760
Male	177 (36.34%)	885	456 (32.39%)	2280
Race				
Asian	12 (2.46%)	60	27 (1.92%)	135
Black	36 (7.39%)	180	106 (7.53%)	530
Hisp	32 (6.57%)	160	116 (8.24%)	580
Other	24 (4.93%)	120	128 (9.09%)	640
White	383 (78.64%)	1915	1031 (73.22%)	5155
Age				
10-14	46 (9.45%)	237 (9.73%)	176 (12.5%)	830 (11.79%)
15-19	247 (50.72%)	1161 (47.68%)	711 (50.5%)	3449 (48.99%)
20-24	194 (39.84%)	1037 (42.59%)	521 (37%)	2761 (39.22%)

- Overuse of vertical lines `|`.
- Inconsistent decimal places (e.g., `487.0` vs `1408.00` vs `106 (7.5%)`).
- Parentheses and percentages formatted inconsistently (`63.7%`, `67.61 %`, `7.5%`).
- Tiny font (`\tiny`) makes the table unreadable.
- Caption is vague and not self-contained.
- Numbers not aligned by decimal point.
- Extra zeros (`2280.000`, `135.000`).
- Percentages outside math mode.
- Floating specifier `[h]` instead of `tbp`.

2.3.5 Figures

Use vector graphs, not raster graphs (unless you have to, e.g., screenshots). Save the code that generates the figures so the figures can be improved easily.

Tips on LaTeX figures:

- Use `tbp` for floating locations; avoid `h`.
- Use LaTeX package `graphicx`.
- Make it self-contained with an informative caption.
- Captions should be located below the figure unless the journal specifies otherwise.
- For line plots with different groups, use different line pattern to distinguish them, not only color, so that readers can tell the difference if printed in black/white. Same for different dots (symbols) on plots.
- Use colorblind friendly colors (especially avoid red/green).
- Keep the right aspect ratio when necessary (e.g., basketball court; map; pp-plot).
- Remove extra margins.
- Keep the ratio when resizing (e.g., `width = \textwidth`)
- Name the figure files appropriately.

2.3.6 References

BibTeX is a reference management tool for formatting lists of references that can be used together with LaTeX to generate a reference list. Non-referenced references are not to be cited. All referenced references are to be listed. This nice feature is made possible by the package `natbib`. We need to collect references in BibTeX format and save them in a bib database (`.bib`) file. The display styles of the references are controlled by bib style (`.bst`) files. Many journals have their own bib style files available for download. One can construct a customized bib style easily with the help of `custom-bib`.

An alternative to BibTeX and `natbib` is `biblatex`. Most journals, however, use BibTeX and `natbib`, so we focus on that here.

A reference is cited in the manuscript through its key by `\citep{}` for parenthetical citations or `\citet{}` for textual citations, where the key is placed inside the curly brackets. The key is used to cite or cross-reference the bibliographic entry in a `.tex` document. Variations `\citep*{}` and `\citet*{}` prints all authors. Sometimes, `\citeauthor{}` and `\citeyear{}` can be useful when only author(s) or year is needed. The key of the cited references is put in the parentheses.

For `\citep{}`, multiple keys separated by commas can be put in the same parentheses for citing multiple references. Two optional arguments are allowed to `\citep[][]{}`. For example, `\citep[see, e.g.,][p. 26]{}` could be useful when a specific page (or section/chapter) is being referenced as an example.

In general, to compile a tex file with bibtex references into a pdf document, one needs to run `pdflatex` first, then `bibtex`, and then `pdflatex` twice to get the references correct. A simpler solution is `latexmk -pdf`. In my practice, I always have a `Makefile` and use `make` to smartly automate the compiling process. See,

for example, Anthony's thesis repo.

Tips on preparing BibTeX databases:

- Devise a good naming convention for reference keys and stick to it.
- Keep the bib database sorted and formatted tidy. (No repeated entries.)
- Title: Capitalize first letters of notional words (not form words).
- Use Google Scholar to get the bibtex source of a reference, but be sure to **quality control** the google output for missing fields and errors.
- Protect capitalization of words with special meanings in curly braces. (e.g., `{B}ayesian`, `{M}arkov Chain {M}onte {C}arlo`)
- Protect capitalization of initial words after a colon in titles and journals.
- Use title style for jornal/book titles.
- For book chapters or proceeding articles, use `@incollection` instead of `@article`, and fill the `booktitle` and `editor` fields.
- Separate pages numbers with double dashes and no other spaces (e.g., `pages = {110--118}`).
- Books need to have publisher and address fields.
- For preprints, always check if they have been published recently.
- Use the `note` field to show information that should always be shown,
- All references without page numbers or volume number should be checked.
- Keep bib key style consistent.

2.3.7 Cross-referencing

Define a label for each object and refer to it by its label.

Tips on cross-referencing:

- Devise a good naming convention for labels and stick to it.
- Use different label prefixes for different types of objects (e.g, `eq:` for equations, `sec:` for sections, `tab:` for tables, `fig:` for figures, `alg:` for algorithms, etc.)
- Labels within the source(s) for a single document must be unique.
- Prevent referencing numbers from starting at a new line (e.g., use `Table~\ref{tab:simulation}`; note the tilde).
- Watch warnings from compiling logs for undefined labels or multiply defined labels and fix them.
- Use package `xr` for cross-document referencing (and labels must be unique across documents).

2.4 Command Line Interface

On Linux or MacOS, simply open a terminal.

On Windows, several options can be considered.

- Cygwin (with X): https://x.cygwin.com
- Git Bash: https://www.gitkraken.com/blog/what-is-git-bash

The new Windows OS provides a Windows Subsystem for Linux. As the name suggests, it aims to provide a Linux system on a Windows computer. It might be worth trying out.

To jump start, here is a tutorial: Ubuntu Linux for beginners.

At least, you need to know how to handle files and traverse across directories. The tab completion and introspection supports are very useful.

Here are several commonly used shell commands:

- `cd`: change directory; `..` means parent directory.
- `pwd`: present working directory.
- `ls`: list the content of a folder; `-l` long version; `-a` show hidden files; `-t` ordered by modification time.
- `mkdir`: create a new directory.
- `cp`: copy file/folder from a source to a target.
- `mv`: move file/folder from a source to a target.
- `rm`: remove a file a folder.

Chapter 3

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 4

Footnotes and citations

4.1 Footnotes

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

4.2 Citations

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

For example, we are using the **bookdown** package [Xie, 2025] (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 5

Blocks

5.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{5.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (5.1).

5.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 5.1.

Theorem 5.1. *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.

5.3 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 6

Sharing your book

6.1 Publishing

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

6.2 404 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.

6.3 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
```

Bibliography

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1): 486 – 507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

D J Caplan, Y Li, W Wang, S Kang, L Marchini, HJ Cowen, and J Yan. Dental restoration longevity among geriatric and special needs patients. *JDR Clinical & Translational Research*, 4(1):41–48, 2019. doi: 10.1177/2380084418799083.

George D Gopen. *Expectations: Teaching Writing from the Reader's Perspective*. Pearson, 2004.

George D Gopen and Judith A Swan. The science of scientific writing. *American Scientist*, 78(6):550–558, 1990.

Maxine Hairston and Michael L Keene. *Successful Writing*. W. W. Norton & Company, 5 edition, 2003.

J. Jiao, Z. Tang, M. Yue, P. Zhang, and J. Yan. Cyberattack-resilient load forecasting with adaptive robust regression. *International Journal of Forecasting*, 38(3):910–919, 2022. doi: 10.1016/j.ijforecast.2021.06.009.

Abby Y. Z. Lau and Jun Yan. Bias analysis of generalized estimating equations under measurement error and practical bias correction. *Stat*, 11(1):e418, 2022. doi: 10.1002/sta4.418.

Jean-Luc Lebrun and Justin Lebrun. *Scientific Writing 3.0: A Reader and Writer's Guide*. World Scientific, 2021.

Yan Li, Kun Chen, Jun Yan, and Xuebin Zhang. Regularized fingerprinting in detection and attribution of climate change with weight matrix optimizing the efficiency in scaling factor estimation. *Annals of Applied Statistics*, 17(1): 225–239, 2023. doi: 10.1214/22-AOAS1624.

Alex Oshima and Ann Hogue. *Writing Academic English*. Longman, 2000.

Michael Price and Jun Yan. The effects of the NBA COVID bubble on the NBA playoffs: A case study for home-court advantage. *American Journal of Undergraduate Research*, 18(4):3–15, 2022. doi: 10.33697/ajur.2022.051.

Yihui Xie. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, 2016.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2025. URL https://github.com/rstudio/bookdown. R package version 0.43.