Guideline to Writing a Master's Thesis in Statistics

Björn Andersson, Shaobo Jin and Fan Yang-Wallentin

Department of Statistics, Uppsala University

May 5, 2014

Contents

1 Introduction

The aim of this document is to help master's students develop effective technical skills in writing a master's thesis in statistics. The contents are meant to reflect the System of Qualifications in the Higher Education Ordinance. Recommendations and guidelines regarding the structure and content of a master's thesis are given. Section 2 describes a typical outline for a master's thesis and Section 3 gives recommendations about language, formatting, mathematical notation and tables and figures. In Section 4, some notes about the rules of conduct when writing a master's thesis are provided.

2 The Structure of a Master's Thesis

A master's thesis is an independent scientific work and is meant to prepare students for future professional or academic work. Largely, the thesis is expected to be similar to papers published in statistical journals. It is not set in stone exactly how the thesis should be organized. The following outline should however be followed.

Title Page

The title page should contain the title of the thesis, your name, the year, your supervisor's name, the Department of Statistics designation and the university logo. Two templates, one for LaTeX and one for MS Word, have been constructed for general usage. The most important thing to consider for the title page is to choose a suitable title for the thesis. The title should be brief yet give sufficient information about the topic of the thesis and should serve to attract further reading.

Abstract

An abstract should summarize the core parts of the thesis. Specifically, the abstract should include the motivation for the study, statement of the research problem, the methodology used and the results and conclusions of the study. For statistics journals the recommended length can vary quite a lot, for a journal following the APA (American Psychological Association) style guide under 120 words is required but for an American Statistical Association journal the requirement typically

is under 200 words. We recommend that the abstract should be no longer than 200 words. The abstract is a miniature thesis and should be written as such, hence you should not just copy parts of the thesis and stitch the parts together. By and large the abstract will determine whether or not your thesis will be granted further study by the reader so it is essential to write a good abstract. The abstract should avoid formulae and references if it is possible to do so and should be self-contained, that is the reader should be able to read only the abstract and still understand what the thesis concerns and its conclusions (American Psychological Association, 2001). It is recommended to start writing the abstract only after the rest of the thesis has been finalized (Dahmström, 2011).

Keywords

A few keywords should be listed describing the contents of the thesis, for usage by indexes to designate the area of the thesis. The keywords should number ten or fewer and terms already included in the title should be avoided (Higham, 1998).

Table of Contents

A table of contents should be provided, listing all the main sections in the thesis starting with the introduction. The main sections should be numbered but the references section, the acknowledgements section and the appendices section should be unnumbered. Note that a list of figures and tables is not required.

Introduction

The introduction should prepare the reader for the remaining parts of the thesis. Often it contains useful background information of the topic of the thesis and a description of the research problem. Previous research done in the area should be mentioned but unnecessary details should be avoided. The purpose and the goal of the study should be clearly stated and the motivations behind the study should be given. In case the specific research questions do not require detailed definitions better suited to the main text, the research questions should be included in the introduction. Otherwise state the specific research questions in the main part of the study. If investigating a theoretical property of an estimator or a similar topic, try to connect the study to an area of application. After

reading the introduction the reader should be convinced that your thesis is important and deserved of attention. Typically, an introduction is concluded with a description of the outline of the paper where each section is briefly described.

The Main Part of the Thesis

This part of the thesis can be structured in many different ways and which structure to use depends on the area studied and personal preferences. Often it is made up of three sections: background, methodology and results. Such a division is however not the only possible structure of this part of the thesis. For example, you may also have sections describing each method used in the paper with the headline indicating the area and then have a section about the study design and a results section. As a student you are advised to use your own discretion in deciding which structure is best suited for your thesis.

The main part of the thesis should contain a description of previous research in the area and current established knowledge (Dahmström, 2011). The requisite concepts and variables are to be defined. If not given in the introduction, the definitive research questions with specific hypotheses should be stated. Furthermore the methodology used in the study should be presented. This part can vary depending on the subject matter. If the thesis consists of an empirical study the methods used to collect the data and how the selection of individuals and variables was conducted must be given (Dahmström, 2011). If the thesis consists mainly of simulations, the simulation setup and software used should be detailed and their usage be justified. The statistical methods of the thesis should be presented and a theoretical background provided. The results of the study should be rigorously presented. Include tables and figures which are illustrative and relevant. Omit tables and figures that overlap in content with each other. Interpret the results and explain how they relate to the hypotheses of the study.

Two example outlines of the main part of the thesis, one for an empirical study and one for a simulation study, are given below as a reference point on how such theses can be structured. The outlines are a guideline to how the thesis can be structured but it is not required that the thesis is structured precisely in this way.

- Empirical study

  - Research question
  - Model
  - Data
  - Results

- Simulation study
  - Research question
  - Methodology
    * Model
    * Estimation methods
    * Simulation design
  - Results

Discussion

In the discussion, the results should be compared to the hypotheses stated beforehand and the conclusions from the study should be presented and put in relation to previous research. Any possible sources of error should be stated and the possibility of generalizing the results should be made clear. Focus on the unexpected or most interesting parts of the study conducted yet be careful to not over-interpret or overstate the results of the study. Emphasize the most important conclusions of the thesis and try to suggest future research in light of the results and conclusions.

Acknowledgements (optional)

The acknowledgements section is not required. It provides an opportunity to appreciate the individuals and organizations who assisted your work, such as your supervisor(s) and institution(s) who funded the work. Usually it is located between the sections **Discussion** and **References**. Acknowledgements should be written in the first voice.

References

By references, we refer to both in-text citations and a reference list. A reference list section is required. It provides the sources from which you found relevant studies and enables the readers to find those sources. Title the section ”References” or ”Bibliography”. The reference list is placed after **Acknowledgements** (or **Discussion** if no individual acknowledgements section) and before **Appendix**. The reference list should contain all the sources which are cited in the thesis. There are many types of references such as books, articles, unpublished manuscripts, technical reports, dissertations, on-line resources and so on.

There are two main citation styles. The Harvard style is the name and year system which is commonly used in statistical journals. The Vancouver style is a numbering system, which is widely used in medical journals. Both styles consist of in-text citations and a reference list.

Harvard Style

Harvard style is the most widely used reference style in statistical journals. We recommend the Harvard style unless your project is designated as medical statistics. Readers are referred to American Psychological Association (2001) for an extensive description of the Harvard style. In the Harvard style, the author's surname and the year of publication are implemented as an in-text citation. Two examples are ”Wilson (1927) proposed...” and ”The Akaike information criterion (Akaike, 1974) was...”. In the former example, the source is directly cited. The author's name is followed by parentheses with the year of publication in them. The latter case is an example of indirect reference. Both author's name and the year of publication are placed in the parentheses, separated by a comma. However, such a comma is not required. ”The Akaike information criterion (Akaike 1974) was...” is also commonly used.

Direct references and indirect references have several features in common.

- If there are multiple references by the same author in the same year, order them alphabetically by titles. A lower case letter is added after the year of publication without spaces.
- If there are two authors for one work, spell both surnames separated by an ”and”.
- If there are more than two authors for one work, there are two common methods:

1. At the first citation appearing in the text, write out all the authors but in subsequent appearances write only the first author followed by "et al.". However, according to this rule if there are six or more authors only write the first author followed by "et al." for all appearances of the reference.
2. Write the first author followed by "et al." for all the appearances of the reference.

The difference between direct citing and indirect citing is the way in which they handle parentheses. Some major differences are listed below.

- Direct reference
  - If there are multiple references by the same author at different years, years of publication are placed chronologically in the same parentheses separated by commas.
  - Multiple types of references are separated by commas except the last two. An ”and” is used to separate the last two entries.
  - For instance, ”Confidence distribution is discussed in Bickel (2006), Schweder and Hjort (2002, 2003), Singh and Xie (2011a, 2011b) and Xie et al. (2011).”.
- Indirect reference
  - If there are multiple references by the same author at different years, use a comma to separate the name from the years and commas to separate years. Works have to be cited chronologically.
  - Multiple types of references are separated by semi-colons.
  - For instance, ”Confidence distribution (Bickel, 2006; Schweder and Hjort, 2002, 2003; Singh and Xie, 2011a, 2011b; Xie et al., 2011) is becoming a hot topic.”.

For a reference list, different journals have different reference styles. The elements in the list are the same, but the order of elements and minor details vary for different journals. Some commonly used styles are the ”apalike” style, the ”plain” style and the ”acm” style, just to mention a few. What we used in this guide is the ”apalike” style. For LaTeX users, ”apalike” style is directly available.

Different types of sources have different elements in the reference list. Among them, books and journal articles are the most commonly cited materials. The elements for books and articles are listed below.

- Books:

  Author's surname, Initials with dots, Year of publication in parentheses. *Title*. Publisher, Place of publication, edition if the book has more than one edition.

  For example:

  Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, 3rd edition.

  In the case of multiple authors, names are separated by commas while an "and" is used between the last two authors without a comma. The title of the book has to be italic. Place of publication should be the name of a city and/or a state.

- Journal articles:

  Author's surname, Initials with dots, Year of publication in parentheses. Title of the article. *Title of the journal*, Volume(Issue):Pages.

  For example:

  Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1-26.

  Parentheses are not required for the year of publication. The title of the journal has to be italic. The number of volume can also be bold. The number of issue is not required.

Vancouver Style

If your work belongs to the subject of medical statistics, the Vancouver style is recommended. In the Vancouver style, references are numbered consecutively using Arabic numerals by the order of the first appearance in the thesis. Readers are referred to Patrias (2007) for a detailed description. Arabic numerals can be placed in parentheses, square brackets, superscripts or a combination of square brackets and superscripts. For example, "Katz (2) found...", "Katz [2] found...", "Katz ^{2} found..." and "Katz $^{[2]}$ found...". If multiple references have to be cited, use commas without spaces to separate non-inclusive numbers. Hyphens without spaces are used for inclusive numbers. For

example, ”A series of experiments (1,3-7,9) showed...”.

The elements for books in the reference list are the same as the Harvard style, but the elements of journal articles are different. The publication month has to be included if possible.

- Books:

  Author’s surname, Initials with dots. Title (with edition if the book has more than one edition). Place of publication: Publisher; Year of publication.

  For example:

  Lehmann E. L., Romano J. P.. Testing Statistical Hypotheses. 3rd ed. New York: Springer; 2005.

  In the case of multiple authors, names are separated by commas. However, if there are six authors, list the first 6 authors and use ”et al.” for the others. The title of the book is not italic.

- Journal articles:

  Author’s surname, Initials with dots. Title of the article. Title of the journal. Publication year and month;Volume(Issue):Pages.

  For example:

  Efron B.. Bootstrap methods: another look at the jackknife. Ann Statist. 1979 Jan;7(1):1-26.

  In the case of multiple authors, the rules are the same as the format for books. Journal abbreviations have to be used. Vancouver style does not use the full journal names. Make sure you have the correct journal abbreviations. No spaces are needed after the semi-colon and the colon.

Appendices (optional)

An appendix is not required. However, whatever is not suitable for the main text but relevant should be placed in the appendix. For example, mathematical proofs are not required but recommended to be placed in the appendix. Tables and figures which cannot be omitted but interrupt people from reading should be placed in an appendix.

Appendices should be numbered using capital letters in alphabetical order. For instance, ”Appendix A”, ”Appendix B” and so forth.

3 Some General Guidelines

In the following, a number of recommendations will be given regarding formatting, language, mathematical notation and how to deal with tables and figures. Consistency is the most important aspect to consider when writing the thesis, so be sure to maintain one set of rules that encompasses the entire thesis.

3.1 Language Recommendations

When communicating through writing you should always be aware of which audience you are targeting and what it is you want to communicate to this audience. Most people who read a master's thesis in statistics will have a fairly solid background in statistics and as such the writing should be adjusted to suit such a group of people particularly well. It is however important that the thesis will be understandable for a person not so familiar with statistics. For example, future employers - who are likely to read your thesis - may not be so well-acquainted with statistical terminology. You should not assume that the reader is familiar with everything that you are familiar with.

In general, it is recommended to write in standard English. However, you should avoid complex, long sentences and unnecessarily complicated words. The main object is to communicate results and ideas, not to display a deep vocabulary or prowess in writing. You should write clearly and concisely and not bury the reader in details nor over-simplify matters. Make sure to spell check and carefully proofread the thesis. The thesis should be read through in its entirety multiple times before submission.

Below we list a few specific recommendations about how to write and how not to write.

- Data is the plural form of datum, hence you should write "data *are*", "the data *were*" etc., and not "data *is*" or "the data *was*".

- Avoid contractions, i.e. write "will not", "should not" etc. and not "won't" or "shouldn't".

- We recommend using the word 'significant' only when talking about statistical significance. Use synonyms in other cases in order not to confuse the reader.

- Write the thesis in present tense, i.e. write "the results *show* that" and not "the results *showed* that". When referencing specific results from other papers, either present tense or past tense may be used. However, for well-known facts, present tense should be used.

- When referencing specific sections, theorems or definitions, capitalize them. E.g. you can write "In Section 1 we derived Theorem 2 from Definition 3". For information on how to reference math expressions, see Section 3.2.

- When using acronyms, write them out the first time they are used with the acronym in parenthesis immediately afterwards. I.e., write "Structural equation modelling (SEM) is a statistical technique".

- Most abbreviations should be avoided, however it is fine to use common Latin abbreviations like "e.g." and "i.e.".

3.2 Mathematics

Mathematical expressions are probably needed in order to describe the methodology clearly. There are several matters that you should be aware of. The following is adapted from Higham (1998) and Knuth et al. (1989).

In-line Equations or Displayed Equations

There are two ways to express your mathematical symbols, that is, in-line equations and displayed equations. For example, the sentence "Let $T = T(X_1, ..., X_n)$ be an unbiased estimator of the location parameter." consists of an in-line equation and the sentence "Let

$$T = T(X_1, ..., X_n)$$

be an unbiased estimator of the location parameter." consists of a displayed equation. Whether a mathematical expression should be placed in-line or be displayed depends on the purpose and its complexity. An equation which is worth special attention or which will be numbered has to be

displayed. If an expression is too long to be fit in a single line, you have to display it. If in-line math symbols look nice, you do not have to display them.

Sometimes an equation has to be broken into several lines due to the complexity. There is no unified rule regarding where you should break the equation. Three widely used rules are breaking before binary operations, breaking after binary operations and repeating the operator symbol before and after the break. Binary operators are the mathematical operators which involve two elements. Addition, subtraction, multiplication and division are typical binary operators. When a displayed equation has to be broken, usually break it after a binary operator. When an in-line equation has to be broken, usually break it before a binary operator.

Reference

If you have to reference a displayed mathematical expression, it is better to describe the nature of the expression. You should not capitalize the description, unless it is the first word in the sentence. For example, ”This is guaranteed by inequality (3)” is more informative than ”This is guaranteed by (3)”.

Punctuation

Both in-line and displayed mathematical expressions are parts of a sentence. Hence you have to punctuate them as well.

Miscellaneous

There are some more rules that you have to pay attention to.

- It is not required to number all the displayed equations.
- A sentence should not start with a mathematical symbol, if it is possible.
- Spell out integers under 10 except when referring to measurements, exact figures, references to sections et al, or similar.
- Some commonly used handwritten symbols, such as $\forall$ and $\exists$ should be avoided in the thesis.

- Avoid expressions which are too tall in a paragraph. For example, p_i/p_j is better than $\frac{p_i}{p_j}$.
- Any symbols should be defined before you use them or just after the first appearance but in the same sentence.

3.3 Formatting

The thesis should meet the following formatting requirements.

Page Style

The thesis must be produced by a word processor (Latex, Microsoft Word or similar processors) and printed in A4 format. Single column style should be used.

Margins

Margins should be 1 inch all around. In LaTeX this can be done by adding

```
\usepackage[margin=1in]{geometry}
```

to the preamble.

Font

Text must be black. Several choices of fonts are acceptable.

- Times New Roman 12pt
- Arial 12pt
- Computer modern 12pt (LaTeX default)

Spacing

Main text must be written with 1.5 line spacing. However, the captions, footnotes and appendices should be single spaced.

Numbering

Number your thesis consecutively using Arabic numerals. Page number starts with 1 on the introduction page. The title page, the abstract page and the table of contents page are not numbered and not counted.

Headlines

Capitalize all the words in the headlines. Some exceptions are articles, coordinating conjunctions and prepositions. However, the first and last words are always capitalized regardless of the types. Do not place the headlines at the bottom of a page.

3.4 Tables and Figures

Where to Place

Tables and figures can be placed either in the text or on a page without text. If tables and figures are placed in the text, they cannot be wrapped by text. Hence, text should be either above or below tables and figures. Tables and figures have to be centered horizontally. If they are placed on a page without text, they should be centered vertically as well. Do not place too many tables and figures in the main text. It distracts readers. Relevant but less important tables and figures can be placed in appendices. If a figure or table is included in the thesis, it should also be referenced in the text.

Numbering

Tables and figures must be numbered throughout the thesis independently. Abbreviations should be avoided. Spell out the words and name them consecutively as Table 1, Table 2, etc. Double numbering is also accepted such as Table 1.1 and Table 1.2, where the first number reflects the section number and the second number indicates the sequence. Capitalize the first letter when you reference them.

Caption

Tables and figures must be self-explanatory. Hence a detailed caption is needed, conveying all the necessary information. Single spacing is used. Where to put captions is different for tables and figures. Table captions should be placed above the table body, while figure captions should be placed below the graph.

Miscellaneous

- If a table is too wide to be fit in a page, rotate the table and put it on a page of its own.
- If a table is too large, you can reduce the font size. But the text should still be readable.

4 Plagiarism in Scientific Writing

A very important consideration is that of ethics in scientific writing. A master's thesis is an independent scientific work and will be treated as such. It is crucial to be aware of what you can and cannot do when writing a scientific work. As a general rule, it is never allowed to copy an entire sentence or more without properly acknowledging that it is a direct quote. I.e., if you want to re-state precisely what another person has written you should put this extract in quotation marks and provide the reference. Failure to do so is considered plagiarism in scientific writing. It is recommended to provide a minimum of quotations in the thesis. Instead of quotations you should adapt the text you use as a source and give a reference. If references are not given adapted text is also considered plagiarism. An exception to the recommendation about not using quotations is when stating theorems and equations. These may be necessary to re-state in their entirety but be sure to provide the reference if doing so. After submitting the thesis, the document will be analysed by a computer program in order to ensure that no disallowed copying has occurred.

References

American Psychological Association (2001). *Publication manual of the American Psychological Association*. American Psychological Association, Washington, D.C., 5th edition.

Dahmström, K. (2011). *Från datainsamling till rapport*. Studentlitteratur, Lund.

Higham, N. J. (1998). *Handbook of Writing for the Mathematical Sciences*. The Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition.

Knuth, D. E., Larrabee, T., and Roberts, P. M. (1989). *Mathematical writing*. Mathematical Association of America, Washington, D.C.

Patrias, K. (2007). *Citing Medicine: the NLM style guide for authors, editors, and publishers*. National Library of Medicine, Bethesda (MD), 2nd edition. Available from `http://www.nlm.nih.gov/citingmedicine`.