TU DORTMUND
Fakultät Statistik
Prof. Dr. Jörg Rahnenführer
Dr. Philipp Adämmer
Dr. Andrea Bommert
M.Sc. Hendrik Dohme

Summer semester 2021

Guidelines for the preparation of project reports for INTRODUCTORY CASE STUDIES

Layout and scope

The choice of the final layout is up to each student, but the following rules should be followed:

- use justification as text alignment (extends each line of your text to both left and right margin)
- leave sufficient space for corrections (at least 2.5 cm at all four edges)
- use 1.5 line spacing
- the font size must be sufficiently large (e.g. 12pt for Times New Roman)
- use uniform font throughout the report
- the main part of the report (without cover page, table of contents, Appendix and bibliography) may comprise a maximum of 15 pages, the Appendix a maximum of 3 pages
- use page numbering, the numbering starts after the table of contents

Note:

- The choice of word processing program is up to you. However, LaTeX is recommended. `BibLaTeX` is recommended for citing and for organizing the bibliography.
- Make sure that no words, figures or tables are written beyond the margin (see warnings in LaTeX). When using `BibLaTeX`, make sure that the results are meaningful (e.g. abbreviations, German umlauts).
- Avoid placing individual lines at the beginning or end of a paragraph on a different page than the remaining lines (so-called „widow“ or „orphan lines“). In Latex this can be achieved by including `\usepackage{nowidow}` in the preamble of the document.

Structure

The following list provides a basic framework for the structure of a scientific report, which is mandatory for the writing of all reports in this course. The additionally listed key points describe the structure of the different sections. Subsections can be created to structure the individual sections.

- Cover page, it must include:
  - course title
  - topic of the project

  - name of the lecturers
  - name of the author and names of the other group members as well as the group number
  - date

- Table of contents
  - Overview of all sections (including subsections, if any), the Appendix and the bibliography (but not the table of contents itself), each with the corresponding number or letters (see Section Outline below), and with the page number on which the corresponding section begins

  Note:
  - The table of contents has to be on the second page.

- 1 Introduction
  - Short general introduction to the topic
  - Brief description of content and goal of the project
  - Brief explanation of the approach to problem solving (in terms of content and methodology)
  - Possible short presentation of the central results
  - Overview of the individual sections

  Note:
  - The section overview should be filled with real content. Example: Specify *which* statistical methods are described instead of just writing *that* statistical methods are described. Rule of thumb: The overview should not fit to a report of a different project.

- 2 Problem statement
  - Description of the dataset
    - context and method of data collection (planned experiment, observational study, questionnaire etc.)
    - type and size of the sample (complete survey, stratification, etc.)
    - description of all variables (meaning, units, etc.)
    - discussion of data quality (missing values, measurement scale (nominal, ordinal, numerical etc.), measurement accuracy, etc.)
  - Description of the objectives of the project (both content-related and statistical objectives)

  Note:
  - Always specify a reference (source) for the data.
  - Explain the data collection process in detail, if you know it. If necessary, also describe any pre-processing steps that have been performed in comparison to the reference you have specified.
  - List all variables present in the data set. Variables used in the report must be described in more detail. Variables that are not considered further in the report are only mentioned (if a data set contains many additional variables, it is sufficient to mention the number and their subject areas). Variables may be renamed and

translated compared to the original data set. Abbreviations may also be introduced. Meaningless abbreviations such as $X_1, X_2, \ldots$ should be avoided.
  - More complicated variables must be explained and interpreted (e.g. how is a score constructed, or what do high and low values mean).
  - While the introduction contains only a short description of the content and methodological goals, here a detailed explanation of these goals is given.

- 3 Statistical methods
  - description of the statistical methods, models, etc. that are used, including their properties and the assumptions on which they are based (mathematical formulas are also required here)
  - details of the tools that are used (software including version number, statistical tables, etc.)

  Note:
  - All methods used for data analysis (and only these) should be described.
  - Motivate each method: Indicate what it can be used for. If you describe several methods that serve the same purpose (e.g. description of the central location), explain the differences (e.g. median is more robust than arithmetic mean, note that for this statement robustness has to be explained as well).
  - All methods must be explained accurately in terms of language and mathematics.
  - Explain all components of the formulas you have specified. Example: If the formula $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ is specified, it must be explained, what $x_1, \ldots, x_n$ and n mean in terms of content. Furthermore, it must be indicated what $x_1, \ldots, x_n$ and n must fulfill (measurement level, range of values, etc.).
  - Methods that you have already described in a previous report can be assumed to be known and can therefore be omitted.
  - Make sure that the methods you describe in this section are really those that you use in the analysis: Therefore, read carefully the help pages for software that you have not implemented yourself.
  - For each statistical method, you must provide a literature reference in which the method is described as you have specified it. Your own notation does not have to match that of the reference (see Section References below).
  - Also explain how the results of the methods are interpreted. Examples: What do high and low values of a variable mean? When can you assume, on the basis of a plot, that assumptions are fulfilled or not fulfilled?

- 4 Statistical analysis
  - if required, checks of the underlying assumptions of the statistical methods used
  - detailed presentation of the results, illustrated with tables and figures
  - interpretation of the results with regard to the problem statement

  Note:
  - The first subsection (except for the first report) must contain a short descriptive analysis of the presented data. All variables must be analyzed univariately. Often, bivariate analyses are also useful. The insights gained in the following subsections should also be set in relation to the results of the description.

  - Describe comprehensibly how you got from the data to the results, e.g. what happened with missing values.
  - A „red thread“ helps to understand your results. Therefore, explain what the respective analyses are used for and make connections between the analyses (Example: „This analysis serves to check the assumptions for ...“). At the beginning of each section or subsection, briefly state what you are going to investigate in it. Then carry out the analysis. Finally, formulate a short conclusion regarding the question that was addressed.
  - Interpretation means objective conclusions based on the data. Subjective comments should be included only in the „Summary“ section.

- 5 Summary
  - short repetition (restatement) of the research question and the data material of the project
  - brief presentation of the most important results
  - discussion of the results (possible conclusions, warning of misinterpretations, etc.)
  - outlook (open questions, possible further investigations, etc.)

  Note:
  - The Sections Introduction and Discussion should be readable on their own. This means in particular that *no* knowledge from the other sections can be assumed.
  - Answer all questions. Non-significant test results are also results.
  - Indicate the statistical methods you used to obtain your results. In statistical tests, mention the significance level.
  - A critical review of the results may be appropriate (e.g. multiple test level). However, you should not question all results.

- Bibliography
  - list of all journal articles, books and other publications used for the analysis (this includes e.g. also software manuals) and web pages
  - literature references are sorted alphabetically by the last name of the first author

  Note:
  - Example of a textbook specification:
    Hastie T., Tibshirani R., and Friedman J. (2009): *The Elements of Statistical Learning*, 2nd Edition, Springer, New York.
  - Example of a journal article:
    Rahnenführer, J. and Futschik, A. (2003): Cost-effective screening for differentially expressed genes in microarray experiments based on normal mixtures, *Austrian Journal of Statistics* **32**(2), 225–238.
  - For web pages, the date of the query must be specified. Example:
    Robert Koch Institut (2020): *The Robert Koch Institute*. URL: `https://www.rki.de/EN/Content/Institute/institute_node.html` (visited on 2nd October 2020).
  - In the bibliography all authors are listed (and not abbreviated with `et al.`)
  - All works listed in the bibliography must be cited in the text and vice versa.

- Appendix
  - Additional important tables and figures

Note:

- Reference to parts of the Appendix should always be made by indicating the page number, on which the table or figure can be found.
- Whether tables or figures are placed in the section „Statistical analysis“ or in the Appendix should be decided according to comprehensibility and clarity. It should be possible to read the evaluation without having to turn several pages to the corresponding graphic.

Outline

- Longer sections should be divided into subsections with meaningful headings. In addition, the text should be structured into paragraphs and, if necessary, by section headings. There should be considerably more space between two paragraphs than between lines within a paragraph.
- All sections (and, if necessary, the subsections) of the main body are given a number, whereby a more than three-level structure can be confusing and therefore is unusual. A two-level structure is recommended.
- In case of an outline in the Appendix (which is not mandatory, e.g. if the attachment is short), letters are used.
- Table of contents and bibliography are given neither a number nor a letter.
- The sections „Introduction“ to „Summary“ are numbered with the numbers 1 to 5. Subsections of these sections are numbered 1.1, 1.2, ..., 2.1, 2.2, etc.. The sections „Bibliography“ and „Appendix“ are not numbered. If the section „Appendix“ is divided into subsections, these subsections are numbered A, B, C, etc.
- The table of contents resulting from the outline should already give an overview of the contents by the subsections. Rule of thumb: The overview should not be compatible with a report on another project. However, the overview should not be too detailed.

Figures and Tables

- All figures and tables must have a meaningful caption (example for a table caption: „Location and dispersion measures for the variable ‚age‘, separated by gender“). Redundant headings in figures should be avoided.
- Both figures and tables are numbered consecutively, for example: “Figure 1: Title or explanation of the figure” or “Table 1: Title or content of the table”
- Figures and tables in the Appendix must also be numbered and must include a caption.
- Each figure and table must be referred to at least once in the text (this also applies to those in the Appendix). The corresponding number of the figure or table is used (Example: “In Table 1 the relationship between ... is shown.”)
- A figure and a table typically appear on the page or the following page of the first reference in the text, but never prior to that.
- Figures and tables should be self-explanatory. This includes that all axes are labeled. Do not forget the units. Additional relevant information can be displayed e.g. in a legend. Information that is necessary for understanding but cannot be placed within the figure or table can be provided in the subtitle.

- The size of the caption of the figures should approximately correspond to the font size of the report.
- If colors are used, colors that are easy to distinguish should be chosen. If the report is later printed in black and white, grayscales are recommended, for example.
- Table format usually is: numbers right-aligned, text left-aligned, otherwise content centered in a certain way if it increases readability. All numbers in a column must have the same number of decimal digits (add zeros if necessary). If reasonable, the same number of decimal digits should be used in all columns of a table. Variables that take only integer values are displayed without decimal digits.
- Do not display variables with different units together in one plot (on the same axis).
- All figures and tables (including those in the Appendix) must be explained in the text. First, it should be briefly mentioned what is shown in a figure or table. Example: „In Table 1, the results of the 2017 Bundestag election are shown separately by federal states." Afterwards, it must be explained which insights can be gained from the figure or table. Example: „Party X achieves its best result in federal state Y." One should not assume that a figure or table would obviously show something; all findings must be recorded in text form.
- Use vector graphics (e.g. .pdf).

References

- Within the report, citations to a bibliographical reference are made in the form `author (year)` or `(author, year)`. If there are two authors, both names must be given; if there are more than two authors, the short form `Author 1 et al. (year)` or `(author 1 et al., year)` must be used. If there are several works of an author from the same year, the references are distinguished by letters after the year: (2005a), (2005b) etc.

  Example for the above mentioned publications:
  - Rahnenführer and Futschik (2003) and (Rahnenführer and Futschik, 2003)
  - Hastie et al. (2009) and (Hastie et al., 2009)
- For books, the corresponding section or page must be specified. Example: "[...] The bias-variance decomposition of the expected prediction error of a regression fit is found in Hastie et al. (2009, p. 223)."
- A reference must be given for all information that is not general knowledge and that is not taken from the project description.
- A reference must also be given for the software used, whereby the publishers often indicate how to refer to the software. When using the software `R`, all packages used must also be cited that are not part of the base system of `R`.
- Lecture notes, Wikipedia, etc. are not acceptable references for statistical methods.
- References must not be copied literally (word by word), but the ideas must be formulated in your own words. The own formulation must not be too close to the one of the original reference. For formulas, it is important that not the notation of each reference is exactly copied, but that a consistent notation is used in the whole report.
- If available, select references that are available online (within the university network).

Mathematical formulas

- Mathematical formulas and symbols (whether as seperated formulas or in running text) must be distinguished from the rest by a different font (*mathfont* in LaTeX).
- Once selected, their names should be retained throughout the report.
- It is important that different types of measurement are clearly separated from each other: Upper case letters for random variables, lower case letters for observations.
- Mathematical formulas can also be numbered (using "(number)"), so that they can be referred to in the text.
- Larger formulas should be separated from the main text (not in running text).
- Function names must not be written in mathfont (e.g. use $\exp(x^2)$ instead of $exp(x^2)$ or $\min\{0, x\}$ instead of $min\{0, x\}$.
- Components of formulas in the running text are also written in mathfont. Examples: $\sum_{i=1}^{n} x_i$ is the sum of n values (not n values), t-test, p-value.
- Too many indexes are confusing.

Phrasing and wording

- The report (except for the summary) is written in present tense. The summary is written in a consistent past tense. Important: Findings that are still valid are also written in the present tense (e.g. there is an empirical correlation of 0.3 between X and Y).
- Regarding language, precise and factual/knowledge-based formulations should be chosen. Experience style, colloquial language, filler words, too many word repetitions and long nested sentences must be avoided. The text must be fluently readable.
- Avoid constructions with „should“ (e.g. use „will be studied“ instead of „should be studied“).
- Pay attention to logic of language. Example: „therefore“ indicates that there is a causal relationship.

Additional remarks:

- Citations must be marked as such (quotation marks, exact reference to the book/article and page number, possibly also indentation of the text) and must not consist of longer text sections. Copying of text sections is not permitted and leads directly to failing the course!
- The use of a specific formatting is not mandatory (besides rules mentioned above). However, it is important that a form once selected is applied consistently throughout the entire report.
- A good report is characterized by the fact that it is correct in content and form and that the question at hand is answered sensibly using appropriate procedures. It is not a question of writing down everything about the data set at hand. Rather, the aim is to present the central findings of the research question in a clear, precise, and correct manner.

- Before submitting your report, you should proofread the entire report at least once. It can also be helpful to proofread the reports within the group.