# Appendix F: Writing a Data Analysis Report

***BACKGROUND:*** In the 21st century, "team science," in which individuals possessing different disciplinary expertise integrate their skills to formulate and address subject matter questions, is the primary mechanism by which the advance of knowledge takes place. Statisticians are key members of such multidisciplinary teams, and almost every PhD statistician will engage in such collaborations with domain science experts. Collaborations with subject matter experts lead to inspiration for novel methodological research by statisticians whose primary responsibility is to engage in statistical research (mainly but not exclusively in academia), as, often, the need for new statistical methods to handle nonstandard challenges for which existing approaches are not approrpraiate becomes apparent in the course of collaborative projects. Collaborative work is the primary activity for statisticians in industry, research institutions, and government. Frankly, statistics as a discipline would not exist except for the challenges in collection, analysis, and interpretation of data that arise in other fields, so by nature relies on such collaborations to advance.

To be an effective collaborator, a statistician must possess outstanding oral and written communication skills. It is critical that the statistician is able not only to determine appropriate statistical methods to use and to carry out analyses using them, but to interpret the results in the context of the subject matter and to communicate them to his or her nonstatistician collaborators in a way that they can understand. Typically, such collaborators may have some familiarity with basic statistical models and methods at the level of an introductory statistics course or of particular methods that are predominant in their areas of expertise, but not of the more advanced and specialized models and methods that are often required to address the questions. Thus, the statistician must be adept at explaining the rationale for his or her choice of models and methods and why more familiar methods may not be appropriate.

When the questions of interest require longitudinal data models and methods, collaborators may not be familiar with the methods you as a statistician might choose. Collaborators may be familiar with, for example, classical linear regression and possibly logistic regression, and they may have heard of more complex methods for longitudinal and clustered data, but their knowledge of the latter is usually superficial. In some subject matter areas, classical analysis of variance methods may still be favored, which the collaborators believe they understand, but they usually do not appreciate the limitations of these methods and why you as the statistician may recommend and use more modern approaches. In fact, collaborators may sometimes lobby vigorously to use methods other than those that you feel

are appropriate because the journals in their area may not accept a paper for publication reporting on the results of a study if it does not use the "accepted" methods.

One key mechanism by which we as statisticians communicate the results of an analysis we have designed to answer specific scientific questions posed by our collaborators is through a formal data analysis report. Such a report documents systematically what was done and why and explains the results and their interpretation in terms of the subject matter. Here, we give some tips on how to structure and write a good data analysis report.

***AUDIENCE:*** Ordinarily, the intended audience for a data analysis report is your collaborators. It should be written primarily for them and not for you or other statisticians with your level of statistical knowledge. This audience wants to understand what you did, why you did it, and what the conclusions are and, while interested in the features of the methods you used and why you used them, does not want to see technical details, lots of symbols and specialized terminology, and computer code and output.

However, there are other interested audiences. You yourself a key audience, because you will want to have a careful and detailed record of what you did and why you did it if you need to revisit a study in the future and get up to speed quickly if new questions arise or if you are working on another project in which similar issues and questions are to be addressed. You may also have reason to share the report with other statisticians facing similar challenges. Thus, the report must also contain a sufficient level of detail for you and other statistician colleagues.

***BASIC STRUCTURE:*** A good report usually follows this basic structure:

1. Introduction
2. Main Body
3. Summary/Conclusions
4. Appendix
5. References (if needed; see below).

This organization makes it easy for readers with different interests to find information presented at a level appropriate for their backgrounds. We now describe what would ordinarily go into each of these sections.

***INTRODUCTION:*** Even if everyone who will be interested in the report is deeply familiar with the problem, data, and questions, it is always important to provide the following, so that the report stands entirely on its own:

- Subject matter background – what is broad scientific context and what are the challenges and unresolved issues? Here, the report should give a short description of the subject matter problem and why it is important in the domain science area.
- A brief summary of the study carried out to address the challenges.
- A statement of each of the specific scientific questions to that will be addressed.
- A "high-level" summary of the conclusions of the data analysis in the context of the subject matter area.
- A brief roadmap for the rest of the report indicating what can be found in each subsequent section.

***MAIN BODY:*** The main portion of the report can be organized in whatever way makes the most sense to you given the nature of the study and questions. Here is one standard way.

- Detailed summary of the data. For continuous longitudinal data, at the very least, spaghetti plots of the data, perhaps separately by natural groupings (e.g., treatment group), should be presented. If the data are very large, random samples can be plotted. For discrete data, other summaries (e.g., for binary data, sample proportions at each time point) can be presented in tables. Missing data and any other features should be noted. Any information on what led to these should be discussed and implications for interpretation of summaries and plots should be noted (e.g., plots of mean outcome at different time points may not be based on the same numbers of individuals).
- If addressing each question involves a different statistical model, you may wish to have a separate section for each question in which you present the question, the model in which it will be addressed, and the analysis, and the results. If all questions are addressed within the same statistical model framework, you may wish to present the model first in its own dedicated section and then have a separate section for each question, stating the question, the analysis, and the results. You should use your best judgment as to what makes it easiest for a reader.

- For each statistical model, you should present a description of the basic features of the model, an explanation for why it is appropriate, and a summary of the assumptions it embodies and the extent to which those assumptions are satisfied for the data at hand. In particular, for longitudinal analysis, you need to explain why specialized statistical models and methods are required and why other, more familiar methods are not appropriate. You also need to explain the basic features of the model and how the question(s) can be stated in terms of the model. Modeling choices and assumptions and your rationale for them should be clearly stated; you might include plots or other summaries that support the model and assumptions. These should be related back to the subject matter; so, for example, if a model assumes constant within-individual variance of the outcome, why that assumption makes sense in the subject matter context and/or is supported by the data should be stated.

  In general, investigators are not interested in seeing lots of equations, formulæ, matrices, and mathematical symbols. Some who are more sophisticated statistically may be able to tolerate some symbols and equations, but many who are not well-versed in statistics will prefer a description of the model that is mainly in words, with perhaps a "hand wavy" equation or two that helps give a sense of the framework but is not precise. Be sure to define clearly any symbols you do use. Any statistical terminology and concepts that may be unfamiliar to the investigators should be defined and explained. So, for example, don't just say "within-individual correlation" without explaining what that term/concept represents from a subject matter point of view and why it is important.

  Describe the method used to fit the model and mention the software that was used (investigators will want to state this in any papers they write reporting on the study). The investigators will likely not be familiar with the methods, so provide an description in nontechnical terms of the basic premise of the method; e.g., "this is similar to least squares for fitting a conventional linear regression model, but takes into account the correlation among longitudinal outcomes on the same subject."

- For a given question, state how the question can be represented in terms of the model. Present relevant numerical results and interpret them in the context of the subject matter. So, for example, do not give a p-value and say "so we reject $H_0$." Explain results in terms of the science, and present all important information. Never present an estimate without an associated standard error, and comment on the quality and precision of the results. If there are graphs or other data summaries that shed light on a result, present/discuss them.

- For each analysis, point out any limitations or caveats. For example, if results are predicated on certain assumptions embodied in the model for which there is not substantial support, comment on how robust a conclusion might be to violation of those assumptions.

- ***Do not*** include code or raw output! It is fine to summarize results in a table, but the table should not just be the raw output from software. All columns and entries should be explained in a table caption, with additional explanation if needed in the text. Code and output belong in an appendix to the report; see below. Never instruct or expect readers to go look at these; everything that investigators need to read should be in the main report.

***SUMMARY/CONCLUSIONS:*** As with any report on any topic, there should be a section providing an overarching summary of the objectives and results. This final section should present the scientific questions again, the conclusions of the analysis, and the interpretation of them in terms of the subject matter. Discussion of the implications of the results for the science; any additional observations or findings that, while not directly related to the questions, seem interesting; and possible future studies suggested by the analyses and results can be given here.

***APPENDIX:*** An appendix to the report contains technical details and supporting information. Typical things that would be presented include:

- A detailed description of all statistical models, with all symbols precisely defined, and precise statements of the assumptions that were made. This can be written for a statistician with general knowledge of statistics but perhaps not deep familiarity with the models and methods used.

- A precise, technical explanation of the methods that were used, including how they were implemented (e.g., software used, with any special options or assumptions noted).

- Code implementing the methods; it is prudent to document all code with extensive descriptive comments, which will make it easier for interested readers to understand it (and for you to understand what you did if you revisit the project or want to use it as an example of what you wish to do in a future project). Output should also be included; if this is voluminous, you may want to present only the key parts of the output, but at the very least all important output should be included.

- Additional data summaries, tables, figures, that might be of interest but are not directly relevant to the results.

It is fine to refer readers to the appendix in the main part of the report for more information and details, but looking at the appendix should not be required. All necessary information for investigators to understand what was done and the results should be in the main report. As above, never refer investigators to output or code.

***MISCELLANEOUS:*** Some additional items:

- If any literature (papers, books, software documentation) is cited, there should be a References section with full information on each, in a consistent format.
- It goes without saying that a report should be typed.
- There should be no misspellings or grammatical errors; always spell check any report before finalizing it!
- The entire report should be organized in a logical fashion, with section headings that make it easy for a reader interested in a particular result or question to locate that portion.
- Keep the writing straightforward and to the point, but not to the point that it is so brief that important information is obscured or missing. Check for run-on sentences and avoid language that is too flowery and wordy. Do not use terminology or words that are unfamiliar to a likely reader unless it is necessary, and, if you do, define them. If you use acroynms, present the entire term the first time you use it and define the acronym; e.g., “generalized estimating equation (GEE).”
- The narrative should flow naturally and “tell a story,” so should be easy to follow. To not go off on tangents regarding details; place details that are not central to the flow of the narrative in the appendix and refer to them.

Overall, good, clear writing is essential. Good report writing, both the writing itself and the knack for organizing the information in a sensible, logical way, is a skill that some people are born with but most people must learn. Use each report you write as an opportunity to develop and hone your skill. Good report writing skills will serve you well in not only collaborative work but in writing statistical research papers.