Statistical Consulting

Karen Lamb, David Price, Jono Tuke, and Nicole White

2024-11-29

Table of contents

Preface

Our goal in this book is to create a resource for statistical consultants - both new and old in the tooth. Often these skills “go without saying”, but once any two statistical consultants get together, there will be lots of

> “Oh I like that - I might steal it”

So we thought, let’s create an online version of that. Hence, if you feel that you have

- advice,
- tips,
- tricks, or
- horror stories from the front-line,

then please feel free to add.

Remember that these chapters are not **THOU SHALL**, but more, *maybe you could.*

1 Introduction

So you have your degree, master, PhD in statistics and have been asked by a researcher to help her analyse her data - now what. You know your Kruskal-Wallis from your KS-test and are proficient at R/Python/Julia/`<INSERT LANG HERE>`, but the researcher is talking about sample sizes, grants, microarrays, bone samples, surveys, etc. etc. etc. and you do not know where to start. Welcome to the wonderful, exciting, scary world of statistical consulting. So before we start, let's introduce ourselves.

1.1 Prof. Karen Lamb

1.2 Assoc. Prof. David Price

1.3 Dr Jono Tuke

1.4 Dr Nicole White

Part I

Part 1: Management skills (people, projects and money)

2 Effective communication as a statistical consultant

2.1 Introduction

One of the most challenging things about embarking on a career as a statistical consultant is learning how to communicate effectively with others, whether they be statisticians, researchers from a different discipline, government officials or members of the public. Almost all statisticians receive requests for assistance with problems involving data analysis or applying statistical methods, even if not consultants. Yet, effective communication skills are not an essential part of statistical training, with few undergraduate or postgraduate statistical programs offering this as part of the curriculum [REF]. While mathematics, statistical design, methodology, programming in statistical software, analysing data and interpreting findings are, of course, essential for statisticians, communication skills (whether oral or written) are rarely emphasised. Where taught, communication skills often focus on presentation skills (e.g., delivering a talk at a conference setting) or providing a pitch (e.g., how to pitch your research or work to a general audience) [REFS]. Although useful, these courses offer training for very specific aspects of the job. Little training is offered on the critical communication skills statisticians require day-to-day, such as clearly describing statistical designs, methods or findings to those not trained in statistics, or writing concise reports avoiding non-technical language. These (sometimes termed 'soft') skills are argued to be difficult to teach or something that should be learned 'on the job'. However, this type of training is typically offered to other professionals. For example, an essential component of training for General Practitioners (GPs) is on developing skills to effectively interact with patients, with practice scenarios to help refine these skills. Therefore, these skills can be taught and practised before beginning employment.

> ℹ Note
>
> [TO WRITE: PARAGRAPH ON EXISTING COMMUNICATION TRAINING FOR STATISTICIANS]

In this chapter, we outline some guidance on communicating effectively as a consultant statistician from our own experience, and that of others, within the discipline. This chapter will primarily focus on oral communication, although we will describe how written communication may assist with some aspects. This chapter is not intended to replace any formal training.

However, it offers advice on our thoughts on essential communication skills which may help those embarking on a career in statistics.

2.2 Establishing trust and managing expectations

Initial requests for statistical support can occur in many ways, from a colleague bumping into you in the corridor with a “quick question”, a surprise phone call from a new client or collaborator, a (brief or lengthy!) e-mail, or through a formal process (e.g., online inquiry form). Irrespective of how the request is placed, our advice is to arrange to meet any new client or collaborator, preferably in person, at a dedicated place and time to have a focussed discussion about their query. Within the Methods and Implementation Support for Clinical and Health research Hub (MISCH) at the University of Melbourne, most requests are received via an online inquiry form (REF) which our dedicated MISCH manager monitors. Our MISCH manager then issues an e-mail to the person requesting assistance which contains standard information to help manage expectations and understand our processes (see #sec-email). This e-mail contains information about how our team provides support and our costing model (see #sec-agreements). This ensures that the new client should be aware of what help we can offer in advance of the meeting to help ensure the first meeting goes smoothly.

2.3 Preparation for the initial meeting

Ask the new client to provide you with written information about the study or problem in advance (e.g., study outline or protocol, article on similar work undertaken elsewhere), ideally with a clear research question which may be refined in the initial meeting. If this information is provided, it is essential that you review it in advance of the meeting and outline any thoughts or questions you have. This preparation is an important part of establishing trust. Do not underestimate the time that you may need to prepare for a meeting. In addition to reading any materials provided by the client, you may wish to undertake some further preparatory work yourself. While this should not be overly onerous, given that projects may change scope during or after the initial meeting, this will provide you with more confidence about questions you want to ask when you first meet. As a guide, the Methods and Implementation Support for Clinical and Health research Hub (MISCH) offers up to 4 hours of preparation time for new consultations [REF: MISCH collaboration agreement]. Three or 4 hours is often not necessary. However, we typically try to allocate at least an hour prior to the initial meeting to familiarise ourselves with the study topic.

2.4 The initial meeting

It is essential to be courteous when meeting new or existing clients and collaborators. Be sure to introduce yourself and ask for the new client to tell you about themselves. It is important to determine the client's level of knowledge of study design and statistics to be able to communicate at an appropriate level. Be mindful that you may have to try another approach if it is apparent that the client does not appear to understand you. The first meeting should be used to understand the problem to be addressed and what the statistical needs are. Be aware that it is not necessary to solve the problem in the first meeting. Indeed, you should not be forced into providing answers during the meeting if you are unsure of the best approach. It is perfectly acceptable to acknowledge that the problem is an interesting one that you have not encountered before and you will consult with your team and colleagues before following up with suggestions.

2.5 Valuing the experience you bring

2.6 Meeting in person

> ℹ Note
>
> [WRITE MORE ABOUT WHY TO MEET IN PERSON] ADD ONLINE MEETING TECHNIQUES AS BECOMING MORE COMMON

2.7 Communicating with different clients

2.8 Guiding communication pathways

2.9 Bad practices and how to avoid them

2.10 Negotiating: the 'no', 'not right now' or 'not like that'

3 Graphical communication

- Clear, Simple, Easy to read
- Know your audience
    - Including 'who' will be looking at it, and whether for paper vs presentation
- Reduce unnecessary detail (e.g., don't need both x-axis and colour)
- Don't include aesthetics that don't map to anything (e.g., colour without legend)
- Choose the 'right' plot type
    - Bars v dot, should lines join dots (i.e., do you want the reader to interpolate?), stacked v dodged bars, etc…
- Think about the message you want the figure to convey
    - Including unit of inference – e.g., aim to show data/individual-level variation (show quantiles), vs. inference on a quantity (show CI)
    - Facet order – do you want reader to compare (e.g.) heights of bars left-to-right? Or location of time series/points top-to-bottom?
- Don't mislead reader (e.g., constrain y-axis limits to exaggerate differences, y-axis discontinuity to show all)
- Avoid trying to make the figure "do too much" – may make it hard to interpret for naïve reader
- Use some annotation to improve readability/interpretation
- 'Better' figure types
    - Donut plot rather than pie
    - Beeswarm or violin plot in place of boxplot
    - Points with error bars in place of detonator plots
- Publication ready figures
    - setting up scripts to automate figure output, format, DPI, etc. according to journal requirements?
    - clear captions that describe all features – tell the reader how to read the plot

3.1 Good references

- Zaloumis et al. (2015)
- Gordon and Finch (2015)
- Rougier, Droettboom, and Bourne (2014)

Part II

Part 2: Technical skills

4 Project structure

So welcome to the opinionated chapter on project structure. First a comment, I am not stating that this is the only way to organise a project, feel free to have your own system, as long as you have a system. This is my system as I write. It will change as I learn from my mistakes. What I want you to do is not follow it slavishly, but thing about the underlying principles and how you want to implement them.

4.1 Why a project structure?

So first the why. Why do you need a project structure? Well, as you become more experienced as a statistical consultant, then you will get more and more projects. This creates two problems:

1. You may be working on more than one project at a time, and you need to be able to move between them cleanly.
2. You may not come back to a project for many months. The longest so far for me returning to a previous analysis was 10 years.

In both cases, having a simple standard structure, means that you can get upto speed easily. This gives us stats consulting Rule 1

Stats consulting Rule 1

- Be organised to look after future you.

In the following, I will outline how to set up your folder stucture in R, and also guidance on why I use this structure.

4.2 Setting up a project stucture in R

First we will set up the project using the `usethis` package:

```r
usethis::create_project("~/Desktop/2025-penguin-analysis")
```

So a couple of things to notice, with my naming of the project, in this case

```
2025-penguin-analysis
```

First I have given the year that I did the analysis. This is useful, I have a massive archive of my previous analysis stored online, and it is often easiest to indentify when the analysis was done from previous emails with the collaborator. Once I have the year, then this reduces my search space. Next, I try to give it a memorable name to remind me of the analysis. In this case, we will assume that we are analysing the Palmer penguins dataset - this should be rememberable enough. I will also show you later how to add this info into your reports in Section 4.5.

So before we create any more of the structure, lets consider our goal- think self-contained (Figure 4.1). In the Kei-tora gardening contest, the contestents make a self-contained garden on the back of a trailer. Replace garden with analysis and trailer with folder. You should be able to send your folder to a colleague and they can reproduce your analysis. Nothing missing.

```r
knitr::include_graphics(here::here("figs", "kei-tora.png"))
```

DoctorPopular.RSS
@DocPop

The Kei-tora Gardening Contest is an annual event put on by the Japan Federation of Landscape Contractors to see who can build the best looking garden on the back of a Kei truck. Here are some of the winners from the past 12 years. zoenren-osaka.jp/activity.html

12:52 AM · Jun 11, 2021

Figure 4.1: The Kei-tora gardening contest

So now we are going to add some sub-folders for out analysis:

```r
fs::dir_create("raw-data")
fs::dir_create("data")

fs::dir_create("figs")
fs::dir_create("tabs")

fs::dir_create("r-code")

fs::dir_create("reports")
```

So, let's break down these subfolders.

First, notice that I have two subfolders for data:

- `raw-data`, and

- `data`.

The subfolder `raw-data` is for the data that I get from the collaborator - this is dated and named (Section 4.3) and is then left alone. I can read it into R, but I do not change it. I repeat: I. Do. Not. Change. It. You will be tempted, a quick adjustment will make your like easier. If you need to do this, then you change it, save the new file in the subfolder `data` and record what you did in your README file (more about this later Section 4.5). Our goal is to make our analysis reproducible and open - unrecorded tinkering of data is neither.

The subfolder `data` is for any data structure that you make yourself. While we are talking about data. Let's talk about `.RData`:

> Rant 1: Stop using .RData
>
> Do not save `.RData` when you quit R and do not load `.RData` when you fire up R. The data you use in your analysis should be purposely loaded by you, not left as a side-effect.

The next two subfolders are `figs` and `tabs`. These are for all of the figures and tables from the analysis. I keep these seperate so that I can keep them easily, and also it makes it easy to share with the collaborators.

So now we get to some code file, these are put in the subfolder `r-code`. If I have python code, then we have a subfolder `py-code`, and a subfolder for julia: `julia-code`. I like to keep these seperate, but it is not necessary.

The next and final subfolder for now is `reports`. This is where I write my reports for the collaborator.

Now we have some subfolders, we can now populate them with some files. So lets talk about naming files.

4.3 Naming things

It is a good idea to keep you file names informative. Here are some guidelines for different types of files.

Data: So all data I receive is given a name of the form

```
2025-02-25-penguin-data.xlsx
```

So let's break this down. First the filename starts with a date entry of the form

```
year-month-day
```

This is so that the files are ordered in my folder structure, so that at a glance I can work out which is the latest files.

Analysis: If I decide to not use targets (Chapter 5), then I will name my analysis files in the order they should be run, and also informative about what they do

```
01-read-in-penguin-data.R
02-clean-penguin-data.R
03-fit-log-reg-penguin.R
```

Again, this is so that in the future if I get new data or change an analysis, then I can see the dependencies and know what needs to be rerun.

HERE::HERE

So you are going to start worrying about all these subfolders and working directories. Well worry no more, as we are going to use projects, then we can use the ever so useful

```
here::here()
```

command.
What does it do, just give the place in your directory where your project is

```
here::here()
```

```
[1] "/Users/jonathantuke/Dropbox/01-projects/statistical-consulting-book"
```

So if you want your data file in the `raw-data` folder, then you can use

```
here::here("raw-data", "2025-02-25-penguin-data.xlsx")
```

```
[1] "/Users/jonathantuke/Dropbox/01-projects/statistical-consulting-book/raw-data/2025-02-2
```

Move your project to somewhere else in your computer - it will still work.

4.4 HERE

```
# fs::dir_tree()
```

4.5 README file

5 Using targets

6 Summary

In summary, this book has no content whatsoever.

References

Gordon, Ian, and Sue Finch. 2015. “Statistician Heal Thyself: Have We Lost the Plot?” *Journal of Computational and Graphical Statistics* 24 (4): 1210–29. https://doi.org/10.1080/10618600.2014.989324.

Rougier, Nicolas P., Michael Droettboom, and Philip E. Bourne. 2014. “Ten Simple Rules for Better Figures.” *PLOS Computational Biology* 10 (9): e1003833. https://doi.org/10.1371/journal.pcbi.1003833.

Zaloumis, Sophie G., Freya J. I. Fowkes, Alysha De Livera, and Julie A. Simpson. 2015. “Presenting Parasitological Data: The Good, the Bad and the Error Bar.” *Parasitology* 142 (11): 1351–63. https://doi.org/10.1017/S0031182015000748.

A Resources

A.1 Email example

A.2 Agreements

MISCH collaboration agreement