#### Adventures in statistical consultation

#### Sean van der Merwe

University of the Free State, South Africa

10 May 2019

DON'T PANIC





### Overview

What do I do and why?

Staying efficient

Conclusion

# University internal consultation

- I spend the majority of most days working on projects brought to me by researchers at the UFS
- Most of the clients are busy with a research masters degree
  - Second biggest category is PhD students
  - Then personnel who are engaged in research projects
  - And then there are fringe cases
- The fringe cases and personnel projects are the most interesting
- Departments that have approached me this year include:
  - Soil, Crop, and Climate Science; Consumer Science; Microbial Biochemical And Food Biotechnology; Plant Sciences; Institute for Groundwater Studies; Genetics:
  - Centre for Development Studies; Student Affairs; Student Counselling And Development; Urban and Regional Planning; Social Responsibility Enterprises; Business Management Quantity Surveying; Centre for Teaching and Learning;

#### Time breakdown

- 40% is interacting with clients
  - Explaining p-values,  $\chi^2$  tests, science in general
  - Asking questions like:
    - What is your research problem?
    - Why is your data not connected to your research problem?
    - When do you need the results anyway?
- 30% is interacting with data
  - Separating data and metadata
  - Cleaning data, solving problems, creating more metadata,
  - Dealing with strange formats, etc.
- 20% is writing code
  - This is the fun part in most cases
- and 10% is pulling out my hair
  - I try to do this in private, mostly



10 May 2019

# Summary of projects so far

- Almost half of the 30+ projects so far involve a survey!
  - This is not a good thing.
  - A bad survey tells you only about the people that answered the survey. A good survey tells you only about people like the people that answered the survey. To get information about the general population you need a designed experiment. A survey can be raised to the level of a designed experiment with careful planning and execution, but I've not had the privilege of seeing this in person.
- At least 80% come to me only after collecting data
- Less than 20% require anything even remotely interesting from me
  - I am generally just required to make the assessors happy





### But what about the exceptions?

- A few projects are really fun and make all the p-value calculations worth it, let's look at a few
  - Let's start with a meta-analysis
    - It's kinda like a survey, but you're surveying research reports instead of people
    - In this example we were studying the effects of nitrogen and phosphorous fertiliser on Biological nitrogen fixation (BNF)
    - Tendai summarised a lot of studies, converting the results to a single scale. She also calculated variance estimates of the results, allowing us to combine studies in a sensible way by weighting them appropriately
  - Other fun projects involve biplots or tree diagrams
  - The nicest ones are where the data is collected systematically on a large scale
  - But the most impactful has probably been a bit of VBA code I wrote for the CTL



### Overview

What do I do and why?

Staying efficient

Conclusion

# And by efficient I mean lazy

- I hate doing things manually
- So every time I do something new I work hard to make sure that the next time I do it it's much easier
- That means making notes, explaining what I do properly, including links as I go, commenting code,
- and most importantly:
- Writing generalised code
  - The job of a statistician is to turn data into information and then into knowledge
  - As a consulting statistician, my focus is on the first part,
  - BUT neither the data nor the information belong to me,
  - SO neither have any place in my code



### Surveys are all the same

- The majority of surveys are created by novices
- Most follow the same setup:
  - Way too many demographic questions
  - Likert scale questions
  - Yes / no questions
  - Comment / other? questions
  - Other scales that people make up
- If you know which questions are which and what the scales are, then each time you need to analyse a new survey you can just take a generic template and delete the parts you don't need



10 May 2019

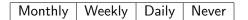
### The hard part

 The biggest task each time I get a survey is to create sheets of metadata (clients assume this happens magically)

Qnum	Qpart	Qshort	Qdesc
Q1	Demographic	Consent	Consent
Q2	Demographic	Age	Age
Q3	Demographic	Gender	Gender
Q4	Demographic	Ethnicity	Ethnic Group
Q5	Demographic	WorkTime	How long have you
Q6	Likert	Vision	I fully understand
Q7	Likert	Informed	I received information
Q8	Likert	WhyHappen	I was given enough UNIVERSITY OF THE UNIVERSITY

### The more systematic the better

- I create a list of comparisons that need to be tested and loop through them
  - Better than coding each manually because no code duplication makes debugging much easier, and it's easy to add or remove comparisons
- I create a list of scale transformations that need to happen
  - I implement them after loading the data
  - Much better than modifying the data because I can make changes quickly and easily without having to revert previous work
- I create a list of final scales in the correct ordering
  - I regularly have to explain that this is not a good ordering:





#### Markdown

- I have done every single consultation project that I've worked on this year in R Markdown
- The amount of time and effort I've saved by doing this can be measured in days, perhaps weeks
- It means no copy-paste of graphs or tables, no redoing of formatting
- It means that if something changes in the data that changes all the graphs and all the tables and all the p-values then fixing my report takes a few seconds
- Even my results (like Significant vs Not significant) are automatic
  - the only exception is any interpretation of results (practical implications) which is generally done with or by the client anyway

#### Overview

What do I do and why?

Staying efficient

Conclusion

### The point

- Consultation is fun because I regularly encounter new challenges to overcome
- Consultation is important because I streamline the scientific process
- Statisticians help at the observation phase of the scientific method by helping people understand what they observe
- And they help at the testing phase by helping people to stop talking so much nonsense



10 May 2019