STATS 101: a modern introduction

Dr Ian Hunt

[1]Manager, Statistical Consulting Service
Monash University
ihunt@bunhill.co.uk; ian.hunt@monash.edu; statisticalconsulting@monash.edu

November 15, 2020

1 Introduction and reality check

2 Session 1: Estimation and Sampling Variance

3 Session 2: Inference

4 Session 3: Data Analysis

Course outcomes

1. Be realistic.
2. Use a computer to properly analyse scientific data (which entails going beyond Microsoft Excel).
3. Make justifiable and reproducible statistical inferences using real data.
4. Plan ahead for the statistical analysis of data for journal articles.

Table of Contents

1 Introduction and reality check

2 Session 1: Estimation and Sampling Variance

3 Session 2: Inference

4 Session 3: Data Analysis

Be realistic

- *"Methodology, like sex, is better demonstrated than discussed, though often better anticipated than experienced."*
  Leamer (1983, Lets take the con out of econometrics, page 40)
- Analysing your PhD data is going to take you **hundreds of hours**.
- This is the *start of a journey.*

Philosophy of Applied Statistics?

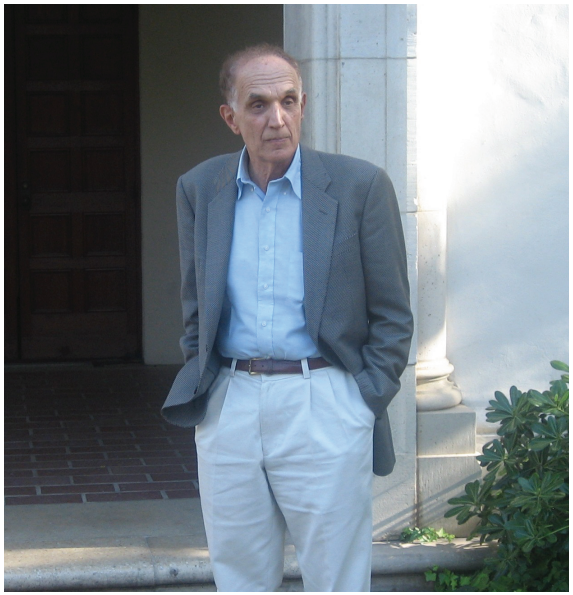

Figure 1: Bradley Efron.

- For over 50 years, Bradley Efron has been a leading researcher in the famous Stanford University Statistics department.
- Bootstrapping, named from Efron's favourite Baron Munchausen story, was initially treated with scepticism. We now take it for granted, though he probably doubts that we should.

"I have a feeling that statisticians are cynics, because you realise how much of the stuff that you are told is true in the world is actually just that month's accident that worked out, or that month's disaster that happened. Appreciating how much randomness there is in everyday experience helps a lot." Efron (2010)

*Who am I?

- My name is Dr Ian Hunt and I have managed the statistical consulting service at Monash since Feb 2019.
  - I have been a statistical consultant for 20 years, and I am a Chartered Statistician (CStat) and Fellow of the Royal Statistical Society. I hold post-graduate degrees, with honours, from Université Catholique de Lille (France), the London School of Economics (London), City University (London) and Otago (New Zealand).
- I am fighting to save one-to-one consulting for HDR students. Short courses are not enough on their own.
- My accent in English, French and Spanish is terrible.

*One-to-one consulting cannot be replaced by courses

Course feedback from October 2020: courses can help but are not enough.

- *"Fore me- while 3 hrs seems long- that was because I found the concepts quite tricky- I realise I need more time- and I cannot be the only one. Any research is greatly enhanced and more comprehensive (& publishable) with robust stats- WE NEED one on ONE stat help - I know I do as I write my first PhD paper."*
- *"I need the ability to ask individual questions. If the university wants meaningful research from Ph.D. students, it MUST make stats support a priority and ensure all students have access to one on one support and guidance to ensure robust methods are used and manuscripts progress to acceptance."*
- *"Almost everything on myDevelopment is a total waste of time. But this stats course was really good—and I'm really looking forward to the next one (a modern introduction). Ian has a great teaching style directed at major take away issues and then encourages attendees to take the next step through further reading."*
- *"Ian clearly knew exactly what he was talking about and this was the most interest I have ever had in statistics - it has always been a guilty gap in my knowledge that I've avoided but Ian presented it in a lucid and understandable way that intrigued me rather than scared me. However, I don't think that this single session is anywhere near enough to consider that I can now 'do' stats. It was a good introduction, but this is an entire discipline and I need clear guidance to wade through it."*
- *"Further to Ian's question about whether there were any students with maths-anxiety (and poor schooling) present – I'm one! As such, I thought it might be useful to know that I found the first session really informative and the readings very enlightening (though they were very slow going due to my lack of familiarity with the concepts). In particular the p-value paper really got me understanding that it isn't as hard as it sounds. However, because the ideas were so unfamiliar, it took me close to the whole day to read ... I thought it might be uncommon for PhD students to admit to how challenging some of this material can be – or maybe to the level of deficit some of us have."*

I am not perfect

Negative course feedback from October 2020

- *"While the lecturer was very knowledgeable, for someone who has no foundation in statistics, this introductory course was a mess, with the lecturer assuming that the students have a lot of assumed knowledge and understandings all his technical terms. In additions, the lecturer spoke way too fast, was mumbling, had confusing slides, and did not explain any of the terms that he used in a way that actually made sense for someone taking a 101 course!"*

Finally: keep these ideas in mind

- Data summaries come before complex models.
- Lump and split data groups dynamically.
- Be flexible with "data types".
- Keep naming conventions simple and recognisable.
- Comment on any code and analysis steps that you take.
- Re-use your code and functions.
- Start with simple analysis and *then* add complexity.
- Test code to ensure you get the right answers.
  - Compare your R code with what you get in SPSS, for example.
  - Use fake data or small subsets for testing.

Table of Contents

1 Introduction and reality check

2 Session 1: Estimation and Sampling Variance

3 Session 2: Inference

4 Session 3: Data Analysis

Topics and resources (Session 1)

- Gentle introduction to R and RStudio
  - *R for Statistics* notes (especially chapter 2).
- Overview of hypothesis testing
  - *Statistical Inference for Research* notes (chapter 2).
- Random variables and probability.
  - R code developed in class.
- Simulation and resampling.
  - R code developed in class and *Statistical Inference for Research* notes (chapter 1).
- Means, the central limit theorem and normal distributions.
  - R code developed in class and the example from Spiegelhalter (2019, chapter 7).

Table of Contents

1 Introduction and reality check

2 Session 1: Estimation and Sampling Variance

3 Session 2: Inference

4 Session 3: Data Analysis

Topics and resources (Session 2)

- Theoretical sampling variance.
  - *Statistical Inference for Research* (chapter 1).
- Bootstrap sampling variance
  - *Statistical Inference for Research* (chapter 1) and R code from class.
- t-tests, p-values and confidence intervals.
  - *Statistical Inference for Research* (chapter 3) and R code from class.
- Size and Power.
  - *Statistical Inference for Research* (chapter 3) and R code from class.
- Multiple hypothesis testing.
  - *Statistical Inference for Research* (chapter 4) and R code from class.

Table of Contents

1 Introduction and reality check

2 Session 1: Estimation and Sampling Variance

3 Session 2: Inference

4 Session 3: Data Analysis

Topics and resources (Session 3)

In this session we will use the notes called *data analysis*.

- Data summaries.
- Histograms.
- QQ-plots.
- Boxplots and "error bars".
- Correlation measures, outliers and ranks.
- Mann-Whitney/Wilcoxon tests.

Course outcomes

1. Be realistic.
2. Use a computer to properly analyse scientific data (which entails going beyond Microsoft Excel).
3. Make justifiable and reproducible statistical inferences using real data.
4. Plan ahead for the statistical analysis of data for journal articles.

Bibliography I

Efron, B. (2010). Editorial interview. *Significance (Royal Statistical Society)*.

Leamer, E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*, 31–40.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Penguin UK.