

# Preparations for Statistical Research

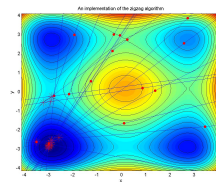
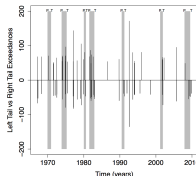
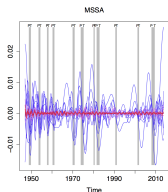
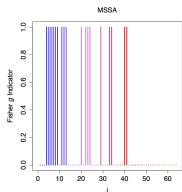
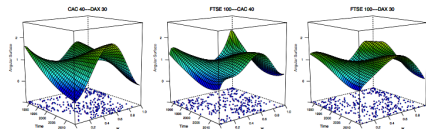
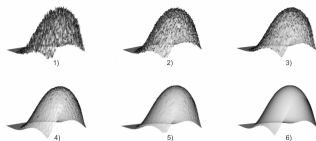
Miguel de Carvalho

Lecture 1—Introduction

# Introduction

## About Me

- I am an applied mathematical statistician with a variety of research interests including, *inter alia*, Applied Statistics, Biostatistics, Econometrics, Risk Analysis, Statistics of Extremes.



- More details on my research can be found on

<http://www.maths.ed.ac.uk/~mdecarv/>

# Introduction

## What is Today's Class About?

- Today's lecture will offer an **introduction to research** in Statistics.
- Virtually any business can profit from rigorous, critical statistical thinking, often referred to nowadays under the buzzword '**Data Science**.'
- Statistical research is both an enterprise of academic and industry interest. A wealth of decisions while you are reading this paragraph are being made on the basis of statistical research by policy-makers, investors, *etc.*

*"**Statistical thinking** will one day be as necessary a qualification for efficient citizenship as the ability to read and write."*

*Samuel Wilks, quoting H. G. Wells.*



- We have reached that day long ago.

# Introduction

## Best Practices

- Best practices of statistical research include:
  - **Reproducibility** of the statistical analysis.
  - **Clarity** about *exactly how* the analysis has been conducted.
  - **Assessment** of strengths and limitations with the analysis.
  - **Checking** and assessing possible consequences of assumptions.
  - **Audit** your data (Can your data be trusted?).
  - ...
- To deliver all this you need to master theory and applications of modern statistical methods, to program effectively and be proficient with computing, to have great writing skills—among other things.

# Introduction

## Structure of this Course

- This course will be organized as follows:
  - Week 1: Preparations for statistical research.
  - Week 2: Critical statistical thinking.
  - Week 3: Simulation Studies.
  - Week 4: How to model it?
  - Week 5: Communicating findings.
  - Week 6: Professional ethics for statisticians.

# Choosing a Good Statistical Problem

How and Where do I Start?



## Quiz: How to choose a good statistical problem?

- To choose a good scientific problem, consider *regularly*:
  - $\text{input}_1$ : Reading **scientific papers** and **research monographs**.
  - $\text{input}_2$ : Attending **research talks** and **conferences**.
  - $\text{input}_3$ : Discussing ideas, concepts, and methods with scientists and applied professionals.
- Talks, papers, and collaborations with peers from other fields can also contribute in some cases to push-forward the limits of our discipline or of other disciplines.

# Introduction

## But What is Statistical Research?

- Statistical research is the activity of mapping all these inputs into new concepts, new approaches that translate into
  - fresh perspectives,
  - new knowledge,
  - game-changers, and
  - paradigm shifts,

in science and / or industry.

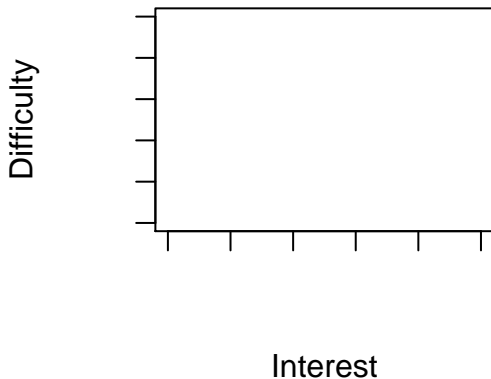
- Conceptually,

$$(\text{input}_1, \text{input}_2, \text{input}_3, \dots) \mapsto (\text{new concepts, new methods, } \dots)$$

# Introduction

## 'Difficulty-Interest' Diagram

When choosing a research problem keep in mind the following 'Difficulty-Interest' diagram:





# Introduction

## What is?

- Ideally your problem or approach should be ‘interesting’ *but* ‘non-trivial’.
- The right problem to tackle depends among other things on:
  - The stage of your career.
  - The team you are involved in.
  - Constraints (e.g. subject of your thesis; fixed time to finish your thesis).
- Never sacrifice ‘Interest’ in favor of ‘Difficulty.’ Always keep in mind that *not* all difficult problems are interesting.
- Alon (2009) provides further guidance with regards to the latter tradeoff.
- But let’s get back to the inputs.

# Input<sub>1</sub>: Reading

How does a statistical paper look like? Here is an example:

<http://onlinelibrary.wiley.com/doi/10.1111/rssa.12338/epdf>



J. R. Statist. Soc. A (2017)

## Non-parametric evidence of second-leg home advantage in European football

Gery Geenens and Thomas Cudihy  
UNSW Sydney, Australia

[Received January 2017. Final revision October 2017]

**Summary.** In international football (soccer), two-legged knockout ties, with each team playing at home in one leg and the final outcome decided on aggregate, are common. Many players, managers and followers seem to believe in 'second-leg home advantage', i.e. that it is beneficial to play at home on the second leg. A more complex effect than the well-documented usual home advantage, it is more difficult to identify, and previous statistical studies have not proved conclusive about its actuality. As opposed to previous research, the paper addresses the question from a purely non-parametric perspective, which is not based on any particular model specification which could orientate the analysis in one or the other direction. Along the way, the paper reviews the well-known shortcomings of the Wald confidence interval for a proportion, suggests new non-parametric confidence intervals for conditional probability functions, revisits the problems of bias and bandwidth selection when building confidence intervals in non-parametric regression and provides a novel bootstrap-based solution to them. Finally, the new intervals are used when analysing game outcome data for the UEFA (Union of European Football Associations) Champions and Europa Leagues from 2009–2010 to 2014–2015. A slight second-leg home advantage is evidenced.

**Keywords:** Confidence intervals; Football; Home advantage; Non-parametric regression; Undersmoothing

### 1. Introduction

'Home field advantage' in sport is well established (Schwartz and Barsky, 1977; Courneya and

# Input<sub>1</sub>: Reading

- Title.
- Abstract / Summary.
- Keywords.
- Bulk of the paper:
  - **Introduction**  
What is your problem and your solution? Why are they interesting? How does it connect with other developments.
  - **Methods**  
Description of statistical methodology (proposed / employed).
  - **Simulation Study**  
Numerical assessment of performance of methods in a simulation setting.
  - **Data Application**  
Showcase methods in a real data context.
  - **Discussion**  
Conclude and comment on shortcomings and extensions.

# Input<sub>1</sub>: Reading

What Data are Typically used in a Statistical Paper?

- Sometimes the goal is on capitalizing on interesting data applications (model second-leg home advantage, model Brexit, forecast bitcoin price dynamics, ...).
- Other times authors simply illustrate their novel methodology on a and widely studied data set (may facilitate comparison with previously proposed methods).
- Examples of the latter datasets are available from the R package datasets.



Figure: Source: BBC

# Input<sub>1</sub>: Reading

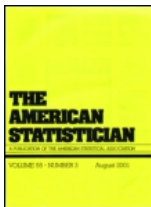
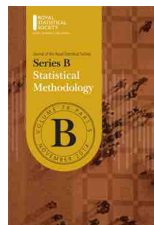
- In terms of readership, Statistical journals can be categorized as:
  - Broad Readership.
  - Application-Specific Readership.
  - Methodology-Specific Readership.
- For a list of Statistical journals see

[https://en.wikipedia.org/wiki/List\\_of\\_statistics\\_journals](https://en.wikipedia.org/wiki/List_of_statistics_journals)

- Some journals are more highly-regarded by the community than others.
- A common score for ranking journals the so-called **impact factor**.
- Although widely used, this metric has been heavily criticized; some advocate the use of **citation-exchange counts** (Varin et al., 2016).
- Scores do not read or scrutinize scientific papers. Thus, no score will ever be a substitute to how a community itself perceives the scientific standards of a journal.

# Input<sub>1</sub>: Reading

## Some Broad-Readership Statistical Journals



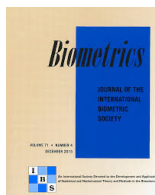
# Input<sub>1</sub>: Reading

Some Journals with Application-Specific Readerships



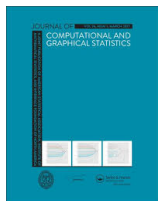
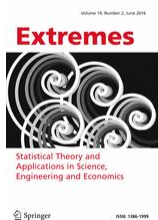
## Journal of Business & Economic Statistics

A Publication of the American Statistical Association  
Volume 25 Number 4 October 2007



# Input<sub>1</sub>: Reading

Some Journals with Methodology-Specific Readerships

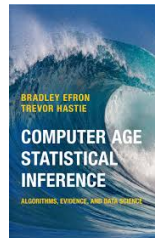
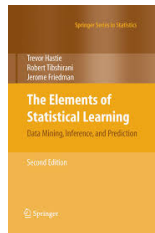
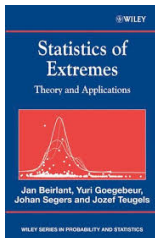
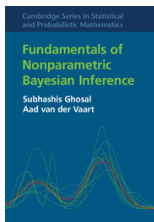




# Input<sub>1</sub>: Reading

## Research Monographs *versus* Textbooks

- A **research monograph** is a book devoted to research-level developments on a certain field; here are some examples



- **Textbooks** are books used for 'introductory courses'; they are appropriate for consulting and getting introduced to a subject (but not not the most useful references in terms of research).
- **Science divulgation books** are great for inspiring the next cohort, for sharing with the society the latest scientific developments (but not the most useful references in terms of research).

# Input<sub>1</sub>: Reading

## Where Can I Start My Literature Search?

- I will focus on two search engines, but there are many options available.
- You can use Google Scholar (<https://scholar.google.co.uk/>)



☒ Articles (☒ include patents) ☐ Case law


- Another option is JSTOR (<https://www.jstor.org>)



# Input<sub>1</sub>: Reading

## Where Can I Start My Literature Search?

- To be considered for publication, papers need to go under a peer review scrutiny (next lecture).
- Often researchers make submitted articles available in the repository arXiv (<https://arxiv.org/>).



Cornell University  
Library

arXiv.org

Open access to 1,349,224 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems and Control Sciences.

Subject search and browse:

02 Jan 2018: [1991–2017 submission rate statistics](#) are now available.  
See cumulative "What's New" pages. Read [robots beware](#) before attempting any automated download

### Physics

- [Astrophysics](#) ([astro-ph](#) [new](#), [recent](#), [find](#))  
includes: [Astrophysics of Galaxies](#); [Cosmology and Nongalactic Astrophysics](#); [Earth and Planetary Astrophysics](#); [High Energy Astrophysical Phenomena](#); [Instrumentation and Methods for Astrophysics](#)
- [Condensed Matter](#) ([cond-mat](#) [new](#), [recent](#), [find](#))  
includes: [Disordered Systems and Neural Networks](#); [Materials Science](#); [Mesoscale and Nanoscale Physics](#); [Other Condensed Matter](#); [Quantum Gases](#); [Soft Condensed Matter](#)
- [General Relativity and Quantum Cosmology](#) ([gr-qc](#) [new](#), [recent](#), [find](#))  
includes: [General Relativity and Quantum Cosmology](#)

M. de Carvalho

Lecture 1

19 / 25

# Input<sub>1</sub>: Reading

Where Can I Start My Literature Search?

- There are some social networks over which researchers share manuscripts, including Researchgate and Mendeley



- Software for managing references includes Papers3, Mendeley, Zotero, etc.



## Input<sub>2</sub> and Input<sub>3</sub>: Listening and Discussing

- Attending a Statistics Seminar or a Colloquium. Here is one example:

### Colloquium

- *Friday, Feb 2nd.*
- *Room and time: TBA.*
- *Prof. Sofia Olhede (UCL, Department of Statistical Science).*

- Or attending a Statistics conference. Here is one example:

### Conference (World Meeting of International Society for Bayesian Analysis)

24–29 June 2018, Edinburgh



<https://bayesian.org/isba2018/>

# Guidance on Extra Readings

## Roadmap

Here are some extra readings.



### **How do I start? How do I make progress?**

Hamada and Sitter (2004) provides some general advice for early-career researchers.



### **Rules for effective statistical practice**

Kass et al. (2016) discusses best principles of statistical practice.

# Guidance on Extra Readings

## Roadmap



**How can I become a professional researcher?**

Altman et al. (2017) is new researchers' survival guide.

# Guidance on Extra Readings

## FAQ in Statistics

Here are some recurrent FAQ on Statistics that have some interesting answers:



### **What is Data Science?**

Cleveland (2001) provides a unified view on what we now call Data Science; published a long-time ago by the *International Statistical Review*.



### **What are the differences between Statistics and Machine Learning?**

Breiman (2001) offers his view on the two cultures.



# Summary

## Roadmap

- Statistical research is an inquiry of interest for a wealth of players in industry and in the academy.
- In order to
  - Keep up with most recent developments,
  - Develop cutting edge research,
  - Be able to contribute with novel and influential ideas, concepts, and methods,

↪ it is important to understand what are the sources and channels along which these are broadcast.
- The next lecture will be devoted to critical statistical thinking.

- Alon, U. (2009), “How To Choose a Good Scientific Problem,” *Molecular Cell*, 35, 726–728.
- Altman, N., Banks, D., Hardwick, J., Roeder, K., Craigmile, P., Hardin, J., and Gupta, M. (2017), *The IMS New Researchers’ Survival Guide*, Institute of Mathematical Statistics.
- Breiman, L. (2001), “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author),” *Statistical Science*, 16, 199–231.
- Cleveland, W. S. (2001), “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics,” *International Statistical Review*, 69, 21–26.
- Hamada, M. and Sitter, R. (2004), “Statistical Research: Some Advice for Beginners,” *The American Statistician*, 58, 93–101.
- Kass, R. E., Caffo, B. S., Davidian, M., Meng, X.-L., Yu, B., and Reid, N. (2016), “Ten Simple Rules for Effective Statistical Practice,” *PLOS Comput Biol*, 12, e1004961.
- Varin, C., Cattelan, M., and Firth, D. (2016), “Statistical Modelling of Citation Exchange Between Statistics Journals,” *Journal of the Royal Statistical Society: Ser. A*, 179, 1–63.