Topic 1 – Introduction

ENVX1002 Introduction to Statistical Methods

Dr. Floris van Ogtrop
The University of Sydney

Feb 2024

THE UNIVERSITY OF SYDNEY

Acknowledgement of Country

I would like to acknowledge the Traditional Owners of Australia and recognise their continuing connection to land, water and culture. I am currently lecturing on the land of the Wangal people of the Eora Nation and pay my respects to their Elders, past, present and emerging.

I further acknowledge the Traditional Owners of the country on which you are on and pay respects to their Elders, past, present and future.

Emergency procedures

- In the unlikely event of an **emergency**, we may need to evacuate the building.
- If we need to **evacuate**, we will ask you to take your belongings and follow the green exit signs.
- We will move a safe distance from the building and maintain physical distancing whilst waiting until the emergency is over.
- In some circumstances, we might be asked to remain inside the building for our own safety. We call this a **lockdown** or **shelter-in-place**. More information is available at www.sydney.edu.au/emergency.

Health and safety advice

Illness & absence

- If you become ill during the semester, or need to stay at home or need to be absent for a period, please notify your unit of study coordinator.
- If illness impacts assessment, use the usual mechanisms including simple extensions and special consideration to arrange reasonable adjustments.

Student support

- Visit the Student life, wellbeing and support webpage to find out about the student services, resources and events available to support you while you study:
  - Health and wellbeing
  - Academic Support
  - Personal support
  - Getting connected
- Questions about getting started this semester? Come visit us at a Welcome Hub in Anderson Stuart or Carslaw Building.

Safer Communities Office

Support and case management for people who have experienced sexual misconduct, domestic/family violence, bullying/harassment or issues relating to modern slavery.

Contact the team:

- 8:30 am to 5:30 pm Monday to Friday, Sydney local time
- phone: +61 2 8627 6808
- email: safer-communities.officer@sydney.edu.au.
- campus: Level 5, Jane Foss Russell building, City Road, Darlington Campus

Make a report:

Visit the website to make a complaint or disclosure of sexual misconduct to the University.

Support services

- **The Office of Educational Integrity**: Report anonymously or seek advice: educational.integrity@sydney.edu.au
- **Learning Hub**:
  - The Learning Hub (Academic Language and Learning) offers workshops, online resources and individual consultations on study and writing skills.
  - The Learning Hub (Mathematics) offers bridging courses, drop-in services and online resources.
- **Library**:
  - Check out the Library's online resources and referencing and citation styles.
  - You can also chat with a Peer Learning Advisor about your studies, including referencing questions
- **Counselling and mental health support**: The University's Counselling and Psychological Services provide self and time-management workshops and online resources.
- **Special Arrangements and Consideration**: Apply for special consideration if impacted by short-term illness or misadventure
- **Disability Services**: Register for Disability Support
- **Student organisations**:
  - SRC (undergraduate students)
  - SUPRA (postgraduate students)

Do you have a disability that impacts on your studies?

You may not think of yourself as having a ‘disability’, but the definition under the Disability Discrimination Act (1992) is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities.

The types of disabilities we see include:

> Anxiety // Arthritis // Asthma // Autism // ADHD Bipolar disorder // Broken bones // Cancer // Cerebral palsy // Chronic fatigue syndrome // Crohn’s disease // Cystic fibrosis // Depression Diabetes // Dyslexia // Epilepsy // Hearing impairment // Learning disability // Mobility impairment // Multiple sclerosis // Post-traumatic stress // Schizophrenia // Vision impairment

and much more.

In order to get assistance, students need to register with Disability Services. It is advisable to do this as early as possible. Please contact us or review our website to find out more.

Disability Services Office (02) 8627 8422

Academic integrity

- **Academic integrity** refers to behaving honestly, ethically and responsibly in relation to all elements of your study at the university, including assessments.
- Always submit your own work, sit your own tests, and take your own examinations.
- Acknowledge any contributions in your assignment which are not your original thoughts, ideas or words.
- Academic Honesty Education Module – all commencing students must complete by census date. Continuing students can self-enrol at any time.

What is academic dishonesty?

The following are some behaviours that are academically dishonest:

- **Plagiarism** (this is the most common form)
- **Collusion** or illegitimate co-operation
- **Recycling** (using your own work from previous assessments)
- **Cheating**, including **contract** cheating
  - sharing questions or accessing solutions on online “help sites”
  - receiving coaching from a private tutoring company on how to complete an assignment
  - asking someone else to write your assignment (for payment or not)
- **Exam cheating** (using prohibited materials, working with others)
- **Fabrication** or falsification of sources, data or results

What are the consequences?

- The University has strong mechanisms for detection of potential academic dishonesty.
- Suspected breaches are reported to the faculty educational integrity team for investigation.
- The University is deeply committed to ensuring the integrity of its educational programs and treats integrity breaches seriously. As a result, the academic consequences for cheating are numerous.
- You may:
  - need to resubmit a task with a mark penalty or
  - receive a 0 for the assessment or even the unit of study
  - be suspended or even excluded from your studies for serious misconduct

Understanding contract cheating

Commercial cheating services are **ILLEGAL** in Australia. Illegal cheating services offer to:

- Sell you essays, assignments, study notes or exams
- Ask you to upload previous work from your course
- Sit exams on your behalf

If you use cheating services, you can face disciplinary action in accordance with USYD's policies. Resulting action can include:

- Failing the unit of study or course
- Suspension or exclusion from your studies
- Losing your professional accreditation
- Being blackmailed by cheating service operators
- For international students, losing your visa

Be aware of illegitimate services

- Be aware of any services that are not affiliated with the University.
- In the online environment, malicious organisations masquerading as “online help sites and platforms” are preying on students.
  - These organisations may pressure you to pay for online assistance, then turn to blackmail when you change your mind.
  - Essays or solutions bought from the internet are usually poor quality, badly written and often wrong.
  - You won’t acquire the skills and knowledge required for your degree, making it difficult to complete further assessments.

Tip

As a student, you can contact the Office of Educational Integrity to report something anonymously or seek advice.

Sitting proctored tests

- Exams and mid semester test will be face-to-face unless we experience circumstances that prohibit this.
- ProctorU software is used to monitor your conduct during an exam. Incidents are flagged to the University and reviewed for breaches of academic integrity.
- The exam will be compromised if you:
  - Use prohibited materials (e.g., headphones, mobile phones, etc)
  - Communicate or collude with others
  - Seek help via a third party, the university's sites or help sites

To ensure success, we recommend the following tips:

1. Sit directly in front of the camera
2. Review the online test support site on Canvas
3. Know what materials are permitted during the exam.
4. Have your ID ready
5. Don't wear headphones, either wired or unwired

Contact information

Floris van Ogtrop - unit coordinator

- Room 306, Level 3, Biomedical Bldg, Australian Technology Park, Eveleigh
- Ph: 02 8627 1024
- Email: floris.vanogtrop@sydney.edu.au

Teaching staff

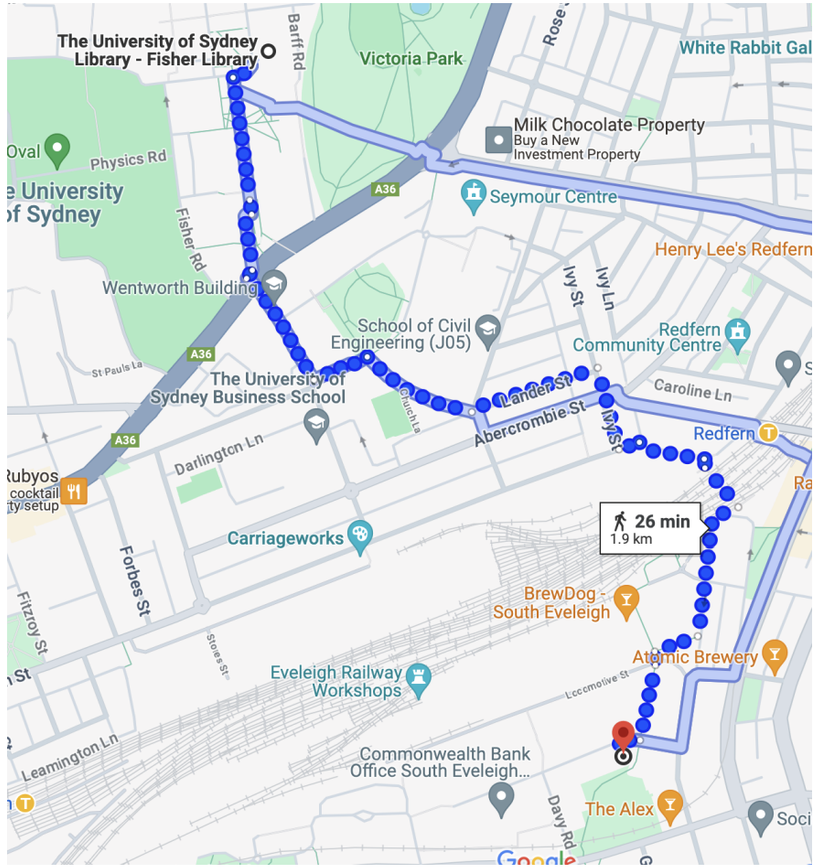
- A/Prof. Floris van Ogtrop - Module 1 (Weeks 1 - 4, 13)
- Dr. Aaron Greenville - Module 2 (Weeks 5 - 8) aaron.greenville@sydney.edu.au
- Dr. Liana Pozza - Module 3 (Weeks 9 - 12) liana.pozza@sydney.edu.au
- Tutors & Demonstrators: mainly Honours/Masters/PhD students or PhD graduates

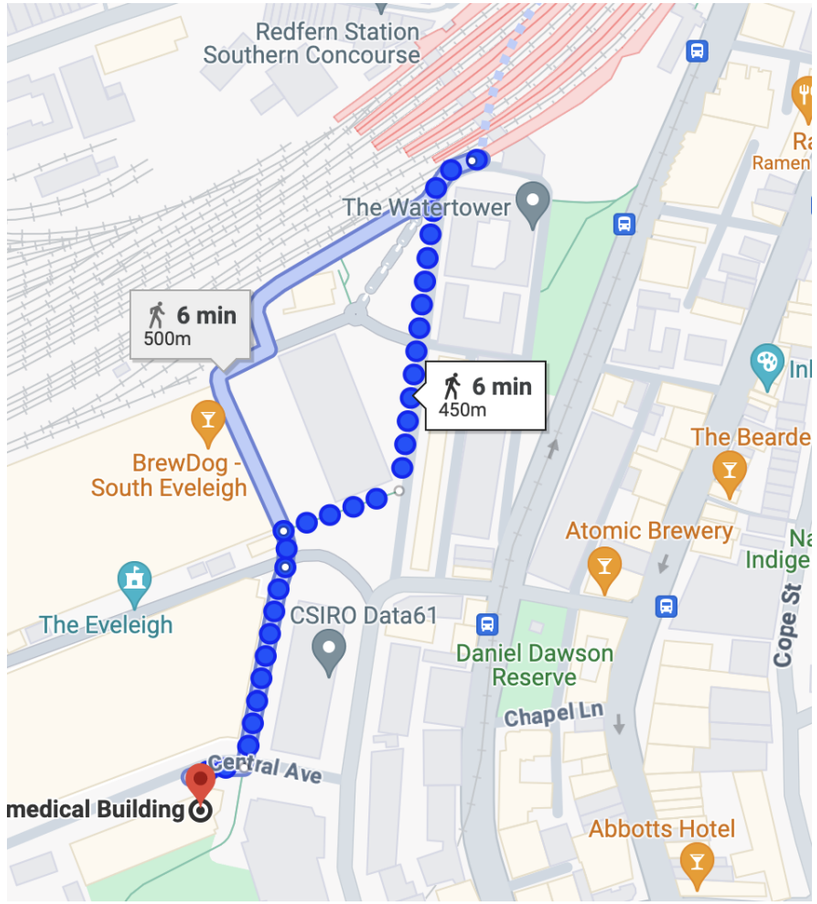
Timetable and locations

1. **Lecture** (recorded) - Monday 12pm ABS Lecture Theatre 1040 and Tuesday 9 am ABS Lecture Theatre 1130
2. **Tutorials** – 1 hour per week self guided study using provided video to complete in allocated time prior to practical
3. **Computer Labs** (Australia Technology Park) – 2 hour computer lab directly following self-study; see personal timetable

Australian Technology Park

- We have programmed in 30 minutes travel time and if you complete the tutorial prior to the scheduled time, you will have an extra hour.
- There is a shuttle service that runs intermittently.





Learning outcomes

- **LO1**. Apply Statistical Tools Using R and Excel: Demonstrate proficiency in utilizing R and Excel to effectively process, describe, and analyse data from simple experiments, showing skill in applying these tools to real-world data sets.
- **LO2**. Evaluate Probability Concepts and Calculations: Evaluate and interpret the concept of probability, proficiently calculate probabilities by correctly applying probability laws and theoretical results and assess their significance in experimental contexts.
- **LO3**. Synthesize Knowledge for Experimental Inference: Integrate understanding of experimental inference to discern and select the most appropriate statistical test (including 1-sample, 2-sample, chi-square, and non-parametric tests) for various types of experimental data, showing the ability to tailor statistical approaches to specific research questions.
- **LO4**. Model Relationships Using Linear and Non-Linear Functions: Construct and apply both linear and non-linear models to describe relationships between variables using R and Excel, demonstrating creativity in developing models that effectively represent complex data patterns.
- **LO5**. Communicate Statistical Findings Effectively: Articulate statistical and modelling results clearly and convincingly in both written scientific reports and oral presentations, working effectively as an individual and collaboratively in a team, showcasing the ability to convey complex information to varied audiences.

UoS outline

- Week 01 - **Data:** Introduction and Scientific Method
- Week 02 - **Data:** Exploratory Data Analysis
- Week 03 - **Data:** Normal and discrete distributions
- Week 04 - **Data:** Sampling distributions
- Week 05 - **Inference:** 1-sample tests
- Week 06 - **Inference:** 2-sample tests
- Week 07 - **Inference:**Non-parametric tests 1
- Week 08 - **Inference:**Non-parametric tests 2
- Week 09 - **Modelling:** Describing relationships
- Week 10 - **Modelling:** Linear functions
- Week 11 - **Modelling:** Linear functions – multiple predictors
- Week 12 - **Modelling:** Non-linear functions
- Week 13 - **Revision** Past exam questions.

Assessment

Projects

- Project 1 (individual 10%): **Exploring data - week 5 - 500 words**
- Project 2 (individual 10%): **Making decisions about data - week 10 - 800 words**
- Project 3 (group 10% + Peer assessment 5%): **Modelling relationships in data - due week 13 - 5 minute group presentation**

In class tests - practicals

- Early Feedback Quiz (individual 5%): **Describing data with R - Week 3 - 15 minutes**
- Coding and data skills evaluation (individual 15%): **Analysing data with R - Week 08 - 50 minutes**

Exam

- Final exam (individual 45%): **MCQ + SAQ Questions - Exam Week - 2 hours**

Attendance

- Lecture attendance is not compulsory but strongly recommended as we do make interactive classes. Lectures will be recorded.
- **A minimum 80% attendance** is required in practicals
- Practical attendance is **very important** for group work as well as learning.
- There is a **“statistically significant”** correlation between class attendance and performance in this unit.

Reference material

- Lecture slides on Canvas
- Statistics texts: see reading list on Canvas
  - Mead R, Curnow RN, Hasted AM (2002) ‘Statistical methods in agriculture and experimental biology.’ [e-book – see Reading List in Canvas]
  - Quinn GP, Keough MJ (2002) ‘Experimental design and data analysis for biologists.’ [e-book – see Reading List in Canvas]
  - Murray, L. (2010) ‘Biostatistical Design and Analysis Using R: a practical guide.’ [e-book – see Reading List in Canvas]
- Unit of Study Notes: See Canvas
- There are heaps of online free texts and online resources that use R - see for example https://leanpub.com/os
- chatGPT, rtutor.ai, co-pilot in RStudio

Need help?

- `Ed discussion` - Q&A discussion: This is the quickest way to get help
- Drop-in sessions: TBA
- Demonstrators: We have a great teaching team of honours and postgraduate students who are all good with R and Stats! Please remember though, they are also learning to teach and so may not have an answer for everything!

Applied statistics units

Core for most of you

- 1st year: Introduction to statistical methods (ENVX1002)
- 2nd year: Applied statistical methods (ENVX2001)

Elective statistics units offered in SOLES

- 2nd or 3rd year: Statistics in the natural sciences (ENVX3002)
- Honours: Experimental Design and Data Analysis (SCIE4002)

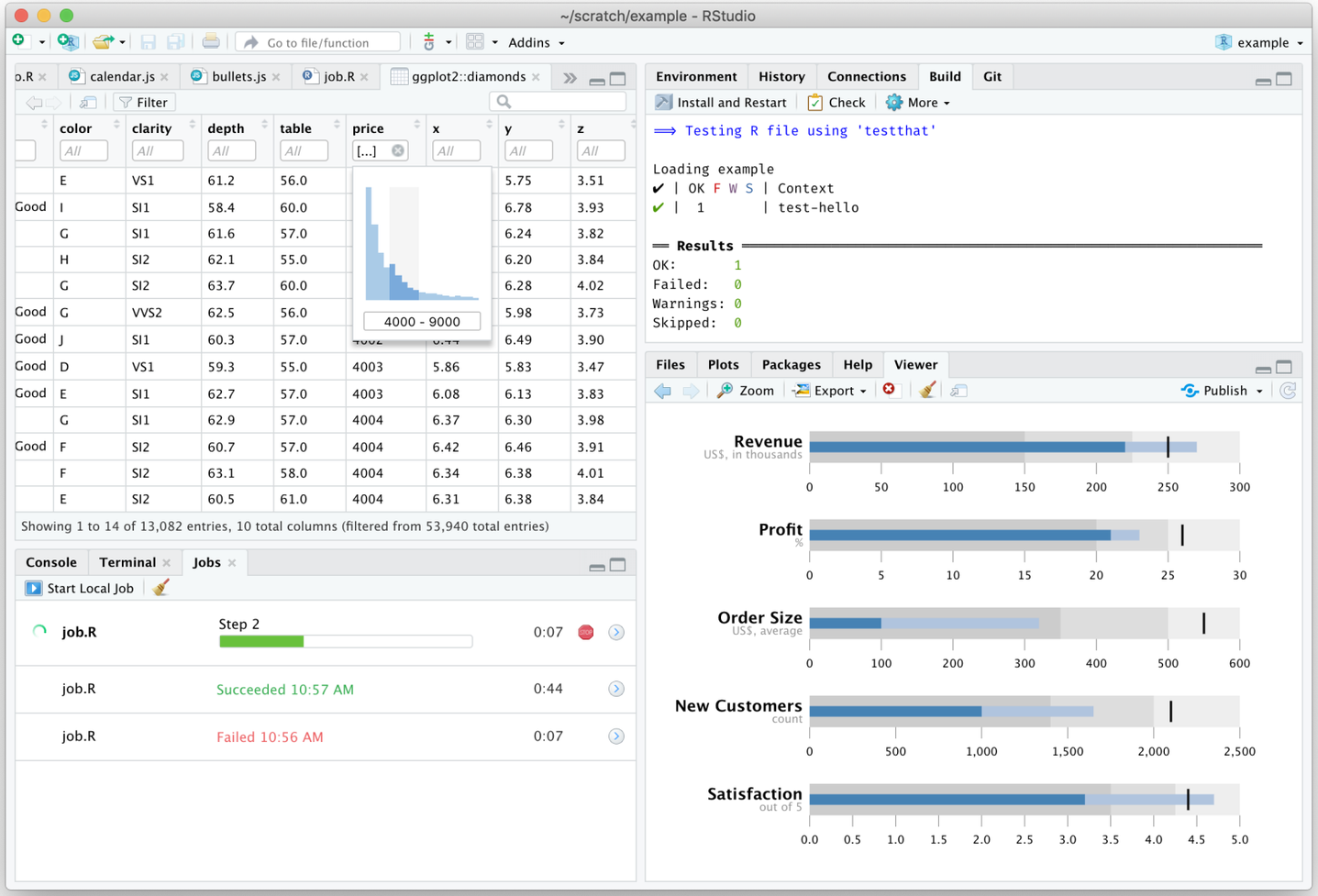
Statistical software packages

- R is Free and Open Source
- R has become a main stream and very powerful statistical tool
- Download from http://cran.r-project.org/
- $>$ 6,000 packages for specialised tasks or modelling

We will use the text editor RStudio:

- Save scripts
- Syntax highlighting of R commands
- Document creation
- Code sharing
- Projects

RStudio

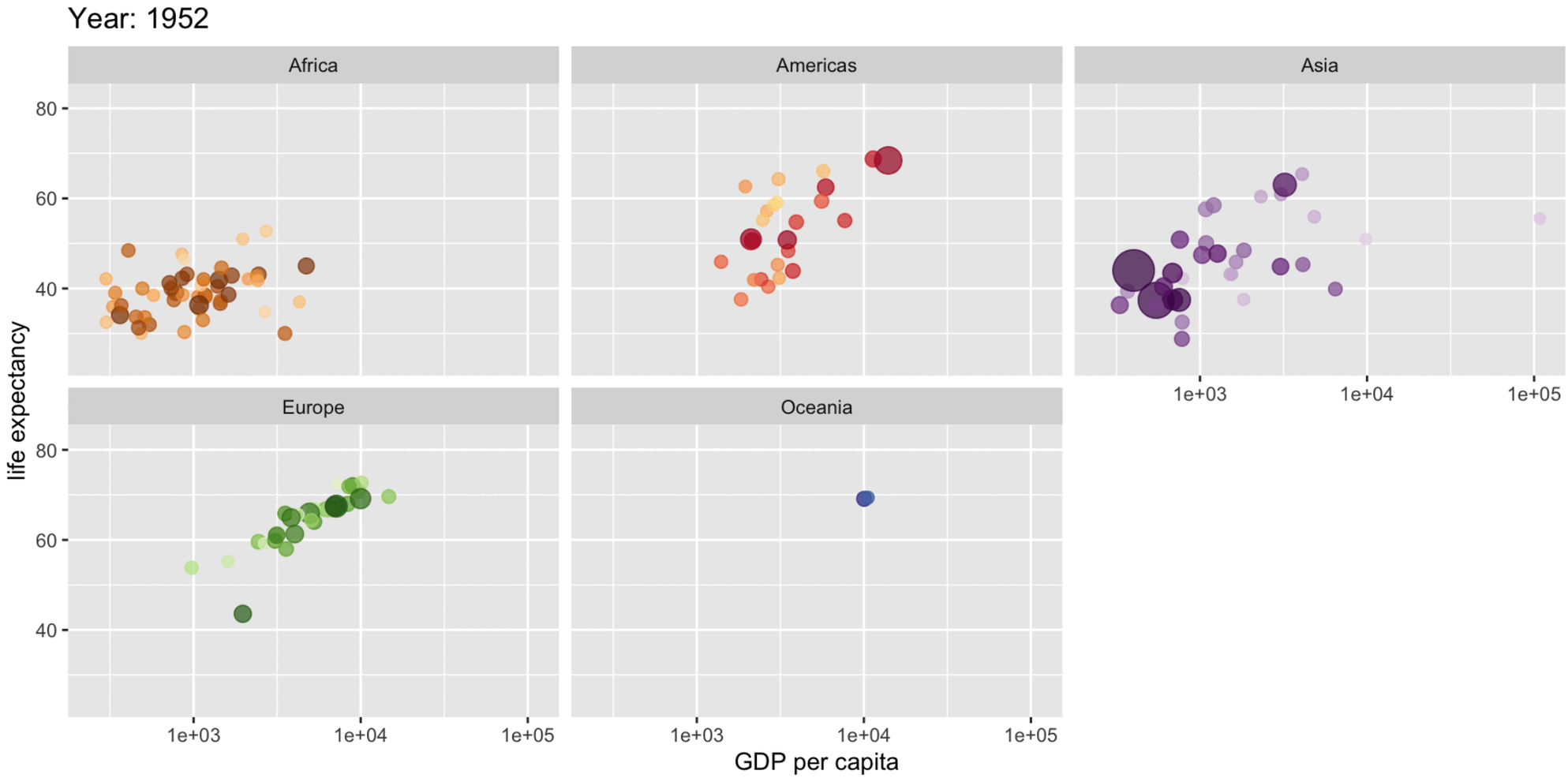

~/scratch/example - RStudio
Go to file/function
Addins
example
o.R
calendar.js
bullets.js
job.R
ggplot2::diamonds
Filter
color
clarity
depth
table
price
x
y
z
4000 - 9000
Showing 1 to 14 of 13,082 entries, 10 total columns (filtered from 53,940 total entries)
Console
Terminal
Jobs
Start Local Job
job.R
Step 2
0:07
job.R
Succeeded 10:57 AM
0:44
job.R
Failed 10:56 AM
0:07
Environment
History
Connections
Build
Git
Install and Restart
Check
More
Testing R file using 'testthat'
Loading example
OK F W S | Context
1 | test-hello
Results
OK: 1
Failed: 0
Warnings: 0
Skipped: 0
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Revenue
US$, in thousands
Profit
%
Order Size
US$, average
New Customers
count
Satisfaction
out of 5

Intro to R

Take a look at this link showing the basics before your first practical this week.

Gapminder demo in R

▶ Code



Topic 1 – The importance of statistics and the scientific method

“Scientists seek to answer questions using rigorous methods and careful observations. These observations collected from the likes of field notes, surveys, and experiments form the backbone of a statistical investigation and are called data. Statistics is the study of how best to collect, analyze, and draw conclusions from data…”

Diez, D.M., Barr, C.D. and Cetinkaya-Rundel, M., 2012. OpenIntro statistics (pp. 174-175). Boston, MA, USA:: OpenIntro. Page 8

Topic 1 – The importance of statistics and the scientific method

- Importance of statistics and more generally data science
  - why?
  - big data, data science
- Scientific method
  - traditional scientific method
  - alternate types of science
    - empirical/mechanistic

Learning outcomes

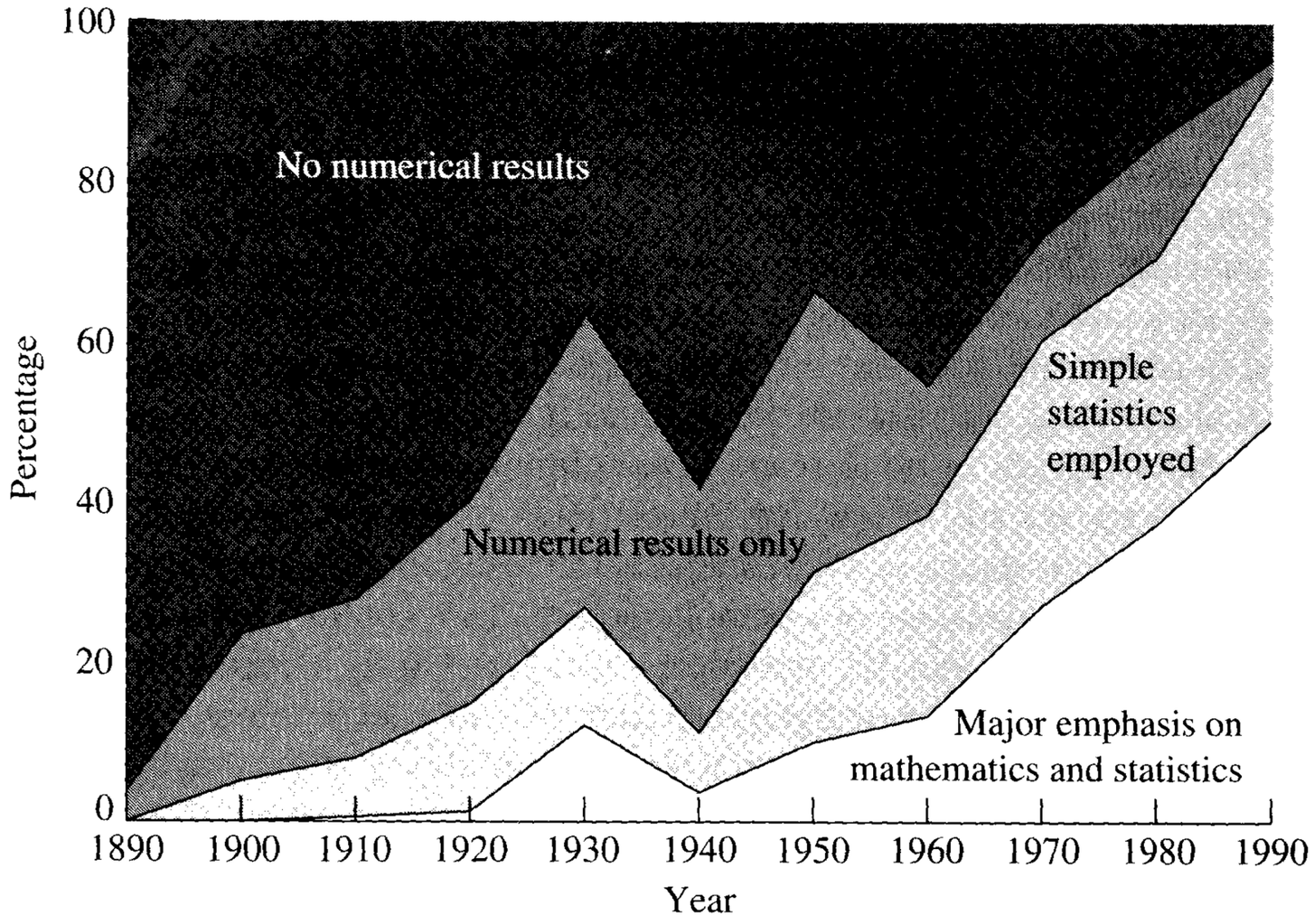
At the end of this topic, you should be able to

- Understand the importance of statistics in society
- Describe the scientific method

Importance of statistics?

z_amir- stock.adobe.com

Importance of statistics?



Sokal & Rohlf (1995) p5

Importance of statistics?

- Big Data started maybe 20 years ago.
- The first large language model ChatGPT (Generative Pre-training Transformer) came out at the end of last year.
- It is changing how we teach you.

Arthead - stock.adobe.com

Importance of statistics?

- How do we interpret statistics
- Language machines

Importance of statistics?

What happens when we get it wrong: Personalised medicine to tailor chemotherapy

Importance of statistics?

Reproducible research

The Annals of Applied Statistics

Info | Current issue | All issues | Search

Ann. Appl. Stat.
Volume 3, Number 4 (2009), 1309-1334.

← Previous article | TOC | Next article →

Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology

Keith A. Baggerly and Kevin R. Coombes

Full-text: Open access

Enhanced PDF (1017 KB)

Abstract | Article info and citation | First page | References | Supplemental materials

Abstract

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

Importance of statistics?

Reproducibility

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

Importance of statistics?

- Gap minder link
- Watch 1st 7 min (or even stay on and watch the whole thing if you like!)

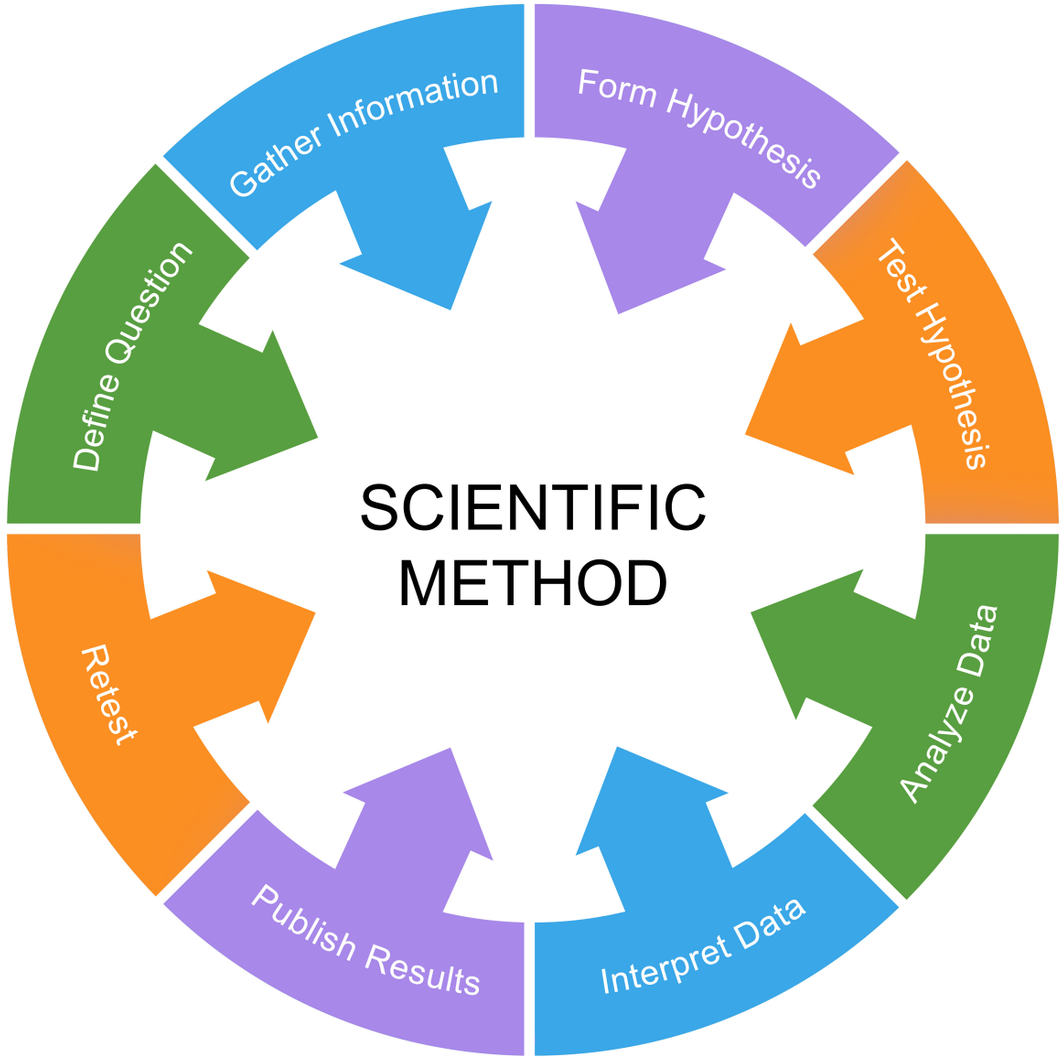
Scientific method

- “Scientific method refers to the body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge”.
  - Testable hypotheses; in the form of a prediction
  - The more you smoke the higher chance of lung cancer
  - Testing these with data
- Scientific method defined in Wikipedia

Scientific method

- Different types of science
  - Controlled experiments – classical scientific method
  - Observational studies – collecting data, abundance wildlife species
  - Modelling (empirical/mechanistic) – empirical: based on data and is a statistical model, mechanistic model is based on underlying physical principles; thermodynamics, physics
  - Model development (mechanistic) – role of stats is in model testing
  - Methodology development – different ways to measure things – role of stats is in testing different methodologies

Scientific method



mybaitshop - stock.adobe.com

Scientific method

- **Observation**: This is the initial stage where a phenomenon or a set of data is observed. The observations must be objective, measurable, and reproducible.
- **Question**: Based on the observations, a specific, clear, and concise question is formulated. This question should be answerable through further investigation.
- **Research**: This involves gathering information and existing knowledge about the topic. This step helps in understanding what is already known and what gaps in knowledge exist.
- **Hypothesis**: A hypothesis is a tentative explanation for the observations and question. It is a statement that can be tested through further experiments and analysis. A good hypothesis is clear, specific, and testable.

Scientific method

- **Experimentation**: Experiments are designed to test the hypothesis. This involves manipulating one or more variables (independent variables) to observe the effect on other variables (dependent variables). Experiments should be controlled, meaning that all other variables are kept constant to ensure that only the variable of interest is affecting the outcome.
- **Data Collection**: Data is collected during the experiments. This data must be objective, measurable, and reproducible. Proper data collection is critical to ensure the validity of the experiment.
- **Analysis**: The data collected from the experiments are analyzed to determine whether they support or refute the hypothesis. Statistical methods are often used in this analysis to determine the significance of the results.
- **Conclusion**: Based on the analysis, a conclusion is drawn. If the data supports the hypothesis, it may become a theory. If the data does not support the hypothesis, the hypothesis may be revised or rejected, and new hypotheses may be formulated.

Scientific method

- **Reporting and Publication**: The findings, along with the methodology, are shared with the scientific community through publications, presentations, or reports. This allows other scientists to review, replicate, and build upon the work.
- **Replication**: Other scientists may attempt to replicate the findings to verify the results. Replication is a key part of the scientific process as it helps to confirm the validity of the findings.
- **Theory Development**: If a hypothesis is repeatedly validated and supported by a significant body of evidence, it may contribute to the development of a scientific theory. A theory is a well-substantiated explanation of some aspect of the natural world that is supported by a large body of evidence.

Ice blocks?

- Why is an ice cube interesting?
- How might we apply the scientific method to ice?
- What might be some criticisms of the scientific method?

Koya979 - stock.adobe.com

Scientific method

- Gap minder link
- From about 50 minutes on wards is about big data & mechanistic models

References

- Quinn & Keough (2002). Sections 1.1-1.2, pages 1-7.

Thanks!