Lecture 01b – Reproducible Science

ENVX1002 Introduction to Statistical Methods

Januar Harianto
The University of Sydney

Feb 2025

THE UNIVERSITY OF SYDNEY

Importance of statistics

> Never leave a number all by itself. Never believe that one number on its own can be meaningful. If you are offered one number, always ask for at least one more. Something to compare it with.

– Hans Rosling (1948-2017)

Why learn statistics?

All of science (and industry) are increasingly data-driven and computational:

- **Research** papers are filled with statistical analyses
- **Business** and **policy** decisions are based on data analytics
- Environmental **policies** are guided by statistical models
- **Medical** treatments are evaluated using statistical methods

Most of you are majoring in a field that will require you to understand and use statistics in some form.

Benefits

Even if you don’t become a data scientist, statistics will help you to:

1. **Evaluate claims critically**
    - Understand and analyse data in your field
    - Make informed decisions based on evidence
2. **Communicate effectively**
    - Create compelling data visualisations and reports
    - Present findings clearly to different audiences
3. **Solve real-world problems**
    - Design and analyse experiments properly
    - Make evidence-based predictions and identify trends

The joy of stats

200 countries, 200 years, 4 minutes

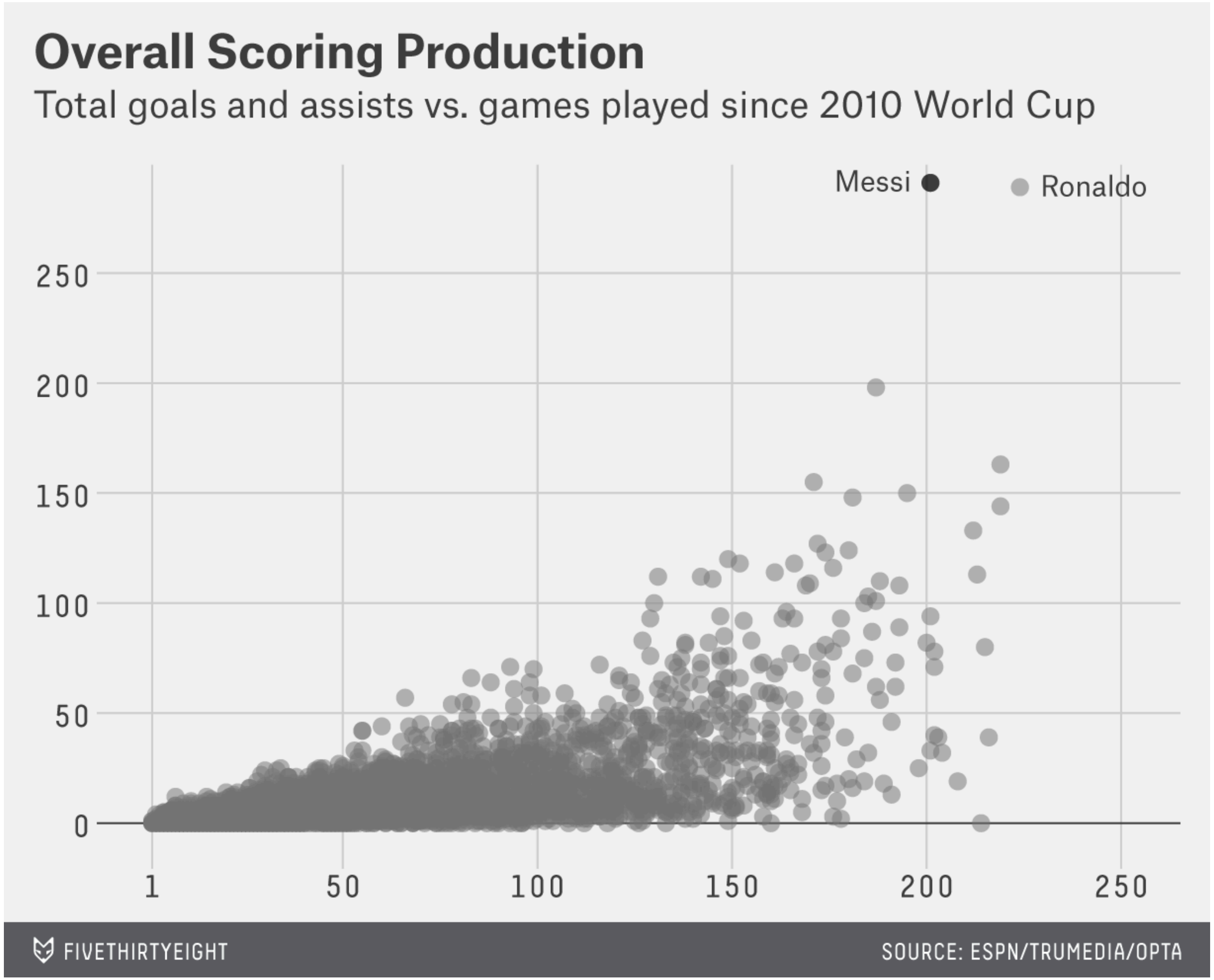
In your own time: The best stats you’ve ever seen

Lionel Messi is impossible

> It’s not possible to shoot more efficiently from outside the penalty area than many players shoot inside it. It’s not possible to lead the world in weak-kick goals and long-range goals. It’s not possible to score on unassisted plays as well as the best players in the world score on assisted ones. It’s not possible to lead the world’s forwards both in taking on defenders and in dishing the ball to others. And it’s certainly not possible to do most of these things by insanely wide margins.
>
> **But Messi does all of this and more.**

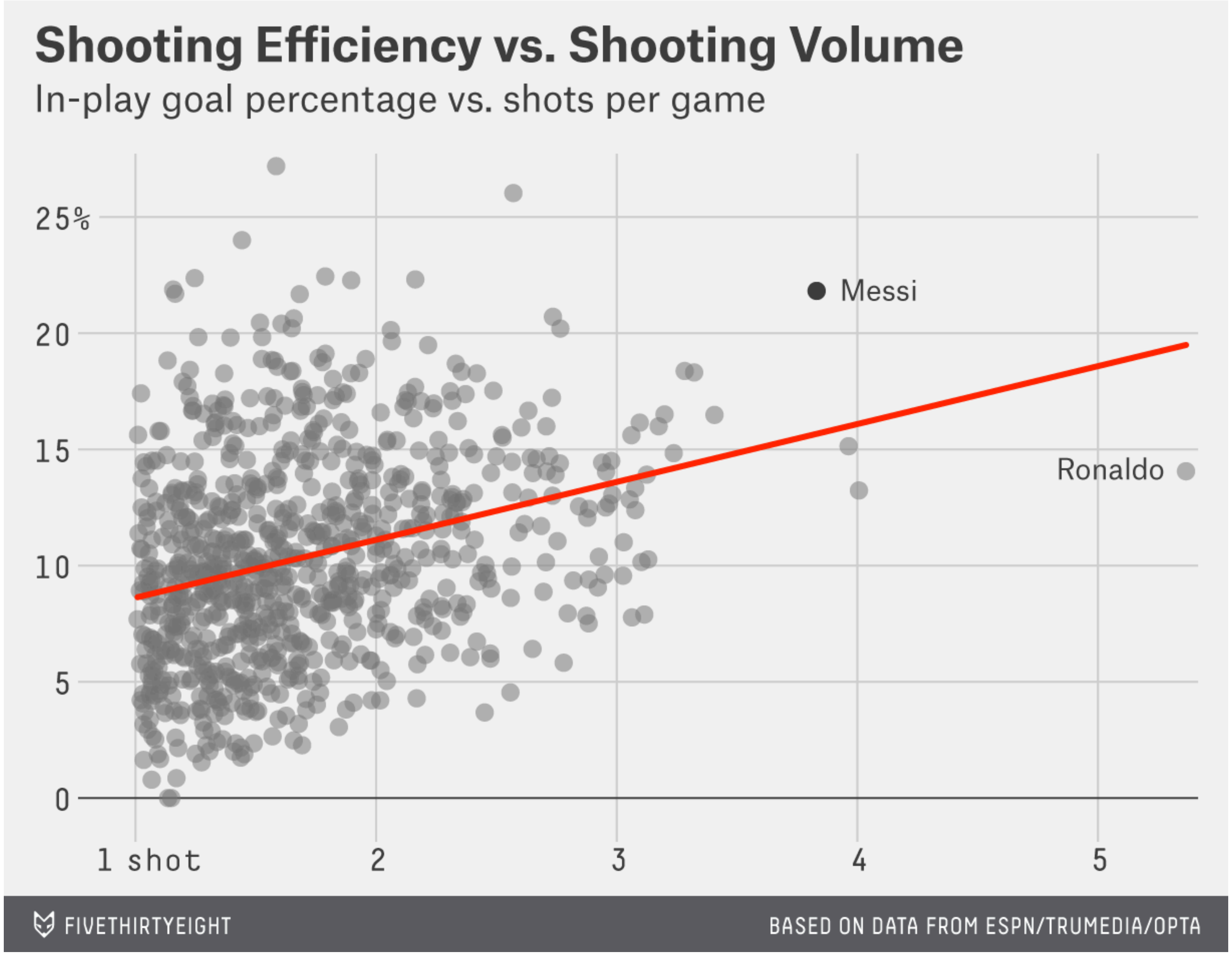
Messi playing for Argentina

Lionel Messi is impossible



Source: fivethirtyeight

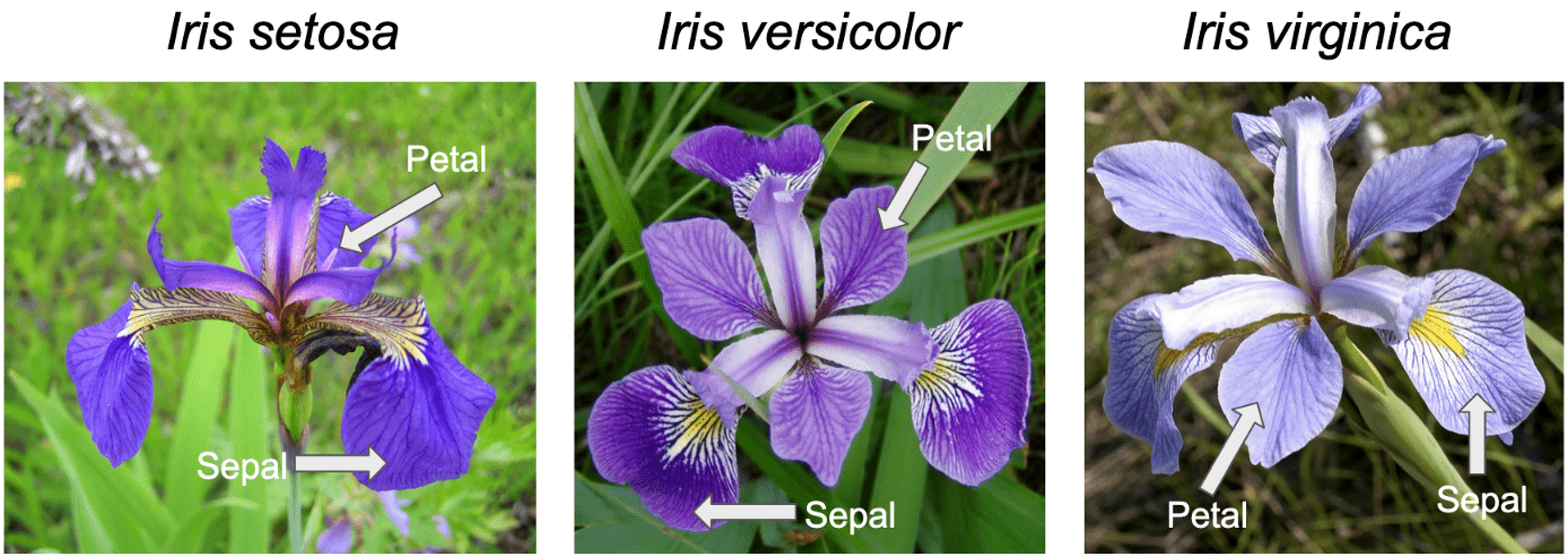
Lionel Messi is impossible



Source: fivethirtyeight

Serious stats

Which sepal length is longer?



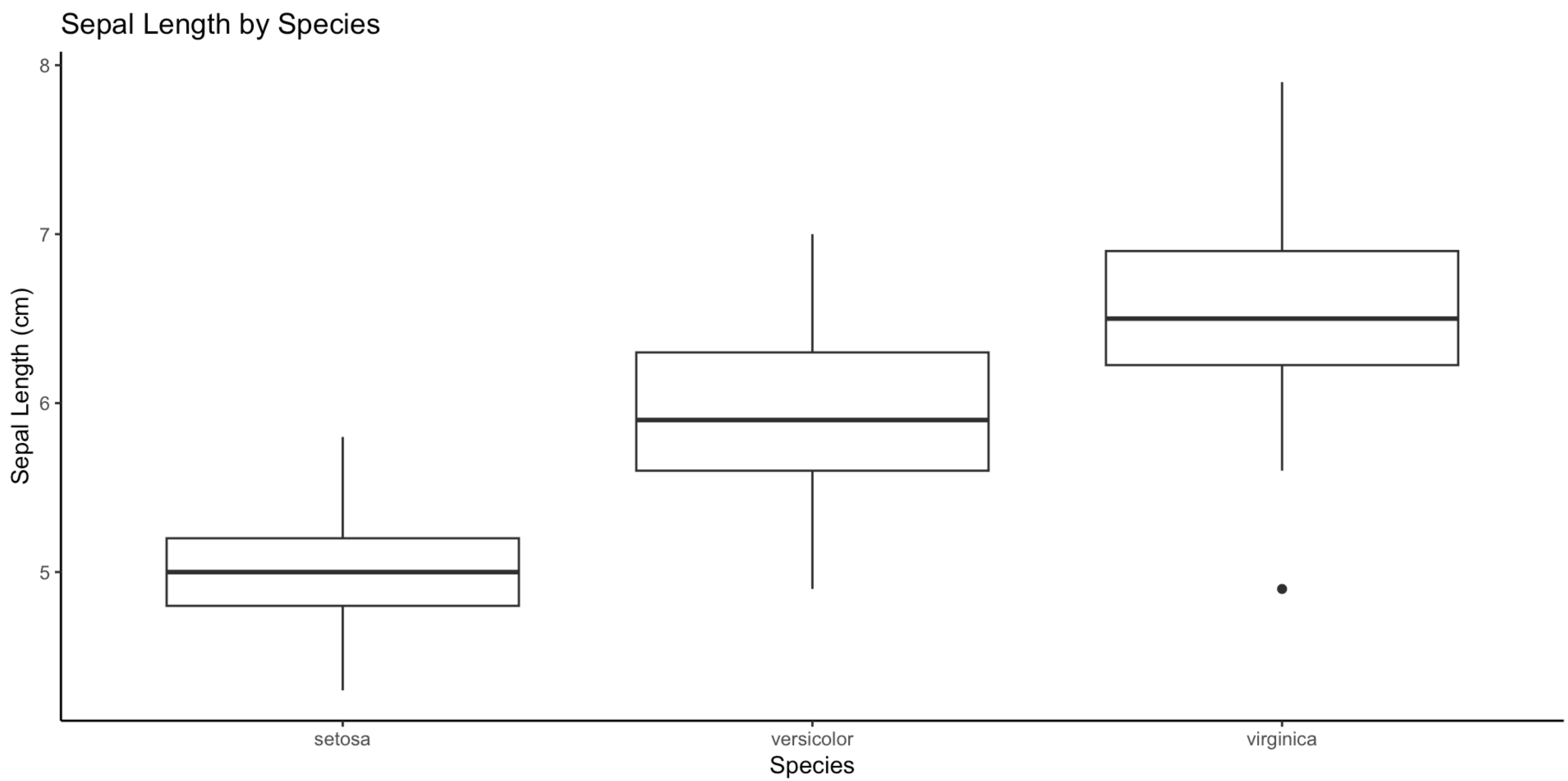
Source: Embedded Robotics

Note: common dataset used in statistics and machine learning.

Serious stats

Visualise

▶ Code



Serious stats

Infer

We use formal statistical tests to determine if differences are **statistically significant** so that we can make **inferences** about the population based on the sample data – part of the **scientific method**.

▶ Code

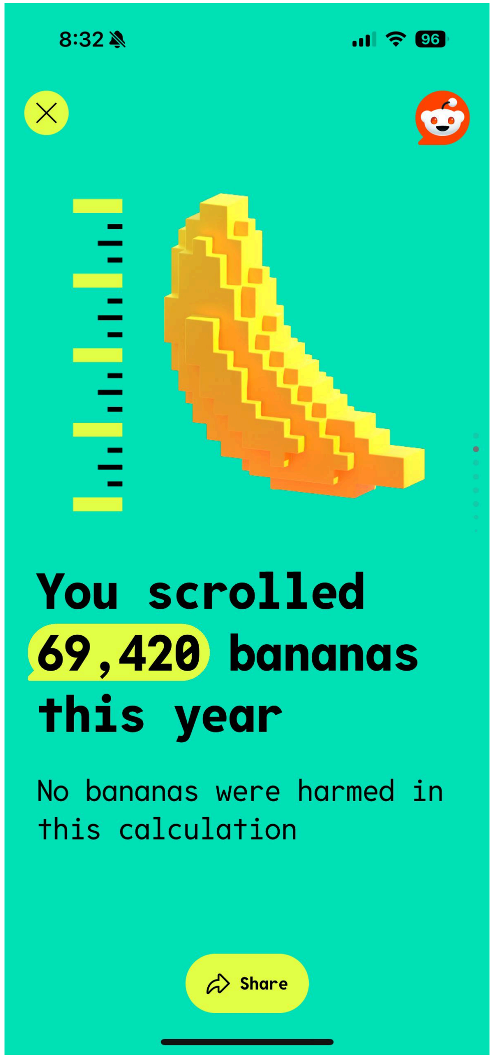
Table 1: One-way ANOVA results comparing sepal length between iris species

Effect	DFn	DFd	F	p	p<.05	ges
Species	2	147	119.265	1.67e-31	*	0.619

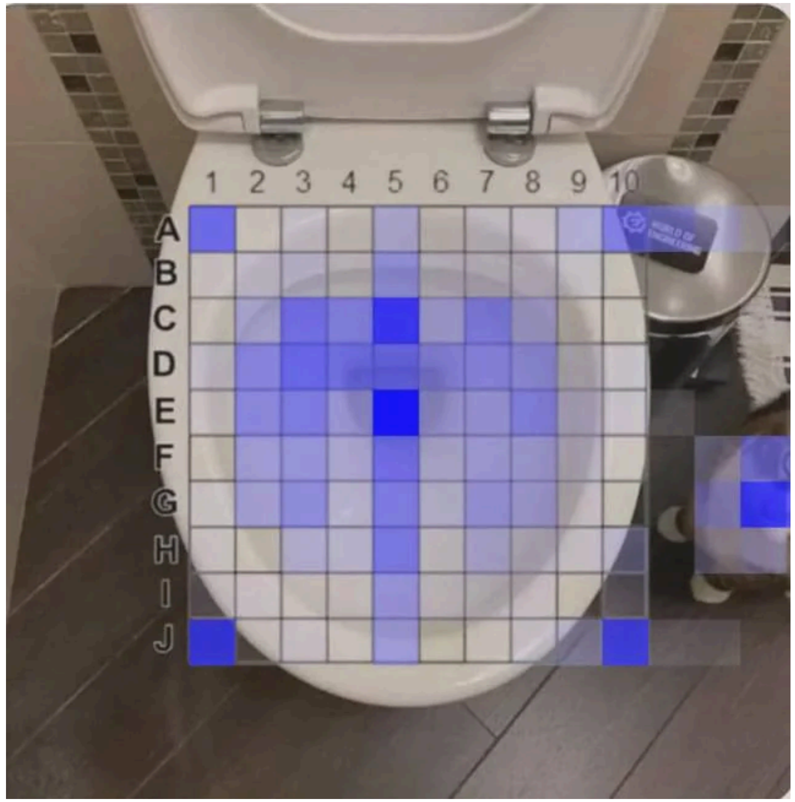
Scientific reporting

> A one-way ANOVA revealed significant differences in sepal length between species (ANOVA, $F(2, 147) = 119.26$, $p < .001$).

Not *always* formal



Source: You scrolled 69,420 bananas this year



Source: "Guys where do you pee?"

The *beauty* of statistics – formal hypothesis testing is not always required to make a point!

Understanding data types

Types of data

We encounter different types of data in our everyday lives:

1. **Numeric (Quantitative)**
    - **Continuous**: Can take any value (e.g., height: 175.5 cm, temperature: 23.4°C)
    - **Discrete**: Whole numbers only (e.g., number of students: 25, floors in a building: 10)
2. **Categorical (Qualitative)**
    - **Nominal**: Categories without order (e.g., eye color, country of birth)
    - **Ordinal**: Categories with order (e.g., satisfaction: poor → fair → good → excellent)

When data types blur

The same data can be treated in different ways:

- Age can be numeric (25.5 years) or categorical (young/middle/old)
- Temperature can be numeric (23.4°C) or categorical (cold/warm/hot)
- Ratings can be numeric (1-5) or ordinal (poor to excellent)

How we treat the data depends on our research question:

- What comparison makes sense for our goals?
- What level of detail do we need?
- What statistical tests are appropriate?

Why data types matter

Understanding data types helps us:

1. **Choose appropriate analyses**
    - Different statistical tests for different data types
    - Avoid incorrect conclusions
2. **Create effective visualisations**
    - Bar plots for categorical data
    - Scatter plots for continuous data
3. **Communicate results clearly**
    - Use appropriate summary statistics
    - Present data in meaningful ways

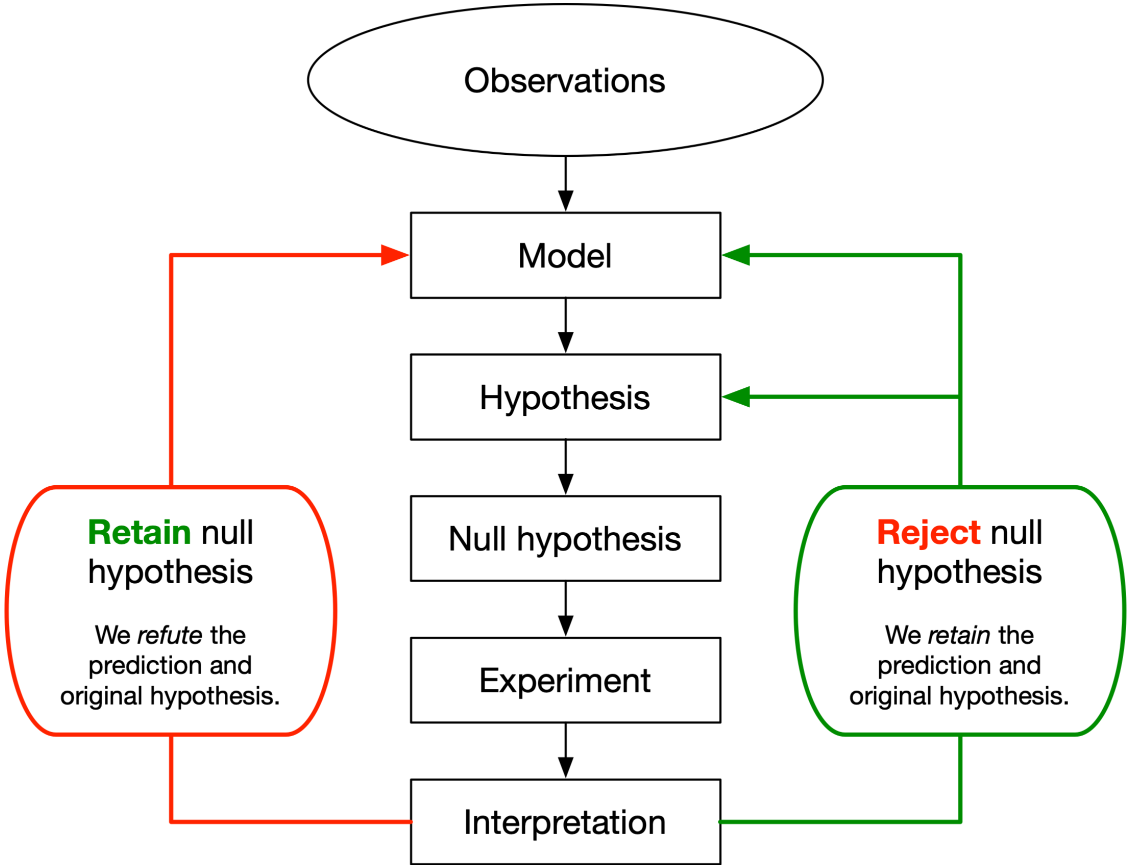
The scientific method

> *The man of science has learned to believe in justification, not by faith, but by verification.*

– Thomas Huxley (1825-1895)

Science as an enterprise

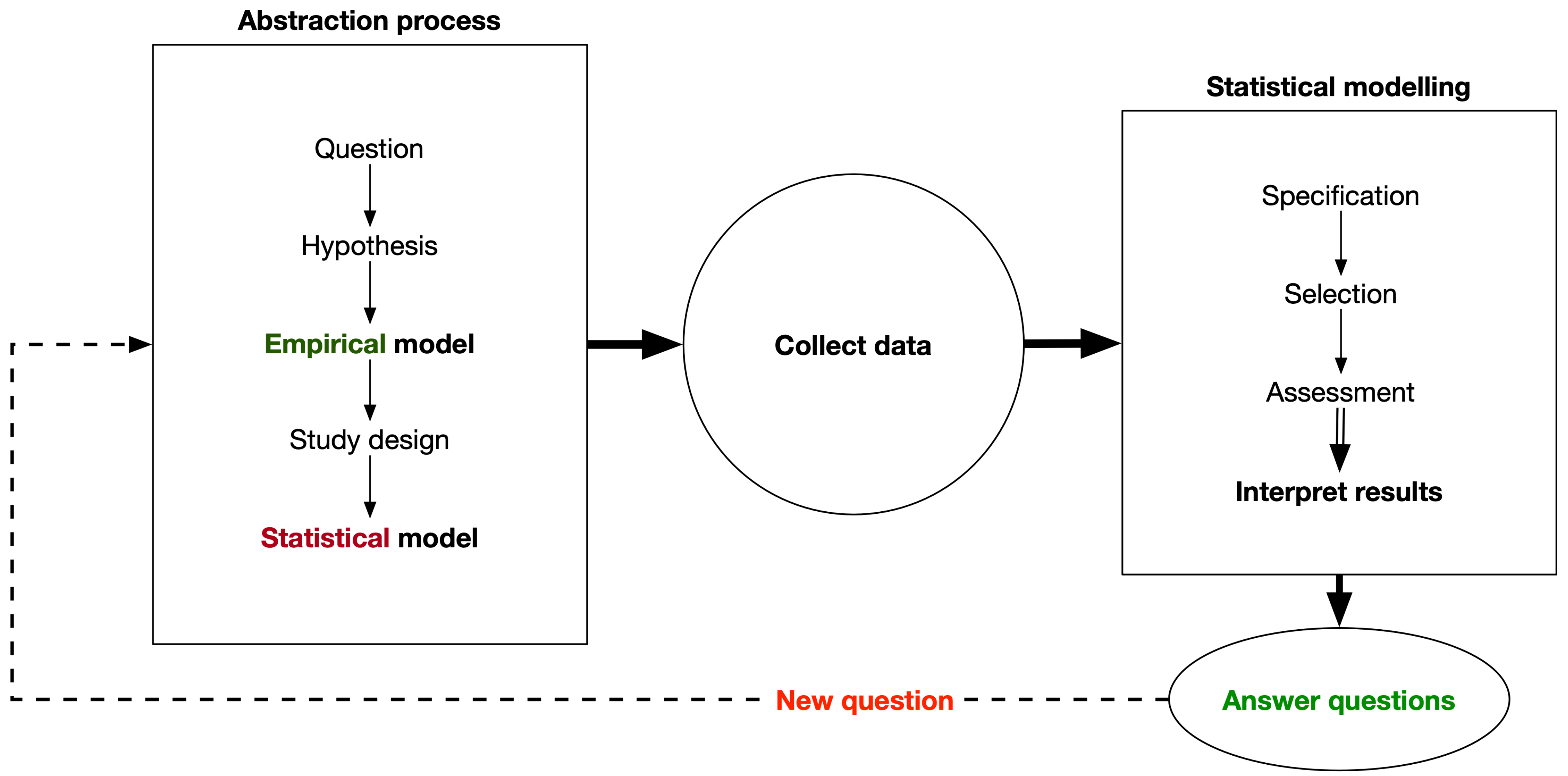
- The scientific method – fundamental to centuries of scientific progress
- If you discover something (**or not**), it should be possible for others to verify your findings independently
- Your findings should be **reproducible** and **replicable**



The logical framework by Underwood (1997)

No single method

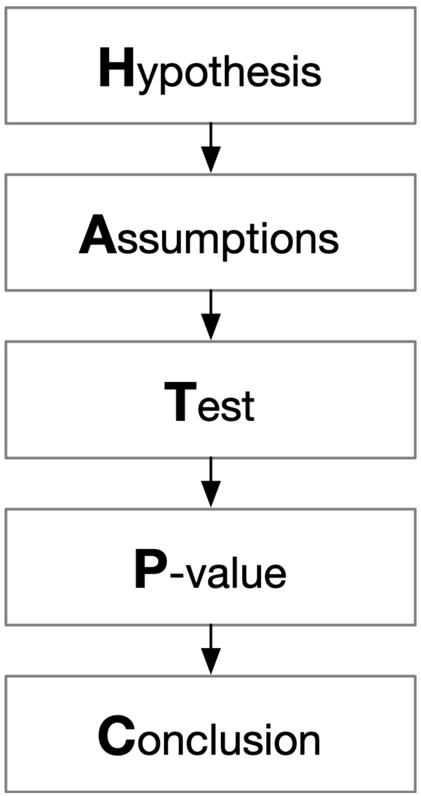
Variation of the scientific method exist– it is a **framework** that guides the process of scientific inquiry



No single method

HATPC

Hypothesis – **A**ssumptions – **T**est statistic – **P**-value – **C**onclusion



You will see some variation of the HATPC in your first-year units – a common framework for **report writing**

We will follow this framework in the future (easiest to apply)

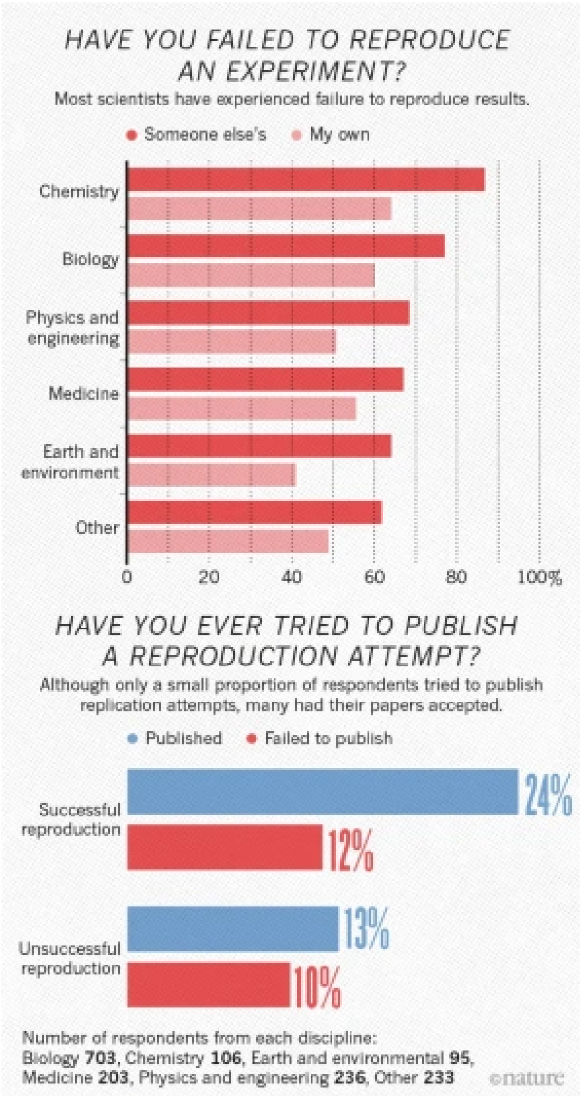
Key principles

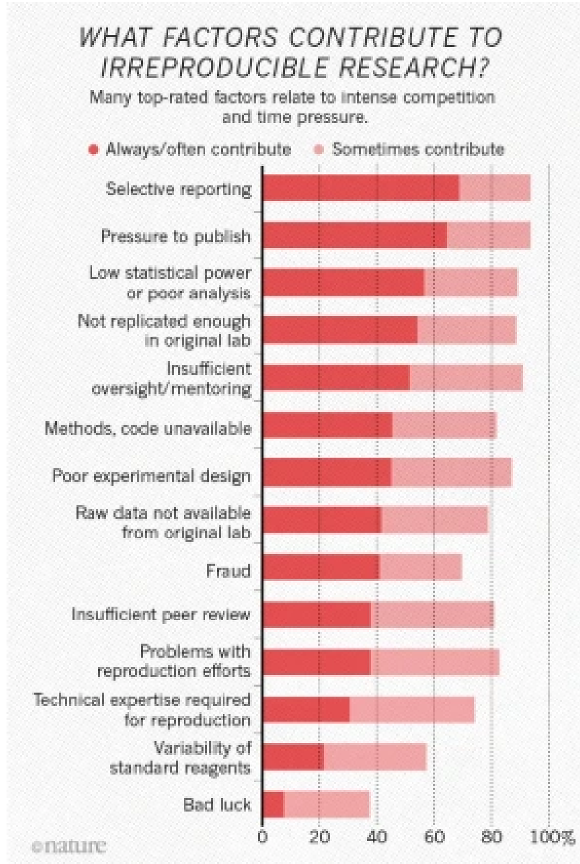
1. **Observation**: Identify a phenomenon of interest that can be measured
2. **Question**: Formulate a question that can be answered by collecting data
3. **Research**: Review the literature to understand what is already known – your question may already have been asked by someone else. This step helps in understanding what is already known and what gaps in knowledge may exist.
4. **Hypothesis**: Formulate a **testable** hypothesis – something that can be assessed using data collection and analysis
5. **Experiment**: Design an experiment to test the hypothesis
6. **Data collection**: Collect data
7. **Analysis**: use statistical methods to analyse the data and determine if results are statistically significant or demonstrate a pattern
8. **Conclusion**: Interpret the results and draw conclusions. If the results are not significant, this is still a valid conclusion!

Reproducibility crisis

The Scientific Method does not guarantee that all research is reproducible or replicable – sometimes the tools we use hinder, rather than help, reproducibility.

From **Nature** (including image sources):





> More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.

Statistical analysis, experimental design and data issues are the main factors affecting research reproducibility.

Reproducibility and replicability

Key definitions

- **Reproducibility**: the ability to re-run an analysis and obtain the same results
- **Replicability**: the ability to obtain the same conclusions using a *different* dataset or study population

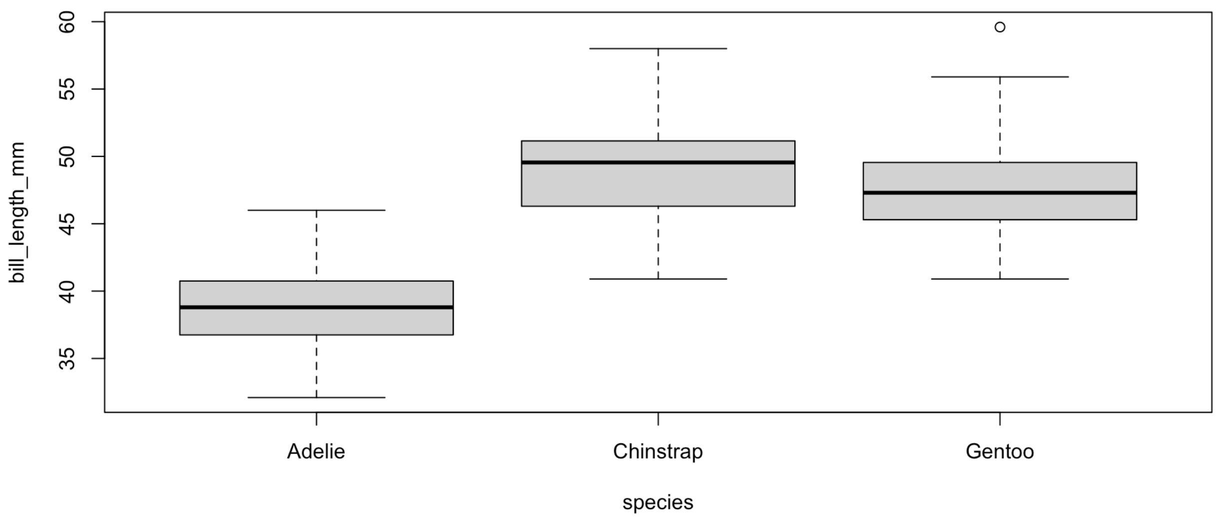
Scientific findings should be **both** reproducible and replicable – the tools that we use *should* facilitate this in the most efficient way possible.

Reproducibility

How would you explain to someone how to reproduce this plot…

- in Excel? Check the **guide**
- In SPSS? Check the **guide**
- In R?

▶ Code



An over-simplification

Those without programming knowledge will *still* struggle to understand and use the two lines of R code shown above.

Pre-requisites

- Understanding of the R programming language
- Knowing how to debug (i.e. read error messages or "play" with code)
- It takes time, but the payoff is worth it – all programming languages follow similar principles and you will find others easier to learn, even if not for statistics…

References

- Quinn & Keough (2002). Sections 1.1-1.2, pages 1-7.
- Underwood AJ (1997) Experiments in Ecology: Their Logical Design and Interpretation using Analysis of Variance. Cambridge University Press, Cambridge.

Thanks!