Lancaster University

Quantitative Research Methods

January 23, 2023

Professor Patrick Rebuschat

p.rebuschat@lancaster.ac.uk

Our plan for today

Lancaster University

Homework assignment

Quantitative research (contd)

- The PPDAC cycle
- The data science workflow

Second steps in R

- Solution to homework task
- Practical handout

Our schedule

Lancaster University

Date	Topics
Jan 16	Introduction to quantitative research methods using R
Jan 23	Data management and data wrangling
Jan 30	Exploratory data analysis
Feb 6	Data visualization
Feb 13	No class, please complete the mid-term assignment (everybody)
Feb 20	Significance testing. Hypothesis tests for continuous variables: two groups.
Feb 27	Tests for discrete variables: Analysing contingency tables
Mar 6	Correlation and linear regression
Mar 13	Analysis of Variance (ANOVA) and tests for *N* groups
Mar 20	Multiple regression

Lancaster University

Reminder from last session

Two basic strategies in quantitative research

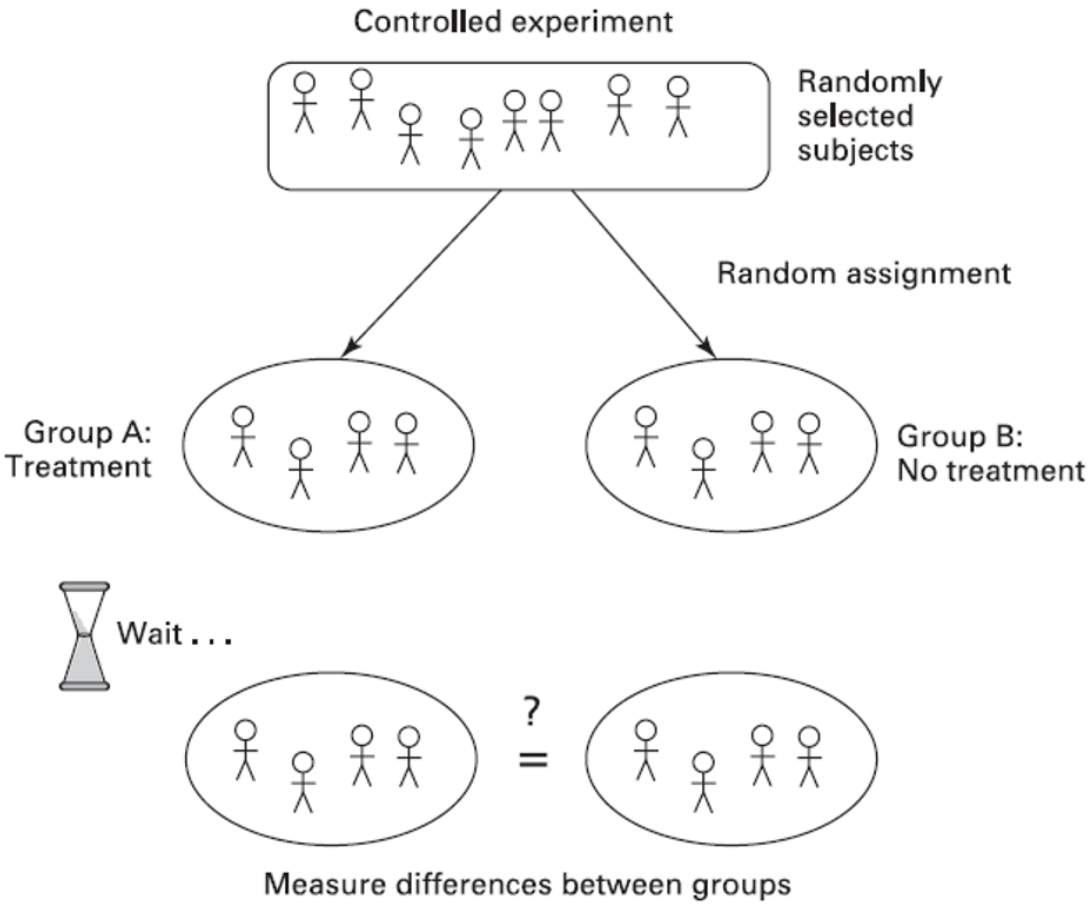
Lancaster University

1. **Observational research**
   - Simply observe and describe our phenomena of interest
   - Example: Systematic reviews, corpus analyses, correlational research, etc.
2. **Experimental research**
   - Systematically manipulate variables of interest and see what effect our manipulation has in the world
   - Example: Experimental studies in the lab or in the wild

Example: Experimental research

Lancaster University

Reproduced from L



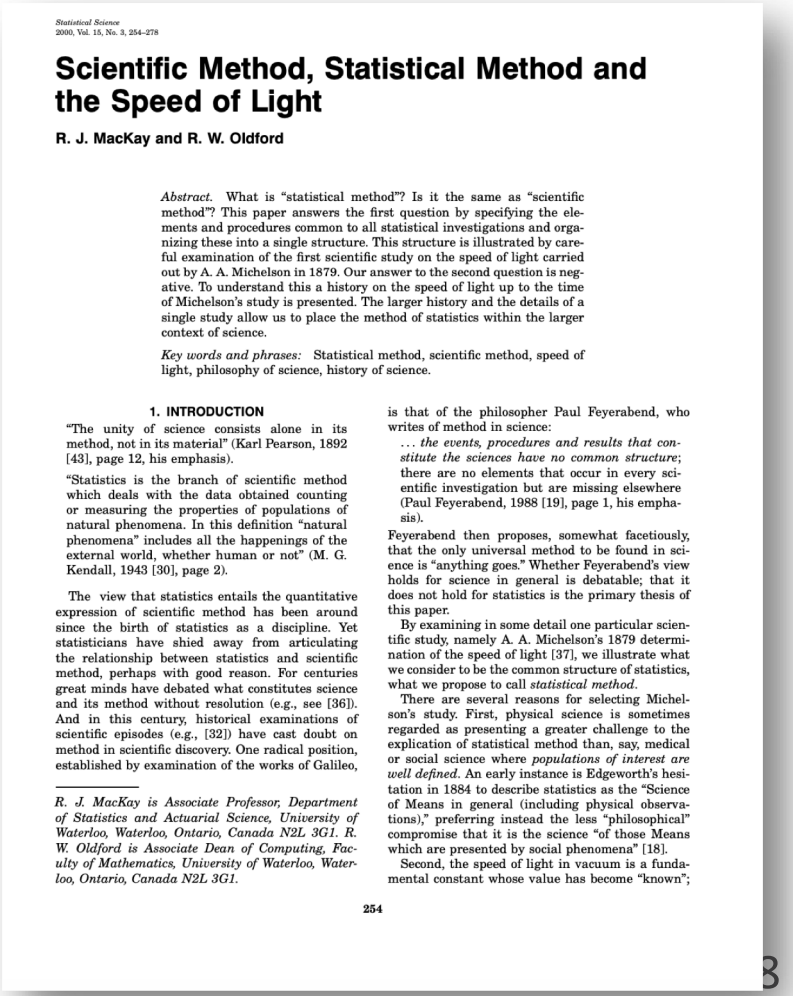
The research cycle

The statistical method: PPDAC (MacKay & Oldford, 2000)

Lancaster University

Statistical method

- "Elements and procedures common to all statistical investigations"
- Can be represented as a series of five interconnected stages: PPDAC.



Statistical Science
2000, Vol. 15, No. 3, 254–278

Scientific Method, Statistical Method and the Speed of Light

R. J. MacKay and R. W. Oldford

Abstract. What is "statistical method"? Is it the same as "scientific method"? This paper answers the first question by specifying the elements and procedures common to all statistical investigations and organizing these into a single structure. This structure is illustrated by careful examination of the first scientific study on the speed of light carried out by A. A. Michelson in 1879. Our answer to the second question is negative. To understand this a history on the speed of light up to the time of Michelson's study is presented. The larger history and the details of a single study allow us to place the method of statistics within the larger context of science.

Key words and phrases: Statistical method, scientific method, speed of light, philosophy of science, history of science.

1. INTRODUCTION

"The unity of science consists alone in its method, not in its material" (Karl Pearson, 1892 [43], page 12, his emphasis).

"Statistics is the branch of scientific method which deals with the data obtained counting or measuring the properties of populations of natural phenomena. In this definition "natural phenomena" includes all the happenings of the external world, whether human or not" (M. G. Kendall, 1943 [30], page 2).

The view that statistics entails the quantitative expression of scientific method has been around since the birth of statistics as a discipline. Yet statisticians have shied away from articulating the relationship between statistics and scientific method, perhaps with good reason. For centuries great minds have debated what constitutes science and its method without resolution (e.g., see [36]). And in this century, historical examinations of scientific episodes (e.g., [32]) have cast doubt on method in scientific discovery. One radical position, established by examination of the works of Galileo, is that of the philosopher Paul Feyerabend, who writes of method in science:

> *. . . the events, procedures and results that constitute the sciences have no common structure*; there are no elements that occur in every scientific investigation but are missing elsewhere (Paul Feyerabend, 1988 [19], page 1, his emphasis).

Feyerabend then proposes, somewhat facetiously, that the only universal method to be found in science is "anything goes." Whether Feyerabend's view holds for science in general is debatable; that it does not hold for statistics is the primary thesis of this paper.

By examining in some detail one particular scientific study, namely A. A. Michelson's 1879 determination of the speed of light [37], we illustrate what we consider to be the common structure of statistics, what we propose to call *statistical method*.

There are several reasons for selecting Michelson's study. First, physical science is sometimes regarded as presenting a greater challenge to the explication of statistical method than, say, medical or social science where *populations of interest are well defined*. An early instance is Edgeworth's hesitation in 1884 to describe statistics as the "Science of Means in general (including physical observations)," preferring instead the less "philosophical" compromise that it is the science "of those Means which are presented by social phenomena" [18].

Second, the speed of light in vacuum is a fundamental constant whose value has become "known";

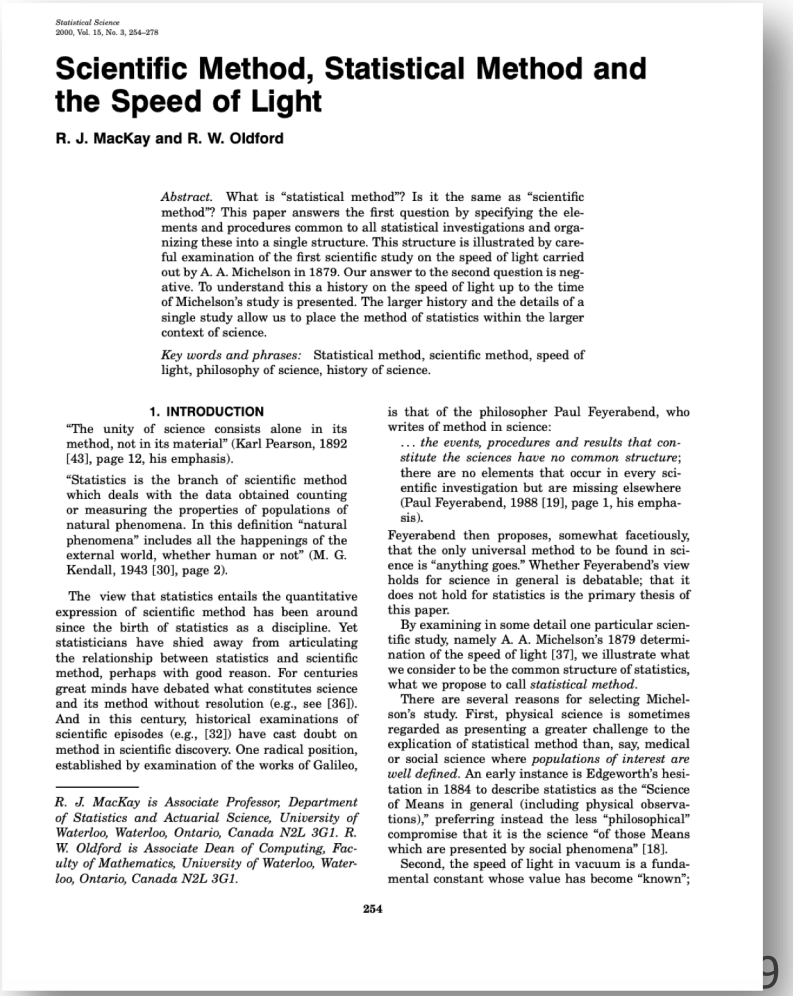
R. J. MacKay is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. R. W. Oldford is Associate Dean of Computing, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

The statistical method: PPDAC (MacKay & Oldford, 2000)

Lancaster University

Five interdependent stages:

1. Problem
2. Plan
3. Data
4. Data analysis
5. Conclusion



Statistical Science
2000, Vol. 15, No. 3, 254–278

Scientific Method, Statistical Method and the Speed of Light

R. J. MacKay and R. W. Oldford

Abstract. What is "statistical method"? Is it the same as "scientific method"? This paper answers the first question by specifying the elements and procedures common to all statistical investigations and organizing these into a single structure. This structure is illustrated by careful examination of the first scientific study on the speed of light carried out by A. A. Michelson in 1879. Our answer to the second question is negative. To understand this a history on the speed of light up to the time of Michelson's study is presented. The larger history and the details of a single study allow us to place the method of statistics within the larger context of science.

Key words and phrases: Statistical method, scientific method, speed of light, philosophy of science, history of science.

1. INTRODUCTION

"The unity of science consists alone in its method, not in its material" (Karl Pearson, 1892 [43], page 12, his emphasis).

"Statistics is the branch of scientific method which deals with the data obtained counting or measuring the properties of populations of natural phenomena. In this definition "natural phenomena" includes all the happenings of the external world, whether human or not" (M. G. Kendall, 1943 [30], page 2).

The view that statistics entails the quantitative expression of scientific method has been around since the birth of statistics as a discipline. Yet statisticians have shied away from articulating the relationship between statistics and scientific method, perhaps with good reason. For centuries great minds have debated what constitutes science and its method without resolution (e.g., see [36]). And in this century, historical examinations of scientific episodes (e.g., [32]) have cast doubt on method in scientific discovery. One radical position, established by examination of the works of Galileo, is that of the philosopher Paul Feyerabend, who writes of method in science:

> . . . *the events, procedures and results that constitute the sciences have no common structure*; there are no elements that occur in every scientific investigation but are missing elsewhere (Paul Feyerabend, 1988 [19], page 1, his emphasis).

Feyerabend then proposes, somewhat facetiously, that the only universal method to be found in science is "anything goes." Whether Feyerabend's view holds for science in general is debatable; that it does not hold for statistics is the primary thesis of this paper.

By examining in some detail one particular scientific study, namely A. A. Michelson's 1879 determination of the speed of light [37], we illustrate what we consider to be the common structure of statistics, what we propose to call *statistical method*.

There are several reasons for selecting Michelson's study. First, physical science is sometimes regarded as presenting a greater challenge to the explication of statistical method than, say, medical or social science where *populations of interest are well defined*. An early instance is Edgeworth's hesitation in 1884 to describe statistics as the "Science of Means in general (including physical observations)," preferring instead the less "philosophical" compromise that it is the science "of those Means which are presented by social phenomena" [18].

Second, the speed of light in vacuum is a fundamental constant whose value has become "known";

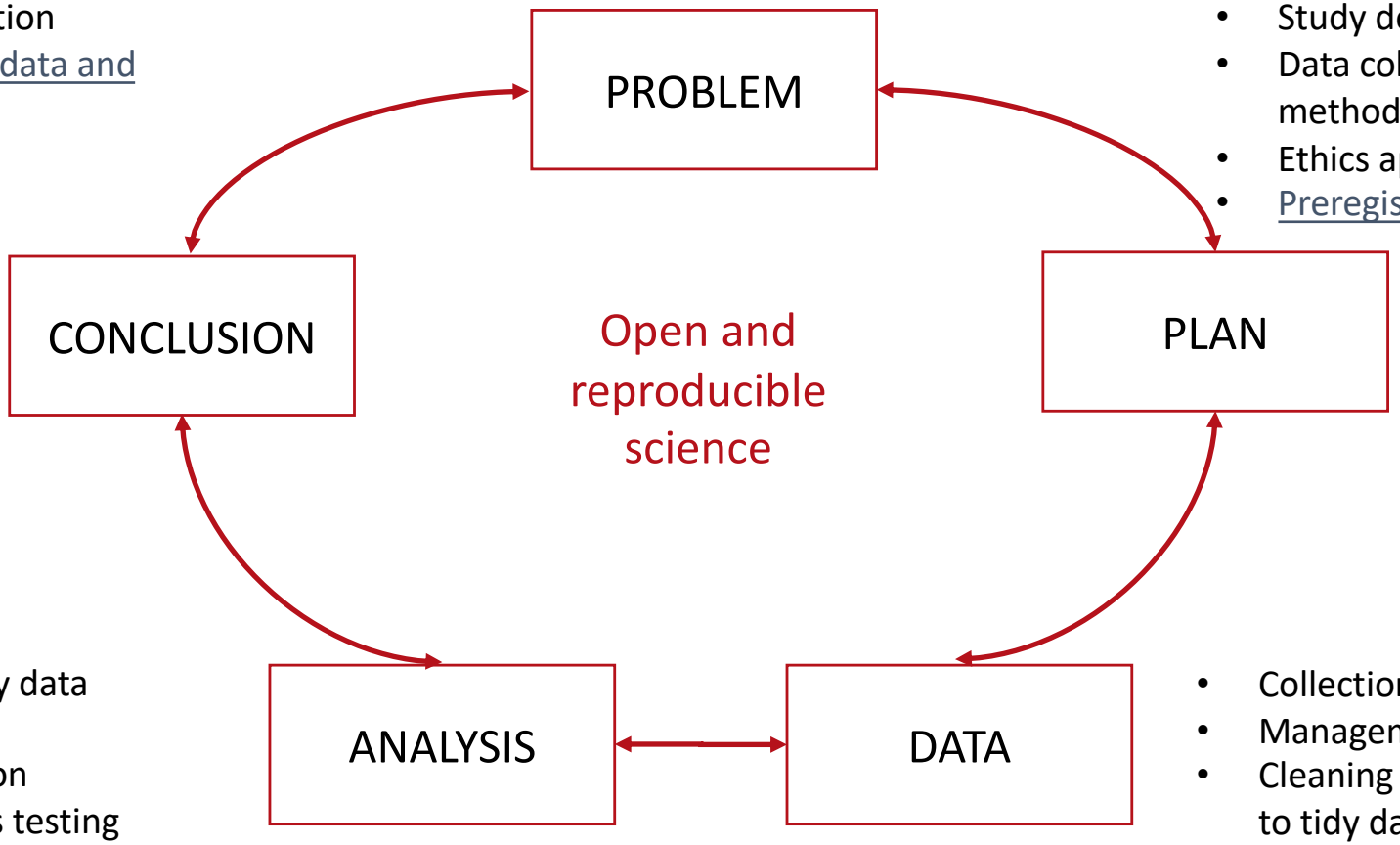
R. J. MacKay is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. R. W. Oldford is Associate Dean of Computing, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

The PPDAC Cycle

Adapted from Spiegehalter (2019)

- Understanding and clearly defining the problem
- Deciding on strategy to answer the question(s)?

- What to measure and how?
- Study design?
- Data collection methods?
- Ethics approval
- Preregistration?

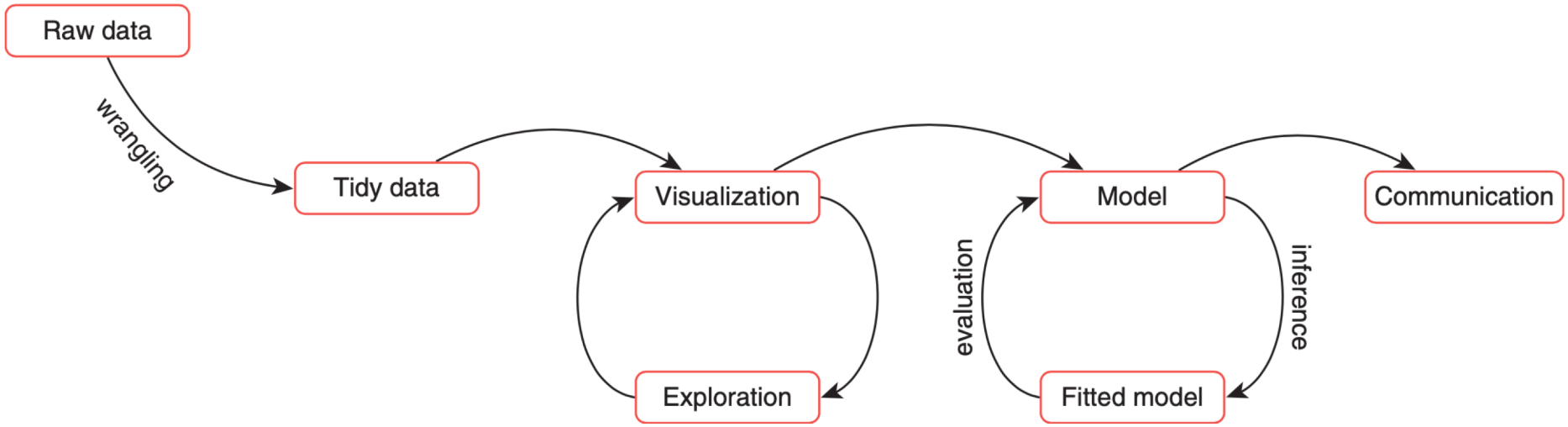


- Collection
- Management
- Cleaning (from raw to tidy data)

- Exploratory data analysis
- Visualization
- Hypothesis testing

- Interpretation
- Conclusion
- Dissemination
- Sharing of data and scripts
- New ideas

The Data Science Workflow



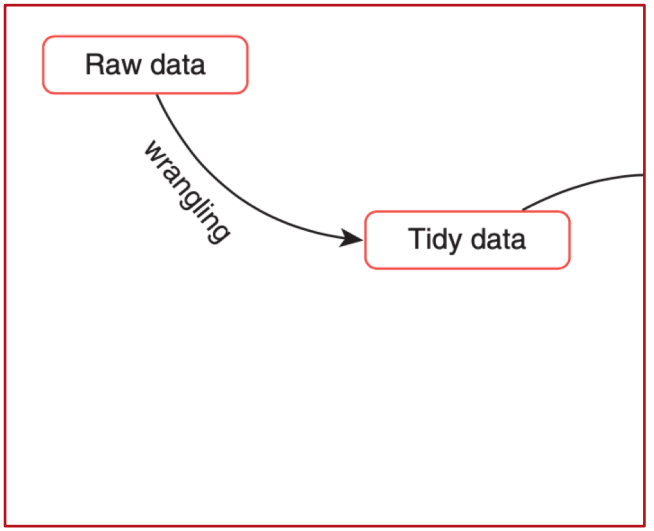
- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.

Reproduced from Andrews (2021)

Step 1: Data wrangling

Lancaster University

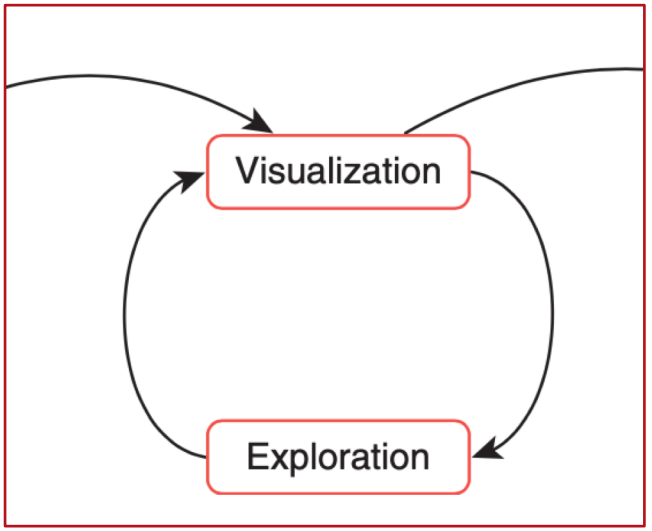
- Data can be messy (e.g., there might be missing values). Before analyzing the data, we need to ”tidy” it.
- Data wrangling: The process of preparing raw data for further analysis.



Step 2: Data exploration and visualization

Lancaster University

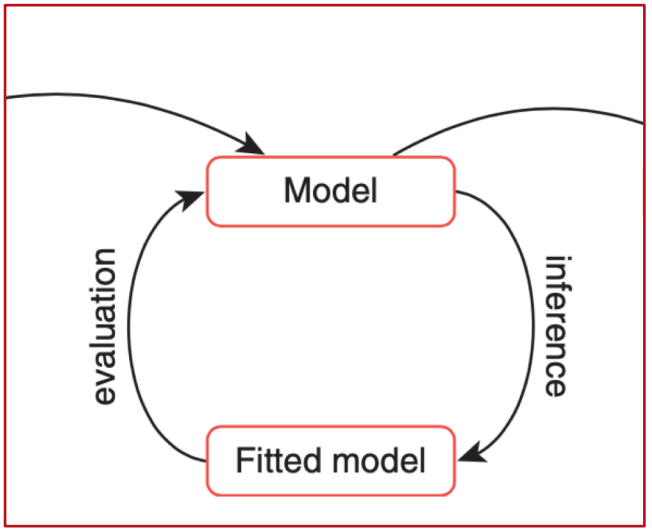
- Once data is tidy, we use computational tools and statistical methods for data exploration and visualization.
- The aim is to discover potentially interesting patterns and behaviors in the data.
- We will focus on this in sessions 3 and 4.



Step 3: Statistical modeling

Lancaster University

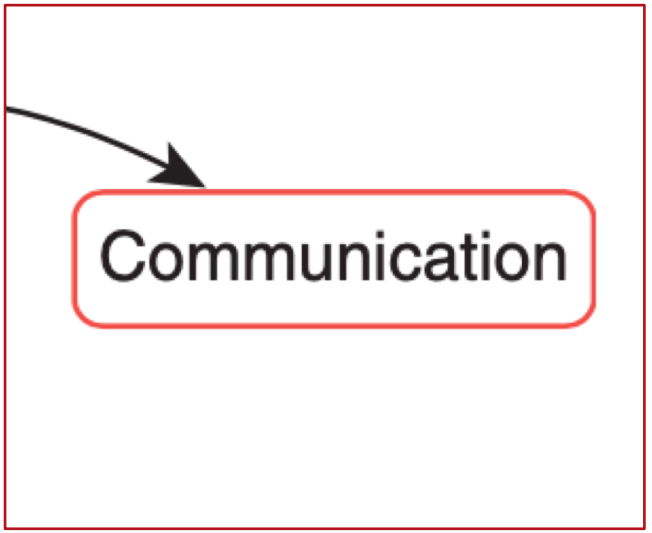
- The exploratory analysis (step 2) leads us to a probabilistic model of the data.
- This is a model of the phenomenon that generated the data.
- Modeling involves statistical inference and model evaluation.



Step 4: Communication

Lancaster University

- We disseminate our results (e.g., presentations, publications, corpora).
- But: Science should be **open, transparent and reproducible**, so we should also share materials, data and scripts via freely available platforms (e.g., OSF, IRIS).
- This way, others can replicate our studies and build on our research.



OSF

Lancaster University

Second steps in R

Solution to the homework

Lancaster University

The following table displays the scores of students in two foreign language exams, one administered at the beginning of term, the other at the end of term.

Student ID	Exam 1	Exam 2
Elin	93	98
Spencer	89	96
Crystal	75	94
Arun	52	65
Lina	34	50
Maximilian	50	68
Leyton	46	58
Alexandra	62	77
Valentina	84	95
Lola	68	86
Garfield	74	89
Lucy	51	70
Shania	84	90
Arnold	34	50
Julie	57	67
Michaela	25	37
Nicholas	72	90

Solution to the homework

Lancaster University

1. What are the mean scores for exam 1 and exam 2?
2. What is the difference between the two means?
3. What are the mean scores for the two exams if you remove extreme values (the top and bottom 20%) from each?

Student ID	Exam 1	Exam 2
Elin	93	98
Spencer	89	96
Crystal	75	94
Arun	52	65
Lina	34	50
Maximilian	50	68
Leyton	46	58
Alexandra	62	77
Valentina	84	95
Lola	68	86
Garfield	74	89
Lucy	51	70
Shania	84	90
Arnold	34	50
Julie	57	67
Michaela	25	37
Nicholas	72	90

Solution to the homework

Lancaster University

1. Based on the previous step (with outliers removed): What is the difference between the two means now? Please round the value before reporting the result.
2. Can you do steps 3 and 4 in a single command?

Student ID	Exam 1	Exam 2
Elin	93	98
Spencer	89	96
Crystal	75	94
Arun	52	65
Lina	34	50
Maximilian	50	68
Leyton	46	58
Alexandra	62	77
Valentina	84	95
Lola	68	86
Garfield	74	89
Lucy	51	70
Shania	84	90
Arnold	34	50
Julie	57	67
Michaela	25	37
Nicholas	72	90

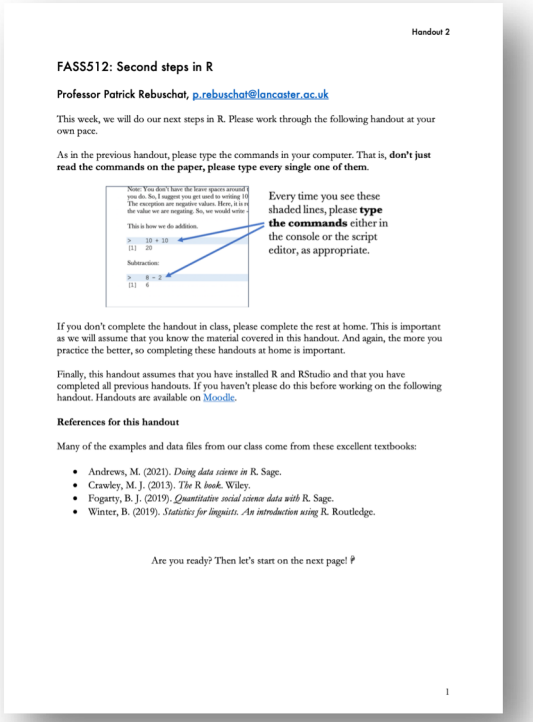
Lancaster University

Check the R script on Moodle, download the following file:

“RScript: Commands used to answer homework task”

Handout 2

1. Scripts
2. Installing and loading packages
3. Working directories and clean workspaces
4. Loading data (various formats)
5. Examining datasets
6. Closing your R session



Handout 2

FASS512: Second steps in R

Professor Patrick Rebuschat, p.rebuschat@lancaster.ac.uk

This week, we will do our next steps in R. Please work through the following handout at your own pace.

As in the previous handout, please type the commands in your computer. That is, **don't just read the commands on the paper, please type every single one of them**.

Every time you see these shaded lines, please **type the commands** either in the console or the script editor, as appropriate.

If you don't complete the handout in class, please complete the rest at home. This is important as we will assume that you know the material covered in this handout. And again, the more you practice the better, so completing these handouts at home is important.

Finally, this handout assumes that you have installed R and RStudio and that you have completed all previous handouts. If you haven't please do this before working on the following handout. Handouts are available on Moodle.

References for this handout

Many of the examples and data files from our class come from these excellent textbooks:

- Andrews, M. (2021). *Doing data science in* R. Sage.
- Crawley, M. J. (2013). *The* R *book*. Wiley.
- Fogarty, B. J. (2019). *Quantitative social science data with* R. Sage.
- Winter, B. (2019). *Statistics for linguists. An introduction using* R. Routledge.

Are you ready? Then let's start on the next page!

1

Lancaster University

Questions?

Lancaster University

Quantitative Research Methods

January 23, 2023

Professor Patrick Rebuschat

p.rebuschat@lancaster.ac.uk