Quantitative methods and statistics (In Sport and Exercise Science)

Daniel Hammarström

Contents


```
file.create('./docs/.nojekyll') # this makes hosting on github possible
```

```
## [1] TRUE
```

Introduction

Welcome to the course **Quantitative methods and Statistics (IDR4000)**. The course aims to give students an overview of methodological aspects within the field of sport and exercise-physiology. Specifically, planning, conducting and analyzing research projects with human participants will be covered. These course notes covers *almost* the entire course through the combination of video lectures, tutorials and references to the course literature and external resources.

INTRO VIDEO LECTURE

0.1 Practical information

0.1.1 Learning objectives

Learning objectives can be read in Norwegian here.

0.1.2 Learning strategies

The course will include lectures, laboratory exercises, computer exercises, seminars and student presentations. Lectures will be held on-line (zoom), as pre-recorded in this book and in-person on campus. Due to the current pandemic, you are required to do laboratory exercises in your cohort. Computer exercises require that you have special computer software installed on your computer. The software is free (see specific chapters in these course notes).

Assignments will be presented in this text with information on how to hand them in. The whole course is evaluated based on a portfolio (see below).

0.1.3 Course evaluation

As a student you can contribute to the quality of the course by engaging in course evaluation throughout the course. You will be asked to answer a pre-course questionnaire about your *expectations* and a post-course questionnaire about your *experiences*. You are also welcomed to take part in systematic discussions during the course about the quality of teaching and course material. With these notes I want to underline the importance of student participation in the continuous development of the course (and program) teaching/learning qualities.

0.1.4 Lecturers and course administration

In order of appearance

- Daniel Hammarström (daniel.hammarstrom@inn.no), is responsible for course administration and will be teaching statistics and molecular methods.
- Kristian Lian, Ingvill Odden and Lars Nymoen will act as teacher assistants in organizing methods in the physiology lab.
- Stein Olaf Olsen will act as a teacher assistant in the molecular lab.
- Prof. Carten Lundby will cover aspects CO2 re-breathing techniques (physiology).
- Prof. Finnur Dellsén will cover philosophy of science.
- Prof. Stian Ellefsen will teach molecular methods.

0.1.5 Updates, notifications and general communication

These course notes **will be** updated during the course. General information and last minute changes will be posted on Canvas, make sure to check it as part of your daily study routine.

0.1.6 Literature

A full list of recommended literature can be found here. Literature will be referenced in specific sections in these course notes.

0.1.7 Grades

The course is graded pass/fail.

0.1.8 Language

My (Daniel) first language is Swedish, I'm sure most of you will understand what I'm talking about. However, due to the fact that we accept international students to the program, most written communication and some lectures will be in English. You are not expected to write in English, it is however possible!

0.2 Assignments and Portfolio exam

The course is based on several assignments. Some of these assignments are to be handed in as part of a portfolio exam upon which your grade is based.

Assignments that are due during the course (arbeidskrav) are expected to be further improved after feedback from fellow students and teachers before inclusion in your portfolio.

The table below shows all assignments that are part of the course. Some are not to be included in the portfolio and some assignments are group assignments (see Table).

Assignment	Due date	Included in portfolio	Group assignment
Descriptive statistics, reliability and validity	2021-09-10	Yes	Yes
Study designs	2021-10-01	Yes	No
Extraction and analysis of DNA	2021-10-15	Optional[a]	Yes
Extraction of RNA and analysis of qPCR experiments	2021-10-	Optional[a]	Yes
Extraction and analysis of Protein	2021-10-	Optional[a]	Yes
Regression models and prediction from data	2021-10-	No	Yes
Drawing inference from statistical models	2021-10-	No	Yes
Statistical power and sample size calculations	2021-11-	Yes	Yes
Analyzing repeated measures experiments	2021-11-	Yes	No
Philosophy of science[b]	2021-11-	Yes	No

[a] Select one laboratory assignments for your portfolio exam. [b] This assignment is presented in connection with lectures.

In addition to arbeidskrav/assignments, smaller assignments and quizzes are presented in this book, but you are not required to do them to pass the course.

0.3 Other information

1 Introduction to data science

1.1 About data in the world of sport and exercise

Data is everywhere. Most of us walk around with a data collection device in our pockets all the time. This device (your mobile phone), records and store data about you all throughout the day. Such data are the basis of the *quantified self movement*[1] that have grown in popularity as capabilities to record data from daily

[1] Read more about the quantified self movement in this Wikipedia article

life has become better. People interested in quantifying their personal life does so for different reasons, but often with the intent to improve their health[2].

Much of these kind of data are readily available to us due to the fact we are protected by data privacy policies and regarded as personal data[3]. With some effort you yourself can get your data out of your iphone to explore, for example, your daily step count. I discovered that my phone(s) has been collecting data for me since 2016 and I tend to walk less steps on Sundays compared to Saturdays (see Figure 1).

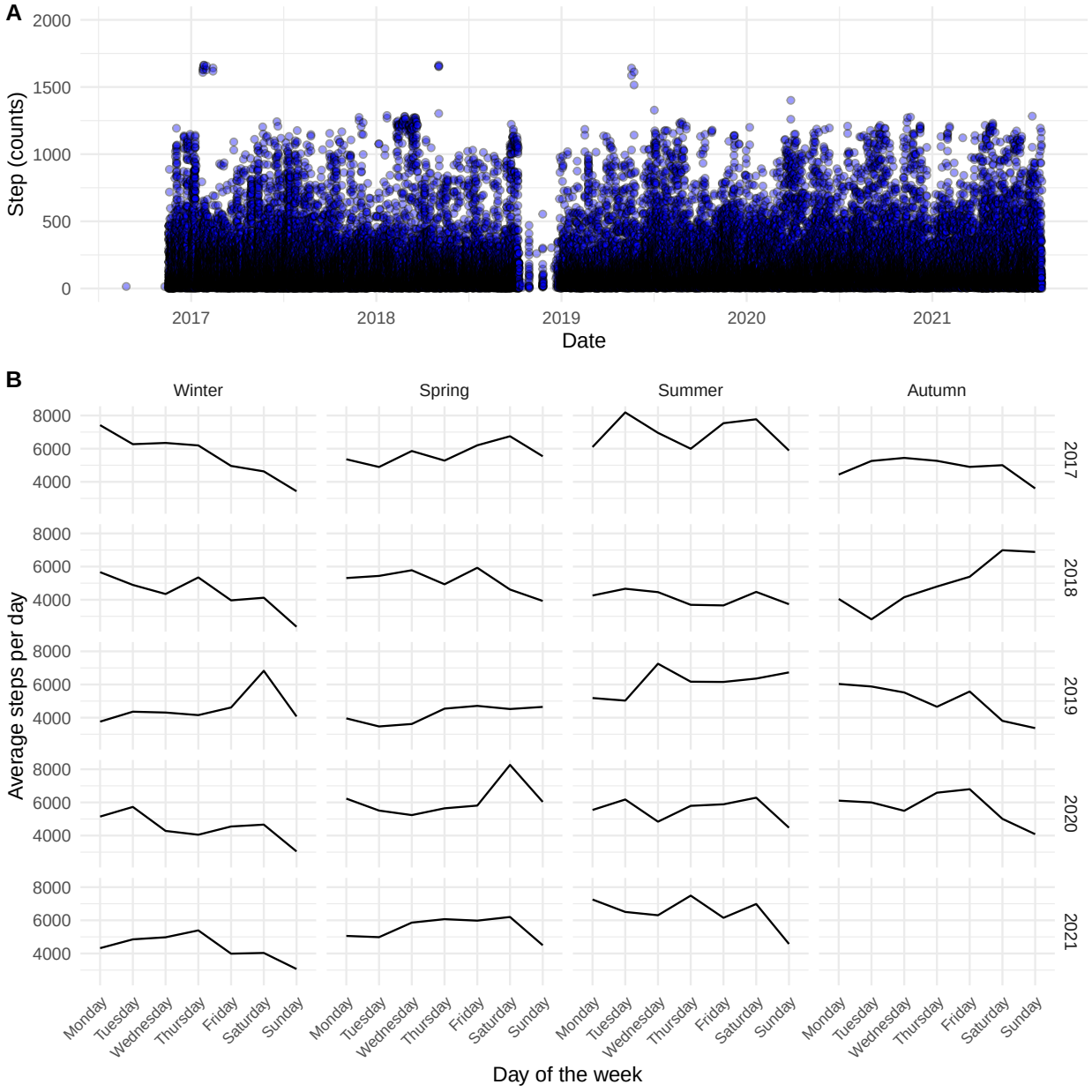

Figure 1: Step count data from my iPhone displayed as all avalable data points (A, after data cleaning) and average step per weekday, per year and season (B).

Data are also collected and stored in publicly available databases. Such databases are created for the purpose

[2]See this website for intriguing examples
[3]See e.g. Apples Privacy Policy.

to store specific types of data, such as soccer[4] or biathlon results[5], or biological information such as gene sequences[6]. Even data from scientific studies are now days often publicly available[7] meaning that we can perform scientific studies on unique data sets without collecting the data ourselves.

The above examples shows that there are abundance of data around and available to us. The problem is that it is hard understand all this data. This is where data science and data literacy comes in. In the world of sport and exercise, regardless if you are interested in doing scientific investigations, coach a soccer-team or individual athletes or help patients recover from surgery using exercise therapy, you are faced with the problem of handling and make sense of data. Some of the key skills and deeper understanding about data science are very much transferable between such areas of practice.

> **Think about the literature!** Spiegelhalter (The Art of Statistics, in the introduction chapter) talks about how statistics has evolved towards the broader field of data science. In data science, statistical theory and methods are just parts of the problem solving cycle. Try to think about how you would use the PPDAC cycle as a exercise coach and a scientist. What are the similarities and differences?

One broader aim of this course is for you to develop skills to better understand data.

1.2 Replication and Reproducibility

In scientific research, replication is a way to confirm scientific claims. When a result can be confirmed by an independent group of researchers, the claim is likely "more true". Many results will however never be possible to replicate due to the size of trials, costs and urgency of the research question. A recent example could perhaps be the many vaccine trials performed to develop a vaccines against COVID-19[8]. Other examples concern studies with unique study populations, such as large scale epidemiological studies [Peng et al., 2006], but the same could be said to be true for unique investigations in sport and exercise science.

When studies are not likely to be *replicated*, *reproducibility* of the analyses and results has been suggested to be a minimum standard for scientific studies. Reproducibility means that given the same datas, similar results or conclusions can be drawn by independent researchers [Peng et al., 2006].

Peng et al. [Peng et al., 2006] suggests that a *fully reproducible* study has

- Available data.
- Computer code (software) that produces the results of the study.
- Documentation that describes the software and data used in the study, and
- ways to share the data and code.

The above principally relates to the trust we can place in scientific results. However, the minimum standard of reproducibility has advantages also for the individual researcher (or master student)! When working with reproducible methods we will develop ways of documenting and automating our analyses. This will make it easier to collaborate with others. And, as it turns out, your most frequent collaborator is you, in the future!

A reproducible data analysis means that you will make it explicit and transparent. In a traditional data analysis, most activities are in the "black box". In order to avoid bias [Ioannidis, 2005], the "black box" needs to be opened and you need to actively make transparent decisions all along the analytic pipeline

[4] understat.com stores match specific data from major leagues. Data are available through software packages such as `worldfootballR`

[5] biathlonresults.com/ hosts results from the international biathlon federation. An example of analyzed data can be seen here.

[6] Ensembl and the National center for biotechnology information are commonly used databases in the biomedical sciences.

[7] We published our raw data together with a recent paper (Mølmen et al 2021 doi: 10.1186/s12967-021-02969-1.) together with code to analyze it in a public repository.

[8] https://www.evaluate.com/vantage/articles/news/snippets/its-official-covid-19-vaccine-trials-rank-among-largest

[Leek and Peng, 2015]. This pipeline preferably involves the whole problem solving cycle described by Spiegelhalter [**?**]. However the tools that we will learn about in this course focuses primarily on the steps from the experimental design to presentation of statistical results [Leek and Peng, 2015]. These steps includes data collection (and storage), data cleaning, exploratory data analysis, statistical modelling and statistical inference (and communication) [Leek and Peng, 2015].

1.3 Tools in data science

Ways to interpret and make sense of data involves different methods. These methods are now days often implemented in computer software. This means that when you as a practitioner (scientist, coach, analyst …) want to understand data, you have to master some kind of computer software. The most common software used to understand data is probably Microsoft's Excel. You can do amazing stuff with Excel! In the world of sport and exercise Excel has been used in such diverse activities such as scientific investigations, planning and recording training for Olympic medalists[9] and scheduling appointments.

For scientific research, most people use additional software to do statistical analyses. If you have spent time in higher education you have probably heard about SPSS, Stata or Jamovi. These are all specialized software used for statistical analyses.

The above mentioned tools can all be used as part of a fully reproducible workflow. However, there are software solutions that actually suits this requirement better than others. Going back to the description of reproducible science as made by Peng et al. [Peng et al., 2006], we want software where analyses can be

- Human- and computer-readable, meaning that we want to be able to write scripts, or computer programs that execute the analyses.
- Documented, meaning that along the code we want to be able to describe what the code does.
- Available and able to share with other, meaning that we analyses can be run on open and free software to maximize ability to share them.

This means that the software that we would prefer should be run using scripts (as opposed to point and click) and be free of charge (and open source, as opposed to expensive and proprietary). These criteria can be fulfilled when we use software that is written around the R language (although alternatives exists[10]).

R is a computer language that is especially well suited for reproducible data analysis. As users are able to contribute software extensions, also called packages, many specialized software implementation exists for different tasks, such as creating figures or analyses of specific data. Around R, people have been developing auxiliary software to enable reproducible data analysis. The negative part of all these opportunities is that using R requires some effort. The learning curve is steep!

Even though you might not use R ever again after this course, making and effort trying to learn it will let you know something about programming, capabilities of modern data science, statistical analysis and software/computers in general. These areas are all part of our modern society and are very much transferrable regardless of what computer language we are talking about.

In the next chapter of these course notes we will go through installing and starting up R.

1.4 Storing data in spreadsheets

Above, I mentioned spreadsheets like Excel. These are indeed great, but not great for reproducible science or data analysis. This is because they are not easily documented and scripted. The data is actually part

[9] The amount of time used by different coaches to create their own specific coaching software really makes many of them amateur software engineers. See for example this training journal from swedish orienteering.

[10] In addition to R, Python offers a free open source environment for reproducible analyses. The choice between the two are matter of taste.

of the analysis. Another danger with spreadsheets (like MS Excel) is that it re-formats your data. This is such a big problem for scientists that we have apparently started renaming genes this. Errors are frequent in spreadsheets, not only because renaming [**?**], but also because of bad formatting of formulas [**?**]. These are both reasons for using spreadsheets only what they are best used for: data input and storage.

> **Think about the literature** Broman and Woo[**?**] gives several pointers on how to use spreadsheets for data input and storage. Think your experince with Excel, what is the most common mistake you made when handling data in spreadsheets?

1.5 References and footnotes

2 Recording and storing data in spreadsheets

3 Creating your first graph

R is an excellent environtment for scientific graphs. There are three main systems for graphical output from R. The first is included in base R.

4 Creating your first table

5 Writing your first reproducible report

References

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124. URL https://doi.org/10.1371/journal.pmed.0020124https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/pdf/pmed.0020124.pdf.

J. T. Leek and R. D. Peng. Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549):612, 2015. ISSN 0028-0836. doi: 10.1038/520612a. URL http://www.nature.com:80/polopoly_fs/1.17412!/menu/main/topColumns/topLeftColumn/pdf/520612a.pdf.

R. D. Peng, F. Dominici, and S. L. Zeger. Reproducible epidemiologic research. *Am J Epidemiol*, 163 (9):783–9, 2006. ISSN 0002-9262 (Print) 0002-9262 (Linking). doi: 10.1093/aje/kwj093. URL http://www.ncbi.nlm.nih.gov/pubmed/16510544.