# Week 1: Introduction & Organization MATH-516 Applied Statistics

Linda Mhalla

2024-02-19

### Section 1

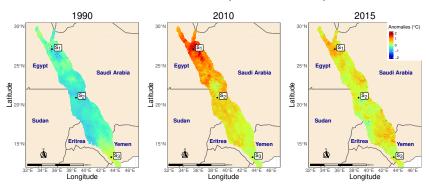
Introduction

# Why Applied Statistics

- Data are everywhere!
  - Large amount of data that come with uncertainty and variability
  - We want to learn something about these data, but what?

# Why Should I Learn Applied Statistics?

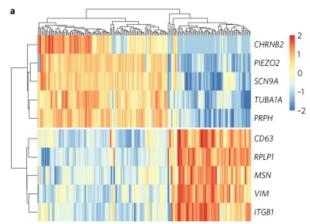
To understand spatial and temporal (climate change) dependencies



Source: Castro-Camilo et al. (2020)

# Why Should I Learn Applied Statistics?

### To understand variability in gene expressions



Source: Schwartzentruber et al. (2018)

### Job of a Statistician

Understand a complex world by describing it in relatively simple terms that capture essential aspects of its structure, and provide us some idea of how uncertain we are about that knowledge

- think about uncertainty and bias (anticipate and reduce it)
- build models emulating nature
  - inference about the models leads to conclusions about nature but what if the model is a poor representation of nature?
- provide interpretable models allowing for rational conclusions
  - prediction vs. information extraction
  - all models are wrong ⇒ critical model validation
- estimate variation (⇒ confidence intervals, significance)
- draw conclusions from data
- traditional role: statisticians invited to analyze existing data
  - e.g. does the existing data contain the desired information?
- modern role: collaborative step-by-step
  - from acquisition of data to presentation of results
  - interdisciplinary communication

# Domains of Application

- actuarial science
- biostatistics (medicine, pharma, genetics, etc.)
- business
- chemometrics
- econometrics
- epidemiology
- finance
- geostatistics
- machine learning and AI
- official statistics (demography, surveys, etc.)
- psychology
- quality control
- reliability
- physics
- signal processing
- ..

# Why Models?

#### We build models in order to

- understand the nature,
- ② predict the future, and
- ③ control the world. [or was it rule the world?]
  - Patrick Winston (former director of the AI lab at MIT)
- is the main goal of applied statistics
  - interpretation
  - parsimony
- is the main goal of Al
  - average accuracy
- is just to slam the message home

All models are wrong, but some are useful.

George Box

### Section 2

Organization

## Organization

- This course: a taste of real world problems and challenges for future statisticians
- Emphasis on models and inference: we overview several techniques of statistical modelling, and discuss real life problems
- Project based evaluation:
  - you will be challenged to use these tools to learn from the data
- This course is problem-driven, and hence you are responsible for understanding what are the appropriate models to analyze the data and to implement these computationally
  - reproducibility of the analysis, and effective and rigorous communication of your analyses are assumed acquired from MATH-517

### Classes

#### Lectures

Teacher: Linda Mhalla

Time: Monday 13:15-14:00 and Wednesday 13:15-14:00 (biweekly)

• Place: MA A1 10

#### **Exercises**

Teacher: Almond StöckerTime: Monday 13:15-16:00

• Place: CM 0 12

A schedule can be found here

### Prerequisites

Learning Prerequisites (from the course book):

- REQUIRED COURSES
  - Regression Methods
  - Statistical Computation and Visualization (MATH-517)
- RECOMMENDED COURSES
  - Time Series
  - Statistical Inference

### Prerequisites

Learning Prerequisites (from the course book):

- REQUIRED COURSES
  - Regression Methods
  - Statistical Computation and Visualization (MATH-517)
- RECOMMENDED COURSES
  - Time Series
  - Statistical Inference

Learning Prerequisites (my strong personal recommendation):

- required course:
  - Statistical Computation and Visualization (MATH-517)
- somewhat helpful courses:
  - Regression Methods
  - Time Series
  - Statistical Inference

### Content

- Week 1: Intro
  - Project 1: Snow Data
- Week 3: Poisson Regression
  - Project 2: Climate-related Disasters
- Week 5: Logistic Regression
  - Project 3: TBD
- Week 7: Linear Mixed Models
  - Project 4: Calorie Expenses in the Human Body
- Week 9: Causal Inference
  - Project 5: Tubingen Datasets
- Week 11: Extreme Value Theory
  - Project 6: The Vargas Tragedy or Rainfall data in Switzerland
- Week 12: Statistical Consulting
  - Project 7: TBD
- Week 14: Oral Exam
  - discussing your submitted projects

**Project deadlines**: Project assigned on (Monday of) Week X has a deadline on Sunday evening of Week X+1, i.e. there are always 2 weeks per project.

# **Project Submission**

R/RStudio, Markdown/Quarto (or eventually LaTeX) and GitHub will be needed

• submissions are made through GitHub Classroom (see dedicated tutorial on the MATH-517 course website)

### **Evaluation**

The grade will reflect on the quality of the projects, which are expected to

- identify questions of interest
  - some will be provided during the description of the project
- choose appropriate models to analyze the data
  - demonstrate understanding of the models used
- implement the models in R
  - though this is probably not displayed in the report
- critically evaluate shortcomings of your models (model diagnostics)
  - a good solution may be to provide more than one model at first and eventually compare those
- use a final model to answer the questions of interest

### **Evaluation**

It is imperative that the final report is

- readable
  - figures need to have self-explanatory captions, appropriate font size, and be generally of a decent quality
  - there should be no code in the report, unless it significantly improves clarity of the report (e.g. R table instead of a Latex table is permitted for simplicity) and even in such a case it has to be verbally explained around any code chunk what it actually does
- reproducible
  - i.e., the R Markdown file can be run again on a different machine inside a copy of your Github repo
  - code is well commented

This makes projects iterative work, where most of the work done (e.g. data exploration and model selection) is underrepresented in the final report

# Report Writing

### Some (paraphrased) quotes:

- If a work is not compiled into a report, it does not exist. If the report is not readable and reproducible, the work is useless.
- Think about what you want to write and then write it as clearly and economically as possible. That is all there is to academic writing.

# Report Writing

- The length of the reports should be between 8 and 15 pages, all included
- The reports should include:
  - An abstract
  - An introduction presenting the problem
  - A description of the data
  - An exploratory analysis of the data (with some relevant plots)
  - A statistical description of the method(s) used to analyse the data
  - An interpretation of the results (with some relevant plots)
  - A conclusion

### **Evaluation**

- 7 projects in total (for you to choose from)
  - specific data and tasks to perform
  - done individually, but exchange of ideas (but not the code) is encouraged
- 5 projects will form your chosen portfolio
  - Project 1 is mandatory
  - at least one from Projects 2-3
    - you will get a detailed feedback on this one
  - at least one from Projects 4-5
  - at least one from Projects 6-7
- Project 1 (linked heavily to MATH-517) gets a grade of its own (not provided), the rest will be graded during the final oral examination
- ullet This course is "without withdrawal" (submit Project  $1 \equiv {\sf commit}$ )

# References (for the 1st half of this course)

- Wood (2017) Generalized Additive Models: an Introduction with R (2nd ed.)
  - even though mainly about GAMs, this book has a short and practical exposition to linear models and GLMs that has a value of its own
  - computational flavor
- Davison (2003) Statistical Models
  - nice reference due to the content breadth; is self-contained, but no R
    code
- Gelman & Hill (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models
  - focuses very much on interpretation
  - eloquent/lengthy and not always to the point (or precise)
- Wickham & Grolemund (2017) R for Data Science
  - useful guide to tidyverse, i.e., data exploration and manipulation

### A Quick Poll

What are your expectations from this course?

- want/need to get out of it
- Click here to answer

