Humanities Data Analysis

alessandra.demulder@uantwerpen.be |
benjamin.nagy@uantwerpen.be

MA DTA | 2003FLWDTA | 2024-2025

Overview

- Introduction to statistics for Humanities students
  - Focused on intuition and understanding (rather than on mathematics and derivation)
  - "Traditional stats 101": basic procedures, no fancy statistics
  - Pragmatic analysis of engaging, real-world datasets from the Humanities
- 5 sessions (3 ECTS) :
  - Session on Fridays (obligatory), 10:15-13:15, S.R.213
    - 18/10, 25/10, 08/11, 15/11, 22/11
- Format
  - Interactive, modular
    - Divided in 3 blocks of 40'-45', followed by 10'-15' breaks
    - Slides session (*theory with lots of examples*) or notebook session (*practice*)
  - Homework exercises: optional, highly recommended (but not graded)

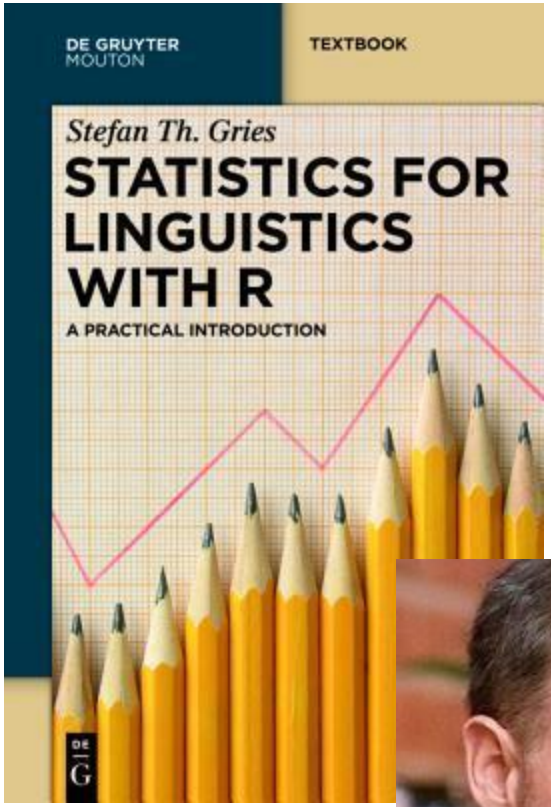
Evaluation

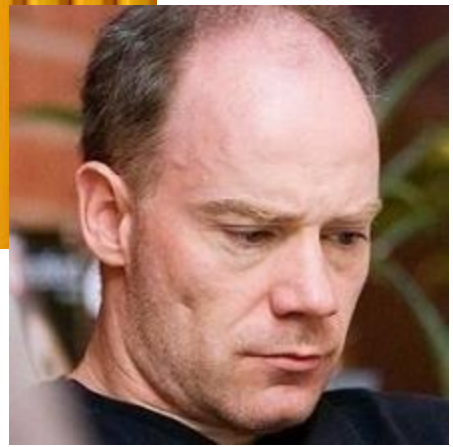
- Final evaluation: written exam
  - During January exam sessions (no resits in January, resits in August/September)
  - Open book: all resources allowed, except for external help (Chat(GPT), discussion, etc.)
  - Additional analysis of known and unknown datasets
  - Very similar to homework style
  - See test exam from last year (?)

Handbook [optional]

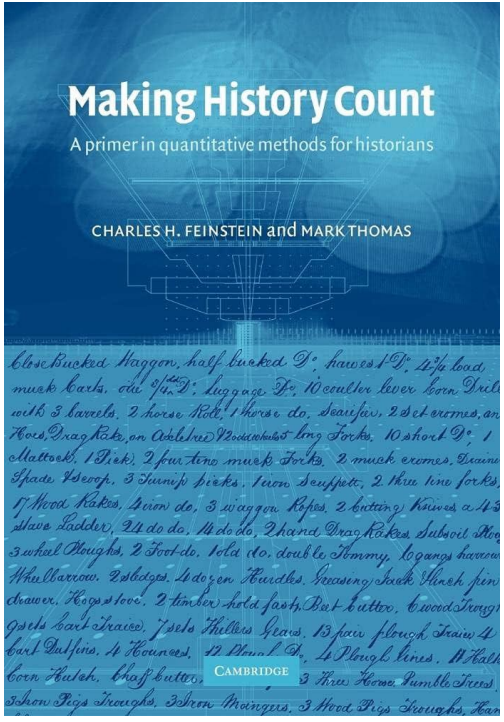
Stefan Th. Gries, *Statistics for Linguistics with R. A Practical Introduction*. 2nd ed. De Gruyter Mouton (2013)

- Well-known textbook by leading figure in the field of corpus linguistics
- Complementary with teaching materials (but different order and emphases)
- We will refer to it in a consistent way
- Gries more verbose: no manual recalculations (due to lack of time)
- We will do more than just linguistics and we won't use R
- Focus: **class materials** (Gries = optional background reading)





Handbooks [extra optional]



- No use of any programming languages
- Written for undergraduate students with no background in statistics
- Very accessible way to grasp (most of) the concepts and methods we'll be implementing
- Loads of examples for thesis inspiration?

How to structure your course folder?

edu.nl/djreg

UAntwerpHDA24 (*top directory*)

- Session 1 (shared weekly)
  - slides, notebooks, ...
- Session 2 (shared weekly)
  - slides, notebooks, ...
- datasets
  - arthur
    - manuscripts.csv
    - manuscripts.xlsx
  - chat
    - chat.tsv
    - readme.txt

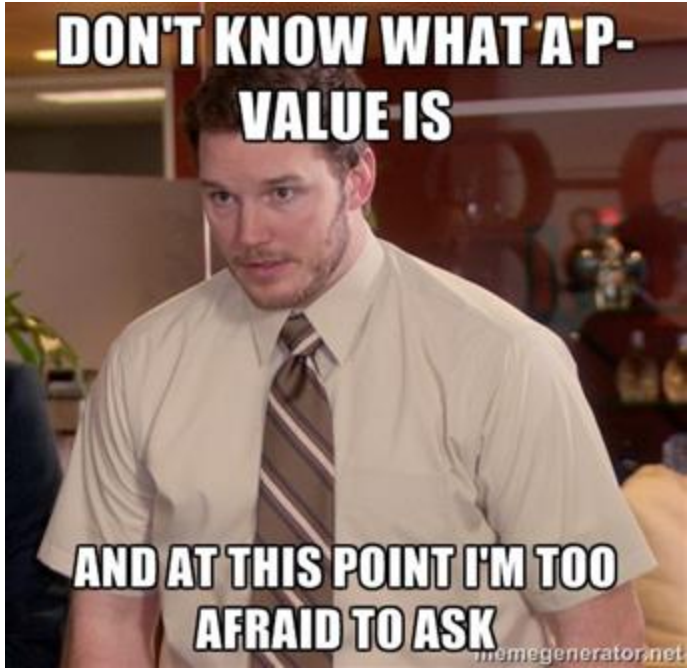
Relative paths from within the notebooks:

`../datasets/subfolder/...'

Course materials

One folder per session:

- Slides (PDF)
- Class notebooks (.ipynb)
- Homework notebooks (.ipynb)



What is statistics (not)?

And how is it different from machine learning?

The application of quantitative (*computational*) methods to (*observed*) data, to make scientifically informed claims about a phenomenon in the real world

- Our data is typically a (noisy, imperfect) sample from a larger population
- Because we work with a sample, there will be uncertainty
- We have to acknowledge that uncertainty when formulating claims
- Statistics is scientific, precisely because it gives us a nuanced way to describe uncertainty

Similarities between statistics and machine learning?

Although they adopt different goals and terminology, both are essentially about fitting a function f_Θ that maps an input (X) to an output (Y), using parameters Θ:

$$f_\Theta(X) \rightarrow Y \text{ (or } \hat{Y})$$

In statistics, we mostly care about the parameters Θ used inside *f*: these are what we want to "estimate", these are what we would like to "infer" from the observed sample (X, Y). In machine learning, we mostly care about the predicted $\hat{Y}$ (and how different it is from Y). Which precise Θ is used for this, is less important.

Differences machine learning and statistics?*

(*This is a caricature)

Machine learning	**Statistics**
Emphasis on *prediction* of unseen, out-of-sample data	Emphasis on *description* of observed, in-sample data
Emphasis on performance	Emphasis on understanding
Lots of code	Not so much code
Complex models (e.g. neural nets)	Simple model (e.g. linear regression)
Results more essential than theory	Theory more essential than results
Terminology (e.g. categorical feature)	Terminology (e.g. factor)

Definitions

- Statistics is (a part of) data science
  - "Data science is statistics on a Mac" (recent term!)
  - Data science = **visualization + statistics + machine learning** (cf. content of Jake van der Plas, *Python Data Science Handbook*)
- The practice (art?) of reasoning in the absence of (complete) certainty
- Inference from observed data: infer a set of "explanations" for the data that can be ranked by plausibility