

# Statistics for Linguists

2023-10-18

Adam J.R. Tallman

# Course management

- Moodle
  - Slides, homeworks, textbook, other optional readings, databases for exercises
- Project
  - You are not graded for homeworks,
  - You must: write a project description in December
  - You must: write a draft of your paper
  - You must: write a final paper

# Textbooks

## The Seven Pillars of Statistical Wisdom

STEPHEN M. STIGLER



## How to do Linguistics with R

Data exploration  
and statistical analysis

Natalia Levshina

John Benjamins Publishing Company

# Starting

- Download R
  - <http://lib.stat.cmu.edu/R/CRAN/>
- Download R Studio
  - <https://posit.co/download/rstudio-desktop/>

# What this course does

- Using R
- Descriptive statistics
  - Aggregating and summarizing data (e.g. sums, means etc.)
- Inferential statistics
  - Making inferences about the validity of a hypothesis from data
- Exploratory data analysis
  - Figuring out what is going on with data, with plots and machine learning etc.

# Some terms you may have heard

- P-values
  - Counterintuitive and often misinterpreted
  - Used in 'null hypothesis testing'
  - Becoming controversial for actual stati
- Linear models ("Regression")
  - When you have to continuous variables and they change together
- Multivariate linear model
  - When you have more tan two quantitative variables that change together

# Other concepts

- Binominal / logistic models
  - Predicting one of two outcomes from continuous data
- Interactions
  - Where the strength and type of relationship between two variables depends on another variable
- Multilevel models
  - Looking at relationships between lots of groups at once allowing different aspects of the model to vary depending on the group

# Bad practices

- data dredging:
  - grouping the data in different ways until it looks convincing
- p-hacking:
  - doing statistical tests on data in different ways until a desirable p-value (e.g. below 0.05)
- causal salad:
  - throwing in lots of variables without a theory and consideration of causal structure



# ‘Best’ (or better) practices

- Practice preregistration instead of data dredging
  - Say what you think the hypothesis is going to be in some report prior to publication and make a distinction between confirmatory and exploratory analysis
- Causal inference instead of causal salad
  - Use statistical models to test specific causal relations