85-309:
Statistical Concepts and Methods for Social and Behavioral Science

Spring 2023

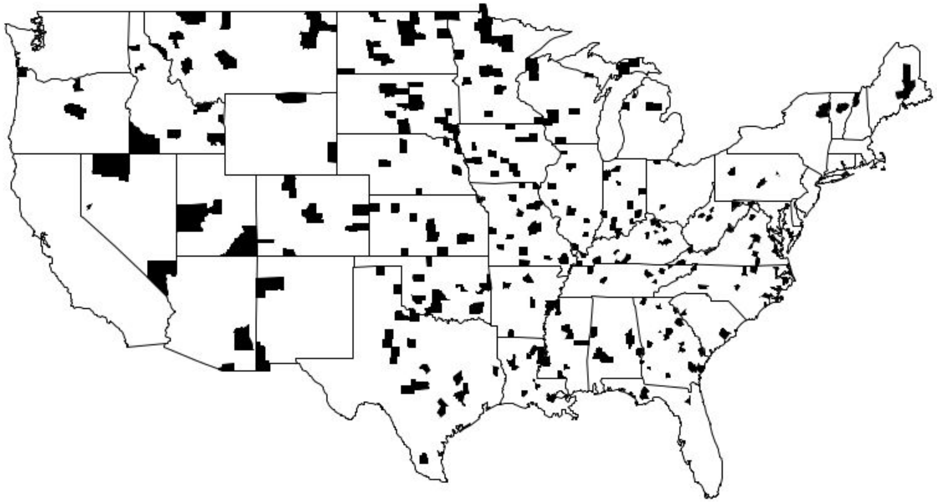
Instructor: Javier Rasero
(call me Javi)

Why is statistics important for yinz?

Why is statistics important for yinz?

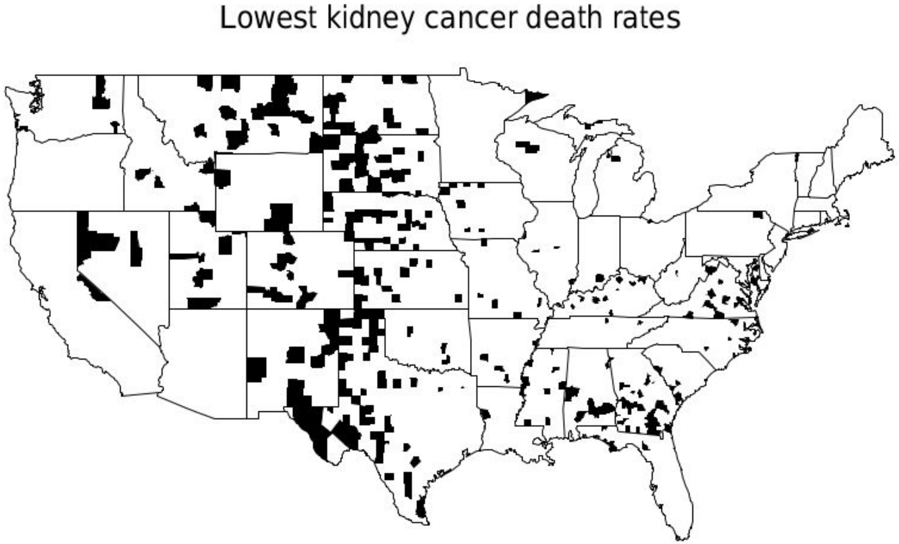
Look at this map of the **highest** cancer death rates (1980-1989):
Why is that there are many counties in the Great Plains shaded compared to near the coasts?

Highest kidney cancer death rates



Why is statistics important for yinz?

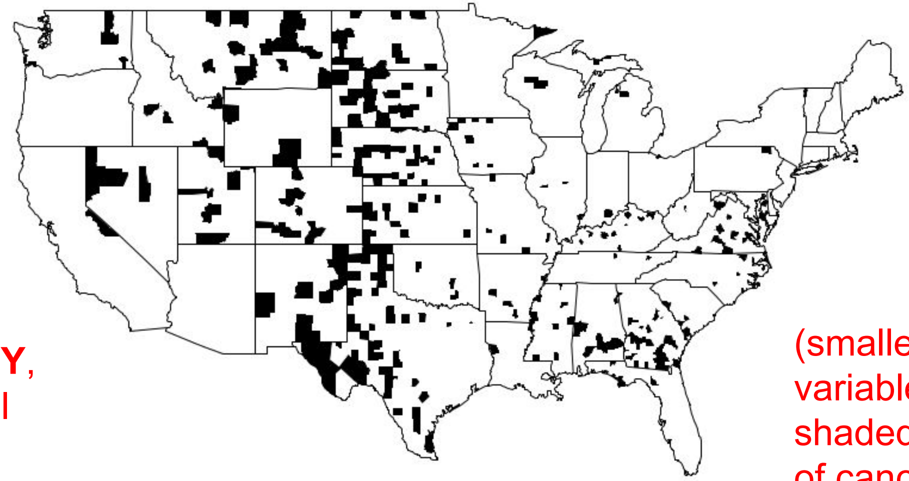
Look now at the map of **lowest** cancer death rates (1980-1989):
Why there are still many more counties in the Great Plains shaded compared to near the coasts?



Why is statistics important for yinz?

Look now at the map of **lowest** cancer death rates (1980-1989):
Why there are still many more counties in the Great Plains shaded compared to near the coasts?

Lowest kidney cancer death rates



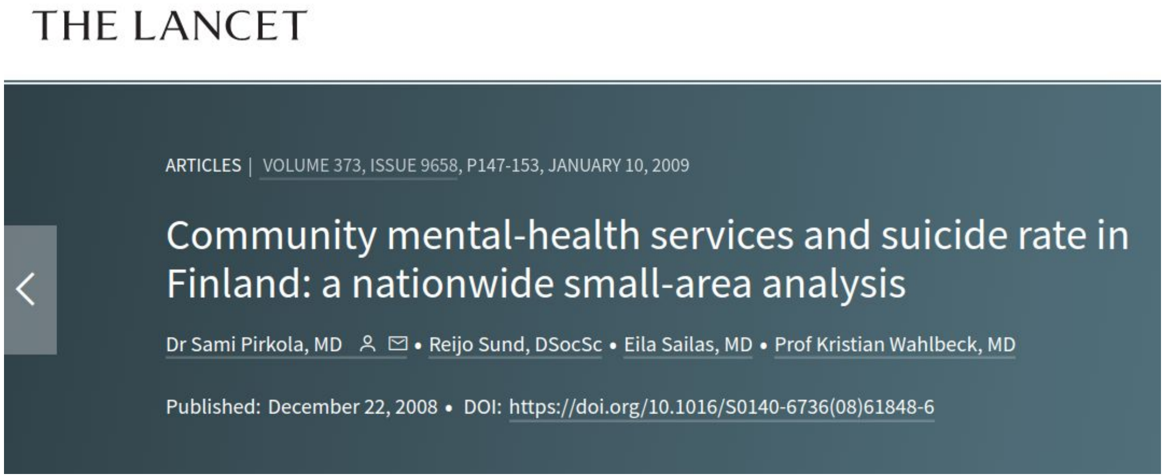
Explanation: **VARIABILITY**, a very important statistical concept!

(smaller counties are much more variable, and hence likely to be shaded, even if the true probability of cancer in these counties is nothing special!)

Why is statistics important for yinz?

- Study design: You need to at least understand the basics of stats to be good at designing psychological studies.
- Understanding research: if you really want to understand the psychology, you need to be able to understand what other people did with their data, which means understanding a certain amount of statistics.
- Performing research: In generals not many lab can afford paying money to an statistician, so chances are that you’ll need to be pretty self-sufficient if you want to do psychological research.

Example



THE LANCET

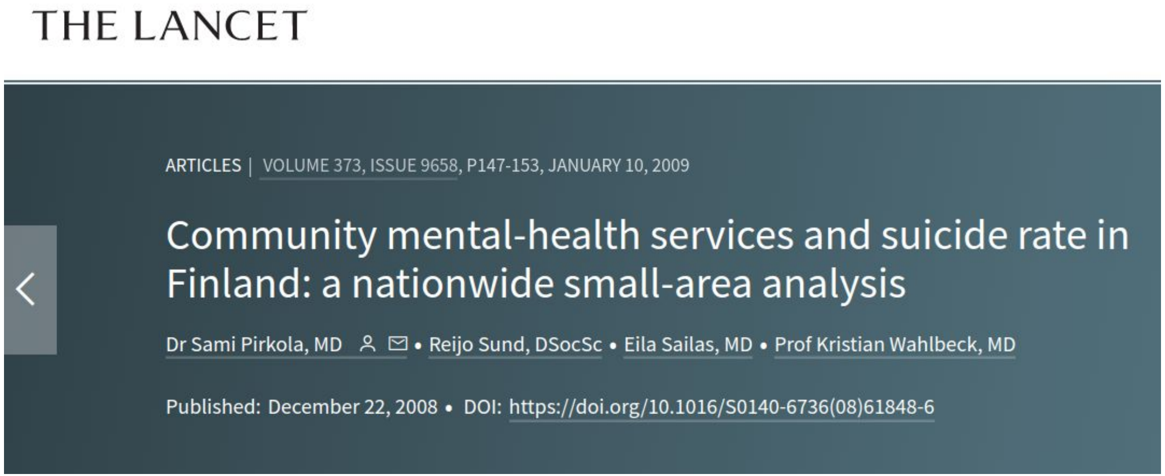
ARTICLES | VOLUME 373, ISSUE 9658, P147-153, JANUARY 10, 2009

Community mental-health services and suicide rate in Finland: a nationwide small-area analysis

Dr Sami Pirkola, MD • Reijo Sund, DSocSc • Eila Sailas, MD • Prof Kristian Wahlbeck, MD

Published: December 22, 2008 • DOI: https://doi.org/10.1016/S0140-6736(08)61848-6

Example: setting up an hypothesis!



THE LANCET

ARTICLES | VOLUME 373, ISSUE 9658, P147-153, JANUARY 10, 2009

Community mental-health services and suicide rate in Finland: a nationwide small-area analysis

Dr Sami Pirkola, MD • Reijo Sund, DSocSc • Eila Sailas, MD • Prof Kristian Wahlbeck, MD

Published: December 22, 2008 • DOI: https://doi.org/10.1016/S0140-6736(08)61848-6

Mental-health services have been transformed from hospital-centred to integrated community-based services. Has this transformation been effective?

Example: let’s collect some data to answer this!

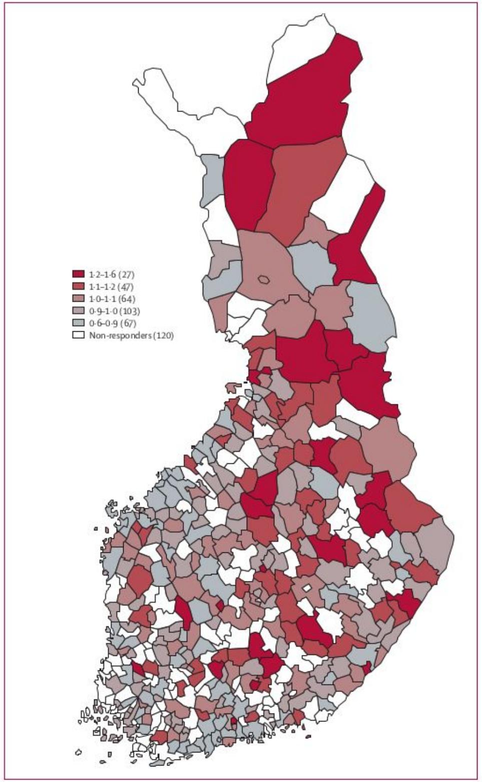
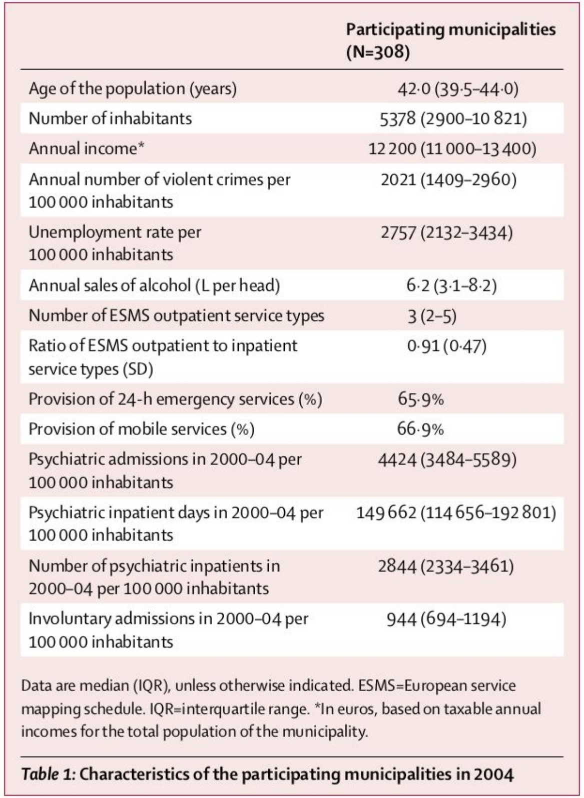


Figure 1: Local age-adjusted and sex-adjusted suicide risk between 2000 and 2004

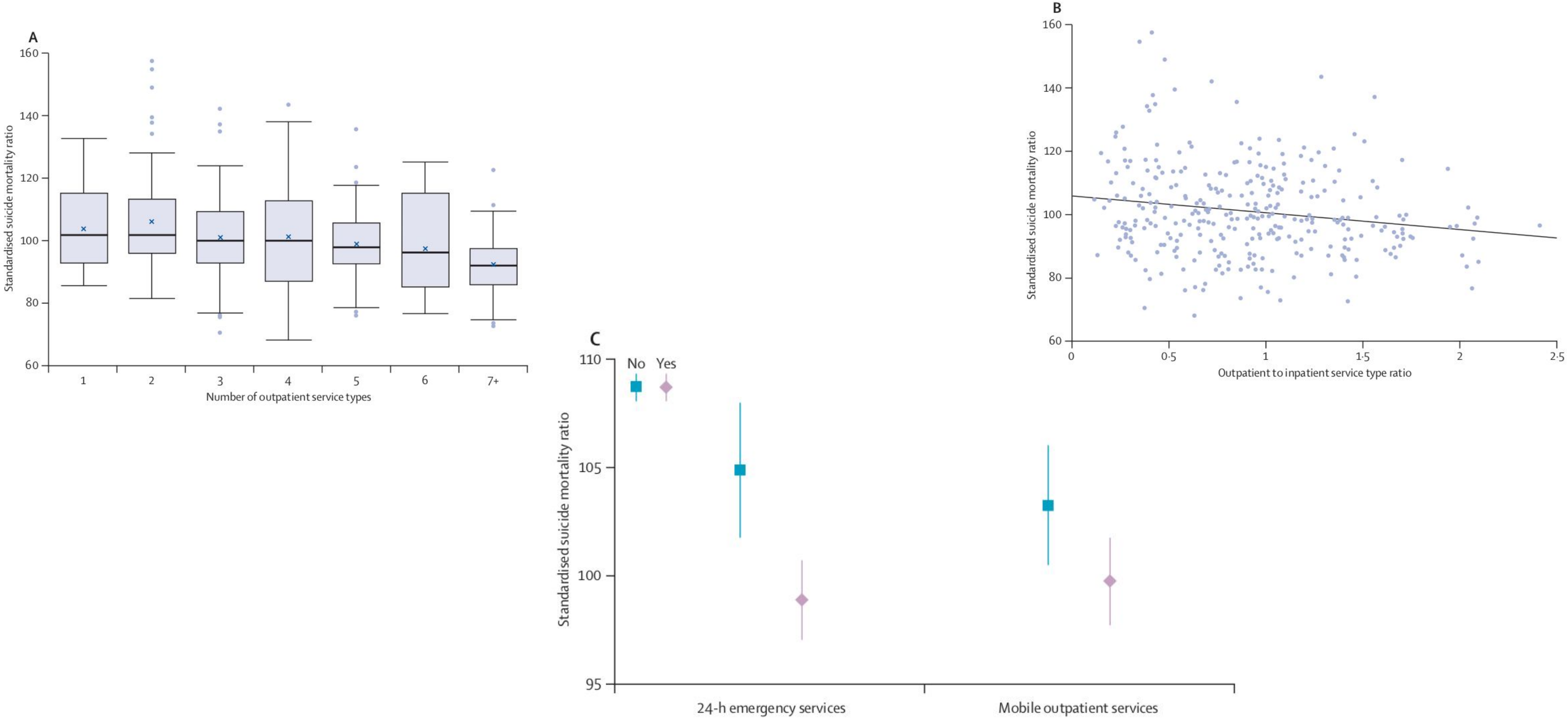


	Participating municipalities (N=308)
Age of the population (years)	42·0 (39·5–44·0)
Number of inhabitants	5378 (2900–10 821)
Annual income*	12 200 (11 000–13 400)
Annual number of violent crimes per 100 000 inhabitants	2021 (1409–2960)
Unemployment rate per 100 000 inhabitants	2757 (2132–3434)
Annual sales of alcohol (L per head)	6·2 (3·1–8·2)
Number of ESMS outpatient service types	3 (2–5)
Ratio of ESMS outpatient to inpatient service types (SD)	0·91 (0·47)
Provision of 24-h emergency services (%)	65·9%
Provision of mobile services (%)	66·9%
Psychiatric admissions in 2000–04 per 100 000 inhabitants	4424 (3484–5589)
Psychiatric inpatient days in 2000–04 per 100 000 inhabitants	149 662 (114 656–192 801)
Number of psychiatric inpatients in 2000–04 per 100 000 inhabitants	2844 (2334–3461)
Involuntary admissions in 2000–04 per 100 000 inhabitants	944 (694–1194)

Data are median (IQR), unless otherwise indicated. ESMS=European service mapping schedule. IQR=interquartile range. *In euros, based on taxable annual incomes for the total population of the municipality.

***Table 1:* Characteristics of the participating municipalities in 2004**

Example: let’s inspect these data!



Example: let's build a statistical model to test our hypothesis!

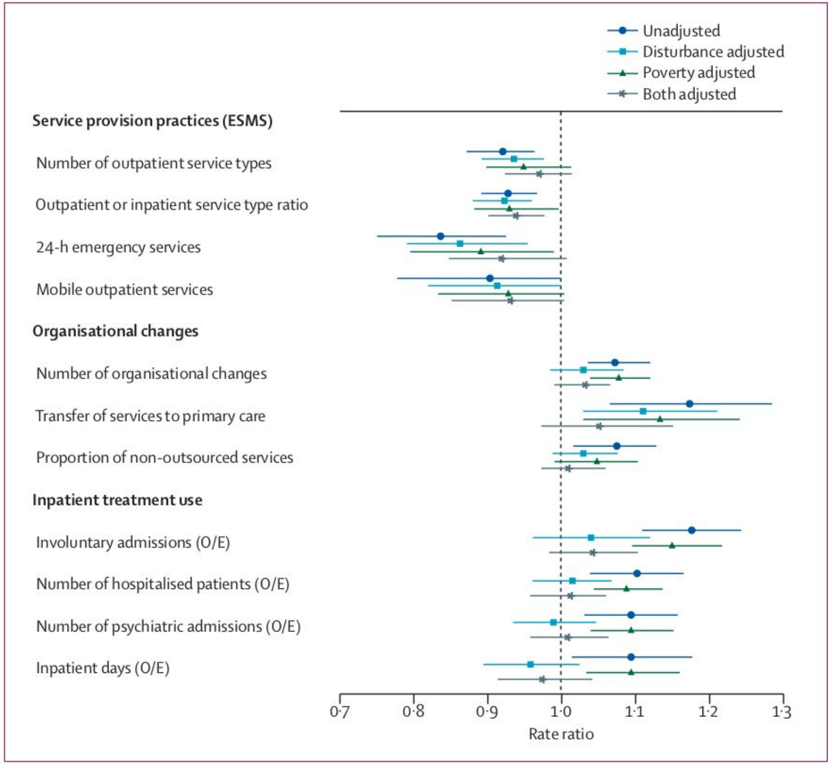


Figure 3: Factors associated with variation in suicide rate at the municipal level

Example: let's build a statistical model to test our hypothesis!

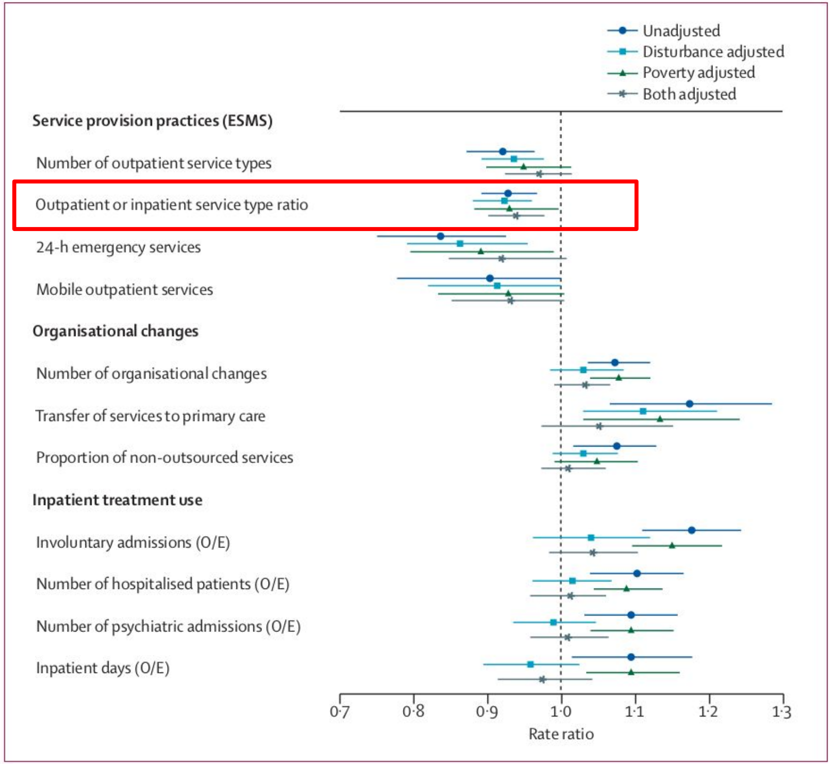


Figure 3: Factors associated with variation in suicide rate at the municipal level

Using statistics to understand (and improve!) the world

1. **Identify the question or problem.**

   Hypothesis: “Outpatient services help lower suicide rates”

2. **Collect relevant data.**

   308 subjects from Finland

3. **Analyze these data.**

   Descriptive statistics: e.g. suicide rates vs number of outpatient services

4. **Draw a conclusion.**

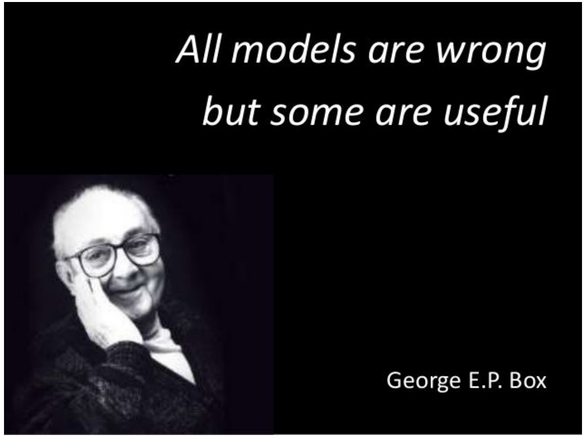
   Are these two variable independent given the data (**Null hypothesis**)?

Keep in mind about statistical methods

Statistics allows us to connect scientific theories to the world…

…However, this will usually show a simplification of the reality...

…Nonetheless, it will still useful!



All models are wrong
but some are useful

George E.P. Box

This course

What you should get from this course

1. To understand the data and know how to explore them.
2. Use statistical software to summarize these data numerically and visually.
3. Ask questions about the data and design a plan to answer them.
4. Build statistical models/tests and understand which ones are more appropriate and why.
5. Make statistical inferences to help answer your initial questions.
6. Present results in a robust and clear way.
7. Understand the claims that others make from data and be able to critique them.

What I expect you to get

- Get statistical intuition.
- Know what to do with your data (e.g. how to visualise and summarise these).
- Get familiar with R programming.
- More importantly, lose any fear for statistics!

Class structure

Classes will begin with a formal lecture. After that and a short break, we will get our hands dirty by walking together through a jupyter-notebook tutorial, running in a R statistical language environment. This will allow you to put into practice what we have learned and more importantly, set a basis for you to be able to complete the assignments. Lab assignments will be posted at the end of the class. The deadline will be one week after the completion of **each phase.**

Assessment and Grading

- Assignments (Homework): 20%
- First exam: (20%)
- Second exam: (20%)
- Final project: 40%
- Participation (10%) (Extra credit)

Assessing your understanding

Assignments (Homework): The objective of the assignments is to give you a first-hand experience with statistics and data analysis. For that we will be using jupyter-notebooks, running in a R statistical language environment. You may start working on the assignments during the class session, but note that these are designed to take more than just the class time, so you will probably need to continue working on them in order to submit before the due date. The assignments will also train you for the exams and the final project.

Assessing your understanding

Exams: There will be two exams, the first one halfway through the course (phase 1 and 2) and the second one at the end (phase 3 and 4). Each exam will consist of jupyter-notebooks with theory and practice questions. Students will have **24h** for their completion.

Assessing your understanding

Project: The objective of the project is to give you independent applied research experience using real data and statistical methods. The goal will be to synthesize what you have learned in the labs over the course of the semester, and to show that you understand which analyses are appropriate and interesting for which kinds of data.

Course structure

Phase 1: Getting familiar with the software and data:

- Week 1: Introduction to R and data visualization using ggplot.
- Week 2: Descriptive statistics (Look and understand the data). Introduction to data transformation using dplyr.

Phase 2: Probability, sampling and introduction to statistical inference

- Week 3: Probability and distributions.
- Week 4: Parameter estimation.
- Week 5: Hypothesis testing.

Course structure

Phase 3: Statistical tools and methods

- Week 6: methods with univariate data (one-sample t-test, single proportions)
- Week 7 and 8: methods with relate bivariate data (e.g. two-sample t-test, correlations)
- Week 9: Simple linear regression (Regression between two variables)
- Week 10-11: General Linear regression.

Phase 4: Question your analyses and results!

- Week 12: Statistics not meeting assumptions (non-parametric testing)
- Week 13: Avoid false positives: Correct for multiple testing.
- Week 14: Do I have too many variables in my regression model?

Resources

Main textbook
(Freely available):

Learning Statistics with R

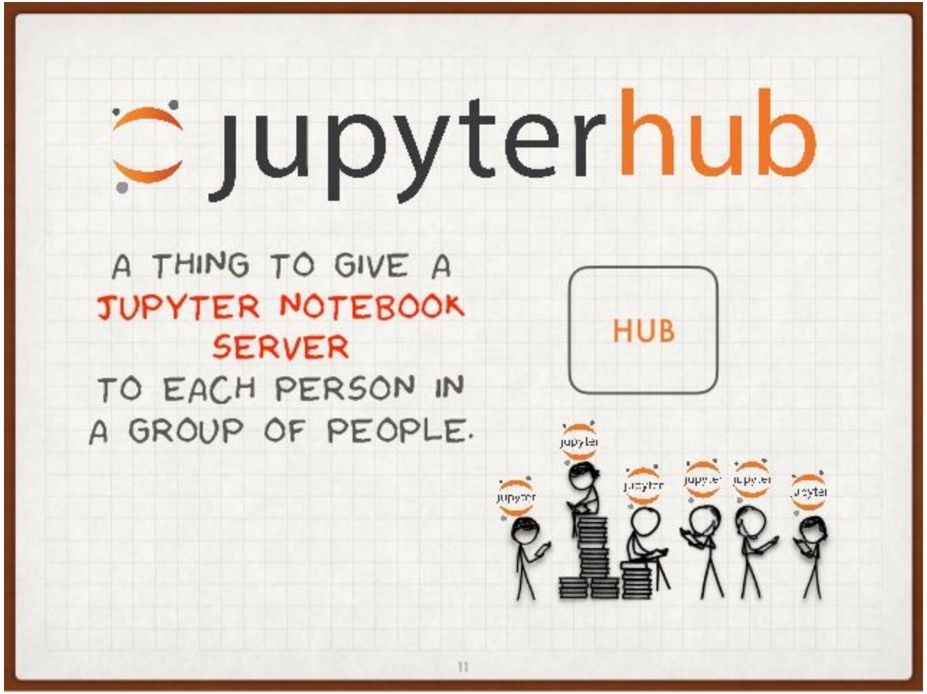
by Danielle Navarro

https://learningstatisticswithr.com/

Software:

R + jupyter

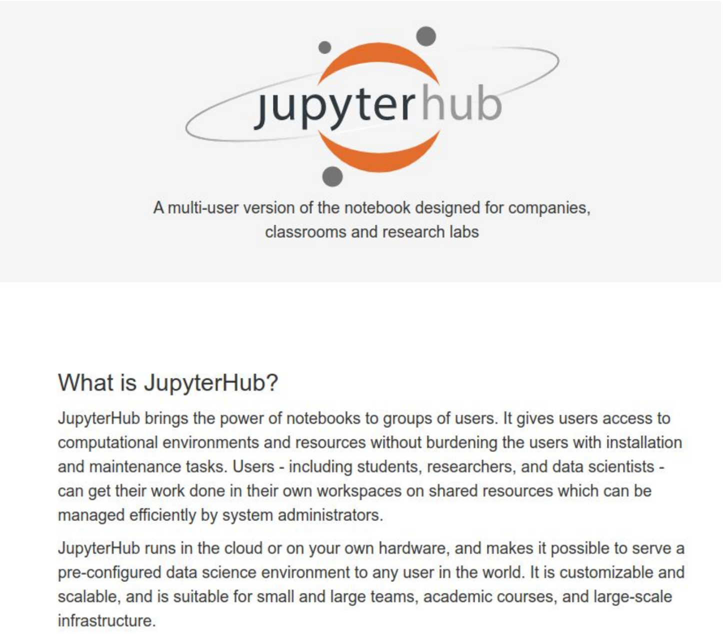
Software: where is going to be run?



jupyterhub

A THING TO GIVE A JUPYTER NOTEBOOK SERVER TO EACH PERSON IN A GROUP OF PEOPLE.

https://geohackweek.github.io/Introductory/05-Jupyter-tutorial/



A multi-user version of the notebook designed for companies, classrooms and research labs

What is JupyterHub?

JupyterHub brings the power of notebooks to groups of users. It gives users access to computational environments and resources without burdening the users with installation and maintenance tasks. Users - including students, researchers, and data scientists - can get their work done in their own workspaces on shared resources which can be managed efficiently by system administrators.

JupyterHub runs in the cloud or on your own hardware, and makes it possible to serve a pre-configured data science environment to any user in the world. It is customizable and scalable, and is suitable for small and large teams, academic courses, and large-scale infrastructure.

Let’s practice a bit!