YData: An Introduction to Data Science

Lecture 29: Correlation

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- Assignment 08 due today at midnight
- Assignment 09 posted on Friday; due Thursday 4/15
- Project 2 due on Friday 4/16
- We've cut back the assignments to reduce the workload
- Two more assignments to go!

High level view

Intro, Cause and Effect	Lectures 1–2
Python, Tables, Visualization	Lectures 3–13
Probability and Distributions	Lectures 14–17
Hypothesis Testing and Causality	Lectures 18–20
Midterm exam	—
Confidence and the Normal Distribution	Lectures 23–28
Regression and Classification	Lectures 29–37
Final exam	—

Prediction

Guessing the Future

- Based on incomplete information
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

(DEMO)

Association

Two Numerical Variables

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible "shape" in the scatter
  - Linear
  - Non-linear

Visualize, then quantify

(DEMO)

Correlation Coefficient

The Correlation Coefficient *r*

- Measures **linear** association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*

(DEMO)

Definition of *r*

Correlation Coefficient (r) =

average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

(DEMO)

Watch Out For ...

- Nonlinearity
- Outliers
- Ecological correlations

(DEMO)