dads 2009|2010 spring

Theme: ***Data Mining for Architecture and Urban Planning***

Lecture IV
Statistics Primer

Ceyhun Burak Akgül, PhD in EE

Ahu Sökmenoğlu, M. Arch.

In this lecture

- Objectives of Statistics
- Measurement and Randomness
- Data Tables
- Histograms
- Centrality and Variability Measures
- Correlation
- …

Objectives of Statistics

- Descriptive statistics
  - Summarize **tangible facts** about a **population**
- Inferential statistics
  - Explain the **machinery** producing a **population**
  - Estimate a **parameter** of the machinery
  - Predict what this machinery would produce under like circumstances

Objectives of Statistics

- **Descriptive statistics**
  - Summarize tangible facts about a population
- Inferential statistics
  - Explain the machinery producing a population
  - Estimate a parameter of the machinery
  - Predict what this machinery would produce under like circumstances

This lecture is mainly about descriptive statistics

Measurement

(…)
How to *measure* a season
against the calendar of your absence?

How to *measure* the stream
of my tangled light
in the mountain
of what has been and will be?
(…)

John Berger

Measurement

- **Properties of measurements**
  - Magnitude
  - Equal intervals
  - Absolute zero
- **Scales of measurement**
  - Ratio scale: weight, height
  - Interval scale: temperature (Celcius, Fahrenheit)
  - Ordinal scale: preferences, ratings
  - Nominal scale: eye color, gender

Measurement

Why do measurements differ?

Randomness

Does God play dice?

Data Tables

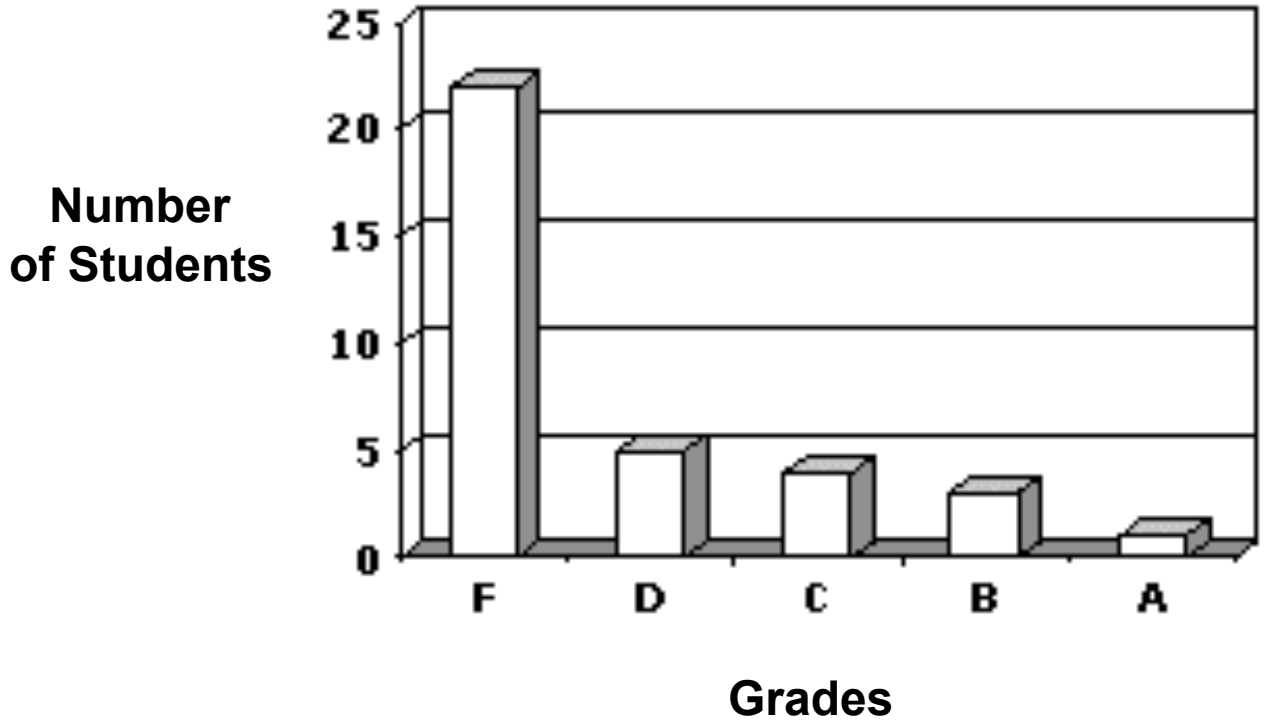
- **Examples**
  - Demographics of student population
  - Financial indicators of a company
  - Climate data
  - …

Data Tables

Let's construct a data table!

	Var 1	Var 2	Var 3	…	Var K
Instance 1	<VALUE>	<VALUE>	<VALUE>	<VALUE>	<VALUE>
Instance 2	<VALUE>	<VALUE>	<VALUE>	<VALUE>	<VALUE>
Instance 3	<VALUE>	<VALUE>	<VALUE>	<VALUE>	<VALUE>
…	<VALUE>	<VALUE>	<VALUE>	<VALUE>	<VALUE>
Instance N	<VALUE>	<VALUE>	<VALUE>	<VALUE>	<VALUE>

Histograms – 1/3



Histograms – 2/3

- What can you do with a histogram?
  - Did the population succeed in general
  - Percentage of people who got an A
  - Percentage of people who got a C or higher
  - …

Histograms – 3/3

- Conditional histograms
  - Consider an additional variable that adds to the description of the population (gender, age, …)
  - You can identify a subgroup w.r.t. to the additional variable and construct the histogram out of the subgroup
  - This is a conditional histogram

Centrality Measures* – 1/2

- Mean
- Median
- Mode

** See whiteboard*

Centrality Measures – 2/2

Mean, Median, Mode: Which one to use?

- Mean
  - stable measure
  - descriptive for symmetric data
  - ratio or interval scale
- Median
  - suitable when the histogram is skewed of there are outliers
  - Ratio, interval, or ordinal data
- Mode
  - You have no other choice for nominal data

Variability Measures* – 1/2

- Range
- Interquartile range
- Variance
- Standard deviation

** See whiteboard*

Variability Measures – 2/2

Range, IQR, Variance, Std. Dev.: Which one to use?

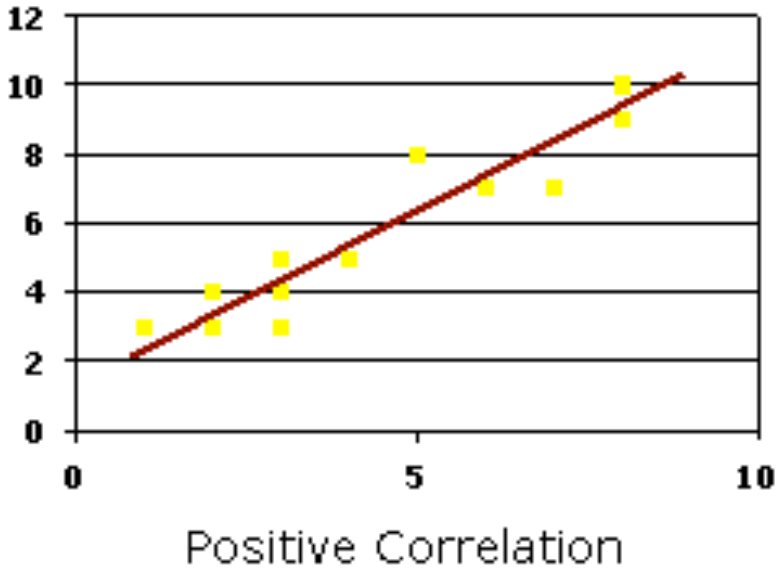
- Range
  - Sensitive to outliers
- Interquartile range (IQR)
  - Good option when data is skewed
- Variance
  - Use standard deviation instead
- Standard Deviation (Std. Dev.)
  - Stable
  - Good option when data is symmetric

Correlation – 1/4

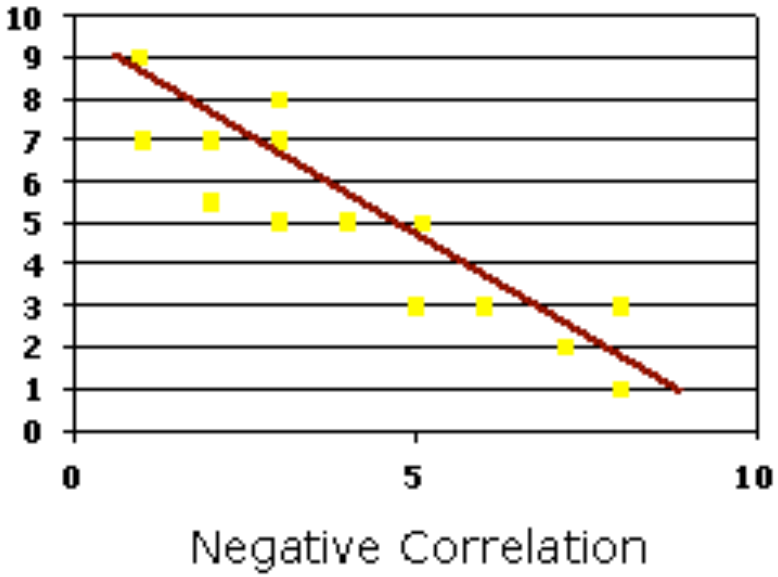
Correlation is one of the many possible ways to quantify how a (random) quantity vary w.r.t. another.

Correlation – 2/4

Correlation is one of the many possible ways to quantify how a (random) quantity vary w.r.t. another.



Positive Correlation



Negative Correlation

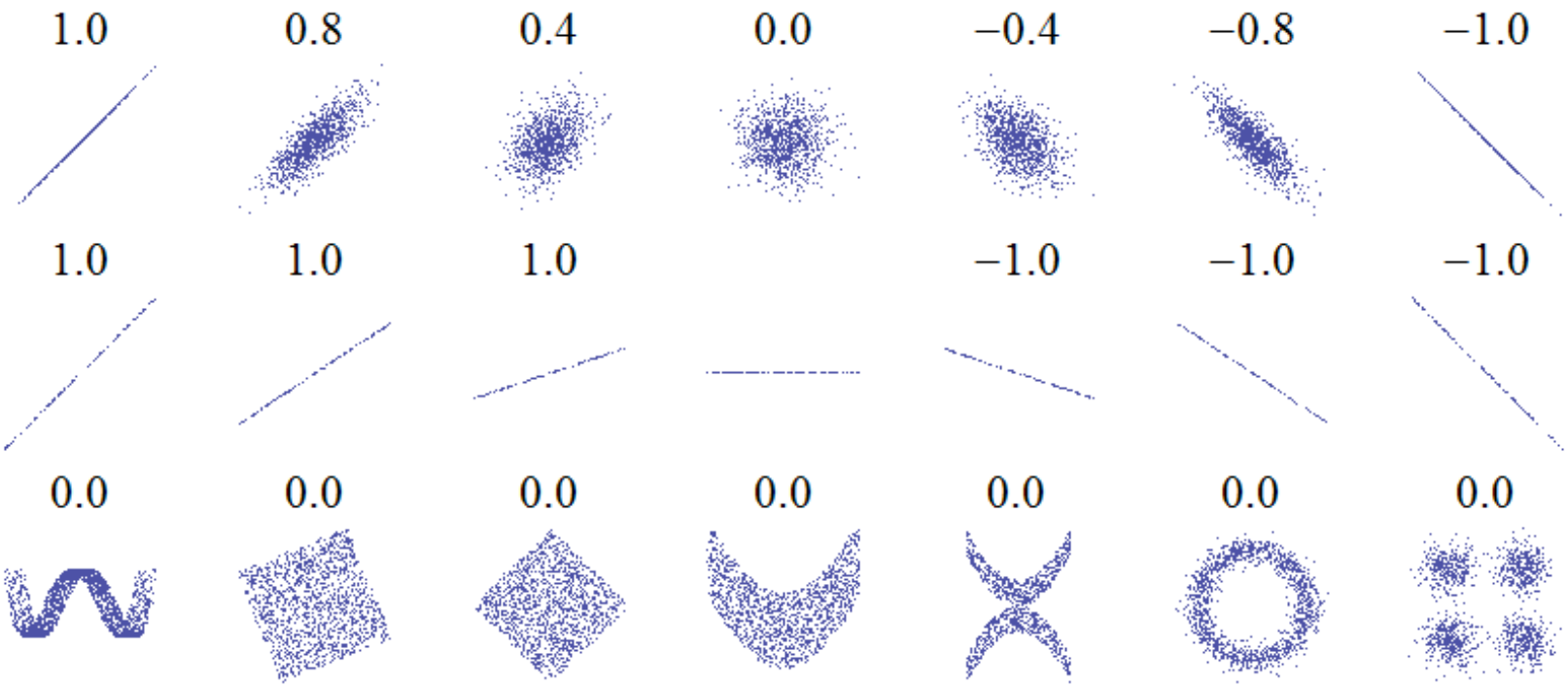
Correlation – 3/4

Correlation is one of the many possible ways to quantify how a (random) quantity vary w.r.t. another.

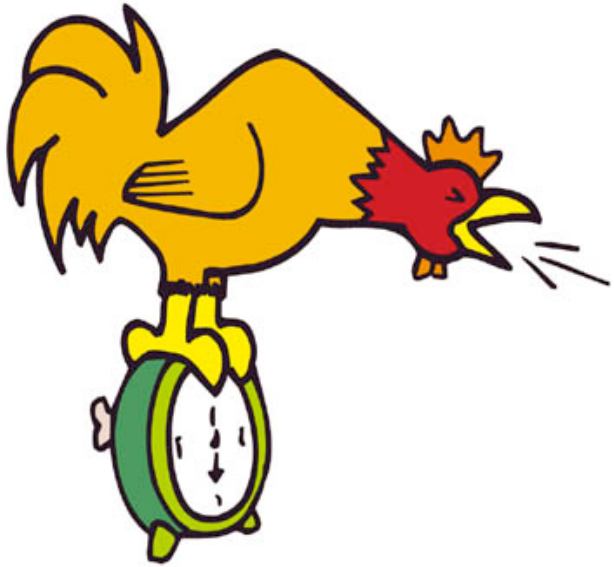
How to compute correlation?

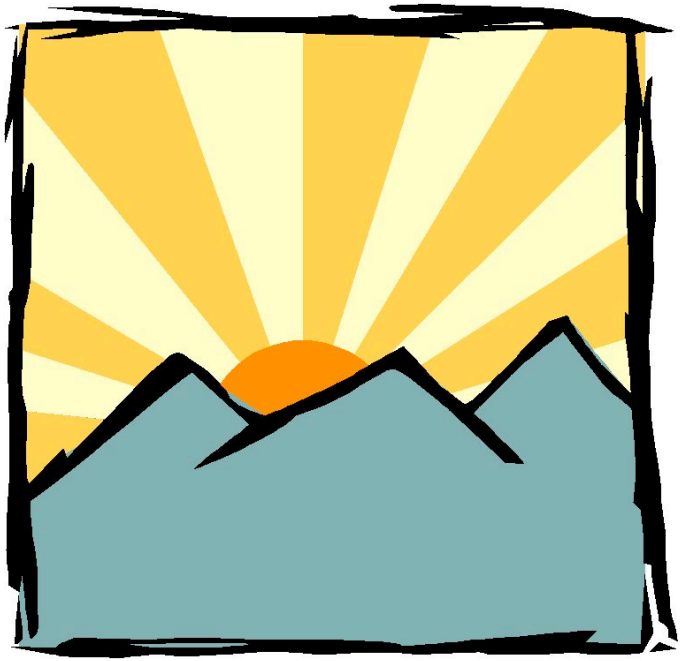
Pearson's correlation coefficient – *see whiteboard.*

Correlation – 4/4



Correlation does not imply causality!





Post hoc ergo propter hoc?

Where are we?

week	date	studio
1	**9-Feb**	**-**
2	**16-Feb**	**Introduction: Data Mining in General**
3	**23-Feb**	**Concepts in Data Mining**
4	**2-Mar**	**Data Mining Applications in Context**
		Introduction to Semester Project
5	**9-Mar**	**Statistics Primer**
6	16-Mar	A Broad Picture of Data Mining Tools
		Jury Meeting; Semester Project's first concepts & ideas
7	23-Mar	Regression and Classification
8	30-Mar	Clustering, Exploratory Data Analysis, and Visualization
		Semester Project's review
9	6-Apr	Semester Project's review
10	13-Apr	Semester Project's review
11	20-Apr	Jury Meeting; Presentations
12	27-Apr	Semester Project's review
13	4-May	Semester Project's review
14	11-May	Jury Meeting; Final Presentations

Let's talk about the last week's assignment

Think of the "City" as a concept:

- Designate a set of attributes related to the city
- Instantiate the "city" concept with several examples
- Specify the attributes of your "city" examples

What kind of knowledge descriptions can you extract with your chosen set of attributes?

Do the reverse*