Applied Data Analytics

Statistics — Basics & location

What do we mean by "data"?

Hans-Martin von Gaudecker and Aapo Stenhammar

Computer Science Definition

A collection of values or information that can be processed by a computer.

- Unstructured data (images, videos, text documents, ...)
- Structured data (commonly: tables with rows and columns)

Statistical Definition

The actual values of variables obtained from samples or populations. These values can be numerical or categorical.

- Sample vs. Population
- Numerical vs. Categorical

What could freight data look like?

- Notation often $x_{k,i}$ where k indexes the variable and i indexes the observation.
- So i could be the name of the vessel
- $x_{1,i}$ could be the owner of the vessel $\in \{\text{Mærsk}, \text{MSC}, \text{CMA CGM}, \ldots\}$:
- $x_{2,i}$ could be the number of containers on the vessel

Tables

ship_name	owner	n_containers
Laura Mærsk	Mærsk	1,926
MSC Flaminia	MSC	2,356

a.k.a. labelled arrays, labelled matrices, Pandas DataFrames

Table columns

ship_name	owner
Laura Mærsk	Mærsk
MSC Flaminia	MSC

ship_name	n_containers
Laura Mærsk	1,926
MSC Flaminia	2,356

a.k.a. labelled vectors, Pandas Series

Only one index to access elements — x_i