Analysis and Visualisation of Complex Agro-Environmental Data

Lesson 03

- Introduction to data analysis
- Descriptive statistics and univariate visualizations
- Visualization modules in Python
- Working examples and exercise

U LISBOA | UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR DE AGRONOMIA

1. **Introduction to data analysis**
   - Data analysis vs. Data Science
   - Introduction to statistics
2. **Descriptive statistics and univariate visualizations**
3. **Visualization modules in Python**
4. **Working examples and exercise**

1. Introduction to data analysis

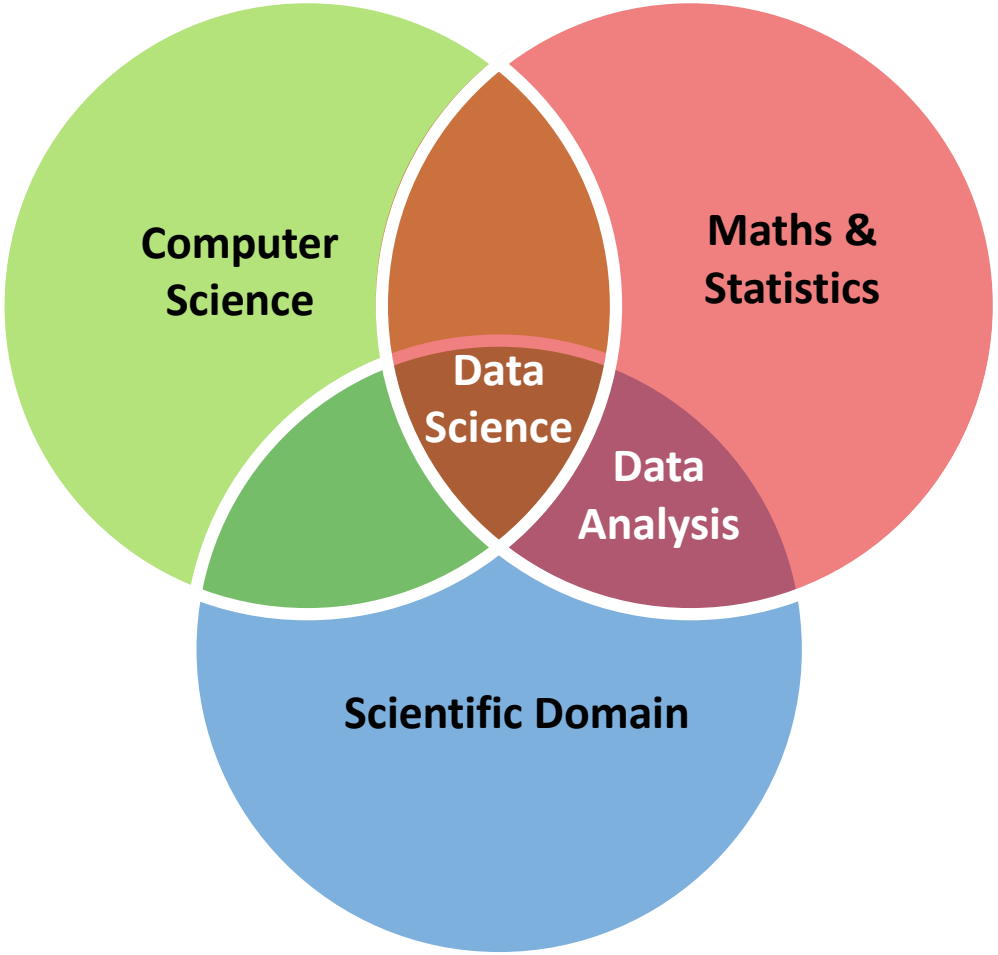
- Data analysis vs. Data Science
- Introduction to statistics

2. Descriptive statistics and univariate visualizations

3. Visualization modules in Python

4. Working examples and exercise

Data Analysis *vs*. Data Science



https://www.springboard.com/blog/data-science/data-analyst-vs-data-scientist/

Data Analysis *vs*. Data Science

- **Data scientists** – More programming and math skills and applications in machine learning

- **Data Analysts** – Better communication and reporting skills, with stronger storytelling abilities

… but no consensus on these differences!

Data Analysis *vs*. Data Science

Data science pipeline

This course

- SQL
- Scrapping
- File Formats
  - CSV
  - JSON
  - XML
- Consulting APIs
- Buying Data
- Distributed Databases

- Missing values and empty data
- Data imputation
- Incorrect types
- Incorrect or invalid values
- Outliers and non relevant data
- Statistical sanitization

- Hierarchical Data
- Handling categorical data
- Reshaping and transforming structures
- Indexing data for quick access
- Merging, combining and joining data

- Exploration
- Building statistical models
- Visualization and representations
- Correlation vs Causation analysis
- Hypothesis testing
- Statistical analysis
- Reporting

- Building Machine Learning Models
- Feature Engineering
- Moving ML into production
- Building ETL pipelines
- Live dashboard and reporting
- Decision making and real-life tests

Source: https://www.youtube.com/watch?v=r-uOLxNrNk8

Data Analysis *vs*. Data Science

Three broad areas of data analysis:

- **Descriptive Statistics** (DS) - focuses on summarizing data through numbers and graphs (e.g. central tendency, dispersion, distribution).
- **Exploratory Data Analysis** (EDA) - focuses on discovering new features in the data (e.g. data visualization; ordination, classification; find ideas for a theory).
- **Confirmatory data analysis** (CDA) – focuses on confirming or falsifying existing hypotheses (e.g. hypothesis testing; test theories that result from EDA).

Data analysis => statistics

1. Introduction to data analysis

- Data analysis vs. Data Science
- Introduction to statistics

2. Descriptive statistics and univariate visualizations

3. Visualization modules in Python

4. Working examples and exercise

What is Statistics?

Branch of mathematics dealing with the **collection**, **analysis**, **interpretation**, and **presentation** of data.

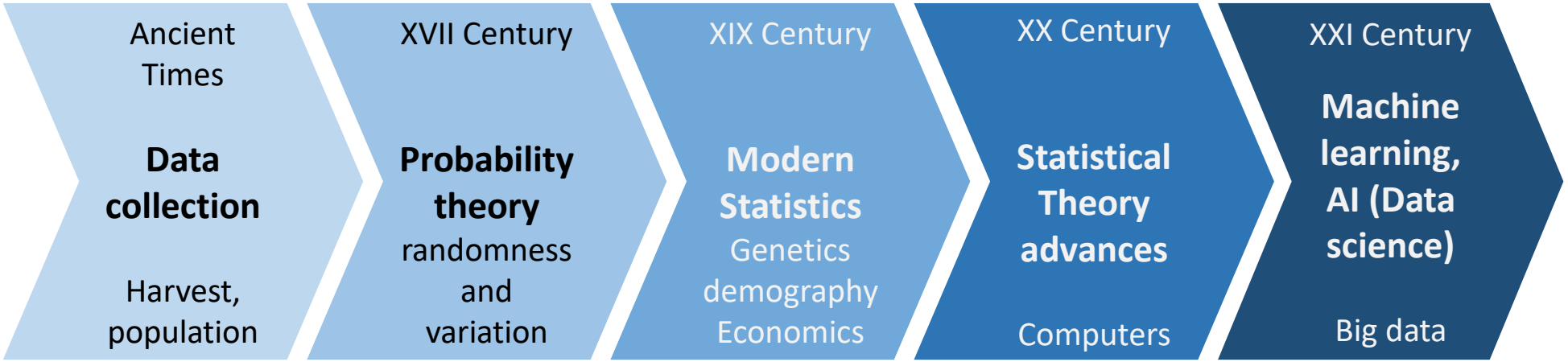
A field in constant evolution:

- New fields of application
- New types of data (e.g. new sensors)
- Advances in computing:
  - ✓ storage capacity (big data)
  - ✓ computing power

→ New analytic methods

What is Statistics?

Statistics Milestones



Common challenge: how to collect and understand **data**!

Which statistics?

3 important questions to adopt the right statistical approach to our problem:

1. Where do data come from?

2. Are the data independent and identically distributed (i.i.d)?

3. What is the type of research study we are conducting?

Question 1: data origin

Where do data come from?

Two main data origins:

- **“Designed” Data Collection** - Data from studies designed to address a particular research objective (involve sampling individuals from populations or manipulative experiments).
- **Organic / Process data** – generated by a computerized information system or a variety of sensors – usually “Big data” and generated over time => requiring significant computational resources to turn data into formats that are suitable for analysis.

Question 1: data origin

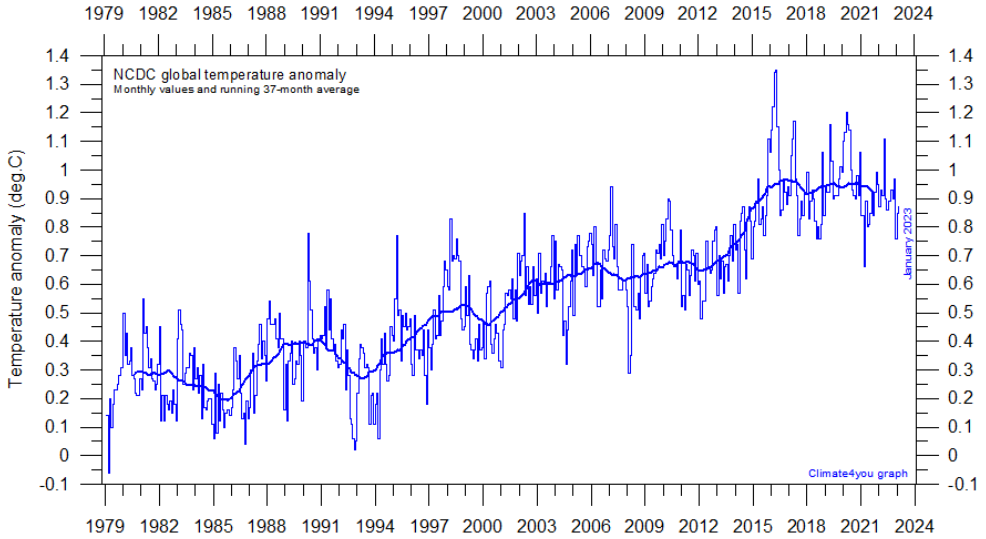
Designed Data Collection – examples

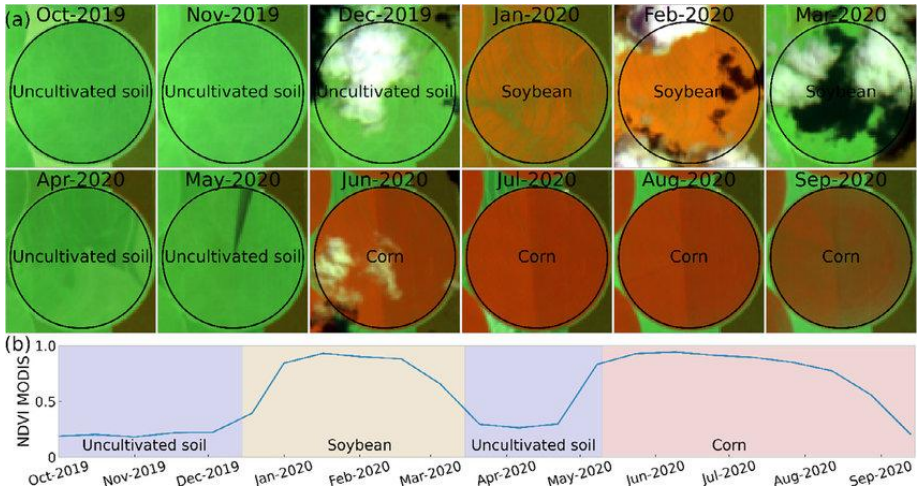
- Data from agriculture field trials
- Data from animal or plant field samplings
- Opinion surveys

Question 1: data origin

Organic / Process data – examples

- Long term environmental data (climate, pollutants) derived from sensors
- Satellite imagery time series.
- Web browser activity





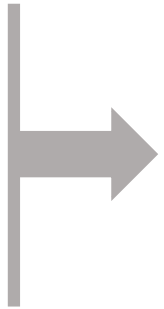
Question 2: statistical properties of data

Are the data i.i.d?

- **i = independent** - observations are independent from all the other observations?
- **id = identically distributed** - do the values observed arise from a common statistical distribution?

Examples of non-i.i.d data:

- Closer sites tend to show similar environmental conditions;
- Individuals of the same sex tend to have more similar biometric measurements



Need specific statistical approaches

Question 3: types of research studies

- **Exploratory** *versus* **Confirmatory**
- **Comparative** *versus* **Non-comparative**
- **Observational** *versus* **Experiments**

Question 3: types of research studies

Exploratory ***VS.*** ***Confirmatory***

- Collect and analyse data without first pre-specifying question
- Visualizations, multivariate methods
- Focuses on type II error (or false negative) – find trends from noise.
- Often involve ***process data***.

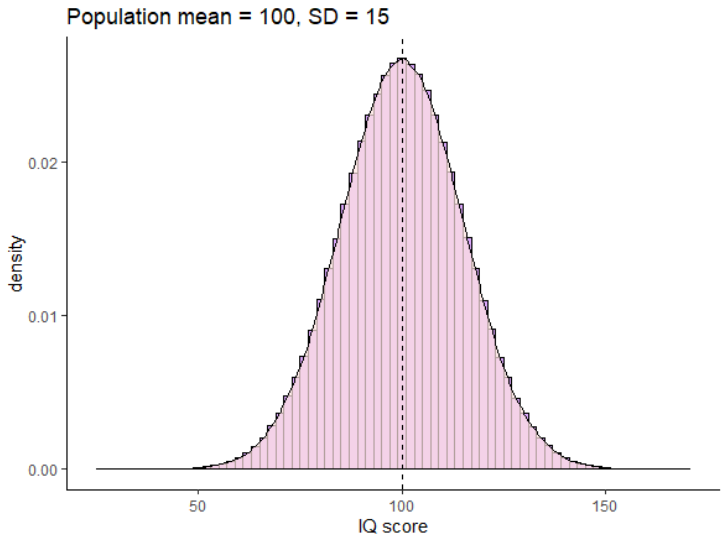
- Collect and analyse data with pre-defined questions in mind.
- Test *a priori* falsifiable hypothesis
- Hypothesis testing methods
- Focuses on type I error (or false positive) – prove trends.
- Often involve ***designed data***

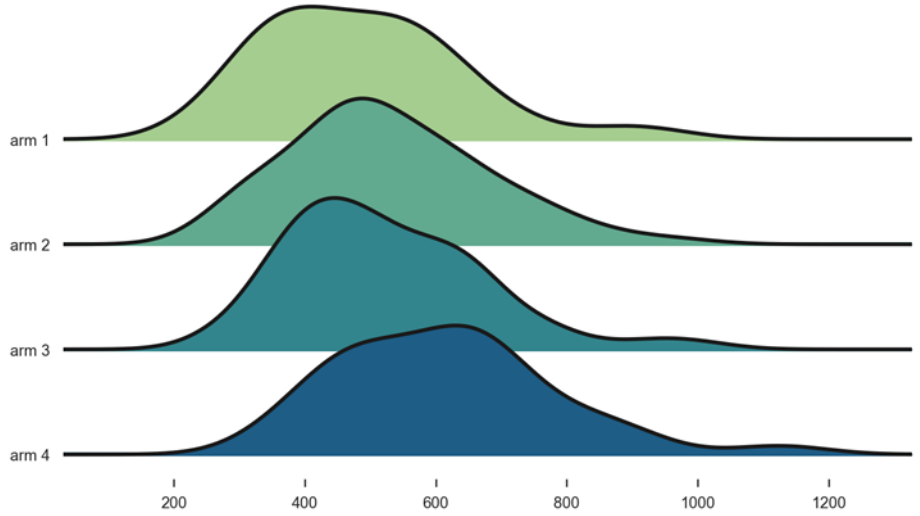
Question 3: types of research studies

Non-comparative - estimating or predicting single parameters (e.g. population size).

VS.

Comparative – compare observations (e.g. harvest production) from different "strata" or "treatments" (e.g. fertilizers, year, region).



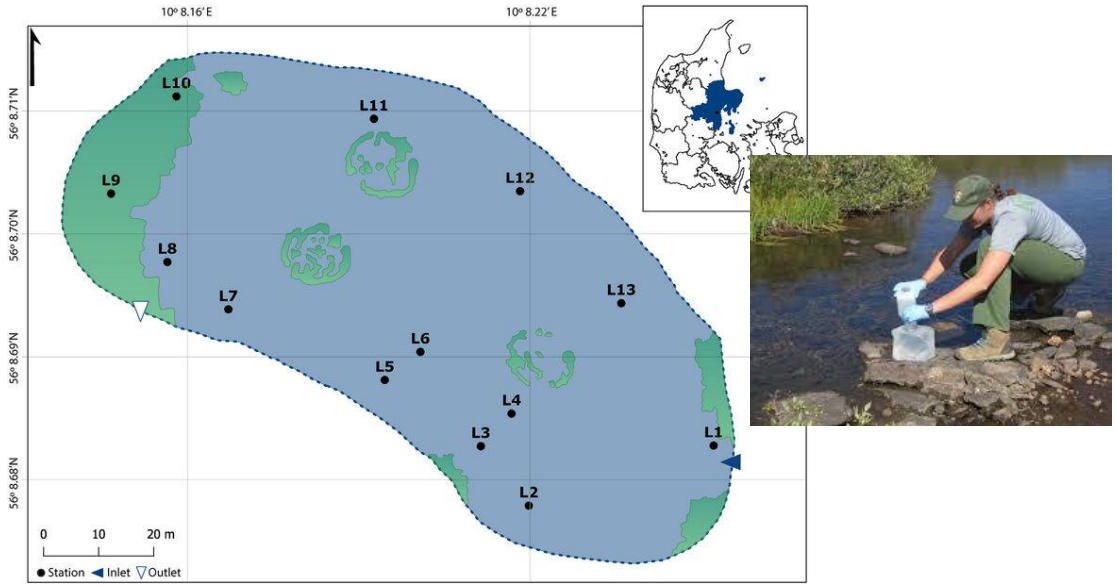


Question 3: types of research studies

Observational – based on empirical observations (across natural gradients).

VS.

Experiments – involve manipulation at different units in different ways (treatments).



1. Introduction to data analysis
   - Data analysis vs. Data Science
   - Introduction to statistics

2. Descriptive statistics and univariate visualizations

3. Visualization modules in Python

4. Working examples and exercise

Descriptive statistics

- Summary statistics (categorical and quantitative data)
- Statistical distributions (continuous or discrete data)
- Based on univariate analyses

Summary Statistics

Categorical variables

Summary statistics:

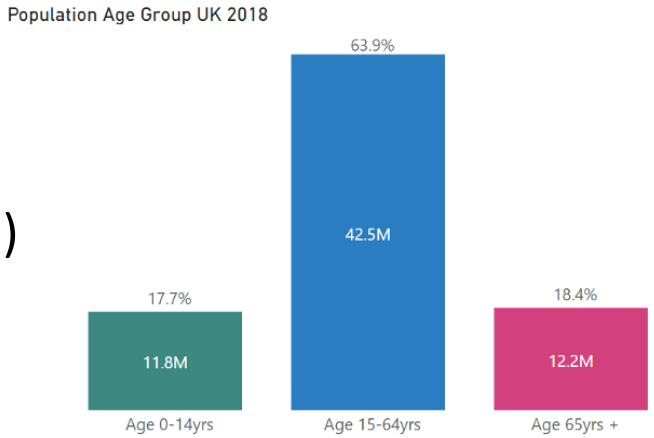
- Data type
- Number of classes or levels
- Measures of central tendency: Mode (most frequent class or level)
- Frequency table (number or proportion of cases per class)

Summary Statistics

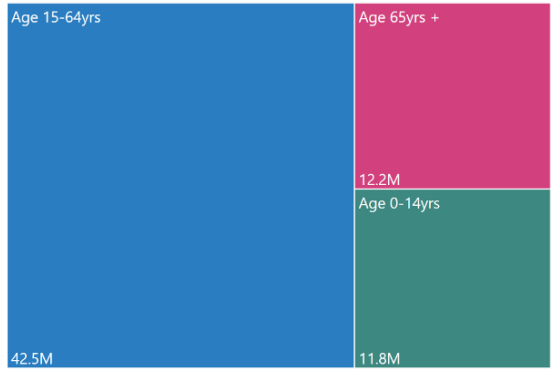
Categorical variables

Common visualization types:

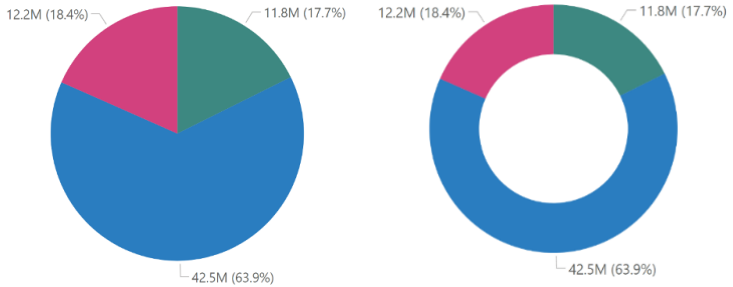
Barplots (unstacked or stacked)



Tree maps



Pie or donut charts



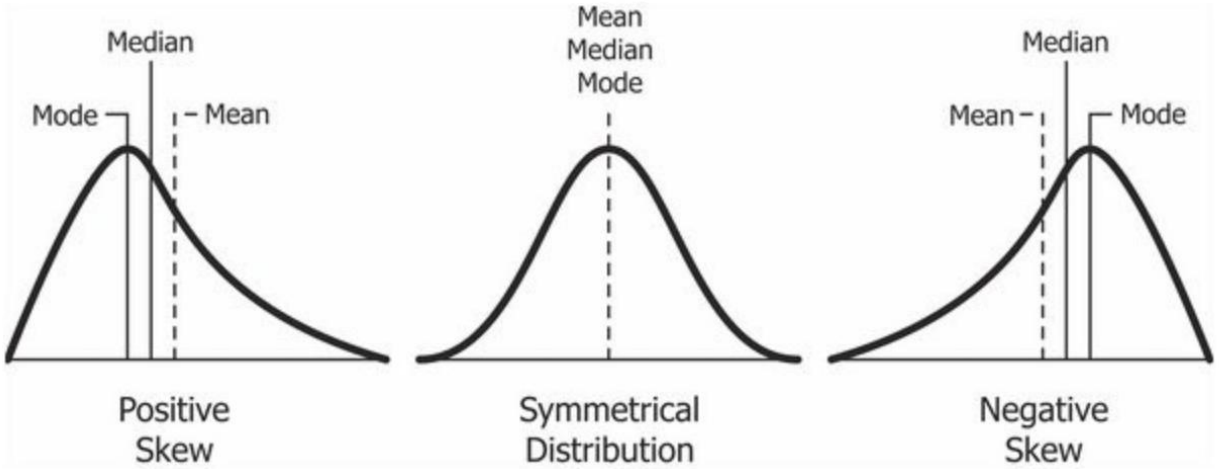
Waffle charts

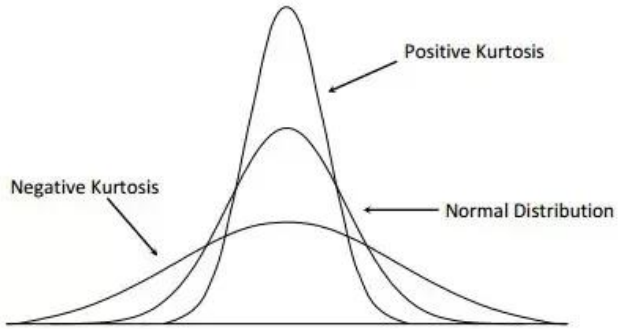
Summary Statistics

Quantitative variables

Summary statistics:

- **Data type** (discrete or continuous)
- Measures of **location** and **central tendency**:
  - ✓ Mean/average
  - ✓ Mode (only for discrete quantitative data)
  - ✓ Median
  - ✓ Percentiles/quantiles/quartiles
- Measures of **dispersion**
  - ✓ Standard deviation/variance
  - ✓ Skewness
  - ✓ Kurtosis





Statistical distributions

Mathematical function that gives the probabilities of the different possible outcomes for an experiment.

Some common distributions:

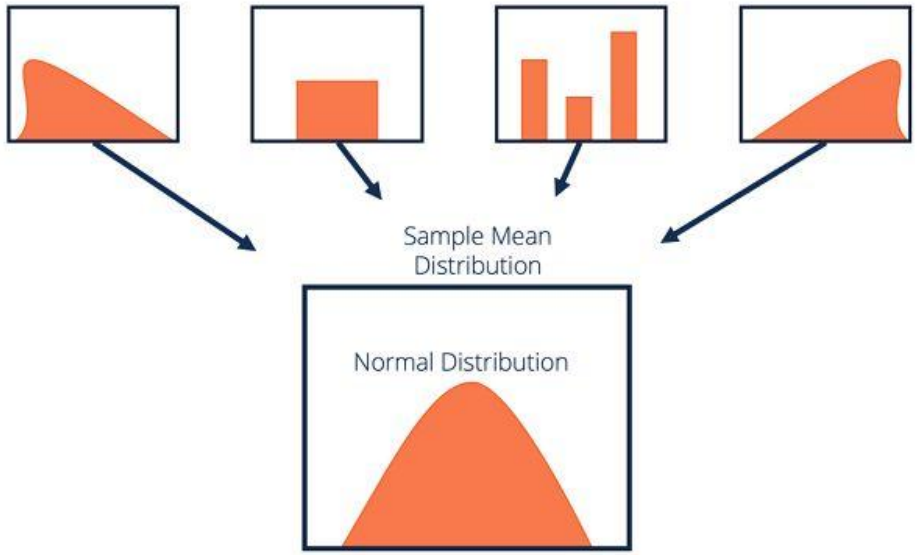
- **Uniform distribution** (discrete or continuous) – Equal probability outcomes (ex. dice roll outcome)
- **Binomial distribution** – deals with binary outcomes (ex. flipping a coin multiple times and counting heads - or tails)
- **Poisson distribution** – deals with the frequency with which an event occurs within a specific interval (ex. animal counts on a fixed time interval).
- **Exponential distribution** – often concerned with the amount of time until some specific event occurs (ex. the amount of time a cell phone battery lasts).
- **Normal distribution** – symmetric, unimodal, and asymptotic, with equal mean, median, and mode. It approximates many natural phenomena (ex. people's height).

Statistical distributions

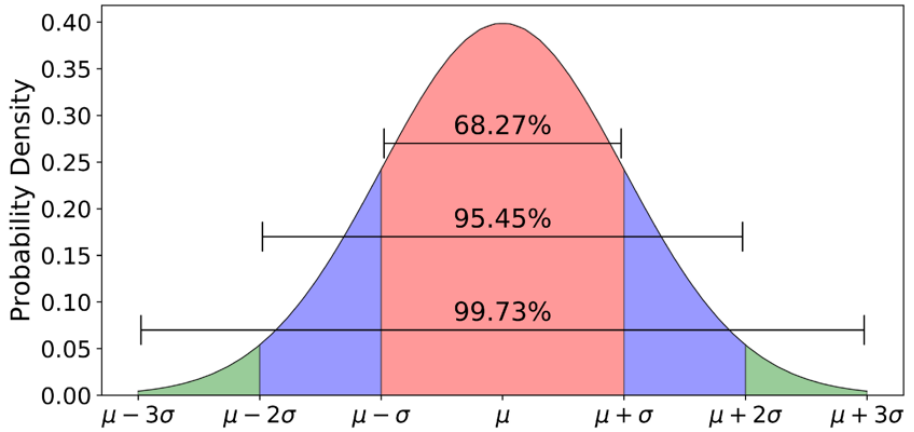
More on Normal Distributions

Central Limit Theorem:

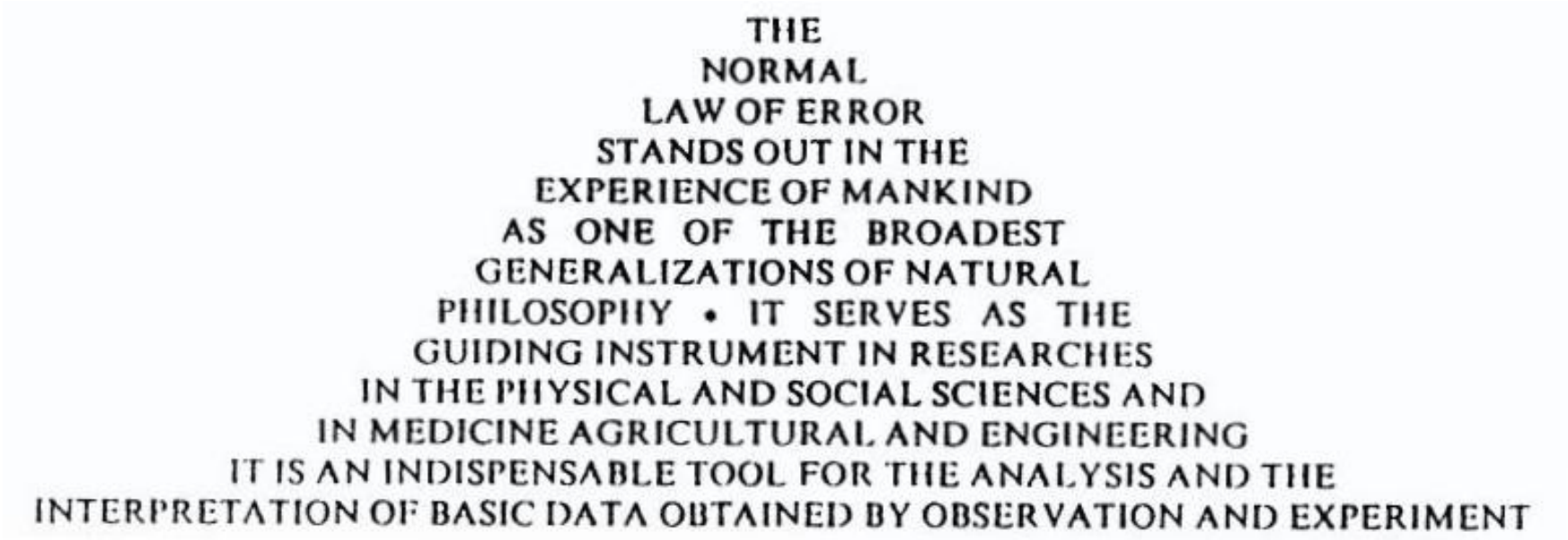
If random samples with replacement are taken from a population + i.i.d. observations => the distribution of the sample means will be approximately normally distributed, even if the population is not normally distributed.



The 68-95-99.7 Rule:



Statistical distributions



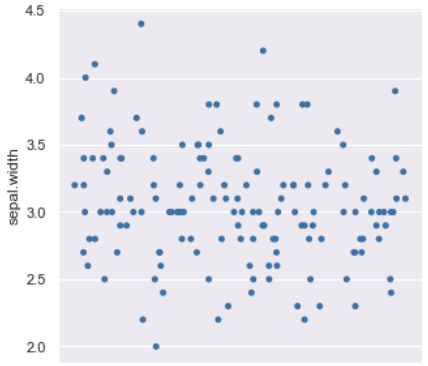
W. J. Youden

Descriptive Statistics – Univariate data

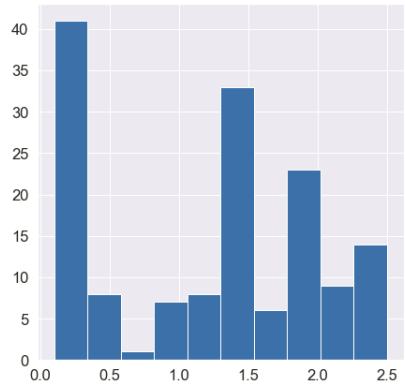
Quantitative variables

Common visualization types:

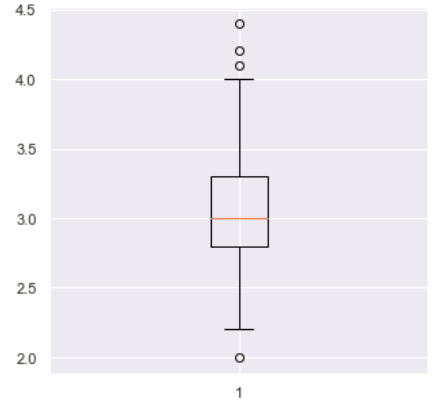
Strip plot



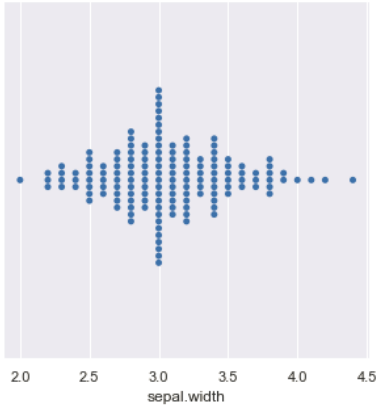
Histogram



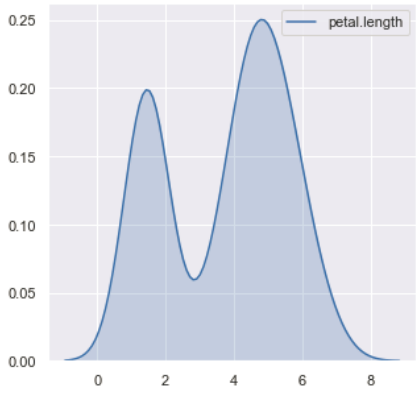
Boxplot



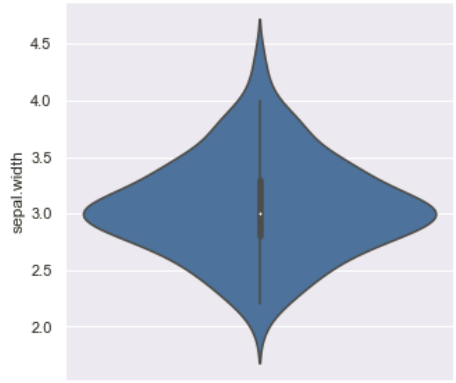
Swarm plot



Density plot

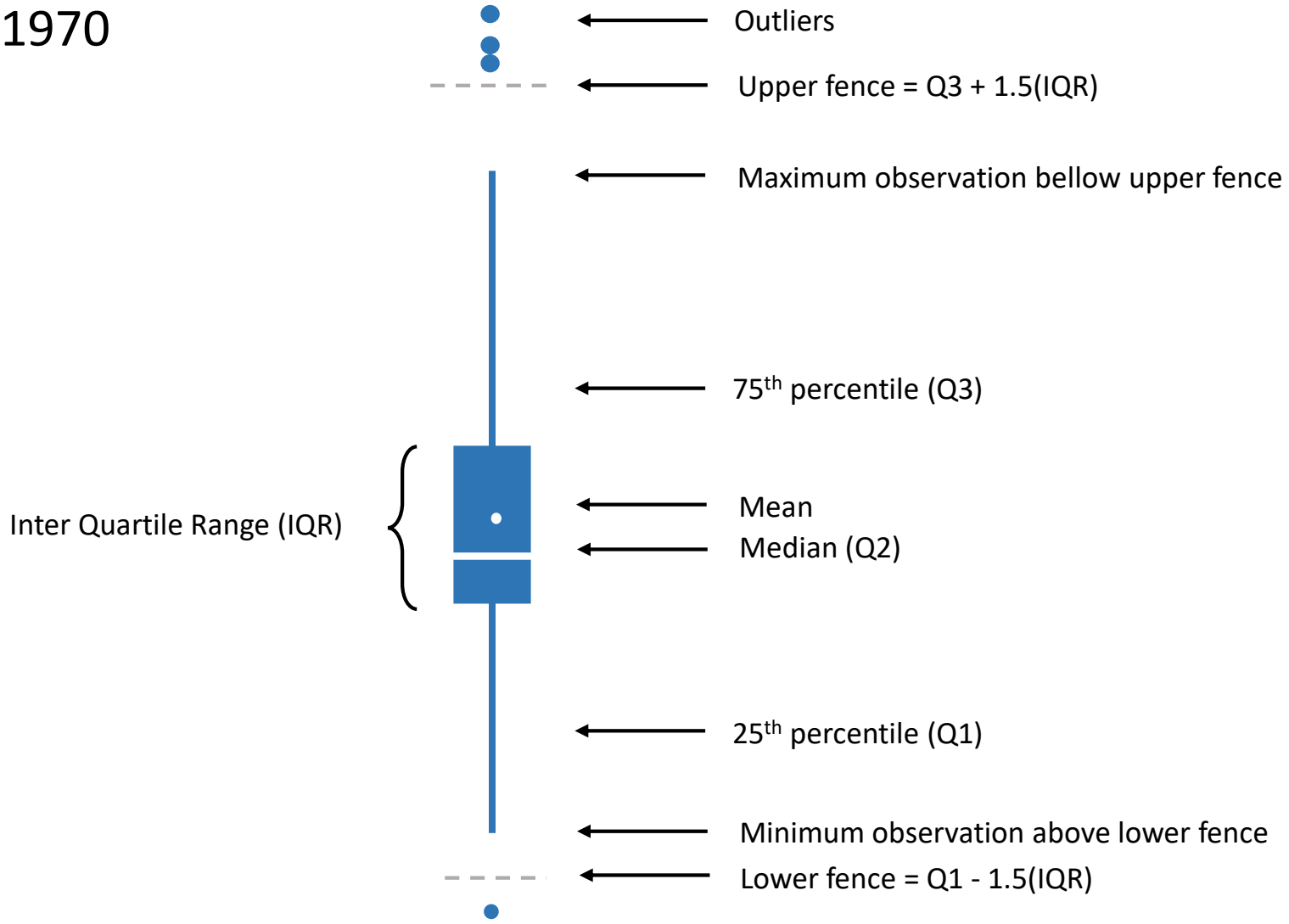


Violin plot



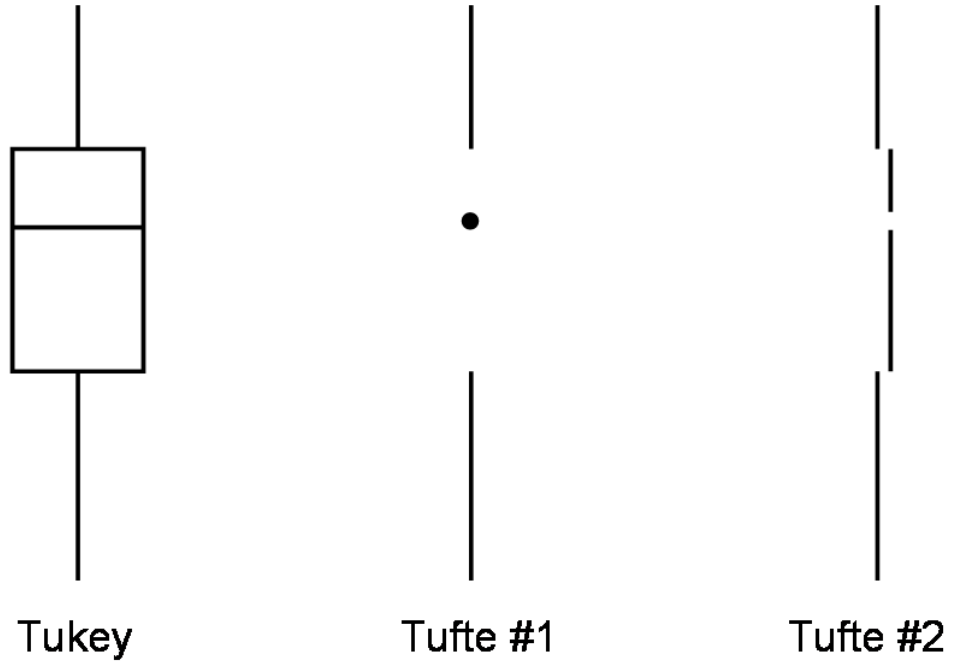
Representing distributions - boxplot

John Tukey, 1970



Representing distributions - boxplot

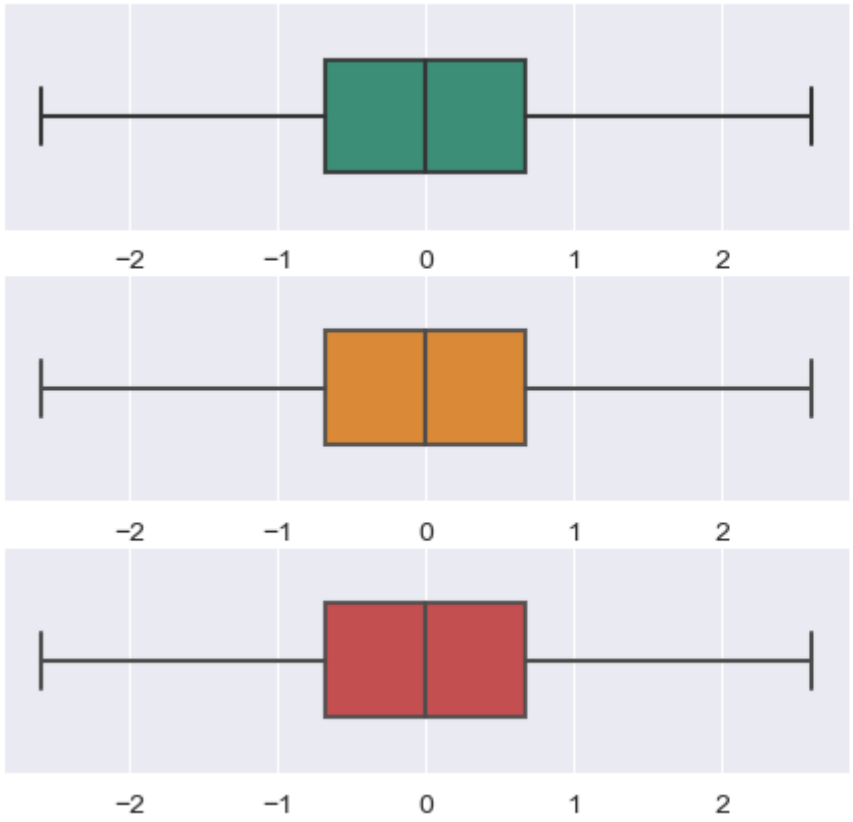
Maximize data-ink

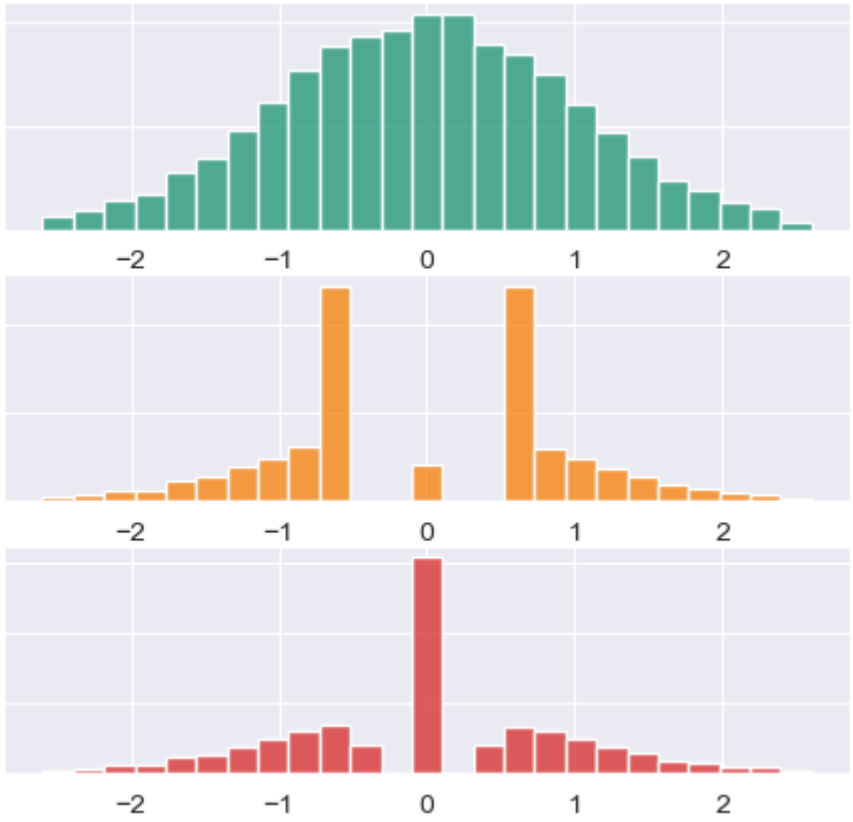


Comment!

Representing distributions - boxplot

Drawbacks (and again, some benefits of visualizations)

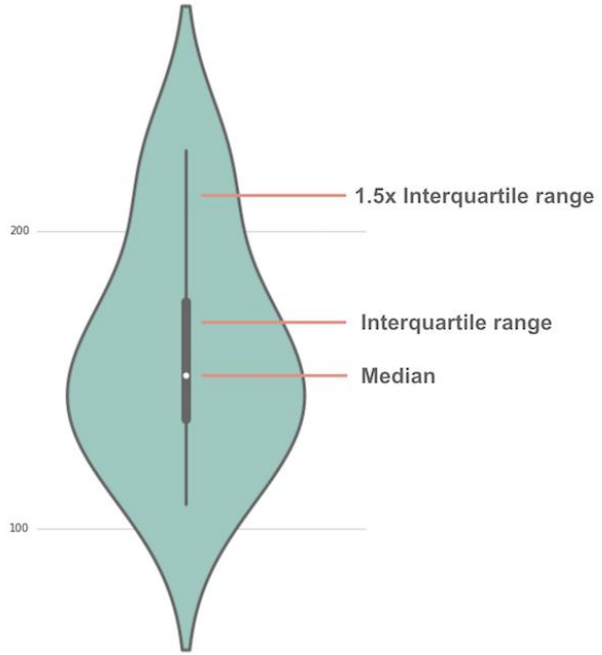




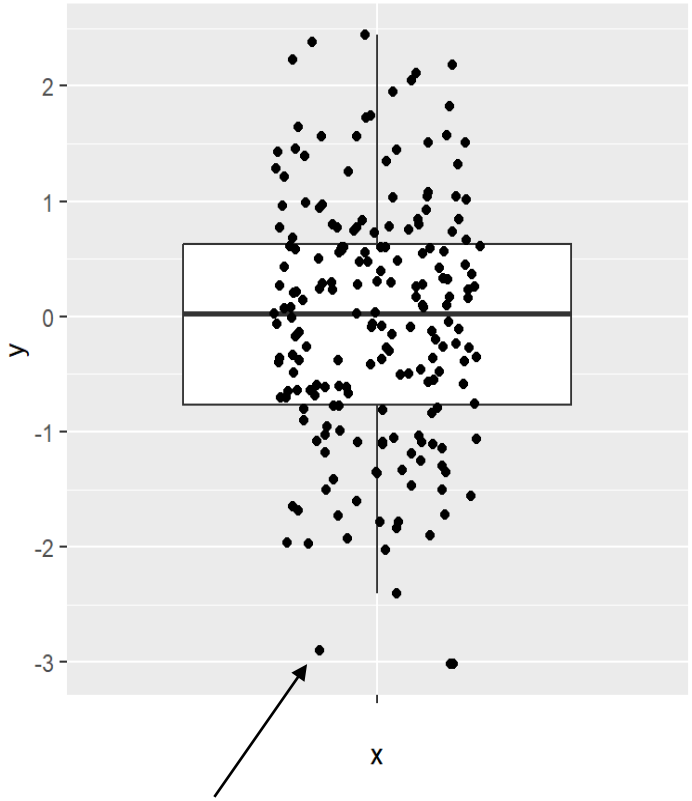
Representing distributions - boxplot

Alternative versions

Violin plots:

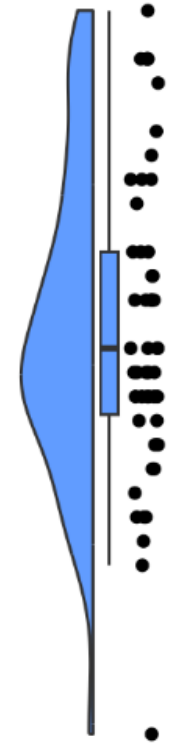


Jittered boxplots:



Jitter plot: randomly assigned x-values within a given strip width

Rain cloud plots:



1. Introduction to data analysis
   - Data analysis vs. Data Science
   - Introduction to statistics
2. Descriptive statistics and univariate visualizations
3. **Visualization modules in Python**
4. Working examples and exercise

Visualization modules in python

Library	Interactive Features	Syntax	Main Strength and Use Case
Matplotlib	Limited	Low-level	Highly customized plots
Seaborn	Limited (via Matplotlib)	High-level	Fast, presentable reports
Bokeh	Yes	High- and low-level, influenced by grammar of graphics	Interactive visualization of big data sets
Altair	Yes	High level, declarative, follows grammar of graphics	Data exploration, and interactive reports
Plotly	Yes	High- and low-level	Commercial applications and dashboards

https://learnpython.com/blog/python-libraries-data-visualization/

Visualization modules in python

Main advantages:

- Widely used in the Python scientific computing community
- Comprehensive, versatile, accessible and customizable
- Good documentation (https://matplotlib.org/)
- Universal tool that plugs into many back ends
- Many popular modules use it as a dependency (e.g. Seaborn, Pandas, SciPy, Statsmodels…)

Main disadvantages:

- Steep learning curve
- Users need to have good Python skills
- Users need to understand the syntax of Matplotlib (based on MATLAB).

Visualization modules in python

Matplotlib architecture - 3 layers:

Scripting layer

Offers the users some easy-to-use methods for generating plots in only a few lines of code (submodule **Pyplot**).

Artist layer

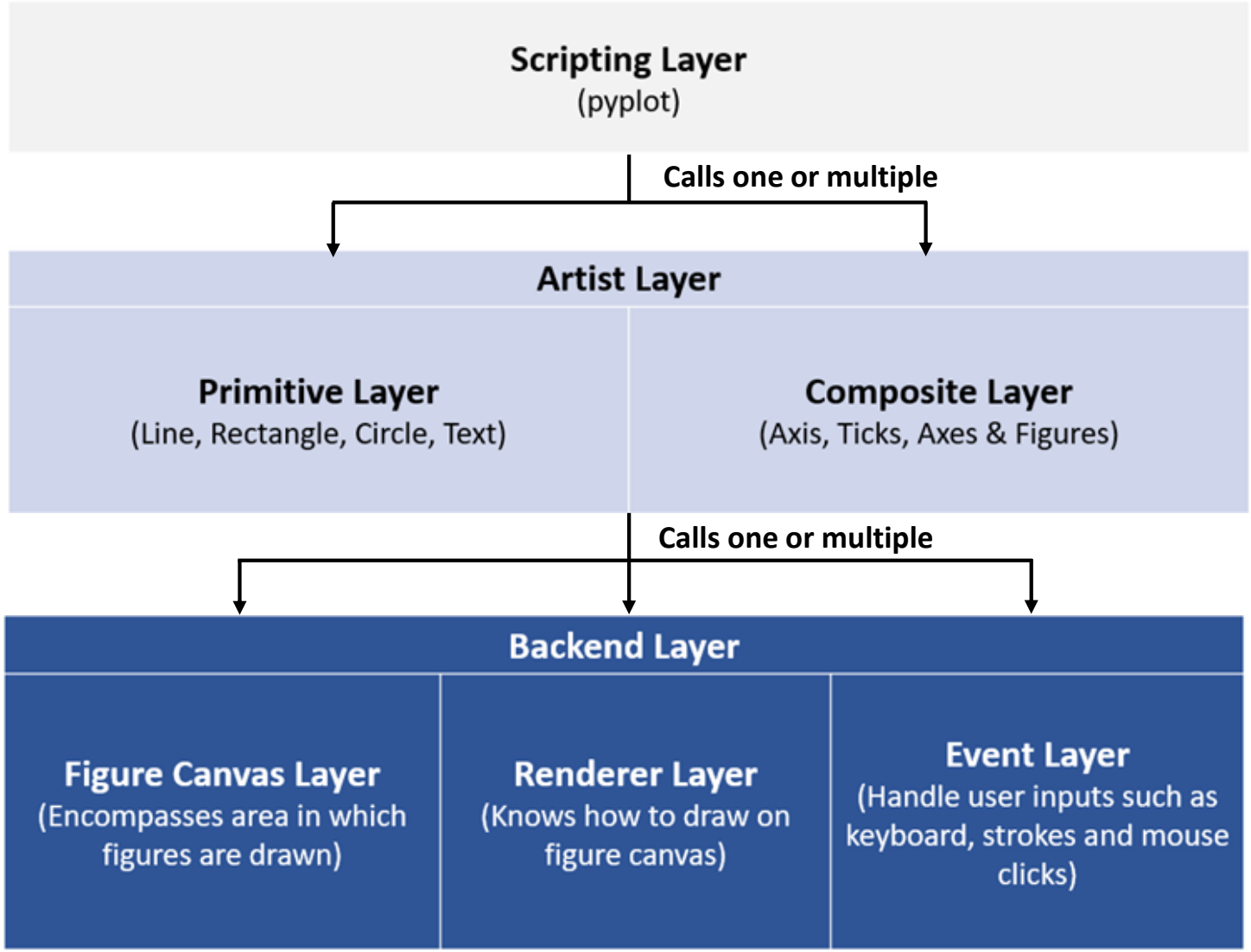
Includes every small piece of a visual component, from the lines to the labels, allowing users to fully customize their results.

Backend layer

The most complex and low-level layer, often not used by users.

Visualization modules in python

Matplotlib architecture



Visualization modules in python

Further readings:

https://www.aosabook.org/en/matplotlib.html

https://www.datacamp.com/tutorial/matplotlib-tutorial-python

Visualization modules in python

Main advantages:

- It builds on top of matplotlib making it more user-friendly
- Integrates closely with pandas data structures
- Very good documentation (https://seaborn.pydata.org/)
- It provides very attractive default statistical plots with much less effort than with matplotlib

Main disadvantages:

- It can be slow and memory-intensive for large or complex datasets,
- Less customizable

Further reading:

Visualization modules in python

Further readings:

https://seaborn.pydata.org/tutorial/introduction.html

https://www.geeksforgeeks.org/introduction-to-seaborn-python/

1. Introduction to data analysis
   - Data analysis vs. Data Science
   - Introduction to statistics
2. Descriptive statistics and univariate visualizations
3. Visualization modules in Python
4. **Working examples and exercise**

Working examples and exercise

https://github.com/isa-ulisboa/greends-avcad-2025/tree/main/examples

Univariate_distributions.ipynb

Pandas.ipynb

Case study 1: EFI+ Project Database (mediterranean subset)

https://github.com/isa-ulisboa/greends-avcad-2025/blob/main/examples/CaseStudy1_univariate_analysis_visuals.ipynb

- ***Rows***: river fish sampling sites in western mediterranean catchments of Europe;
- ***Columns***: 3 types of variables: (1) natural environmental variables; (2) anthropogenic pressure variables; (3) fish occurrences.
- **Main purpose**: relate river human disturbance with fish-based parameters (occurrence, richness, community composition) – find fish-based indicators of river disturbance.

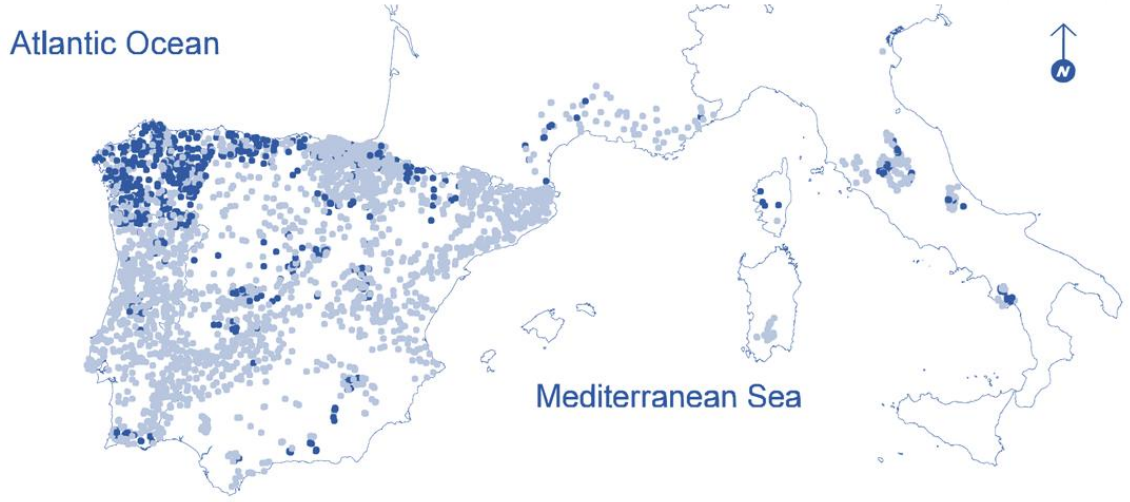
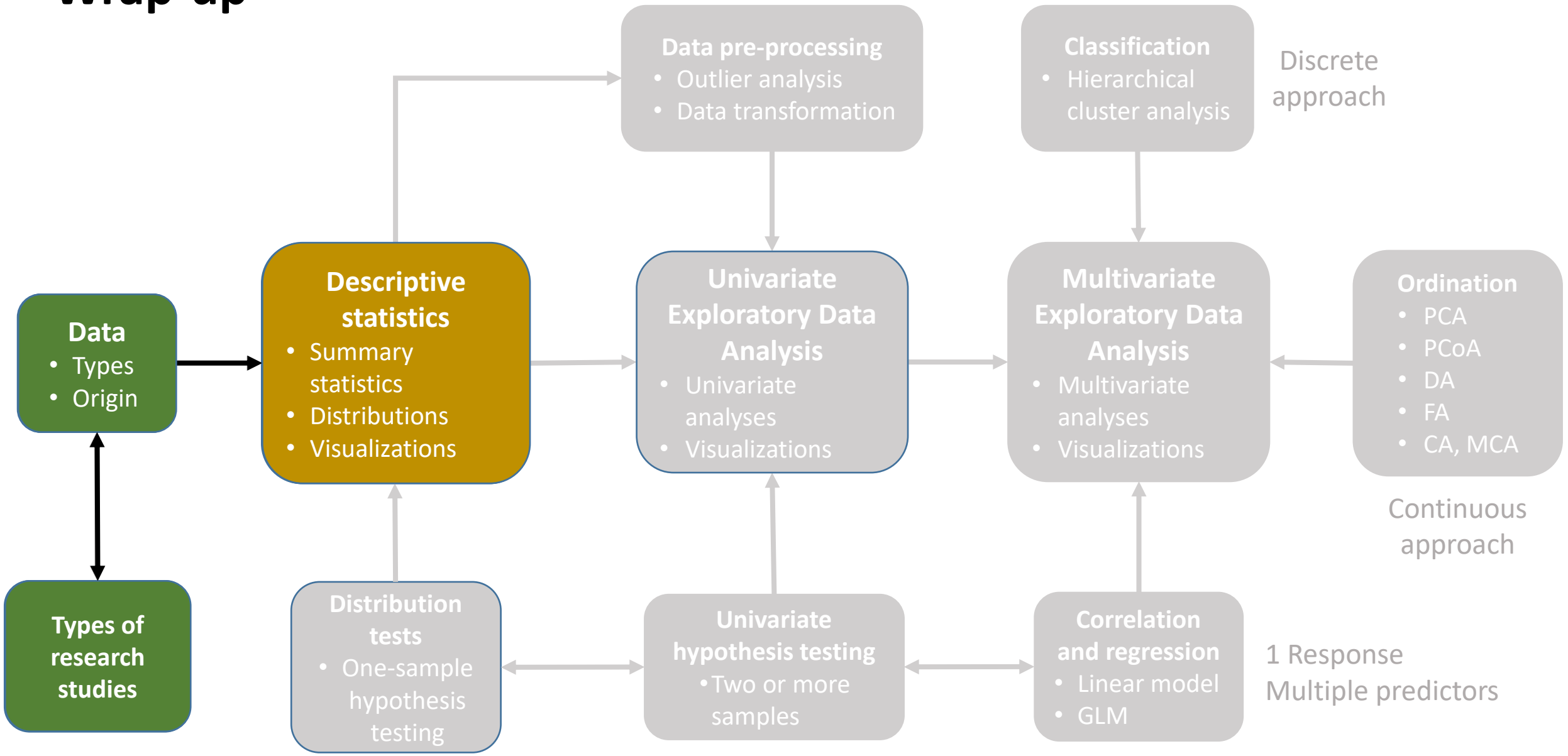


Fig. 2. Locations of sample sites (black dots are least-disturbed sites used for model calibrations).

https://cordis.europa.eu/project/id/44096/reporting/es

Wrap-up


Data pre-processing
• Outlier analysis
• Data transformation
Classification
• Hierarchical cluster analysis
Discrete approach
Data
• Types
• Origin
Descriptive statistics
• Summary statistics
• Distributions
• Visualizations
Univariate Exploratory Data Analysis
• Univariate analyses
• Visualizations
Multivariate Exploratory Data Analysis
• Multivariate analyses
• Visualizations
Ordination
• PCA
• PCoA
• DA
• FA
• CA, MCA
Continuous approach
Types of research studies
Distribution tests
• One-sample hypothesis testing
Univariate hypothesis testing
• Two or more samples
Correlation and regression
• Linear model
• GLM
1 Response
Multiple predictors

Exercise 3

- Use the Univariate_analysis.ipynb and the dataset in EFIplus_medit.zip to plot **strip plots**, **histograms** and **boxplots** (and any additional plot that you feel appropriate) of **Annual Mean Temperature** (temp_ann) at each of the **four catchments with the highest number of fish sampling sites**. Try to fit each type of graph in a single window (4 graphs per window - check how to do it in previous examples I gave, which are available in github).
- You may change the settings in order to follow the best practices of data visualization (the ones I gave in the second lesson or other that you feel are also important).
- Have a deeper look at the three types of plots and evaluate the pros and cons of each type as univariate visualizations.
- I also challenge you to construct a plot that shows how the mean value of temp_ann varies with the size of random samplings of sites. Take 1000 random samples with replacement of increasing sample sizes (e.g. 10, 50, 100, 150, 200, 250, 300, 500 and 1000 observations), compute the mean Temp_ann of each sample and use an appropriate visualization to show how many samples will we need to have a good estimate of the population mean