CMU Portugal Academy

CMU Portugal
Advanced Training Program

Foundations of Data Science

DAVID SEMEDO
RAFAEL FERREIRA
NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Email: df.semedo@fct.unl.pt

CMU Portugal Academy

This Week's Topics

1. STATISTICS AND PROBABILITY

2. DATA LOADING, PREPARATION AND PROCESSING WITH PANDAS

3. DATA VISUALIZATION WITH PANDAS AND SEABORN

CMU Portugal Academy

What will you accomplish today?

- Load and Analyze a Dataset
- Data Pre-processing
- Data Analysis - Plotting

Use-Case Datasets

- Potential datasets:
  - Cost of Living Index by Country – Link
  - Red Wine Quality - Link
  - House Rent Prediction – Link
  - Stellar Classification – Link
  - Sleep Efficiency – Link
  - Healthcare Diabetes Dataset - Link

CMU Portugal Academy

01

Statistics and Probability

Carnegie Mellon Portugal | TÉCNICO LISBOA | Ciências ULisboa | NOVA SCHOOL OF SCIENCE & TECHNOLOGY

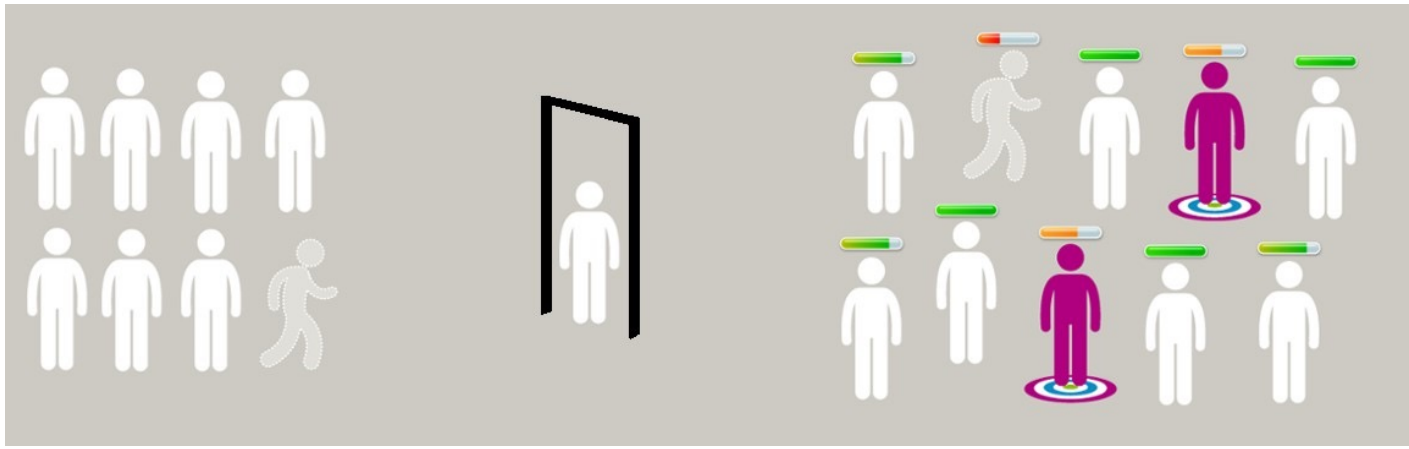
CMU Portugal Academy

Data Science and Uncertainty

- What is Data Science?
- Where does Uncertainty come from?

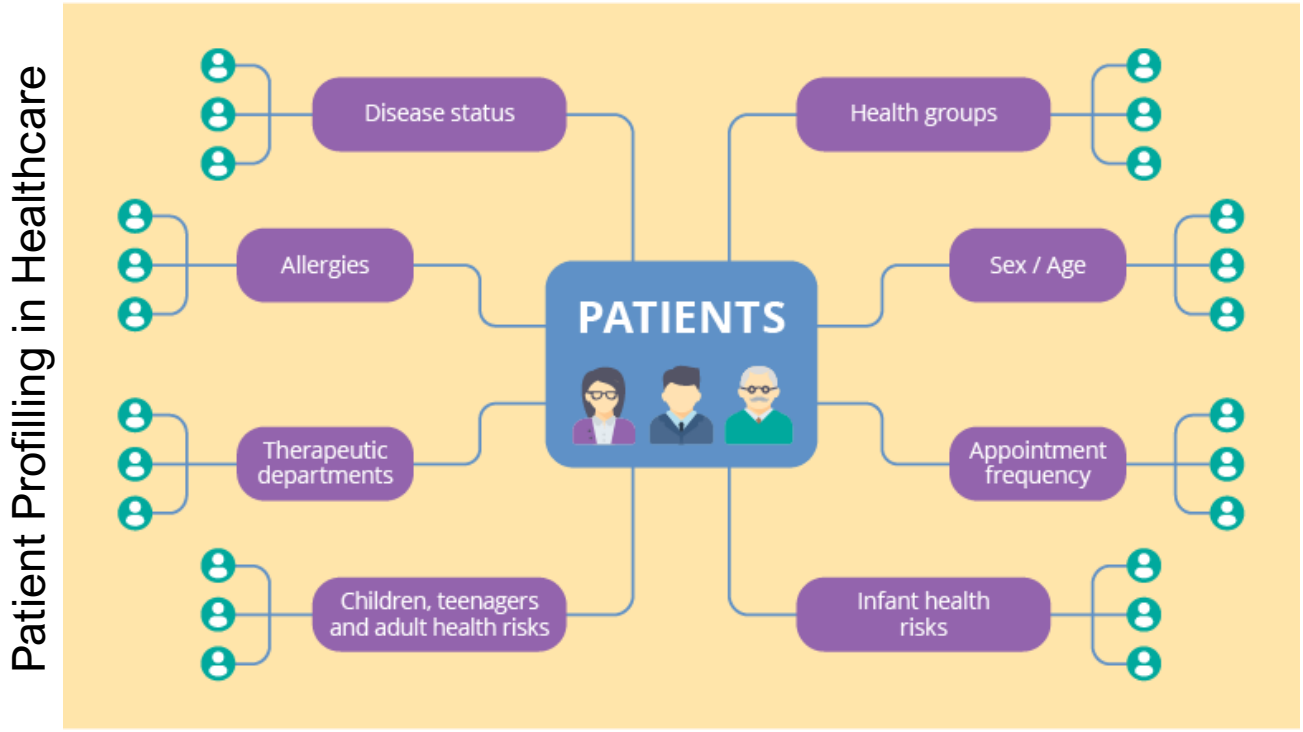
Data Science and Uncertainty

Churn prediction



Source: https://medium.com/@fatihfidan/customer-churn-analysis-55b6ebc8ca68

Data Science and Uncertainty



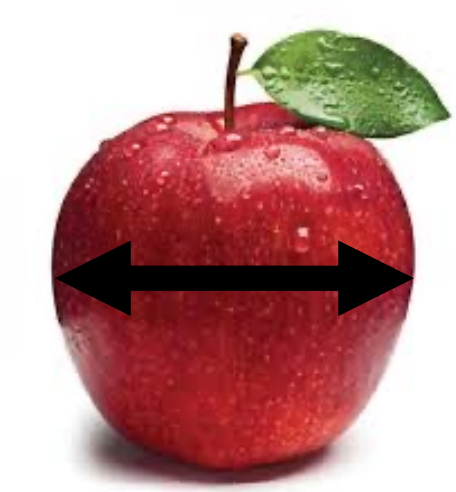
Source: https://www.scnsoft.com/healthcare/assembling-the-puzzle-of-a-patient-health-profile-in-crm

Data Science Lifecyle

We want to make decisions, from data, under uncertainty!

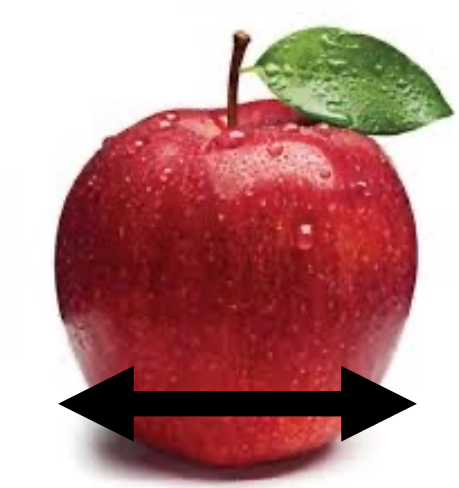
CMU Portugal Academy

1) Descriptive statistics

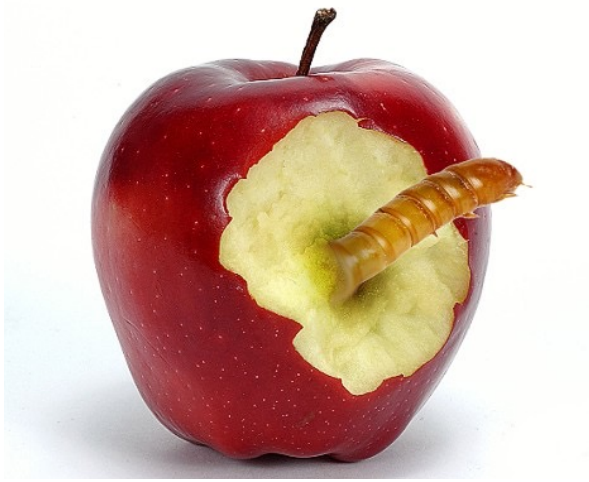


CMU Portugal Academy

1) Descriptive statistics

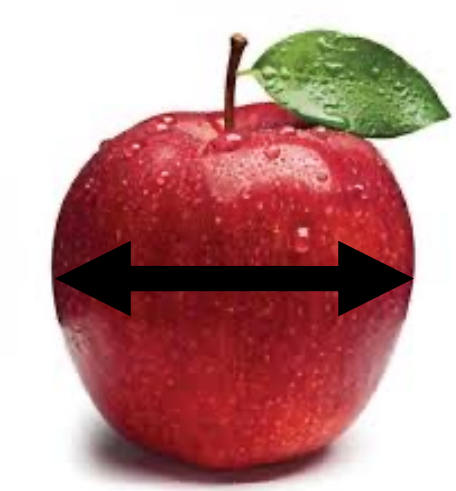


2) Exploratory

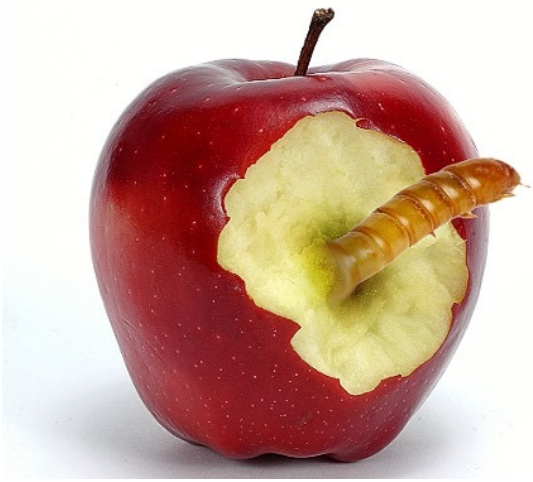


CMU Portugal Academy

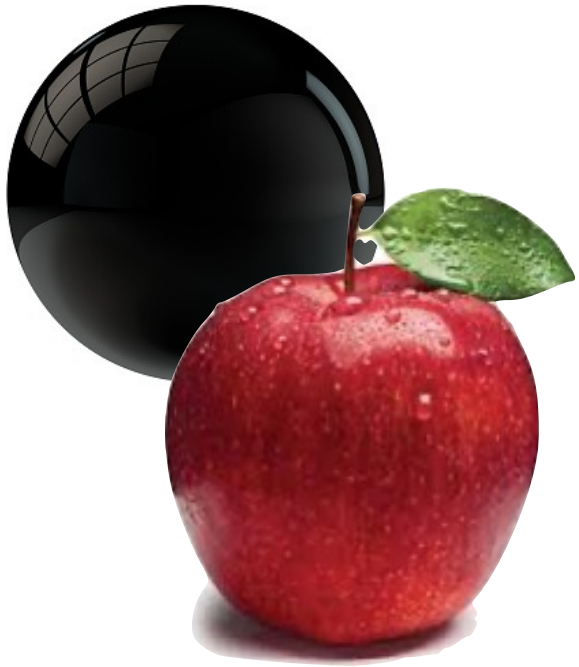
1) Descriptive statistics



2) Exploratory



3) Inferential statistics



CMU Portugal Academy

Descriptive Statistics

- Informational coefficients that summarize a data set
- Central tendencies vs. variability (spread)
- Data Distribution
- Univariate, Bivariate, Multivariate

Dataset Example

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

Dataset Example

Attributes

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

Data object

Data object = record, individual, point, event, observation, vector, entity

Attribute = field, feature, variable, dimension, characteristic

Types of Variables: Categorical or Quantitative

Discrete set of categories

Humans' eye colors

Energy Rating

CMU Portugal Academy

Categorical variables can be: nominal or ordinal

Humans' eye colors

Energy Rating

Types of Variables: Categorical or Quantitative

Humans' eye colors

Energy Rating

Take values for which arithmetic operations make sense

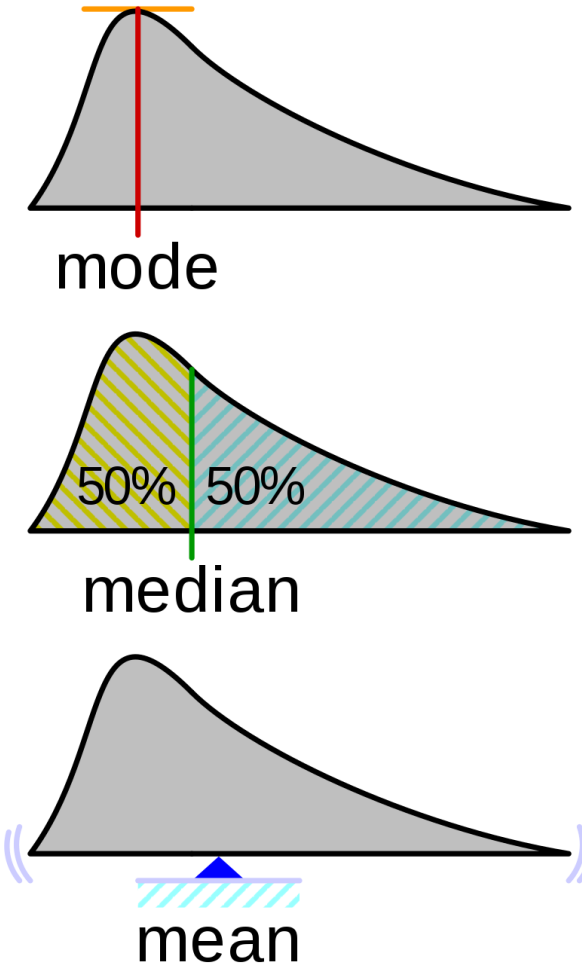
Quantitative variables can be in an interval or ratio

Humans’ eye colors

Energy Rating

CMU Portugal Academy

Descriptive Statistics



CMU Portugal Academy

Statistical Sample Moments

Mean – center

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance – Spread

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standardized

Skewness – Dispersion asymmetry

$$m_3 = \frac{1}{n\sigma^3}\sum_{i=1}^{n} (x_i - \bar{x})^3$$

Kurtosis – Tail “heaviness”

$$m_4 = \frac{1}{n\sigma^4}\sum_{i=1}^{n} (x_i - \bar{x})^4$$

Descriptive Statistics – Mean

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum x_i$$

It gives an idea of the "center" of the distribution

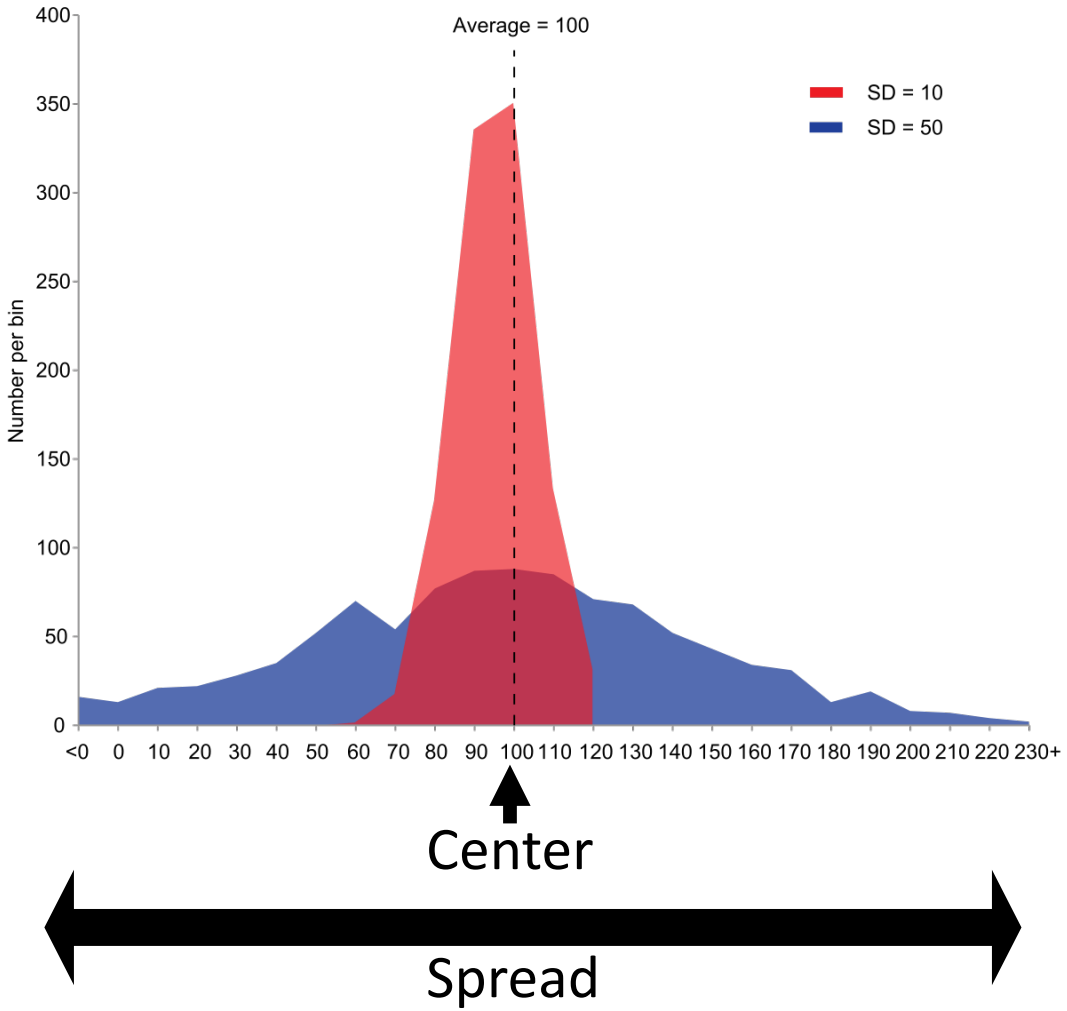
CMU Portugal Academy

Descriptive Statistics – Standard Deviation

Variance: $$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

$$= \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

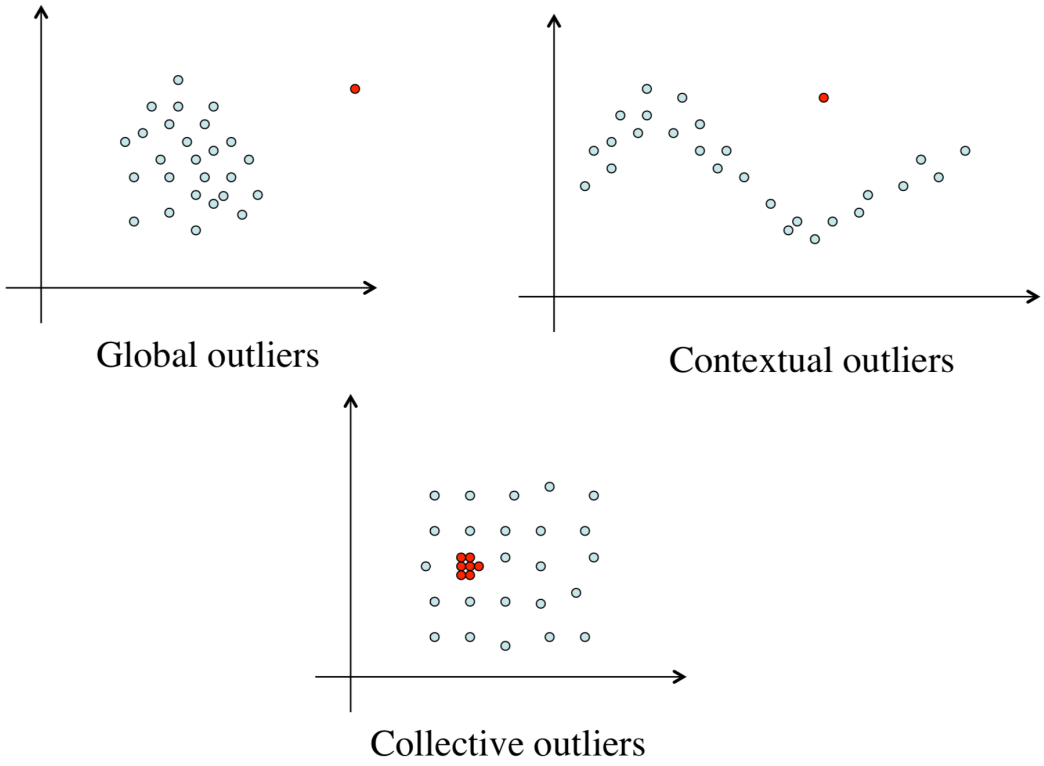
$$s = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$$



https://en.wikipedia.org/wiki/Statistical_dispersion

Mean and Variance – Do not tell the full story!

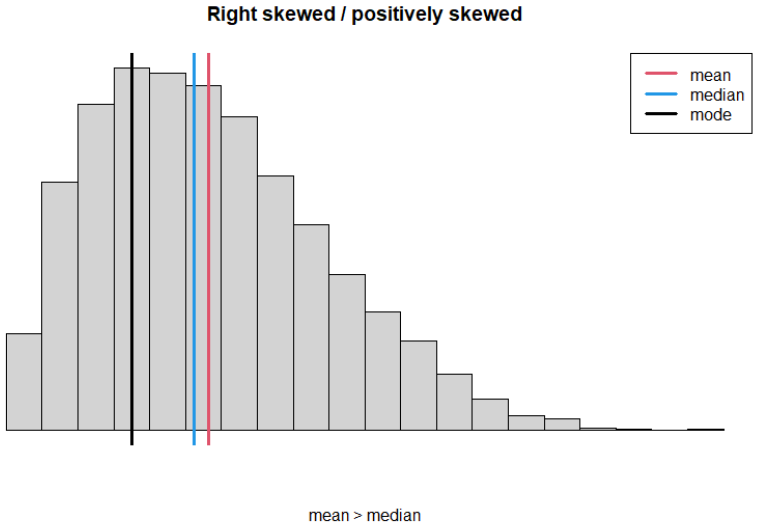
- Not robust to outliers

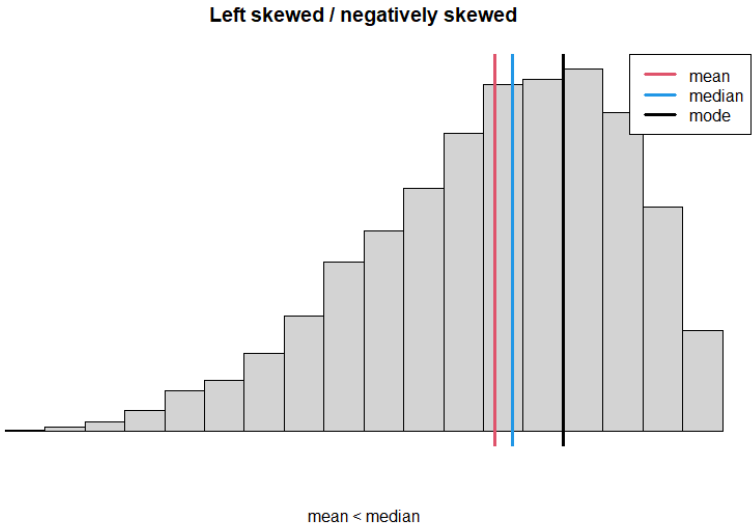


CMU Portugal Academy

Mean and Variance – Do not tell the full story!

- Not robust to outliers
- Inadequate for skewed distributions

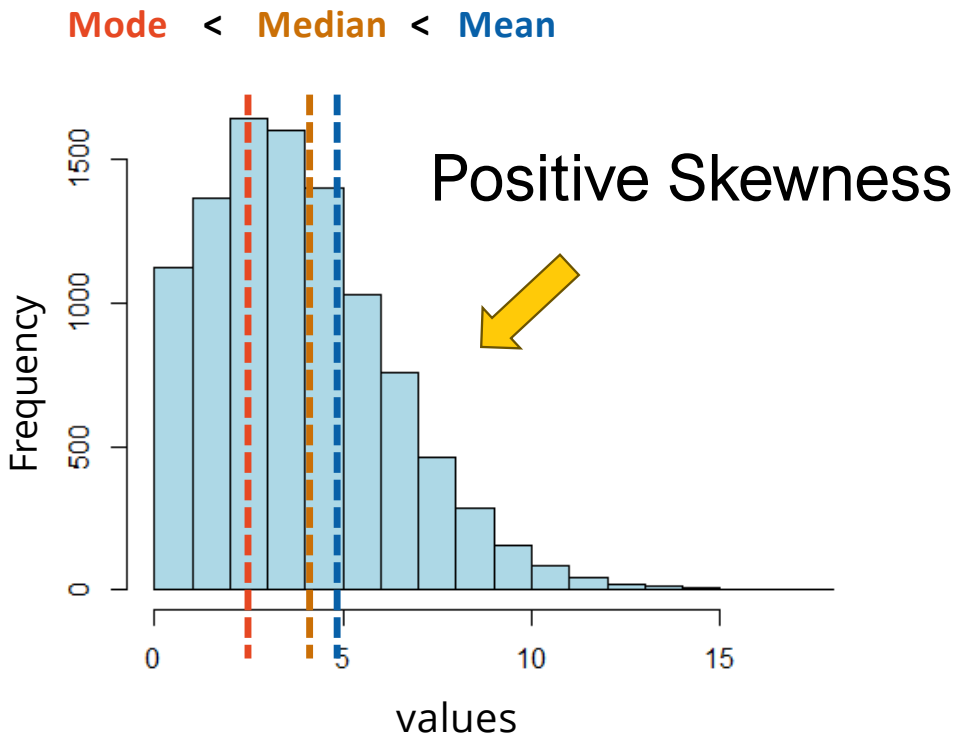




Skewness -3rd Statistical Sample Moment

Measures dispersion asymmetry

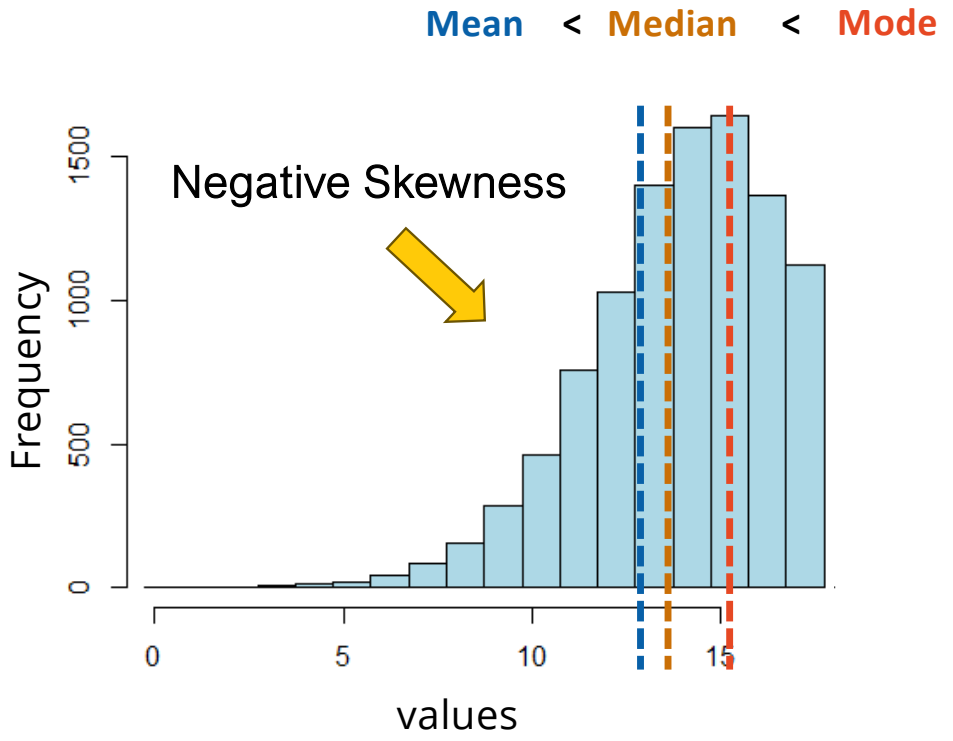
$$m_3 = \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$



Skewness -3rd Statistical Sample Moment

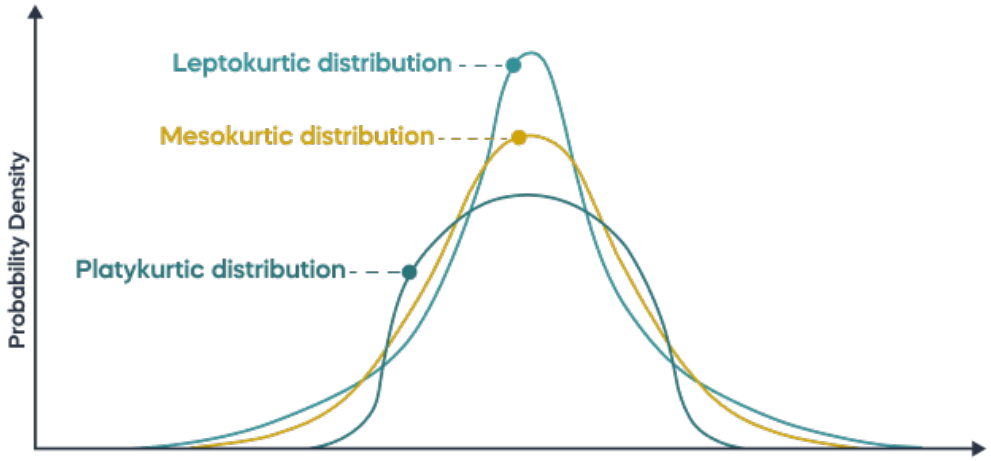
Measures dispersion asymmetry

$$m_3 = \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$



Kurtosis - 4th Statistical Sample Moment

Kurtosis – Tail “heaviness” – How often outliers occur.



Source: https://365datascience.com/calculators/kurtosis-calculator/

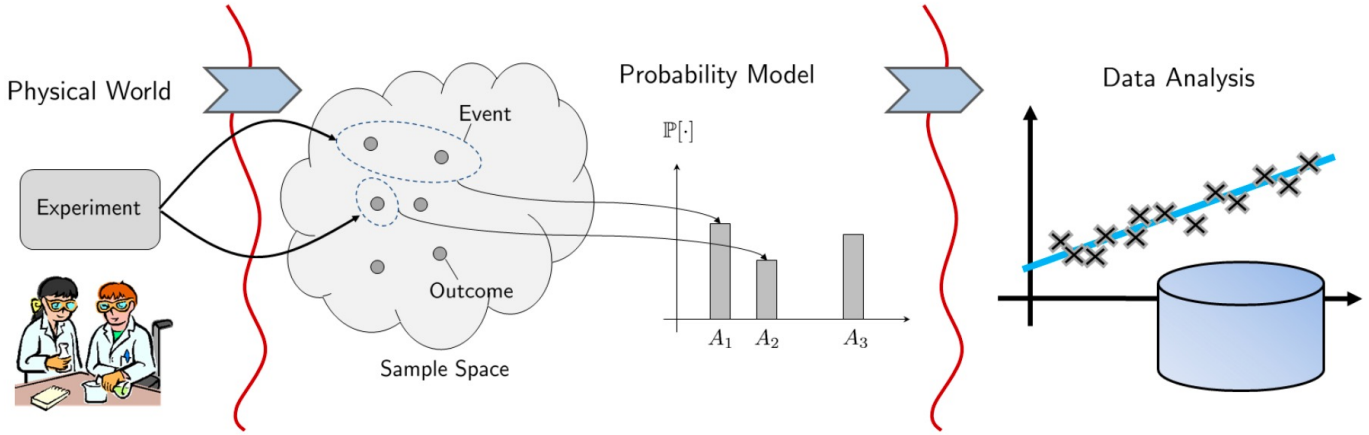
$$m_4 = \frac{1}{n\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Basic Probability Concepts

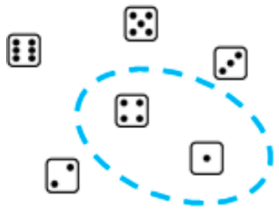
- **Frequentist**: The relative frequency of an outcome
- **Bayesian**: A subjective belief

Basic Probability Concepts

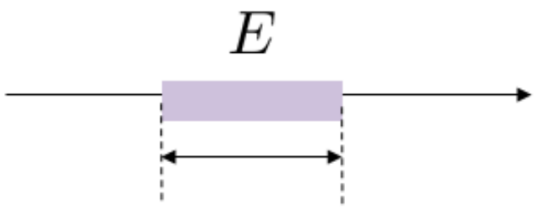
- **Sample space**: Set of all possible outcomes
- **Event space**: Collection of all possible events. One or a combination of outcomes
- **Probability model**: Measures the size of the event.



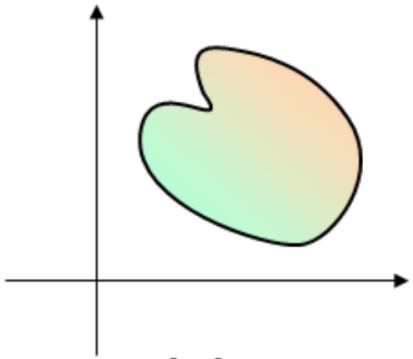
Basic Probability Concepts



$\mathbb{P}[E] = \text{count}$



$\mathbb{P}[E] = \text{length}$



$\mathbb{P}[E] = \text{area}$

CMU Portugal Academy

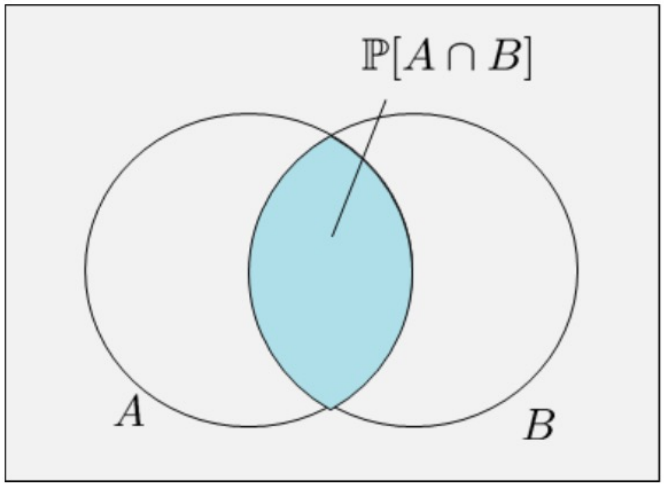
Probability Axioms

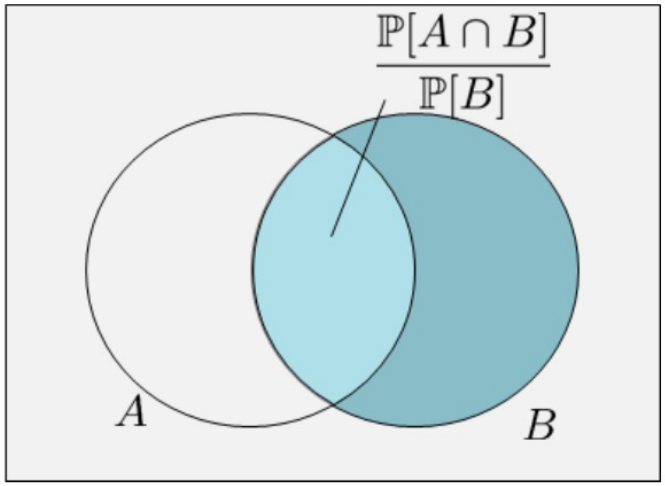
- Non-negativity: $\boldsymbol{P[E] \geq 0}$, for any possible event in $\boldsymbol{\Omega}$.
- Normalization: $\boldsymbol{P[\Omega] \geq 0}$
- Additivity: For any disjoint set of events $S = \{E_1, E_2, E_3\}$,

$$\boldsymbol{P[S] \geq \sum_{i=1}^{\infty} P[E_i]}$$

Conditional Probability – Bayes Theorem

- $\boldsymbol{P[A|\,B] = \frac{P[A \cap B]}{P[B]}}$, with $\boldsymbol{P[B] > 0}$





Probability Independence

- $\boldsymbol{P[A \cap B] = P[A] \cdot P[B]}$

- According to the conditional probability rule,

$$\boldsymbol{P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A] \cdot P[B]}{P[B]} = P[A]}$$

CMU Portugal Academy

Deriving Bayes Theorem

- $P[A|B] = \frac{P[A \cap B]}{P[B]}$ and $P[B|A] = \frac{P[B \cap A]}{P[A]}$
- $P[A|B] \cdot P[B] = P[B|A] \cdot P[A]$
- Bayes Theorem: $\underbrace{P[A|B]}_{\text{Posterior}} = \frac{\overbrace{P[B|A]}^{\text{Likelihood}} \cdot \overbrace{P[A]}^{\text{Prior}}}{\underbrace{P[B]}_{\text{Marginal}}}$

$$\longrightarrow \quad P[B] = \sum_i P[B|A_i] \cdot P[A_i]$$

Likelihood: How probable is **B**, if the hypothesis **A** is true?
Prior: How probable is the hypothesis **A**, before observing **B**?
Marginal: How probable is **B**, under all possible hypothesis A_i?
Posterior: How probable is the hypothesis, given the evidence **B**?

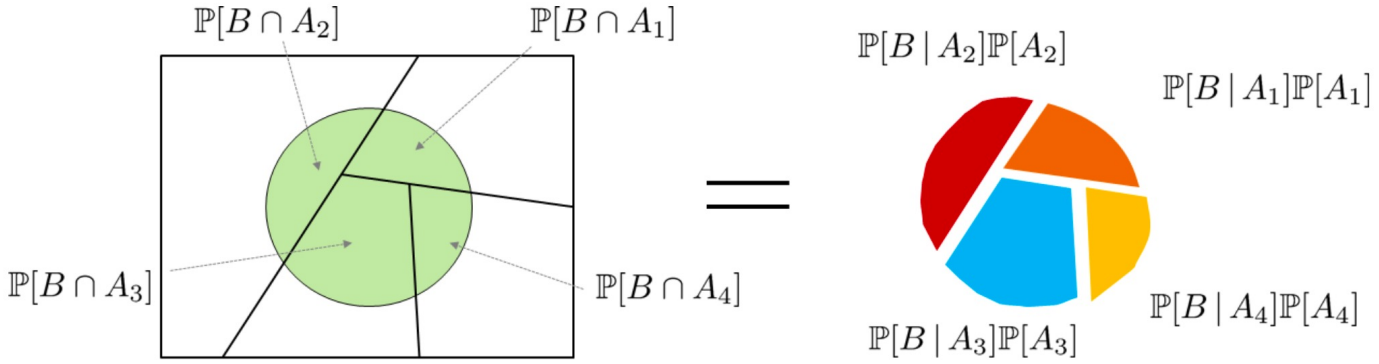
Law of Total Probability

- Bayes Theorem: $\underbrace{P[A|B]}_{\text{Posterior}} = \frac{\overbrace{P[B|A]}^{\text{Likelihood}} \cdot \overbrace{P[A]}^{\text{Prior}}}{\underbrace{P[B]}_{\text{Marginal}}}$

Likelihood: How probable is **B**, if the hypothesis **A** is true?
Prior: How probable is the hypothesis **A**, before observing **B**?
Marginal: How probable is **B**, under all possible hypothesis A_i?
Posterior: How probable is the hypothesis, given the evidence **B**?

$$P[B] = \sum_i P[B|A_i] \cdot P[A_i]$$

A_i are disjoint events



CMU Portugal Academy

Law of Total Probability

- Bayes Theorem: $P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B]}$

Likelihood Prior

Posterior

Marginal

Likelihood: How probable is **B**, if the hypothesis **A** is true?
Prior: How probable is the hypothesis **A**, before observing **B**?
Marginal: How probable is **B**, under all possible hypothesis A_i?
Posterior: How probable is the hypothesis, given the evidence **B**?

$$P[B] = \sum_i P[B|A_i] \cdot P[A_i]$$

A_i are disjoint events

$$P[A_j|B] = \frac{P[B|A_j] \cdot P[A_j]}{\sum_i P[B|A_i] \cdot P[A_i]}$$

Decomposing an event into smaller events

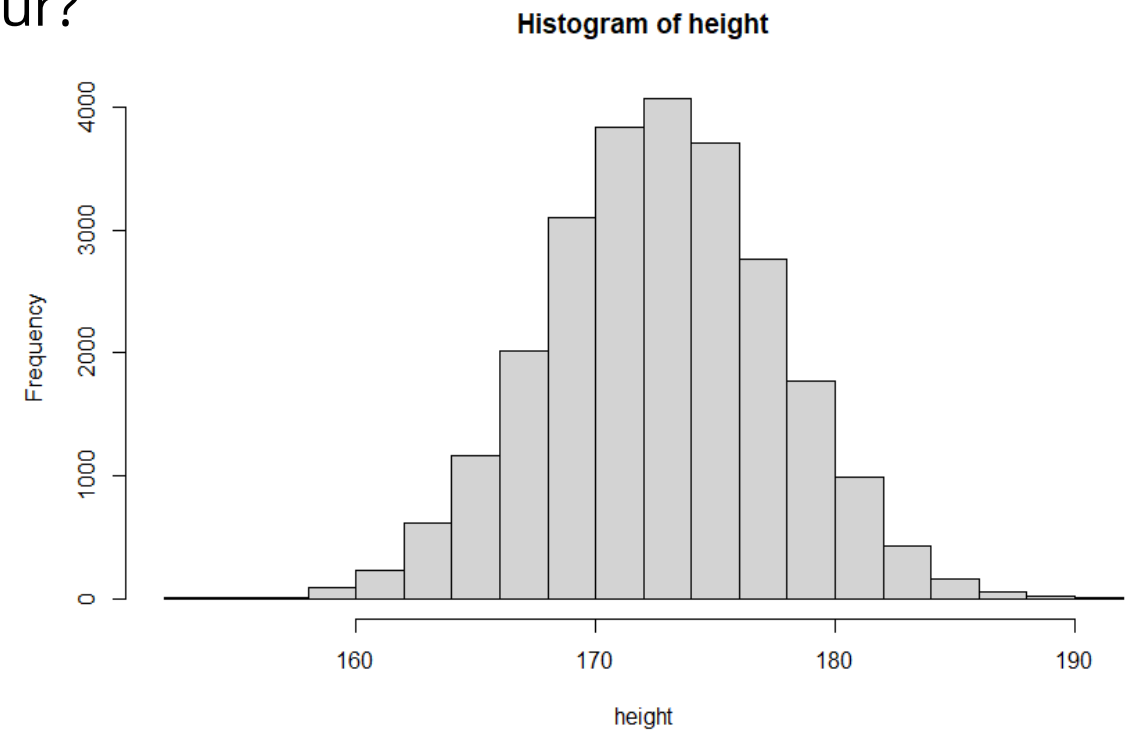
CMU Portugal Academy

Distributions

How often do certain values occur?

Discrete
- Bernouli (coin flip)
- Binomial (n coin flips)
- Geometric
- Poisson

Continuous
- Uniform
- Exponential
- Gaussian/Normal



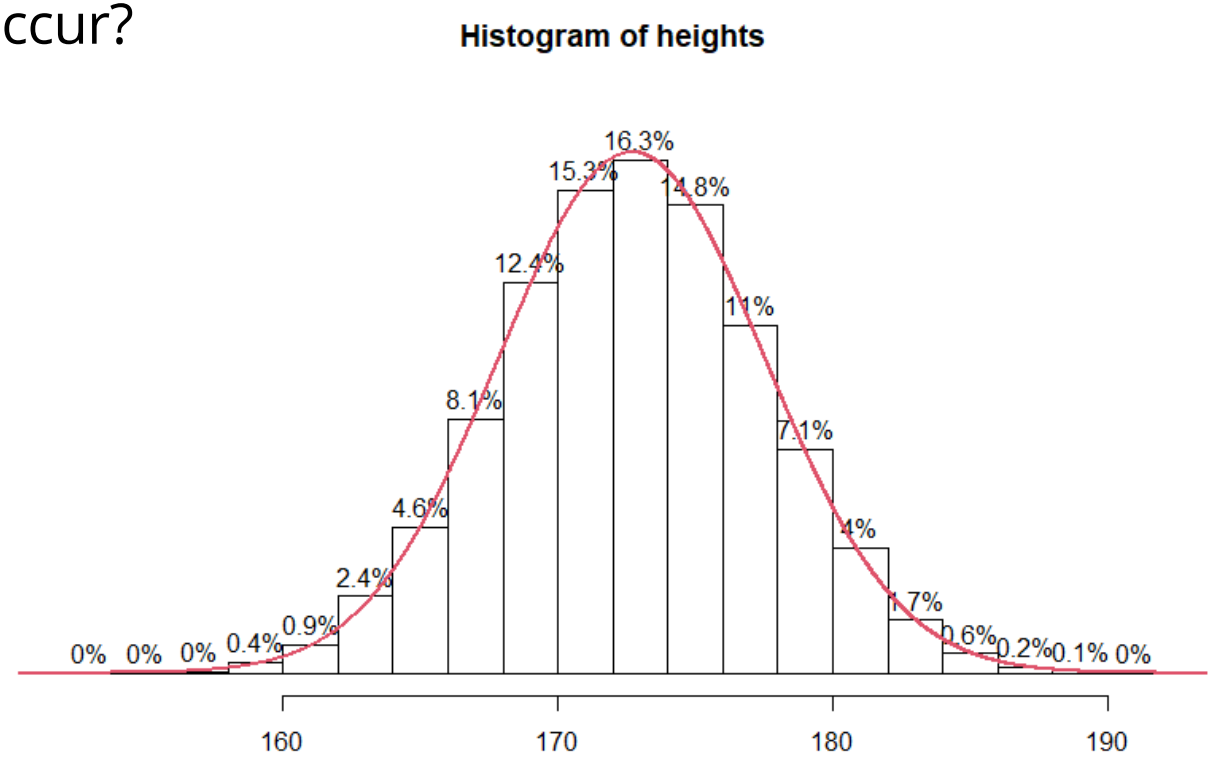
CMU Portugal Academy

Distributions

How often do certain values occur?

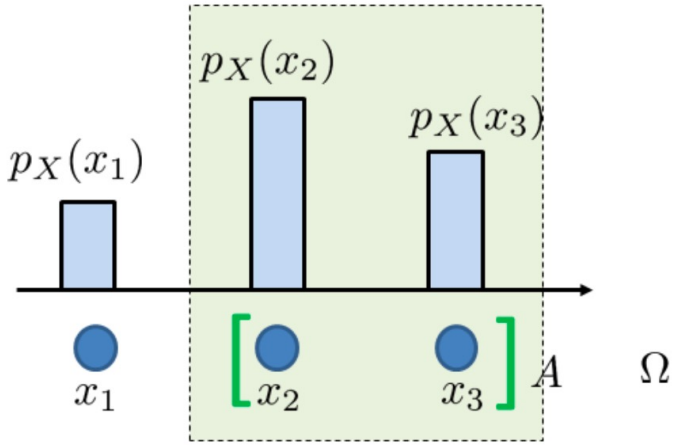
Discrete
- Bernouli (coin flip)
- Binomial (n coin flips)
- Geometric
- Poisson

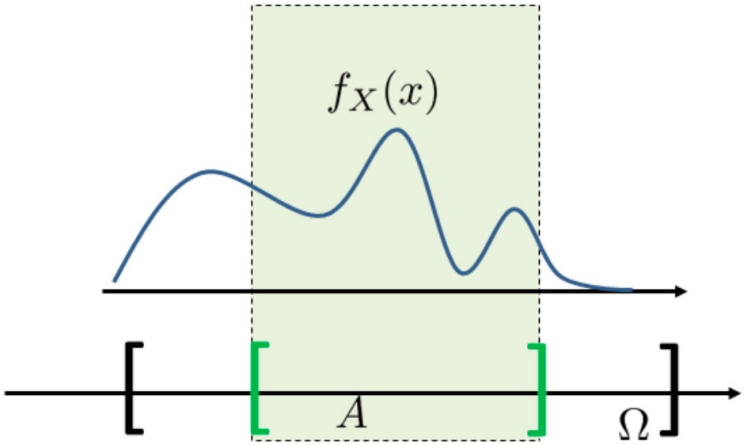
Continuous
- Uniform
- Exponential
- Gaussian/Normal



CMU Portugal Academy

Probability Mass vs Densify Functions

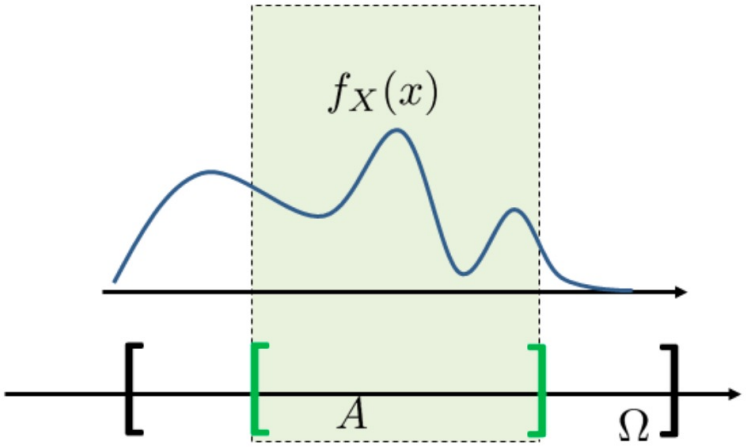




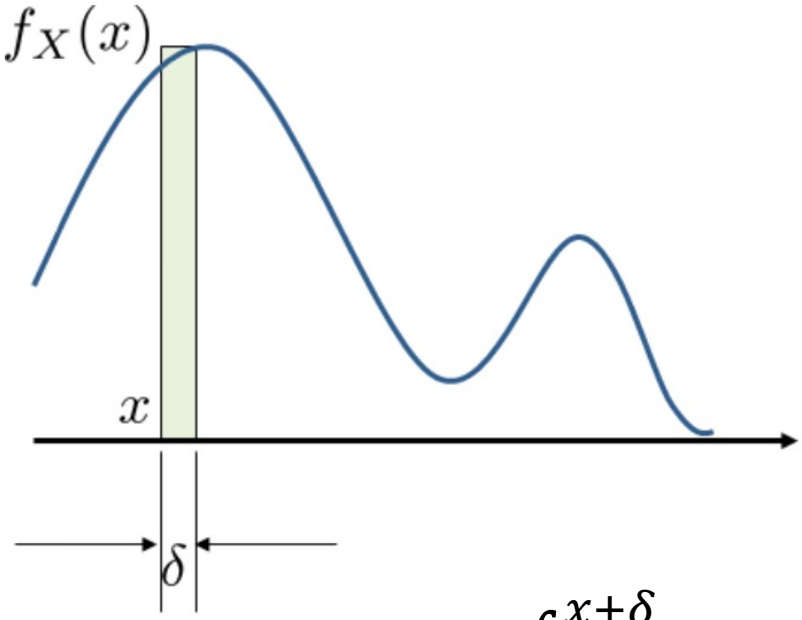
Probability Mass Function (PMF)
Discrete Random Variables

Probability Density Function (PDF)
Continuous Random Variables

From Density Functions to Probability



The probability is the area under the PDF

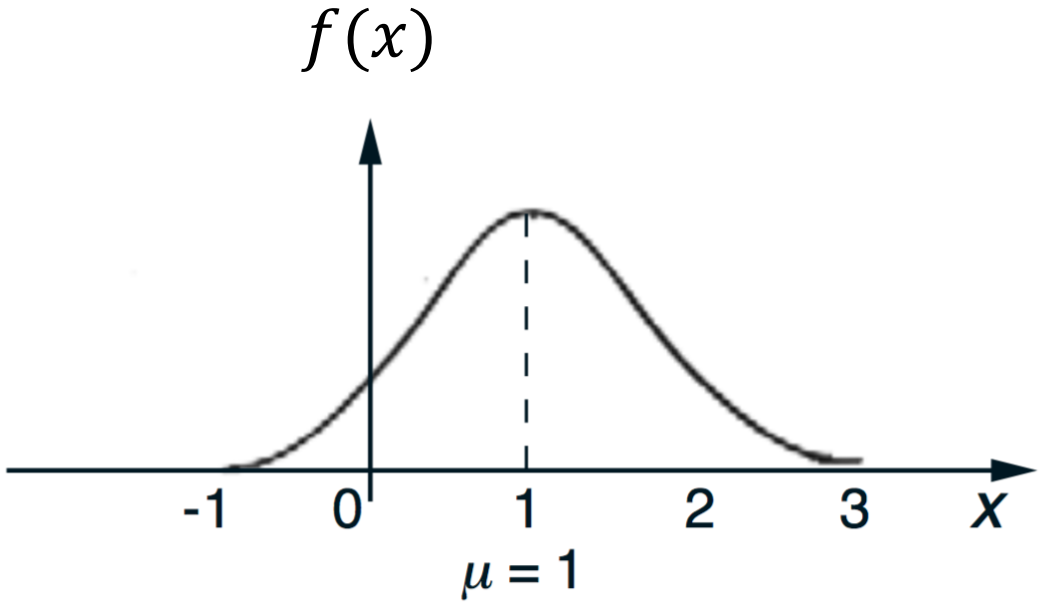


$$P[x \leq X \leq x + \delta] = \int_{x}^{x+\delta} f_X(x)$$

Gaussian/Normal Distribution $X \sim \mathcal{N}(\mu, \sigma^2)$

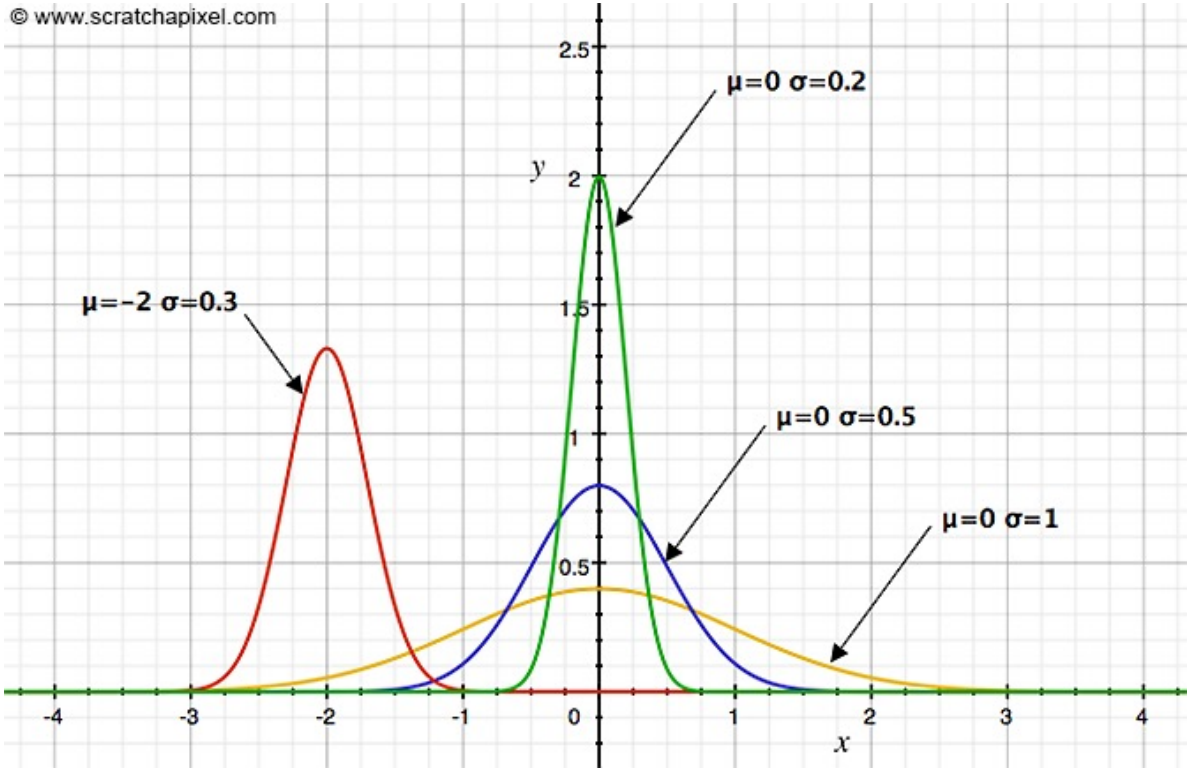
Probability Density Function $f(x)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Gaussian/Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

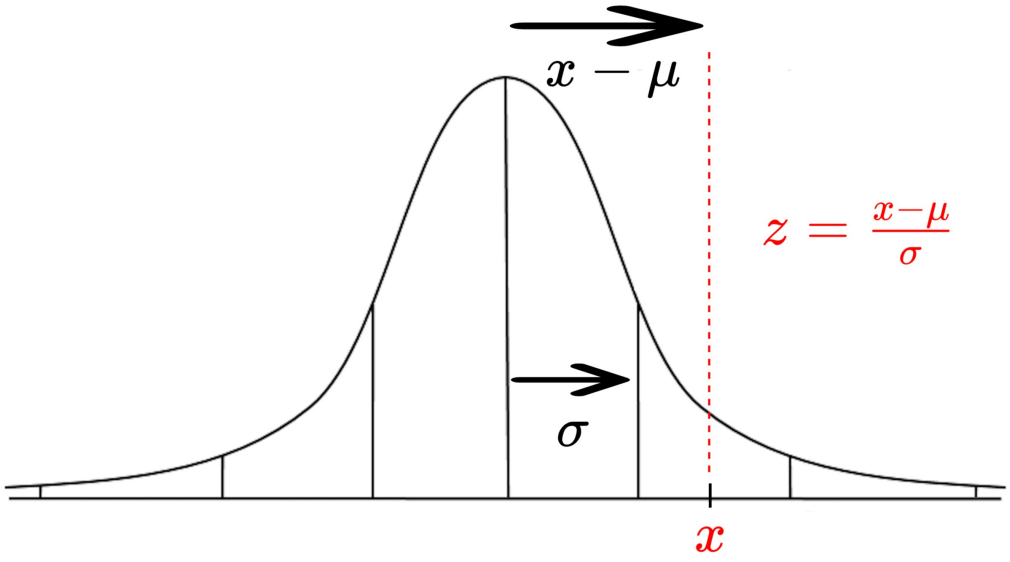


Gaussian/Normal Distribution - Standardization

Turn any normally distributed variable $X \sim \mathcal{N}(\mu, \sigma)$ into $\mathcal{N}(0,1)$ by standardizing it:

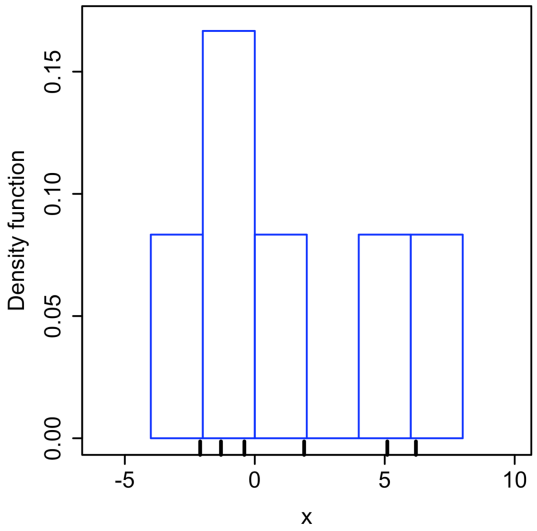
$$Z = \frac{X - \mu}{\sigma}$$

The z-score (standard score) is the number of standard deviations that an observed value x is away from the mean of a reference distribution

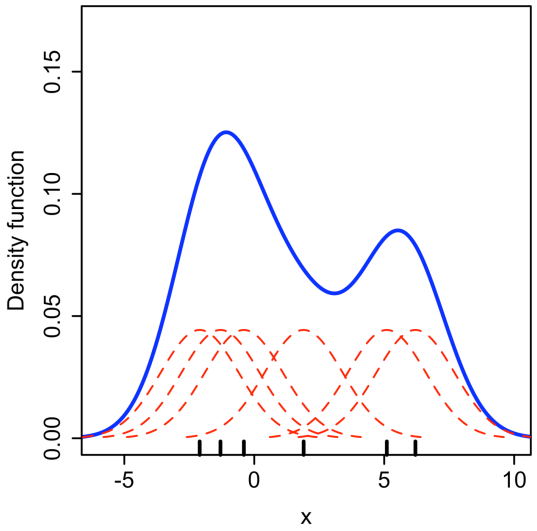


Fitting data to a distribution

Histogram

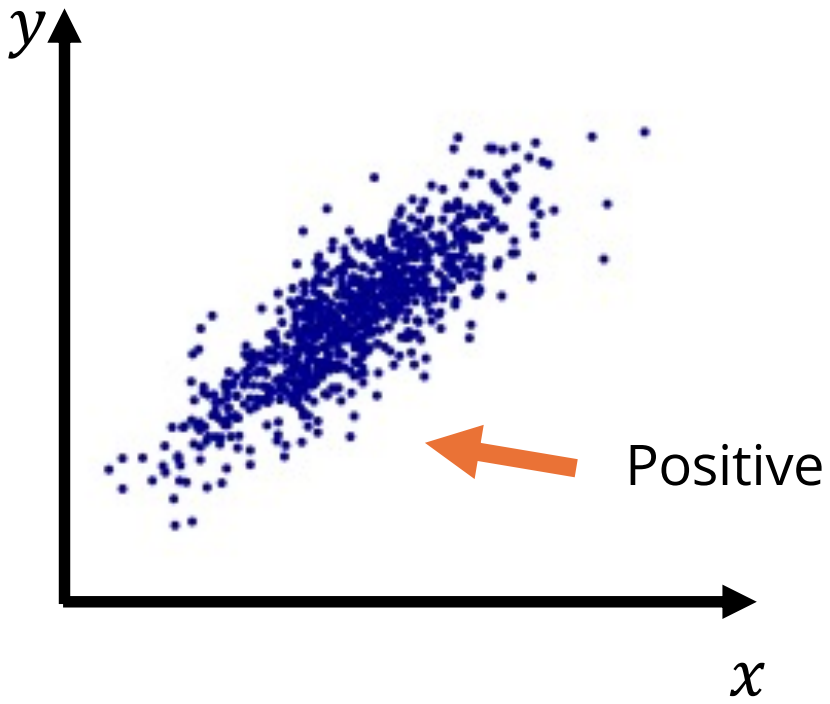


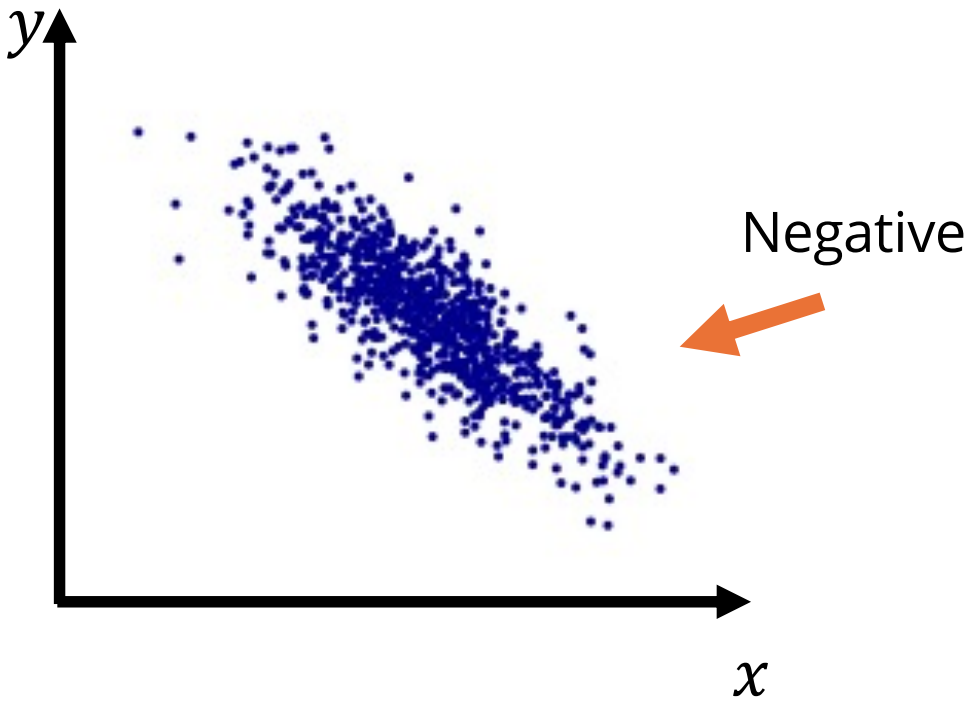
Kernel Density Estimation



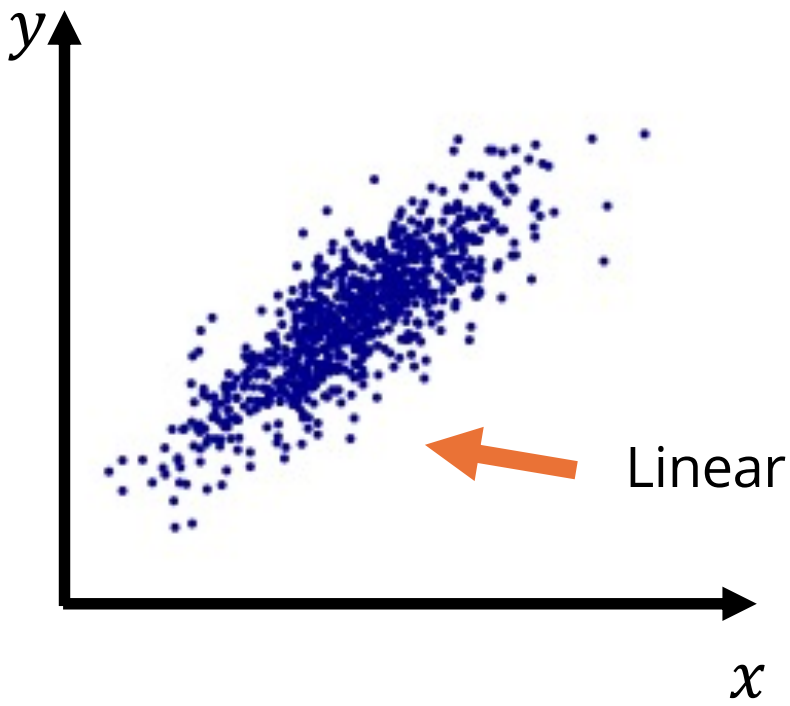
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html
https://en.wikipedia.org/wiki/Kernel_density_estimation

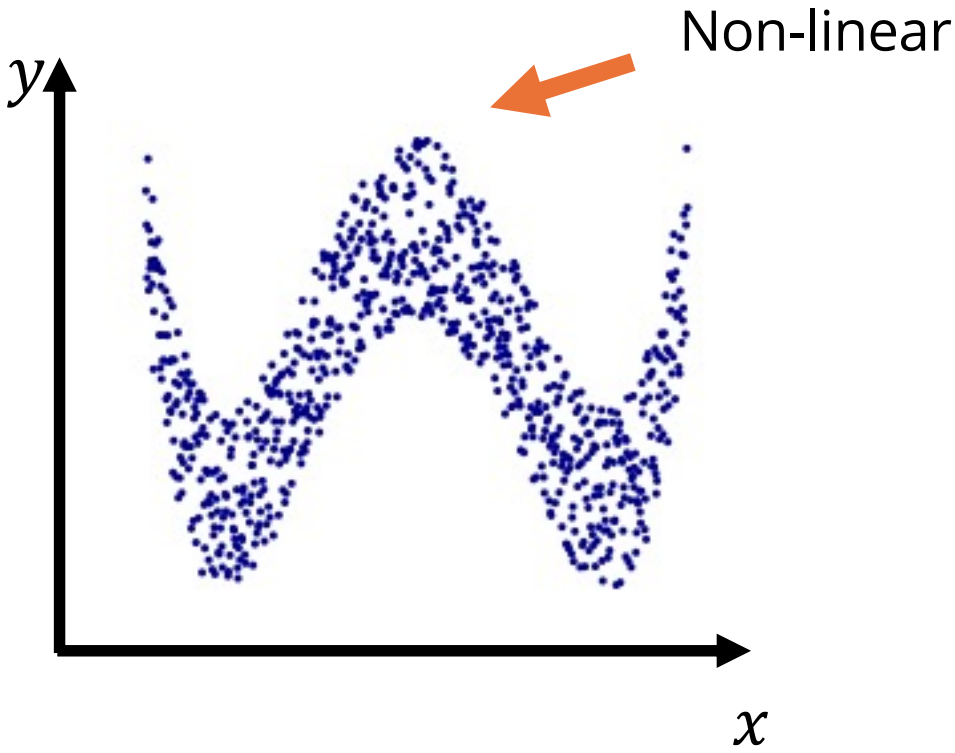
Correlation





Correlation





CMU Portugal Academy

Pearson Correlation Coefficient r

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

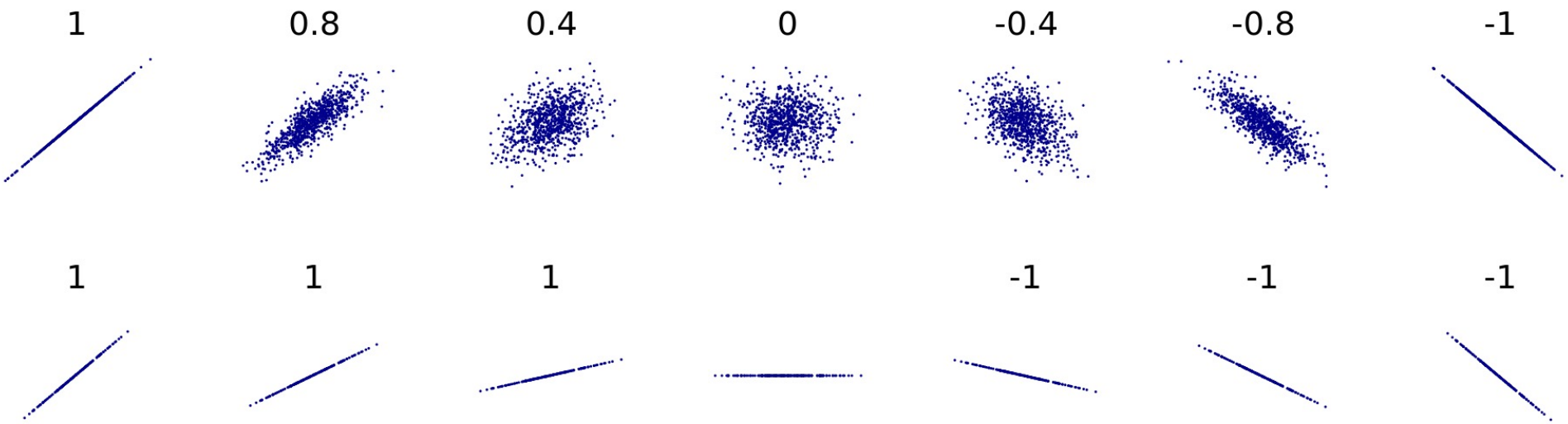
standardized x values

standardized y values

Therefore r does not change with linear transformations

CMU Portugal Academy

Pearson Correlation Coefficient r

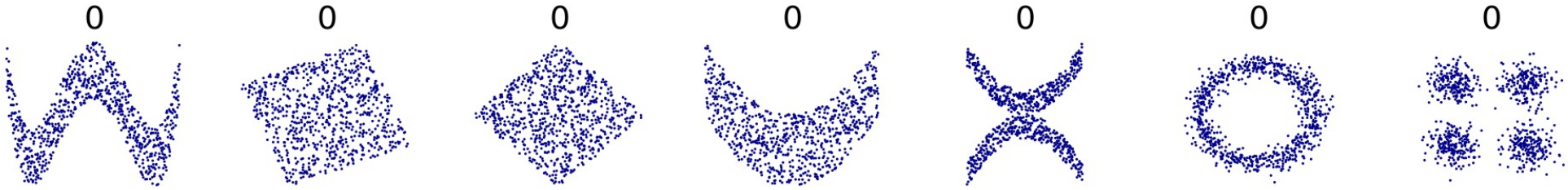


Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

CMU Portugal Academy

Pearson Correlation Coefficient r

Correlation is an incomplete description of bi-variate data!



Shouldn't be used:

- Outliers
- Non-linear relationships
- Subgroups

Data Visualization!!

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

CMU Portugal Academy

02

Data Loading, Preparation and Processing with Pandas

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

CMU Portugal Academy

Pandas


```
$ conda install pandas
```

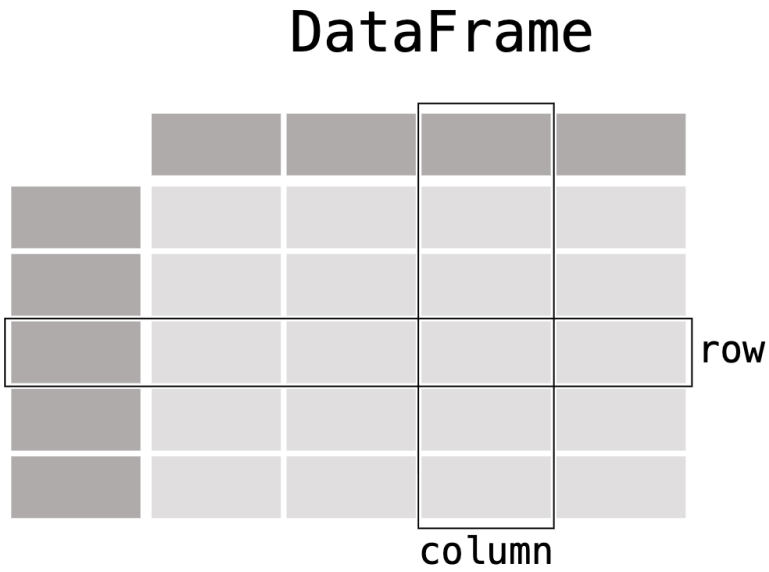
- Python Library for Data Analysis
- Manipulating Data Tables
- Highly flexible and versatile!
- Leverages NumPy for numerical operations and Plotting Libraries.

CMU Portugal Academy

Pandas

DataFrame: A 2-dimensional data structure that can store data of different types.

Similar to a spreadsheet or SQL table.



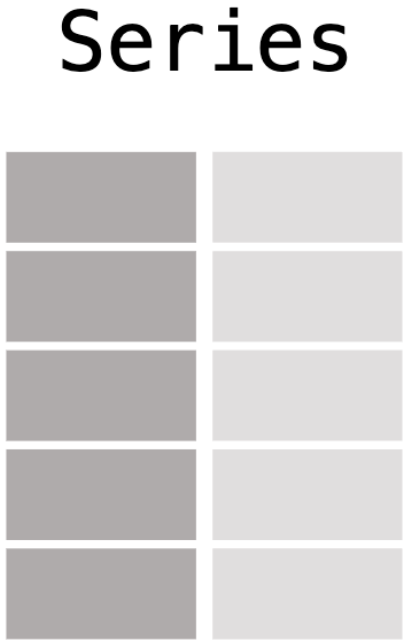
CMU Portugal Academy

Pandas

Series: Corresponds to a single column of a DataFrame.

Columns do not have labels

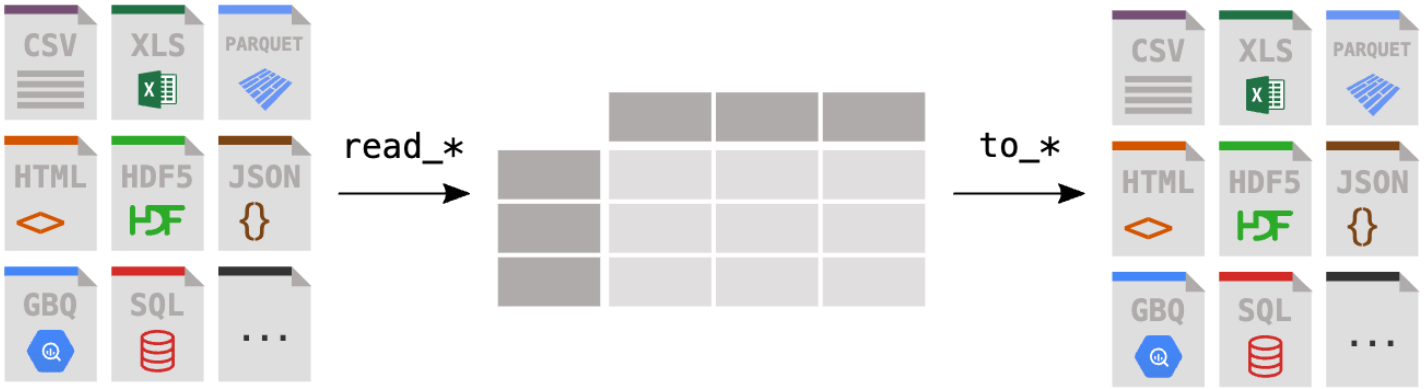
Rows have labels.



CMU Portugal Academy

Pandas

We can import data from multiple formats to a DataFrame, and export it back



Loading a dataset in CSV – Mental Illness Dataset

Comma Separated Values (CSV)

```
Entity,Code,Year,Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized,Depressive dis
Afghanistan,AFG,1990,0.22320578,4.996118,4.713314,0.70302314,0.12770003
Afghanistan,AFG,1991,0.22245377,4.9892898,4.7021,0.7020688,0.123255946
Afghanistan,AFG,1992,0.22175121,4.9813457,4.683743,0.700792,0.11884415
Afghanistan,AFG,1993,0.22098725,4.9769583,4.6735487,0.70008695,0.11508888
Afghanistan,AFG,1994,0.22018303,4.977782,4.67081,0.6998978,0.11181468
Afghanistan,AFG,1995,0.2194088,4.978228,4.6681,0.6997684,0.10850699
Afghanistan,AFG,1996,0.21846454,4.981489,4.6657586,0.6996502,0.10526882
Afghanistan,AFG,1997,0.21728611,4.9875927,4.665013,0.69959736,0.10153082
Afghanistan,AFG,1998,0.21607415,4.9968576,4.6682405,0.69976467,0.09805014
Afghanistan,AFG,1999,0.21506761,5.004257,4.6735573,0.700054,0.095722646
Afghanistan,AFG,2000,0.21451464,5.0084476,4.6768804,0.70025194,0.09487554
Afghanistan,AFG,2001,0.21431455,5.0038724,4.675786,0.70012945,0.09454673
Afghanistan,AFG,2002,0.21418841,4.994327,4.6720624,0.69977945,0.09481004
Afghanistan,AFG,2003,0.21409267,4.9817224,4.6718807,0.69960576,0.09498258
Afghanistan,AFG,2004,0.21406065,4.9731607,4.6727514,0.6995014,0.09553074
Afghanistan,AFG,2005,0.21409707,4.9658794,4.6736326,0.6994395,0.095989875
Afghanistan,AFG,2006,0.214259,4.96284,4.6780405,0.69953,0.09687539
Afghanistan,AFG,2007,0.21459809,4.9597197,4.682732,0.6995203,0.09869648
Afghanistan,AFG,2008,0.21502665,4.9531674,4.688263,0.69948006,0.100904115
Afghanistan,AFG,2009,0.21545461,4.9492145,4.695408,0.6994876,0.10303231
Afghanistan,AFG,2010,0.21580303,4.946899,4.701994,0.6994379,0.10496491
Afghanistan,AFG,2011,0.21621753,4.945379,4.715658,0.69939154,0.10716891
Afghanistan,AFG,2012,0.21681084,4.9445314,4.73926,0.69936365,0.109071545
Afghanistan,AFG,2013,0.21744443,4.9443674,4.7667584,0.6993666,0.11139344
Afghanistan,AFG,2014,0.21797295,4.944629,4.7931795,0.69947726,0.11349482
```

Loading a dataset in CSV – Mental Illness Dataset

```
dataset_path = "datasets/1- mental-illnesses-prevalence.csv"
df = pd.read_csv(dataset_path)
```

[11] ✓ 0.0s Python

```
df
```

[8] ✓ 0.0s Python

	Entity	Code	Year	Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized	Depressive disorders (share of population) - Sex: Both - Age: Age-standardized	Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized	Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized	Eating disorders (share of population) - Sex: Both - Age: Age-standardized
0	Afghanistan	AFG	1990	0.223206	4.996118	4.713314	0.703023	0.127700
1	Afghanistan	AFG	1991	0.222454	4.989290	4.702100	0.702069	0.123256
2	Afghanistan	AFG	1992	0.221751	4.981346	4.683743	0.700792	0.118844
3	Afghanistan	AFG	1993	0.220987	4.976958	4.673549	0.700087	0.115089
4	Afghanistan	AFG	1994	0.220183	4.977782	4.670810	0.699898	0.111815
...	...	...	...	...	...	...	...	...
6415	Zimbabwe	ZWE	2015	0.201042	3.407624	3.184012	0.538596	0.095652
6416	Zimbabwe	ZWE	2016	0.201319	3.410755	3.187148	0.538593	0.096662
6417	Zimbabwe	ZWE	2017	0.201639	3.411965	3.188418	0.538589	0.097330
6418	Zimbabwe	ZWE	2018	0.201976	3.406929	3.172111	0.538585	0.097909
6419	Zimbabwe	ZWE	2019	0.202482	3.395476	3.137017	0.538580	0.098295

6420 rows × 8 columns

Pandas – Basic Indexing

Supported Indexing Operations:

- By column (or columns) names;
- By row number (`iloc`)
- By row and column labels (`loc`)
- By a filtering condition, over all or subset of columns

More on this topic later!

CMU Portugal Academy

Carnegie Mellon Portugal | TÉCNICO LISBOA | Ciências ULisboa | NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Hands-On Session!

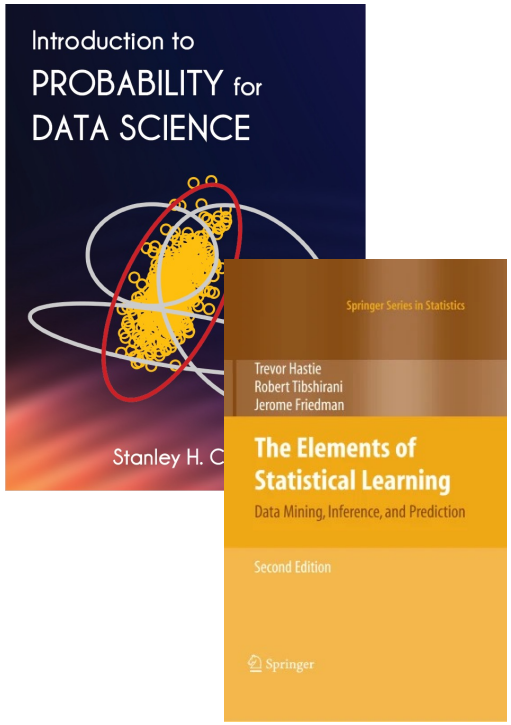
Course Shared Folder

CMU Portugal Academy

More on Statistics, Probability and Inference

- **Introduction to Probability for Data Science**, Stanley H. Chan, Michigan Publishing

- **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd edition, Springer



CMU Portugal Academy

CMU Portugal
Advanced Training Program

Foundations of Data Science

DAVID SEMEDO
RAFAEL FERREIRA
NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Email: rah.ferreira@campus.fct.unl.pt