



# L02 Visualization & Statistics

D. Schneider, B. Stuhr, J. Haselberger

## Data & Data Science

- Data itself are simple collections of different information (time series, object series, images, ...)
- Data Science reveals information within the data (using various methods, e. g. Clustering)

## Machine Learning & Deep Learning

- ML uses algorithms to analyze data, learn from that, and make informed decisions based on what is learned
- DL is a subset of ML. While both belong to the category of AI, DL is what drives the most human-like AI

## Python

- Python as one of the leading interpreted languages (high-level) for Data Science
- Google Colab (and much more) to work online in an agile way
- Python uses “call by object reference”
- Different types with different methods and functions available
- Large number of libraries available

# Agenda

Theory (120 min)

Break (15 min)

Exercise (60 min)

L02.1 (20 min)

- Visualization
- Matplotlib
- Scatter plot
- Histogram

L02.2 (30 min)

- Mean
- Quantiles
- Mode
- Standard deviation
- Variance

L03.3 (20 min)

- Gaussian distribution
- Covariance
- Sets and Spaces

L02.4 (30 min)

- Probability
- Conditional probability
- Bayes' Theorem

L02.5 (20 min)

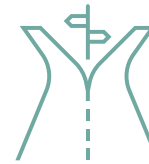
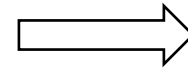
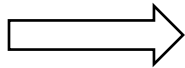
- Linear Algebra
- Eigenvalues and Eigenvectors

E02.1 (60 min)

- Home assignment

## L02.1 Visualization

- As a Data Scientist, you need to be able to create visual analytics and present them to your team members, your boss, but also to yourself
- Data visualization is a **modern form of visual communication**. It involves the creation and study of visual representation of data. Which is used to **make the decision-making process** and helps to **quickly understand the analyses** presented visually, so that anyone can grasp difficult concepts or identify new patterns.



### Visual data exploration (Data mining) [[deOliveira2003](#)]

1. Data without hypotheses about the data
2. Interactive, usually undirected search for structures, trends, etc.
3. Visualization of the data, which provides information (hypothesis) about the findings

### Confirmative Analysis

1. Data with hypotheses about it
2. Examination of the hypothesis's key features
3. Visualization of data that provides insight into the hypothesis to be rejected or accepted

### Presentation

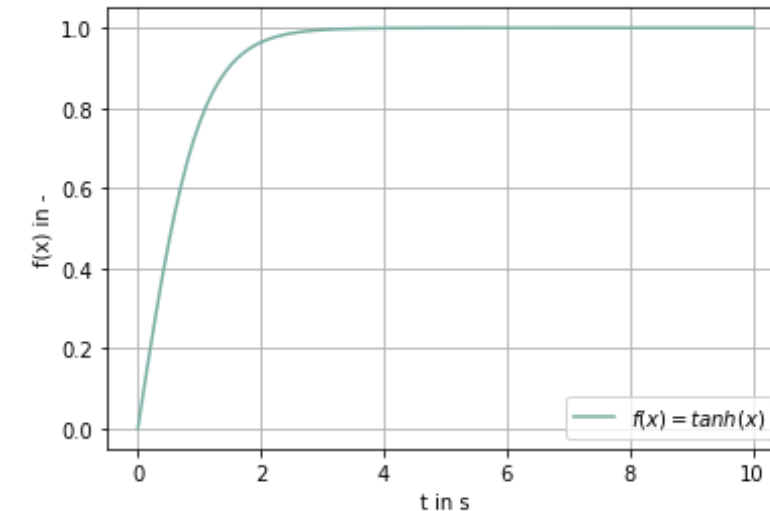
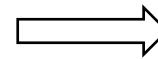
- Present existing information that was previously extracted
- The right choice of a suitable presentation technique is key



- [Matplotlib](#) is the state-of-the-art library for scientific figures and plots
- Since the community is large, a lot of [impressive examples](#) exists (using domain specific packages as seaborn, ...)
- Figures can be easily exported for further presentation purposes

```
# Data preparation
x = np.linspace(0,10,100)
y = np.tanh(x)

# Plot and configuration
fig = plt.figure()
plt.plot(x, y, c='#79AEA3', label=r'$f(x) = \tanh(x)$')
plt.legend()
plt.grid()
plt.xlabel('t in s')
plt.ylabel('f(x) in -')
fig.savefig('tanh.png',bbox_inches='tight', transparent=True)
```



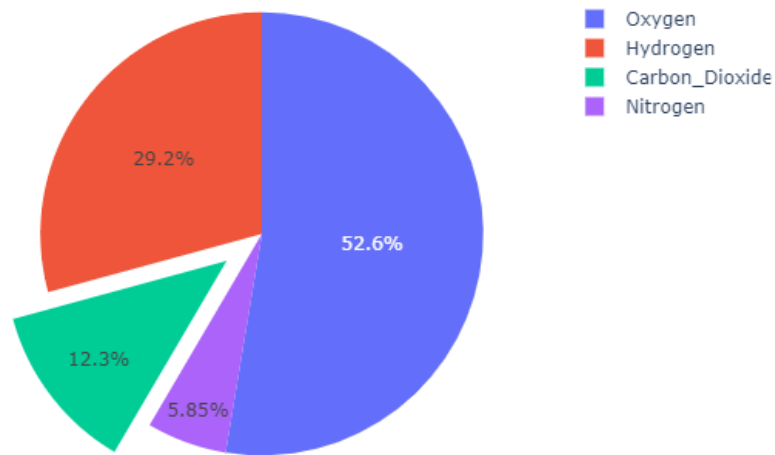
- High-end engines available for Data Scientists such as [Plotly Dash](#), [Tableau](#), [Cuxfilter](#) which we will not discuss in our lecture

### What makes a good scientific graphic?

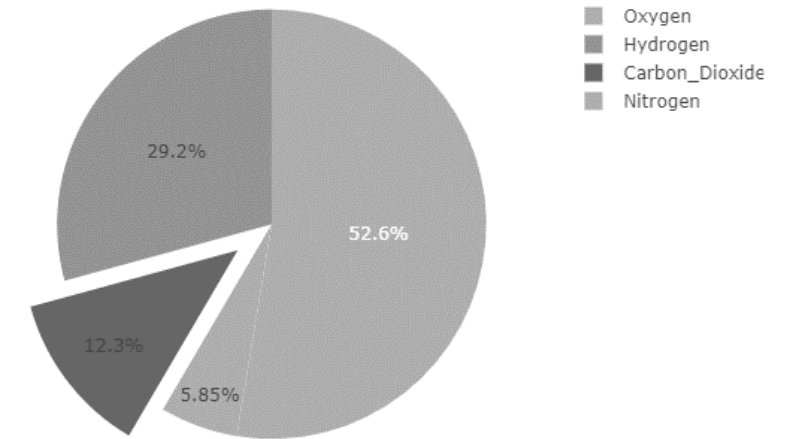
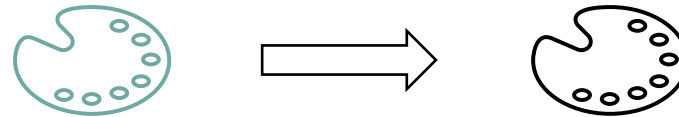
1. Show the data clearly
  - Showing the data clearly, including meaningful text information and labels
  - Use standard SI-units
2. Use simple designs and graphs
  - A graph with a simple design is often sufficient and better than a complex, overloaded design
3. Use alignments on common scale
  - Use a single linear scale, when possible
  - Use grid points to structure the plot
  - Common x-axis
4. Keep the visual encoding transparent
  - Readers must decode the diagram
  - Decoding best when the task is facilitated by clever choices in the design of the diagram
5. Use standard graphs
  - Scatterplot, Histogram, Boxplot, Time series plot, Bar chart, ...

# L02.1 Visualization

- [Color schemes](#) are important to distinguish between the different quantities in the data



Pie chart from [\[Plotly\]](#)

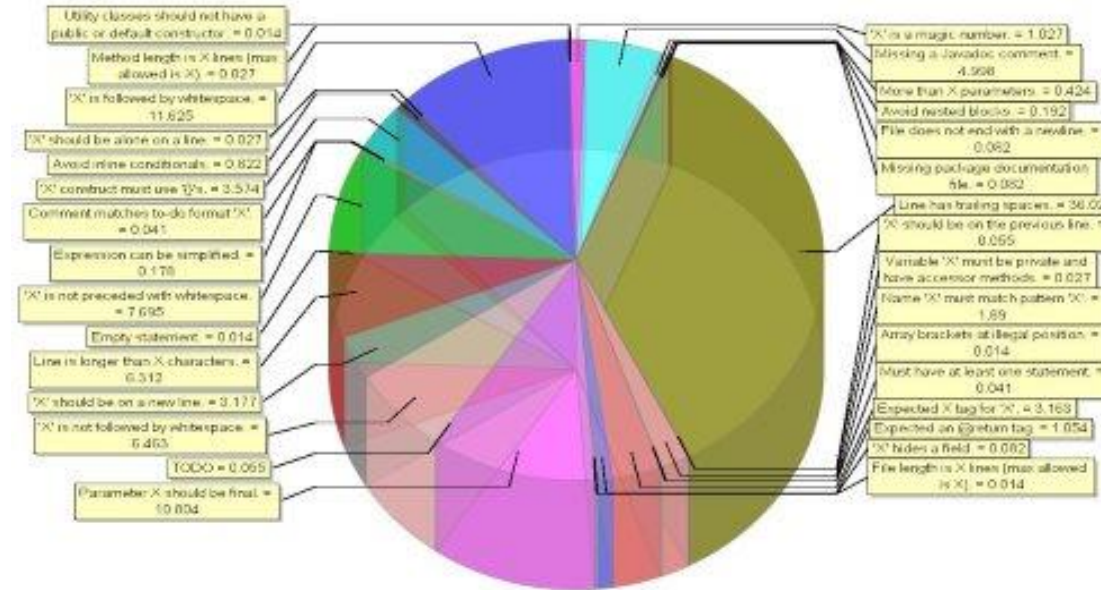


Pie chart grayscale

- Keep in mind, that publications often printed in grayscale
- Make sure that the main features of the diagram remain separable



## L02.1 Visualization

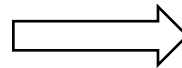


Overloaded pie chart (bad example)

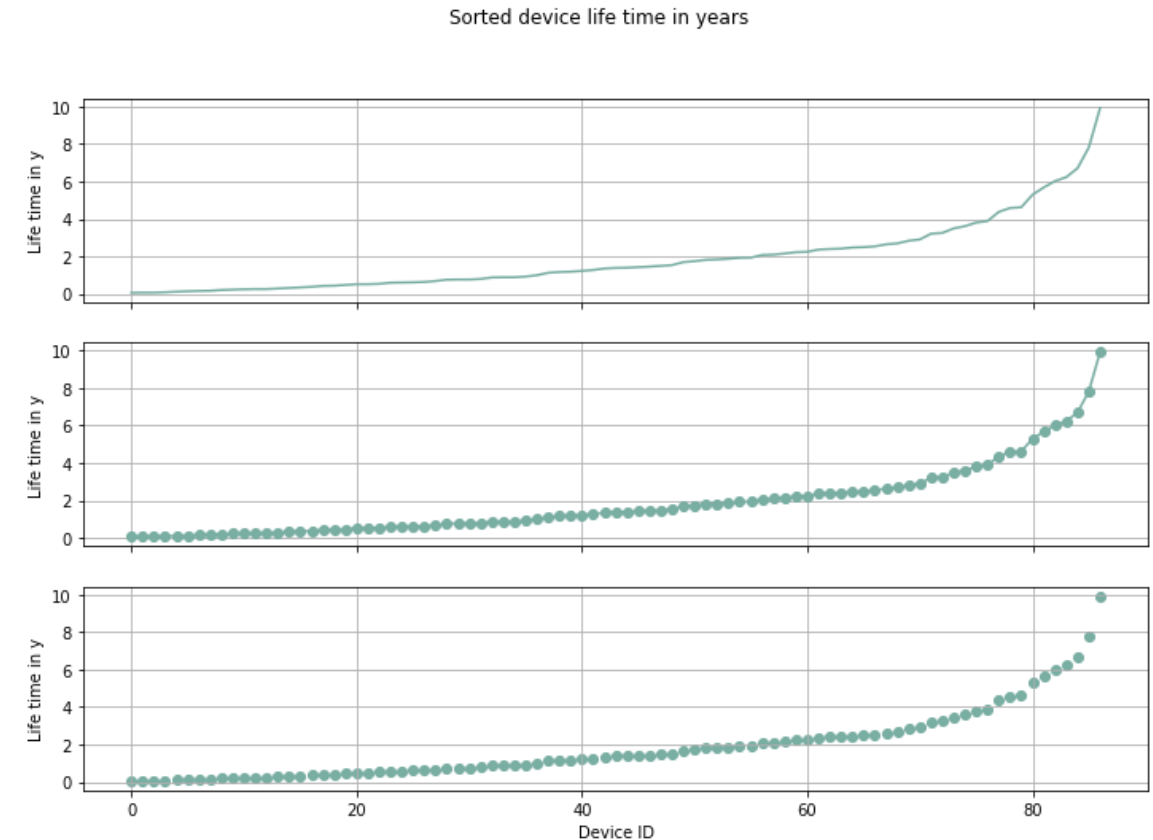


- The data “refrigerator” data set for the consists of  $n=87$  observations
- Contains the lifetime in years for each device
- Sorted in ascending order

```
y = np.array([0.05,0.06,0.06,0.08,0.11,0.13,0.15,0.16,0.20,  
             0.22,0.24,0.25,0.25,0.28,0.31,0.34,0.37,0.42,  
             0.43,0.47,0.51,0.51,0.53,0.59,0.60,0.61,0.63,  
             0.68,0.75,0.76,0.76,0.79,0.87,0.88,0.88,0.92,  
             0.99,1.12,1.16,1.18,1.22,1.27,1.35,1.38,1.39,  
             1.42,1.45,1.49,1.53,1.69,1.74,1.81,1.83,1.87,  
             1.92,1.93,2.07,2.09,2.15,2.22,2.24,2.36,2.39,  
             2.41,2.47,2.49,2.53,2.64,2.69,2.83,2.90,3.21,  
             3.25,3.49,3.61,3.80,3.88,4.37,4.58,4.62,5.29,  
             5.68,6.02,6.23,6.71,7.82,9.93])  
x = np.arange(y.size)
```



```
fig, ax = plt.subplots(3, figsize=(12,8), sharex=True)  
ax[0].plot(x,y)  
ax[1].plot(x,y, marker='o')  
ax[2].scatter(x,y)
```



## Scatter plot

- A scatter plot is a graphical representation of **observed pairs of values of  $n$  statistical characteristics**
- These pairs of values are plotted in a Cartesian coordinate system

```
fig, ax = plt.subplots(1, figsize=(8,8))

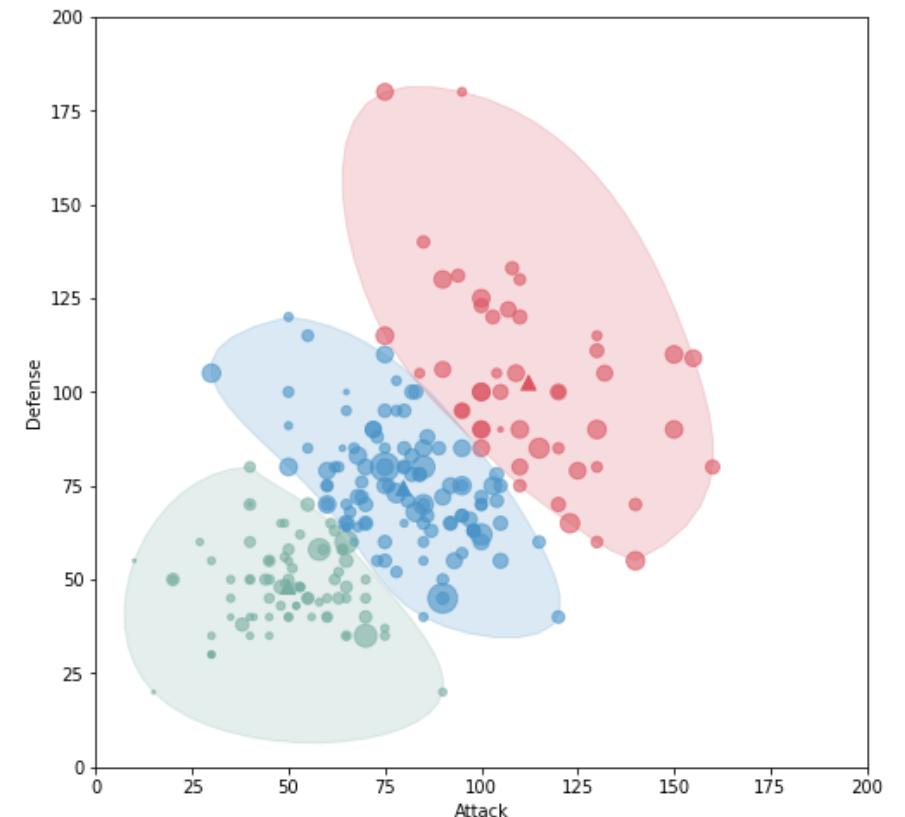
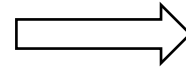
s = [s**2/100 for s in df['HP']]

plt.scatter(df.Attack, df.Defense, c=df.c, alpha=0.6, s=s)
plt.scatter(cen_x, cen_y, marker='^', c=colors, s=70)

for i in df.cluster.unique():
    # get the convex hull
    points = df[df.cluster == i][['Attack', 'Defense']].values
    hull = ConvexHull(points)
    x_hull = np.append(points[hull.vertices,0],
                       points[hull.vertices,0][0])
    y_hull = np.append(points[hull.vertices,1],
                       points[hull.vertices,1][0])

    # interpolate
    dist = np.sqrt((x_hull[:-1] - x_hull[1:])**2 + (y_hull[:-1] -
    y_hull[1:])**2)
    dist_along = np.concatenate(([0], dist.cumsum()))
    spline, u = interpolate.splprep([x_hull, y_hull],
                                   u=dist_along, s=0)
    interp_d = np.linspace(dist_along[0], dist_along[-1], 50)
    interp_x, interp_y = interpolate.splev(interp_d, spline)

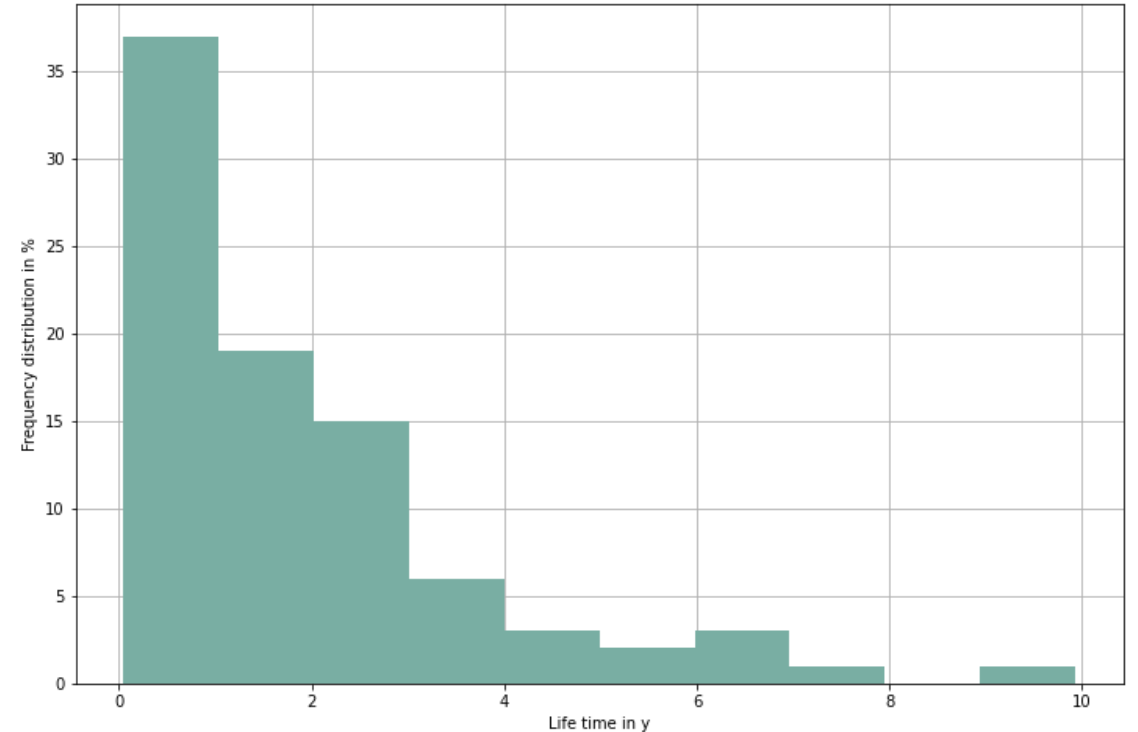
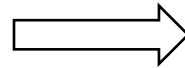
    # plot shape
    plt.fill(interp_x, interp_y, '--', c=colors[i], alpha=0.2)
```



## Histogram

- This representation (area-proportional) is useful when the values are divided into different categories
- The bar areas are therefore plotted **proportionally** to the **relative frequencies** on the ordinate, so that the **sum of the partial areas equals one**
- The number of bins depends on your specific problem

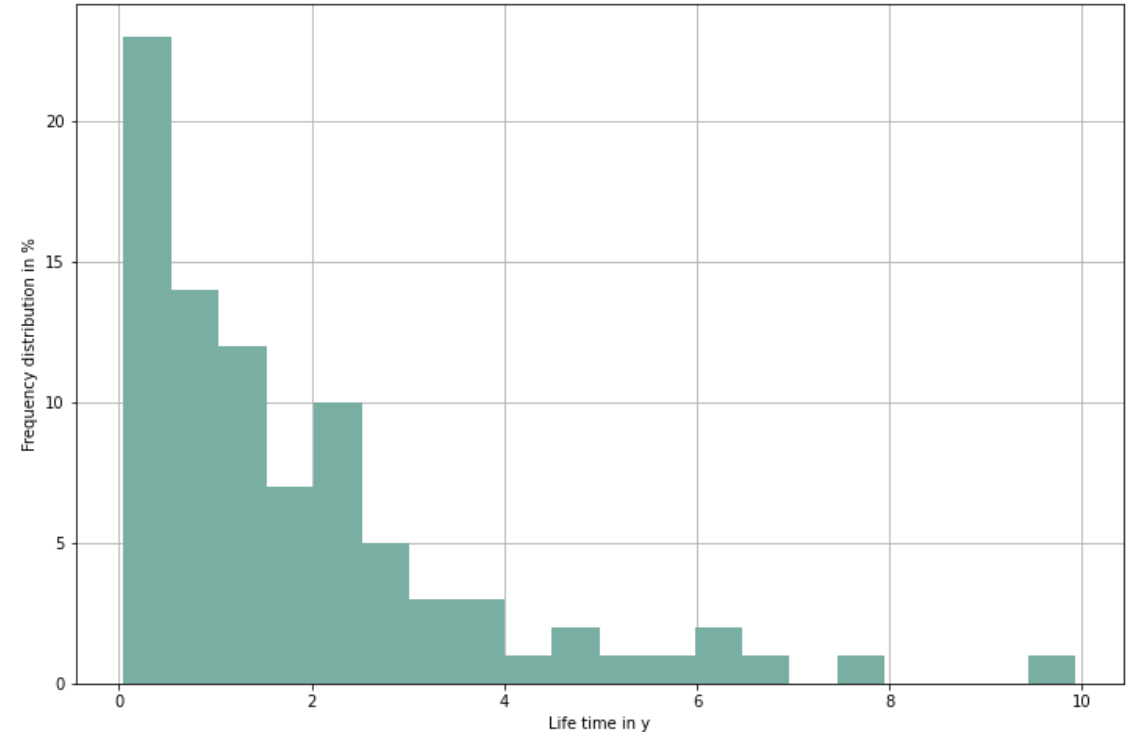
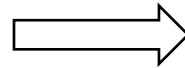
```
fig = plt.figure(figsize=(12,8))
plt.hist(y, bins=10, zorder=10)
plt.grid()
plt.ylabel('Frequency distribution in %')
plt.xlabel('Life time in y')
fig.savefig('cooling_hist.png', bbox_inches='tight',
transparent=True)
```



## Histogram

- Choose the number of partial areas (`bins`, or `nbins`) specified to your actual problem
- The difference between `bins=10` and `bins=20` shows more details in the data distribution

```
fig = plt.figure(figsize=(12,8))
plt.hist(y, bins=20, zorder=10)
plt.grid()
plt.ylabel('Frequency distribution in %')
plt.xlabel('Life time in y')
fig.savefig('cooling_hist.png', bbox_inches='tight',
transparent=True)
```



## Histogram

- How do you find an **approximate** value for the number of bins for large data sets in an analytical way?

### Sturge's Rule:

$$K = 1 + 3.22 \log_n$$

$$K = 1 + 3.22 \log(87) = 7.24 \approx 7$$

Sturge's rule works best for continuous data that is **normally distributed** and **symmetric**

### Rice's Rule:

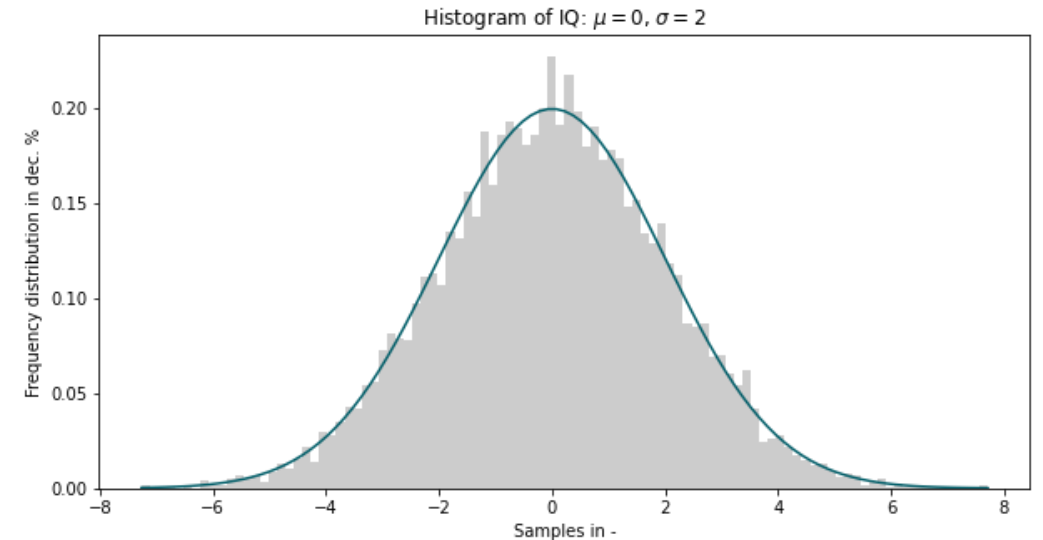
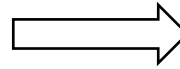
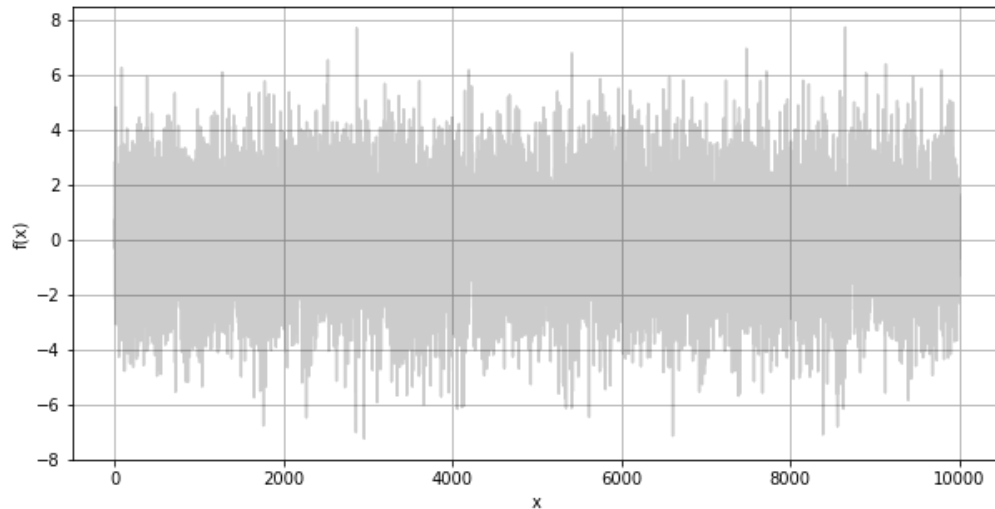
$$K = 2\sqrt[3]{n}$$

$$K = 2\sqrt[3]{87} = 8.86 \approx 9$$

- Histogram should contain all data including outliers
- Choose between 5 and 20 bins (rule of thumb)
- The larger the data set, the larger the number of bins
- The exact number of bins is usually a judgment call

## Objective statistic parameters

- Further information about the data are often hidden
- Statistical parameters such as the mean, standard deviation can be extracted from the data



- **Gaussian normal distribution**  $X \sim \mathcal{N}(\mu, \sigma)$  describes the expected value ( $\mu$ ) and standard deviation  $\sigma^2$
- We will use the Gaussian normal distribution for the classification process in the latter

## Nomenclature

- Scalar values:  $x \in \mathbb{R}^1$

$$x = 1$$

- Vectors:  $\mathbf{x} \in \mathbb{R}^{1 \times n}$

$$\mathbf{x} = [1, 2, 3, 4]$$

- Matrices:  $\mathbf{X} \in \mathbb{R}^{m \times n}$

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

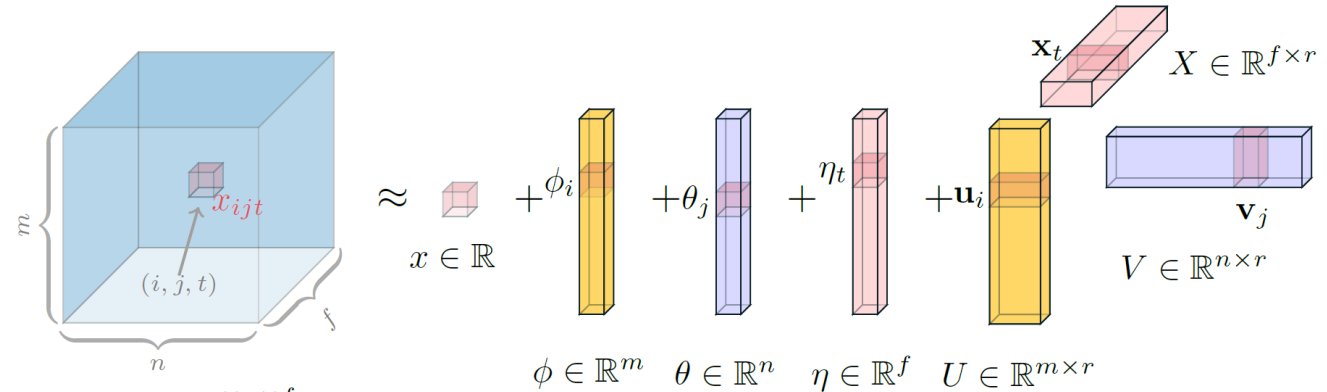
- Tensors:  $\mathcal{X} \in \mathbb{R}^{m \times n \times f}$

$$\mathcal{X} = \begin{bmatrix} [1 & 2] & [3 & 4] \\ [5 & 6] & [7 & 8] \end{bmatrix}$$

- Classes:  $\mathbf{C} = \{c_i, c_r\} \ i \in \mathbb{R}^{N-1}$  with  $c_r = \text{rejection class}$

$$N := \{\text{cats}, \text{dogs}, c_r\}$$

- Classification process:  $\Theta : \mathbf{X} \rightarrow \mathbf{C}$





### Arithmetic mean

- The arithmetic mean ( $\bar{x}$ ) is a positional measure that calculates the center, the centroid, of (ordered) metric data<sup>1</sup>
- The calculation of the arithmetic mean requires distance information of the data, which is only available for metric data
- It is the most frequently used positional measure for metric data

Arithmetic mean:

$$\bar{x} = \frac{1}{n} \sum_{i=0}^{n-1} x_i$$

- The arithmetic mean can also be used to assess the symmetry or skewness of a distribution

<sup>1</sup>Numerical values that are interval-scaled are referred to as metric values. For these, the distance between all values is always the same - so there is just as much distance between 1 and 2 as between 42 and 43



Is the mean meaningful?

Example 1: Size of 4 randomly selected persons

$$x = \{1.89, 1.92, 1.78, 1.82\}$$

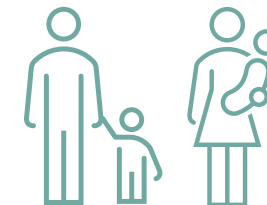
$$\bar{x} = \frac{1}{4} \sum_{i=0}^3 x_i = \frac{7.41}{4} = 1.825$$



Example 2: A survey of 8 people about the number of their children

$$x = \{3, 0, 2, 2, 1, 3, 1, 1\}$$

$$\bar{x} = \frac{1}{8} \sum_{i=0}^7 x_i = \frac{13}{8} = 1.625$$



### Harmonic mean

- The harmonic mean ( $\bar{x}_h$ ) is used, for instance, when you want to weight integers harmonically
- It is also important to indicate the method of determination, so that others know which mean value was calculated

Harmonic mean for numbers:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

- Example: Averaging two numbers with a difference of 20

$$x = \{5, 25\}$$

$$\bar{x} = \frac{30}{2} = 15$$

$$\bar{x}_h = \frac{2}{\frac{1}{5} + \frac{1}{25}} = \frac{2}{\frac{6}{25}} = \frac{25}{3} \approx 8.33$$

### Weighted Harmonic mean

- The weighted harmonic mean ( $\bar{x}_{h,w}$ ) is used if the characteristic expression is described **as a ratio of two units** e. g. €/l or km/l

Weighted harmonic mean:

$$\bar{x}_{h,w} = \frac{\sum_{j=0}^{m-1} x_j n(x_j)}{\sum_{j=0}^{m-1} n(x_j)} = \frac{\sum_{j=0}^{m-1} x_j n(x_j)}{\sum_{j=0}^{m-1} \frac{x_j n(x_j)}{x_j}}$$



## Weighted Harmonic mean

- Example: Assume that a train runs through a route of 600 km once at 60 km/h and once at 120 km/h. Determine the average velocity of the train

$$\bar{x}_{h,w} = \frac{60+120}{2} = 90$$

$$\bar{x}_{h,w} = \frac{\frac{\text{km}}{\text{h}} + \frac{\text{km}}{\text{h}}}{-} = 2 \frac{\text{h}}{\text{km}}$$

$$\bar{x}_{h,w} = \frac{10 \text{ h} \cdot 60 \frac{\text{km}}{\text{h}} + 5 \text{ h} \cdot 120 \frac{\text{km}}{\text{h}}}{15 \text{ h}} = 80 \frac{\text{km}}{\text{h}}$$



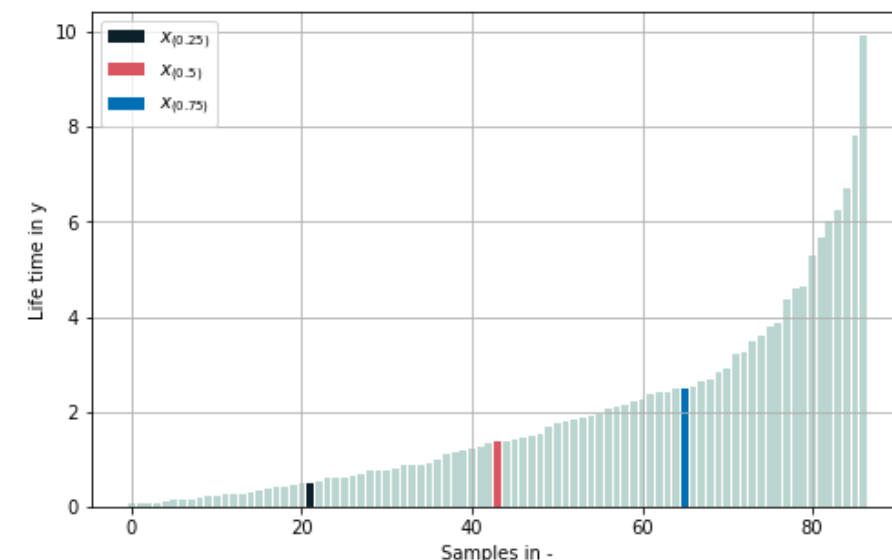
## Quantile

- A quantile is a measure of position in statistics (Lagemmaß)
- A certain proportion of the values (of a random sample) is smaller than the quantile, the rest is larger
- For instance, the 25 % quantile ( $p=0.25$ ) is the value for which 25 % of all values are below this value,  $n - p$  values are above this value
- The median  $x_{(0.5)}$  (0.5-quantile) divides a sorted list of values into two equal parts

p-quantile:

$$x_{(p)} = \begin{cases} \frac{1}{2} (x_{| (np)} + x_{| (np+1)}) & \text{if } (np) \text{ even} \\ x_{| (\lfloor np \rfloor + 1)} & \text{else} \end{cases}$$

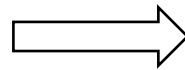
$\lfloor \cdot \rfloor$  means that we always complete the value between the parenthesis, no matter how close it is to the next highest value (floor)



### Mode

- The mode is also a location parameter
- It is defined as the **most frequent value occurring in the data set** and **can be ambiguous**
- For example, if you rank 10 students, the mode corresponds to the number of most frequently achieved grades

$$x = \{2, 1, 4, 1, 1, 4, 3, 5, 2, 5\}$$



Grade	Occurrence
1	3
2	2
3	1
4	2
5	2



## Mode on classified data

- If the data are classified, we can determine the mode by **using the frequency densities**  $d$  of the classes itself
- Thus, we need to determine the densities using the statistic parameters given in the data set
- Using the **class width**  $w_c$  to normalize the spreading
- **Modal class**  $x_M$  is the class with the highest density

Mode:

$$x_{\text{mod}} = x_M^l + \frac{d_M - d_{M-1}}{2d_M - d_{M-1} - d_{M+1}} (x_M^u - x_M^l)$$

Exam Points	Grade	Absolute frequency ( $n_c$ )
0–20	5	57
20–30	4	93
30–37	3	92
37–46	2	29
46–51	1	3
$\Sigma$		274

Relative frequency:

$$f_{r,c} = \frac{n_c}{n}$$

Frequency density:

$$d_c = \frac{f_{r,c}}{w_c}$$

Class width:

$$w_c = x_c^u - x_c^l$$





## Mode on classified data

1. Determine the relative frequencies for each class

$$f_{r,5} = \frac{n_c}{n} = \frac{57}{274} = 0.208$$

2. Determine the class width for each class

$$w_5 = x_5^u - x_5^l = 20 - 0 = 20$$

3. Determine the frequency densities

$$d_c = \frac{f_{r,5}}{w_5} = \frac{0.208}{20} = 0.010$$

4. Find the modal class

$$x_M = \max d_c = \mathbf{3}$$

5. Determine the mode

$$x_{\text{mod}} = 30 + \frac{0.048 - 0.034}{2 * 0.048 - 0.034 - 0.012} * (37 - 30) = 31.96$$

Exam Points	Grade	Abs frequency ( $n_c$ )	Rel. frequency ( $f_{r,c}$ )	Frequency densities $d_c$
0-20	5	57	0.208	0.010
20-30	4	93	0.339	0.034
30-37	3	92	0.336	0.048
37-46	2	29	0.106	0.012
46-51	1	3	0.011	0.020
$\Sigma$		274	1.000	

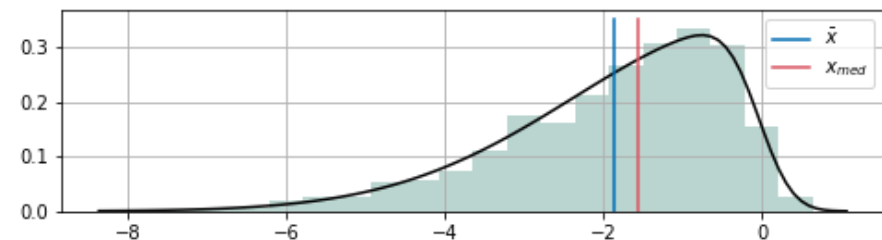
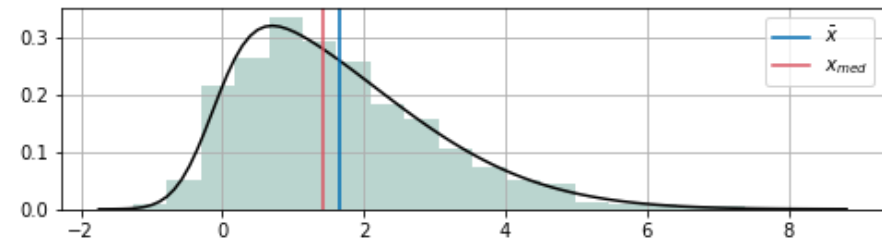
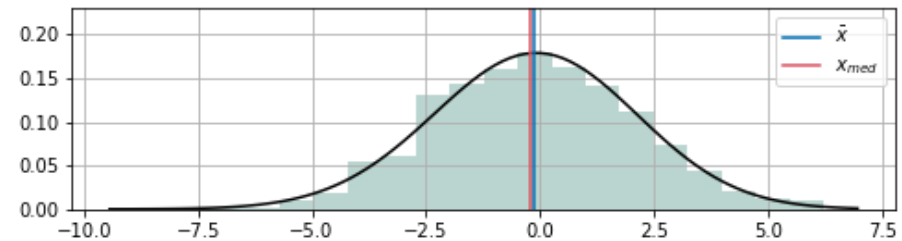
### Skewness of a distribution

- For metric characteristics, the arithmetic mean, median and mode can also be used to assess the symmetry or skewness of a distribution

Pearson Mode skewness:

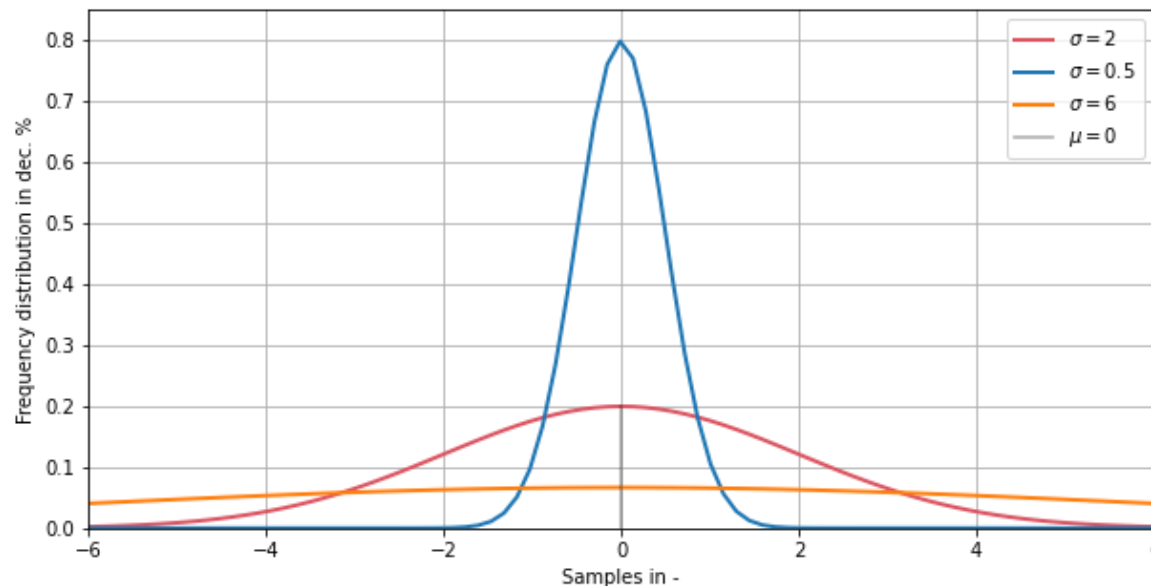
$$s = \frac{\bar{x} - x_{\text{mod}}}{\sigma}$$

- Symmetrical distribution  $\bar{x} = x_{(0.5)}$
- Left side, positive skew  $\bar{x} > x_{(0.5)}$
- Right side, negative skew  $\bar{x} < x_{(0.5)}$



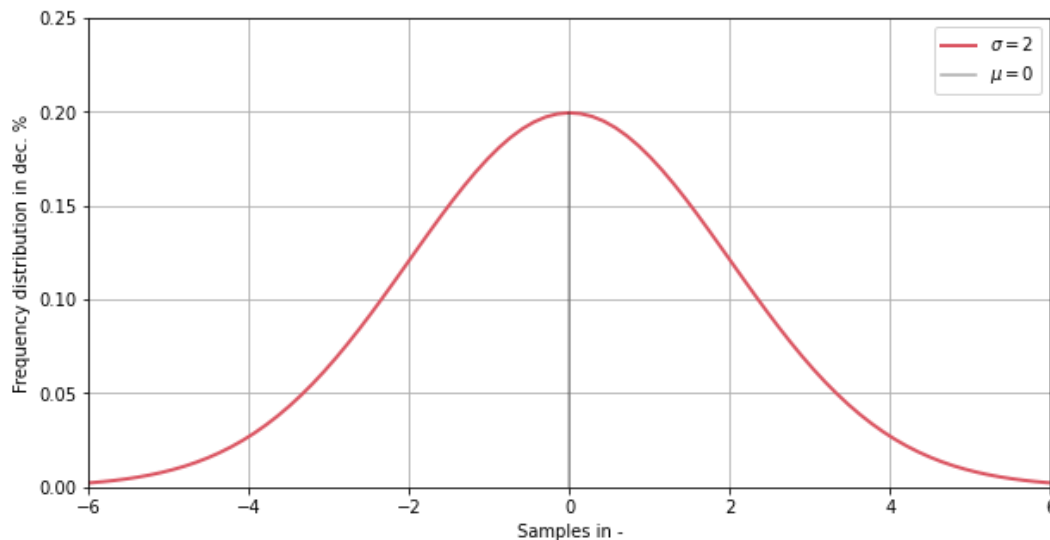
### Measures of dispersion

- **Standard deviation**, **variance**, and **range** are among the measures of dispersion in descriptive statistics
- They are also called **measures of dispersion** and are used to describe the dispersion of values of a sample around a position parameter
- They are used to describe how much a data set (sample) fluctuates around a mean value



### Standard deviation

- The standard deviation ( $\sigma$ ) indicates the spread of a variable around its mean value
- Thus, the standard deviation indicates how much the individual values scatter around the mean value



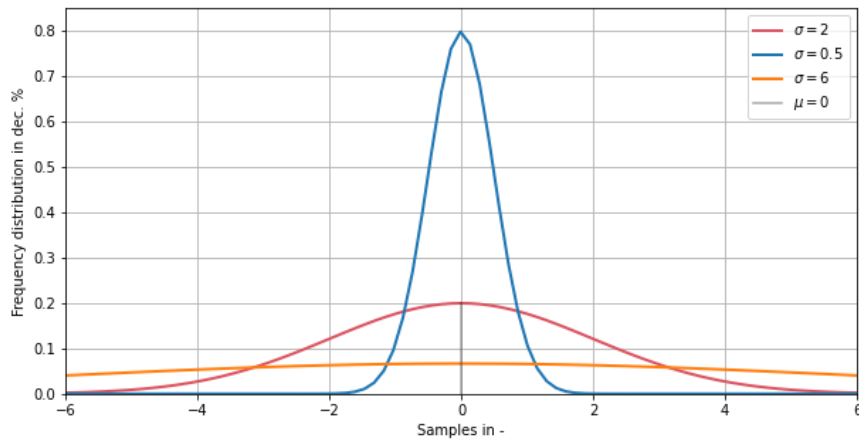
#### Standard deviation:

$$\sigma_s = \sqrt{\frac{1}{n-1} \sum_{i=0}^{n-1} (x_i - \mu)^2} = s_p$$

$$\sigma_p = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (x_i - \mu)^2}$$

### Variance

- For the mean absolute deviation, the simple absolute deviations are used to measure the dispersion of a given data set
- For the **variance** (`var`), the **squared deviations are used**, thus larger distances to the mean value are more strongly taken into account in this way



#### Variance:

$$\sigma_s^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \mu)^2$$



The difference between variance and standard deviation is that std measures the average distance from the mean and the var measures the squared average distance from the mean

### What's about probabilities?

- **probability density functions** (pdf,  $f_x$ ) are used for the construction of **probability distributions** by using integrals
- They are used for the investigation and **classification of probability distributions**
- The **expected value**<sup>1)</sup> ( $E(\cdot)$ ) of a random variable  $\mathbf{x}$  with pdf  $f_x$  is given by

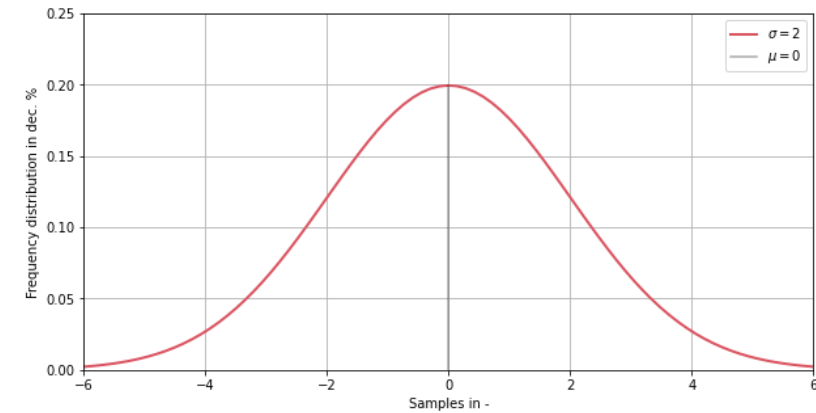
$$E(\mathbf{x}) = \int_{-\infty}^{+\infty} x f_{\mathbf{x}}(x) dx$$

- Thus, the **expected value** ( $\mu$ ) of the random variable is given as

$$\mu = E(\mathbf{x})$$

- Then the variance of the random variable is given by

$$\text{Var}(\mathbf{x}) = E((\mathbf{x} - \mu)^2) = \int_{-\infty}^{+\infty} (\mathbf{x} - \mu)^2 f_X(\mathbf{x}) d\mathbf{x}.$$



<sup>1)</sup>The expected value ( $E(x)$ ) of a random variable describes the number that the random variable takes on average

### Gaussian normal distribution (1D vs. 2D)

- Density function 1-dimensional:  $X \sim \mathcal{N}(\mu, \sigma) \rightarrow p(x; \mu, \sigma^2)$

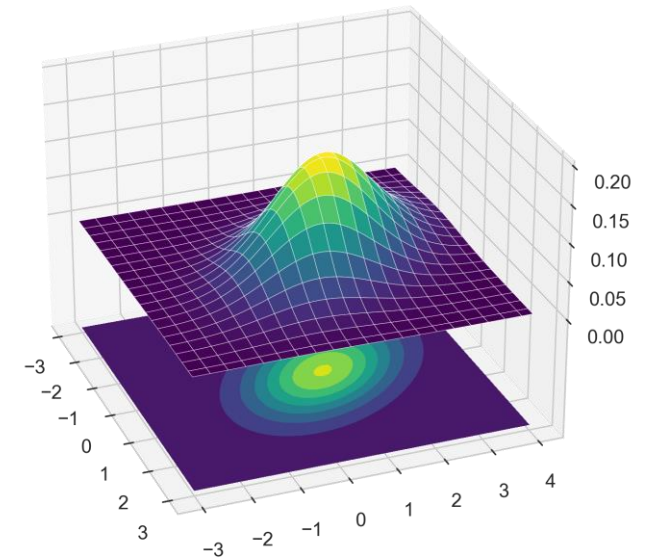
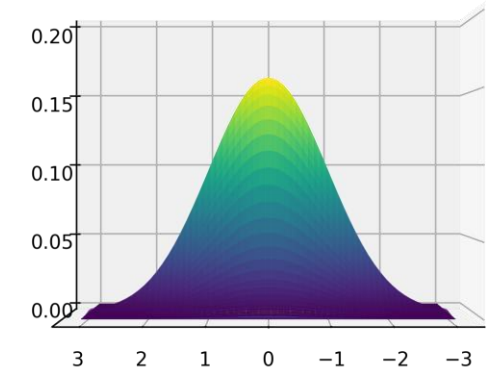
Gaussian distribution:

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

- Density function  $n$ -dimensional:  $X \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow p(\mathbf{x}; \boldsymbol{\mu}; \boldsymbol{\Sigma})$

Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$



### Covariance

- The **covariance matrix** and the **expected value** vector are the **most important parameters of a probability distribution**
- The expected value ( $E(\mathbf{x})$ ) of a random variable describes the number that the random variable takes on average
- The covariance matrix as a matrix of **all pairwise covariances** of the elements of the random vector  $\mathbf{x}$  contains information about its dispersion and about correlations between its components

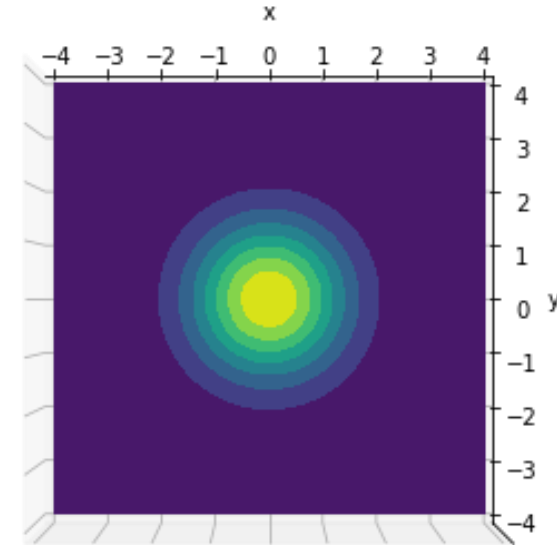


### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite
- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 = \sigma_y^2 = \sigma^2$

$$\mu = [0, 0]^T$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} = \mathbf{Id}\sigma^2$$

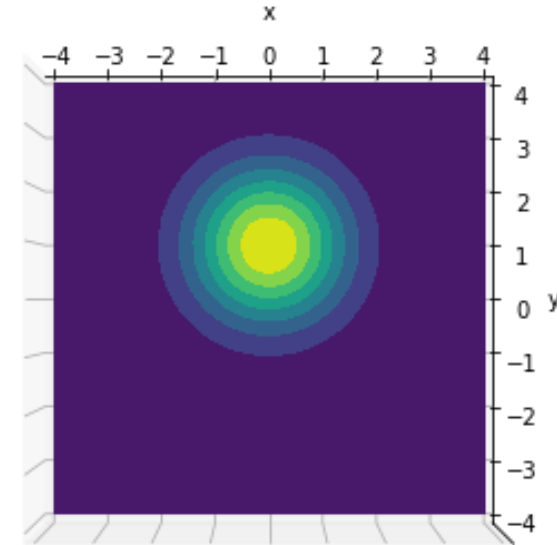


### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite
- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 = \sigma_y^2 = \sigma^2$

$$\mu = [0, 1]^T$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} = \mathbf{Id}\sigma^2$$



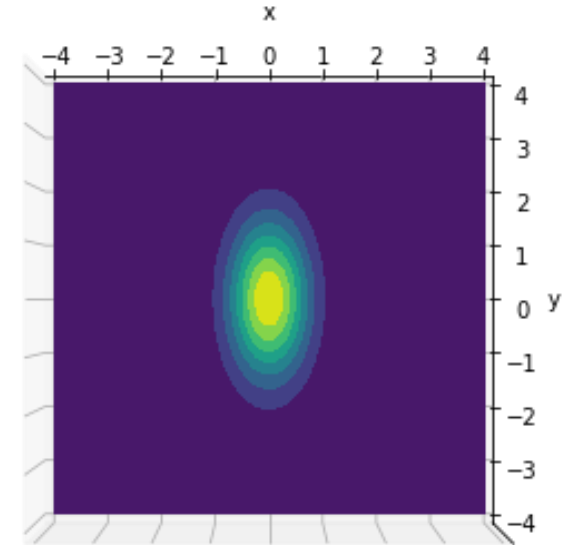
### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite

- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 \neq \sigma_y^2$

$$\mu = [0, 0]^T$$

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 1 \end{bmatrix}$$

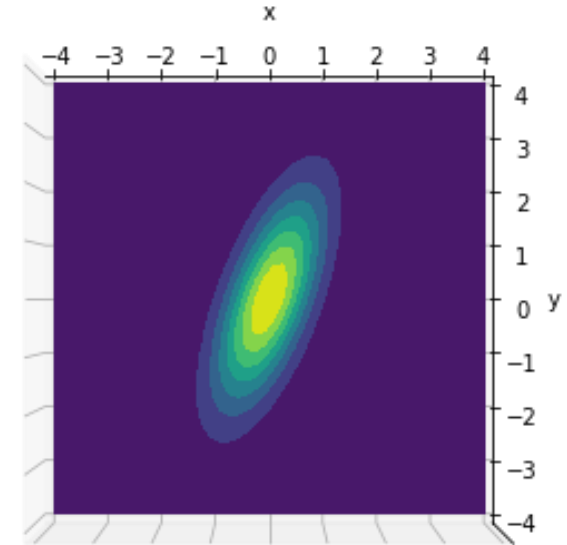


### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite
- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 \neq \sigma_y^2$

$$\mu = [0, 0]^T$$

$$\Sigma = \begin{bmatrix} 0.5^2 & 0.8^2 \\ 0 & 1 \end{bmatrix}$$

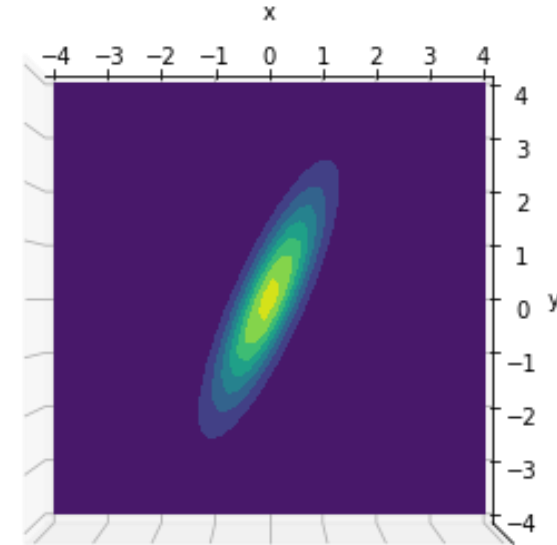


### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite
- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 \neq \sigma_y^2$

$$\mu = [0, 0]^T$$

$$\Sigma = \begin{bmatrix} 0.5^2 & 0.8^2 \\ 0.44^2 & 1 \end{bmatrix}$$

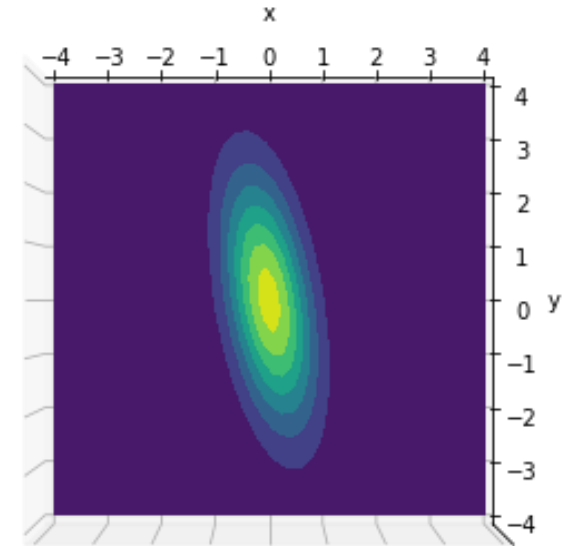


### Covariance

- The covariance (cov,  $\Sigma$ ) is in probability theory and statistics a square matrix giving the **covariance between each pair of elements** of a given random vector  $\mathbf{x}$
- The matrix is in any case **symmetric** and positive semi-definite
- Example: 2-dimensional gaussian distribution with  $\sigma_x^2 \neq \sigma_y^2$

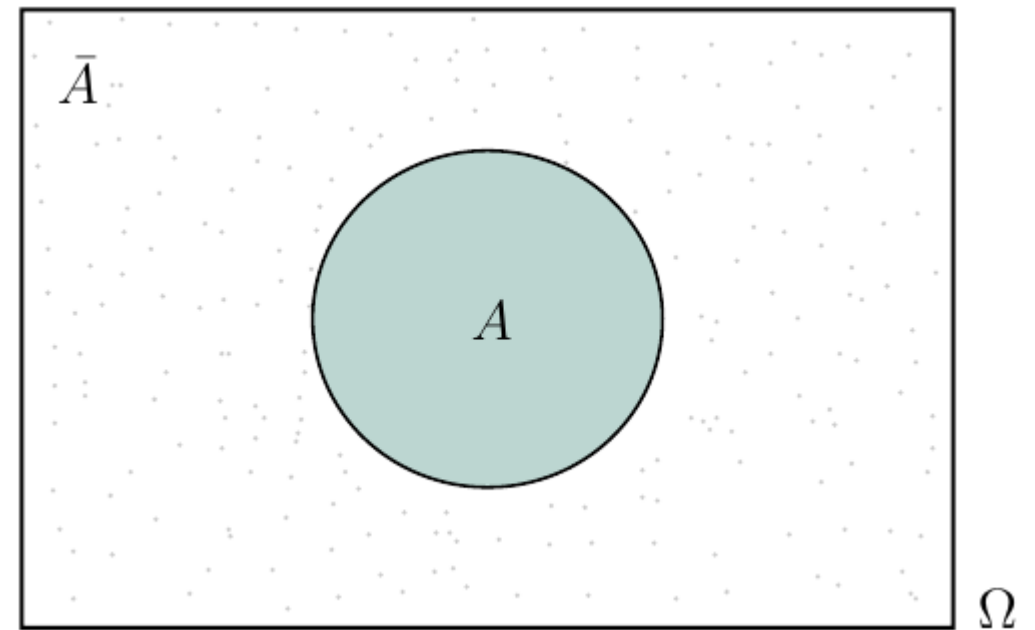
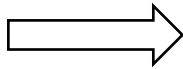
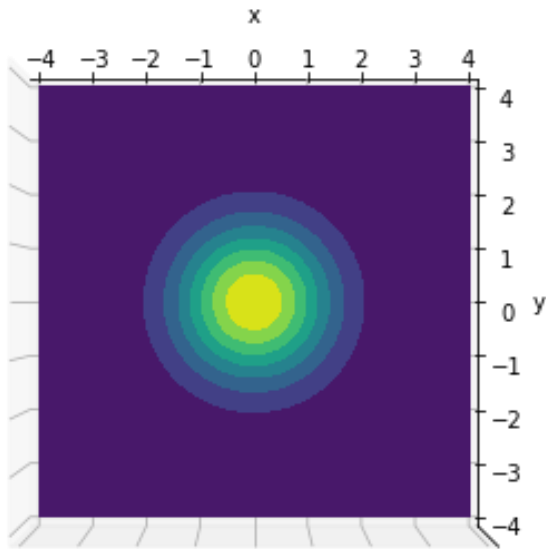
$$\mu = [0, 0]^T$$

$$\Sigma = \begin{bmatrix} 0.5^2 & 0 \\ -(0.77^2) & 2 \end{bmatrix}$$



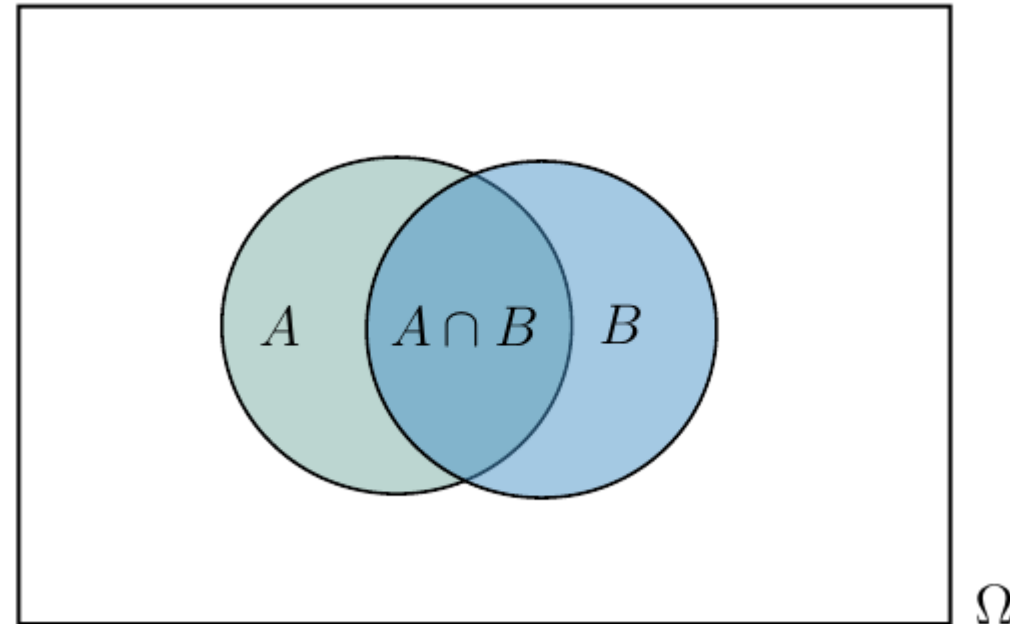
## L02.3 Sets and Spaces

- Events  $(A, \bar{A})$  are subsets of the result set  $\Omega$  (analogy to the entire space in the given figure)
- The event  $A$  **complementary** to an event  $\bar{A}$  occurs exactly when  $A$  does not occur



## L02.3 Sets and Spaces

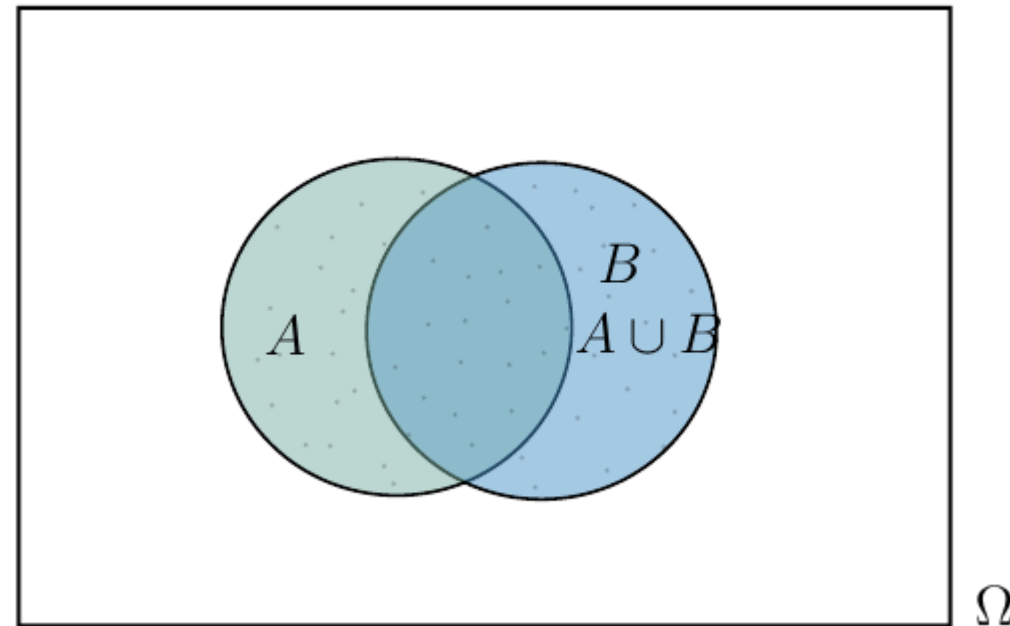
- The **average** of two events  $A \cap B$  consists of all elementary events, that belong to both  $A$  and  $B$
- The average event  $A \cap B$  occurs when  $A$  **and**  $B$  occur together





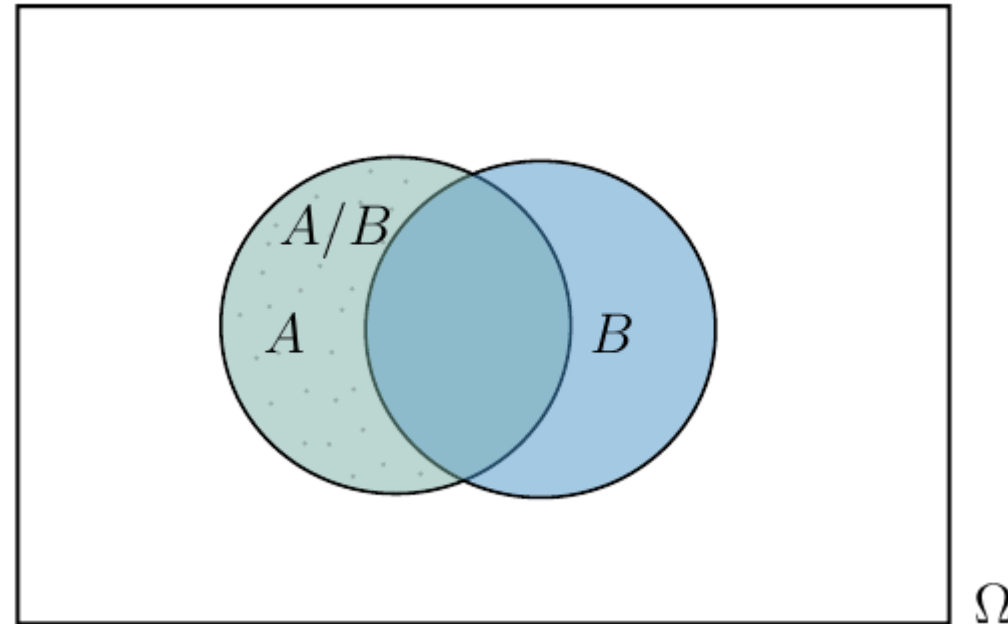
## L02.3 Sets and Spaces

- The **unification** of two events  $A \cup B$  includes all elementary events belonging to A **or** to B
- The unification event  $A \cup B$  occurs when A or B occurs



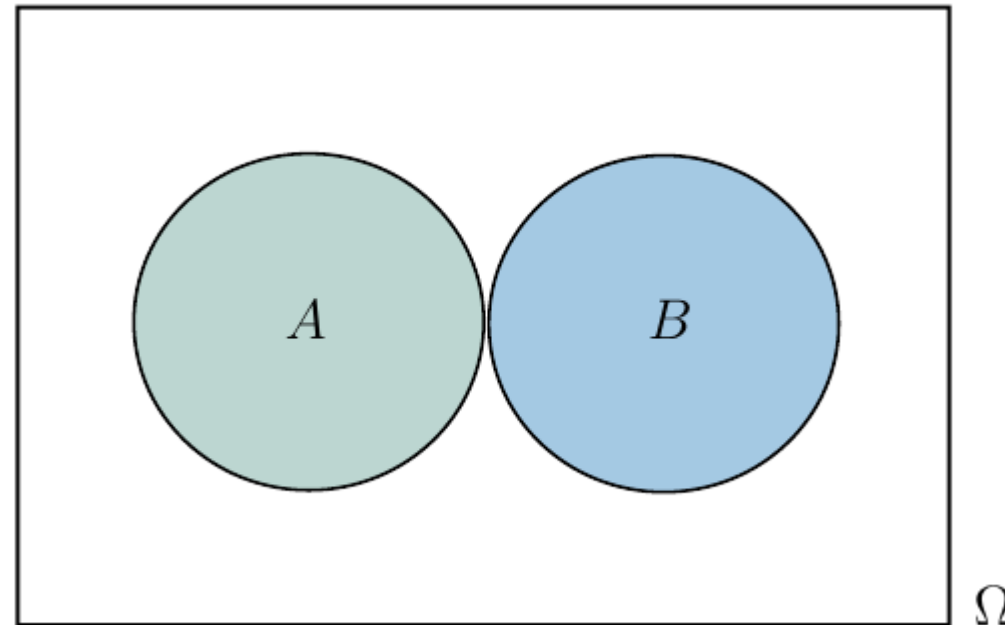
## L02.3 Sets and Spaces

- The **difference** of two events  $A/B$  consists of all elementary events belonging to  $A$  **but not** to  $B$
- The difference event  $A/B$  occurs exactly if  $A$  but not  $B$  occurs.



## L02.3 Sets and Spaces

- Two events  $A$  and  $B$  are called **incompatible** or disjoint if there is no elementary event leading to the occurrence of both  $A$  and  $B$
- The **disjoint** events  $A$  and  $B$  have no elementary event in common, their intersection is empty:  $A \cap B = \emptyset$



### Mathematics

#### Commutative law:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

#### Associative law:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

#### Distributive law:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

### Representation

- It is possible to describe each event by set-theoretic operations and represent it in  $\Omega$
- The resulting set  $\Omega$  does not contain itself and the empty set  $\emptyset$
- In order to be able to assign a probability to each event later, a set system (so-called **event algebra**,  $\mathcal{A}$ ) must be created, in which **all possible subsets, also  $\emptyset$  and  $\Omega$  are contained**  $\mathcal{A} = \{\Omega, \emptyset, c_i\}$
- $\mathcal{A}$  differs from the result set  $\Omega$  in that, that  $\mathcal{A}$  consists of subsets of  $\Omega$  and not of its elements



$\mathcal{A}$  is a set of events, which contains for any number of events also their complements, averages and unions

### Power set (Potenzmenge)

- A power set  $\mathcal{P}(A)$  of the set  $A$  **is the set of all subsets**
- If we know how many subsets we have, we know (for finite events) the maximal number of possible events

Power set<sup>1</sup>:

$$|\mathcal{P}(A)| = 2^n$$

**?** How many possible solutions are available for the algebra  $\mathcal{A} = \{1, 2, 3\}$ ?

$$|\mathcal{P}(A)| = 2^3 = 8$$

$$\mathcal{P}(A) = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{\Omega\}, \{\emptyset\}\}$$

<sup>1</sup> without proof, taken from [\[De2005\]](#)

### Limits

- In case of infinitely many discrete events and in the continuous case, the result set is not finite, although countable (countably infinite)
- The event algebra (and thus power set) is then not finite, since it contains infinitely many subsets
- One makes do with constructing an event- $\sigma$ -algebra which does not contain all subsets of  $\Omega$ , but only the part necessary for the consideration of the random experiment. so that all basic set operations can be performed. can be performed

$$\mathcal{A} = \{\Omega, \emptyset, c_i\}$$



$\mathcal{A}$  is a set of events, which contains for any number of events also their complements, averages and unions

### Probability space

- Build mathematical framework to **represent and analyze phenomena in data** and/or experiment
- To quantify how likely it is that the outcome of the experiment belongs to a particular set of outcomes (events), we assign a probability (or a **measure**) to the event
- We can characterize the experiment by constructing a **probability space**  $(\Omega, \mathcal{A}, P)$  consists of:
  - **Sample space**  $\Omega$ : contains all possible solutions/outcomes

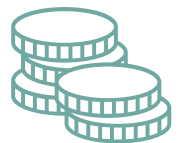
$$\Omega := \{\text{heads}, \text{tails}\}$$

- **Set of events**  $\mathcal{A}$ :  $\sigma$ -algebra assumption

$$\mathcal{A} := \{\{\text{heads or tails}\}, \{\text{heads}\}, \{\text{tails}\}, \{\emptyset\}\}$$

- **Probability measure**  $P$ : assigns probabilities to the events

$$P(A \cap B) = P(A)P(B)$$



### Conditional probability

- Conditional probability allows us to **update probabilistic models when additional information are available**
- Given a probability space  $(\Omega, \mathcal{A}, P)$  where the outcome of the experiment belongs to a certain event  $S \in \mathcal{A}$
- This outcome affects how likely it is for any other event  $S' \in \mathcal{A}$
- The updated probability of each event is known as the conditional probability of  $S'$  given  $S$   $P(S'|S)$

$$\begin{aligned} P(S'|S) &= \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} \Big| \cdot \frac{\text{total}}{\text{total}} \\ &= \frac{\text{outcomes in } S' \text{ and total } S}{\text{total outcomes in } S} \end{aligned}$$

Conditional probability:

$$P(S'|S) = \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} = \frac{P(S' \cap S)}{P(S)}$$





### Conditional probability

$$P(S'|S) = \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} = \frac{P(S' \cap S)}{P(S)}$$

- German train company provides data set with 101 samples about “air conditioning (AC) performance during the seasons”
- The company is interested in the probability that an AC failure occurs in summer
- Using the conditional probability

$$\mathbf{x} = \begin{cases} \{P(AC, \text{Winter}) = \frac{41}{101}\}, \\ \{P(AC, \text{Summer}) = \frac{5}{101}\}, \\ \{P(\overline{AC}, \text{Winter}) = \frac{2}{101}\}, \\ \{P(\overline{AC}, \text{Summer}) = \frac{53}{101}\} \end{cases}$$

$$P(\overline{AC}, \text{Summer}) = \frac{\frac{53}{101}}{\frac{5}{101} + \frac{53}{101}} = 0.914$$

$$P(\overline{AC}, \text{Winter}) = \frac{\frac{2}{101}}{\frac{2}{101} + \frac{41}{101}} = 0.046$$

### Conditional probability via Contingency table

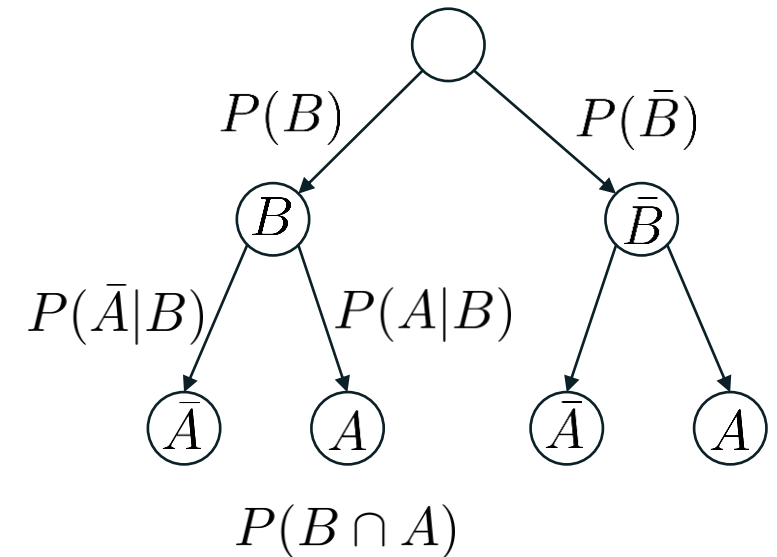
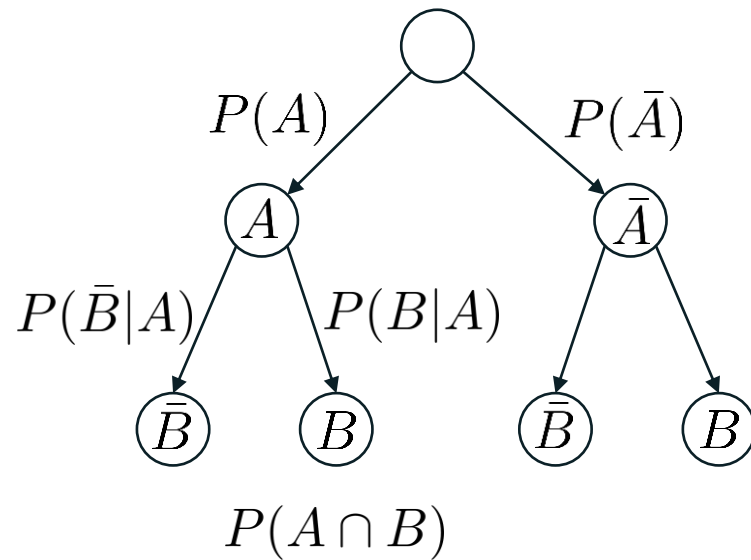
- The event algebra  $\mathcal{A}$  of two events  $A$  and  $B$  can be decomposed into the four subsets
  1.  $P(A \cap B)$
  2.  $P(\bar{A} \cap B)$
  3.  $P(A \cap \bar{B})$
  4.  $P(\bar{A} \cap \bar{B})$
- Each outcome belongs to exactly one subset
- The absolute frequencies or the probabilities of the events are the inner cells of a four-cell table

	$A$	$\bar{A}$	
$B$	$P(A \cap B)$	$P(\bar{A} \cap B)$	$P(B)$
$\bar{B}$	$P(A \cap \bar{B})$	$P(\bar{A} \cap \bar{B})$	$P(\bar{B})$
	$P(A)$	$P(\bar{A})$	

### Conditional probability

- Conditional probabilities can be used to determine the intersection of several events in a structured way
- We can express the probability of the intersection of two events  $\{A, B\} \in \mathcal{A}$  as:

$$\begin{aligned}P(A \cap B) &= P(B \cap A) \\P(A)P(B|A) &= P(B)P(A|B) \\P(B|A) &= \frac{P(B)P(A|B)}{P(A)}\end{aligned}$$



### Bayes' Theorem

- Thomas Bayes (1702-1761) dealt with the problem of the relationship between  $P(A | B)$  and the inverse probability  $P(B | A)$
- One assumes the known value  $P(B | A)$ , but is interested in the value  $P(A | B)$
- The Bayes' Theorem goes back to the definition of conditional probabilities

- **a priori** probability: The probability that a hypothesis is true, before any evidence is available (initial guess)
- **conditional** probability: The probability that a certain event will occur after another event has already occurred
- **a-posteriori** probability: The probability that a hypothesis is true, after the occurrence of a certain event has been considered

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Likelihood**

Probability for B under condition that A has occurred (test detects feature correctly)

**Prior**

Probability of occurrence of event A (feature was present)

**Marginal**

Probability that B occurs



### COVID test

- Humans have a certain disease (D) or are healthy (H) with probability

$$P(D) = \frac{20}{10^5} = 0.0002 \quad P(H) = 1 - P(D) = 0.9998$$

- A COVID-test (T) can detect the disease with a probability of 95 %

$$P(T|D) = 0.95$$

- False-positive-rate of 1 %

$$P(T|H) = 0.01$$

- What is the probability for an arbitrary human to have the disease given a positive result of the test?

$$P(D|T) = ?$$



### COVID test

- Using Bayes' Theorem to solve the problem

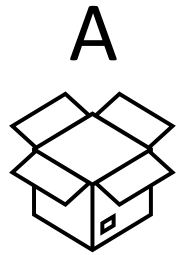
$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D)+P(H)P(T|H)} =$$



$$\frac{0.0002 \cdot 0.95}{0.0002 \cdot 0.95 + 0.9998 \cdot 0.01} = 0.0186 \approx 2\%$$

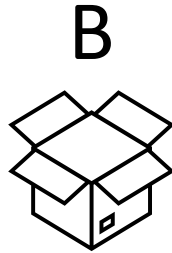
Bayes' Theorem:



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

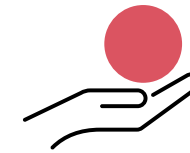
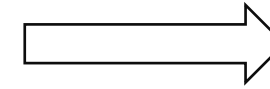
## Bayes' theorem example (urn model)



80 red balls   
20 blue balls 



10 red balls   
90 blue balls 



From where do we get the red ball?  
→ Go with the highest probability!

- $P(A) = P(B) = 0.5$
- $P(\text{red}|A) = 0.8$
- $P(\text{blue}|A) = 0.2$
- $P(\text{red}|B) = 0.1$
- $P(\text{blue}|B) = 0.9$

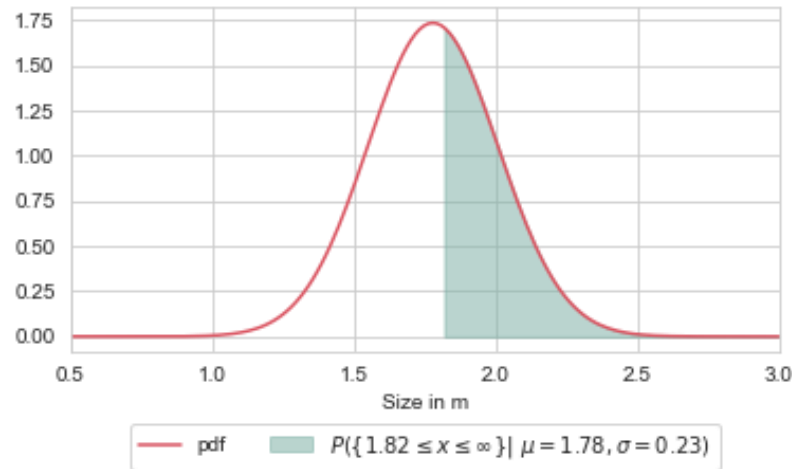
$$P(A|\text{red}) = \frac{P(\text{red}|A)P(A)}{P(\text{red})} = \frac{P(\text{red}|A)P(A)}{P(\text{red}|A)P(A) + P(\text{red}|B)P(B)}$$

$$P(A|\text{red}) = \frac{0.8 \cdot 0.5}{0.8 \cdot 0.5 + 0.1 \cdot 0.5} = \frac{8}{9} \qquad P(B|\text{red}) = \frac{1}{9}$$

## L02.4 Probabilities and Bayes

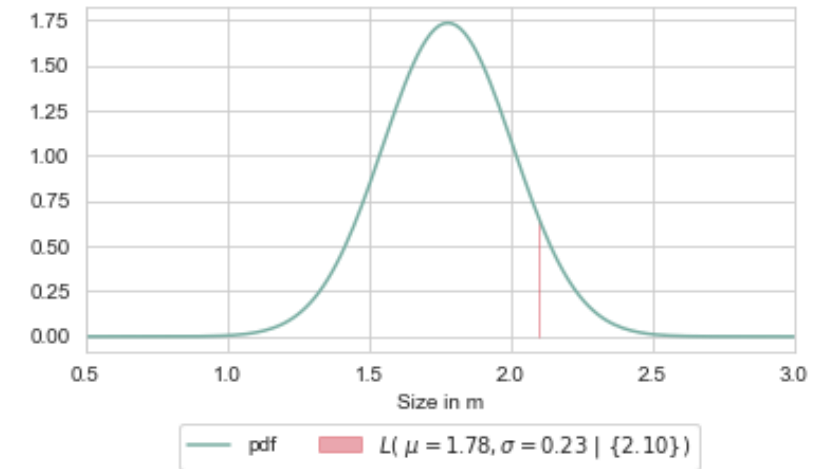
Difference between likelihood and probability  $X \sim \mathcal{N}(1.78, 0.23)$

Probability



The **probability** of meeting a person larger than 1.82 m

Likelihood



The **likelihood** of being a person with 2.10 m



Z-normalization transforms any gaussian distribution to standard gaussian distribution  $X \sim \mathcal{N}(0, 1)$





### Z-Score

1. Determine z-score based on given value, mean and standard deviation (standardization)
2. Go to the table and read the corresponding z-score
3. Determine the resulting probability

Z-normalize:

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

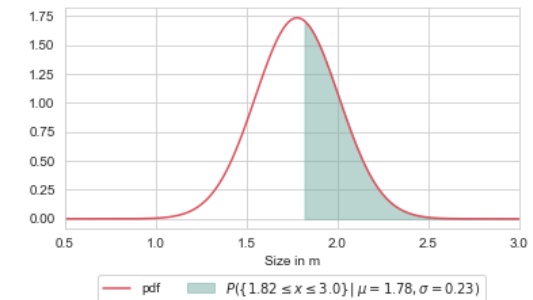
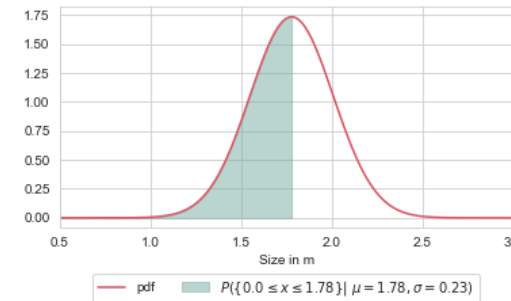
- **Example 1:** Probability to meet a person below 1.78 m

$$Z = \frac{1.78 - 1.78}{0.28} = 0 \rightarrow P(z = 0) = 0.5$$

- **Example 2:** The person is taller than 1.82 m

$$Z = \frac{1.82 - 1.78}{0.28} = 0.14 \rightarrow P(z = 0.14) = 1 - 0.556 = 0.443$$

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987



### Eigenvalues and Eigenvectors

- The multiplication of a matrix  $\mathbf{A}$  with a vector  $\mathbf{v}$  results in a vector again

$$\underbrace{\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}}_{\mathbb{R}^{2 \times 2}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbb{R}^{2 \times 1}} = \underbrace{\begin{bmatrix} A_{1,1}x_1 + A_{1,2}x_2 \\ A_{2,1}x_1 + A_{2,2}x_2 \end{bmatrix}}_{\mathbb{R}^{2 \times 1}}$$

- For square matrices  $\mathbb{R}^{n \times n}$  there are **certain vectors** that can be multiplied by the matrix, so that you get the same vector as a result, only multiplied by a factor ( $\lambda$ )

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \lambda \neq 0$$

- Such a vector is called an Eigenvector  $\mathbf{v}$  and the factor is called the Eigenvalue  $\lambda$  of a matrix and can be found by solving the following equation

$$\begin{aligned}\mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{A}\mathbf{v} - \lambda\mathbf{v} &= \mathbf{0} \\ (\mathbf{A} - \lambda\mathbf{Id})\mathbf{v} &= \mathbf{0}\end{aligned}$$

Eigenvalues:

$$(\mathbf{A} - \lambda\mathbf{Id})\mathbf{v} = \mathbf{0}$$

### Eigenvalues and Eigenvectors

- How to find Eigenvalues and Eigenvectors of the matrix  $\mathbf{X}$ ?
  1. Get characteristic polynomial ( $P$ ) of  $\mathbf{X}$
  2. Zero of  $P$  equals to the eigenvalues  $\lambda$  of  $\mathbf{X}$
  3. Normalize eigenvalues and sort in descending order
  4. Solving  $(\mathbf{A} - \lambda \mathbf{Id}) = \mathbf{0}$  returns eigenvectors  $\mathbf{v}$
  5. Sorting  $\lambda$  and  $\mathbf{v}$  in descending order

Characteristic polynomial:

$$P = |(\mathbf{X} - \lambda \mathbf{Id})| = \det(\mathbf{X} - \lambda \mathbf{Id})$$



## Eigenvalues and Eigenvectors

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$

1. Finding the char. polynomial ( $P$ ) of  $\mathbf{X}$

$$P(\mathbf{X}) = \begin{bmatrix} 1 - \lambda & 2 \\ 4 & 3 - \lambda \end{bmatrix}$$

$$\begin{aligned} \det(P(\mathbf{X})) &= \\ (1 - \lambda)(3 - \lambda) - 4 \cdot 2 &= 0 \\ (1 - \lambda)(3 - \lambda) - 8 &= 0 \\ 3 - \lambda - 3\lambda + \lambda^2 - 8 &= 0 \\ \lambda^2 - 4\lambda - 5 &= 0 \end{aligned}$$

2. Finding the zeros of ( $P(\mathbf{X})$ ). For 2<sup>nd</sup> order polynomials, we use the **pq-equation**, otherwise we have to use the **polynomial division**

$$\lambda_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}$$

$$\lambda_{1,2} = \frac{4}{2} \pm \sqrt{\left(\frac{4}{2}\right)^2 + 5}$$

$$\lambda_1 = 5, \lambda_2 = -1$$



## Eigenvalues and Eigenvectors

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}, \lambda_1 = 5, \lambda_2 = -1$$

1. Finding the Eigenvectors  $\mathbf{v} = [\mathbf{v}(\lambda_1) \ \mathbf{v}(\lambda_2)]^T$  for each Eigenvalue by solving

$$(\mathbf{X} - \lambda \mathbf{Id})\mathbf{v} = \mathbf{0}$$

$$\begin{bmatrix} 1 - \lambda_{1,2} & 2 \\ 4 & 3 - \lambda_{1,2} \end{bmatrix} \underbrace{\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}}_{\mathbf{v}(\lambda_{1,2})} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

2. We get a system of linear equations which we can solve using common tools

$$\begin{array}{l} \lambda_1 \quad \begin{array}{l} \text{(I)} \quad \left\{ \begin{array}{cc|c} -4v_1 & 2v_2 & 0 \\ 4v_1 & -2v_2 & 0 \end{array} \right. \\ \text{(II)} \quad \left\{ \begin{array}{cc|c} -4v_1 & 2v_2 & 0 \\ 4v_1 & -2v_2 & 0 \end{array} \right. \end{array} \\ \{-4v_1 + 2v_2 = 0 \rightarrow v_1 = \frac{1}{2}v_2 \\ \mathbf{v}(\lambda_1) = \begin{bmatrix} \frac{1}{2}v_2 \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} \end{array}$$

$$\begin{array}{l} \lambda_2 \quad \begin{array}{l} \text{(I)} \quad \left\{ \begin{array}{cc|c} 2v_1 & 2v_2 & 0 \\ 4v_1 & 4v_2 & 0 \end{array} \right. \quad | \cdot \frac{1}{2}(\text{I}) \\ \text{(II)} \quad \left\{ \begin{array}{cc|c} 1v_1 & 1v_2 & 0 \\ 4v_1 & 4v_2 & 0 \end{array} \right. \quad | \cdot -4(\text{I}) \end{array} \quad \begin{array}{l} \{v_1 + v_2 = 0 \rightarrow v_2 = -v_1 \\ \mathbf{v}(\lambda_2) = \begin{bmatrix} -v_2 \\ v_2 \end{bmatrix} = v_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{array} \end{array}$$



# Break



HA02.4.iypnb



[www.hs-kempten.de/ifm](http://www.hs-kempten.de/ifm)