Statistics and Sensor Data Fusion

1. Descriptive Statistics

Descriptive Statistics

Goal: Suitable graphical and numerical representation of the data

1.1 Basic Concepts of Data Acquisition

1.2 Analysis of Univariate Data

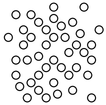
- Elementary Concepts
- Measures of Central Tendency
- Measures of Dispersion

1.3 Analysis of Bivariate Data

- Analysis of Correlation
- Regression Analysis
- Time Series Analysis

1.1 Basic Concepts of Data Acquisition

Basic Concepts of Data Acquisition



Statistical Unit and Population:

- A single "object" is called statistical unit or entity

  Example: Robotics student XYZ taking the course "Statistics and Sensor Data Fusion" in the winter term 2023

- The set of "all objects" under consideration is called population

  Example: Entire class of robotics students taking the course "Statistics and Sensor Data Fusion" in the winter term 2023

Basic Concepts of Data Acquisition

Variables and Occurrences:

- Variables are those properties of a statistical unit one is interested in
- Variables are also called attributes or characteristics
- Variables of the statistical unit “robotics student XYZ” may be
  - matriculation number
  - citizenship
  - civil status
  - age in years
- An occurrence is the specific **value** a variable takes on

Basic Concepts of Data Acquisition

Example:

Matr. Nr.	Citizenship	Civil Status	Age
3313019	UK	single	21
3635302	Nigeria	single	19
3214177	Russia	married	21
3312008	India	single	18

- **Population:** All students (entire table)
- **Statistical Unit:** One student (single row)
- **Variables and Occurrences:**
  - Matr. Nr.: 3313019, 3635302, 3214177, 3312008
  - Citizenship: UK, Nigeria, Russia, India
  - Civil Status: single, married
  - Age: 18, 19, 21

Basic Concepts of Data Acquisition

Key Attributes:

- An attribute or a minimal combination of attributes that identifies a statistical unit **uniquely** is called a key attribute (compare primary/alternate keys in a relational database)

Possible Occurrences:

- Sometimes it is recommended or even necessary to also think of possible occurrences, even though they might not be represented in the current data sample at hand

Question: What are the key attributes and possible occurrences with respect to the previous example?

Basic Concepts of Data Acquisition

For the applicability of statistical procedures, the scale of a variable or attribute is of particular importance:

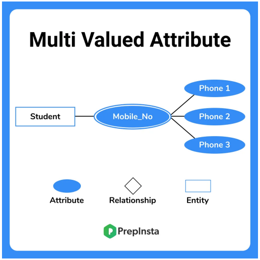
Scale of Variables:

- Categorical or nominal scale: Only different "labels"
  - Language
  - Citizenship
- Ordinal scale: Sorting is possible $(<, >, =)$
  - Army rank
  - Placement in a sporting event (e.g. top ten)
- Cardinal or metric scale: Calculation is possible $(+, \cdot)$
  - Stock and commodity prices
  - Age, weight, length

Basic Concepts of Data Acquisition

Multivalued Variables:

- Sometimes a variable can take on **two or more values** at the same time
- **Example:** A student
  - can have more than one citizenship
  - can speak more than one language
- Such variables are called multivalued variables or multivalued attributes



Question: Which scale is compatible with multivalued variables?

https://prepinsta.com/dbms/multivalued-attributes/

Exercise

Give **three more examples** for nominally, ordinally and cardinally scaled variables, respectively. For each of them, indicate at least three possible occurrences.

1.2 Analysis of Univariate Data

Analysis of Univariate Data

Frequently, descriptive statistics starts with raw data:

- The occurrences of the variables characterizing the statistical units are written down in form of a **list** with no specific order
- Since the same data record (row) may occur more than once, raw data actually constitutes a so-called **multiset** or **bag**
- In the case of univariate raw data, the list is **one-dimensional**, i.e. only a single variable or attribute is under consideration
- For a population consisting of n statistical units in total, this would result in the list

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array} \quad \text{or} \quad (x_1, x_2, \ldots, x_n)$$

Analysis of Univariate Data

Raw Data – Example:

In total $n = 10$ German pupils are asked about their final grades (German school system uses the grades $1, 2, 3, 4, 5, 6$)

- Raw data: $(2, 3, 1, 1, 5, 3, 6, 3, 3, 1)$
- Here, there are $k = 5$ different occurrences: 1, 2, 3, 5, 6
- The occurrence "4" is not present in the data sample, but theoretically possible
- The occurrences "1" and "3" appear several times

Analysis of Univariate Data

From Raw Data to Sorted Lists:

- In a sorted list, the raw data is sorted with respect to a specific criterion, e.g. **in ascending order**

$$\underbrace{(2,3,1,1,5,3,6,3,3,1)}_{\text{raw data}} \quad \longrightarrow \quad \underbrace{(1,1,1,2,3,3,3,3,5,6)}_{\text{sorted list}}$$

- For the generation of sorted lists, data of at least **ordinal scale** is required

- In the so-called list of occurrences, every occurrence is written down **only once** in ascending order

$$(a_1, a_2, a_3, a_4, a_5) = (1,2,3,5,6)$$

In order not to loose information, we have to complement the list of occurrences by the associated frequencies

Analysis of Univariate Data

Univariate Frequencies:

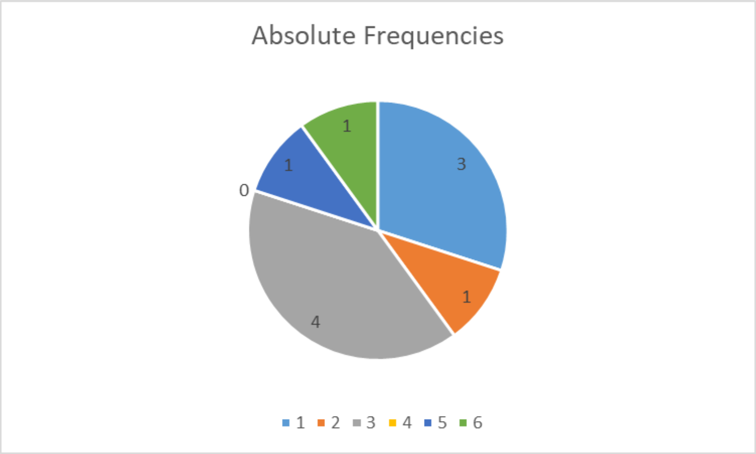
- The absolute frequency $h(a_j)$ of the occurrence a_j tells us **how often** a_j appears in the raw data
- The relative frequency $f(a_j) = h(a_j)/n$ of occurrence a_j tells us the **fraction** with which a_j occurs in the raw data
- With respect to our previous example, we get

a_j	1	2	3	4	5	6	Σ
$h(a_j)$	3	1	4	0	1	1	**10**
$f(a_j)$	0.3	0.1	0.4	0	0.1	0.1	**1**

- Observe that we have also considered the possible occurrence "4", which is actually not present in the raw data

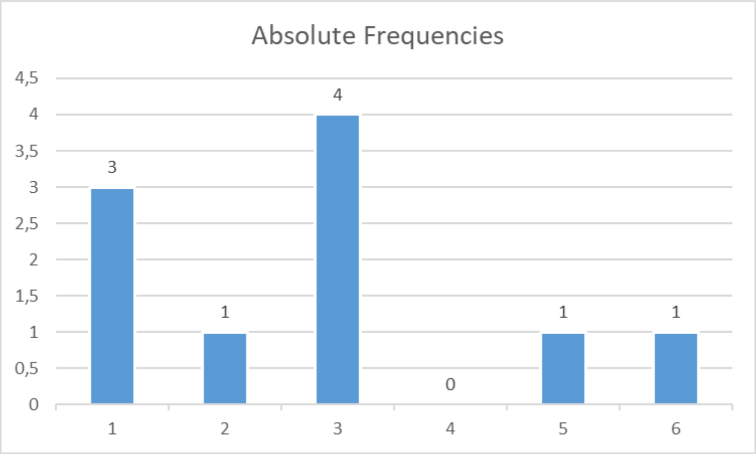
Analysis of Univariate Data

Representation of Frequencies – Pie Chart:



Analysis of Univariate Data

Representation of Frequencies – Bar Chart:

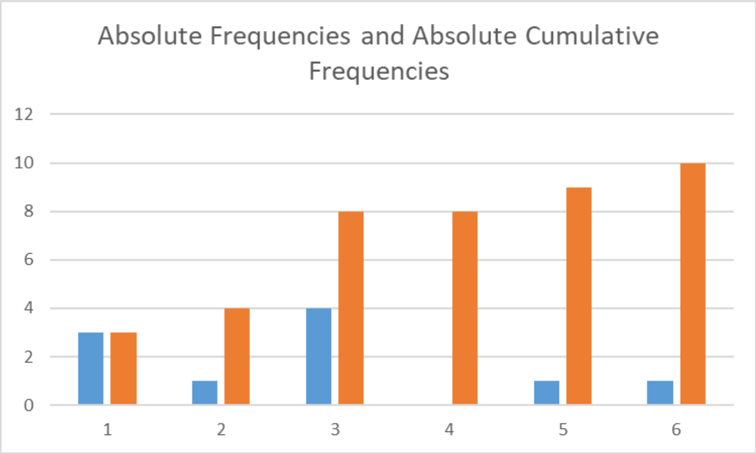


Analysis of Univariate Data

Cumulative Frequencies:

- With n the number of univariate statistical units $(x_1, \ldots, x_n)$ and k the number of different occurrences $(a_1, \ldots, a_k)$ of the variable under consideration, we have
  - the absolute frequencies $h_j = h(a_j)$
  - the relative frequencies $f_j = f(a_j) = h(a_j)/n$
- Based on h_j and f_j, we calculate
  - the absolute cumulative frequencies
    $$H_j = h_1 + h_2 + \cdots + h_j$$
  - the relative cumulative frequencies
    $$F_j = f_1 + f_2 + \cdots + f_j = H_j/n$$
- Generally, it holds that $H_k = n$ and $F_k = 1$

Absolute and Cumulative Frequencies – Example:



Analysis of Univariate Data

Cumulative Frequency Distribution:

- The absolute cumulative frequency distribution is given by

$$H(x) = \begin{cases} 0 & \text{if} \quad x < a_1 \\ H_j & \text{if} \quad a_j \leq x < a_{j+1} \\ n & \text{if} \quad x \geq a_k \end{cases}$$

- By dividing through the total number of statistical units n, we obtain the relative cumulative frequency distribution

$$F(x) = \frac{H(x)}{n} = \begin{cases} 0 & \text{if} \quad x < a_1 \\ F_j & \text{if} \quad a_j \leq x < a_{j+1} \\ 1 & \text{if} \quad x \geq a_k \end{cases}$$

- The function $F(x)$ is also called empirical distribution function

Exercise

During the uptime of a automated fabrication facility, the total number of objects produced within one hour was recorded as follows:

Number of objects	3	4	5	6	7	8	9	10
Absolute frequency	10	15	30	30	25	20	15	5

Compute $F(3)$, $F(5.5)$ and $F(10)$ of the empirical distribution function $F(x)$.

How many hours did the data acquisition take at least?

Analysis of Univariate Data

Numerical Representation of Univariate Data:

Frequently, the representation of a data sample in terms of a **few meaningful indicators** is desirable.

For this purpose, univariate data can be **numerically** represented by

- measures of central tendency

  "Where are the data located primarily?"

- measures of dispersion

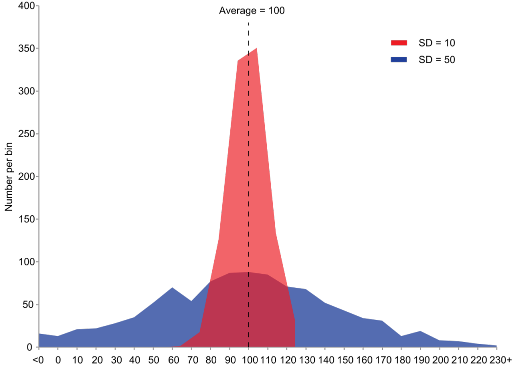
  "How are the data dispersed around the center?"

- measures of concentration

  "Is the focus of the data only on a few statistical units?"

Analysis of Univariate Data

Central Tendency and Dispersion – Example:



Two populations with the same mean but different dispersion. The blue population is much more dispersed than the red population.

By JRBrown - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10777712

Analysis of Univariate Data

Measures of Central Tendency:

- Measures of central tendency describe **where** the data are located primarily (compare the center of mass in physics)
- The whole data sample $(x_1, \ldots, x_n)$ is summed up into a **single value**
- Important measures of central tendency are
  - the mode
  - the median
  - the mean

Which measure is applicable depends on the **scale** of the variable!

Analysis of Univariate Data

Mode:

- The mode $\bar{x}_{mod}$ of a data sample $(x_1, \ldots, x_n)$ is the **most frequent value**

- **Example:** For the data sample $(2, 3, 1, 1, 5, 3, 6, 3, 3, 1)$ we have the occurrences and frequencies according to

a_j	1	2	3	4	5	6
h_j	3	1	4	0	1	1

$\Longrightarrow \quad \bar{x}_{mod} = 3$

- **Advantage:** The mode can already be used with nominal data

- **Disadvantage:** It only makes sense if the value is **unique**

a_j	1	2	3	4	5	6
h_j	4	1	4	1	0	4

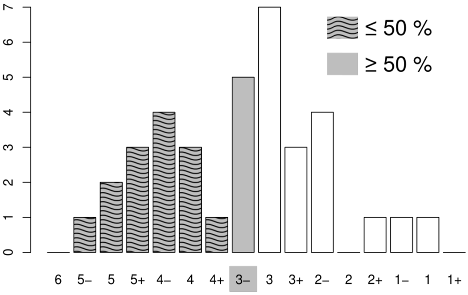
$\Longrightarrow \quad \bar{x}_{mod} = ?$

Analysis of Univariate Data

Median:

- The median $\bar{x}_Z$ is the value that separates the lower half of the data sample $(x_1, \ldots, x_n)$ from the higher half, in the sense that **at most 50%** of the values are smaller and **at most 50%** of the values are larger than the value of the median

- **Example:**



Analysis of Univariate Data

Median:

- If the data sample $(x_1, \dots, x_n)$ is sorted in ascending order, the median $\bar{x}_Z$ is defined by

$$\boxed{\bar{x}_Z = \begin{cases} x_{\frac{n+1}{2}} & \text{if} \quad n \text{ odd} \\ x_{\frac{n}{2}} & \text{if} \quad n \text{ even} \end{cases}}$$

- An **alternative choice** could be

$$\bar{x}_Z = \begin{cases} x_{\frac{n+1}{2}} & \text{if} \quad n \text{ odd} \\ x_{\frac{n}{2}+1} & \text{if} \quad n \text{ even} \end{cases}$$

- **Advantage:** The median can be used with ordinal data
- **Disadvantage:** It is generally **not unique** for even n

Analysis of Univariate Data

In the case of **ordinal data**, the concept of the median can be generalized to the so-called *p*-quantile:

Generalization of the Median – p-Quantile:

- The *p*-quantile with $0 < p < 1$ generalizes the idea of the median for a population of n statistical units
- Here, **at most** $p \cdot n$ of the values in data sample are smaller and **at most** $(1 - p) \cdot n$ of the values in the data sample are larger than the value of the p-quantile
- Important **special cases** of the p-quantile are given for
  - $p = 0.25$ first quartile Q1
  - $p = 0.5$ second quartile Q2 = median
  - $p = 0.75$ third quartile Q3

Exercise

Determine the quartiles Q1, Q2, Q3 of the following data samples:

(a) $(x_1, \ldots, x_6) = (2, 7, 8, 11, 13, 17)$

(b) $(x_1, \ldots, x_7) = (1, 5, 66, 234, 440, 489, 500)$

(c) $(x_1, \ldots, x_{10}) = (1, 3, 5, 66, 111, 234, 440, 489, 500, 777)$

Analysis of Univariate Data

The most frequently applied measures of central tendency are given by **mean values**, where the computation of mean values is only possible for cardinal data:

Mean Values – Overview:

- Arithmetic mean

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

- Geometric mean

$$\boxed{\bar{x}_G = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}}$$

- Harmonic mean

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

Analysis of Univariate Data

Arithmetic Mean:

- The arithmetic mean $\bar{x}$ of a data sample $(x_1, \ldots, x_n)$ given by the expression

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

  is often simply called "the mean"

- For a given data sample $(x_1, \ldots, x_n)$, the arithmetic mean solves the **minimization problem**

$$\min_c \sum_{i=1}^{n} (x_i - c)^2$$

- In contrast to mode and median, the resulting value of the arithmetic mean **does not have to occur** in the data sample

Analysis of Univariate Data

Geometric Mean:

- The geometric mean $\bar{x}_G$ according to

$$\boxed{\bar{x}_G = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}}$$

  is only defined for **positive values**

- **Geometric Idea:** Search for the side length s of a square which has the same area as a given rectangle according to

$$A = a \cdot b$$

- **Approach:** $A = s^2 \implies s = \sqrt{a \cdot b}$

- Generally, the geometric mean solves the equation

$$\bar{x}_G^n = x_1 \cdot x_2 \cdot \ldots \cdot x_n$$

Analysis of Univariate Data

Arithmetic Mean vs. Geometric Mean:

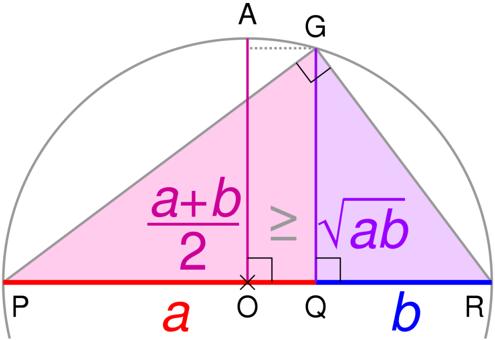


Illustration of the inequality of arithmetic and geometric mean:

$$\bar{x} \geq \bar{x}_G$$

Analysis of Univariate Data

Harmonic Mean:

- The harmonic mean $\bar{x}_H$ defined by

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

  is typically applied in situations where the **average rate** is considered

- **Example/Exercise:** Student Hans walks 2 km from home to college at a speed of 5 km/h. At arrival he realizes that he forgot his papers and runs back at a speed of 15 km/h.

  What is the average speed of Klaus?

Analysis of Univariate Data

Measures of Central Tendency: *When to use which parameter?*

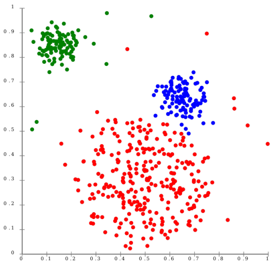
- **Nominal data** $\rightarrow$ mode
- **Ordinal data** $\rightarrow$ median
- **Cardinal data**: If the aggregate is
  - ... the result of a sum $\rightarrow$ arithmetic mean
  - ... the result of a product $\rightarrow$ geometric mean
  - ... the result of a fraction with unknown denominator $\rightarrow$ harmonic mean

For the arithmetic, geometric and harmonic mean of a data sample $(x_1, \ldots, x_n)$, the following **inequality** holds:

$$\boxed{\min\{x_1, \ldots, x_n\} \leq \bar{x}_H \leq \bar{x}_G \leq \bar{x} \leq \max\{x_1, \ldots, x_n\}}$$

Analysis of Univariate Data

Measures of Dispersion:



- Measures of dispersion describe how the data are **dispersed** around a central tendency
- Computation of dispersion parameters is only possible for cardinal data
- The most important dispersion parameters are
  - the range
  - the mean deviation
  - the empirical variance and standard deviation

Analysis of Univariate Data

Range:

- The range or width w of a data sample $(x_1, \ldots, x_n)$ is just the **difference** between the **largest** and the **smallest value**
- With the **sorted list of occurrences** $(a_1, \ldots, a_k)$, we obtain

$$\boxed{w = \max\{a_1, \ldots, a_k\} - \min\{a_1, \ldots, a_k\} = a_k - a_1}$$

- **Example:**

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = (11, 5, 7, 11, 18, 21, 5)$$

$$\Longrightarrow (a_1, a_2, a_3, a_4, a_5) = (5, 7, 11, 18, 21)$$

$$\Longrightarrow w = 21 - 5 = 16$$

Analysis of Univariate Data

Mean Deviation:

- The mean deviation or mean absolute deviation mad of a data sample $(x_1, \ldots, x_n)$ is defined by the expression

$$\boxed{\text{mad} = \frac{1}{n}\sum_{i=1}^{n} |x_i - m|}$$

- In the above expression, the value m corresponds to a **measure of central tendency**, e.g. the arithmetic mean $\bar{x}$

- Taking into account the occurrences a_j and the absolute frequencies h_j or the relative frequencies f_j, it holds that

$$\text{mad} = \frac{1}{n}\sum_{i=1}^{n} |x_i - m| = \frac{1}{n}\sum_{j=1}^{k} h_j \cdot |a_j - m| = \sum_{j=1}^{k} f_j \cdot |a_j - m|$$

Analysis of Univariate Data

Empirical Variance and Standard Deviation:

- The empirical variance s^2 of a data sample $(x_1, \dots, x_n)$ with respect to its arithmetic mean $\bar{x}$ is defined as

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2}$$

- The so-called standard deviation s is given by

$$\boxed{s = \sqrt{s^2}}$$

- Taking into account the occurrences a_j and the absolute frequencies h_j or the relative frequencies f_j, it holds that

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k} h_j \cdot (a_j - \bar{x})^2 = \sum_{j=1}^{k} f_j \cdot (a_j - \bar{x})^2$$

Analysis of Univariate Data

Relative Dispersion:

- The introduced measures of dispersion yield **absolute values**
- In order to obtain **relative values**, we also have to take the magnitude of the observed values into account
- For this purpose, measures of relative dispersion are generally defined in the form of a **ratio**

$$\frac{\text{absolute dispersion}}{\text{mean}}$$

- When inserting the **standard deviation** s and the **arithmetic mean** $\bar{x}$, we obtain the so-called coefficient of variation

$$\boxed{v = \frac{s}{\bar{x}}}$$

Exercise

Consider the data sample $(x_1, \ldots, x_{10})$ given by the table

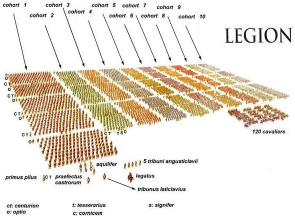
a_j	1	2	3	4	5	6
h_j	1	2	3	2	1	1
f_j	0.1	0.2	0.3	0.2	0.1	0.1

Calculate

(a) mode, median, arithmetic, geometric and harmonic mean

(b) range, mean absolute deviation (w.r.t. arithmetic mean), empirical variance, standard deviation and coefficient of variation

Analysis of Univariate Data



Merging Aggregates:

- Sometimes, the results of a statistical investigation may be delivered in m **groups** or **cohorts** $G_1, \ldots, G_m$
- For each group G_ℓ, we know the size of the group n_ℓ, the arithmetic mean $\bar{x}_\ell$ and the empirical variance s_ℓ^2
- Based on these groupwise results, we obtain the overall result by merging aggregates according to

$$\boxed{n = \sum_{\ell=1}^{m} n_j} \qquad \boxed{\bar{x} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot \bar{x}_\ell}$$

$$\boxed{s^2 = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot s_\ell^2 + \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot (\bar{x}_\ell - \bar{x})^2}$$

Exercise

Consider a statistical survey where the results of three groups G_1, G_2, G_3 are given according to

$$G_1: \quad n_1 = 10,\ \bar{x}_1 = 100,\ s_1^2 = 400$$

$$G_2: \quad n_2 = 5,\ \bar{x}_2 = 120,\ s_2^2 = 144$$

$$G_3: \quad n_3 = 20,\ \bar{x}_3 = 60,\ s_3^2 = 100$$

and calculate the overall values for $\bar{x}$ and s^2.

1.3 Analysis of Bivariate Data

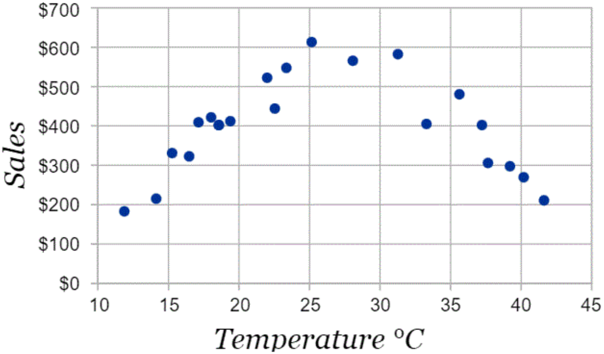
Analysis of Bivariate Data

Bivariate or Two-Dimensional Data:

- Up to now, we have considered univariate or one-dimensional data samples of the form $(x_1, \dots, x_n)$, where the occurrences of **only one variable** have been recorded
- Bivariate or two-dimensional data refers to a collection of **pairs**
$$((x_i, y_i))_{i=1}^{n} = ((x_1, y_1), \dots, (x_n, y_n))$$
where the values of two variables x and y are recorded **simultaneously**
- **Typical questions are:** Do the two variables depend on each other and if so, in which way and to what extent?
- Important bivariate concepts in this context are correlation and regression

Analysis of Bivariate Data

Bivariate Data – Example:



Total daily sales of an ice-cream seller versus the top temperature for various days during the summer holidays.

https://www.mathsisfun.com/definitions/bivariate-data.html

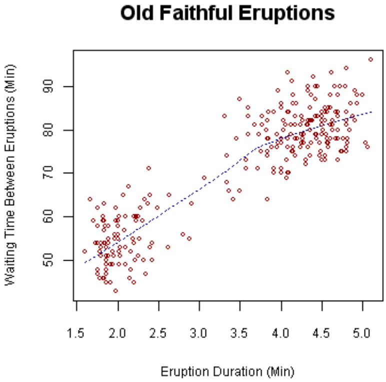
Analysis of Bivariate Data

Scatter Plot:

- A scatter plot is a type of diagram using Cartesian coordinates to display the values of two variables for a set of data
- The data is displayed as a **collection of points**, where the positions on the x- and y-axis are determined by the values of the two variables under consideration
- Based on the shape of the collection of points, it is possible to visually identify **correlations** between the variables
- Scatter plots are a suitable graphical representation of bivariate data if identical points are nearly impossible

Analysis of Bivariate Data

Scatter Plot – Example:

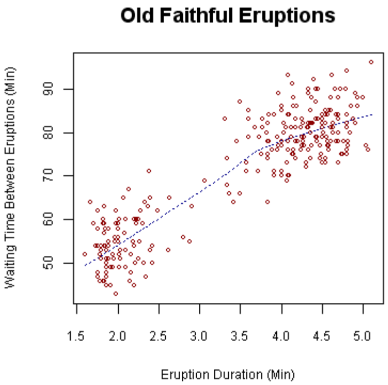


Waiting time between eruptions and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA.

Public Domain, https://commons.wikimedia.org/w/index.php?curid=646999

Analysis of Bivariate Data

Scatter Plot – Example:



Observations:

- The longer the waiting time the longer the following eruption
- There are **clusters** on both ends

Analysis of Bivariate Data

Correlation:

- The concept of correlation describes a relationship between two variables which are at least of **ordinal scale**
- **Example:** Taller persons are usually heavier
- **Attention:** Correlation is just a statistical observation, there is not necessarily a cause-symptom relationship
- **Example:** The residents of a rural town have noticed over the last hundred years that the decline of the birth rate in their town goes hand in hand with a reduction of stork population

Analysis of Bivariate Data

At first, we look at a correlation parameter for data of at least ordinal scale:

Spearman's Rank Correlation Coefficient:

- Consider the bivariate data sample

$$((x_i, y_i))_{i=1}^{n} = ((x_1, y_1), \ldots, (x_n, y_n))$$

  where x_i and y_i are occurrences of variables of **ordinal scale**

- **Idea:** Every x_i receives a rank $R(x_i)$ and every y_i receives a rank $R(y_i)$ with respect to their **position** in the sorted list of x-values and y-values

- As a result, each pair (x_i, y_i) can be associated with the pair $(R(x_i), R(y_i))$ of the respective ranks within the data sample

Analysis of Bivariate Data

Spearman's rank correlation coefficient indicates wheter the ranks of x_i and y_i run in the same direction:

Spearman's Rank Correlation Coefficient – Possible Cases:

1. The ranks of x_i and y_i run in the **same direction**

   $\rightarrow$ evidence for positive correlation

2. The ranks of x_i and y_i run in the **opposite direction**

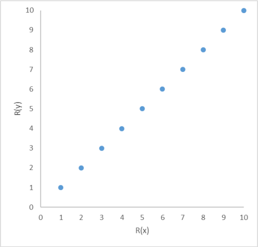
   $\rightarrow$ evidence for negative correlation

3. There is **no relationship** between the ranks of x_i and y_i

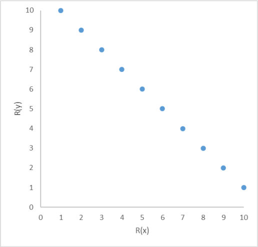
   $\rightarrow$ evidence for uncorrelated variables

Analysis of Bivariate Data

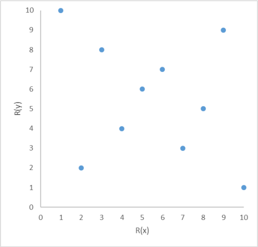
Spearman's Rank Correlation Coefficient – Examples:



$r_S = 1$



$r_S = -1$



$r_S \approx 0$

Spearman's rank correlation coefficient r_S measures both **strength and direction** of the correlation.

Analysis of Bivariate Data

Computing Spearman's Coefficient:

Based on the bivariate data sample $((x_i, y_i))_{i=1}^n$ and the associated ranks $((R(x_i), R(y_i)))_{i=1}^n$, Spearman's rank correlation coefficient r_S is defined by

$$\boxed{r_S = 1 - \frac{6 \cdot \sum_{i=1}^n (R(x_i) - R(y_i))^2}{(n-1) \cdot n \cdot (n+1)}}$$

By construction, it holds that $-1 \leq r_S \leq 1$.

Special Cases:

- $r_S = 1$ (ranks run into **exactly the same direction**)
- $r_S = -1$ (ranks run into **exactly the opposite direction**)
- $r_S = 0$ (ranks run **without any relationship**)

Exercise

Calculate Spearman's rank correlation coefficient for the variables "Age" and "Points in test":

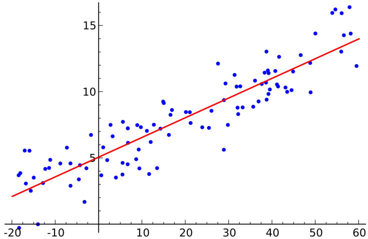
Student No.	1	2	3	4	5	6	7	8	9	10	11
Age (years)	38	47	44	51	35	29	22	14	12	19	9
Points in test	39	34	31	48	46	23	17	12	16	28	10

Analysis of Bivariate Data

Now we introduce a correlation parameter for cardinal data:

Coefficient of Correlation of Bravais-Pearson:

- The coefficient of correlation of Bravais-Pearson is a measure for the **linear relationship** between two cardinal variables
- It measures how well the bivariate data can be approximated by a so-called linear trend line or regression line



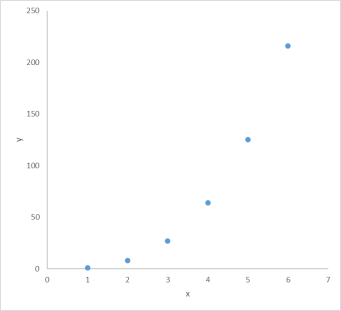
By Sewaqu - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11967659

Analysis of Bivariate Data

Difference to Spearman's Coefficient – Example:

Consider the bivariate data sample according to

$$((x_1, y_1), \ldots, (x_6, y_6)) = ((1,1),(2,8),(3,27),(4,64),(5,125),(6,216))$$



Question: What would be the value of Spearman's coefficient r_S?

Analysis of Bivariate Data

Computing the Coefficient of Bravais-Pearson:

Based on the bivariate data sample $((x_i, y_i))_{i=1}^n$ and the arithmetic means $\bar{x}$ and $\bar{y}$, the coefficient of correlation of Bravais-Pearson r is defined by

$$\boxed{r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}}$$

By construction, it holds that $-1 \leq r \leq 1$.

Special Cases:

- $r = 1$ (data **exactly located on a line** with positive slope)
- $r = -1$ (data **exactly located on a line** with negative slope)
- $r = 0$ (data show **no linear relationship** at all)

Exercise

Determine the coefficient of correlation for the variables “Inflation” and “Jobless rate”:

Year	2001	2002	2003	2004	2005
Inflation (%)	2	3	3	2	5
Jobless rate (%)	4	7	2	3	4

Analysis of Bivariate Data

Empirical Covariance:

- The empirical covariance COV is an auxiliary measure for the **joint variability** of two cardinally scaled variables
- For a bivariate data sample $((x_i, y_i))_{i=1}^n$, the empirical covariance between the variables x and y is defined by

$$\boxed{\mathrm{COV}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}$$

- Using the above definition of the covariance, an **alternative calculation** of the coefficient of correlation r is given by

$$r = \frac{\mathrm{COV}(x, y)}{s_x \cdot s_y}$$

with s_x and s_y the **standard deviation** of x and y

Analysis of Bivariate Data

Relationship Between the Coefficients of Spearman and Bravais-Pearson:

Replacing x_i and y_i by their ranks $R(x_i)$ and $R(y_i)$ and inserting the ranks into the formula for the coefficient of Bravais-Pearson, one obtains an **alternative expression** for Spearman's coefficient:

$$r_S = \frac{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right) \cdot \left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} \left(R(y_i) - \overline{R(y)}\right)^2}}$$

If the ranks are **unique**, it holds for the mean ranks that

$$\overline{R(x)} = \overline{R(y)} = \frac{n+1}{2}$$

Analysis of Bivariate Data

Alternative Calculation of Spearman's Coefficient – Example:

												Σ
x_i	38	47	44	51	35	29	22	14	12	19	9	
y_i	39	34	31	48	46	23	17	12	16	28	10	
$R(x_i)$	4	2	3	1	5	6	7	9	10	8	11	66
$R(y_i)$	3	4	5	1	2	7	8	10	9	6	11	66
$R(x_i) - \overline{R(x)}$	-2	-4	-3	-5	-1	0	1	3	4	2	5	
$R(y_i) - \overline{R(y)}$	-3	-2	-1	-5	-4	1	2	4	3	0	5	
$(R(x_i) - \overline{R(x)})^2$	4	16	9	25	1	0	1	9	16	4	25	110
$(R(y_i) - \overline{R(y)})^2$	9	4	1	25	16	1	4	16	9	0	25	110
$(R(x_i) - \overline{R(x)}) \cdot (R(y_i) - \overline{R(y)})$	6	8	3	25	4	0	2	12	12	0	25	97

$$r_S = \frac{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right)\left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right)^2}\sqrt{\sum_{i=1}^{n} \left(R(y_i) - \overline{R(y)}\right)^2}} = \frac{97}{110} = 0.88$$

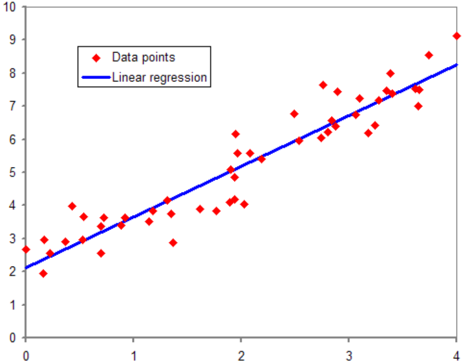
Analysis of Bivariate Data

Regression Analysis:

- In contrast to correlation analysis, in regression analysis the two variables x and y are assigned **different roles**
- The first variable x is assumed to be the **cause**, while the second variable y is assumed to be the **symptom**, i.e. one assumes a functional relationship according to $y = f(x)$
- **Example:** The longer the engine is running (x), the less gas remains in the tank (y) – The variable y (remaining gas) is considered to depend on the variable x (engine run time)
- **Goal:** Find the so-called regression function $f(x)$ which relates the two variables according to $y = f(x)$ in an optimal fashion

Analysis of Bivariate Data

Linear Regression – Example:



Linear trend line (blue) for 50 data points (red) around the line representing the regression function $y = f(x) = \frac{3}{2}x + 2$

By Amatulic at English Wikipedia (same as Anachronist on Wikimedia) - Transferred from en.wikipedia to Commons. Transfer was stated to be made by User:anachronist., Public Domain, https://commons.wikimedia.org/w/index.php?curid=3337769

Analysis of Bivariate Data

Computing a Linear Trend Line:

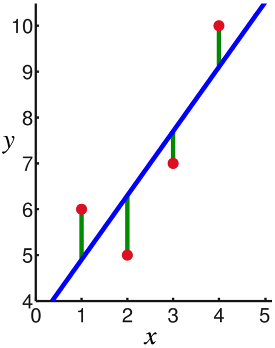
- We are searching for a linear function of the type
$$y = f(x) = ax + b$$
with **minimal distances** to the data $((x_1, y_1), \ldots, (x_n, y_n))$

- The resulting linear trend line is determined by the so-called coefficients of regression a and b

- The task of determining the coefficients of regression a and b for $f(x) = ax + b$ results in the **optimization problem**
$$Q(a, b) = \sum_{i=1}^{n} \big(\underbrace{y_i - f(x_i)}_{\text{residual}}\big)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \rightarrow \text{min!}$$
where the above minimization is carried out by means of varying both a and b

Linear Regression by Minimizing the Squared Residuals:

In linear regression, the deviation (green) of an observation (red) from an underlying relationship (blue) between a **dependent variable** y and an **independent variable** x is called residual.



The goal is to find the linear trend line according to

$$\boxed{f(x) = ax + b}$$

which minimizes the sum of the squared residuals (method of least squares).

By Oleg Alexandrov - self-made with MATLAB, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=4099808

Analysis of Bivariate Data

Solving the minimization problem **analytically** provides the formulas for the coefficients of regression a and b:

Computing the Coefficients of Regression:

- For the slope a of the linear trend line we get

$$\boxed{a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}}$$

- The intersection with the y-axis (y-intercept) b is given by

$$\boxed{b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i\right)}$$

Analysis of Bivariate Data

Computing the Coefficients of Regression:

Alternative expressions for the coefficients of regression a and b are given by

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{COV}(x, y)}{s_x^2} = r \cdot \frac{s_y}{s_x}, \quad b = \bar{y} - a\bar{x}$$

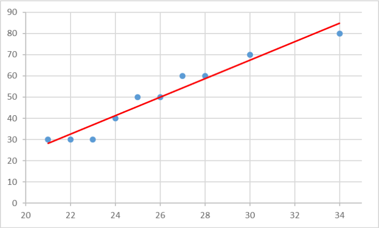
where

- $\bar{x}, \bar{y}$ are the arithmetic means of x and y
- s_x, s_y are the standard deviations of x and y
- r is the coefficient of correlation of Bravais-Pearson

Exercise

A distribution center runs ten shrink wrap machines on different velocities x (m/min), the variable y indicates the number of stops caused by cracks:

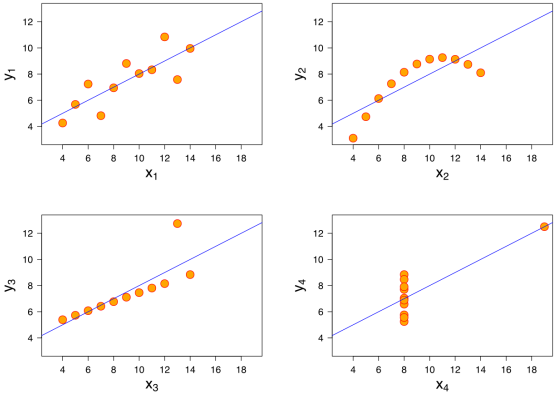
Machine	1	2	3	4	5	6	7	8	9	10
x_i	21	22	23	24	25	26	27	28	30	34
y_i	30	30	30	40	50	50	60	60	70	80



Compute the coefficients of regression a and b.

Analysis of Bivariate Data

Linear Regression – Pitfalls:



The data sets in the Anscombe's quartet are designed to have nearly the same linear regression line but are graphically very different.

Analysis of Bivariate Data

Nonlinear Regression:

- Assuming a linear regression function is not always a suitable hypothesis for the relation between the variables x and y
- If the nature of the problem or the scatter plot suggest another type of regression function (like e.g. logarithmic, exponential, quadratic, cubic, ...), assuming a linear relationship

  $$y = f(x) = ax + b$$

  is no longer justified
- Here, the key task is to identify a proper type of regression function, calculation of the parameters is the subsequent step
- In principle, minimization via least squares also works for **nonlinear regression**, but calculation may be much harder and involve more than two regression parameters

Analysis of Bivariate Data

In the case of monotonously increasing or decreasing regression functions and two regression parameters, a suitable substitution can reduce the problem to linear regression:

Nonlinear Regression – Example:

x_i	1	2	3	4
y_i	8	18	30	51

- **Hypothesis:** The regression function is of the type

$$y = f(x) = ax^2 + b$$

- **Task:** Compute a and b

- **Trick:** Substitute $\tilde{x} = x^2$ to arrive at the new (linear) problem

$$y = f(\tilde{x}) = a\tilde{x} + b \quad \text{where} \quad \tilde{x}_i = x_i^2$$

Analysis of Bivariate Data

Computing the Coefficients of Regression – Example:

					Σ
x_i	1	2	3	4	10
$\tilde{x}_i$	1	4	9	16	30
$\tilde{x}_i - \bar{\tilde{x}}$	-6.5	-3.5	1.5	8.5	
$(\tilde{x}_i - \bar{\tilde{x}})^2$	42.25	12.25	2.25	72.25	129
y_i	8	18	30	51	107
$y_i - \bar{y}$	-18.75	-8.75	3.25	24.25	
$(\tilde{x}_i - \bar{\tilde{x}}) \cdot (y_i - \bar{y})$	121.875	30.625	4.875	206.125	363.5

$$a = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\tilde{x}}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\tilde{x}})^2} = \frac{363.5}{129} = 2.82$$

$$b = \bar{y} - a\bar{\tilde{x}} = 26.75 - 2.82 \cdot 7.5 = 5.6$$

$$\Longrightarrow \quad y = f(x) = 2.82x^2 + 5.6$$

Analysis of Bivariate Data

Quality of a Regression Analysis:

- After calculating the parameters of a regression function, one should evaluate whether the resulting function adequately represents the relationship between the observed variables

- We do so by comparing the empirical values y_i against the corresponding values $f(x_i)$ of the regression function, the so-called estimates

$$\boxed{\hat{y}_i = f(x_i)}$$

- The **difference** between the empirical value y_i and its estimate $\hat{y}_i$ is the so-called residual $\hat{u}_i$ defined by

$$\boxed{\hat{u}_i = y_i - \hat{y}_i}$$

Analysis of Bivariate Data

Quality of a Regression Analysis:

- Based on the residuals $\hat{u}_i$, the coefficient of determination R^2 is a measure how much **reality** (the empirical data) meets a **statistical model** (the regression function):

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- By construction, it holds that $0 \leq R^2 \leq 1$
- **Special Cases:**
  - $R^2 = 1$: **Perfect regression function**, all residuals are equal to zero, all raw data are hitting the curve
  - $R^2 < 0.1$: The obtained regression function is **nonsense**

Analysis of Bivariate Data

Coefficient of Determination – Example:

According to the previous example, we have obtained the quadratic regression function

$$y = f(x) = 2.82x^2 + 5.6$$

The calculation of the coefficient of determination R^2 yields

					Σ
x_i	1	2	3	4	
y_i	8	18	30	51	
$\hat{y}_i = f(x_i)$	8.42	16.88	30.98	50.72	
$\hat{u}_i^2 = (y_i - \hat{y}_i)^2$	0.1764	1.254	0.96	0.078	2.4696
$(y_i - \bar{y})^2$	351.5625	76.56	10.56	588.1	1026.75

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{2.4696}{1026.75} = 0.9976 \quad \text{(almost perfect)}$$

Analysis of Bivariate Data

Coefficient of Determination – Alternative Formulas:

In the case of a linear regression function according to

$$y = f(x) = ax + b$$

the coefficient of determination R^2 can also be expressed as

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Furthermore, it holds that

$$\boxed{R^2 = a^2 \frac{s_x^2}{s_y^2} = r^2}$$

where r denotes the coefficient of correlation of Bravais-Pearson.

Analysis of Bivariate Data

Key Features of Regression Analysis:

Compared to the analysis of correlation only, regression analysis **additionally** provides

- a **compact description** of the mathematical relationship between the observed variables x and y by means of the obtained regression function $f(x)$
- the possibility to insert **new** (i.e. unobserved) values for the variable x into the regression function $f(x)$ in order to perform **forecasts** with respect to the dependent variable y (interpolation, extrapolation)

The above-mentioned features of regression analysis can be used for time series analysis.

Analysis of Bivariate Data

Time Series Analysis:

- A time series is a time-ordered collection of sequentially observed data pairs

  $$((t_i, x_i))_{i=1}^n = ((t_1, x_1), \ldots, (t_n, x_n))$$

  where the first variable t represents **points in time** with

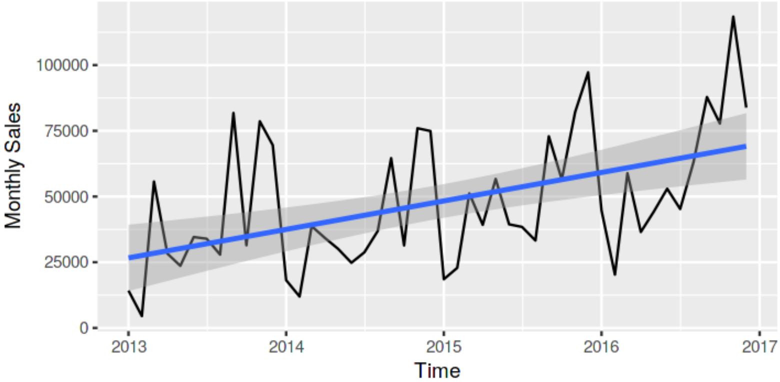
  $$t_1 < t_2 < \ldots < t_n$$

- The second time-dependent variable x is assumed to be **cardinally scaled**

- Analyzing the time series $((t_1, x_1), \ldots, (t_n, x_n))$ can be interpreted as a special kind of **regression analysis**

Analysis of Bivariate Data

Time Series – Example:



Time series (black) showing a long-term trend (blue) obtained as linear regression function.

https://www.mygreatlearning.com/blog/time-series-analysis-and-forecasting/

Analysis of Bivariate Data

Time Series – Scope of Applications:

- **Tracking:** Air surveillance, object tracking in robotics
- **Finance:** Stock market prices, financial liquidity development, currency rates
- **Econometrics:** Gross domestic product, unemployment rate, foreign trade
- **Biometrics:** ECG (electrocardiogram), EEG (electroencephalogram)

In the above fields, frequently multivariate time series with more than one time-dependent variable will be encountered.

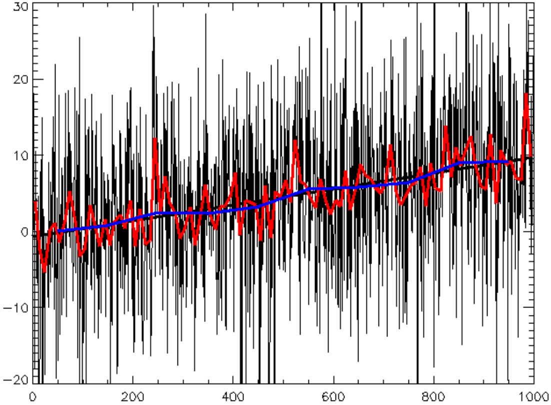
Analysis of Bivariate Data

Time Series – Key Features:

- In many cases, time series show temporal patterns like
  - long-term trends
  - seasonal cycles
  - random fluctuations
- **Approach:** Analyze a time series by means of decomposing the series into its basic components
- **Goal:** When all underlying principles of a time series in terms of the components present are discovered, it is possible to forecast future behaviour

Analysis of Bivariate Data

Time Series – Example:



Time series (black) composed of a linear trend and random fluctuations with different applied filters (red, blue).

Analysis of Bivariate Data

Time Series – Possible Components:

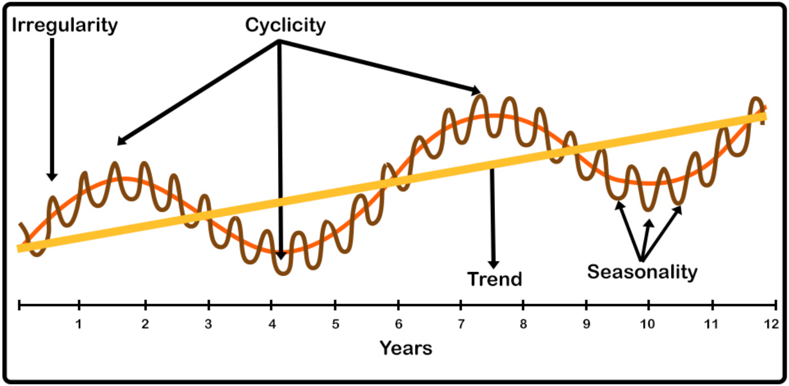
Our working hypothesis in the analysis of a time series $((t_i, x_i))_{i=1}^n$ will be that the time-dependent variable x is the sum of **at most four different components**:

$$\boxed{x = T + C + S + R}$$

- Trend T: **long-term** tendency
- Cycle C: **medium-term** periodic component
- Season S: **short-term** periodic component
- Random Component R: **irregular fluctuations**, day-to-day noise, anything that does not fit into the model

Analysis of Bivariate Data

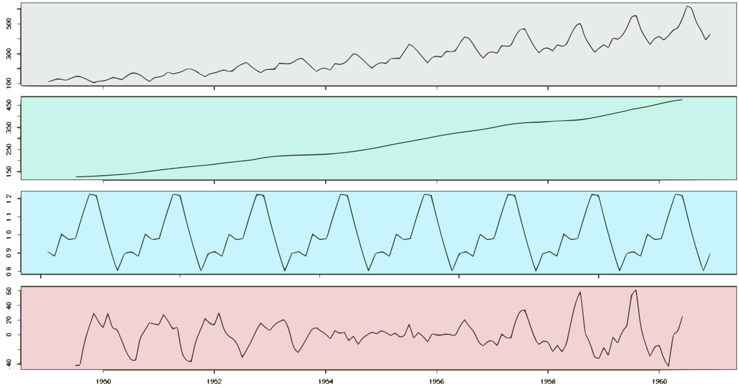
Time Series – Possible Components:



https://www.kaggle.com/general/169703

Analysis of Bivariate Data

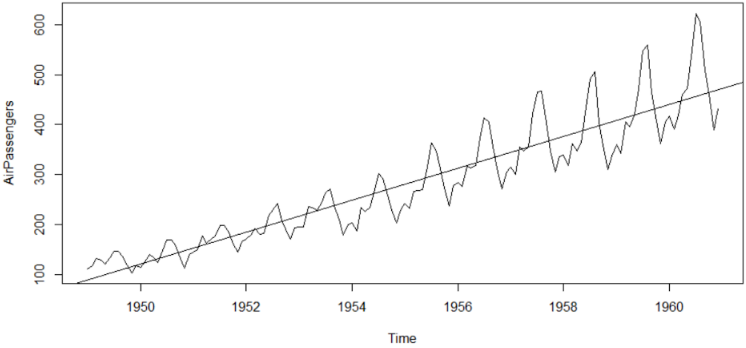
Decomposed Time Series – Example:



https://sthalles.github.io/a-visual-guide-to-time-series-decomposition/

Analysis of Bivariate Data

Time Series and Trend T:



https://www.simplilearn.com/tutorials/data-science-tutorial/time-series-forecasting-in-r

Analysis of Bivariate Data

Component Trend T:

- The trend T describes the **long-term tendency** of a time series, often in terms of a straight line ($\rightarrow$ regression analysis)
- The trend T corresponds to factors responsible for mainstream development and is based on **global phenomena**, e.g.
  - technological progress
  - general rise in living standards
  - depletion of natural resources
- **Examples:**
  - increasing automation of production processes
  - decreasing gasoline consumption per car
  - global warming

Analysis of Bivariate Data

Time Series and Cycle C:

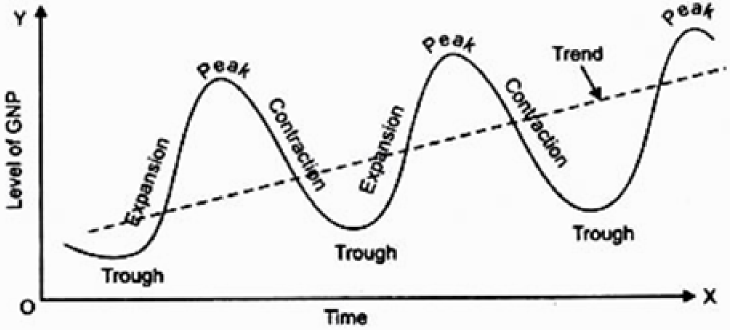


Fig. 13.2. *Cycles with Trend (i.e., Growth)*

www.economicsdiscussion.net

Analysis of Bivariate Data

Component Cycle C:

- The cycle C corresponds to **medium-term oscillations** in a time series, and is sometimes difficult to isolate
- One example is the so-called business cycle of about $7 - 11$ years (Juglar cycle)
- The cycle C is a periodic component with a cycle length or period **larger than** the length of the component season S
- The standardized cycle C shows oscillations around the zero line **without** the trend T

Analysis of Bivariate Data

Standardized Cycle C:

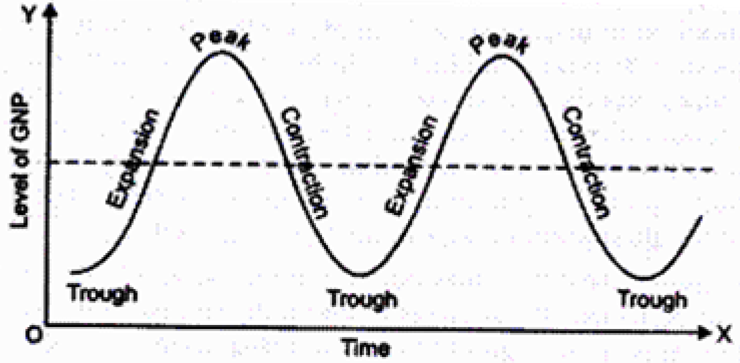
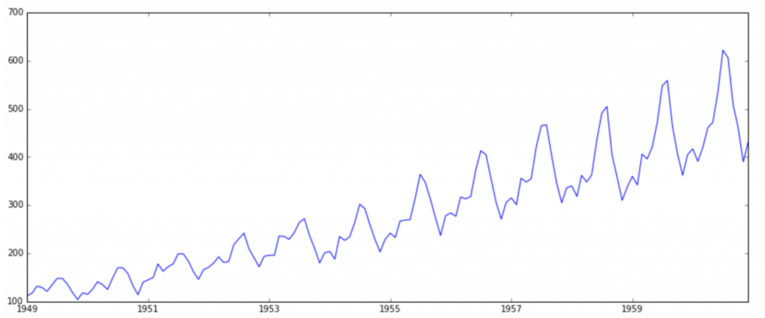


Fig. 13.1. *Four Phases of Business Cycles without Growth Trend*

www.economicsdiscussion.net

Analysis of Bivariate Data

Time Series and Season S:



Question: What could be the length of the season S?

www.analyticsvidhya.com

Analysis of Bivariate Data

Component Season S:

- The season S corresponds to a **short-term oscillation** with a typical pattern and a usually well-known wavelength
- **Examples:**
  - X-mas trade
  - summer clearance sale
  - holiday time
  - winter downturn in construction business
  - daily consumption of electricity
- A typical seasonal period is **one year**, but there are also alternatives (e.g. electric power consumption shows a periodic behaviour on a daily time scale)

Analysis of Bivariate Data

Time Series and Random Component R:

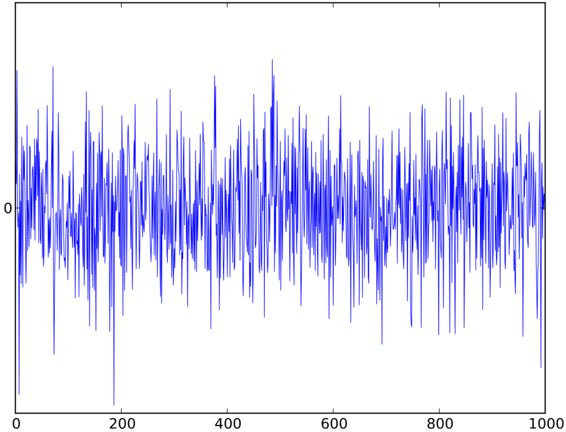
- The random component R includes all deviations of reality from the assumed model consisting of trend, cycle and season:

$$R = x - (T + C + S)$$

- Therefore, it represents **irregular and unpredictable effects**
- If the random component R is significant, reasons could be
  - unpredictable big incidents which overshadow normality (9/11, natural disasters)
  - analysis of time series data was flawed (e.g. wrong model assumptions)
- The standardized random component R fluctuates around the **zero line**, i.e. it is zero on average

Analysis of Bivariate Data

Standardized Random Component R:



Analysis of Bivariate Data

Time Series Analysis – Course of Action:

$$\boxed{x = T + C + S + R}$$

1. Determine the trend T by **regression analysis**
2. Determine the so-called smooth component $G = T + C$ by **filtering** the time series in terms of moving averages
3. Based on T and G, determine the cycle $C = G - T$
4. Determine the seasonal component S by **averaging** over the difference $x - G = S + R$
5. Eventually, the random component R remains as the difference $R = x - T - C - S$ and is hopefully small

Smooth Component – Seasonal Adjustment by Filtering:

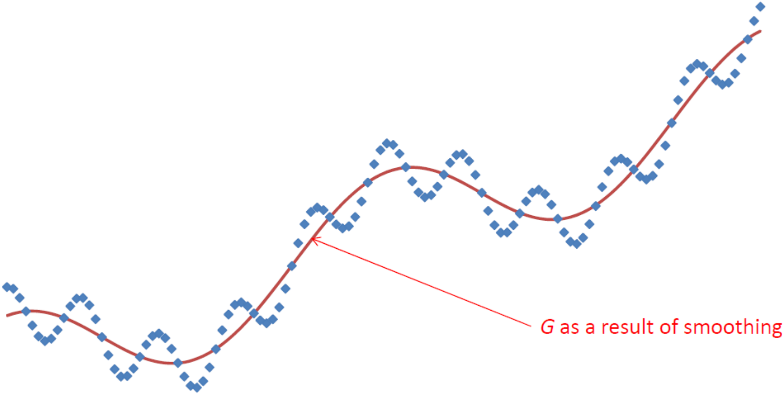
- The combination of trend T and cycle C is called smooth component G of a time series:

$$G = T + C$$

- It is obtained via seasonal adjustment where the influence of predictable seasonal patterns is removed from the time series
- Technically, this is accomplished by **smoothing** or **filtering** the time series in terms of moving averages with a time window of the same size as the season length
- As a result, the smooth component G does not contain seasonal variations anymore

Analysis of Bivariate Data

Smooth Component – Seasonal Adjustment by Filtering:



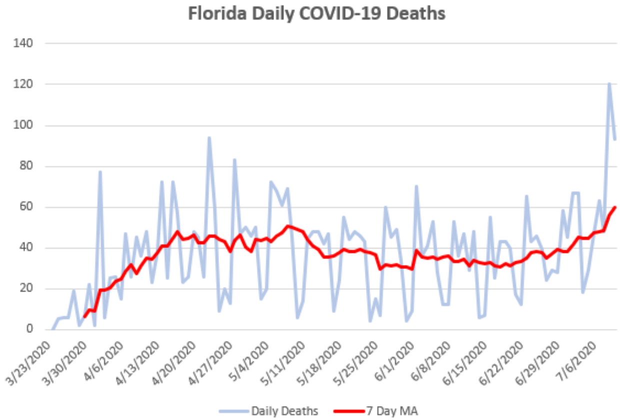
Analysis of Bivariate Data

Smoothing of Time Series by Moving Averages:

- In order to obtain the smooth component G, we will consider **filtering** the time series in terms of moving averages
- During this procedure, the actual value of the time series is replaced by an **average of its neighbouring values** according to a specified time window
- Generally it holds that the more neighbouring values are taken into account (i.e. the larger the size of the time window), the **more smoothing** is achieved
- For obtaining the smooth component G, the size of the time window has to be equal to the season length

Analysis of Bivariate Data

Smoothing of Time Series by Moving Averages – Example:



https://statisticsbyjim.com/time-series/moving-averages-smoothing/

Exercise

Determine the moving averages x_j^* of the following time series according to a time window of length 5:

t_i	0	1	2	3	4	5	6	7	8
x_i	3.00	4.07	5.13	6.20	7.27	3.33	4.40	5.47	6.53

9	10	11	12	13	14	15
7.60	3.67	4.73	5.80	6.87	7.93	4.00

Analysis of Bivariate Data

For computing moving averages, it is necessary to distinguish between time windows of **even or odd order**:

Moving Averages of Odd Order:

- Time series data $((t_i, x_i))_{i=1}^n = ((t_1, x_1), \ldots, (t_n, x_n))$
- We assume an **odd season length** of $2k+1$, $k \in \mathbb{N}$
- For the corresponding moving average x_j^* with a time window of length $2k+1$, it holds

$$x_j^* = \frac{1}{2k+1} \sum_{i=j-k}^{j+k} x_i = \sum_{i=j-k}^{j+k} \underbrace{\frac{1}{2k+1}}_{\text{weight } g_i} x_i, \quad j = k+1, \ldots, n-k$$

- Due to the odd order, the moving average x_j^* is automatically **centered**

Analysis of Bivariate Data

Moving Averages of Even Order:

- Time series data $((t_i, x_i))_{i=1}^n = ((t_1, x_1), \ldots, (t_n, x_n))$
- Now we assume an **even season length** of $2k$, $k \in \mathbb{N}$
- For the corresponding moving average with a time window of length $2k$, it holds

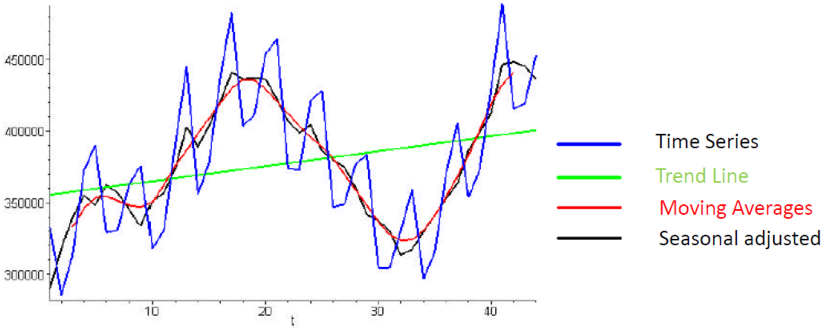
$$x_j^* = \frac{1}{2k}\left(\frac{x_{j-k}}{2} + \frac{x_{j+k}}{2} + \sum_{i=j-k+1}^{j+k-1} x_i\right), \quad j = k+1, \ldots, n-k$$

- Again we take $2k+1$ values into account to center x_j^*, but the first and the last value receive only **half of the regular weight**

Analysis of Bivariate Data

Applying Moving Averages to Time Series – Example:

Number of unemployed in Bavaria



t: Quarters since 1993

Source: Statistisches Bundesamt

Analysis of Bivariate Data

Applying Moving Averages to Time Series – Example:

t_i	x_i	x_i^*
1	6.16	
2	6.37	
3	6.53	
4	6.64	
5	6.72	
6	6.75	
7	6.76	6.33
8	6.75	6.17
9	6.72	5.98
10	6.69	5.80
11	6.66	5.61
12	4.23	5.43

Year 1

t_i	x_i	x_i^*
13	4.23	5.27
14	4.24	5.12
15	4.29	5.01
16	4.38	4.92
17	4.50	4.87
18	4.68	4.86
19	4.90	4.89
20	5.17	4.96
21	5.49	5.08
22	5.86	5.24
23	6.28	5.44
24	4.34	5.67

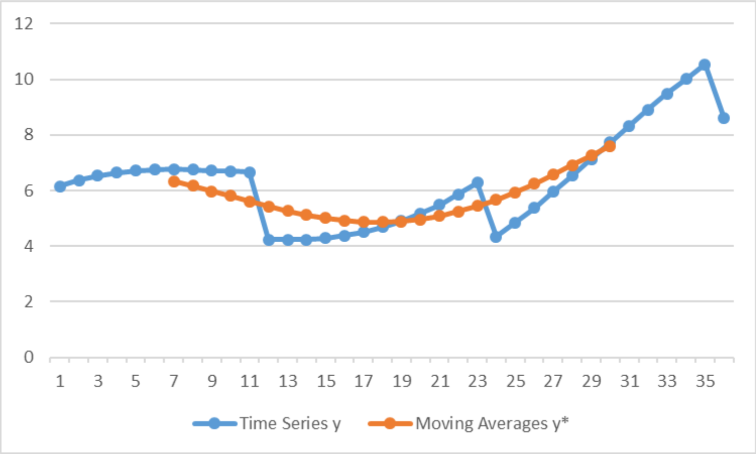
Year 2

t_i	x_i	x_i^*
25	4.85	5.94
26	5.38	6.24
27	5.95	6.57
28	6.53	6.91
29	7.13	7.26
30	7.73	7.61
31	8.33	
32	8.91	
33	9.48	
34	10.02	
35	10.53	
36	8.60	

Year 3

Analysis of Bivariate Data

Applying Moving Averages to Time Series – Example:



Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

- The 12-month moving average x^* obtained by filtering can be considered as an **estimate** for the smooth component:

$$x^* \approx G = T + C$$

- Therefore, the difference

$$x - x^* \approx S + R$$

  cleans the time series from G

- The seasonal component S is determined **by averaging** the values for the difference $x - x^*$ with respect to the same month of year 1, 2 and 3

Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

t_i	x_i	x_i^*	$x_i - x_i^*$
1	6.16		
2	6.37		
3	6.53		
4	6.64		
5	6.72		
6	6.75		
7	6.76	6.33	0.43
8	6.75	6.17	0.58
9	6.72	5.98	0.74
10	6.69	5.80	0.89
11	6.66	5.61	1.05
12	4.23	5.43	-1.20

Year 1

t_i	x_i	x_i^*	$x_i - x_i^*$
13	4.23	5.27	-1.04
14	4.24	5.12	-0.88
15	4.29	5.01	-0.72
16	4.38	4.92	-0.54
17	4.50	4.87	-0.37
18	4.68	4.86	-0.18
19	4.90	4.89	0.01
20	5.17	4.96	0.21
21	5.49	5.08	0.41
22	5.86	5.24	0.62
23	6.28	5.44	0.84
24	4.34	5.67	-1.33

Year 2

t_i	x_i	x_i^*	$x_i - x_i^*$
25	4.85	5.94	-1.09
26	5.38	6.24	-0.86
27	5.95	6.57	-0.62
28	6.53	6.91	-0.38
29	7.13	7.26	-0.13
30	7.73	7.61	0.12
31	8.33		
32	8.91		
33	9.48		
34	10.02		
35	10.53		
36	8.60		

Year 3

Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

Month	Average $x - x^*$
1	-1.07
2	-0.87
3	-0.67
4	-0.46
5	-0.25
6	-0.03
7	0.22
8	0.40
9	0.58
10	0.76
11	0.95
12	-1.27
Σ	-1.71

Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

- In the ideal case, the sum Σ of all averaged differences $x - x^*$ over one season (e.g. one year) equals **zero**
- This standardized case corresponds to an oscillation of the seasonal component around the **zero line**
- If the sum Σ is smaller or larger than zero, the values are **corrected** to eventually obtain the standardized seasonal effects, the seasonal pattern S of the time series
- The seasonal pattern S then takes the form

$$S = \underbrace{(x - x^*)}_{\text{average}} - \frac{\Sigma}{12}$$

Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

Month	Average $x - x^*$	$(x - x^*) - \frac{\Sigma}{12}$
1	-1.07	-0.93
2	-0.87	-0.73
3	-0.67	-0.53
4	-0.46	-0.32
5	-0.25	-0.11
6	-0.03	0.11
7	0.22	0.36
8	0.40	0.54
9	0.58	0.72
10	0.76	0.90
11	0.95	1.09
12	-1.27	-1.13
Σ	-1.71	0

Analysis of Bivariate Data

Seasonal Adjustment by Moving Averages:

The seasonally adjusted time series $x - S$ is eventually obtained as

t_i	x_i	$x_i - S_i$
1	6.16	7.09
2	6.37	7.10
3	6.53	7.05
4	6.64	6.96
5	6.72	6.82
6	6.75	6.65
7	6.76	6.41
8	6.75	6.22
9	6.72	6.01
10	6.69	5.79
11	6.66	5.57
12	4.23	5.36

Year 1

t_i	x_i	$x_i - S_i$
13	4.23	5.15
14	4.24	4.97
15	4.29	4.82
16	4.38	4.69
17	4.50	4.61
18	4.68	4.57
19	4.90	4.54
20	5.17	4.63
21	5.49	4.77
22	5.86	4.96
23	6.28	5.20
24	4.34	5.47

Year 2

t_i	x_i	$x_i - S_i$
25	4.85	5.78
26	5.38	6.11
27	5.95	6.47
28	6.53	6.85
29	7.13	7.24
30	7.73	7.62
31	8.33	7.97
32	8.91	8.38
33	9.48	8.77
34	10.02	9.13
35	10.53	9.45
36	8.60	9.73

Year 3

Seasonal Adjustment by Moving Averages – Summary:

1. **Filter** the original time series x in terms of moving averages x^* with a time window of equal size to the season length
2. **Remove** the smooth component $x^* \approx G = T + C$ from the time series to obtain $S + R$
3. **Average** over the corresponding differences $x - x^*$ to obtain the seasonal averages
4. **Standardize** the seasonal averages so that they oscillate around the zero line to obtain the seasonal pattern S
5. Eventually, determine the seasonally adjusted time series $x - S$ by **removing** the seasonal pattern S from the original time series $x = T + C + S + R$

Exercise

In a tide station, the water level is measured on four consecutive days at 0:00 a.m., 8:00 a.m. and 4:00 p.m., respectively.

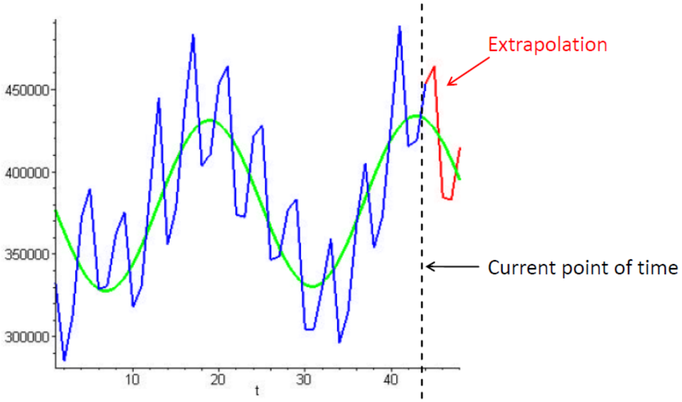
The results of the measurements in cm are as follows:

Day 1	Day 2	Day 3	Day 4
603|723|480	606|720|420	600|660|420	537|660|600

Apply appropriate moving averages to obtain the seasonal pattern and the seasonally adjusted time series.

Analysis of Bivariate Data

An important application of time series analysis is **forecasting future behaviour** by extrapolation:



Analysis of Bivariate Data

Time Series Forecasting by Extrapolation – Steps:

1. Acquire time series data $((t_i, x_i))_{i=1}^{n}$
2. Check that **preconditions** for time series analysis are fulfilled:
   - At least **one full cycle** is contained in time series data
   - At least a **couple of seasons** are contained in time series data
   - Seasonal wavelength is **well-known**
3. Determine T, C and S
4. Forecast the time series by means of **forwarding** T and **repeating** the periodic components C and S:

$$x = T + C + S$$

Analysis of Bivariate Data

Time Series Forecasting – Example:

The acquired time series is given by

t_i	x_i
1	4.00
2	4.64
3	5.20
4	5.69
5	6.07
6	6.36
7	6.56
8	6.69
9	6.78
10	6.85
11	6.94
12	3.06

Year 1

t_i	x_i
13	3.26
14	3.55
15	3.93
16	4.41
17	4.97
18	5.61
19	6.29
20	6.98
21	7.66
22	8.30
23	8.87
24	5.35

Year 2

Analysis of Bivariate Data

Time Series Forecasting – Example:

We want to forecast the time series into the first quarter of year 3:

t_i	y_i	t_i	y_i	t_i	y_i
1	4.00	13	3.26	25	?
2	4.64	14	3.55	26	?
3	5.20	15	3.93	27	?
4	5.69	16	4.41	28	?
5	6.07	17	4.97		
6	6.36	18	5.61		
7	6.56	19	6.29		
8	6.69	20	6.98		
9	6.78	21	7.66		
10	6.85	22	8.30		
11	6.94	23	8.87		
12	3.06	24	5.35		
Year 1		Year 2		Year 3	

Analysis of Bivariate Data

Time Series Forecasting – Example:

The calculation of T, C and S yields

- Trend T (as function of month x):
  $5.00 + 0.05 \cdot x$
- Cycle C (with length 20):
  0.45, 0.70, 0.89, 0.99, 0.99, 0.89, 0.71, 0.46, 0.16, -0.15, -0.45, -0.70, -0.89, -0.99, -0.99, -0.89, -0.71, -0.46, -0.16, 0.15
- Season S (with length 12):
  -1.50, -1.17, -0.83, -0.50, -0.17, 0.17, 0.50, 0.83, 1.17, 1.50, 1.83, -1.83

While the trend T is available in form of a (linear) function, the other components C and S are only available in form of **data**.

Analysis of Bivariate Data

Time Series Forecasting – Example:

The decomposed time series can be extrapolated:

t_i	T	S	C
1	5.05	-1.50	0.45
2	5.10	-1.17	0.70
3	5.15	-0.83	0.89
4	5.20	-0.50	0.99
5	5.25	-0.17	0.99
6	5.30	0.17	0.89
7	5.35	0.50	0.71
8	5.40	0.83	0.46
9	5.45	1.17	0.16
10	5.50	1.50	-0.15
11	5.55	1.83	-0.45
12	5.60	-1.83	-0.70

Year 1

t_i	T	S	C
13	5.65	-1.50	-0.89
14	5.70	-1.17	-0.99
15	5.75	-0.83	-0.99
16	5.80	-0.50	-0.89
17	5.85	-0.17	-0.71
18	5.90	0.17	-0.46
19	5.95	0.50	-0.16
20	6.00	0.83	0.15
21	6.05	1.17	0.45
22	6.10	1.50	0.70
23	6.15	1.83	0.89
24	6.20	-1.83	0.99

Year 2

t_i	T	S	C
25	6.25	-1.50	0.99
26	6.30	-1.17	0.89
27	6.35	-0.83	0.71
28	6.40	-0.50	0.46

Year 3

Analysis of Bivariate Data

Time Series Forecasting – Example:

This eventually yields the desired forecast:

t_i	y_i
1	4.00
2	4.64
3	5.20
4	5.69
5	6.07
6	6.36
7	6.56
8	6.69
9	6.78
10	6.85
11	6.94
12	3.06

Year 1

t_i	y_i
13	3.26
14	3.55
15	3.93
16	4.41
17	4.97
18	5.61
19	6.29
20	6.98
21	7.66
22	8.30
23	8.87
24	5.35

Year 2

t_i	$T + C + S$
25	5.74
26	6.03
27	6.23
28	6.36

Year 3

Analysis of Bivariate Data

Alternative Methods of Filtering:

- By smoothing a time series $((t_i, x_i))_{i=1}^{n}$ in terms of **moving averages** according to

$$x_j^* = \frac{1}{2k+1} \sum_{i=j-k}^{j+k} x_i = \sum_{i=j-k}^{j+k} \frac{1}{2k+1} x_i = \sum_{i=j-k}^{j+k} g_i x_i$$

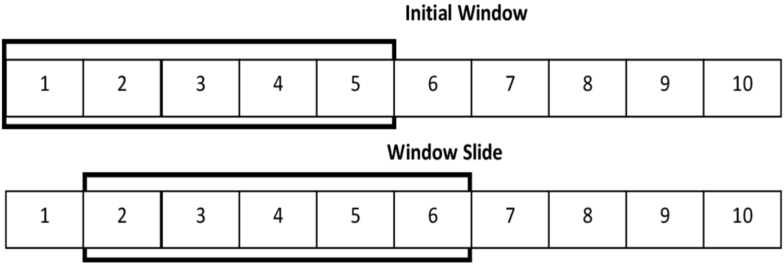
  all values within the time window receive the **constant weight**

$$g_i = \frac{1}{2k+1}$$

- Values **outside** the time window are **not** taken into account, i.e. they receive the weight $g_i = 0$
- This can be considered as a special way of filtering the time series by a sliding time window with specific weights

Analysis of Bivariate Data

Filtering by Sliding Time Window:



H. Hota and R. Handa and A. Shrivas: Time Series Data Prediction Using Sliding Window Based RBF Neural Network, 2017

Analysis of Bivariate Data

Alternative Methods of Filtering:

- Alternative methods of smoothing or filtering can be obtained by means of considering varying weights g_i
- Frequently, the weight g_i of time series data x_i depends on the **distance to the center of the time window**
- An approach for a time window of size $2k+1$ which implements linearly decreasing weights is given by

$$x_j^* = \sum_{i=j-k}^{j+k} g_i x_i = \sum_{i=j-k}^{j+k} \frac{k+1-|j-i|}{(k+1)^2} x_i$$

where the weight g_i is given by

$$g_i = \frac{k+1-|j-i|}{(k+1)^2}$$

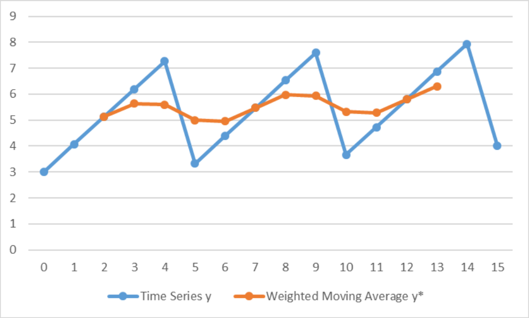
Analysis of Bivariate Data

Example:

t_i	x_i	x_i^*
0	3.00	
1	4.07	
2	5.13	5.13
3	6.20	5.64
4	7.27	5.60
5	3.33	5.00
6	4.40	4.96
7	5.47	5.47
8	6.53	5.98
9	7.60	5.93
10	3.67	5.33
11	4.73	5.29
12	5.80	5.80
13	6.87	6.31
14	7.93	
15	4.00	

$$x_j^* = \sum_{i=j-2}^{j+2} g_i x_i = \sum_{i=j-2}^{j+2} \frac{3-|j-i|}{9} x_i$$

$$x_4^* = \frac{1}{9} \cdot 4.07 + \frac{2}{9} \cdot 5.13 + \frac{3}{9} \cdot 6.20 + \frac{2}{9} \cdot 7.27 + \frac{1}{9} \cdot 3.33$$



Analysis of Bivariate Data

Filtering by Exponential Smoothing:

- Filtering by exponential smoothing is also known as "exponential weighted moving average"
- **Idea:** Control the influence of recent time series data by a smoothing factor α, where it holds that
  - $0 \leq \alpha \leq 1$
  - larger values of α reduce the extent of smoothing
  - typical usage of α is between 0.1 and 0.3
- Computation can be performed in a **recursive manner**:

$$\boxed{x_1^* = x_1 \quad \text{and} \quad x_{j+1}^* = \alpha \cdot x_{j+1} + (1-\alpha) \cdot x_j^*}$$

- This approach is similar to the Kalman filter

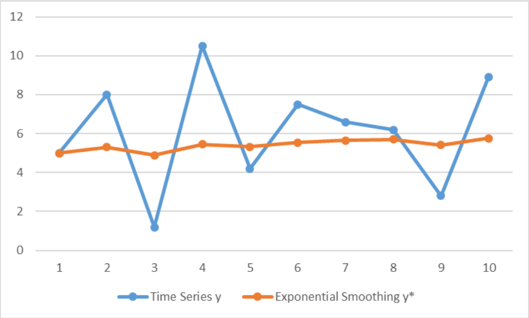
Analysis of Bivariate Data

Example: Exponential smoothing with $\alpha = 0.1$

t_i	x_i	x_i^*
1	5.00	5.00
2	8.00	5.30
3	1.20	4.89
4	10.50	5.45
5	4.20	5.33
6	7.50	5.54
7	6.60	5.65
8	6.20	5.70
9	2.80	5.41
10	8.90	5.76

$$x_{j+1}^* = \alpha \cdot x_{j+1} + (1-\alpha) \cdot x_j^*$$

$$x_3^* = 0.1 \cdot x_3 + 0.9 \cdot x_2^* = 0.1 \cdot 1.20 + 0.9 \cdot 5.30$$



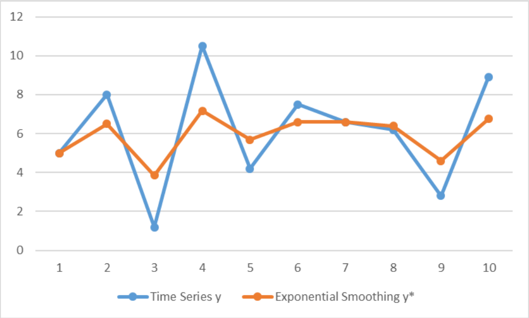
Analysis of Bivariate Data

Example: Exponential smoothing with $\alpha = 0.5$

t_i	x_i	x_i^*
1	5.00	5.00
2	8.00	6.50
3	1.20	3.85
4	10.50	7.18
5	4.20	5.69
6	7.50	6.59
7	6.60	6.60
8	6.20	6.40
9	2.80	4.60
10	8.90	6.75

$$x_{j+1}^* = \alpha \cdot x_{j+1} + (1-\alpha) \cdot x_j^*$$

$$x_3^* = 0.5 \cdot x_3 + 0.5 \cdot x_2^* = 0.5 \cdot 1.20 + 0.5 \cdot 6.50$$



Analysis of Bivariate Data

Variations of Exponential Smoothing:

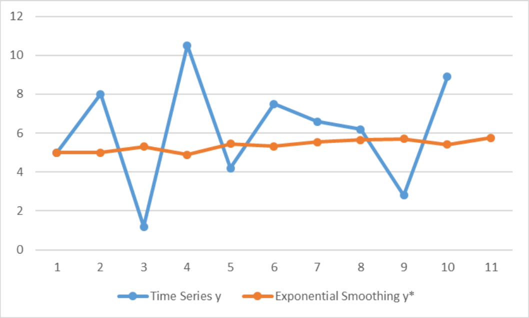
- The explained method of exponential smoothing is called 1-level smoothing
- If the exponentially smoothed time series is **smoothed again**, we obtain 2-level smoothing (and so on)
- Applying smoothing n times results in n-level smoothing or smoothing of order n
- A small change in the procedure of exponential smoothing enables to give a **forecast** of the following value:

$$\boxed{x_1^* = x_1 \quad \text{and} \quad x_{j+1}^* = \alpha \cdot x_j + (1-\alpha) \cdot x_j^*}$$

Analysis of Bivariate Data

Example: Forecast by exponential smoothing for $\alpha = 0.1$

t_i	x_i	x_i^*
1	5.00	5.00
2	8.00	5.00
3	1.20	5.30
4	10.50	4.89
5	4.20	5.45
6	7.50	5.33
7	6.60	5.54
8	6.20	5.65
9	2.80	5.70
10	8.90	5.41
11		5.76

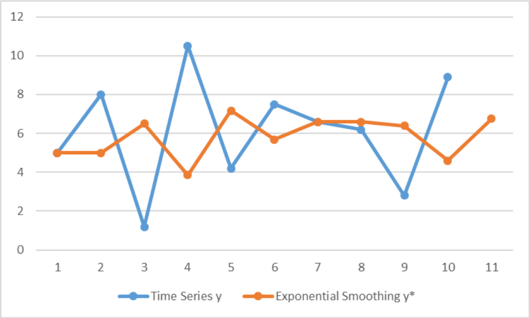


$$x_{11}^* = 0.1 \cdot x_{10} + 0.9 \cdot x_{10}^* = 0.1 \cdot 8.90 + 0.9 \cdot 5.41$$

Analysis of Bivariate Data

Example: Forecast by exponential smoothing for $\alpha = 0.5$

t_i	x_i	x_i^*
1	5.00	5.00
2	8.00	5.00
3	1.20	6.50
4	10.50	3.85
5	4.20	7.18
6	7.50	5.69
7	6.60	6.59
8	6.20	6.60
9	2.80	6.40
10	8.90	4.60
11		6.75



$$x_{11}^* = 0.5 \cdot x_{10} + 0.5 \cdot x_{10}^* = 0.5 \cdot 8.90 + 0.5 \cdot 4.60$$

Exercise

Quarterly production volumes of a garment factory (in mio. pieces) are available for the three consecutive years 2021, 2022 and 2023:

2021	2022	2023
10|12|8|14	14|16|12|18	18|20|16|22

Apply exponential smoothing to the time series with $\alpha = 0.2$ in order to give a **forecast** on the production volume of the first quarter in 2024.