Statistics

Dr. Marc Wellner

Contents

Agenda (1)

Day 1:

8:30 – 10:15	**Introduction**
10:15 – 12:00	**Displaying Descriptive Statistics**
12:00 – 13:00	**Lunch**
13:00 – 14:45	**Calculating Descriptive Statistics I**
14:45 – 16:30	**Calculating Descriptive Statistics II**

Agenda (2)

Day 2:

8:30 – 11:00	**Measuring Concentration**
11:00 – 12:00	**Calculating Price Indexes I**
12:00 – 13:00	**Lunch**
13:00 – 14:30	**Calculating Price Indexes II**
14:45 – 16:30	**Correlation and Regression I**

Agenda (3)

3. Day:	
8:30 – 12:00	**Correlation and Regression II**
12:00 – 13:00	**Lunch**
13:00 – 15:30	**Exercises with Excel**
15:30 – 16:30	**Repetition, Q & A**
4. Day:	**Written examination**

Contents

Introduction – what will you learn here?

In this chapter you will learn about:

- The purpose of statistics – what's in for you?
- Brief overview of the fields of statistics
- The difference between data and information
- Where does data come from?
- What kinds of data can we use
- Different ways of measuring data

Definition of Statistics (1)

- Historically, where does it come from?

  „Statistics“ comes from the Latin word *status*, which means state: It reflects the earliest focus of statistics on measuring things which are related to the state such as the number of:
  - taxable subjects in a kingdom (or state)
  - subjects sent to war
  - …

- Statistics today:

  - Formally (see dictionary):
    “The science that deals with the collection, tabulation and systematic classification of quantitative data, especially as a basis for inference and induction” …. Now that’s a mouthful!

Statistics and for what is it useful

- … Statistics today … more practical ;-):
  - My view of statistics:
    "a way to convert numbers into useful information so that good decisions can be made"
  - … and due to the fact that in today's world we are surrounded by a baggage of data (e.g. gathered via the internet and stored in large data warehouses) statistics becomes more and more important for

    - ourselves: to make the right (and not biased) decisions in daily life (e.g. where do I invest my money? ROI vs. Risk? Stocks, housing, life insurance, …)

    - companies: to achieve a competitive advantage (e.g. what does the customer want? Cross selling and/or up selling? Amazon, Wal Mart, …)

Examples (1)

- Examples in the non–aviation sector:

  Insurances: risk evaluation (which car's needed to have a high insurance?)

  Medicine: effectiveness of new drugs (which cancer therapy works best)

  Banks: credit ratings (sub prime crisis?), stock exchange: forecasting returns and volatilities, portfolio management, fraud detection (credit cards)

  Economics: income distribution (inequality?), evaluation of social programs

  Sports: who is the best player (e.g. in terms of goals, passes, …)

  Manufacturing environment: Quality control → is the process working satisfactorily? (e.g. automotive industry)

Examples (2)

- … some examples in the airline business:

  Market research: surveys on customer satisfaction/behavior and how it is affected (→ conjoint analysis) e.g.: What are the key drivers to book an Airline? Schedule, service, FFP, seats, price, brand, …

  Networkmanagement: Demand forecast, e.g. amount of noshows vs. denied boarding costs, steering of the different booking classes, achievable market share on new routes (market model), fleet assignment, …

  Sales and stations: Revenue forecast, price elasticity, process optimization (what is the optimal lounge size? How many employees are needed at the CKI?)

  Direct marketing: Customer segmentation, e.g.: business vs. leisure travelers, customer life time value, customer equity, share of wallet (→ cluster analysis, regression, …). One to one marketing campaigns: monitoring of effectiveness (control groups)

… statistic helps you to understand paradox results – first example (1)

Simpson Paradox
Example: Test values on reading skills of children differentiated by race

	1981 (average)	2002 (average)	Difference
White	519	527	8
Black	412	431	19
Asian	474	501	27
Mexican	438	446	8
Puerto Rican	437	455	18
Indian	471	479	8
all	506	504	-2

- All groups have increased there reading skills from 1981 to 2002
- Overall the readings skills of the children have decreased!?!

… statistic helps you to understand paradox results – first example (2)

Simpson Paradox
Example: Test values on reading skills of children differentiated by race

	1981 (average)	1981 (share in %)	2002 (average)	2002 (share in %)	Difference
White	519	85	527	65	**8**
Black	412	9	431	11	**19**
Asian	474	3	501	10	**27**
Mexican	438	2	446	4	**8**
Puerto Rican	437	1	455	9	**18**
Indian	471	0	479	1	**8**
all	**506**	**100**	**504**	**100**	-2

→ Reason: the single (group) results are entering the overall result with different weights in the 2 considered years, due to the fact, that the shares of the different races with in the population have changed !

… statistic helps you to understand paradox results – second example (1)

Simpson Paradox
Example: Failure rate for the drivers license differentiated by gender

	male	female
	failure rate	failure rate
1. day	0 %	12,5 %
2. day	33,3 %	50 %
overall	25 %	20 %

- By looking at each day, female always have a higher failure rate
- Looking at the (aggregated) overall result female have a lower failure rate!?!

… statistic helps you to understand paradox results – second example (2)

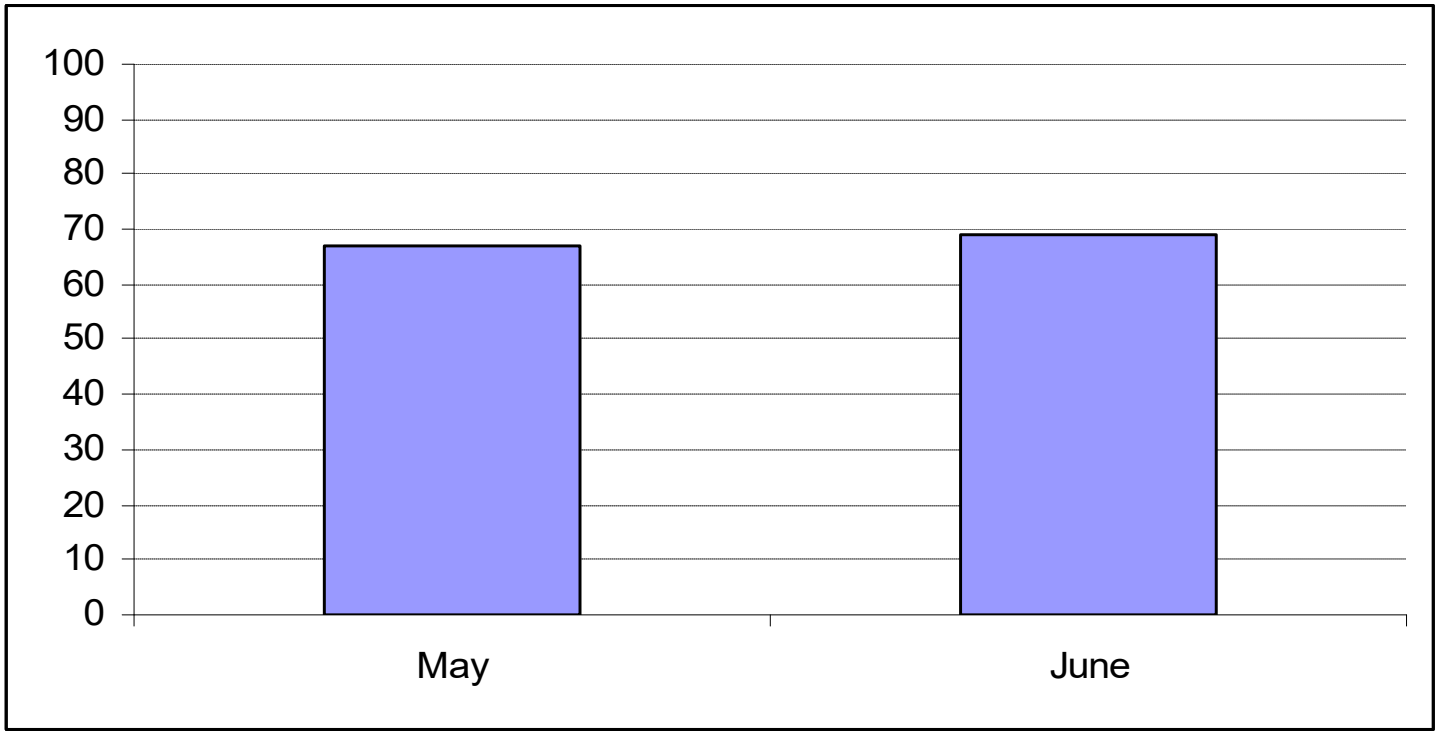
Simpson Paradox
Example: Failure rate for the drivers license differentiated by gender

	male					female				
	failure	total	failure rate	total/ overall	calc. weighted avg.	failure	total	failure rate	total/ overall	calc. weighted avg.
1. day	0	1	0%	25%	0%	1	8	12,5%	80%	10%
2. day	1	3	33%	75%	25%	1	2	50%	20%	10%
overall	1	4	25%	100%	25%	2	10	20%	100%	20%

→ Reason: the single (daily) results are entering the overall result with different weights!

… a good understanding of statistic helps you to be aware (1)

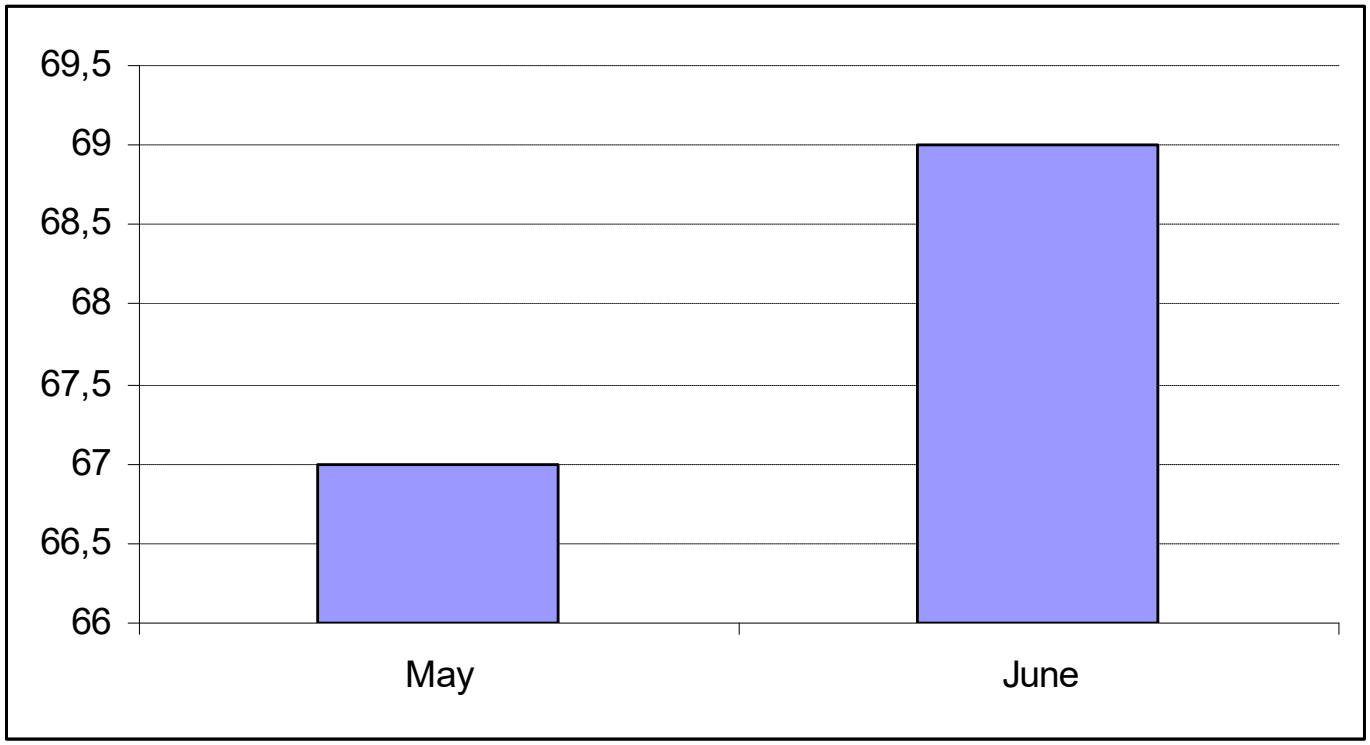
- Seat load factor of an airline in two consecutive months:



→ message: Has only increase slightly

… a good understanding of statistic helps you to be aware (2)

- Seat load factor of an airline in two consecutive months:



→ message: Has increased tremendously (but same figures as in chart before!)

The fields of statistics today

Two basic categories of statistics

descriptive statistics	inferential statistics
→ Allows us to summarize or displays data so that we can quickly obtain an overview	→ Allows us to make claims or conclusions about a population based on a sample of data from that population

- In this course we will focus on descriptive statistics only …
- ... and learn how to perform these techniques by using Microsoft Excel

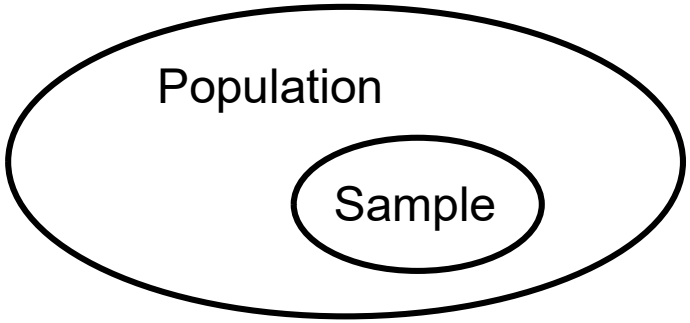
A short excurse to inferential statistics – what is it all about?

- Inferential statistics covers a large variety of techniques that allow us to make actual claims about a population based on a sample of data.

- Example: I'm interested in discovering *in general* who has a longer attention span, Labrador retrievers or, let's say, teenage boys.
  → Problem: Since it's not possible to measure the attention span of *every* teenager and dog, we need to take a sample of each and measure them.

- Relationship between a population and a sample (representative <-> random)

- Requires probability theory



- Further examples: Polls, quality control, credit rating, …

Data, data everywhere and not a drop to drink

- Three building blocks of a statistical study:

- Data is the basic foundation for the field of statistics → the validity of any statistical study hinges on the validity of data!!!
- What's so difficult about data … after all aren't we just talking about numbers?
- Well, we need to learn that data can be classified and measured in several ways, e.g. quantitative vs. qualitative data, grouped vs. non-grouped data, …
- Depending on the type of data different statistical techniques need to be applied

The importance of data (1)

- Definition of data: “The value assigned to a specific observation or a measurement”

- Example: I would like to collect data on my wife’s snoring behavior. There are different ways: I can measure:
  1. how often Astrid snores over 10-minute period
  2. how long the time span between each snore is over a 1-hour period
  3. the length of each snore in seconds
  4. how loud each sore is with a descriptive phrase like
     - “That sounds like a bear just waking up from hibernation”
     - “Wow! That one sounded like an Alaskan seal calling for its young”

  → In each case I’m recording data on the same event in a different form … allowing me to answer different questions via different statistical techniques!

  → the best statistical techniques cannot compensate irrelevant data!

The importance of data (2)

- Example: Table on monthly Airline sales data for the small travel agency "cheap price"
- Data all by its lonesome is not at all useful

Month	Sales ($)
January	15.178
February	14.293
March	13.492
April	12.287
May	11.321

- Using statistical analysis we can generate information, that may be of interest, such as: "Wake up! Something is going wrong. At this rate you will be out of business by early next year!"
- Based on this valuable information, we can make some important decisions about how to avoid this impending disaster

Types of data and level of measurement

- Examples of different data and values:
  - Gender: male vs. female
  - Family status: single, married, divorced, widowed
  - Weight of an adult person: values between 30 and 300 kg
  - Age of a car: values between 1 year and 120 years
  - Education: high school, university, college, apprenticeship, …
  - FFP status: Black, Gold, Silver, Base
  - Types of airplanes: 747-400, 737-800, 320, 319, 380, 340-600, …
  - Seats of airplanes: values between 1 and 550
  - Airline sales area: Asia, Africa & Middle East, Europe, Americas
  - Flight time on the leg FRA-JFK: values between 6 hours and 10 hours
  - Ticket price on the route FRA-JFK vv: values between 399€ and 10.000€
  - Number of Senators in Europe: values between 1 and 150.000
  - …

Types of data and level of measurement

- **Some useful terms**

- **Statistical unit:**
  e.g.: a person, a car, an airplane …

- **Characteristics of a statistical unit:**
  e.g.: age, education, etc. of a person; color, horse power, etc of a car; seats, speed of an airplane ….

- **Value of a characteristic of a statistical unit:**
  23 years, 3 months , high school, university; black, 150 horse power; 200 seats, 900 km/h ….

Types of data

- One way to classify data is:
  - Quantitative: Uses numerical values to describe something of interest
  - Qualitative: Uses descriptive terms to classify something of interest

- Quantitative data can be differentiated into:
  - Discrete data: Finite amount of values (“can be counted”), e.g. number of employees, number of students, number of trucks, …
  - Continuous data: Infinite amount of values (“need to be measured”), e.g. speed of an airplane, height of a building, …

Level of measurement (1)

- Finally data can be classified by the way it is measured, i.e. the scale of measurement:

  - Nominal level of measurement:

    Examples: Gender, religion, nationality, …

    - deals strictly with qualitative data
    - observations are simply assigned to predetermined categories
    - no ranking of categories possible
    - allows no mathematical operations, e.g. in terms of subtraction, in order compare values, only counting makes sense!
    - considered as lowest level of data
    - most restrictive when choosing statistical techniques for the analysis

Level of measurement (2)

- Ordinal level of measurement:

  Examples: School grades, university ranking, movie rankings, hotel categories, …

  - on the food chain of data “the next level up”
  - can be quantitative or qualitative
  - has all the properties of nominal data with the added feature that we can rank order the values from the highest to the lowest
  - does not allow us to make measurements between the categories, e.g. we cannot say a 4 star hotel is four times as good a 1 star hotel
  - still no mathematical operations allowed, e.g. in terms of subtraction, in order to compare values

Level of measurement (3)

- Metrical level of measurement:

  Examples: Age, income, speed, temperature, time, size, …

  - "king of data types"
  - strictly quantitative data
  - allows mathematical operations to compare values, e.g. the difference between the age of two people (25 and 20) provides a meaningful information (5 years older)

  Metrical level of measurement can be further distinguished into:
  - interval level of measurement (e.g. temperature). Has no true 0 point
    → only subtraction and addition lead to meaningful information
  - ratio level of measurement (e.g. age). Has a true 0 point
    → also multiplication and division lead to meaningful information, e.g. "my mother is twice as old as I am.

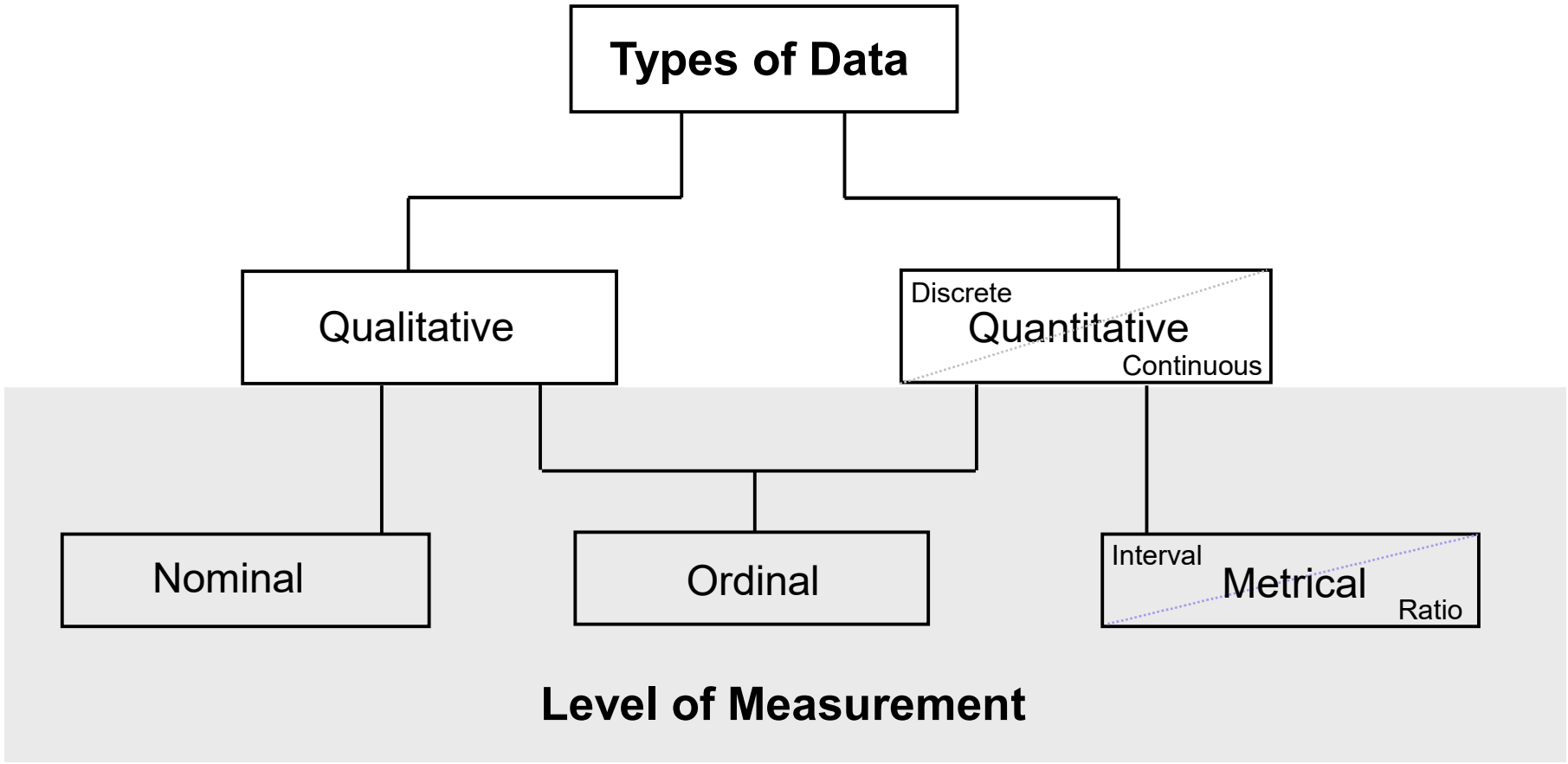
Level of measurement - overview

Level of measurement vs. calculation

Calculation ⇨ / Level ⇩	count	order	math operations (add/multiply)*
Nominal	✓	no	no
Ordinal	✓	✓	no
Metrical (interval/ratio)*	✓	✓	✓

* Distinction between interval level and ratio level is related to the allowed (in terms of usefulness!) mathematical operations

Types of data and level of measurement - overview



Sources of Data – where does all this stuff come from? (1)

- Sources of data can be classified into two broad categories:

  - Primary data:
    - has been collected by the person who eventually uses it
    - drawback of primary data: expensive and time consuming to acquire
    - advantage of primary data: tailored to your needs – allows you to make sure that it is not biased
    - ways to collect primary data:
      - direct observation – “I’ll be watching you!”: E.g. via moderated focus groups
      - experiments – “Who is in control?”: Designed to determine the effectiveness of a treatment, e.g. a new medical drug, via comparison of treatment and control group
      - survey – “Is that you final answer?”: Directly asking the subjects based on a questionnaire … administered e.g. via mail, email or telefone

Sources of Data – where does all this stuff come from? (2)

- Sources of data can be classified into two broad categories:

  - Secondary data:
    - has been collected by somebody else and made available for others to use
    - publishes data on various topics can be found e.g. in the internet:
      - www.census.gov → US Census Bureau provides different types of economic data (people, households, business, industry, …) for the US
      - www.iata.org → IATA provides airline industry data
      - … just google for it … the entries "secondary data download" produce 597.000 results in google ;-)

    - drawback of secondary data: no control over how the data was collected
    - advantage of secondary data: cheap and fast of availability

A short survey ….

1.) Age: _____ in years

2.) Gender: male ○ female ○

3.) Do you have a car? yes ○ no ○
If yes, what is the brand of the car? _____

4.) What hobbies do you have? _____

5.) “The chances that Germany will win the world championship in 2010 are very good”. Please rate this judgment:
disagree ○ partly disagree ○ don’t know ○ partly agree ○ agree ○

6.) How much time do you spend in the internet per week? _____ in minutes

7.) In which country were you born? _____

Some data source at an Airline

- A lot of data at an Airline is stored in a large data warehouse
  - Reservation data
  - Check in data
  - Frequent Flyer Program (FFP) data
  - Feedback data
  - Sales and Revenue data
- For many data there are flexible analysis reports available

Your turn (1)

A1) In order to create a rent index for Frankfurt/Main following data are being collected:

- Rent (monthly in €)
- Living space (in m2)
- Age of the buildings
- Central heating (yes / no)
- Hot water supply (yes / no)
- Location of the flat
- Configuration of the bathroom (high / low)

1. Discuss the collected data regarding the level of measurement!
2. Moreover decide whether the data is discrete or continuous!

Your turn (2)

A2) Classify the following the following data in terms of level of measurement. Explain your choice

- Average monthly rainfall in centimeters for the city of Frankfurt/Main throughout the year
- Education level of the survey respondents: High school, Bachlor's degree, Master's degree, PhD
- FFP status, which passengers show at the check in: Black, Gold, Silver, Base
- Martial status of survey respondents: Single, Married, Divorced
- The year in which the respondent of a survey was born
- Performance rating of employees classified as Above Expectations, Meets Expectations, or Below Expectations
- The country of residence of each respondent in a survey

Your turn (3)

A3) What are the advantages and disadvantages of the following two tables?

Age	Amount
under 10 y.	6
10 until under 20 y.	8
20 until under 30 y.	10
30 until under 40 y.	13
40 until under 50 y.	11
50 until under 60 y.	7
60 until under 70 y.	6
70 until under 80 y.	3
Total	64

Age	Amount
under 20 y.	14
20 until under 40 y.	23
40 until under 60 y.	18
60 until under 80 y.	9
Total	64

Contents

Displaying descriptive statistics – what will you learn here?

In this chapter you will learn about:

- How to construct a frequency distribution for different data, e.g. nominal vs. metrical data
- How to graph a frequency distribution, e.g. with a histogram, pie chart

Frequency distributions

- After having explained various types of data that exist for statistical analysis, here we will explore different ways in which we can present data
- In this chapter we are always considering only one attribute → i.e. the analysis is one dimensional or univarate
- We will learn how to summarize raw data by using frequency distributions
- Absolute frequency: The number of data points which fall into a specific category/class within a frequency distribution
- Relative frequency: The ratio of the absolute frequency to the total number of data points in a frequency distribution.
- Frequency distribution can be presented in tables and/or graphs

Frequency distributions of nominal data (1)

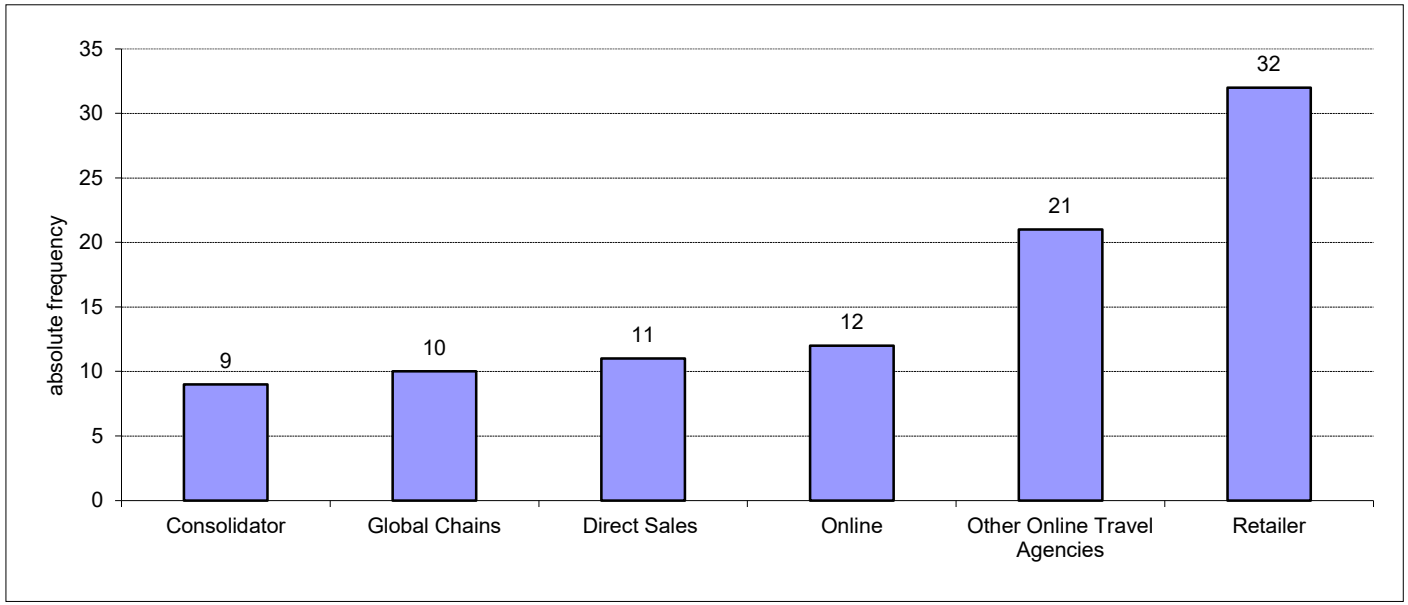
- Example: How do 95 Airline tickets which were bought in one region distribute among the different sales channels?

Sales Channel (1st Level)	absolute frequency	relative frequency
Consolidator	9	9,5%
Global Chains	10	10,5%
Direct Sales	11	11,6%
Online	12	12,6%
Other Online Travel Agencies	21	22,1%
Retailer	32	33,7%
Total	95	100%

→ The frequency distribution tells you that 33,7 % were bought via retailer

Frequency distributions of nominal data (2)

- Graphical presentation as a bar chart



→ The absolute and relative frequency is represented by (proportional to) the height of the bars

Frequency distributions of nominal data (3)

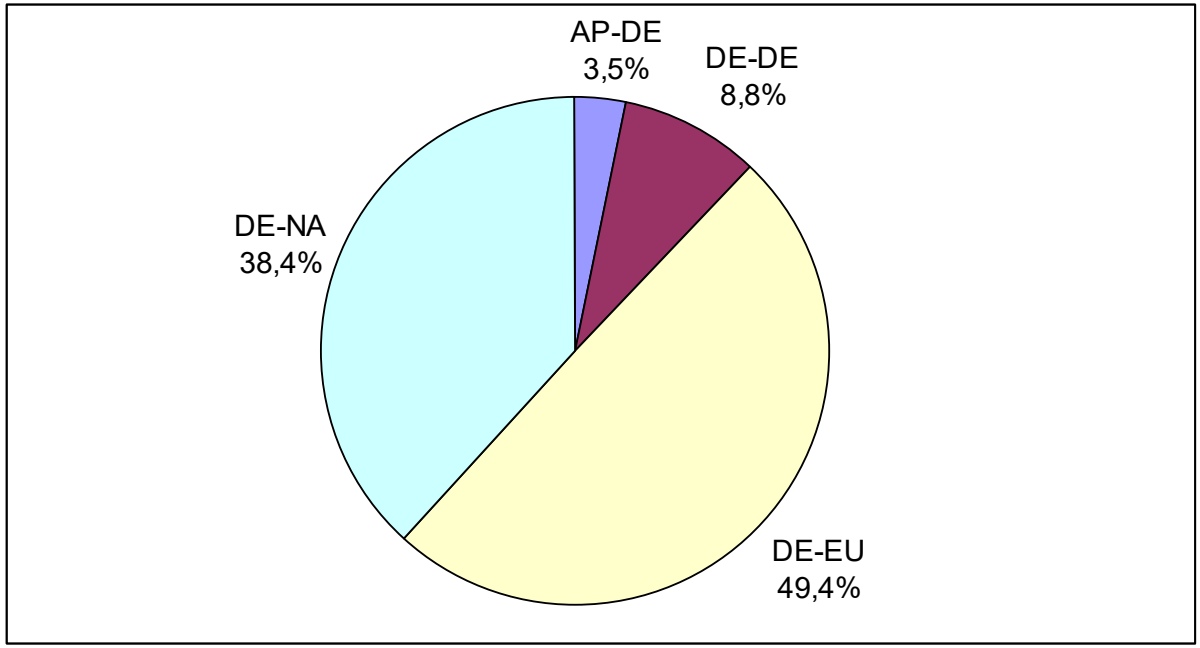
- Example: How do 1.707 bookings of a travel agency distribute with respect to the routing among the area pairs vv?

Area Pair vv	absolute frequency	relative frequency
AP-DE	59	3,5%
DE-DE	150	8,8%
DE-EU	843	49,4%
DE-NA	655	38,4%
Total	1.707	100%

→ The frequency distribution tells you that 8,8 % of the routings are within Germany only

Frequency distributions of nominal data (4)

- Graphical presentation as a pie chart



→ The absolute and relative frequency is represented by (proportional to) the covered areas in the pie chart

Frequency distributions of metrical data (1)

- Example: Assume that a city office generates on several days following revenues (in €): 500, 1.890, 780, 980, 2.000, 1.050, 1.200, 1.450, 800, 1.550, 1.620, 1.850, 920, 1.730, 700, 1.780, 1.810, 1.880, 1.960, 1.980

- By looking at this metrical "raw data" it is difficult to get a feeling about the "typical or usual" daily revenue of the city office

- Transforming this raw data into a frequency distribution helps you to get an overview

- E.g. from the frequency distribution you can see that on 35% of the days the revenue is between 1.800 and 2.100 €

Revenue (in €)	absolute frequency	relative frequency
<300<=600	1	5%
<600<=900	3	15%
<900<=1200	4	20%
<1200<=1500	1	5%
<1500<=1800	4	20%
<1800<=2100	7	35%
Total	20	100%

Frequency distributions of metrical data (2)

- When turning raw data into a frequency distribution it is important to specify the amount and ranges of the frequency intervals or classes

- Five reasonable rules when arranging classes for the frequency distribution:

  1.) From classes of equal size. In the above example covering a range of 300 €

  2.) Make classes mutually exclusive, i.e. overlap between the classes

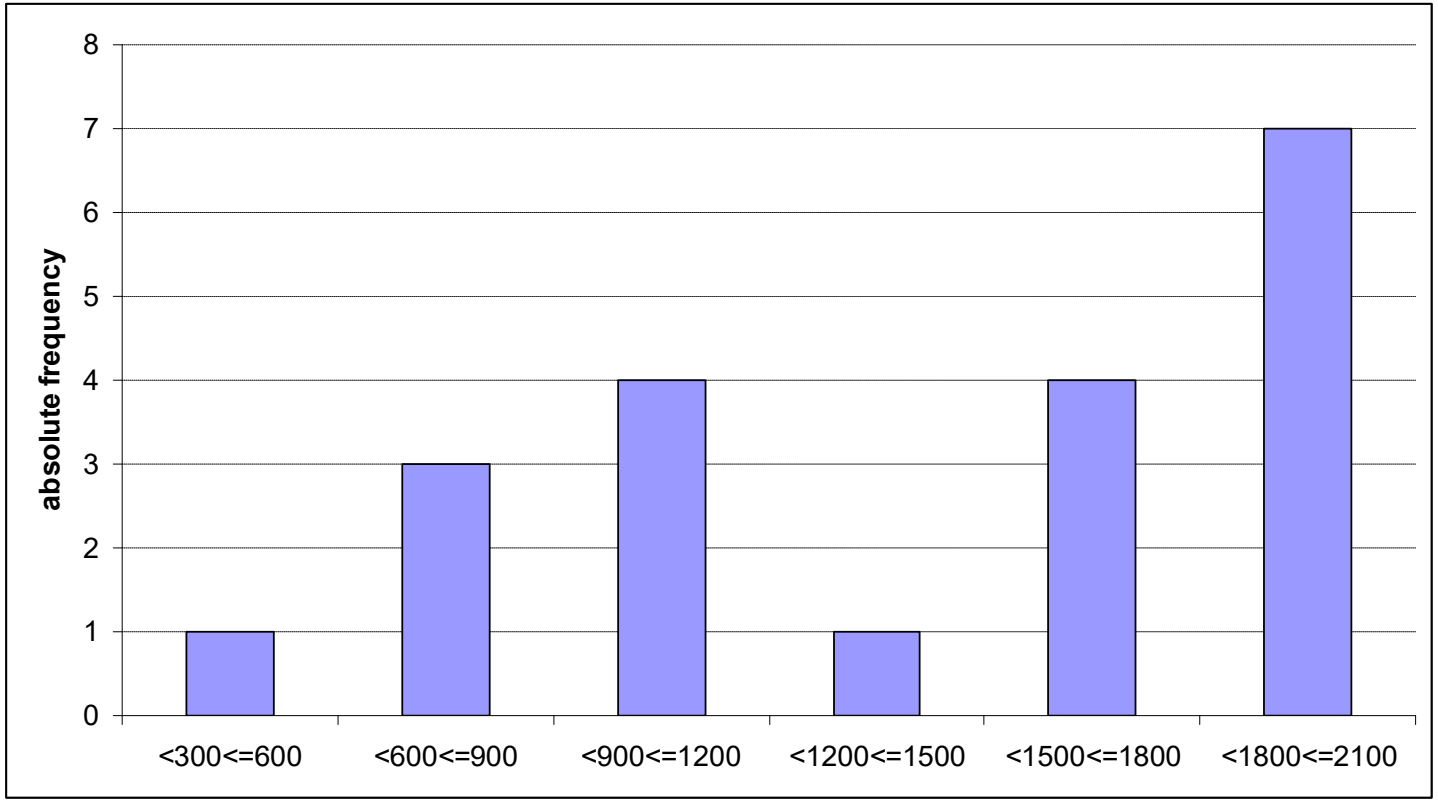
  3.) Do not have to few (e.g. < 3) and not to many (e.g. > 10) classes. I.e. keep the balance between information overflow and information loss.

  4.) If possible (e.g. no outliers) try to avoid open-ended classes, for instance a highest class of 2.100 – over

  5.) Include all data values for the raw data in a class, i.e. the classes should be exhaustive

Frequency distributions of metrical data (3)

- Graphical presentation via histogram



Frequency distributions of metrical data (4)

- Example: How does the yearly overall revenue of one region distribute to the 300 customers

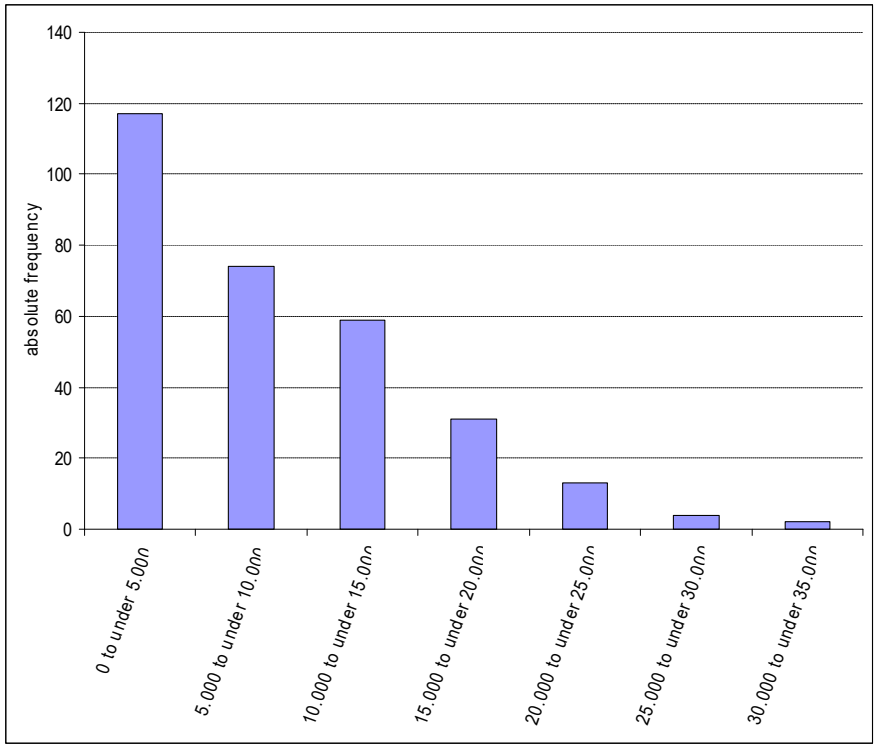
Revenue in € p.a.	absolute frequency	relative frequency	cumulative frequency	cumulative relative frequency
0 to under 5.000€	117	39,0%	117	39,0%
5.000 to under 10.000 €	74	24,7%	191	63,7%
10.000 to under 15.000 €	59	19,7%	250	83,3%
15.000 to under 20.000 €	31	10,3%	281	93,7%
20.000 to under 25.000 €	13	4,3%	294	98,0%
25.000 to under 30.000 €	4	1,3%	298	99,3%
30.000 to under 35.000 €	2	0,7%	300	100,0%
Total	300	1		

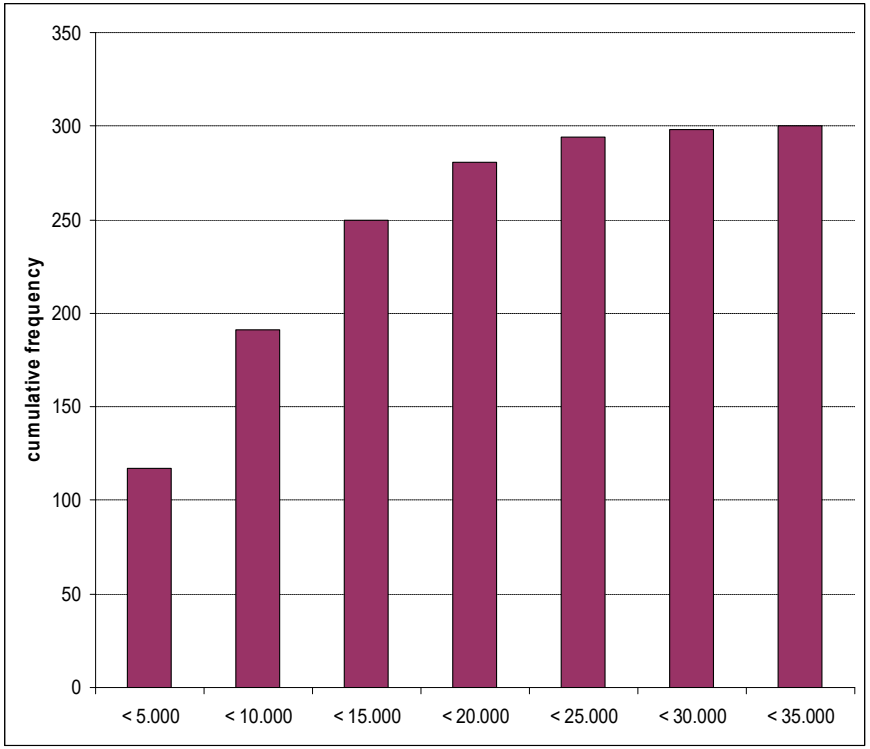
- E.g. 281 customers or 93,7% of all customers have a revenue of less then 20.000 € p.a.

- The cumulative frequency distribution allows you to say how many observation have a value less or equal to a class of interest

Frequency distributions of metrical data (5)

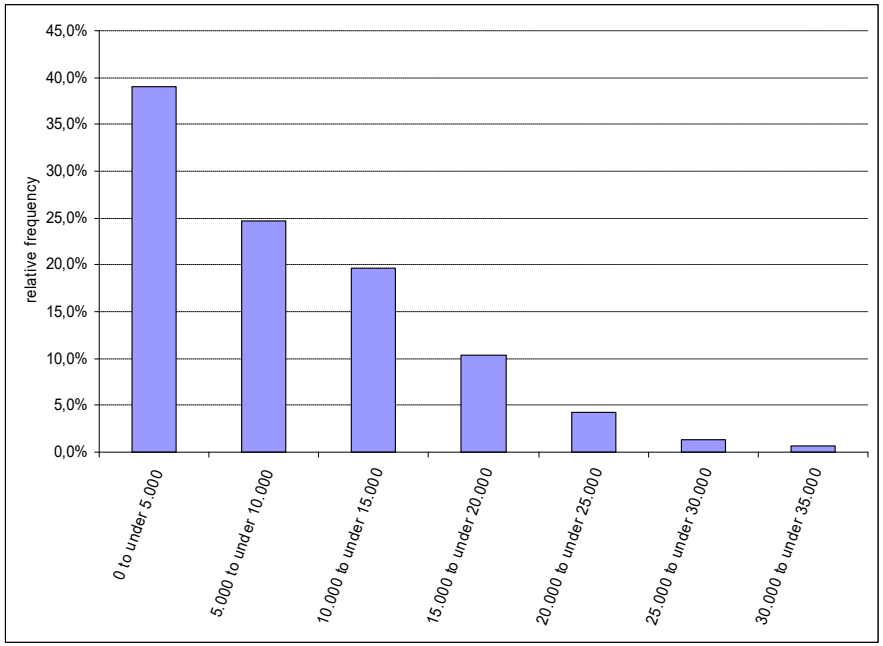
- Graphical presentation of frequency distribution and cumulative frequency distribution

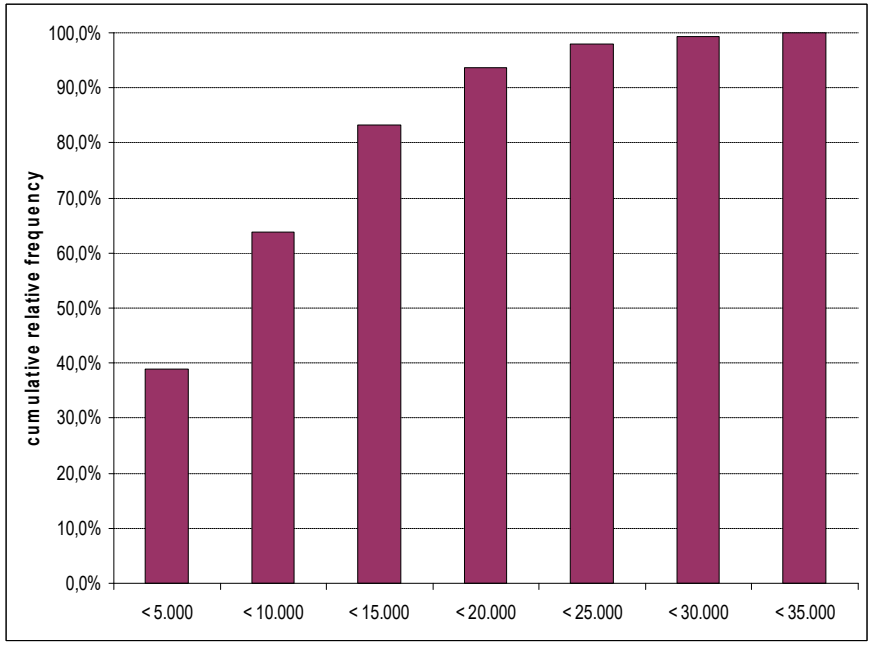




Frequency distributions of metrical data (6)

- Graphical presentation of relative frequency distribution and cumulative relative frequency distribution





Formal presentation of the frequency distribution

- Absolute frequency of a class or value:

$$n(x_i) = n_i$$

- Cumulative absolute frequency:

$$\sum_{j=1}^{i} n_j = n_1 + n_2 + n_3 + \ldots n_i$$

- Relative frequency:

$$h(x_i) = \frac{n_i}{n}$$

- Cumulative relative frequency:

$$\sum_{j=1}^{i} h(x_j) = h(x_1) + h(x_2) + \ldots + h(x_i) = h(x \leq x_i)$$

Your turn (1)

A1) Following table shows the distribution of the duration of member ship (in years) for 100 FFP customers:

Membership duration in years	absolute frequency
0 to under 2	5
2 to under 4	15
4 to under 6	65
6 to under 8	10
8 to under 10	5
Total	100

1.) Calculate the relative frequency, the cumulative frequency as well as the cumulative relative frequency

2.) Graphically display the frequency distribution and the cumulative frequency distribution

Your turn (2)

A2) Following table shows the distribution of different types of airplanes of an airline:

Type of Airplane	absolute frequency
Boing 737	35
Airbus 320	45
Boing 747	20
Airbus 340	15
Total	115

1.) Calculate the relative frequency
2.) Graphically display the frequency distribution

Your turn (3)

A3) The following raw data represents the amount of flights from 36 FFP customers:

Customer	1	2	3	4	5	6	7	8	9	10	11	12
Flights	4	2	1	8	12	7	13	2	5	9	4	3
Customer	13	14	15	16	17	18	19	20	21	22	23	24
Flights	6	7	2	8	10	15	5	8	8	12	7	13
Customer	25	26	27	28	29	30	31	32	33	34	35	36
Flights	2	5	4	2	1	8	12	7	4	3	6	7

1.) Construct a frequency distribution with 5 classes ranging from 1 to 15
2.) Calculate the relative frequency and the cumulative frequency
3.) Display the frequency distribution with a histogram

Your turn (4)

A4) The following raw data represents the advanced booking duration (i.e. the difference between scheduled departure date and booking date) in weeks for all bookings of three specific booking classes Y, M, and W of one specific flight (e.g. UA 400, 24.04.08):

1.) Construct 3 frequency distributions with a useful class allocation
2.) Calculate and display the relative and cumulative relative frequency
3.) Compare the three frequency distributions

	Booking Class				Booking Class		
Week	Y	M	W	Week	Y	M	W
1	21	0	0	28	0	2	11
2	19	0	0	29	0	0	9
3	13	0	0	30	0	1	13
4	10	0	0	31	1	0	9
5	5	0	0	32	0	0	8
6	4	0	0	33	1	3	11
7	3	1	0	34	0	0	12
8	2	12	0	35	0	2	3
9	5	4	0	36	0	0	2
10	7	6	0	37	0	0	5
11	0	7	0	38	0	0	7
12	0	0	0	39	0	0	1
13	0	8	0	40	0	0	3
14	1	0	0	41	0	0	4
15	0	7	0	42	0	0	0
16	2	14	0	43	0	0	6
17	0	4	0	44	0	0	2
18	5	0	0	45	0	0	1
19	0	12	0	46	0	0	4
20	0	0	0	47	0	0	5
21	0	15	0	48	0	0	0
22	0	0	0	49	0	0	0
23	1	4	0	50	0	0	2
24	0	7	1	51	0	0	1
25	1	0	3	52	0	0	0
26	0	3	5	53	0	0	1
27	0	2	7	54	0	0	1

Your turn (5)

A5) The following raw data represents the duration of stay (i.e. the difference between arrival date and departure date at the point of destination) in days for a sample of intercontinental, continental and domestic trips:

1.) Construct 3 frequency distributions with a useful class allocation
2.) Calculate and display the relative and cumulative relative frequency
3.) Compare the three frequency distributions

	Trip		
Days	Interkont	Kont	Domestic
1	2	0	18
2	3	11	20
3	5	16	14
4	2	4	17
5	5	2	5
6	4	1	7
7	14	12	16
8	6	1	4
9	7	4	3
10	8	6	4
11	3	7	1
12	2	0	2
13	7	6	3
14	16	7	1
15	0	7	0
16	2	4	0
17	0	4	0
18	5	0	0
19	2	12	0
20	7	0	0
21	15	11	0

Your turn (6)

A6) The revenue (in 1000 €) of an airline in 10 different markets is given by the following table:

Market	1	2	3	4	5	6	7	8	9	10
Revenue	55	65	49	84	18	105	88	58	12	87

1.) Construct a frequency distribution with 4 classes having a range of 30.000€
2.) Calculate the relative frequency and the cumulative frequency
3.) Display the frequency distribution with a histogram

Contents

Calculating Descriptive Statistics – what will you learn here?

In this chapter you will learn about:

- Understanding central tendency of data
- Calculating different measures of central tendency: mode, median, arithmetic mean and geometric mean
- Understanding dispersion of data
- Calculating different measures of dispersion: range, variance, standard deviation and measures of relative position to identify outlier data values

Descriptive Statistics – Measures of central tendency and dispersion

- In the previous chapter we have demonstrated ways to summarize raw data in a first step, i.e. by using frequency distributions and displaying those via tables or graphically

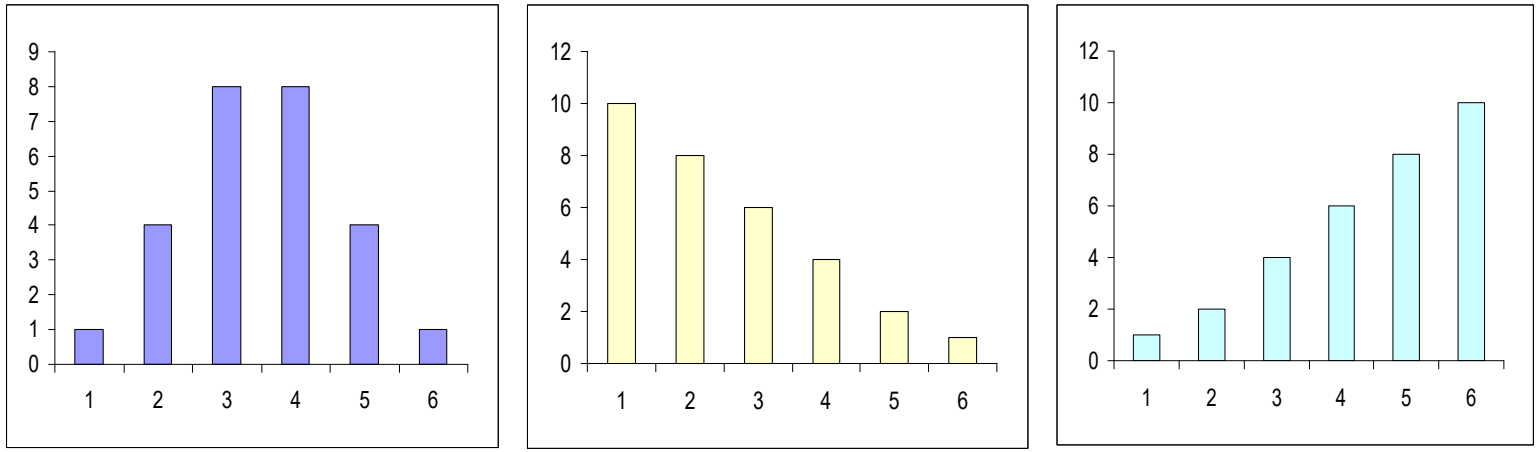
- We will now learn how to summarize the raw data even further by using:

  - Measures of central tendency: Describes with a single value the center point of the data

  - Measures of dispersion: Describes with a single value how the data is spreads around the center point

Descriptive Statistics – Measures of central tendency

- Measures of central tendency: Where is the center of a distribution?
- Three exemplary distributions:



- In the following we will consider four different measures of central tendency:
  - Mode, Median, Arithmetic Mean, Geometric Mean

Mode (1)

- The mode (D) is simply the observation in the data set that occurs the most frequently

- Example: Distribution of grades from 1 (very good) to 6 (unsatisfactory)

Grade	1	2	3	4	5	6	Total
Amount	5	7	14	13	6	3	48

→ In this example the mode equals 3 (D = 3), because the grade 3 occurs most frequently (i.e. 14 times)

- Note that there can be situation, where the are more than one value which occur most frequently (→ not interpretable). Meaningful only when only one value occurs most frequently (i.e. the distribution is unimodal)

Mode (2)

- In case of grouped data (i.e. metrical continuous data which is grouped into classes) the mode is defined as class midpoint with the highest density:

$$f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$$

- Example: Distribution of the revenue from 98 ticket sales

Revenue	Amount of Tickets	Density	Midpoint
0 to under 300 €	10	0,0333	150
300 to under 600 €	32	0,1067	**450**
600 to under 900 €	29	0,0967	750
900 to under 1200 €	12	0,04	1050
1200 to under 1500 €	10	0,0333	1350
1500 to under 1800 €	5	0,0167	1650
Total	98		

→ The mode equals 450 (D = 450), because the midpoint of the class with the highest density equals 450 €

Median (1)

- The median ($\tilde{x}$) is the value in the data set for which half (50%) of the observations are higher and half (50%) of the observations are lower

- To find the median we need to arrange the data values in ascending order and identifying the halfway point:

$$x_{(1)}, x_{(2)}, x_{(3)}, \ldots\ldots, x_{(n)}$$

where:

$x_{(1)} =$ lowest data value

$x_{(n)} =$ highest data value

- If the number of data points is not even, i.e. n is an odd number the median is defined as:

$$\tilde{x} = x_{((n+1)/2)}$$

Median (2)

- Example: For his way from home to work an employee needs on 5 consecutive days 12, 10, 16, 12, 17 minutes. The ordered data values are 10, 12, 12, 16, 17. This leads to:

$$\tilde{x} = x_{((5+1)/2)} = x_{(3)} = 12 \text{ minutes}$$

→ the center point of the ordered data values is 3, hence the median is 12.

→ the data value "12 minutes" is the value in the data set for which half of the observations are to the left (lower or equal) and half of the observations are to the right (equal or higher)

- In case the number of data points is even, i.e. n is an even number the median is defined as the average of the two center points:

$$\tilde{x} = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2)+1)}\right)$$

Median (3)

- Example: The travel agency “Good deal” had following monthly revenues figures (in thousand €) in the last 12 months:

  91, 67, 75, 54, 61, 135, 53, 85, 76, 111, 59, 61.

  The ordered data values are:

  53, 54, 59, 61, 61, 67, 75, 76, 85, 91, 111, 135.

  This leads to:

$$\tilde{x} = \frac{1}{2}\left(x_{(12/2)} + x_{(12/2)+1)}\right) = \frac{1}{2}\left(x_{(6)} + x_{(7)}\right) = \frac{1}{2}(67 + 75) = 71$$

  → the two center points of the ordered data values are 67 and 75, hence the median lies in between, i.e. the median is 71

Median (4)

- In case of grouped data (i.e. metrical continuous data which is grouped into classes) the median is located in class i, in which the cumulative relative frequency exceeds 50% the first time and can be calculated as follows:

$$\tilde{x} = x_i^u + \frac{50 - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$$

where:

x_i^u is the lower bound of the class i in which the median is located

$h(x_i) = \frac{n_i}{n}$ is the relative frequency

$h(x \le x_i) = h(x_1) + h(x_2) + \ldots + h(x_i) = \sum_{j=1}^{i} h(x_j)$ is the cumulative relative frequency

Median (5)

- Example: Monthly salary (in $) of 250 employees

class	salary	class size	abs. frequency	rel. frequency	cum. rel. frequency
1	500 to under 1000	500	6	2,4%	2,4%
2	1000 to under 1500	500	13	5,2%	7,6%
3	1500 to under 2000	500	22	8,8%	16,4%
4	2000 to under 2500	500	32	12,8%	29,2%
5	2500 to under 3000	500	40	16,0%	45,2%
6	3000 to under 3500	500	42	16,8%	62,0%
7	3500 to under 4000	500	39	15,6%	77,6%
8	4000 to under 4500	500	31	12,4%	90,0%
9	4500 to under 5000	500	20	8,0%	98,0%
10	5000 to under 5500	500	5	2,0%	100,0%
			250	100,0%	

→ The median $\tilde{x}$ is given by: $\tilde{x} = 3000 + \frac{50{,}0 - 45{,}2}{16{,}8} \cdot 500 = 3000 + 142{,}86 = 3142{,}86$

Arithmetic mean (1)

- The arithmetic mean (or average) $\bar{x}$ is the most common measure of central tendency. It is calculated by adding all values in our data set and then dividing this result by the number of observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + x_3 + x_4 + \ldots + x_n)$$

- In the previous example where we had 5 data values for the time from home to work (12, 10, 16, 12, 17 minutes) the arithmetic mean is calculated as follows:

$$\bar{x} = \frac{1}{5}(12 + 10 + 16 + 12 + 17) = 13{,}4 \text{ minutes}$$

Arithmetic mean (2)

- In case of grouped data the arithmetic mean is calculated by using weights according to the absolute or relative frequencies:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \quad \text{or} \quad \bar{x} = \sum_{i=1}^{k} x_i h(x_i)$$

- Example: Distribution of grades from 1 (very good) to 6 (unsatisfactory)

Grade	1	2	3	4	5	6	Total
Amount	5	7	14	13	6	3	48

→ The mean is calculated as follows: $\bar{x} = \frac{1 \cdot 5 + 2 \cdot 7 + 3 \cdot 14 + 4 \cdot 13 + 5 \cdot 6 + 6 \cdot 3}{48} = 3{,}35$

Arithmetic mean (3)

- If we have data which is grouped into classes, the calculation is based on the midpoint of each class

- Example: Monthly salary (in $) of 250 employees

class	salary	class size	abs. frequency	rel. frequency	cum. rel. frequency
1	500 to under 1000	500	6	2,4%	2,4%
2	1000 to under 1500	500	13	5,2%	7,6%
3	1500 to under 2000	500	22	8,8%	16,4%
4	2000 to under 2500	500	32	12,8%	29,2%
5	2500 to under 3000	500	40	16,0%	45,2%
6	3000 to under 3500	500	42	16,8%	62,0%
7	3500 to under 4000	500	39	15,6%	77,6%
8	4000 to under 4500	500	31	12,4%	90,0%
9	4500 to under 5000	500	20	8,0%	98,0%
10	5000 to under 5500	500	5	2,0%	100,0%
			250	100,0%	

→ $\bar{x} = 750 \cdot 0{,}024 + 1250 \cdot 0{,}052 + 1750 \cdot 0{,}088 + \ldots + 5250 \cdot 0{,}02 = 3108\$$

Geometric mean (1)

- If we want to calculate the average of growth rates over time, the geometric mean is the correct measure to calculate central tendency
- The geometric mean (GM) is defined as follows:

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots x_n}$$

  i.e. the data values (growth rates) are multiplied and then the n^{th} root has to be taken

- The geometric mean can only be calculated if all data values are positive

Geometric mean (2)

- Example: Yearly revenue the travel agency “happy holiday”

year	revenue in mil. €	growth rate in percent	growth factor
2002	2		
2003	2,4	20%	1,2
2004	2,9	21%	1,2083
2005	2,7	-7%	0,931
2006	3,1	15%	1,1481

- Question: What was the average yearly revenue growth?

$$GM = \sqrt[4]{1{,}2000 \cdot 1{,}2083 \cdot 0{,}9310 \cdot 1{,}1481} = 1{,}1158$$

→ The average yearly growth over the 5 years was: $(1{,}1158 - 1) \cdot 100\% = 11{,}58\%$

Geometric mean (2)

- Example: Yearly revenue the travel agency “happy holiday”

year	revenue in mil. €	growth rate in percent	growth factor
2002	2		
2003	2,4	20%	1,2
2004	2,9	21%	1,2083
2005	2,7	-7%	0,931
2006	3,1	15%	1,1481

- Question: What was the average yearly revenue growth?

$$GM = \sqrt[4]{1{,}2000 \cdot 1{,}2083 \cdot 0{,}9310 \cdot 1{,}1481} = 1{,}1158$$

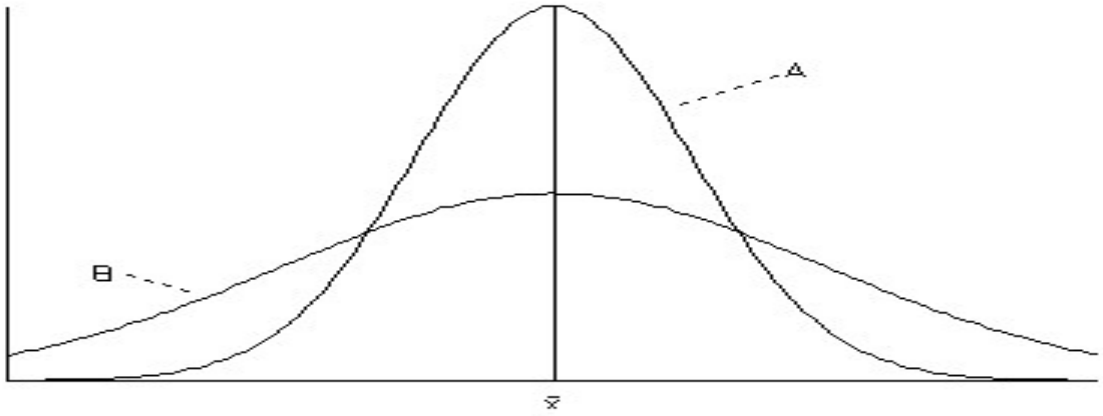
→ The average yearly growth over the 5 years was: $(1{,}1158 - 1) \cdot 100\% = 11{,}58\%$

Descriptive Statistics – Measures of dispersion (1)

- Measures of central tendency only give a limited description of the distribution

  →They do not contain information about how far the individual data values have strayed from the center point, i.e. contain no information about variation

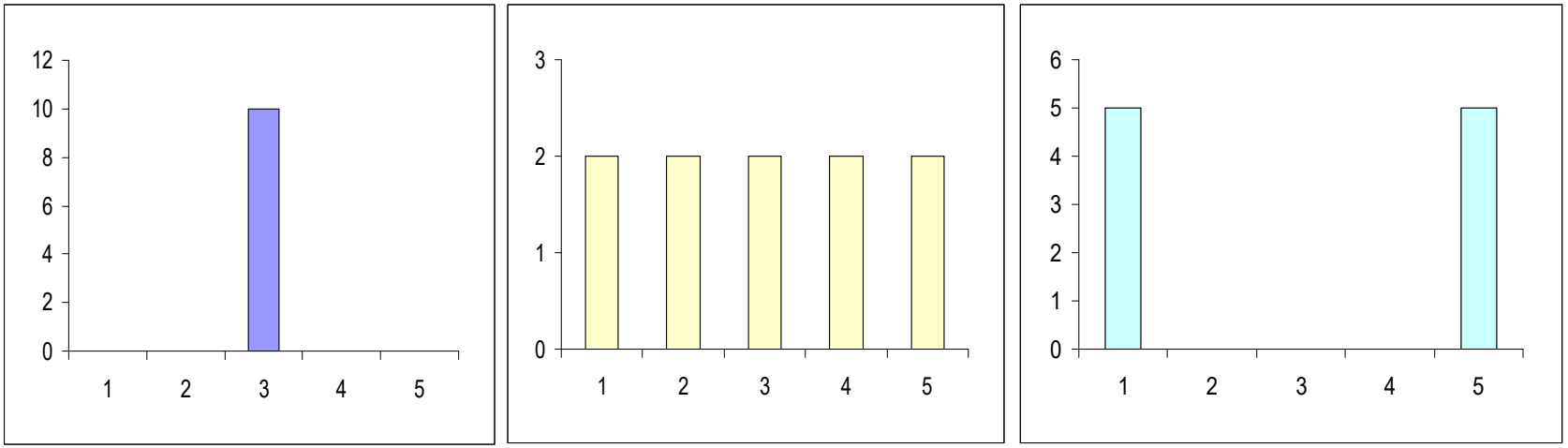
- Measure of dispersion: How far das the data is spread around the center point



→ Distribution A and B have the same arithmetic mean, while has B much higher variation compared to A

Descriptive Statistics – Measures of dispersion (2)

- Example: Three Distributions with the same number of observations, the same arithmetic mean but different dispersions



- Measures of dispersion: The second major category of descriptive statistics
- Along the different measures of central tendency there are different ways to measure dispersion

Range (1)

- The range (R) is the simplest measure of dispersion and is calculated by finding the difference between the highest and lowest value in the distribution:

$$R = x_{(h)} - x_{(l)}$$

- Example: For the monthly revenue figures of travel agency “Good deal” in the last 12 months (in thousand €):

  91, 67, 75, 54, 61, 135, 53, 85, 76, 111, 59, 61.

  the range is:

$$R = 135 - 53 = 82 \text{ thousand €}$$

Range (2)

- In case of grouped data the range is defined by using the lower (upper) boarder of the lowest (highest) class:

$$R = x^{u}_{(h)} - x^{l}_{(l)}$$

- Example: Distribution of the revenue from 98 ticket sales: $R = 1800 - 0 = 1800$

Revenue	Amount of Tickets	Density	Midpoint
0 to under 300 €	10	0,0333	150
300 to under 600 €	32	0,1067	450
600 to under 900 €	29	0,0967	750
900 to under 1200 €	12	0,04	1050
1200 to under 1500 €	10	0,0333	1350
1500 to under 1800 €	5	0,0167	1650
Total	98		

- Disadvantage of range: Influenced only by two values – sensitive to extreme values (e.g. outliers)

Variance and Standard Deviation (1)

- The most common and important measurement of dispersion is the variance (s^2) and the standard deviation which is the square root of the variance (s)
- In case of non grouped date the variance is calculated as follows:

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

- Hence standard deviation is given by:

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

Variance and Standard Deviation (2)

- For the example “time for the way from home to work” that an employee needs on 5 consecutive days (12, 10, 16, 12, 17 minutes) the mean is given by:

$$\bar{x} = \frac{1}{5}(12+10+16+12+17) = 13{,}4\ Minutes$$

- The variance can then be calculated as:

$$\begin{aligned} s^2 &= \frac{1}{5}\left[(12-13{,}4)^2 + (10-13{,}4)^2 + (16-13{,}4)^2 + (12-13{,}4)^2 + (17-13{,}4)^2\right] \\ &= \frac{1}{5}\left[(-1{,}4)^2 + (-3{,}4)^2 + (2{,}6)^2 + (-1{,}4)^2 + (3{,}6)^2\right] \\ &= \frac{1}{5}\left[1{,}96 + 11{,}56 + 6{,}76 + 1{,}96 + 12{,}96\right] = \frac{1}{5}\left[35{,}2\right] = 7{,}04\ Minutes^2 \end{aligned}$$

Variance and Standard Deviation (3)

- Problem when interpreting the variance: Due to taking the square - which is important in order to avoid that positive and negative deviations do not neutralize each other - the dimension of the variance is the square of the original dimension, i.e. minutes^2:
- Hence usage of a measure which is easy to interpret → standard deviation:
- In the example the standard deviation is given by:

$$s = \sqrt{7{,}04\ Minutes^2} = 2{,}65\ Minutes$$

- Interpretation: On average the times for the way from home to work stray by 2,65 minutes around (above and below) the arithmetic mean (13,4 minutes)

Variance and Standard Deviation (4)

- In case of grouped data (i.e. metrical continuous data which is grouped into classes) the variance can be calculated as follows:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$ using the absolute frequency of each class or

$$s^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 h(x_i)$$ using the relative frequency of each class

where x_i is the midpoint of each class. The standard deviation is the square root of the variance: $s = \sqrt{s^2}$

Variance and Standard Deviation (5)

- Example: Monthly salary (in $) of 250 employees → arithmetic mean: 3.108 €

class	salary	class midpoint	abs. frequency	rel. frequency	rel. frequency * class midpoint	midpoint - mean	abs frequency * (class midpoint - mean)^2
1	500 to under 1.000	750	6	2%	18	-2.358	33.360.984
2	1.000 to under 1.500	1.250	13	5%	65	-1.858	44.878.132
3	1.500 to under 2.000	1.750	22	9%	154	-1.358	40.571.608
4	2.000 to under 2.500	2.250	32	13%	288	-858	23.557.248
5	2.500 to under 3.000	2.750	40	16%	440	-358	5.126.560
6	3.000 to under 3.500	3.250	42	17%	546	142	846.888
7	3.500 to under 4.000	3.750	39	16%	585	642	16.074.396
8	4.000 to under 4.500	4.250	31	12%	527	1.142	40.429.084
9	4.500 to under 5.000	4.750	20	8%	380	1.642	53.923.280
10	5.000 to under 5.500	5.250	5	2%	105	2.142	22.940.820
sum			250	100%	3108		281.709.000

Variance can be calculated as follows: $s^2 = \frac{281.709.000}{250} = 1.126.836$

Standard deviation is given by: $s = \sqrt{1.126.836} = 1.061{,}52€$

→ the salary varies on average 1.061,52 € around the arithmetic mean (3.108 €)

Measures of relative position – Quartiles and interquartile measurements (1)

- Quartile and interquartile measurements: Another way of looking at dispersion of data is through measures of relative position, which describe the percentage of the data below a certain point

- The three quartiles divide the data set into four equal segments, after it has been arranged in ascending order:

  - 25% quartile (Q_1): Approx. 25% of the data fall below this quartile

  - 50% quartile (Q_2) i.e. median: Approx. 50% of the data fall below this quartile

  - 75% quartile (Q_3) : Approx. 75% of the data fall below this quartile

Measures of relative position – Quartiles and interquartile measurements (2)

- Based on Q_1 and Q_3 you can calculate interquartile range (IQR), i.e. the spread of the center half (50%) of our data set:

$$IQR = Q_3 - Q_1$$

- Identification of the three quartiles similar to the calculation of the median

- Example (see median): For his way from home to work an employee needs on 5 consecutive days 12, 10, 16, 12, 17, 18 minutes.*

  1. Step: The ordered data values are:

  10 12 12 16 17 18

* Simple visual approach

Measures of relative position – Quartiles and interquartile measurements (3)

2. Step: Find the median, i.e. the 50% quartile (Q_2)

10 12 12 16 17 18 $Q_2 = 14$

3. Step: Find the median of the lower half of the data set, Q_1 (in parenthesis):

(10 12 12) 16 17 18 $Q_1 = 12$ $Q_2 = 14$

4. Step: Find the median of the upper half of the data set, Q_3 (in parenthesis) :

10 12 12 (16 17 18) $Q_1 = 12$ $Q_2 = 14$ $Q_3 = 17$

5. Step: The interquartile range is the difference between Q_3 and Q_1:

$$IQR = Q_3 - Q_1 = 17 - 12 = 5$$

Measures of relative position – Quartiles and interquartile measurements (4)

- Calculation formula* for the quartile:

$$x[q] = \frac{1}{2}\left(x_{nq} + x_{nq+1}\right)$$

where:
- n is the number of observations
- q is the quartile (i.e. 0.25, 0.5, 0.75)
- and the product nq is the corresponding position of the order data values

If the product nq is no integral number, one has to take the next higher integral number <nq> and the corresponding quartile is given by:

$$x[q] = x_{<nq>}$$

- Formally the data values for the 3 quartiles are given by:

$$Q_1 = x[0{,}25] \qquad Q_2 = x[0{,}5] \qquad Q_3 = x[0{,}75]$$

* Leads to slightly different results than the simple visual approach

Measures of relative position – Quartiles and interquartile measurements (5)

- Example: Passenger of top 10 US Airlines in 2007

Top 10 U.S. Airlines, ranked by Passengers in 2007 Passenger numbers in millions (000,000)		
Order	Carrier	Passengers
10	Southwest	101,911
9	American	98,165
8	Delta	72,924
7	United	68,363
6	Northwest	53,678
5	Continental	48,975
4	US Airways	42,172
3	AirTran	23,741
2	Sky West	22,047
1	JetBlue	21,305
Source: Bureau of Transportation Statistics, T-100 Market		

Measures of relative position – Quartiles and interquartile measurements (6)

- To determine $Q_1 = x[0{,}25]$ we first need to calculate $nq = 10 \cdot 0{,}25 = 2{,}5$
  Since nq is not an integral number, we need to take the higher integral number which is 3. Therefore the Q_1 is given by:

$$x[q] = x_{<nq>} = x[0{,}25] = x_3 = 23{,}741$$

- To determine $Q_2 = x[0{,}5]$ we first need to calculate $nq = 10 \cdot 0{,}5 = 5$
  Since nq is an integral number, Q_2 is given by:

$$x[q] = \frac{1}{2}\left(x_{nq} + x_{nq+1}\right) = x[0{,}5] = \frac{1}{2}\left(x_5 + x_6\right) = \frac{1}{2}\left(48{,}975 + 53{,}678\right) = 51{,}327$$

Measures of relative position – Quartiles and interquartile measurements (7)

- To determine $Q_3 = x[0{,}75]$ we first need to calculate $nq = 10 \cdot 0{,}75 = 7{,}5$ Since nq is not an integral number, we need to take the higher integral number which is 8. Therefore the Q_3 is given by:

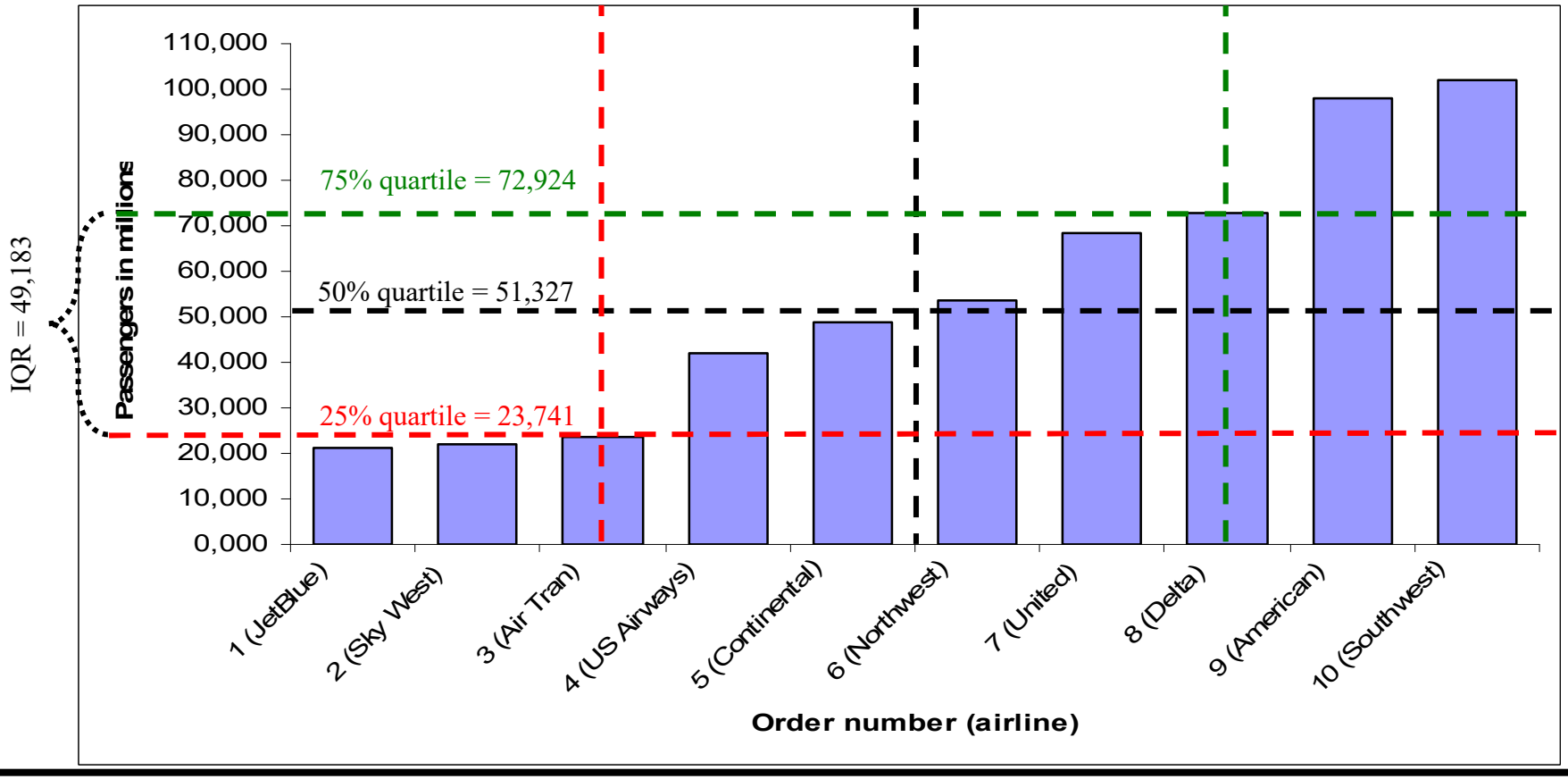
$$x[q] = x_{<nq>} = x[0{,}75] = x_8 = 72{,}924$$

- Finally the interquartile range (IQR) can be calculated as follows:

$$IQR = Q_3 - Q_1 = 72{,}924 - 23{,}741 = 49{,}183$$

Measures of relative position – Quartiles and interquartile measurements (8)

- Graphical visualization



Measures of relative position – Quartiles and interquartile measurements (9)

- In case of grouped data (i.e. metrical continuous data which is grouped into classes) the same logic as for the median applies, i.e. Q_1, Q_2, Q_3 are located in the class in which the cumulative relative frequency exceeds 25%, 50%, 75% the first time and can be calculated as follows:

$$x[0{,}25] = x_i^u + \frac{25 - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$$

$$x[0{,}5] = x_i^u + \frac{50 - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$$

$$x[0{,}75] = x_i^u + \frac{75 - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$$

where:

x_i^u is the lower bound of the class i in which the Q_1, Q_2, Q_3 are located

$h(x_i) = \frac{n_i}{n}$ is the relative frequency

$h(x \le x_i) = h(x_1) + h(x_2) + \ldots + h(x_i) = \sum_{j=1}^{i} h(x_j)$ is the cumulative relative frequency

Measures of relative position – Quartiles and interquartile measurements (10)

- Example: Monthly salary (in $) of 250 employees

class	salary	class size	abs. frequency	rel. frequency	cum. rel. frequency
1	500 to under 1000	500	6	2,4%	2,4%
2	1000 to under 1500	500	13	5,2%	7,6%
3	1500 to under 2000	500	22	8,8%	16,4%
4	2000 to under 2500	500	32	12,8%	29,2%
5	2500 to under 3000	500	40	16,0%	45,2%
6	3000 to under 3500	500	42	16,8%	62,0%
7	3500 to under 4000	500	39	15,6%	77,6%
8	4000 to under 4500	500	31	12,4%	90,0%
9	4500 to under 5000	500	20	8,0%	98,0%
10	5000 to under 5500	500	5	2,0%	100,0%
			250	100,0%	

→ Q_1 is given by: $x[0,25] = 2000 + \frac{25,0 - 16,4}{12,8} \cdot 500 = 2000 + 335,94 = 2335,94$

→ Q_2 is given by: $x[0,5] = 3000 + \frac{50,0 - 45,2}{16,8} \cdot 500 = 3000 + 142,86 = 3142,86$

→ Q_3 is given by: $x[0,75] = 3500 + \frac{75,0 - 62,0}{15,6} \cdot 500 = 3500 + 416,66 = 3916,66$

→ IQR is given by: $IQR = 3916,66 - 2335,94 = 1589,72$

Other Quantiles such as Deciles or Percentiles

- The formulas for the quartile can be used in the same way to calculated **any other quantile** of the distribution such as **deciles**

$$x[0{,}1],\ x[0{,}2],\ x[0{,}3],\ldots,\ x[0{,}9]$$

or **percentiles**

$$x[0{,}01],\ x[0{,}02],\ x[0{,}03],\ldots,\ x[0{,}99]$$

Other Quantiles such as Deciles or Percentiles

- Calculation formula* for any quantile:

$$x[q] = \frac{1}{2}\left(x_{nq} + x_{nq+1}\right)$$

where:
- n is the number of observations
- q is the quantile (e.g. 0.1, 0.15, … ,0.85, 0.99)
- and the product nq is the corresponding position of the order data values

If the product nq is no integral number, one has to take the next higher integral number <nq> and the corresponding quartile is given by:

$$x[q] = x_{<nq>}$$

* Leads to slightly different results than the simple visual approach

Other Quantiles such as Deciles or Percentiles

- **Example:** Salary (in €) of 25 employees

- How much salary do 20% of the employees with the lowest income at most have?

$$nq = 25 \cdot 0.2 = 5$$

$$x[0.2] = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(1760 + 1800) = 1780$$

- How much salary do 10% of the employees with the highest income at least have?

$$nq = 25 \cdot 0.9 = 22{,}5 \quad \rightarrow \quad < nq >= 23$$

$$x[0.9] = x_{23} = 11050$$

Employee	Salary
1	600
2	880
3	1200
4	1430
5	1760
6	1800
7	1800
8	1850
9	2100
10	2220
11	2380
12	2400
13	2580
14	3120
15	3600
16	3680
17	3700
18	5000
19	8000
20	9900
21	9900
22	10100
23	11050
24	12100
25	13000

Other Quantiles such as Deciles or Percentiles

- In case of grouped data (i.e. metrical continuous data which is grouped into classes) the quantiles can be calculated as follows:

$$x[q] = x_i^u + \frac{q - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$$

where:

q is the quantile which is considered (e.g. 0.01, 0.1, …0.9, 0.99)

x_i^u is the lower bound of the class i in which the quantile q is located

$h(x_i) = \frac{n_i}{n}$ is the relative frequency

$h(x \le x_i) = h(x_1) + h(x_2) + \ldots + h(x_i) = \sum_{j=1}^{i} h(x_j)$ is the cumulative relative frequency

Other Quantiles such as Deciles or Percentiles

- Example: Monthly salary (in $) of 250 employees

class	salary	class size	abs. frequency	rel. frequency	cum. rel. frequency
1	500 to under 1000	500	6	2,4%	2,4%
2	1000 to under 1500	500	13	5,2%	7,6%
3	1500 to under 2000	500	22	8,8%	16,4%
4	2000 to under 2500	500	32	12,8%	29,2%
5	2500 to under 3000	500	40	16,0%	45,2%
6	3000 to under 3500	500	42	16,8%	62,0%
7	3500 to under 4000	500	39	15,6%	77,6%
8	4000 to under 4500	500	31	12,4%	90,0%
9	4500 to under 5000	500	20	8,0%	98,0%
10	5000 to under 5500	500	5	2,0%	100,0%
			250	100,0%	

$$x[0,2] = 2000 + \frac{20,0 - 16,4}{12,8} \cdot 500 \quad = 2000 + 140,63 = 2140,63$$

$$x[0,95] = 4500 + \frac{95,0 - 90,0}{8,0} \cdot 500 \quad = 4500 + 312,5 = 4812,5$$

Using quartiles and interquartile measurements to identify any existing outliers in the data set (1)

- Outliers: Are the so called “black sheep” of a data set. These are extreme values whose accuracy is questioned and can cause unwanted distortions in statistical results

- Goal: Identification of any existing outliers in the data set in order to have a detailed look at them and possibly exclude them from any further statistical analysis (e.g. mean) and/or to do sensitivity analysis regarding the robustness of the statistical results (i.e. how does e.g. the mean change with and without the outliers)

- One simple approach: Any values that are greater than (upper bound):

$$Q_3 + 1.5 \cdot IQR$$

or less than (lower bound):

$$Q_1 - 1.5 \cdot IQR$$

are outliers and should be discarded.

Using quartiles and interquartile measurements to identify any existing outliers in the data set (2)

- Example: The travel agency “Good deal” had following monthly revenues figures (in thousand €) in the last 12 months:

  91, 67, 75, 54, 61, 135, 53, 85, 76, 111, 59, 61.

  The ordered data values are:
  53, 54, 59, 61, 61, 67, 75, 76, 85, 91, 111, 115, 135.

  This leads to: $Q_1 = 60 \quad Q_2 = 71 \quad Q_3 = 88 \quad IQR = 28$

  resulting in an upper bound of: $88 + 1.5 \cdot 28 = 130$

  and a lower bound of: $60 - 1.5 \cdot 28 = 18$

  → Hence we would consider the value of 135 thousand € monthly revenue as an outlier

Your turn (1)

A1) Following table shows the distribution of the duration of member ship (in years) for 100 FFP customers:

Membership duration in years	absolute frequency
0 to under 2	5
2 to under 4	15
4 to under 6	65
6 to under 8	10
8 to under 10	5
Total	100

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range

Your turn (2)

A2) The following raw data represents the amount of flights from 36 FFP customers:

Customer	1	2	3	4	5	6	7	8	9	10	11	12
Flights	4	2	1	8	12	7	13	2	5	9	4	3
Customer	13	14	15	16	17	18	19	20	21	22	23	24
Flights	6	7	2	8	10	15	5	8	8	12	7	13
Customer	25	26	27	28	29	30	31	32	33	34	35	36
Flights	2	5	4	2	1	8	12	7	4	3	6	7

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range
3.) How many flights to 85% of the FFP customers at least have

Your turn (3)

A3) The following raw data represents the advanced booking duration (i.e. the difference between scheduled departure date and booking date) in weeks for all bookings of three specific booking classes Y, M, and W of one specific flight (e.g. UA 400, 24.04.08):

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range
3.) Are there any outliers
4.) How early do 95% of all Y, M, W class bookings occur

	Booking Class				Booking Class		
Week	Y	M	W	Week	Y	M	W
1	21	0	0	28	0	2	11
2	19	0	0	29	0	0	9
3	13	0	0	30	0	1	13
4	10	0	0	31	1	0	9
5	5	0	0	32	0	0	8
6	4	0	0	33	1	3	11
7	3	1	0	34	0	0	12
8	2	12	0	35	0	2	3
9	5	4	0	36	0	0	2
10	7	6	0	37	0	0	5
11	0	7	0	38	0	0	7
12	0	0	0	39	0	0	1
13	0	8	0	40	0	0	3
14	1	0	0	41	0	0	4
15	0	7	0	42	0	0	0
16	2	14	0	43	0	0	6
17	0	4	0	44	0	0	2
18	5	0	0	45	0	0	1
19	0	12	0	46	0	0	4
20	0	0	0	47	0	0	5
21	0	15	0	48	0	0	0
22	0	0	0	49	0	0	0
23	1	4	0	50	0	0	2
24	0	7	1	51	0	0	1
25	1	0	3	52	0	0	0
26	0	3	5	53	0	0	1
27	0	2	7	54	0	0	1

Your turn (4)

A4) The following raw data represents the duration of stay (i.e. the difference between arrival date and departure date at the point of destination) in days for a sample of intercontinental, continental and domestic trips:

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range
3.) Are there any outliers
4.) How long are the 10% longest trip durations of the intercont, cont and domestic trips

	Trip		
Days	Interkont	Kont	Domestic
1	2	0	18
2	3	11	20
3	5	16	14
4	2	4	17
5	5	2	5
6	4	1	7
7	14	12	16
8	6	1	4
9	7	4	3
10	8	6	4
11	3	7	1
12	2	0	2
13	7	6	3
14	16	7	1
15	0	7	0
16	2	4	0
17	0	4	0
18	5	0	0
19	2	12	0
20	7	0	0
21	15	11	0

Your turn (5)

A5) The revenue (in 1000 €) of an airline in 10 different markets is given by the following table:

Market	1	2	3	4	5	6	7	8	9	10
Revenue	55	65	49	84	18	105	88	58	12	87

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range
3.) Are there any outliers?
4.) How much revenue do 40% of the markets at least generate?

Your turn (6)

A1) Following table shows the distribution of yearly revenue (in €) of 125 customers:

Revenue p.a.	# customers
A (12000-14000€)	1
B (10000-12000€)	2
C (8000-10000€)	3
D (6000-8000€)	13
E (4000-6000€)	69
F (2000-4000€)	22
G (0-2000€)	15

1.) Calculate the modus, mean, median and range
2.) Calculate the variance, standard deviation and the interquartile range
3.) How much Revenue p.a. is at least generated by the top 10% customers with the highest revenue p.a.

Contents

Concentration measures – what will you learn here?

In this chapter you will learn about:

- Ways to measure concentration, e.g. how competitive is a market in terms of how equally is the market revenue shared across the different competitors.
- How to graphically display concentration by means of the Lorenz Curve
- How to calculate concentration by using the GINI coefficient

Concentration measures – the Lorenz Curve

- The standard deviation is already a measure for concentration:
  → *s* measures the average deviation from perfect equality (i.e. a uniform distribution), since the mean $(\bar{x})$ is the value, which would result when the sum of all data values is distributed equally to all observations

- BUT: More information about concentration can be obtained from a graphical representation of the deviation from the uniform distribution, i.e. comparing the existing inequality (concentration) with perfect equality → **Lorenz Curve**

- Typical questions to be analyzed:

  - Income/wealth concentration: How is income/wealth (the attribute) distributed to all individuals or households (observations)

  - Market concentration: How is the overall market revenue distributed to all existing competitors (companies)? Monopoly, …, competition

- Note, that the Lorenz Curve can only be calculated in case of non negative data values measured on a metrical level

Concentration measures – the Lorenz Curve

- **Example:** Market concentration in the airline industry in 3 countries A, B and C. The following table displays the revenue distribution (in 1000€) of 5 airlines in the three countries.

	Country		
Airline	A	B	C
1 ("airquality")	40	180	60
2 ("cheap buy air"	40	5	50
3 ("only eco air")	40	5	40
4 ("air no delay")	40	5	30
5 ("air premium")	40	5	20
Total	200	200	200

→ same overall market revenue in all 3 countries (total) of 200.000 €.
→ but in country A we have an equal distribution of the overall market revenue to all airlines
→ in contrast in country B the total revenue is strongly concentrated on airline
→ in country C we have a slight concentration towards airline 1 and 2

- **Goal of the Lorenz Curve**: Visualization of concentration/inequality in comparison to an equal distribution

Concentration measures – the Lorenz Curve in case of ungrouped data

- **5 steps to calculate the Lorenz Curve in case of ungrouped data:**

1.) The data values (observations) have to be sorted, i.e. $x_1 \leq x_2 \leq x_3 \ldots \leq x_n$

2.) Calculation of the ratio for the first j statistical units (units of observation):

$$u_j = \frac{j}{n}$$ → cumulative relative frequency of the *first j* statistical units

3.) Calculation of the cumulative relative sum of data values for the *first j* statistical units:

$$p_j = \sum_{i=1}^{j} x_i$$ → cumulative absolute sum of data values for the *first j* statistical units

$$p_n = \sum_{i=1}^{n} x_i$$ → cumulative absolute sum of data values for *all* statistical units („totals“)

$$v_j = \frac{p_j}{p_n}$$ → cumulative relative sum of data values for the *first j* statistical units

Concentration measures – the Lorenz Curve in case of ungrouped data

- **5 steps to calculate the Lorenz Curve in case of ungrouped data:**

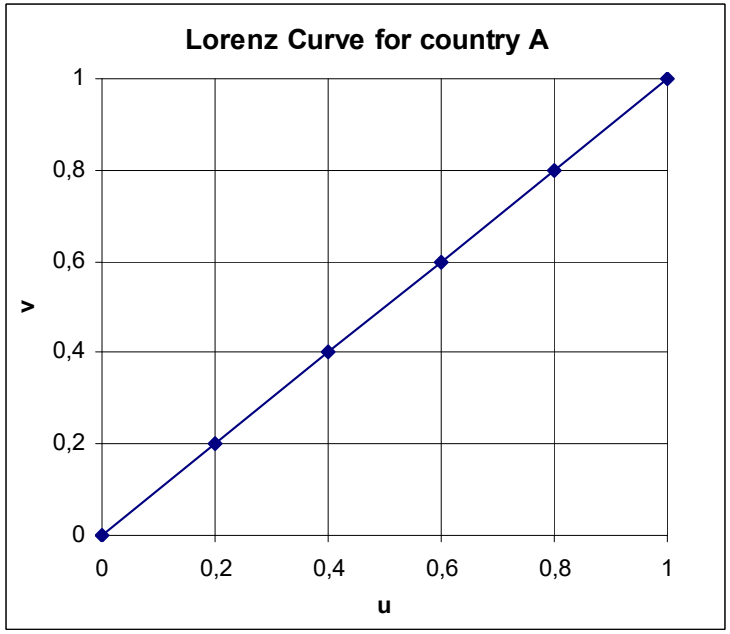
  4.) Graphical representation of the value pairs u_j, v_j, where the cumulative relative sum of the statistical units (u) is displayed on the x-axis and the cumulative relative sum of data values (v) is displayed on the y-axis

  5.) As a comparison we always consider the Lorenz Curve in case of a perfect equal distribution, i.e. the diagonal through 0,0 and 1,1

Concentration measures – the Lorenz Curve in case of ungrouped data

- **Example airline market concentration in country A**

	Country A			
Airline	u_j	x_i	p_j	v_j
1 ("airquality")	0,2	40	40	0,2
2 ("cheap buy air")	0,4	40	80	0,4
3 ("only eco air")	0,6	40	120	0,6
4 ("air no delay")	0,8	40	160	0,8
5 ("air premium")	1	40	200	1

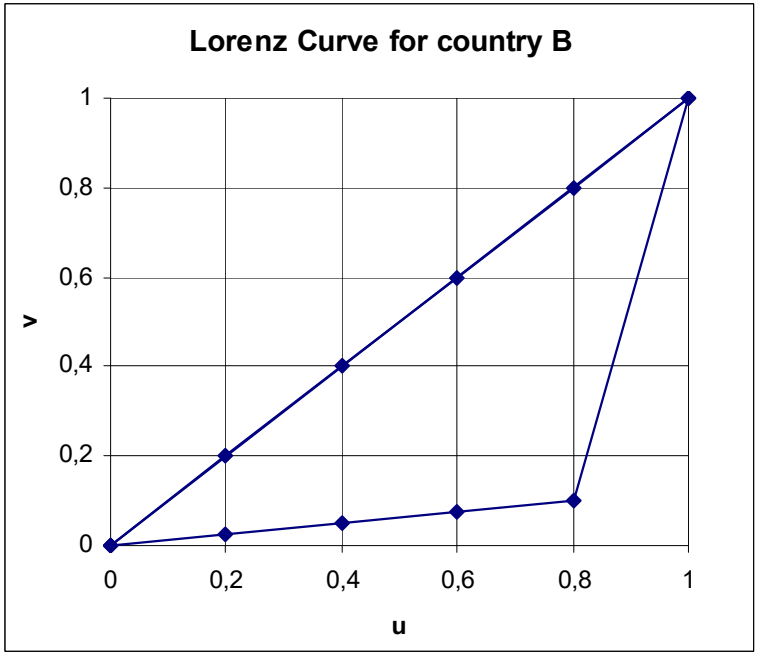


→ perfect equal distribution, i.e. all airlines have the same market share: E.g. 20% of the airlines generate 20% of the total market revenue, 40% of the airlines generate 40% of the total market revenue, 60% of the airlines generate 60% of the total market revenue, 80% of the airlines ...

Concentration measures – the Lorenz Curve in case of ungrouped data

- **Example airline market concentration in country B**

	Country B			
Airline	u_j	x_i	p_j	v_j
2 ("cheap buy air")	0,2	5	5	0,025
3 ("only eco air")	0,4	5	10	0,050
4 ("air no delay")	0,6	5	15	0,075
5 ("air premium")	0,8	5	20	0,100
1 ("airquality")	1	180	200	1,000

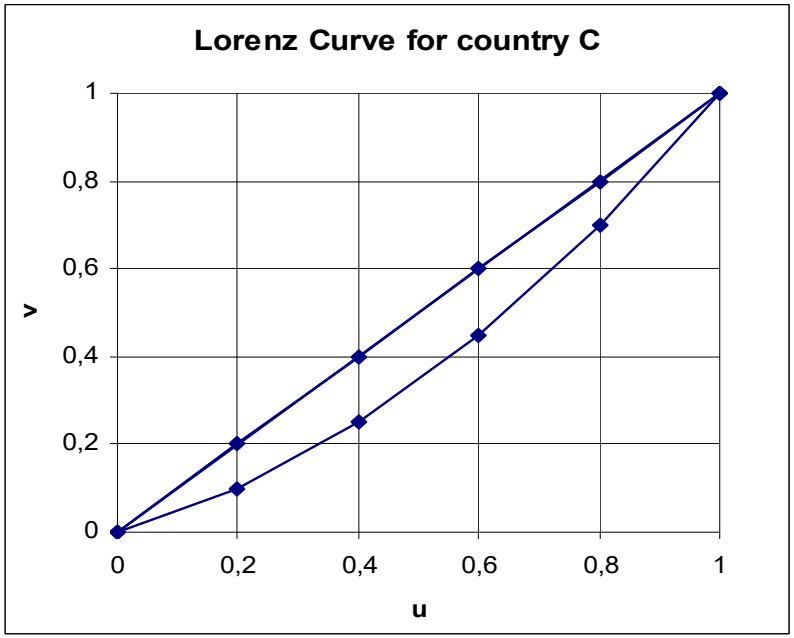


→ strong concentration (inequality): E.g. 40% of the airlines with the lowest revenue generate only 5% of the total market revenue while 20% of the airlines with the highest revenue generate 90% of the total market revenue

Concentration measures – the Lorenz Curve in case of ungrouped data

- **Example airline market concentration in country C**

	Country C			
Airline	u_j	x_i	p_j	v_j
5 ("air premium")	0,2	20	20	0,10
4 ("air no delay")	0,4	30	50	0,25
3 ("only eco air")	0,6	40	90	0,45
2 ("cheap buy air")	0,8	50	140	0,70
1 ("airquality")	1	60	200	1,00



→ some concentration: E.g. 40% of the airlines with the lowest revenue generate only 25% of the total market revenue while 40% of the airlines with the highest revenue generate 55% of the total market revenue

Concentration measures – the Lorenz Curve in case of grouped/classified data

- **5 steps to calculate the Lorenz Curve in case of grouped/classified data:**

1.) Use of absolute frequencies of each class: $n_1, n_2, \cdots, n$

2.) Calculation of the ratio for the first j statistical units (units of observation):

$$u_j = \frac{1}{n}\sum_{i=1}^{j} n_i$$ → cumulative relative frequency of the *first j* statistical units

3.) Calculation of the cumulative relative sum of data values for the *first j* statistical units using the class midpoints (x_i represents the class midpoint)

$$p_j = \sum_{i=1}^{j} n_i \cdot x_i$$ → cumulative absolute sum of data values for the *first j* statistical units

$$p_n = \sum_{i=1}^{n} n_i \cdot x_i$$ → cumulative absolute sum of data values for *all* statistical units („totals“)

$$v_j = \frac{p_j}{p_n}$$ → cumulative relative sum of data values for the *first j* statistical units

Concentration measures – the Lorenz Curve in case of grouped/classified data

- **5 steps to calculate the Lorenz Curve in case of grouped/classified data:**

  4.) Graphical representation of the value pairs u_j , v_j, where the cumulative relative sum of the statistical units (u) is displayed on the x-axis and the cumulative relative sum of data values (v) is displayed on the y-axis

  5.) As a comparison we always consider the Lorenz Curve in case of a perfect equal distribution, i.e. the diagonal through 0,0 and 1,1

Concentration measures – the Lorenz Curve in case of grouped/classified data

- **Example:** Grouped/classified distribution (fictitious) of all airlines according to the number of airplanes which they have in use

number of airplanes	number of airlines
1 bis 2	67
3 bis 5	88
6 bis 10	80
11 bis 20	92
21 bis 30	58
31 bis 50	65
51 bis 100	53
101 bis 500	22
	525

- **Question:** How are the airplanes in use distributed to all airlines? Rather equally or unequally?

Concentration measures – the Lorenz Curve in case of grouped/classified data

- **Step 1 and 2:** Calculation of $u_j = \frac{1}{n}\sum_{i=1}^{j} n_i$, i.e. the cumulative relative frequency of the first j statistical units

class number	number of airplanes	number of airlines	cum. Number of airlines	cum. relative number of airlines (u_j)
1	1 bis 2	67	67	0,128
2	3 bis 5	88	155	0,295
3	6 bis 10	80	235	0,448
4	11 bis 20	92	327	0,623
5	21 bis 30	58	385	0,733
6	31 bis 50	65	450	0,857
7	51 bis 100	53	503	0,958
8	101 bis 500	22	525	1,000

Concentration measures – the Lorenz Curve in case of grouped/classified data

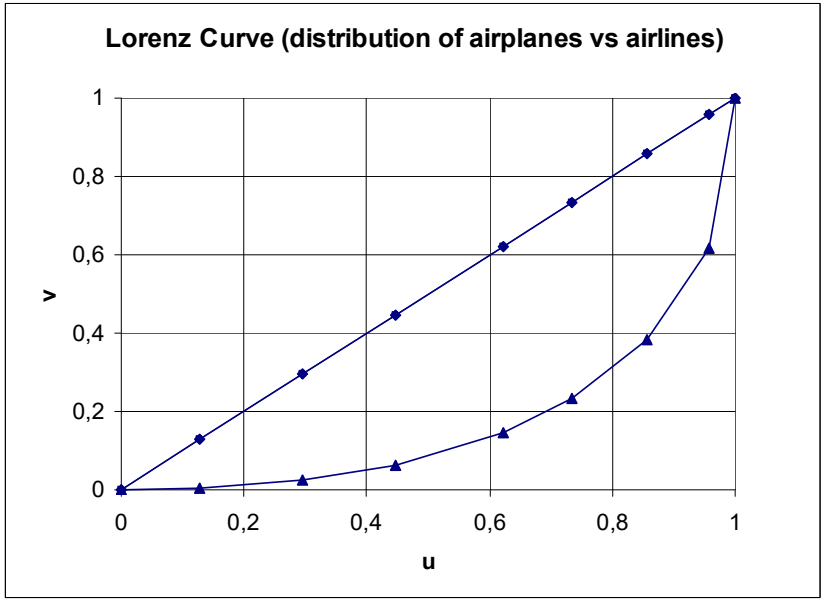
- **Step 3:** Calculation of $v_j = \frac{p_j}{p_n}$, i.e. the cumulative relative sum of data values of the first j statistical units

class number	number of airplanes	class mid point	number of airlines	absolute sum of data values of class i	cum. absolute sum of data values (p_j)	cum. relative sum of data values (v_j)
1	1 bis 2	1,5	67	100,5	100,5	0,006
2	3 bis 5	4	88	352	452,5	0,026
3	6 bis 10	8	80	640	1092,5	0,063
4	11 bis 20	15,5	92	1426	2518,5	0,146
5	21 bis 30	25,5	58	1479	3997,5	0,232
6	31 bis 50	40,5	65	2632,5	6630	0,385
7	51 bis 100	75,5	53	4001,5	10631,5	0,617
8	101 bis 500	300,5	22	6611	17242,5	1,000

Concentration measures – the Lorenz Curve in case of grouped/classified data

- **Step 4 and 5:** Displaying the Lorenz Curve (i.e. the data value pairs u_j, v_j) and comparison with the perfect equal distribution

u_j	v_j
0,000	0,000
0,128	0,006
0,295	0,026
0,448	0,063
0,623	0,146
0,733	0,232
0,857	0,385
0,958	0,617
1,000	1,000



→ The 29,5% smallest airlines own only 2,6% of all airplanes. The 4,2% largest airlines own 38,3% of all airplanes

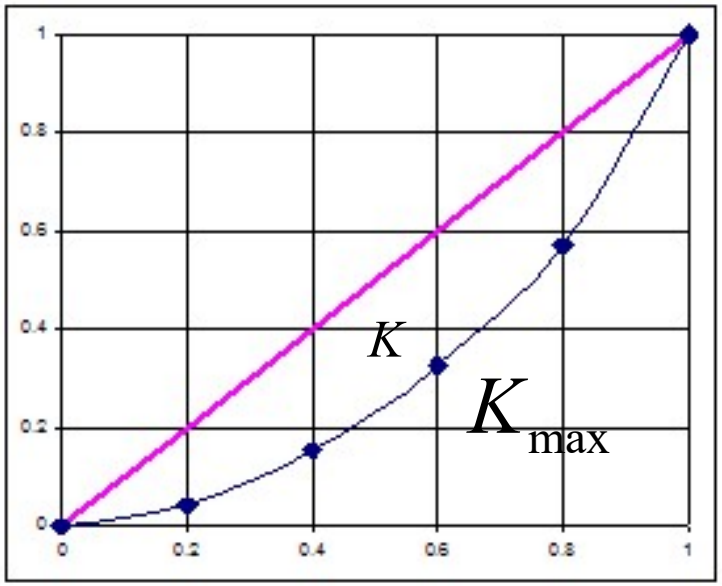
Concentration measures – the Gini coefficient

- The Gini Coefficient is a complementary way of presenting information about inequality.
- It is the ratio of the area between the Lorenz Curve and the line of absolute equality (actual concentration area → numerator) and the whole area under the line of absolute equality (maximum concentration area → denominator).

$$GINI = \frac{act.conc.area}{\max.conc.area} = \frac{K}{K_{\max}}$$

where $K_{\max} = \frac{1}{2}$

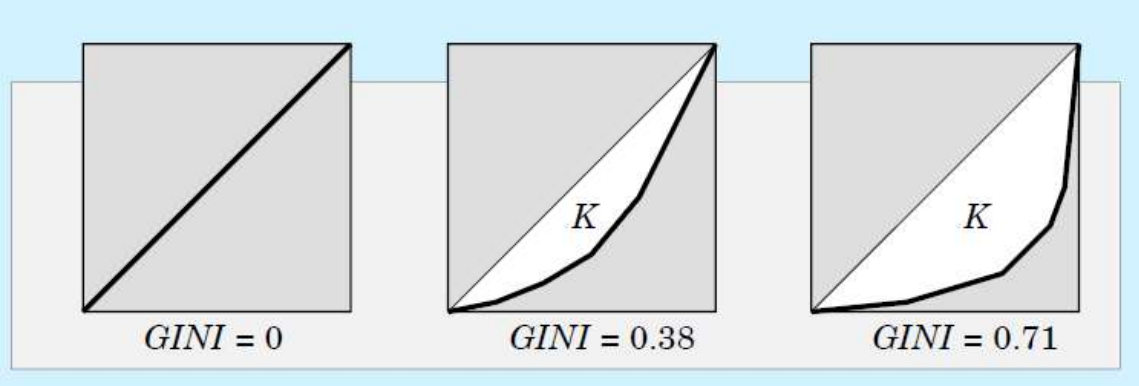
→ The Gini Coefficient is a single aggregated concentration figure which measures the shape ("the size of the belly") of the Lorenz Curve



Concentration measures – the Gini coefficient

- The Gini Coefficient is standardized with extreme values of 0 and 1:
  - 0 implies perfect equality (e.g. company has exactly the same amount of revenue), i.e. the actual Lorenz Curve corresponds with the Lorenz Curve in case of perfect equality (diagonal)
  - 1 implies total inequality (e.g. one has all the revenue and everyone else has nothing), i.e. the actual Lorenz Curve corresponds with the Lorenz Curve in case of perfect inequality (triangle)

- Key thing to bear in mind: The lower (higher) the figure that Gini Coefficient takes, the greater the degree of prevailing equality (inequality).



Concentration measures – the Gini coefficient

- To calculate the Gini coefficient we need to measure the actual concentration area K
- K can be expressed as the difference between

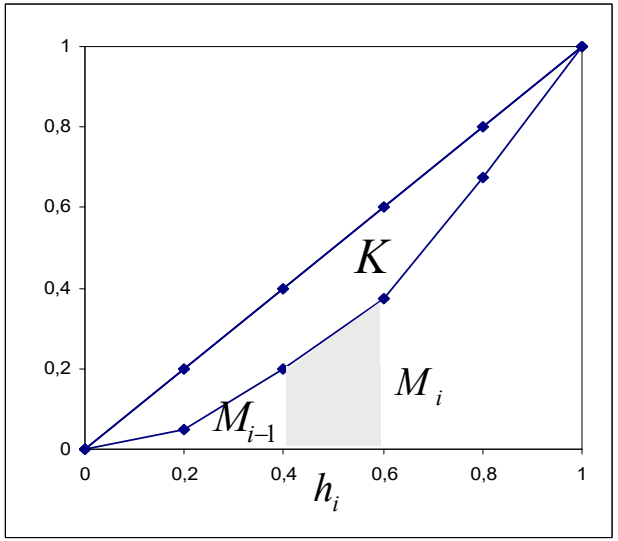
$$K_{\max} = \frac{1}{2}$$

and the area *under* the actual Lorenz Curve

$$\sum_{j=1}^{k} \frac{1}{2}\left(M_{j-1} + M_j\right) \cdot h_j$$

i.e. the calculation formula for K is:

$$K = \frac{1}{2} - \sum_{j=1}^{k} \frac{1}{2}\left(M_{j-1} + M_j\right) \cdot h_j$$



$$= \frac{1}{2}\left(M_{i-1} + M_i\right) \cdot h_i$$

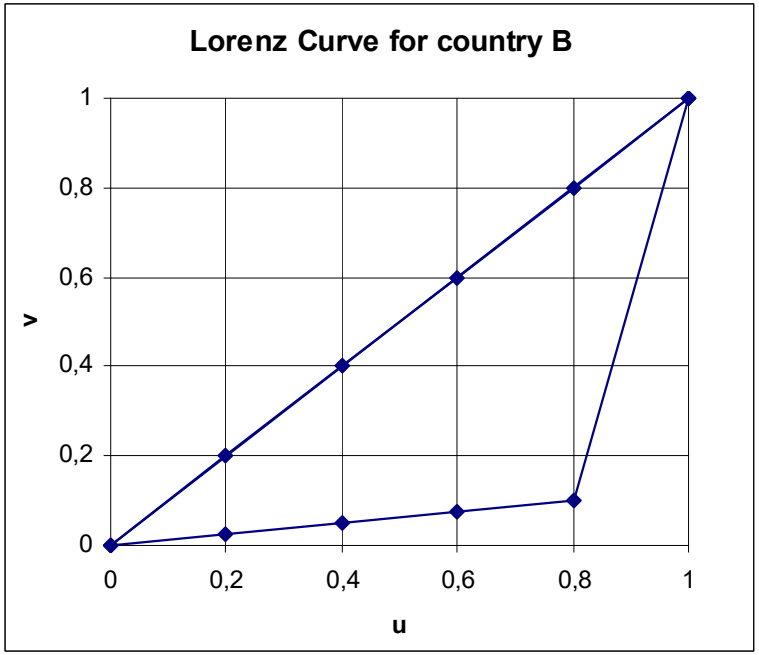
$$= \frac{1}{2}\left(M_{i-1} \cdot h_i + \text{(triangle)} \cdot h_i + M_i \right)$$

Concentration measures – the Gini coefficient

- **Example airline market concentration in country B**

	Country B			
Airline	u_j	x_i	p_j	v_j
2 ("cheap buy air")	0,2	5	5	0,025
3 ("only eco air")	0,4	5	10	0,050
4 ("air no delay")	0,6	5	15	0,075
5 ("air premium")	0,8	5	20	0,100
1 ("airquality")	1	180	200	1,000

$$K = \frac{1}{2} - \left\{ \left(\frac{1}{2}(0 + 0{,}025) \cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}025 + 0{,}05) \cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}05 + 0{,}075) \cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}075 + 0{,}1) \cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}1 + 1) \cdot 0{,}2 \right) \right\} = 0{,}35$$



$$\Rightarrow GINI = \frac{K}{K_{\max}} = \frac{0{,}35}{0{,}5} = 0{,}7$$

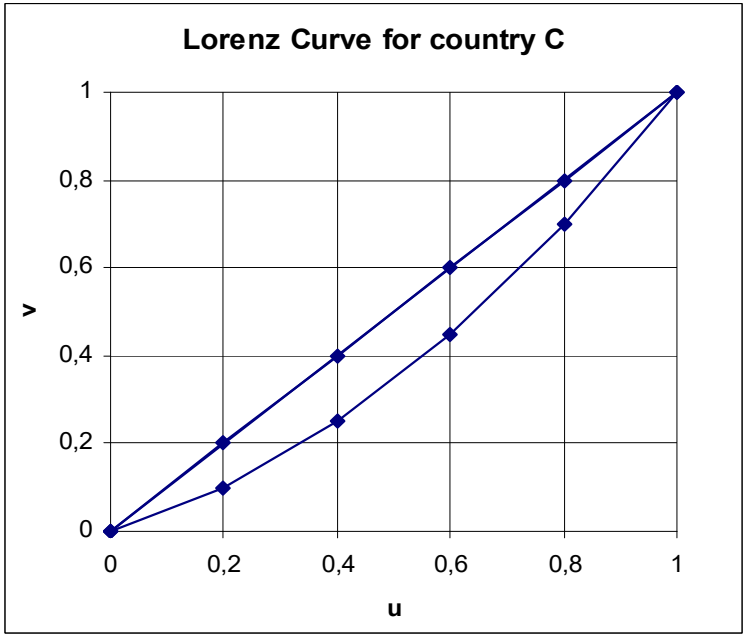
0,7 is close to 1 → high concentration

Concentration measures – the Gini coefficient

- **Example airline market concentration in country C**

	Country C			
Airline	u_j	x_i	p_j	v_j
5 ("air premium")	0,2	20	20	0,10
4 ("air no delay")	0,4	30	50	0,25
3 ("only eco air")	0,6	40	90	0,45
2 ("cheap buy air")	0,8	50	140	0,70
1 ("airquality")	1	60	200	1,00

$$K = \frac{1}{2} - \left\{ \left(\frac{1}{2}(0+0{,}1)\cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}1+0{,}25)\cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}25+0{,}45)\cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}45+0{,}7)\cdot 0{,}2 \right) + \left(\frac{1}{2}(0{,}7+1)\cdot 0{,}2 \right) \right\} = 0{,}1$$



$$\Rightarrow GINI = \frac{K}{K_{\max}} = \frac{0{,}1}{0{,}5} = 0{,}2$$

0,2 is close to 0 → low concentration

Your turn (1)

A1) The following table shows the number cars which were produced in 1995 by different automotive companies.

	amount in thousand
Opel	743
Ford	350
BMW	563
VW	838
Audi	447
Mercedes	550
Porsche	18

1.) Calculate and display the Lorenz Curve.
2.) Calculate the Gini Coefficient.
3.) What share of the total car production is being produced by the approximately 30% of the automotive companies with the lowest production volume?

Your turn (2)

A2) The following table shows Passenger of top 10 US Airlines in 2007.

1.) Calculate and display the Lorenz Curve.
2.) Calculate the Gini Coefficient.
3.) What share of passengers are approximately flying with the 60% of the airlines with the lowest pax volume?
4.) Compare the concentration in the US airline market with the concentration in German automotive market (previous exercise)

Top 10 U.S. Airlines, ranked by Passengers in 2007 Passenger numbers in millions (000,000)		
Order	Carrier	Passengers
10	Southwest	101,911
9	American	98,165
8	Delta	72,924
7	United	68,363
6	Northwest	53,678
5	Continental	48,975
4	US Airways	42,172
3	AirTran	23,741
2	Sky West	22,047
1	JetBlue	21,305
Source: Bureau of Transportation Statistics, T-100 Market		

Your turn (3)

A3) The following table shows the share of aggregate income of US households in different years

Year	Lowest 20%	Next Lowest 20%	Middle 20%	Second Highest 20%	Highest 20%
1968	4,2	11,1	17,5	24,4	42,8
1982	4,1	10,1	16,6	24,7	44,5
1992	3,8	9,4	15,8	24,2	46,9
2001	3,5	8,7	14,6	23,0	50,1

1.) Calculate and display the Lorenz Curve for each year.
2.) Calculate the Gini Coefficient for each year.
3.) How has concentration changed over the years?

Your turn (4)

A4) The following table shows the net income of employees of a company classified according to their income level.

Net income in EUR	Amount of employees
under 1000	800
1000 to under 2000	1600
2000 to under 3000	1900
3000 to under 4000	3300
4000 to under 5000	2000
5000 to under 6000	400

1.) Calculate the necessary values for the Lorenz Curve an draw it.
2.) Calculate the GINI coefficient and give a statement regarding the concentration.
3.) How high is the share of the total net income which is given to the 50% employees with the lowest income?

Contents

Index numbers and price indexes – what will you learn here?

In this chapter you will learn about:

- How to make the development of different single variables comparable by using index numbers
- How to measure the development of a group of variables – such as prices of different goods – over time. E.g. how has the overall price level in an economy developed?

Index numbers

- **Goals of index numbers** is to describe the
  - development over time
  - regional differences

  of a group of variables and to make these developments/differences comparable to other groups of variables

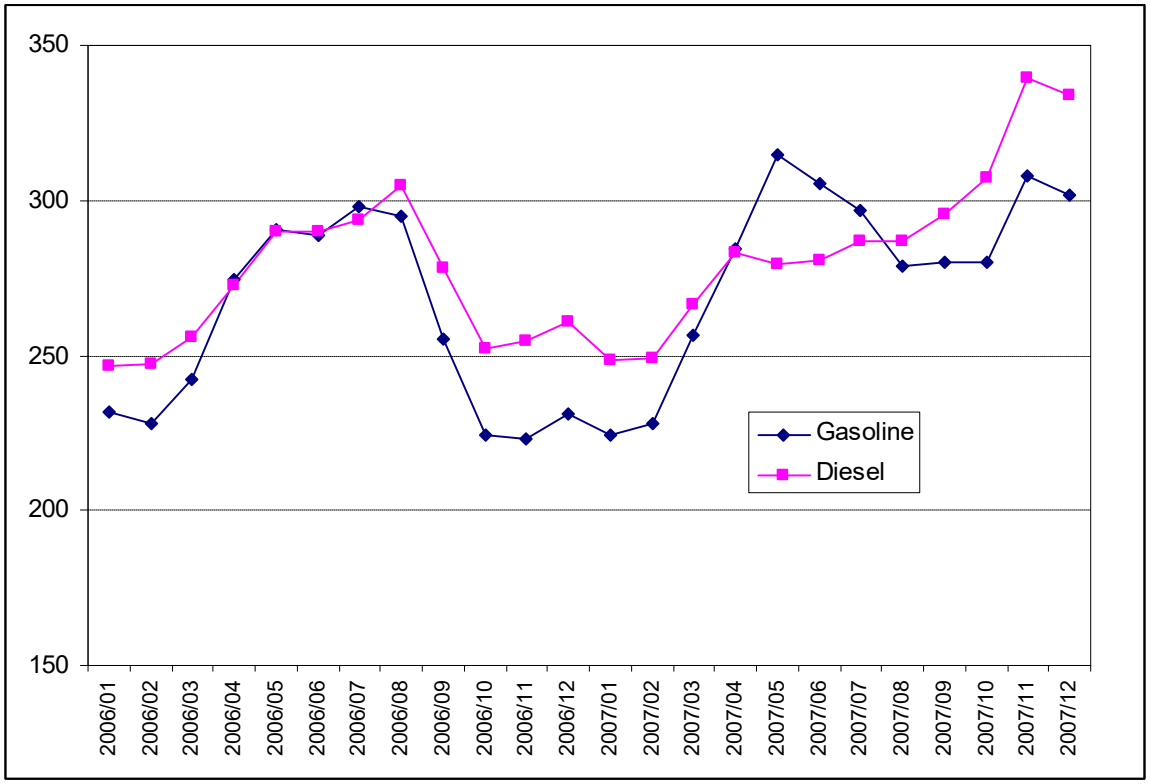
- When only **one variable** is considered this is a straight forward task: Define one observation of the original time series as the basis of the index number, i.e. as 100, and then relate the other index values to observations of the original time series correspondingly:

$$x_t = 100 \cdot \frac{y_t}{y_0} \quad \text{for} \quad y_0 \Rightarrow x_0 = 100$$

Index numbers

- **Example:** Comparison of price development of gasoline and diesel (measured in cents per gallon)

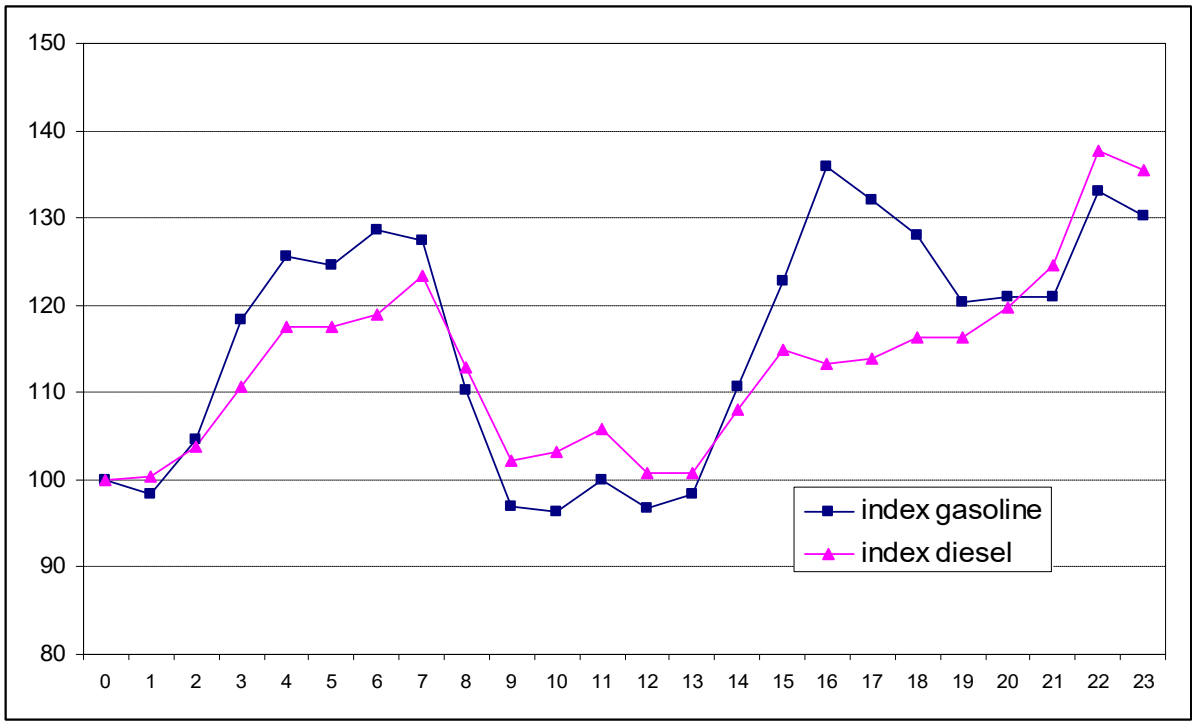
Year/Month	Gasoline	Diesel
2006/01	231,6	246,7
2006/02	228,0	247,5
2006/03	242,5	255,9
2006/04	274,2	272,8
2006/05	290,7	289,7
2006/06	288,5	289,8
2006/07	298,1	293,4
2006/08	295,2	304,5
2006/09	255,5	278,3
2006/10	224,5	251,9
2006/11	222,9	254,5
2006/12	231,3	261,0
2007/01	224,0	248,5
2007/02	227,8	248,8
2007/03	256,3	266,7
2007/04	284,5	283,4
2007/05	314,6	279,6
2007/06	305,6	280,8
2007/07	296,5	286,8
2007/08	278,6	286,9
2007/09	280,3	295,3
2007/10	280,3	307,5
2007/11	308,0	339,6
2007/12	301,8	334,1



Index numbers

- **Question:** Has gasoline or diesel increased more since January 2006 (Basis 100)?

t	index gasoline	index diesel
0	100,0	100,0
1	98,4	100,3
2	104,7	103,7
3	118,4	110,6
4	125,5	117,4
5	124,6	117,5
6	128,7	118,9
7	127,5	123,4
8	110,3	112,8
9	96,9	102,1
10	96,2	103,2
11	99,9	105,8
12	96,7	100,7
13	98,4	100,9
14	110,7	108,1
15	122,8	114,9
16	135,8	113,3
17	132,0	113,8
18	128,0	116,3
19	120,3	116,3
20	121,0	119,7
21	121,0	124,6
22	133,0	137,7
23	130,3	135,4



Diesel has increased by 35% while gasoline only by 30%

Prices index

- **Price index:** Aggregation of different prices for a given class of goods or services, i.e. a basket of goods and services, into one index. In this case a **group of different variables** need to be aggregate into one index
- **Problem of price indexes:** The aggregation typically involves averaging the different prices. Since the average is typically a weighted average the **problem of adequate weights arises**
- The price index is a statistic, designed to help to **compare how the considered prices, taken as a whole, i.e. the price level, differ between time periods**, i.e. the base period (0), and the reporting period (t) for which the index is computed
- Price indices have several **potential uses**:
  - Economics: The index can be said to measure the economy's price level or a cost of living
  - Business: Price indices help companies with business plans and pricing

Price index

- **Example:**

Good	Price		Quantity	
Nr. i	Base period 0	Reporting period t	Base period 0	Reporting period t
1	4	6	5	5
2	6	8	10	15
3	10	12	18	16

Definition:

p_0^i : price of good i in the base period 0

q_0^i : quantity of good i in the base period 0

p_t^i : price of good i in the reporting period t

q_t^i : quantity of good i in the reporting period t

Price index

- Calculation of price indexes for each good

$$I_{0,t}\left(p^i\right)=\frac{p_t^i}{p_0^i}\cdot 100$$

- In the previous example we get following 3 price indexes

$$I_{0,t}\left(p^1\right)=\frac{6}{4}\cdot 100=150\% \qquad I_{0,t}\left(p^2\right)=\frac{8}{6}\cdot 100=133\% \qquad I_{0,t}\left(p^3\right)=\frac{12}{10}\cdot 100=120\%$$

- **Simple construction** of an overall price index to measure the price level development. Simply take the average of the single price indexes:

$$I_{0,t}\left(p\right)=\frac{I_{0,t}\left(p^1\right)+I_{0,t}\left(p^2\right)+I_{0,t}\left(p^3\right)}{3}=134\%$$

Price index

- Generally this simple aggregated price index can be written as:

$$I_{0,t}(p) = \frac{1}{n} \cdot \sum_{i=1}^{n} I_{0,t}(p^i)$$

- This simple aggregate price index is however **unsatisfactory because the average does not take into account the quantities of the considered goods.** Problem of adequate weighting when averaging the prices (→ what is the correct weighted average?)

- Not only consideration of the price vector of each period:

$$p_0^1, p_0^2, p_0^3 \quad \text{vs.} \quad p_t^1, p_t^2, p_t^3$$

but also consideration of the quantities, i.e. the so called basket of goods / commodities:

$$q_0^1, q_0^2, q_0^3 \quad \text{vs.} \quad q_t^1, q_t^2, q_t^3$$

Price index

- But how should the quantities be considered when constructing a weighted average? Two approaches:
  - Price index according to Laspeyres
  - Price index according to Paasche

- The **Laspeyres price index** is defined as:

$$I_{0,t}^{L}(p) = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} \cdot 100$$

here the basket of goods of the base period 0, $q_0^1, \ldots, q_0^i, \ldots, q_0^n$, is used for weighting, i.e. considered as constant

Price index

- In the previous example the Laspeyres price index is given by:

$$I_{0,t}^{L}(p) = \frac{326}{260} \cdot 100 = 125{,}4\%$$

in the base period 0 the value of the basket of goods is 260. In the reporting period t the same basket of goods (q_0^1, q_0^2, q_0^3) would have a value of 326.

→ The ratio of the 2 values quantifies the average change of prices with regards to the quantities of the base period 0

Price index

- The **Paasche price index** is defined as:

$$I_{0,t}^{P}(p) = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i} \cdot 100$$

here the basket of goods of the reporting period t, $q_t^1, \ldots, q_t^i, \ldots, q_t^n$, is used for weighting, i.e. considered as constant

- In the previous example the Paasche price index is given by:

$$I_{0,t}^{P}(p) = \frac{342}{270} \cdot 100 = 126{,}7\%$$

in the reporting period t the value of the basket of goods is 342. In the base period 0 the same basket of goods (q_t^1, q_t^2, q_t^3) would have a value of 270.

→ The ratio of the 2 values quantifies the average change of prices with regards to the quantities of the reporting period t

Price index

- **Comparison** of the two price indexes **Laspeyres** and **Paasche**:

Laspeyres: $I_{0,t}^{L}(p) = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} \cdot 100$ Paasche: $I_{0,t}^{P}(p) = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i} \cdot 100$

Difference

- Both indexes compare market values for a defined basket of goods (0 vs. t):

Laspeyres:
Denominator: Actual market value of the basket of goods in the base period
Numerator: Estimated market value of the basket of goods in the base period, evaluated with prices from the reporting period
Paasche:
Denominator: Estimated market value of the basket of goods in the reporting period, evaluated with prices from the base period
Numerator: Actual market value of the basket of goods in the reporting period

Price index

- **Comparison** of **Laspeyres** and **Paasche** using price indexes for each good:

$$I_{0,t}\left(p^i\right) = \frac{p_t^i}{p_0^i} \cdot 100 \quad \text{for each good I = 1,2,…,n}$$

Both aggregate price indexes (Laspeyres and Paasche) are a **weighted** average of the single price indexes for each good:

- **Laspeyres:**

$$I_{0,t}^L(p) = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} \cdot 100 = \sum_i \frac{p_0^i q_0^i}{\sum_i p_0^i q_0^i} \cdot \frac{p_t^i}{p_0^i} \cdot 100 = \sum_i \frac{w_0^i}{W_0} \cdot I_{0,t}\left(p^i\right)$$

with $w_0^i = p_0^i q_0^i$ and $W_0 = \sum_i p_0^i q_0^i$

→ The weights are the **market value ratios from the base period**

Price index

- **Paasche:**

$$I_{0,t}^{P}(p) = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i} \cdot 100 = \sum_i \frac{p_0^i q_t^i}{\sum_i p_0^i q_t^i} \cdot \frac{p_t^i}{p_0^i} \cdot 100 = \sum_i \frac{g_t^i}{G_t} \cdot I_{0,t}(p^i)$$

with $g_t^i = p_0^i q_t^i$ and $G_t = \sum_i p_0^i q_t^i$

→ In these **weights are quantities from the reporting period** but also **price ratios from the base period**

Price index

- Which aggregated price index is better? Index Problem!

  Laspeyres:

  Advantage: Reflects the pure price development, i.e. there are no biased effects due to changes of quantities in the reporting period

  Disadvantage: The basket of goods gets out dated, i.e. the price changes of the goods are not considered (weighted) according to their current (reporting period) economical importance

  Paasche:

  Advantage: More actuality of the basket of goods. But also price ratios of the base period keep an impact

  Disadvantage: Price index does not only reflect price changes but also changes in quantities can affect this index. The Paasche index underestimates price increases (substitution effects)

Your turn (1)

A1) The following table shows the prices and quantities of a 3 persons household in the base and reporting period.

Example of a 3 persons household					
		base period		reporting period	
i	good	qi0	pi0	qi1	pi1
1	bread	32 kilo	1,90 €/kilo	23 kilo	2,00 €/kilo
2	milk	28 liter	0,54 €/liter	30 liter	0,54 €/liter
3	butter	54 pieces	0,13 €/piece	54 pieces	0,15 €/piece
4	potatos	45 pounds	0,18 €/pound	38 pounds	0,20 €/pound
5	meat	24 kilo	6,25 €/kilo	32 kilo	6,37 €/kilo

1.) How did the price level change for each good from the base period to the reporting period?
2.) Compare the results from a simple (equally weighted) aggregated price index with the results from Laspeyres and Paasche Index.

Your turn (2)

A2) The following table shows the ticket prices and passengers of an airline in the two periods (base and reporting period).

Roundtrip ex FRA	Base Period						Reporting Period					
	Ticket Price			Passengers			Ticket Price			Passengers		
	First	Bus	Eco	First	Bus	Eco	First	Bus	Eco	First	Bus	Eco
MUC		650	550		9.600	48.000		600	500		9.900	49.000
BER		700	600		7.200	36.000		750	550		7.000	37.000
LHR		1.050	900		9.800	47.000		1.000	800		10.000	49.000
MAD		1.750	1.500		4.800	24.000		1.850	1.550		4.500	22.000
JFK	7.350	4.100	3.300	2.400	10.400	36.000	7.500	4.200	3.200	2.200	10.100	37.000
LAX	10.050	6.050	4.950	1.200	5.800	21.000	9.900	5.800	5.100	1.300	6.100	19.000

1.) How did the yields change from the base period to the reporting period if you look at each compartment and city pair combination?
2.) How did the yields change from the base period to the reporting period if you look at the aggregated city pairs?
3.) How did the yields change from the base period to the reporting period is you look at it from an aggregated compartment perspective?
4.) How did the yields change from the base period to the reporting period if you want to give one statement for the whole airline?

Your turn (3)

A3) The following table shows the prices and quantities of a basket with 4 good in the base and reporting period.

	1998		2002	
Good	quantity	price	quantity	price
1	10	40	10	60
2	10	30	8	45
3	5	20	25	30
4	25	80	5	120

1.) How did the price level change for each good from the base period to the reporting period?
2.) Compare the results from a simple (equally weighted) aggregated price index with the results from Laspeyres and Paasche Index.
3.) If all are the same, explain why (even though the basket of goods has changed)

Your turn (4)

A4) The following table shows the yearly passengers of an airline differentiated by traffic type

	passengers		
Year	Interkont	Kont	Domestic
1997	14.503	5.209	4.066
1998	14.025	5.240	4.606
1999	13.566	4.980	4.343
2000	14.437	4.587	4.234
2001	14.780	4.851	3.882
2002	16.459	5.044	3.880
2003	15.964	4.754	3.802
2004	16.609	4.600	3.656
2005	19.043	3.856	3.380
2006	20.408	3.498	3.221
2007	23.848	3.476	3.291

1.) Compare the 3 time series regarding the growth rates by constructing for each time series an index number with basis 1997
2.) What is the average yearly growths rate of the 3 time series?

Your turn (5)

A3) The following table shows the prices (in €) and quantities of a basket with 5 goods (the sports equipment you use) in the base and reporting period. Find out if your hobby (sports) has become more or less expensive.

	2000		2008	
Good	quantity	price	quantity	price
Tennis rackets	4	250	2	300
Tennis balls	50	3	30	4
Soccer shoes	2	120	1	150
Soccer ball	3	50	1	60
Golf set	1	1000	2	800

1.) How did the price level change for each sport equipment from the base period to the reporting period?

2.) Compare the results from a simple (equally weighted) aggregated price index with the results from Laspeyres and Paasche Index. What are the advantages and disadvantages of both indexes?

Contents

Analysis of two variables - Correlation and Regression analysis – what will you learn here?

In this chapter you will learn about:

- How to analyze the (causal) relationship between two variables
- Distinguish between independent and dependent variable
- Learn to calculate a correlation coefficient and a (simple) regression analysis

Correlation and regression analysis

- **Goal:** Determine the relationship between two variables, i.e. how do the two variables relate to one another. Some Examples: Relationship between:
  - Profit and revenue of a company
  - Consumption of fuel and speed of a car
  - Volume of sales and level of advertising of a company
  - Consumption and income of a household
  - Size of a TV and selling price of a TV
  - Size of a sports team payroll and number of games won
  - Level of price and amount of products sold
  - …
- There are always two central questions to be answered:
  1) How strong is the relationship between the two variables?
     → **correlation analysis**
  2) Through which mathematical function can we describe the nature of this relationship between the two variables, i.e. the causal relationship
     → (**simple) regression analysis**

Correlation

- **Correlation:** Measures the strength and the direction of the relationship between two variables x and y.

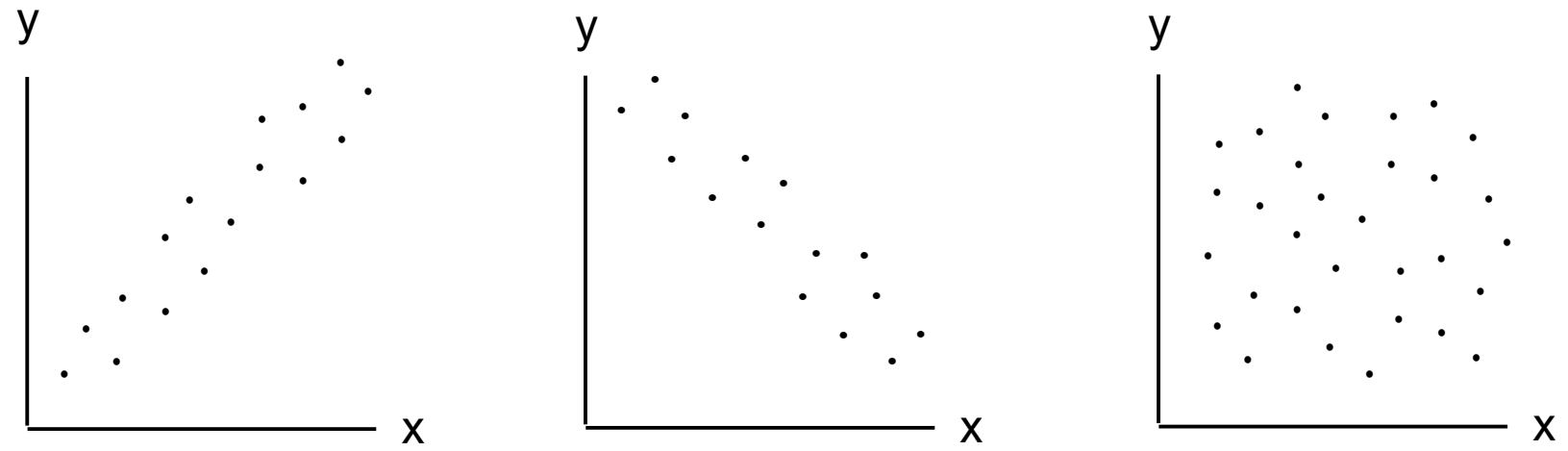
- **Example:** Data from statistics exam. Is there a relation ship between hours studied (x) and exam grade (y)?

Observation (i)	Hours Studied (x)	Exam Grade (y)
1	3	86
2	5	95
3	4	92
4	4	83
5	2	78
6	3	82
7	1	71
8	6	98

The data from the table is considered observation pairs of (x_i, y_i) values. Here we have 8 pairs: $(x_1, y_1) = (3,86), (x_2, y_2) = (5,86), \ldots, (x_8, y_8) = (6,98)$

Correlation

- **Scatter plots** are a nice way to display the observation pairs (x_i, y_i)
- **Different directions of correlation:** The following figures show different directions of correlations in a series of scatter plots, which graphs each observation pairs of (x_i, y_i). The convention is to place the x variable on the horizontal axis and the y variable on the vertical axis



(A) positive correlation (B) negative correlation (C) no correlation

Correlation

- Correlation between two variables x and y can measured by means of following **correlation coefficient r:**

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

where:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ is the **standard deviation of the variable x**

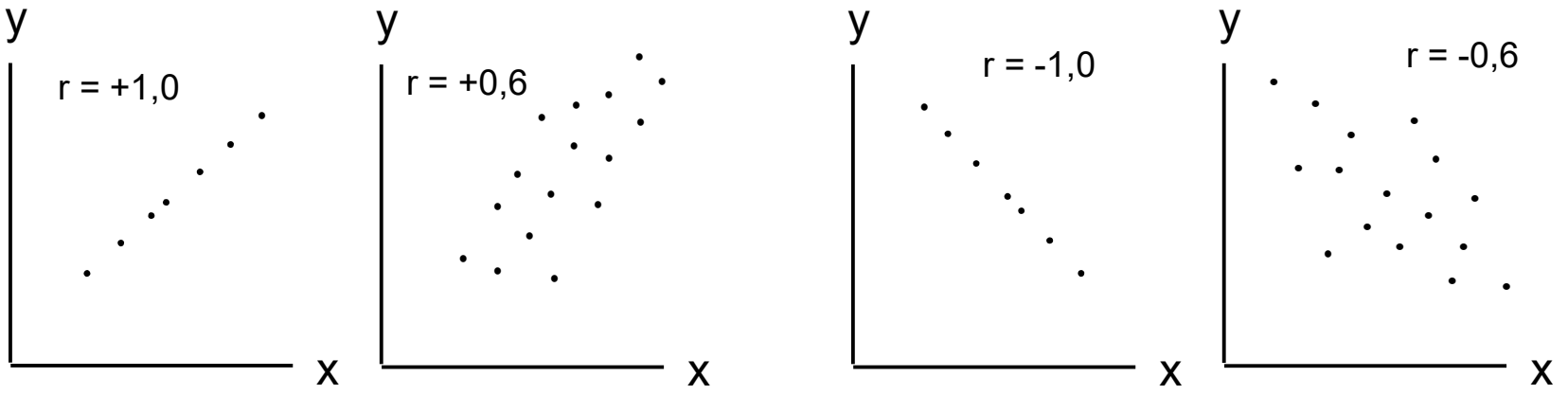
$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$ is the **standard deviation of the variable y**

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$$ is the **covariance of variable x and y**

note that $\bar{x}$ and $\bar{y}$ are the means of the variables x and y

Correlation

- What exactly does the correlation coefficient measure?
  - The correlation coefficient, r, indicates both strength and direction of the relationship between the variable x and y.
  - The values of r range from -1, a strong negative relationship, to +1, a strong positive relationship. When r equals 0 there is no relationship between the variables x and y.
- Strength of the relationship:



Correlation

- **Example:** Calculation of the correlation between hours studied (x) and exam grade (y)

Observation (i)	Hours Studied (x)	Exam Grade (y)	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	3	86	0,25	0,14	-0,19
2	5	95	2,25	87,89	14,06
3	4	92	0,25	40,64	3,19
4	4	83	0,25	6,89	-1,31
5	2	78	2,25	58,14	11,44
6	3	82	0,25	13,14	1,81
7	1	71	6,25	213,89	36,56
8	6	98	6,25	153,14	30,94
	$\bar{x}$	$\bar{y}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	3,50	85,63	1,50	8,47	12,06

Correlation

- **Example:** Calculation of the correlation between hours studied (x) and exam grade (y)

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{12{,}06}{1{,}5 \cdot 8{,}47} = 0{,}949$$

→ There is a strong positive correlation between hours studied and exam grade, i.e. the higher the hours studied are the higher the exam grades are and the lower the exam grades are the lower the hours studied are

- **The correlation coefficient treats both variables equally and thus does not say anything about causality between the two variables:**

→ Note that the correlation coefficient does not make any statement about the causality of the relationship between the two variables x and y, i.e. the correlation coefficient does not say, whether the hours studied affects the exam grades or whether the exam grades affects the hours studied

Correlation

- Instead of displaying the observation pairs (x_i, y_i) via scatter plots the two dimensional frequency distribution can also be displayed in so call **contingency table.**

- $n(x_i, y_i) = n_{ij}$ corresponds to the **absolute frequency** of the observation pair

- The **relative frequency** of the observation pair is given by

$$h(x_i, y_i) = \frac{n_{ij}}{n}$$

	$y_1\ y_2 \dots y_j \dots y_m$	Sum of rows
x_1	$n_{11}\ n_{12} \dots n_{1j} \dots n_{1m}$	$n_{1\bullet}$
x_2	$n_{21}\ n_{22} \dots n_{2j} \dots n_{2m}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i1}\ n_{i2} \dots n_{ij} \dots n_{im}$	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
x_k	$n_{k1}\ n_{k2} \dots n_{kj} \dots n_{km}$	$n_{k\bullet}$
Sum of columns	$n_{\bullet 1}\ n_{\bullet 2} \dots n_{\bullet j} \dots n_{\bullet m}$	n

Correlation

- The marginal distribution correspond to the one dimensional distribution of each of the two variables x and y

$n_{i\bullet}$ absolute frequency of variable x

$h(x_i) = h_{i\bullet} = \frac{n_{i\bullet}}{n}$ relative frequency of variable x

$n_{\bullet j}$ absolute frequency of variable y

$h(y_j) = h_{\bullet j} = \frac{n_{\bullet j}}{n}$ relative frequency of variable y

Correlation

- In case of grouped data of two variable we have to work with contingency tables to calculate the correlation
- Using the class midpoints the standard deviation of x and y as well as the covariance can be calculated as follows:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 \cdot n_{i\bullet}}$$ is the **standard deviation of the variable x**

$$s_y = \sqrt{\frac{1}{n}\sum_{j=1}^{m}\left(y_j - \bar{y}\right)^2 \cdot n_{\bullet j}}$$ is the **standard deviation of the variable y**

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)\cdot n_{ij}$$ is the **covariance of variable x and y**

Correlation

- The means of x and y are calculated as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_{i\bullet}$$

is the **mean of the variable x**

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{m} y_j \cdot n_{\bullet j}$$

is the **standard deviation of the variable y**

- With the above statistics at had the **correlation coefficient r** can then be calculated in the usual manner:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Correlation

- **Example:** Contingency Table of revenue (x) and expenditure for research and development (y) for 50 companies. Calculate whether there is a relationship between x and y

	Expenditure for research and development (y)					
Revenue (x)	0-10	10-20	20-30	30-40	40-50	sum
0-200	2	3	1	0	0	6
200-400	2	6	3	1	0	12
400-600	1	4	5	4	0	14
600-800	0	2	4	3	2	11
800-1000	0	0	1	2	4	7
sum	5	15	15	10	6	50

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_{i\bullet} = \frac{1}{50}\cdot(100\cdot 6+\ldots) = 504$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{m} y_j \cdot n_{\bullet j} = \frac{1}{50}\cdot(5\cdot 5+\ldots) = 24{,}9$$

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot n_{i\bullet}} = \sqrt{\frac{1}{50}\cdot\left[(100-504)^2\cdot 6+\ldots\right]} = 244{,}9$$

$$s_y = \sqrt{\frac{1}{n}\sum_{j=1}^{m}(y_j-\bar{y})^2\cdot n_{\bullet j}} = \sqrt{\frac{1}{50}\cdot\left[(5-24{,}9)^2\cdot 5+\ldots\right]} = 11{,}7$$

$$r = \frac{s_{xy}}{s_x\cdot s_y} = \frac{1922{,}4}{244{,}9\cdot 11{,}7} = 0{,}67$$

→ **Quit strong positive correlation**

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})\cdot(y_i-\bar{y})\cdot n_{ij} = \frac{1}{50}\left[(100-504)\cdot(5-24{,}9)\cdot 2+\ldots\right] = 1922{,}4$$

Simple Regression analysis

- **Simple regression analysis:** Through which mathematical function can we describe the nature of the causal relationship between the two variables x and y

- For this causal relationship we need to define which of the two variables is the **independent variable** and which is the **dependent variable**. It is usual convention that x is the label for the independent variable while y is the label for the dependent variable, i.e. the simple regression analysis evaluates the following causal relationship:

$$x \rightarrow y$$

- With the simple regression analysis we describe the causal relationship between x and y with a **straight line** that fits best the series of observation pairs (x_i, y_i)

Simple Regression analysis

- The equation for a straight line, known as linear equation, takes the following form:

$$\hat{y} = a + b \cdot x$$

where:

$\hat{y} =$ the predicted value of the dependent variable y, given the value of x

$x =$ the independent variable

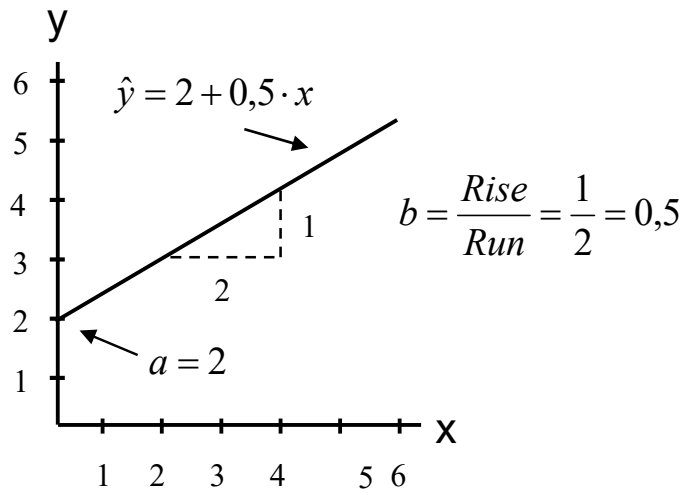
$a =$ the y-intercept for the straight line

$b =$ the slope of the straight line

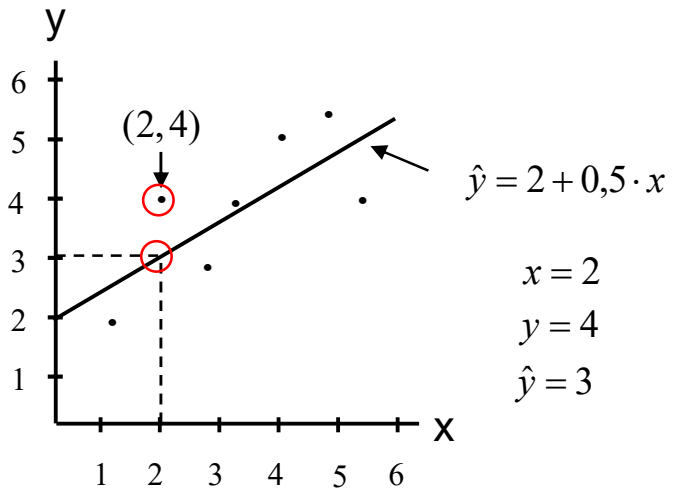
Simple Regression analysis

- **Example** of a line described by the linear equation $\hat{y} = 2 + 0{,}5 \cdot x$

Equation for a straight line



Difference between y and $\hat{y}$



- The y-intercept is the point where the line crosses the y-axis: here $a = 2$
- The slope is shown as the ratio of rise over run of the line: here $b = 0{,}5$
- While $\hat{y}$ represents the predicted value according to the line, y represents an actual data point. E.g. here $(x, y) = (2, 4)$ vs $\hat{y} = 2 + 0{,}5 \cdot x = 2 + 0{,}5 \cdot 2 = 3$

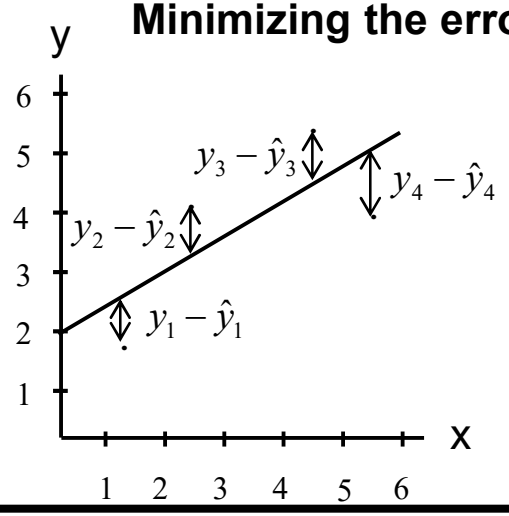
Simple Regression analysis

- If we have a data set of observation pairs, e.g. hours studied (x) and exam grade (y), **how do we find the linear equation (straight line) that best fits the observed data?**

- The **least square method** is a mathematical procedure to identify the linear equation that fits best the data set of observation pairs. The goal is to find the a and b of the linear equation that minimizes the total squared error between the values of y and $\hat{y}$. If we define the error as $(y - \hat{y})$ for each data point, the least square method will minimize the following equation:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - b \cdot x)^2 = Min!$$

Minimizing the error



Simple Regression analysis

- The **least square method**

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-a-b\cdot x\right)^2=Min!$$

leads to following **equations for the calculation of two parameters** a and b:

$$b=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{s_{xy}}{s_x^2}$$

$$a=\bar{y}-b\cdot\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the means of the variables x and y. s_{xy} is the covariance of x and y and s_x^2 is the variance of x

Simple Regression analysis

- **Example:** Calculation of the regression line between hours studied (x) and exam grade (y)

$$b = \frac{s_{xy}}{s_x^2}$$

$$b = \frac{12{,}06}{2{,}25} = 5{,}36$$

$$a = \bar{y} - b \cdot \bar{x}$$

$$a = 85{,}63 - 5{,}36 \cdot 3{,}50$$

$$a = 66{,}86$$

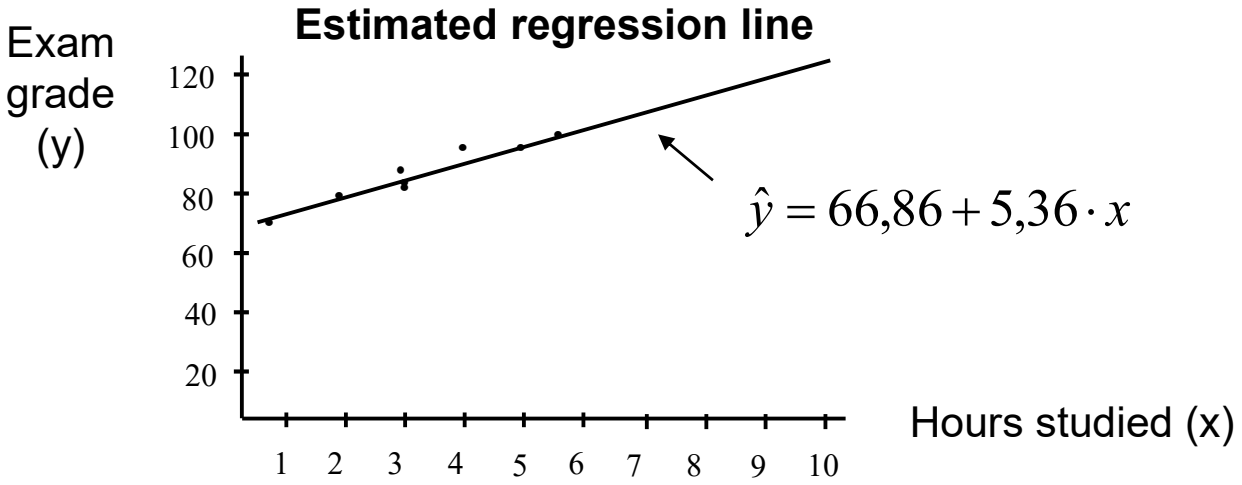
$$\hat{y} = a + b \cdot x$$

$$\hat{y} = 66{,}86 + 5{,}36 \cdot x$$

Observation (i)	Hours Studied (x)	Exam Grade (y)	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	3	86	0,25	-0,19
2	5	95	2,25	14,06
3	4	92	0,25	3,19
4	4	83	0,25	-1,31
5	2	78	2,25	11,44
6	3	82	0,25	1,81
7	1	71	6,25	36,56
8	6	98	6,25	30,94
	$\bar{x}$	$\bar{y}$	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	3,50	85,63	2,25	12,06

Simple Regression analysis

- **Example:** Calculation of the regression line between hours studied (x) and exam grade (y)



- **Interpretation:** If the hours studied increase by one unit, the exam grade increases by 5,36 unites

- The regression line can be used for **prediction**. E.g. if a student invests 5,5 hours time to study for the exam he/she would achieve the following grade:

$$\hat{y} = 66{,}86 + 5{,}36 \cdot x = 66{,}86 + 5{,}36 \cdot 5{,}5 = 96{,}36 \approx 96$$

Simple Regression analysis

- A way of measuring the quality of the regression line is the **coefficient of determination** D :

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

- It represents the percentage of the variation of y which is explained by the regression line $\hat{y}$. For the simple regression the coefficient of determination equals the squared correlation coefficient, i.e. the strength (squared) of the relationship between x and y:

$$D = \left(\frac{s_{xy}}{s_x \cdot s_y}\right)^2 = r^2$$

- D can have values between 0 and 1. If $D = 1$, all of the variation in y is explained by the variable x. If $D = 0$, none of the variation in y is explained by the variable x.

Simple Regression analysis

- **Example:** Calculation of the coefficient of determination for the regression of hours studied (x) and exam grade (y)

$$r^2 = \left(\frac{s_{xy}}{s_x \cdot s_y} \right)^2 = \left(\frac{12{,}06}{1{,}5 \cdot 8{,}47} \right)^2 = 0{,}949^2 = 0{,}901$$

→ In other words, 90,1 percent of the variation of the exam grades can be explained by the variable hours studied

Spurious correlation

- Be aware of **spurious correlation!**

  → i.e. high correlation gives an impression of a worthy link between the two variables, which however is invalid when objectively examined

- **Example:** Relationship between "active vocabulary" (measured by the amount of different word in a written essay about the experiences in their last vacation), x, and the size (measured in centimeters), y, of ten children:

I	1	2	3	4	5	6	7	8	9	10
active vocabulary (xi)	37	30	20	28	35	40	18	22	42	22
size (yi)	130	112	108	114	136	141	105	110	143	109

the correlation coefficient between x and y is: $r = \frac{s_{xy}}{s_x \cdot s_y} = 0{,}951$

→ strong positive correlation between active vocabulary and size!
Does this make sense?!?

Spurious correlation

- No it does not make sense! there is a third important other variable which leads to this “spurious” correlation:

I	1	2	3	4	5	6	7	8	9	10
active vocabulary (xi)	37	30	20	28	35	40	18	22	42	22
size (yi)	130	112	108	114	136	141	105	110	143	109
age (zi)	12	7	6	7	13	14	5	6	14	4

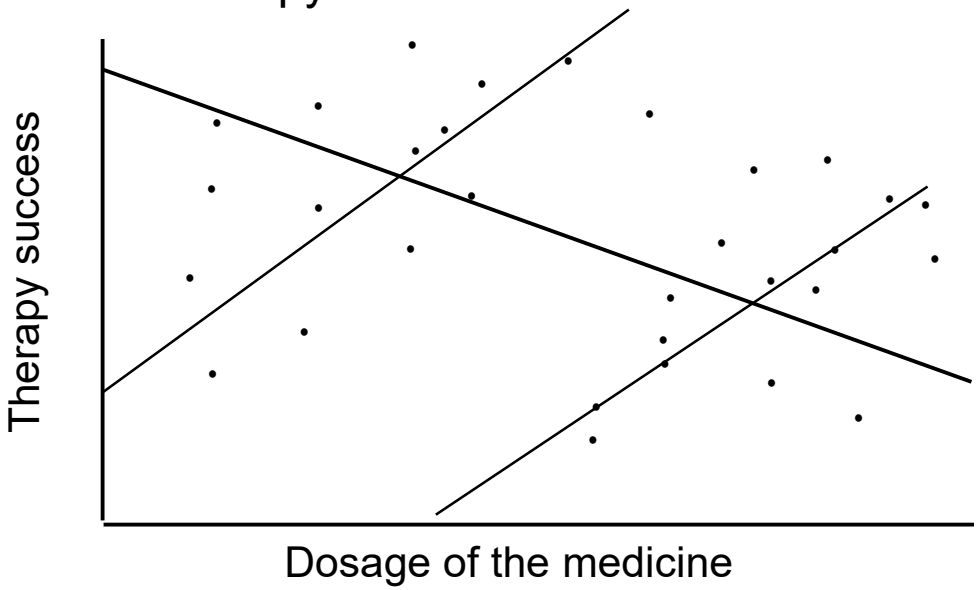
Correlation coefficient between x and z and y and z:

$$r_{xz} = 0{,}946 \qquad r_{yz} = 0{,}983$$

→ A third variable z which is highly correlated with the two variables x and y is leading to the spurious correlation between the two variables

Spurious correlation

- **Another example of spurious correlation:** Relationship between dosage of a medicine and the therapy success



If in this case we would not consider that there are two populations (slightly vs. heavy ill) we would estimate a negative relation between dosage and success

→ Do never only look at the correlation but always also at the scatter plot!

Multiple Regression analysis

- In some cases the variation of the dependent variable y can be explained by more than one independent variable → **multiple regression**
- A multiple (instead of simple) regression looks a follows:

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots + b_n \cdot x_n$$

where:

$\hat{y} =$	predicted value of y, given the values of the x variables
$x_1, x_2, \ldots, x_n =$	independent variables
$a =$	intercept
$b_1, b_2, \ldots, b_n =$	parameters of the different independent variables

Multiple Regression analysis

- Using the **least square method**

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-a-b_1\cdot x_1-b_2\cdot x_2,\ldots,-b_n\cdot x_n\right)^2=Min!$$

  The parameters $a,b_1,b_2,\ldots,b_n$ can be estimated

- Note, when running a multiple regression, it is important the independent variables are very much independent of each other, i.e. do not correlate with each other → low multicollinearity. Otherwise results may be misleading

- In order to run a multiple regression analysis in Excel you need to use the Analysis Tool “Regression” within the Data Analysis Add-In

Multiple Regression analysis

- **Exemplary results of a multiple regression**:
  Analysis of sulfur dioxide concentration in the air (i.e. the y variable) depending on climatic and geographic variables for 41 cities in the United States.

$$\hat{y} = 7{,}4 - 0{,}0557x_1 + 0{,}0006x_2 - 0{,}158x_3$$

Variables:

y = sulfur dioxide concentration in the air
x_1 = average temperature in Fahrenheit during the year
x_2 = number of manufacturing companies with more than 20 employees
x_3 = average wind speed in miles per hour during the year

Give an interpretation of the results? How do the independent variables effect the sulfur dioxide concentration?

Multiple Regression analysis

- **Exemplary multiple regression with Excel**:
  A CEO of a company that produces butter wants to analyze of the impact of price, advertising expenditures and sales visits on the sales quantity of butter the company sells. The CEO uses data on 20 different markets

- Following equation should be estimated:

$$\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$$

where:

y = quantity

x_1 = price

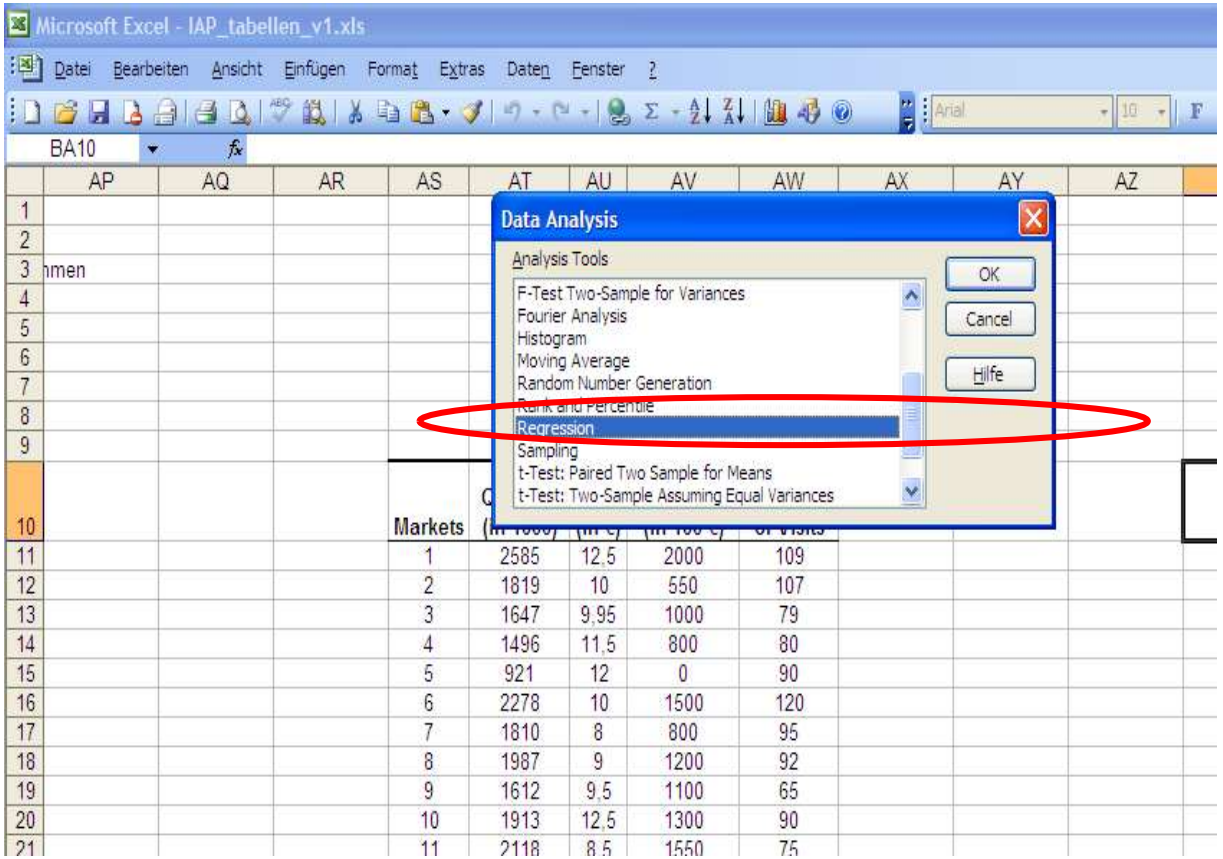
x_2 = advertising

x_3 = visits

Markets	Quantity (in 1000)	Price (in €)	Advertising (in 100 €)	Number of Visits
1	2585	12,5	2000	109
2	1819	10	550	107
3	1647	9,95	1000	79
4	1496	11,5	800	80
5	921	12	0	90
6	2278	10	1500	120
7	1810	8	800	95
8	1987	9	1200	92
9	1612	9,5	1100	65
10	1913	12,5	1300	90
11	2118	8,5	1550	75
12	1438	12	550	106
13	1834	9,5	1980	66
14	1869	9	1600	80
15	1574	7	500	90
16	2597	11	2000	120
17	2026	10	1680	95
18	2016	9,5	1700	92
19	1566	10	1400	65
20	2168	13	1800	90

Multiple Regression analysis

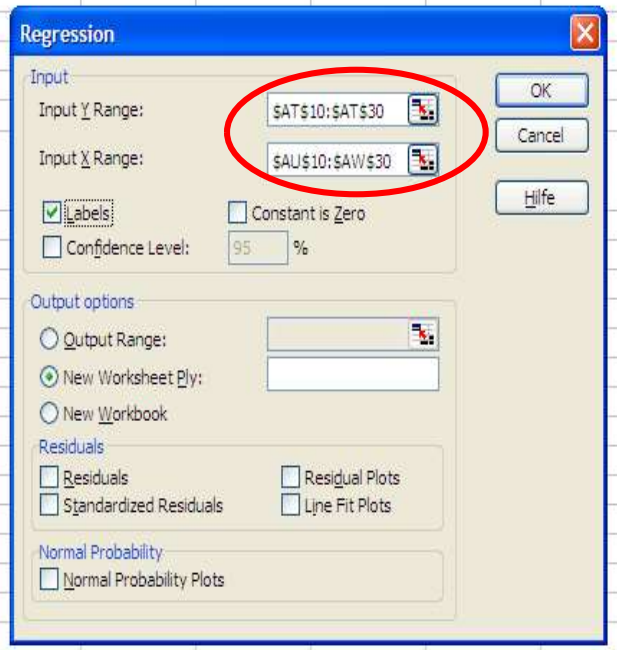
- Excel's Data Analysis dialog box → Regression



Multiple Regression analysis

- Define dependent and independent variables

Markets	Quantity (in 1000)	Price (in €)	Advertising (in 100 €)	Number of Visits
1	2585	12,5	2000	109
2	1819	10	550	107
3	1647	9,95	1000	79
4	1496	11,5	800	80
5	921	12	0	90
6	2278	10	1500	120
7	1810	8	800	95
8	1987	9	1200	92
9	1612	9,5	1100	65
10	1913	12,5	1300	90
11	2118	8,5	1550	75
12	1438	12	550	106
13	1834	9,5	1980	66
14	1869	9	1600	80
15	1574	7	500	90
16	2597	11	2000	120
17	2026	10	1680	95
18	2016	9,5	1700	92
19	1566	10	1400	65
20	2168	13	1800	90



Multiple Regression analysis

- The results of the multiple regression:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,94892264							
R Square	0,90045418							
Adjusted R Square	0,88178933							
Standard Error	135,426589							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	2654400,42	884800,141	48,2433326	3,0733E-08			
Residual	16	293445,778	18340,3611					
Total	19	2947846,2						
	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	446,870061	238,196547	1,876056	0,07901231	-58,084057	951,824179	-58,084057	951,824179
Price (in €)	-35,9611048	20,0448836	-1,79402912	0,09172619	-78,4543594	6,53214976	-78,4543594	6,53214976
Advertising (in 100 €)	0,56586926	0,0544522	10,3920359	1,6045E-08	0,45043574	0,68130277	0,45043574	0,68130277
Number of Visits	11,9249482	1,96438548	6,07057442	1,6215E-05	7,76063707	16,0892594	7,76063707	16,0892594

- Excursus to inferential statistics: Gives information about significance of the parameters
- Simple rule so far: If Abs(t Stat) > 1,7 than significant otherwise no significant impact

$$\hat{y} = 446{,}87 - 35{,}96 \cdot x_1 + 0{,}56 \cdot x_2 + 11{,}92 \cdot x_3$$

Multiple Regression analysis

- **Be careful with multicollinearity** – Example: How does age (x) predict "active vocabulary" (measured by the amount of different word in a written essay about the experiences in their last vacation), y:

- Simple Linear Regression leads to meaningful results :

	Coefficients	*t Stat*	*P-value*
Intercept	11,01695	4,55077	0,00187
age (zi)	2,08898	8,25859	0,00003
R Square	0,8950191		

- Including size, which is highly correlated to age, i.e. we have multicollinearity, leads to misleading insignificant results:

	Coefficients	*t Stat*	*P-value*
Intercept	-22,25308	-0,68128	0,51759
size (yi)	0,37722	1,02135	0,34108
age (zi)	0,69145	0,49696	0,63444
R Square	0,90863462		

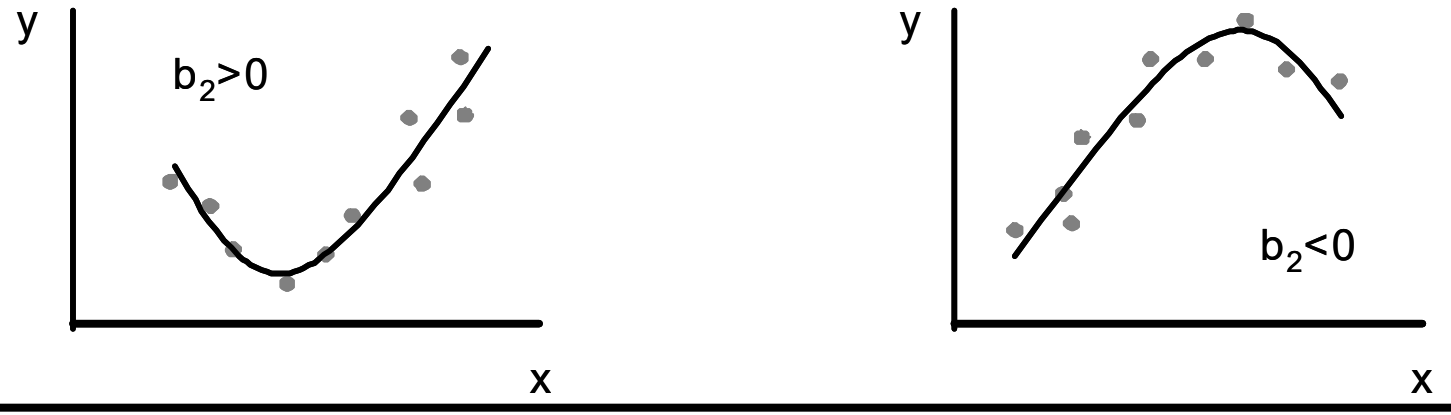
Nonlinear Regression analysis

- In some cases the variation of the dependent variable y can be explained by the independent variable not in a linear but rather in a nonlinear way → **non linear regression**

- **One exemplary non-linear relationship** between y and x could be a quadratic function (polynomial 2. order):

$$\hat{y} = a + b_1 \cdot x + b_2 \cdot x^2$$

- Different shapes of the regression depending on sign of b_2



Nonlinear Regression analysis

- Quadratic functions allow that the impact direction of x on y changes (e.g. from positive to negative)

- **Example**: A Farmer wants to know, how much fertilizer (x) is optimal in order to achieve the maximum output of corn (y). Therefore he does a test on 14 different fields and based on this data he estimates a regression assuming a quadratic relationship between x and y

Field	*Fertilizer (kg/hectare)*	*Corn (kg/hectare)*
1	15	1800
2	30	3600
3	45	6840
4	60	7200
5	75	8100
6	90	8460
7	105	8640
8	120	9000
9	135	9180
10	150	9000
11	165	8640
12	180	8460
13	195	8100
14	210	7740

Nonlinear Regression analysis

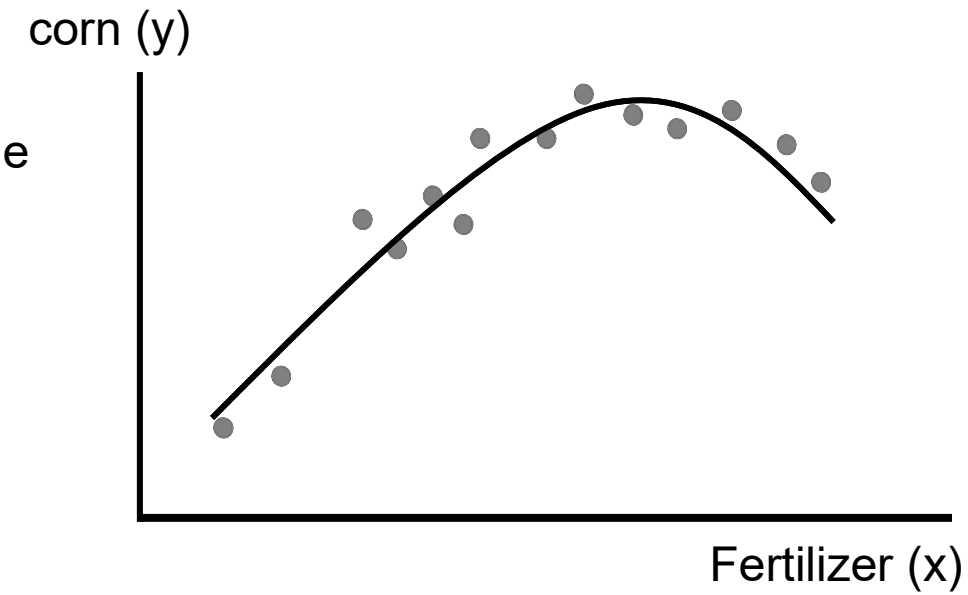
- As a result he gets following equation:

$$\hat{y} = 881 + 123 \cdot x - 0{,}44 \cdot x^2$$

- Using this result, he can determine the optimal amount of fertilizer

$$\frac{d\hat{y}}{dx} = 123 - 2 \cdot 0{,}44 \cdot x = 0$$

$$\Rightarrow x^{opt} = 139{,}8 \; kg / hectare$$



Nonlinear Regression analysis

- **Exemplary non linear Regression with Excel**: The market research of a company has a data about of 31 customer on age (x) and the revenue (y) they generate.

  Goal is to find out whether there is a relationship, and if so, what age group is the most attractive in terms of revenue.

- First a linear regression is considered:

$$\hat{y} = a + b \cdot x$$

- In a second step a non linear regression is estimated

$$\hat{y} = a + b_1 \cdot x + b_2 \cdot x^2$$

KN_R	AGE	REVENUE
1	34	23
2	60	723
3	55	863
4	71	55
5	50	907
6	37	239
7	40	739
8	44	893
9	30	95
10	37	158
11	64	730
12	62	478
13	46	1061
14	53	659
15	40	138
16	40	148
17	45	322
18	57	1123
19	26	145
20	70	333
21	58	823
22	34	205
23	49	1201
24	42	160
25	65	429
26	33	350
27	37	503
28	32	441
29	37	445
30	70	346
31	45	745

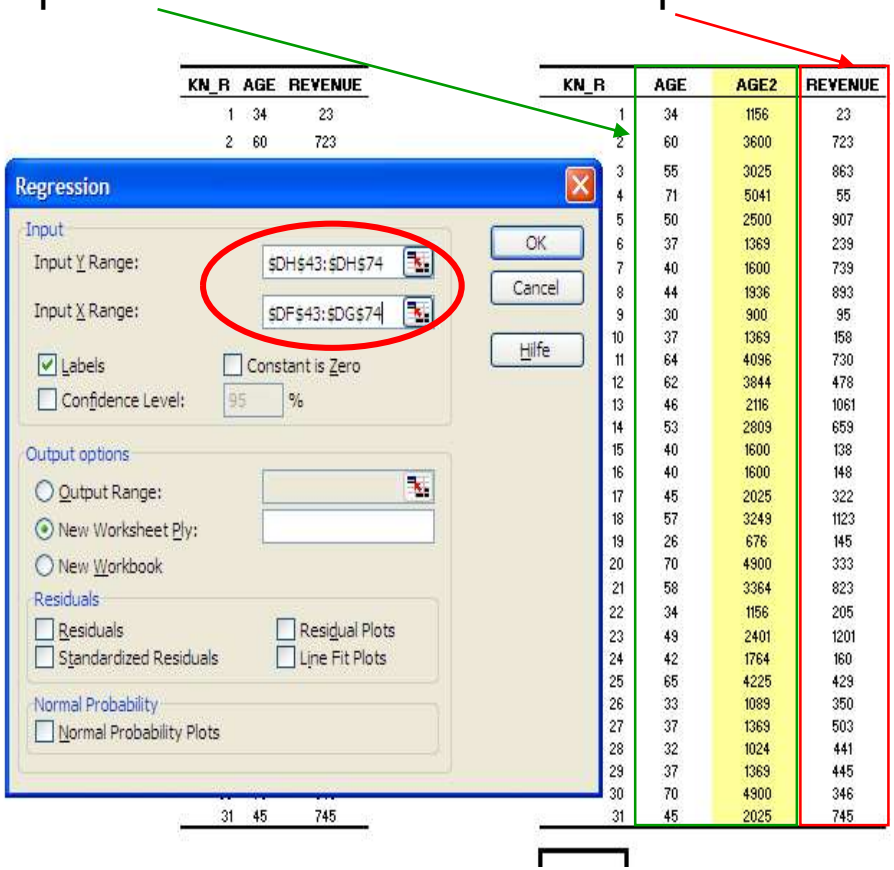
Nonlinear Regression analysis

- **Data preparation:** In order to prepare the data for the non linear (quadratic) regression a further column, age^2 has to be constructed in the data set

KN_R	AGE	AGE2	REVENUE
1	34	1156	23
2	60	3600	723
3	55	3025	863
4	71	5041	55
5	50	2500	907
6	37	1369	239
7	40	1600	739
8	44	1936	893
9	30	900	95
10	37	1369	158
11	64	4096	730
12	62	3844	478
13	46	2116	1061
14	53	2809	659
15	40	1600	138
16	40	1600	148
17	45	2025	322
18	57	3249	1123
19	26	676	145
20	70	4900	333
21	58	3364	823
22	34	1156	205
23	49	2401	1201
24	42	1764	160
25	65	4225	429
26	33	1089	350
27	37	1369	503
28	32	1024	441
29	37	1369	445
30	70	4900	346
31	45	2025	745

Nonlinear Regression analysis

- Define the independent variables and the dependent variable



Nonlinear Regression analysis

- **Comparison** of both model estimates shows that the nonlinear approach fits a lot better than the linear approach:

- Simple Linear Regression results :

	Coefficients	*t Stat*	*P-value*
Intercept	121,1758	0,5347	0,5969
AGE	8,0116	1,7284	0,0946
R Square	0,0934		

- Non Linear Regression results:

	Coefficients	*t Stat*	*P-value*
Intercept	-3054,5812	-4,2303	0,0002
AGE	145,3759	4,7647	0,0001
AGE2	-1,3835	-4,5335	0,0001
R Square	0,4772		

Nonlinear Regression analysis

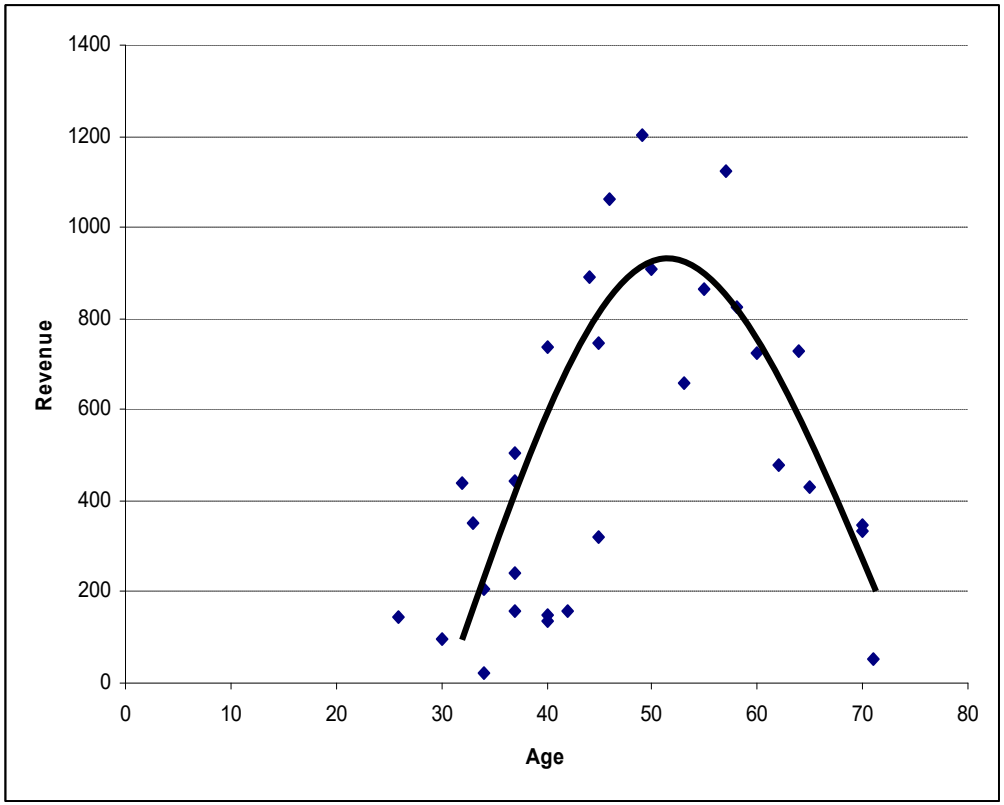
- Looking at the **scatter plot** of age and revenue confirms the nonlinear quadratic relation ship

- The result of the non linear regression indicate a maximum at:

$$\hat{y} = -3054 + 145{,}4 \cdot x - 1{,}38 \cdot x^2$$

$$\frac{d\hat{y}}{dx} = 145{,}4 - 2 \cdot 1{,}38 \cdot x = 0$$

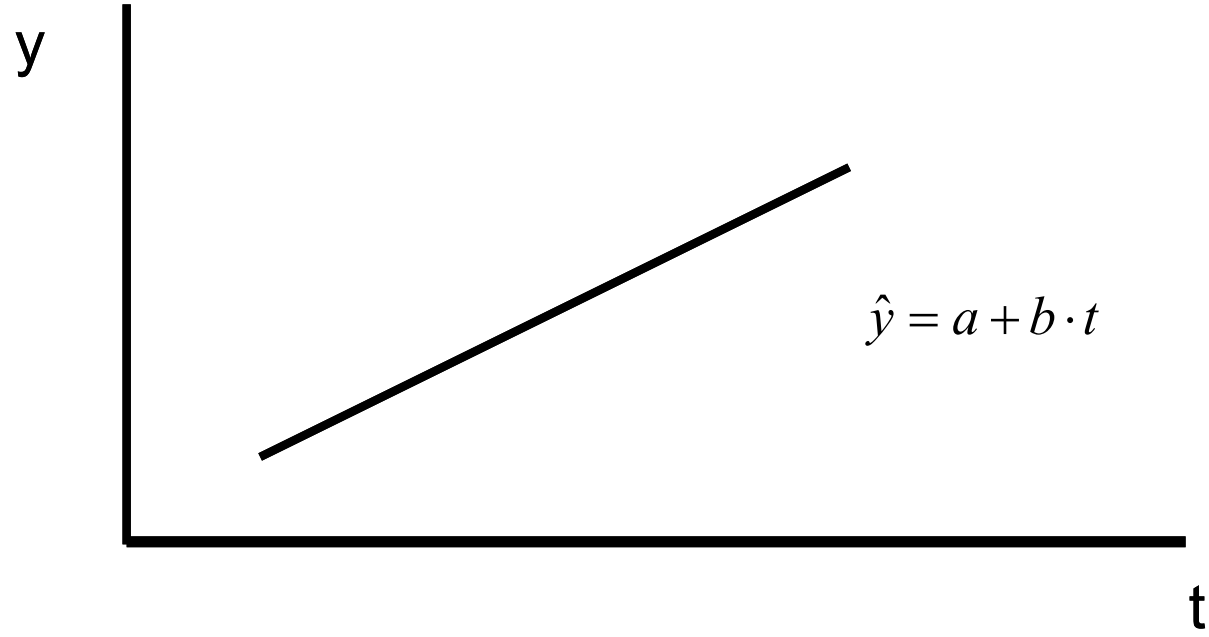
$$\Rightarrow x^{opt} = 52{,}7 \; years$$



Forecasting the trend of a time series by using the Regression analysis

- **Idea of a Trend model:** The development of dependent variable y can explained by the independent variable t, a time component. This model can then be **used to forecast the future development of y**

- Example: Linear trend model:



Forecasting the trend of a time series by using the Regression analysis

- The **linear Trend model** $\hat{y} = a + b \cdot t$ can be estimated analog to the linear regression:

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \qquad a = \bar{y} - b \cdot \bar{x}$$

If we substitute x with t the time component we obtain

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t}) \cdot (y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2} = \frac{s_{ty}}{s_t^2} \qquad a = \bar{y} - b \cdot \bar{t}$$

where $\bar{t}$ and $\bar{y}$ are the means of the variables t and y. s_{ty} is the covariance of t and y and s_t^2 is the variance of t

Forecasting the trend of a time series by using the Regression analysis

- t_i the time component can e.g. be 1988, 1989, ..., 2006

  Since they are numerical values they can be used

  But: for graphical presentation (scale) not appropriate

  Therefore typically transformation of 1988, 1989, ..., 2006 in 1, 2, …, 19 and usage of transformed time component

- **Example:** Linear trend model for GDP development

year	transformed year (t)	GDP
1988	1	878,2
1989	2	933,7
1990	3	1007,2
1991	4	1084
1992	5	1139,6
1993	6	1179,8
1994	7	1214,2
1995	8	1278,1
1996	9	1347,1
1997	10	1406,8
1998	11	1497,6
1999	12	1550
2000	13	1635,5
2001	14	1738,1
2002	15	1892,2
2003	16	2043,5
2004	17	2140,7
2005	18	2129,2
2006	19	2197,1

Forecasting the trend of a time series by using the Regression analysis

- **Example:** Calculation of the linear trend

$$b = \frac{s_{ty}}{s_t^2}$$

$$b = \frac{2267,4}{30} = 75,58$$

$$a = \bar{y} - b \cdot \bar{t}$$

$$a = 1489,1 - 75,58 \cdot 10$$

$$a = 733,29$$

$$\hat{y} = a + b \cdot t$$

$$\hat{y} = 733,29 + 75,58 \cdot t$$

year	transf. year (t)	GDP	$(t_i - \bar{t})^2$	$(t_i - \bar{t}) \cdot (y_i - \bar{y})$
1988	1	878,2	81	5498,0
1989	2	933,7	64	4443,1
1990	3	1007,2	49	3373,2
1991	4	1084	36	2430,5
1992	5	1139,6	25	1747,4
1993	6	1179,8	16	1237,1
1994	7	1214,2	9	824,7
1995	8	1278,1	4	422,0
1996	9	1347,1	1	142,0
1997	10	1406,8	0	0,0
1998	11	1497,6	1	8,5
1999	12	1550	4	121,8
2000	13	1635,5	9	439,2
2001	14	1738,1	16	996,1
2002	15	1892,2	25	2015,6
2003	16	2043,5	36	3326,5
2004	17	2140,7	49	4561,3
2005	18	2129,2	64	5120,9
2006	19	2197,1	81	6372,1
	$\bar{t}$	$\bar{y}$	$\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2$	$\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t}) \cdot (y_i - \bar{y})$
	10	1489,1	30	2267,4

Forecasting the trend of a time series by using the Regression analysis

- The coefficient of determination is:

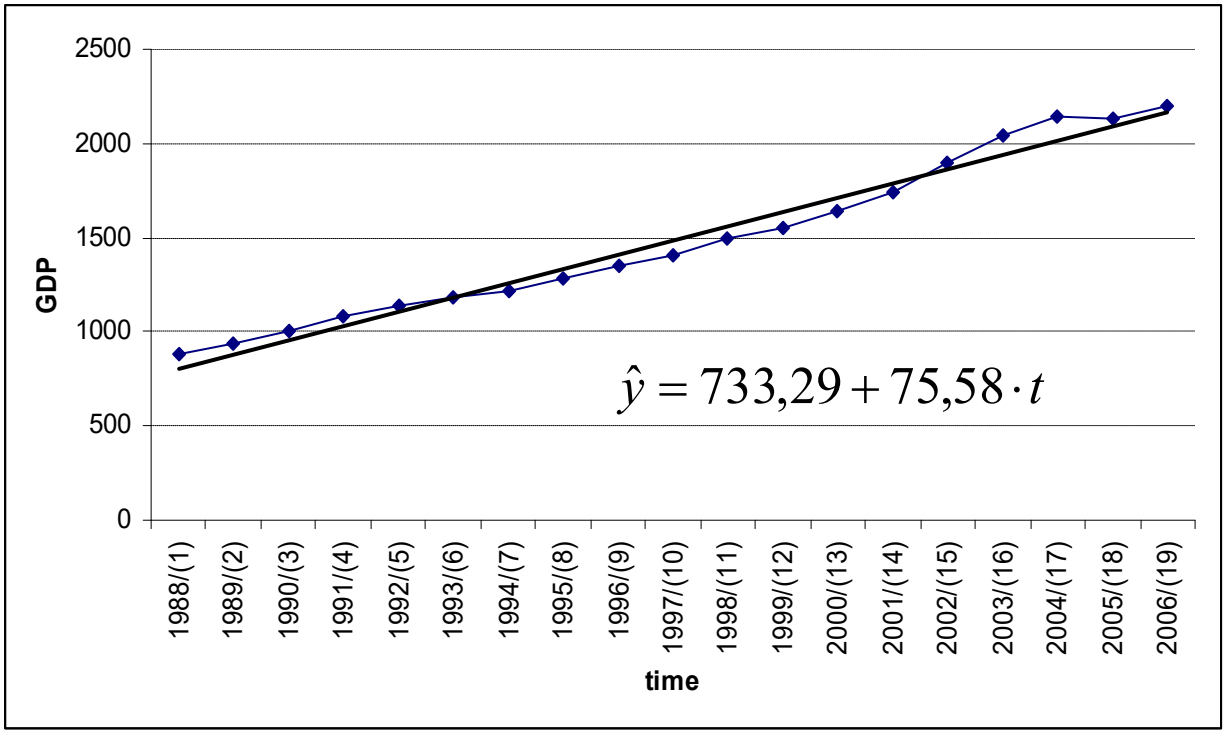
$$D = r^2 = \left(\frac{s_{yt}}{s_y \cdot s_t} \right)^2 = \left(\frac{2267,4}{418,99 \cdot 5,48} \right)^2 = (0,99)^2 = 0,98$$

- The estimated trend for the GDP looks as follows

- For 2009 we would expect a GDP of:

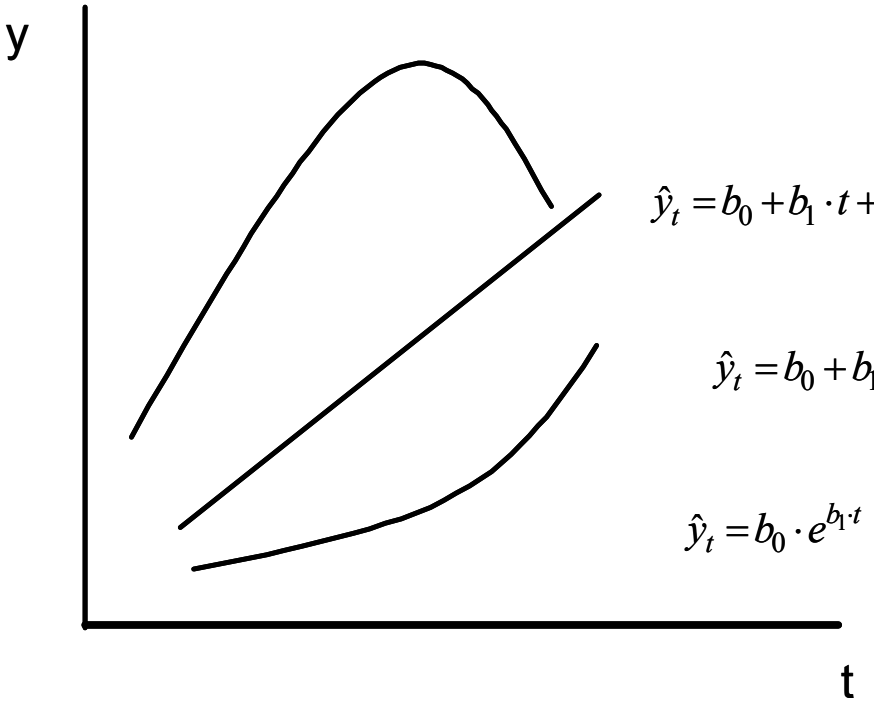
$$\hat{y} = 733,29 + 75,58 \cdot 22$$

$$\hat{y} = 2396,03$$



Forecasting the trend of a time series by using the Regression analysis

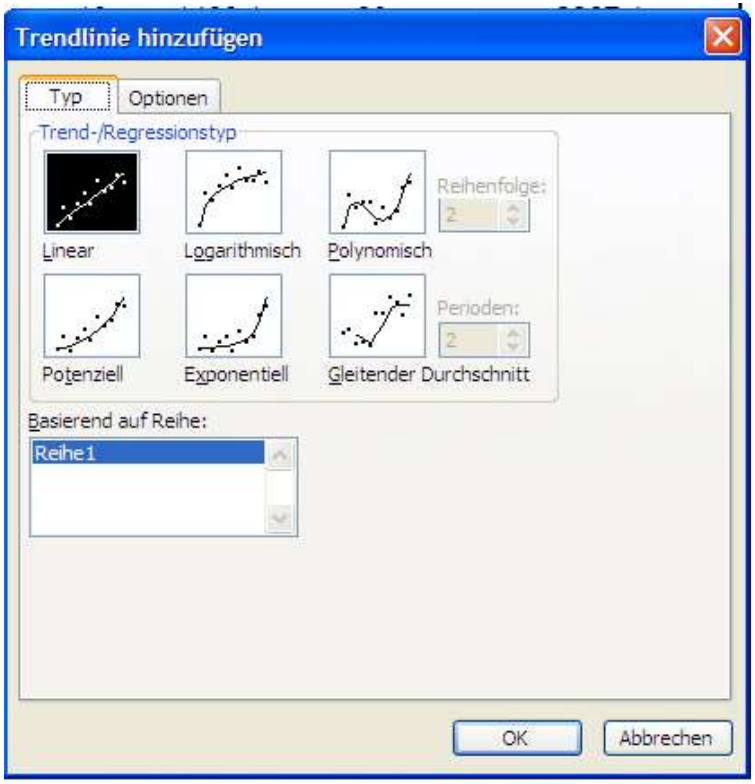
- Other nonlinear trend models can be estimated in Excel



$$\hat{y}_t = b_0 + b_1 \cdot t +$$

$$\hat{y}_t = b_0 + b_1$$

$$\hat{y}_t = b_0 \cdot e^{b_1 \cdot t}$$



Your turn (1)

A1) The following table shows the marketing expenditures and revenues of a company (in 1000 €) over ten month:

Month i	Marketing x_i	Revenue y_i
1	24	201
2	16	184
3	20	220
4	26	240
5	14	180
6	16	164
7	20	186
8	12	150
9	18	182
10	22	210

1.) Analyze the correlation between the two variables?
2.) Specify a meaningful causal relationship between the two variable, estimate the corresponding regression line and interpret the results
3.) What revenue would you expect, if marketing expenditures increase to 30.000€?
4.) Calculate the coefficient of determination and interpret it.

Your turn (2)

A2) The following table shows the average prices (in €) of an economy tickets on the route FRA-BCN and the corresponding amount passenger (only during the 20 working days) over 12 consecutive months:

1.) Analyze the correlation between the two variables?
2.) Specify a meaningful causal relationship between the two variable, estimate the corresponding regression line and interpret the results
3.) Calculate the coefficient of determination and interpret it.
4.) How many passengers would you expect if the average price decreases to 100€?

month	average price	pax per month
1	150	3.447
2	240	3.064
3	188	3.360
4	130	3.967
5	142	3.648
6	260	3.165
7	312	2.864
8	438	2.367
9	209	3.289
10	285	2.987
11	363	2.446
12	342	2.405

Your turn (3)

A3) The following table shows the prices (in $) and the milages of 12 used cars of the same type (VW Jetta):

1.) Analyze the correlation between the two variables?
2.) Specify a meaningful causal relationship between the two variable, estimate the corresponding regression line and interpret the results
3.) Calculate the coefficient of determination and interpret it.
4.) What price would you expect if the used car had 120.000 miles?

milage	price (in $)
21.800	16.000
34.000	11.500
41.700	13.400
53.500	14.800
65.800	10.500
72.100	12.300
76.500	8.200
84.700	9.500
91.000	8.100
101.500	7.900
105.200	8.300
109.100	6.900

Your turn (4)

A4) The following table shows the budget (in mil €) and the number of wins of 12 soccer clubs:

1.) Analyze the correlation between the two variables?
2.) Specify a meaningful causal relationship between the two variable, estimate the corresponding regression line and interpret the results
3.) Calculate the coefficient of determination and interpret it.
4.) How many wins would you expect, if you had a team with a budget of 20 million €

budget (in mil. €)	wins
171	33
108	28
119	30
43	17
58	21
56	16
62	29
43	25
57	19
75	31
32	12
29	18

Your turn (5)

A5) The following table shows age and revenue (in $) of 30 customers:

1.) Specify a meaningful relationship and run a corresponding regression
2.) What revenue would you expect on on average for customers who are 30 years old
3.) How old would you expect the customer with the highest revenue

KN_R	AGE	REVENUE
1	32	11
2	58	362
3	53	431
4	69	27
5	48	454
6	35	120
7	38	370
8	42	447
9	28	48
10	35	79
11	62	365
12	60	239
13	44	531
14	51	330
15	38	69
16	38	74
17	55	562
18	24	73
19	68	167
20	56	412
21	32	103
22	47	601
23	40	80
24	63	215
25	31	175
26	35	251
27	30	220
28	35	223
29	68	173
30	43	373

Your turn (6)

A5) The following table shows consumption (in mil €) of an economy:

1.) Analyze the growth during the considered time period.
2.) What was the average yearly growth
3.) Estimate a linear trend regression.
4.) Based on your results of the trend model what consumption would expect in 2010

year	Consumption
1997	1036,5
1998	1066,4
1999	1108,0
2000	1153,7
2001	1221,0
2002	1321,2
2003	1420,7
2004	1536,3
2005	1588,9
2006	1644,5

Your turn (7)

A6) The following contingency table shows the distribution of age and revenue of 50 customers of a company. Analyze whether there is a correlation between both variables

	Revenue (y)					
Age (x)	0-200	200-400	400-600	600-800	800-1000	sum
10-20	2	2	1	0	0	5
20-30	2	6	4	1	0	13
30-40	1	5	5	4	0	15
40-50	0	1	4	3	2	10
50-60	0	0	1	2	4	7
sum	5	14	15	10	6	50

Contents

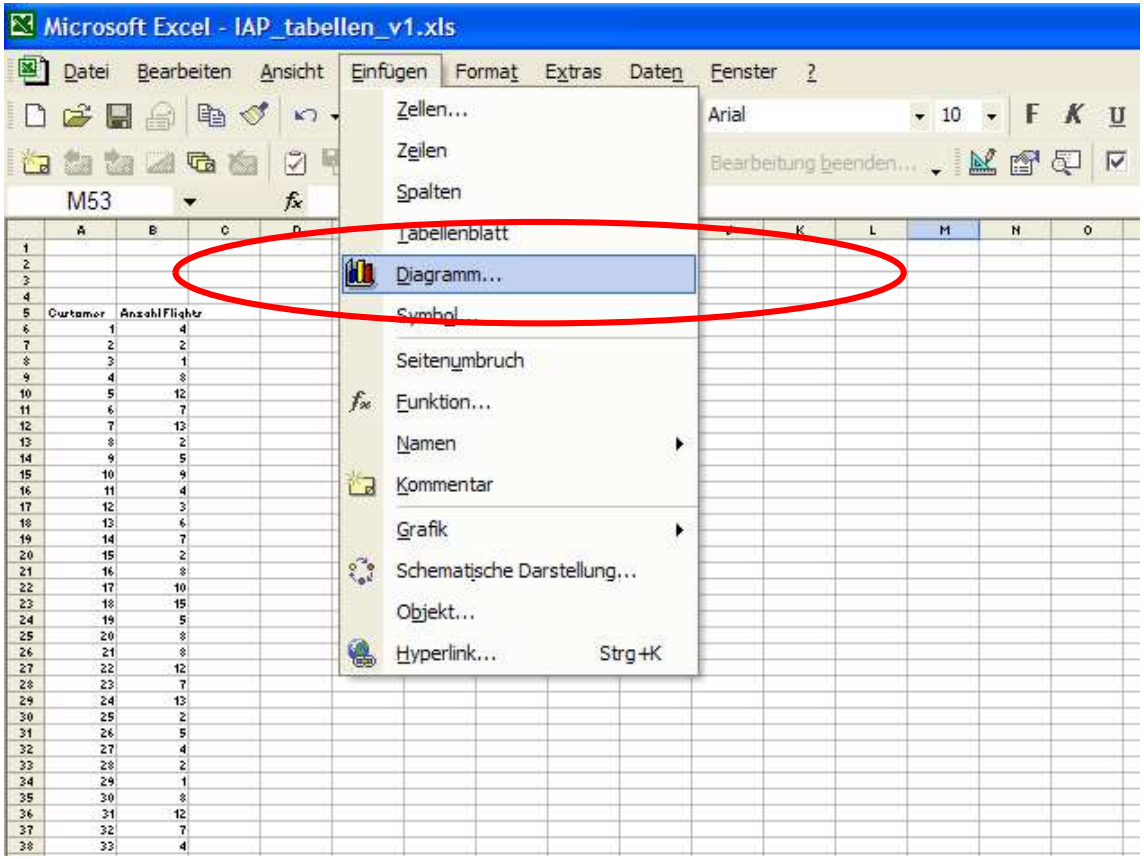
Excel – what will you learn here?

In this chapter you will learn about:

- How to do diagrams
- How to use Excel for displaying frequency distributions
- How to use Excel to calculate descriptive statistics
- How to use Excel for regression analysis

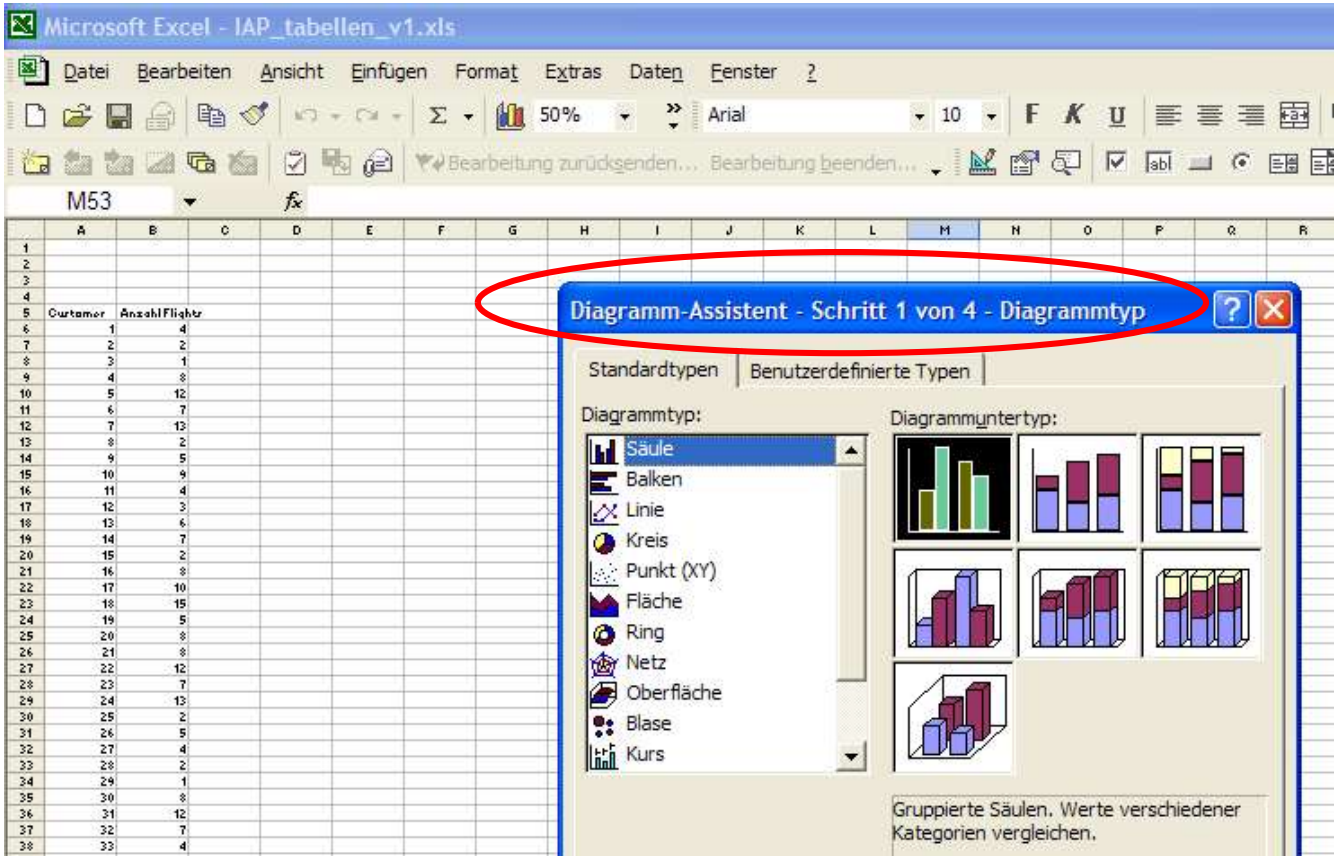
Diagrams (1)

- How to do Diagrams



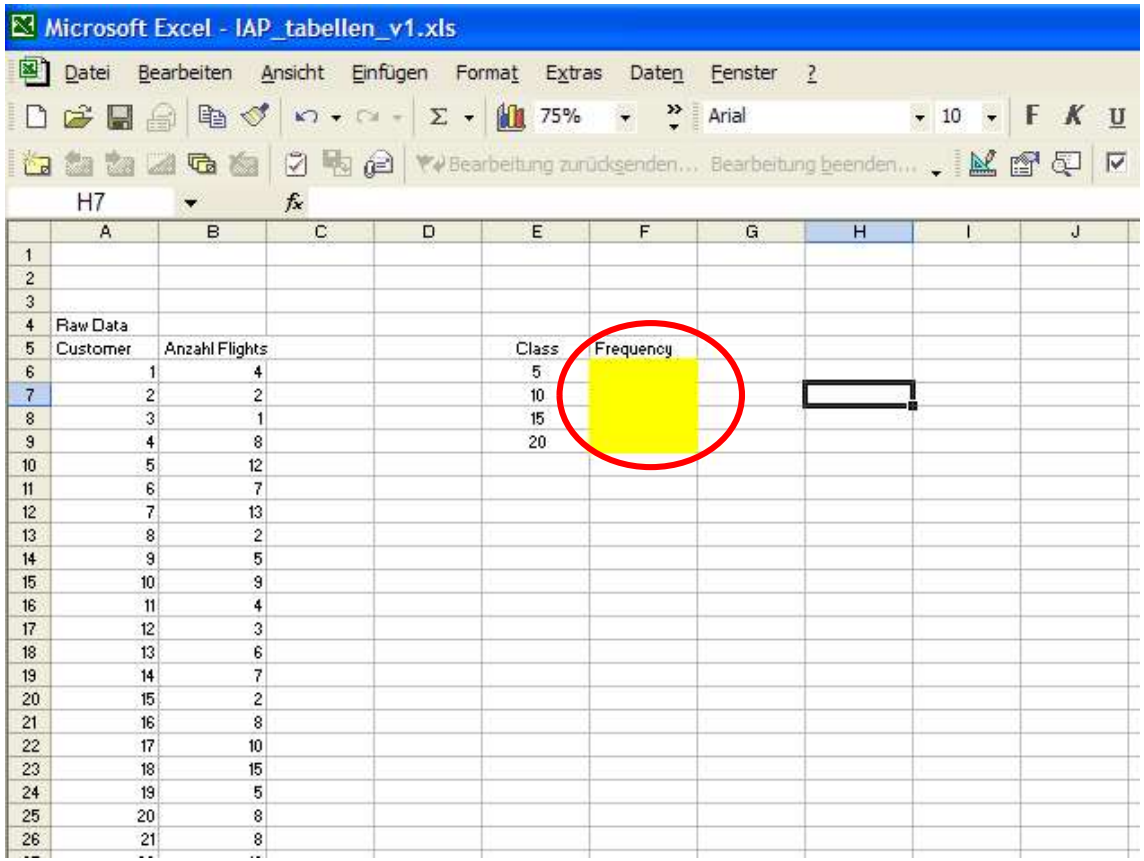
Diagrams (2)

- How to do Diagrams – the diagram assistant



Generating Frequency Distributions out of raw data (1)

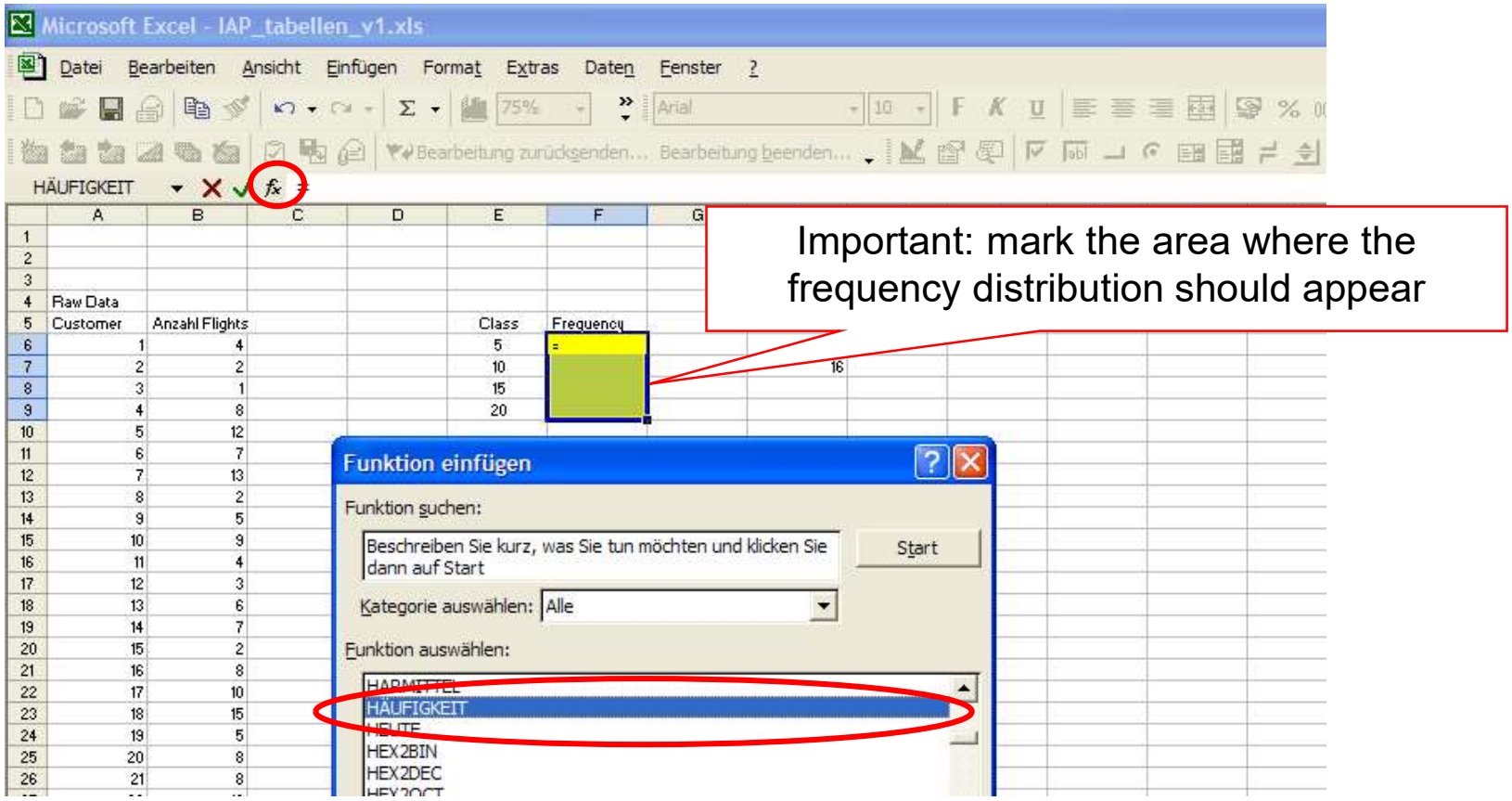
- How to generate frequency distributions



	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4	Raw Data									
5	Customer	Anzahl Flights			Class	Frequency				
6	1	4			5					
7	2	2			10					
8	3	1			15					
9	4	8			20					
10	5	12								
11	6	7								
12	7	13								
13	8	2								
14	9	5								
15	10	9								
16	11	4								
17	12	3								
18	13	6								
19	14	7								
20	15	2								
21	16	8								
22	17	10								
23	18	15								
24	19	5								
25	20	8								
26	21	8								

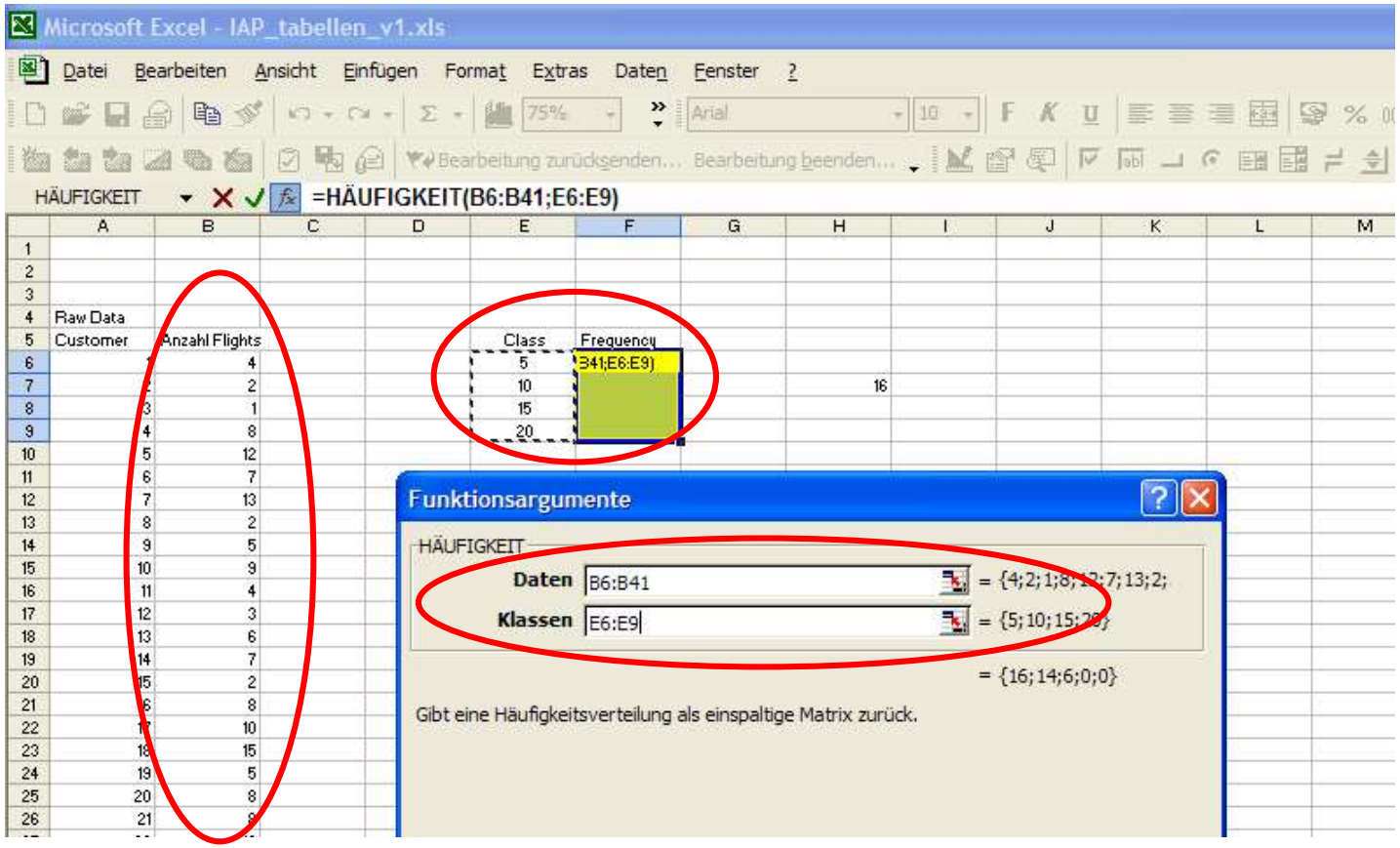
Generating Frequency Distributions out of raw data (2)

- How to generate frequency distributions



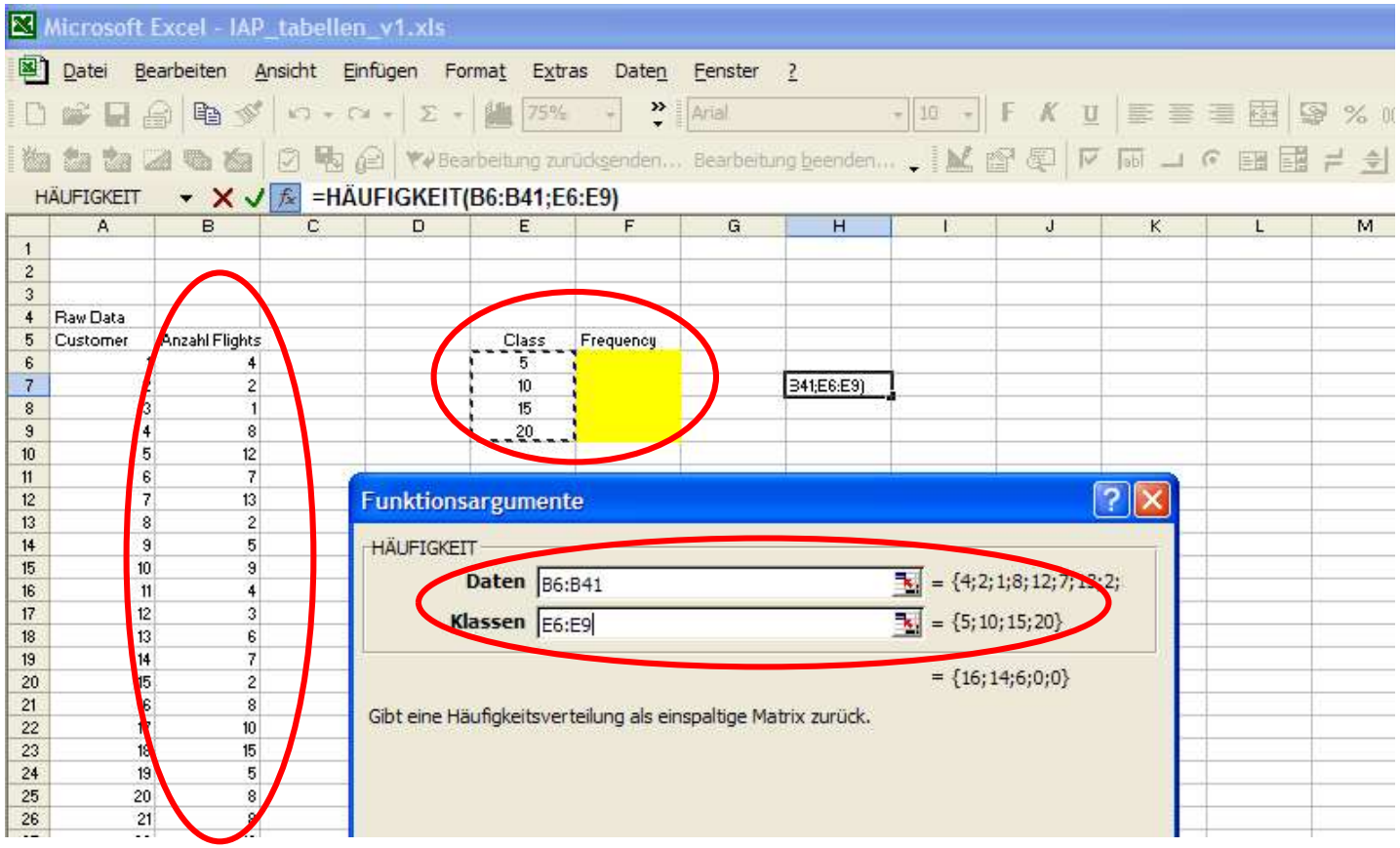
Generating Frequency Distributions out of raw data (3)

- How to generate frequency distributions



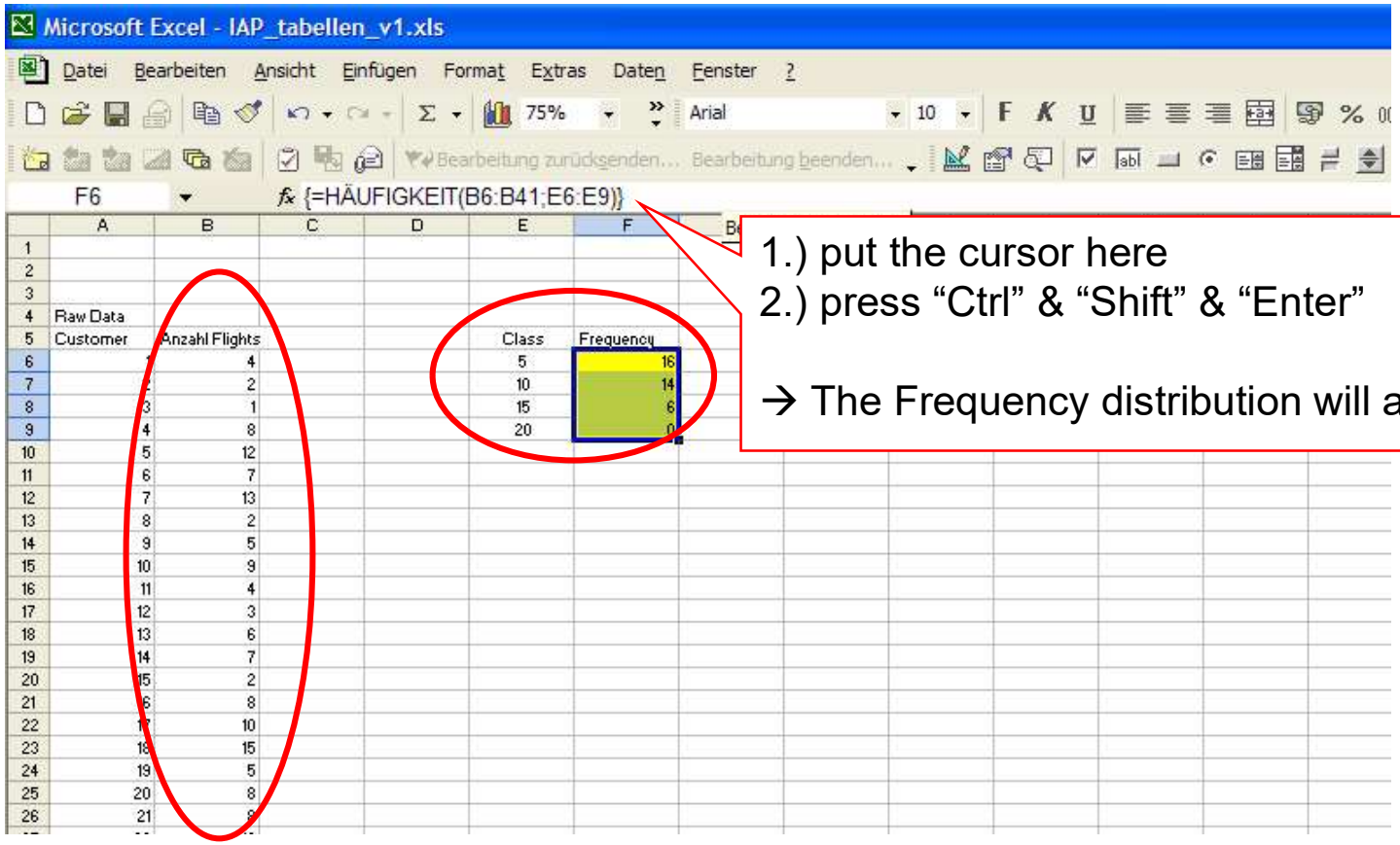
Generating Frequency Distributions out of raw data (4)

- How to generate frequency distributions



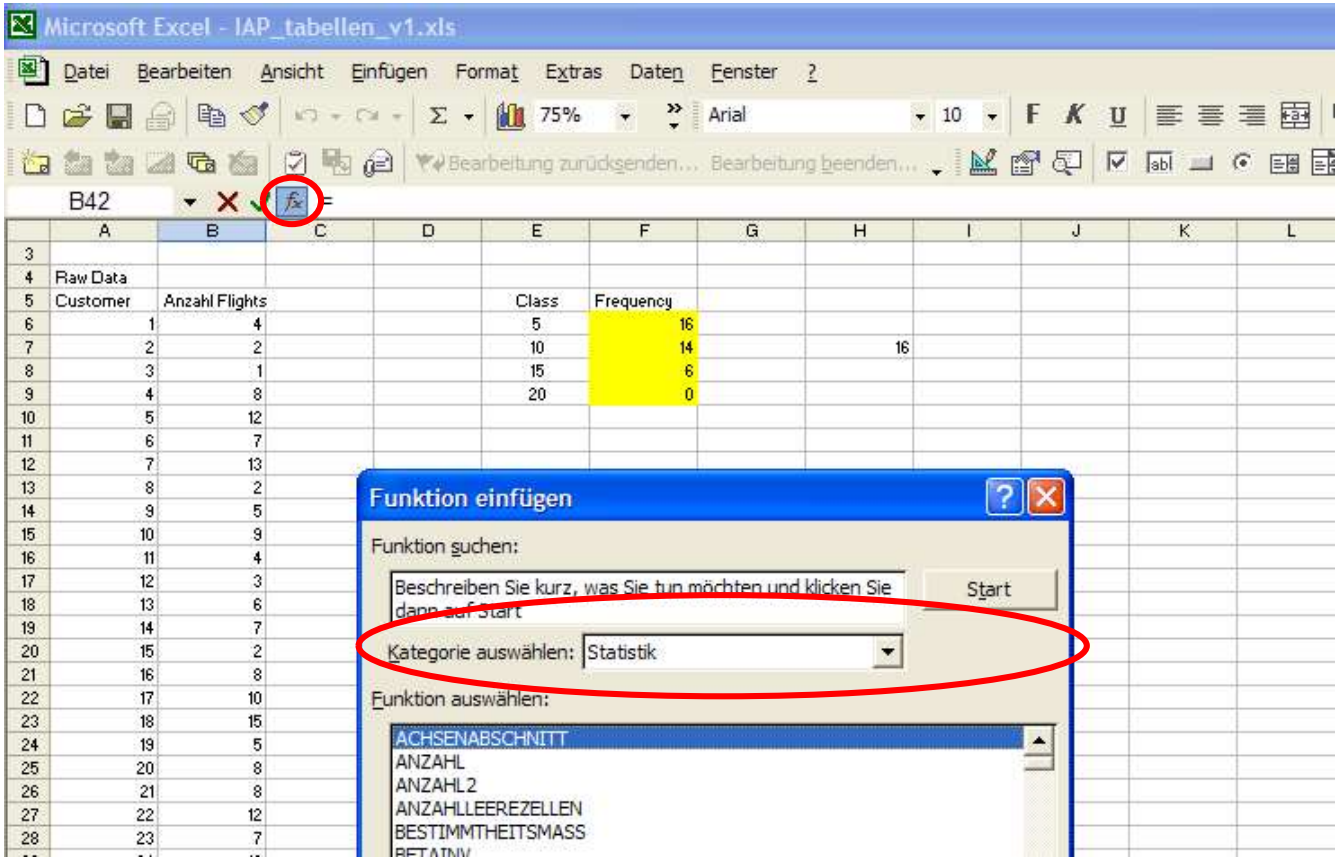
Generating Frequency Distributions out of raw data (5)

- How to generate frequency distributions



Statistical Functions such as Median, Mean, Variance, Quartile, Correlation, … can be found under functions

- Statistical Functions



Statistical Functions such as Median, Mean, Variance, Quartile, Correlation, … can be found under functions

- Some relevant statistical functions in Excel – English vs German

German	English
MAX	MAX
MIN	MIN
MODALWERT	MODE
MEDIAN	MEDIAN
MITTELWERT	AVERAGE
VARIANZEN	VARP
STABWN	STDEVP
QUARTILE	QUARTILE
QUANTIL	PERCENTILE
KOVAR	COVAR
KORREL	CORREL

Excursus (1): Excel calculates the quantiles a little different – How? Example 1. and 3. quartile

- In the previous example
- For $Q_1 = x[0{,}25]$ define:

$$rank = 0{,}25 \cdot (n-1) + 1 \qquad index = Integer(rank) \qquad weight = rank - index$$

$$Q_1 = x[0{,}25] := x_{index} + weight \cdot (x_{index+1} - x_{index})$$

- For $Q_3 = x[0{,}75]$ define:

$$rank = 0{,}75 \cdot (n-1) + 1 \qquad index = Integer(rank) \qquad weight = rank - index$$

$$Q_3 = x[0{,}75] := x_{index} + weight \cdot (x_{index+1} - x_{index})$$

* the 2. quartile (median) is calculated analog

Excursus (2): Excel calculates the quantiles a little different – How? Example 1. and 3. quartile

- Excel uses a slightly different method to calculate the 1. and 3. quartile*
- For $Q_1 = x[0{,}25]$ we get: $rank = 0{,}25 \cdot (10-1)+1 = 3{,}25$

$$index = Integer(3{,}25) = 3 \qquad weight = 3{,}25 - 3 = 0{,}25$$

$$Q_1 = x[0{,}25] := x_3 + 0{,}25 \cdot (x_4 - x_3) = 23{,}741 + 0{,}25 \cdot (42{,}172 - 23{,}741) = 28{,}349$$

- For $Q_3 = x[0{,}75]$ we get: $rank = 0{,}75 \cdot (10-1)+1 = 7{,}75$

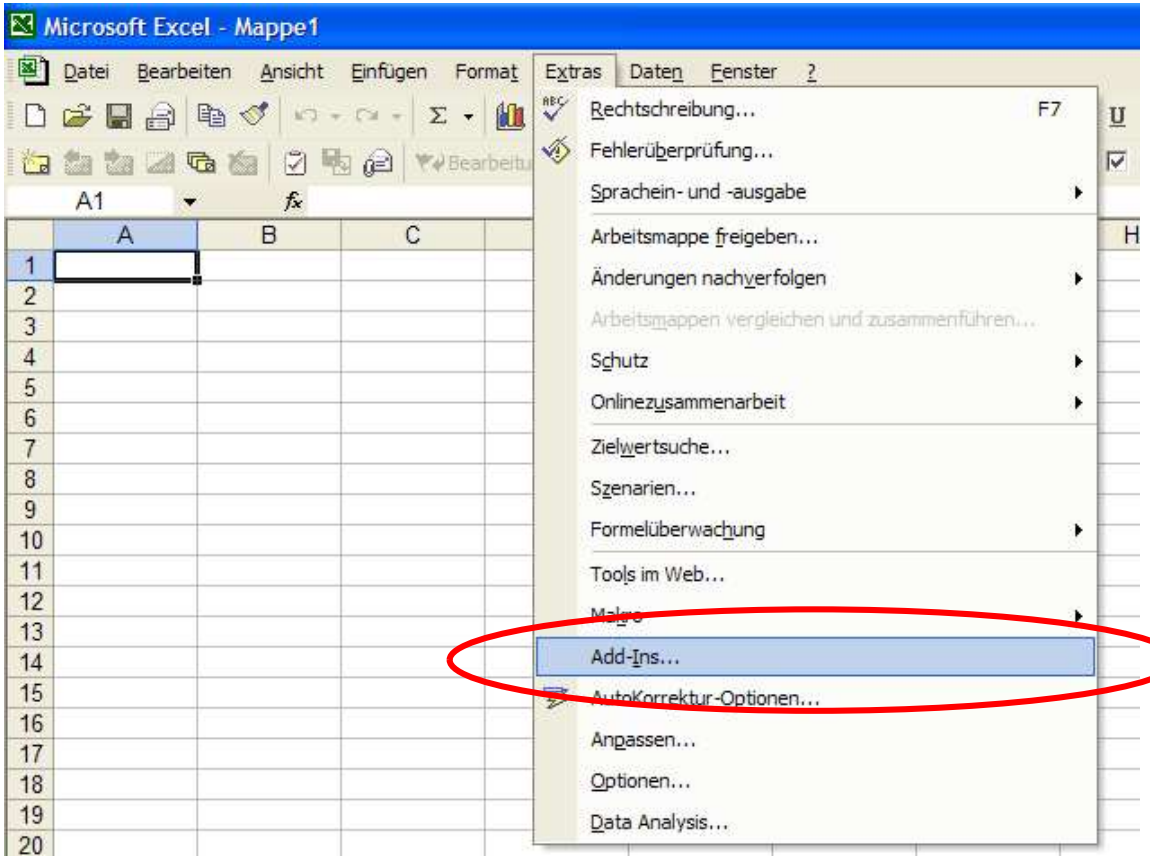
$$index = Integer(7{,}75) = 7 \qquad weight = 7{,}75 - 7 = 0{,}75$$

$$Q_3 = x[0{,}75] := x_7 + 0{,}75 \cdot (x_8 - x_7) = 68{,}363 + 0{,}75 \cdot (72{,}924 - 68{,}363) = 71{,}784$$

* the 2. quartile (median) is calculated analog

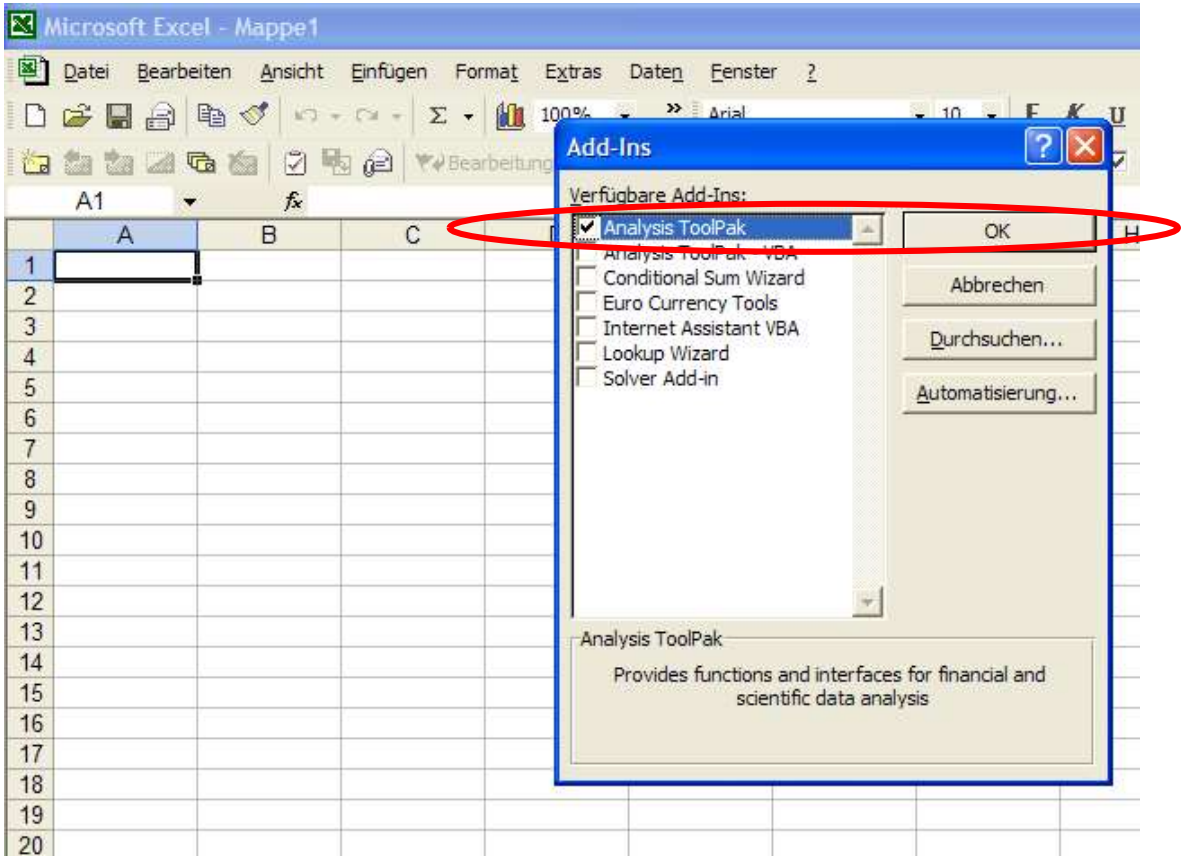
For the Regression Analysis the Data Analysis Add-In is very helpful

- Installing the Data Analysis Add-In



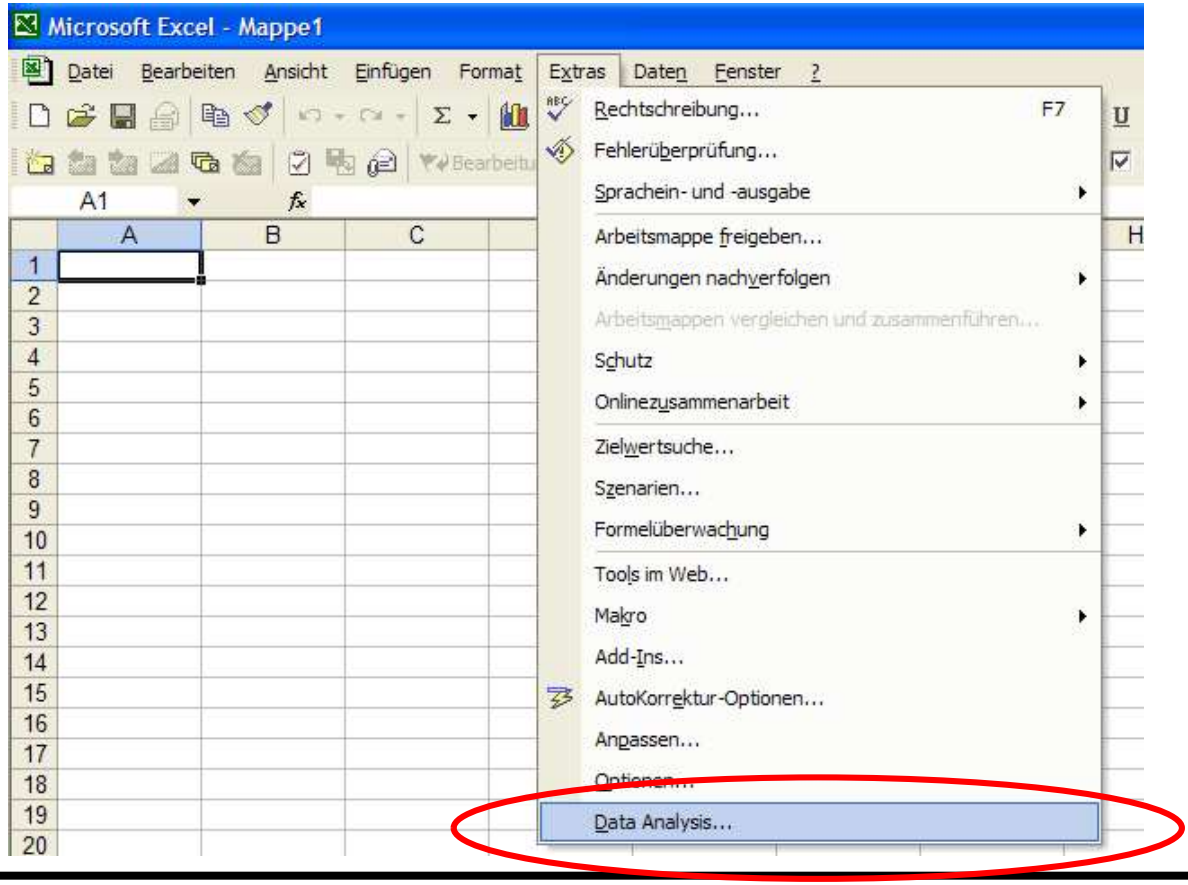
For the Regression Analysis the Data Analysis Add-In is very helpful

- Excel's Add-Ins dialog box



For the Regression Analysis the Data Analysis Add-In is very helpful

- Excel's Tools menu → Data Analysis

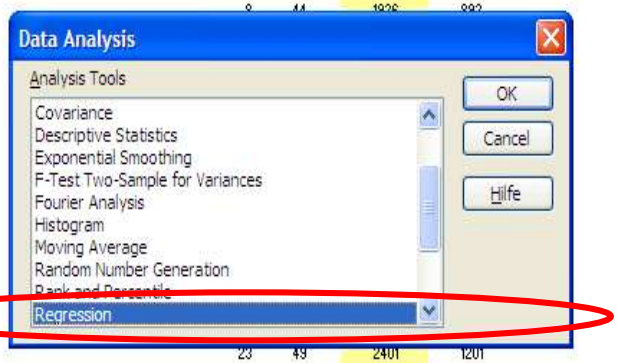


For the Regression Analysis the Data Analysis Add-In is very helpful

- Excel's Data Analysis dialog box → Regression

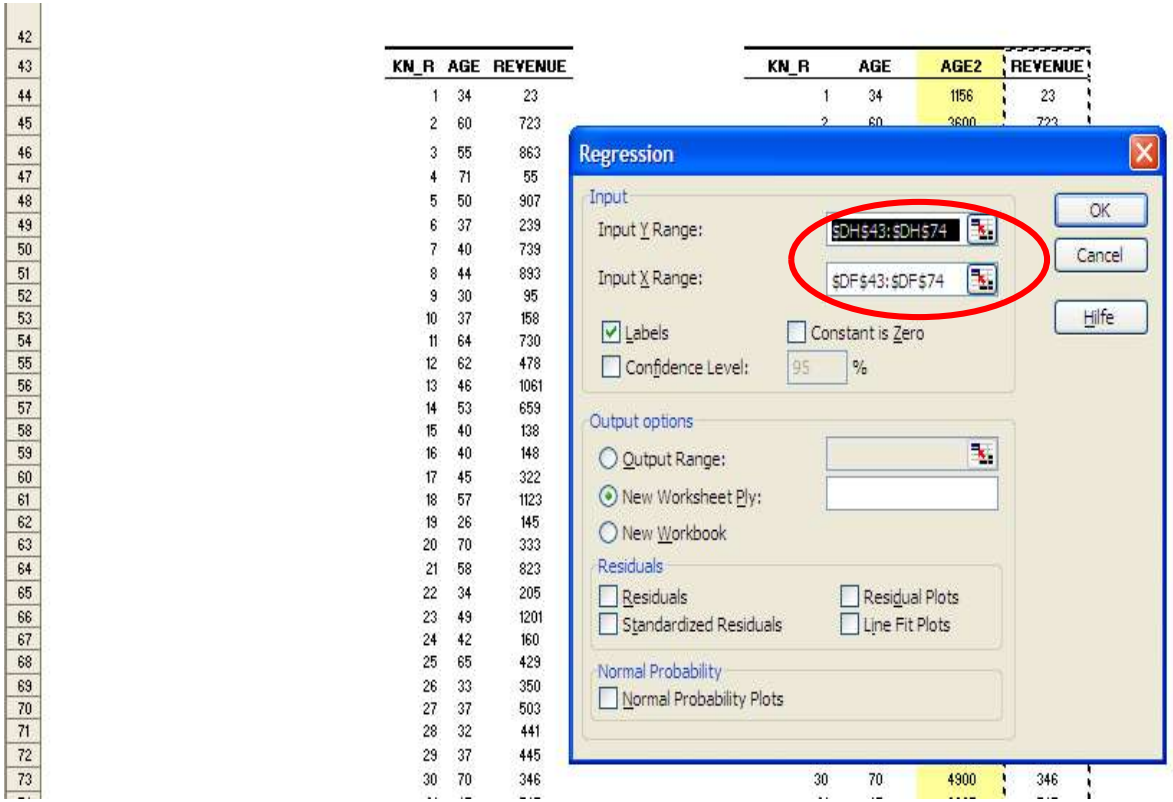
KN_R	AGE	REVENUE
1	34	23
2	60	723
3	55	863
4	71	55
5	50	907
6	37	239
7	40	739
8	44	893
9	30	95
10	37	158
11	64	730
12	62	478
13	46	1061
14	53	659
15	40	138
16	40	148
17	45	322
18	57	1123
19	26	145
20	70	333
21	58	823
22	34	205
23	49	1201
24	42	160
25	65	429
26	33	350
27	37	503
28	32	441
29	37	445
30	70	346

KN_R	AGE	AGE2	REVENUE
1	34	1156	23
2	60	3600	723
3	55	3025	863
4	71	5041	55
5	50	2500	907
6	37	1369	239
7	40	1600	739
23	49	2401	1201
24	42	1764	160
25	65	4225	429
26	33	1089	350
27	37	1369	503
28	32	1024	441
29	37	1369	445
30	70	4900	346



For the Regression Analysis the Data Analysis Add-In is very helpful

- Excel's Data Analysis dialog box → Regression → Specify dependent and independent variables



Contents

Formulas (1)

- Absolute frequency of a class or value:

$$n(x_i) = n_i$$

- Cumulative absolute frequency:

$$\sum_{j=1}^{i} n_j = n_1 + n_2 + n_3 + \ldots n_i$$

- Relative frequency:

$$h(x_i) = \frac{n_i}{n}$$

- Cumulative relative frequency:

$$\sum_{j=1}^{i} h(x_j) = h(x_1) + h(x_2) + \ldots + h(x_i) = h(x \leq x_i)$$

Formulas (2)

- Mode is the observation in the data set that occurs the most frequently
- Mode in case of grouped data. Class midpoint with the highest density:

$$f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$$

- Median, when n is an odd: $\tilde{x} = x_{((n+1)/2)}$
- Median, when n is an even number: $\tilde{x} = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2)+1)}\right)$
- Median in case of grouped data: $\tilde{x} = x_i^u + \frac{50 - h(x \le x_i^u)}{h(x_i)} \cdot \Delta x_i$
- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + x_3 + x_4 + \ldots + x_n)$
- Mean in case of grouped data: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$ or $\bar{x} = \sum_{i=1}^{k} x_i h(x_i)$

Formulas (3)

- Geometric mean: $GM = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots x_n}$
- Range: $R = x_{(h)} - x_{(l)}$
- Range in case of grouped data: $R = x^u_{(h)} - x^l_{(l)}$
- Variance: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Variance in case of grouped data: $s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$ or $s^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 h(x_i)$
- Standard deviation: $s = \sqrt{s^2}$
- Quartiles (Q_1, Q_2, Q_3) if the product nq an integral number: $x[q] = \frac{1}{2}(x_{nq} + x_{nq+1})$
- Quartiles (Q_1, Q_2, Q_3) if the product nq is no integral number one has to take the next higher integral number <nq>: $x[q] = x_{<nq>}$

Formulas (4)

- Quartiles (Q_1, Q_2, Q_3) in case of grouped data: analog Median in case of grouped data
- Interquartile range: $IQR = Q_3 - Q_1$
- Simple approach to determine outliers (bounds): $Q_3 + 1.5 \cdot IQR$ and $Q_1 - 1.5 \cdot IQR$
- Formulas for the Lorenz Curve:

$$u_j = \frac{j}{n}, \qquad p_j = \sum_{i=1}^{j} x_i, \qquad p_n = \sum_{i=1}^{n} x_i, \qquad v_j = \frac{p_j}{p_n}$$

- Formulas for the Lorenz Curve in case of grouped data:

$$u_j = \frac{1}{n}\sum_{i=1}^{j} n_i, \qquad p_j = \sum_{i=1}^{j} n_i \cdot x_i, \qquad p_n = \sum_{i=1}^{n} n_i \cdot x_i, \qquad v_j = \frac{p_j}{p_n}$$

Formulas (5)

- GINI Coefficient: $GINI = \frac{act.\,conc.\,area}{\max.\,conc.\,area} = \frac{K}{K_{\max}}$

  with $K = \frac{1}{2} - \sum_{j=1}^{k} \frac{1}{2}\left(M_{j-1} + M_j\right) \cdot h_j$ and $K_{\max} = \frac{1}{2}$

- Simple index number: $x_t = 100 \cdot \frac{y_t}{y_0}$ for $y_0 \Rightarrow x_0 = 100$

- Laspeyres price index: $I_{0,t}^{L}(p) = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} \cdot 100$

- Paasche price index: $I_{0,t}^{P}(p) = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i} \cdot 100$

Formulas (6)

- Correlation coefficient: $r = \frac{s_{xy}}{s_x \cdot s_y}$

  with $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$

- Simple regression analysis: $\hat{y} = a + b \cdot x$

  with $b = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$ and $a = \bar{y} - b \cdot \bar{x}$

Formulas (7)

- Coefficient of Determination: $D = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}$

  or in case of a simple regression also: $D = \left(\frac{s_{xy}}{s_x \cdot s_y}\right)^2 = r^2$