MBA
AVIATION & TOURISM
MANAGEMENT

FRANKFURT
UNIVERSITY
OF APPLIED SCIENCES

Research Methods

Faculty of Business and Law
Frankfurt University of Applied Sciences

September 2018

LECTURER

Prof. Dr. Dilek Bülbül
Professor of Finance
Frankfurt University of AS

Office: Building 4, Room 406
Email: bulbul@fb3.fra-uas.de

Robert Schorr, M.Sc.
Department of Mathematics
Technische Universität Darmstadt

LECTURE

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

17 September 9:00 - 17:00

Introduction to research (Bülbül)
Thinking about methods
Methods, group work

18 September 9:00 - 12:30

Methods, group work (Bülbül)

18 September 13:30 – 17:00

Analyzing Data, Descriptive Statistics, etc. (Schorr)

LECTURE

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

9:00-10:30

11:00-12:30

12:30-13:30 Lunch

13:30-15:00

15:30-17:00

ESSENTIAL READINGS

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Text books:

- 🕮 Blaxter, L.; Hughes, Ch.; Thight, M. “How to Research” Open University Press, 2010
- 🕮 Bryman, A.; Bell, E. “Business Research Methods, Oxford University Press, 4th Edition, 2015
- 🕮 Heiss, F. “Using R for Introductory Econometrics”, 2016
- 🕮 Maylor, H.; Blackmon, K.; Huemann, M. “Researching Business and Management, 2nd Edition, 2017
- 🕮 Wooldridge, J. “Introductory Econometrics”, 2015

I. How to research

How to research

Characteristics of business and management research

I - Research is original

II - Research is relevant

III - Research has immediacy

How to research
Characteristics of business and management research

I - Research is original

- Make an **original contribution to knowledge**
- Research needs some degree of originality

Originality may come through building on existing knowledge but providing

- New or improved insight or evidence
- New or improved methods for doing research
- New or improved analysis of data
- Etc…

How to research
Characteristics of business and management research

II - Research is relevant

Asking questions to solve a problem of incomplete knowledge, which maybe practical or theoretical

Practical problem

- Real-life-situation: issues observed in a real life setting

Theoretical problem

- Investigate how to apply the theories and models to real-life-setting
- Understand which of competing theories best explains how people or organization behave

How to research

Characteristics of business and management research

III - Research has immediacy

Academic research has different level of immediacy

- Practical problems have high immediacy
- Theoretical problems have low immediacy

Research	Role of this type of research
Basic research	Conducted to increase knowledge with little consideration of future application
Development	Taking an original idea, possibly a basic research project, and looking for application
Commercial	Taking an idea from the possibility of application through to commercial usage

How to research

Different type of research: basic research, development, commercial

Ideas worth spreading

The next step for airlines | Sadiq Gillani | TEDxBerlin

https://www.youtube.com/watch?v=uXeVJDvJ2-c

Advancing Airline Single-Pilot Concept | James Green | TEDxUVU

https://www.youtube.com/watch?v=etIH9PHpVyM

How to research

Different type of research: basic research, development, commercial

Ideas worth spreading

The next step for airlines | Sadiq Gillani | TEDxBerlin

Is there customer demand for innovation?

Passengers willingness to provide private information to get access to innovative services

Determinants for taking the decision to implement innovative ideas

How to research - Theories

Relationship Research and Theory

- Background literature on an topic defines the focus of research and thereby acts as a “theory”.
- Research is guided by theoretical ideas and the aim of research is to make a contribution to theory

Theory	→	**Observations/ Findings**	**… deductive**
Observations/ Findings	→	**Theory**	**… Inductive**

How to research – Deductive Theory

- Theory to be tested before they can considered be valid or useful
- Typically associated with quantitative research approach

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

How to research – Inductive Theory

- Theory is the outcome of research (generation of theory)
- Drawing generalizable inferences (conclusions) out of observations
- Typically associated with qualitative research approach

How to research

How to research – Research question

- Narrowing down the research topic to a manageable scope
- Research question focus on specific areas of your research topic

Research question guide you what you will do

Research question determine…

- …the existing research that you use to support your work
- …the data you will collect
- …the model you will use
- …how you report your research

How to research - Hypothesis

Hypothesis

- A hypothesis is in a sense a research question, but it is not stated as a question and provides an anticipation of what will be found out

- How to develop hypothesis?!
- Let's see an example….

How to research - Hypothesis

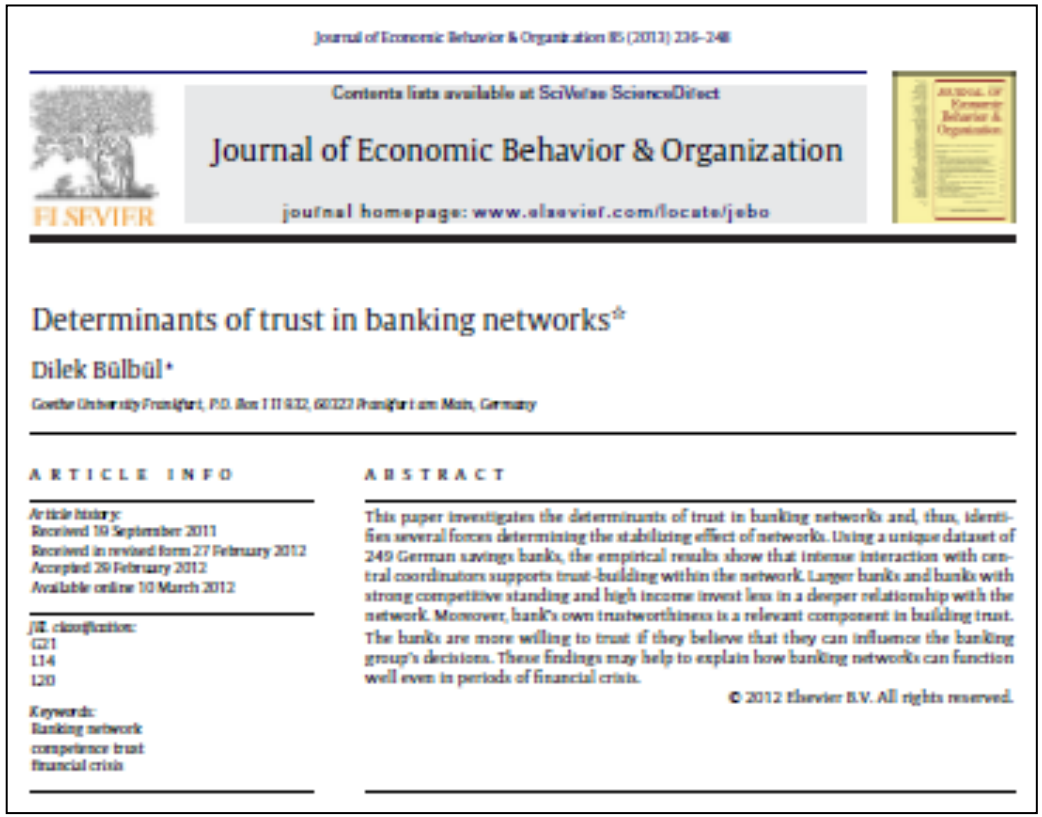
Hypothesis

Drinking sugary drinks daily leads to obesity.

- The dependent variable(s): who or what you expect to be affected
- The independent variable(s): who or what you predict will affect the dependent variable
- What you predict the effect will be.

How to research - Hypothesis

Example *Dilek Bülbül "Determinants of trust in banking networks", Journal of Economic Behavior & Organization, 2013, Vol. 85, S. 236-248,*



Journal of Economic Behavior & Organization 85 (2013) 236–248

Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Determinants of trust in banking networks☆

Dilek Bülbül*

Goethe University Frankfurt, P.O. Box 111932, 60323 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 19 September 2011
Received in revised form 27 February 2012
Accepted 28 February 2012
Available online 10 March 2012

JEL classification:
G21
L14
L20

Keywords:
Banking network
competence trust
financial crisis

ABSTRACT

This paper investigates the determinants of trust in banking networks and, thus, identifies several forces determining the stabilizing effect of networks. Using a unique dataset of 249 German savings banks, the empirical results show that intense interaction with central coordinators supports trust-building within the network. Larger banks and banks with strong competitive standing and high income invest less in a deeper relationship with the network. Moreover, bank's own trustworthiness is a relevant component in building trust. The banks are more willing to trust if they believe that they can influence the banking group's decisions. These findings may help to explain how banking networks can function well even in periods of financial crisis.

How to research - Hypothesis

Example *Dilek Bülbül "Determinants of trust in banking networks", Journal of Economic Behavior & Organization, 2013, Vol. 85, S. 236-248,*

During financial crisis:
financial intermediaries in the interbank market stopped their activities

Financial crisis:
strong need to understand the role of trust in complex and interlinked networks

Uncertainty (no trust) is an important determinant for limited market participation

How to research - Hypothesis

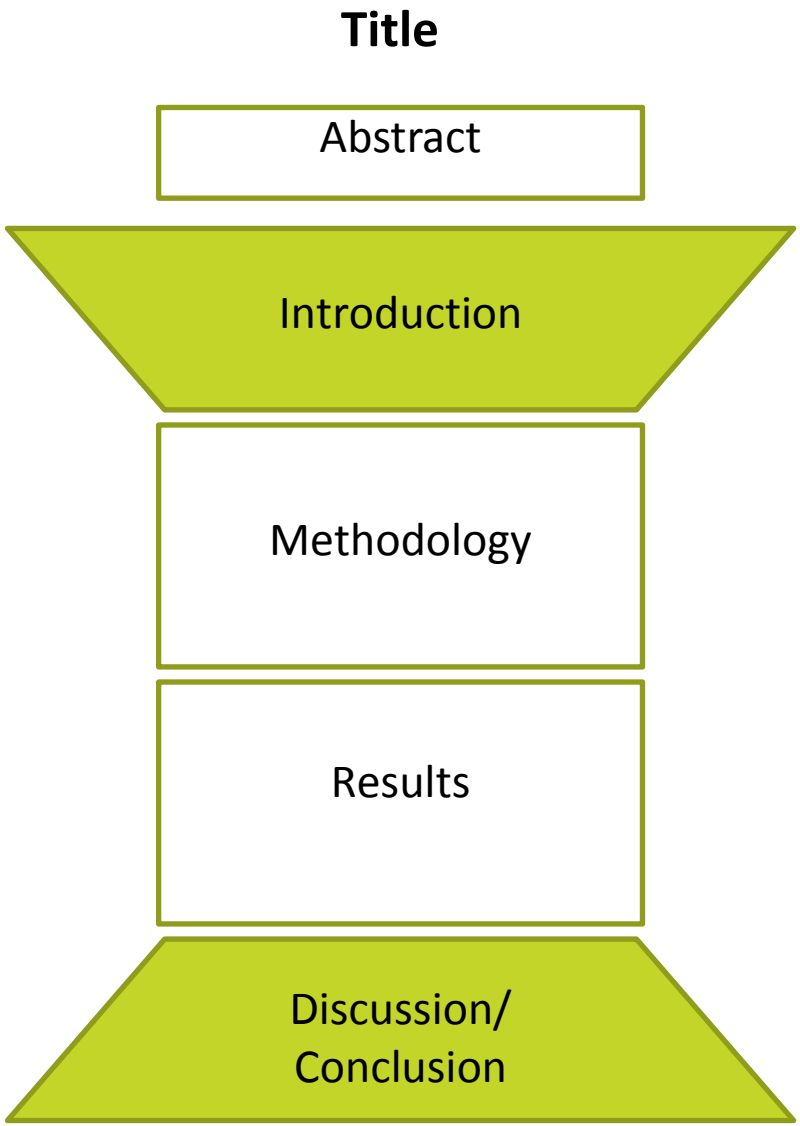
Example *Dilek Bülbül "Determinants of trust in banking networks", Journal of Economic Behavior & Organization, 2013, Vol. 85, S. 236-248,*

Starting Point:
Trust is important for the stability of banking networks
How banking networks can function well even in periods of financial crisis

Research Question:
What are the determinants for trust in banking networks?

Hypotheses:
see Paper

Paper in scientific journal style and format



Max. 150 words

Paper in scientific journal style and format

Identify an interesting research question. Explain why your search question is relevant. What do you want to investigate and why?

What is the current stand of literature? Provide a literature review of relevant scientific literature related to your topic (cite also recent literature), Non-published papers can be part of your literature review if you are convinced by its scientific quality (e.g. publication of relevant institutions such as ECB, BIS, EU, etc.).

Explain your conclusions and interpretations of your literature review and (own) findings.

Paper in scientific journal style and format

Develop hypotheses related to your research question.

Support your arguments by applying appropriate methodology.

Provide your results: test your hypotheses, answer your research question.

Conclude with a conclusion.

Research Example – PhD Project

School of Aviation
PhD Project

An empirical investigation of the relationships between airline service quality, perceived price, service value, passenger satisfaction, airline image, and passengers' behavioral intentions.

This research seeks to investigate the relationship of airline service quality with price, service value, airline image, passenger satisfaction and passengers' behavioral intentions. By acquiring data from both Korean and Australian passengers, the purpose of this research is to study and to understand choice behavior of airline passengers.

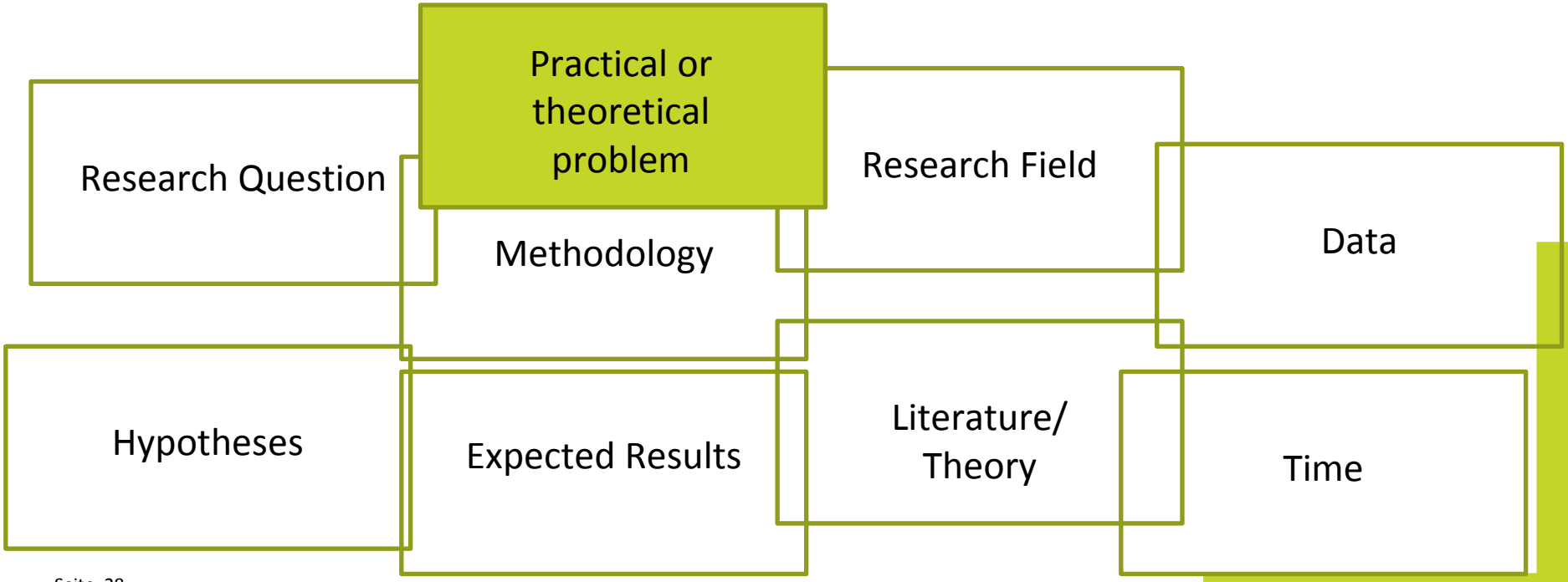
Research Example – Masters Projects

School of Aviation
Masters Projects

- On the flight choice behavior of business purpose passengers in the Australian domestic market
- National culture and its impact on airline corporate culture
- The low cost model evolution: a case study of Jetstar International
- The development of aviation in the Middle East and its influence on the European aviation industry
- ***Predicting pilots risk taking behavior***
- ***Effects of caffeine on pilot performance***
- ***Do Airlines need A380?***

Getting Started - Research Proposal

- Clarify your own ideas
- Document your ideas so that you can discuss with other people
- Starting point for your research and point of reference that you can come back to, should things not progress as you plan



II. Thinking about methods

Thinking about methods

The research method has to follow the research question

The way questions are asked influences what needs to be done to answer them

Your research question should determine your approach and techniques

Thinking about methods

Research

Quantitative or Qualitative

Deskwork or Fieldwork

Research techniques

Documents

Interviews

Observations

Questionnaires

Research approach

Surveys

Experiments

Case Studies

Thinking about methods

Quantitative vs qualitative methods

Quantitative research is empirical research where the data are in the form of numbers.

Quantitative research, with an emphasis on closed-ended research design

Qualitative research is empirical research where the data are not in the form of numbers.

Qualitative research, with an emphasis on open-ended research design

III. Quantitative methods

Surveys
Experiments

Surveys

All surveys ask structured questions with defined types of answers across a set of respondents

Survey enables to collect comparable data from multiple respondents

Surveys are used to gather data about organizations or their members through
(1) questionnaires,
(2) structured interviews,
(3) structured observations

Surveys

You should consider using a survey if…..

- Collect data from a large number of respondents
- Collect the same data the same way across a group of people
- Respondents can understand and answer your questions
  - with minimal explanation (structured interviews)
  - or even without being physically present (questionnaire)

Surveys

You should consider using a survey if…..

- Sensitive information that respondents may only provide anonymously (questionnaire)
- Large number of responses to analyze statistically in order to have a confidence in your findings

Questionnaires

Different ways in which questionnaires can be administered

- Sent by post
- Over the telephone
- Face to face (highly structured interviews)
- Sent over the internet/ email

Pre-Test and Pilot your survey

Pilot your questionnaire before you carry out the full survey, and to modify your questions in the light of the responses you receive

Getting enough people to participate is probably most the biggest barrier to using questionnaires

Questionnaires

Good survey questions are…

Clear

Structure your responses so that it is clear what it means instead making your respondent to interpret them

“seldom” vs “once a year” vs “less than ten times a year”

→ Give precise frame and spell out or avoid technical terms that may be unfamiliar to respondents

Questionnaires

Good survey questions are…

Simple

Avoid multiple questions or general questions.

"How often do you walk or use the bus and train to get to work?"

Vs

"How often do you walk to get to work?"

"How often do you use the bus to get to work?"

"How often do you use the train to get to work?"

→ Make sure that each question is only a single question

Questionnaires

Good survey questions are…

Unbiased

Avoid asking leading questions.

"Are you in favor of raising taxes to waste money on able-bodied people who could work but don't .

→ You are conducting research to find out information, not confirm your own opinions.

Questionnaires

Good survey questions are…

Brief

Avoid long questions, because your respondent may lose interest and skip the question or in interviews where it is difficult for your interviewee to remember the entire question

Questionnaires

Keep in mind….

That you might still get inaccurate answers depending on how you ask the questions and whether your respondent actually knows the answer

"How many times did you go the cinema last year?"

Vs

"On average, how often do you go the cinema monthly?"

Questionnaires

Design your questionnaire…

- (1) Begin with simple questions and
  (2) put difficult questions at the end,
  (3) put an awkward or embarrassing question last

- Divide the questionnaire in sections

- Provide clear and explicit instruction on how to answer the questions

- Work on the layout

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Questionnaires – Probability Sampling

- If you want to generalize your findings the sample must be representative of the population
- Otherwise, the sample is “biased”, over-representing some members of the population and under-representing others
- Select units/respondents randomly from your population
- Each unit/respondent has equal probability of being selected (random selection)

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Questionnaires – Probability Sampling

Techniques	
Simple random sampling	any particular member of the population is equally likely to be selected
Systematic sampling	e.g. selecting every tenth employee on the list
Stratified random sampling	subdivide the population according to some criterion and then apply random sampling, e.g. year of study
Cluster sampling	select entire sample from particular subset (cluster) that is representative of the entire population but not random, e.g. single hall of residence to represent all first-year student houses

Questionnaires – Non-Probability Sampling

- Some units have a greater probability of being selected than other units **(selection bias)**
- Some units have a greater probability of selecting themselves than other units **(self- selection bias)**

- Generalizing of your results not possible but….

→ Lessons learned from the sample

Questionnaires – Non-Probability Sampling

Techniques	
Volunteer recruitment	people who volunteer tend to be different from the general population
Convenience sampling	approaching mates, friends, family. This sample is better suitable for pre-test or pilot
Snowball sampling	Starting with a singe person and expand sample based on contacts known to or suggested by your original respondent
Quota sampling	Define a categories and keep recruiting respondents until you have enough representatives of each category (e.g. 50 men, 50 women; opinion pooling for election research)

Questionnaires – Types of questions

Quantity or Information

In which year did you enroll on the part-time degree?

Category

Have you ever been, or are you now, involved almost full-time in domestic duties (.i.e. as a housewife/househusband)?

Yes, currently ☐ Yes, in the past ☐ Never ☐

Questionnaires– Types of questions

List or multiple choice

Do you view the money spent on your higher education as any of the following?

A luxury	☐	An investment	☐	A necessity	☐
A gamble	☐	a burden	☐	a right	☐
None of these	☐				

Questionnaires – Types of questions

Scale

How would you describe your parents' attitude to higher education at that time? Please tick one of the options below:

Very positive	positive	neutral	negative	very negative	not sure
☐	☐	☐	☐	☐	☐

Questionnaires – Types of questions

Likert Scale

<u>Statement! Not a question!</u>
“My job is like a hobby to me.”

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	not sure
☐	☐	☐	☐	☐	☐

Questionnaires – Types of questions

Ranking

What do you see as the main purpose(s) of your degree study? Please rank all those relevant in order form 1 downwards:

Personal development	☐
Career advancement	☐
Subject interest	☐
Keeping stimulated	☐
Other	☐

Questionnaires – Types of questions

Complex grid or table

How would you rank the benefits of your degree study for each of the following. Please rank each item:

for:	very positive	positive	neutral	negative	very negative	not sure
you						
your family						
your employer						
your community						

Questionnaires – Types of questions

Open-ended

We would like to hear from you if you have any further comments:

Questionnaires – Types of questions

How to code the questions and answers to make it useable for statistical analyses (empirical research)

Remember:

Quantitative research is empirical research where the data are in the form of numbers.

Qualitative research is empirical research where the data are not in the form of numbers.

Questionnaires - Example

Research Project at Frankfurt University of Applied Sciences - HOLM Halal-Logistic

Research Question:

Sind Muslimas und Muslime in Deutschland daran interessiert, Produkte mit einem Halal-Zertifikat zu konsumieren und wie ist ihre Zahlungsbereitschaft?“

Sample:

- 772 respondents were interviewed via an online survey questionnaire during the period from September 2016 – April 2017.
- Self-selected sample (non-probability sample)
- 26 questions
- Target group: consumers from different age groups, professions, and social background
- Controlled distribution of the survey link through Muslim associations.

Questionnaires - Example

Research Project at Frankfurt University of Applied Sciences - HOLM Halal-Logistic

Selection of Questions:

F1 Sind Sie Muslim/a

F2 Religiöse Praxis?

F3 Was verstehen Sie unter halal? Sie können mehr als eine

F4 Wie wichtig ist es Ihnen, Halal-Produkte zu konsumieren?

F5 Bei welchen Produkten achten Sie auf halal?

F21 Geschlecht

F22 Persönliches Jahreseinkommen (netto, nach Abzug der Steuern)

F23 Alter

F26 Ausbildungsabschluss

You will be provided with the raw data
→ exercise statistical analysis

Structured Interviews

The interview method involves questioning or discussing issues with people

Structured Interviews is very similar to questionnaire

- Ask the same question in the same order to every interviewee
- Face-to-face, over the phone, over internet
- Record the answers directly
- Challenge: maintaining consistency

Structured Observations

- Observing behaviors and recording the information to a schedule
- Collect data **indirectly** by observing the traces people leave in the physical environment or other natural setting (e.g. in a shop, shopping behavior)
- Data collected in the natural setting of organizations and people

  - Collect data about sensitive issues or respondents unwilling to answer
  - in questionnaires e.g. respondents may over-report positive, socially desirable behaviors and under-report negative behaviors or attitudes

Experiments

Experiments

- Experiments is a structured process for testing how varying one or more inputs affects one or more outcomes
- Testing cause-and -effect relationship

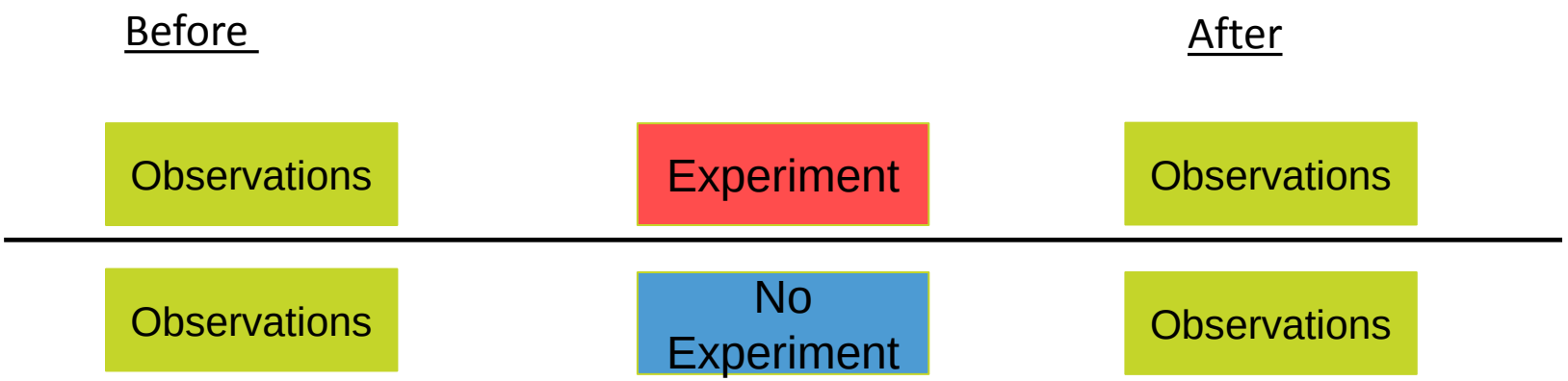
- Experiments typically known from natural and applied sciences
- Experiments are conducted often in controlled settings such as laboratories

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Experiments

- Experimental group/ treatment group receives the treatment, and it is compared against the control group, which does not receive the treatment.
- Experimental manipulation of the independent variable
- Dependent variable is measured before and after the experimental manipulation/ treatment
- Units are assigned randomly to their respective groups

Experiments



Example: medical treatment
to lower cholesterol

Experiments –Laboratory experiments

- Experiment takes place in a laboratory
- Could be classrooms, model of an office, factory or other business settings
- Appropriate when you are investigating basic aspects of how people behave rather than complex social and organizational phenomena

Criticism

- Laboratory settings can be artificial and too simplified compared with the real world
- Tasks people are asked may not closely represent actual task in organizational settings
- Experiments participants are often students rather than managers

→ **Laboratory experiments are high in control, but low in realism**

Experiments –Field experiments

- Takes place in its natural setting, real-life setting
- Workplace (office, shops), households, public spaces etc.

→ **Field experiments are low in control, but high in realism**

Experiments – Quasi experiments

- Takes place in real-life setting
- Control group is compared to a treatment group
- But there is no random assignment
- E.g. comparison company A with company B

Example: Day-care-centers (Kindergarten)

- The deterrence hypothesis predicts that the introduction of a penalty that leaves everything else unchanged will reduce the occurrence of the behavior subject to the fine.
- field study in a group of day-care centers that contradicts this prediction.
- Parents used to arrive late to collect their children, forcing a teacher to stay after closing time.
- We introduced a monetary fine for late-coming parents.

Example: Day-care-centers (Kindergarten)

Experiment

- 10 day-care-centers in Haifa, Israel
- 6 treatment groups, 4 control groups (randomly assigned)
- Opening hours 7:30 - 16:00
- Experiment: 20 weeks

- After 5th week: Introduction of a fine of 2 Euros for parents in the treatment group.
- (Babysitter earn 3€- 4€ per hour, Violation fees for parking. 15€ and violation fees for crossing red signs 200€)
- After 17th week no fine

Example: Day-care-centers (Kindergarten)

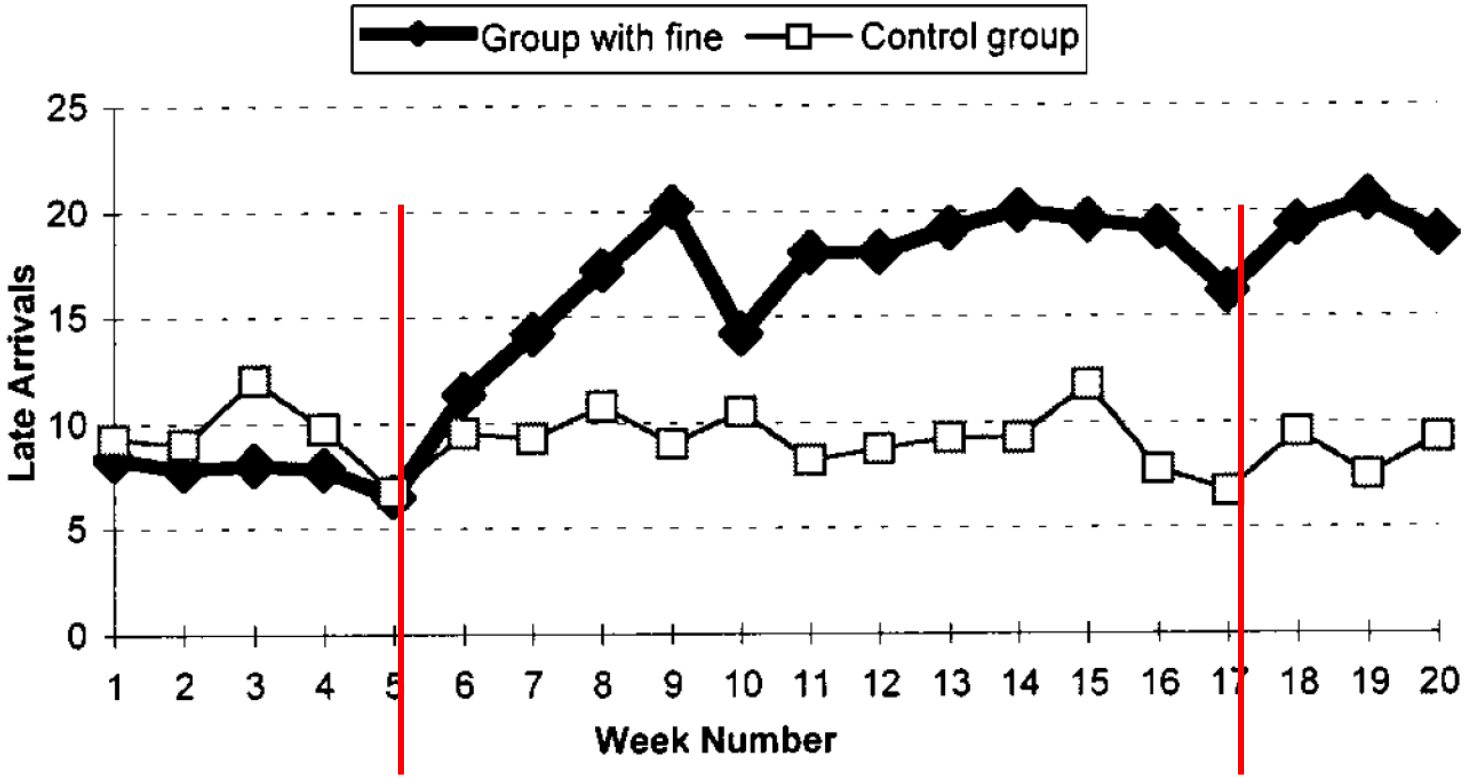


FIGURE 1.—Average number of late-coming parents, per week

Example: Day-care-centers (Kindergarten)

Results of Experiment

- As a result, the number of late-coming parents increased significantly.
- After the fine was removed no reduction occurred.

- We argue that penalties are usually introduced into an incomplete contract, social or private. They may change the information that agents have, and therefore the effect on behavior may be opposite of that expected.
- If this is true, the deterrence hypothesis loses its predictive strength, since the clause “everything else is left unchanged” might be hard to satisfy.

III. Qualitative methods

Qualitative Methods

Qualitative research differ from quantitative research

Multidimensional	Explore different dimensions of social systems
Iterative	Revise your research design according to what you find out in your investigation as it proceeds
Subjective	Researcher brings his or her own mental models which influence the collection and interpretation of these data

Qualitative Methods – Interviews and focus groups

- Oral and personal, non-standardized, using predominantly open questions, and with the interviewer focused on mediation and investigating the response of the interviewee.

- **Focus group** is a particular type of group interview that enables researcher to gather broad data from multiple parties.
- Involves different participants who will provide different perceptions on a topic and engage in reflection on the topic
- Workshop format being more interactive

- **Purposive Sampling:** select the participant to generate the maximum variety in their responses

Qualitative Methods – Example: First Generation Students

Research Project at Frankfurt University of Applied Sciences – Bülbül/Graf/Inowlocki

- Rekonstruktive Analyse der Erfahrungsdarstellung und Bewertung der Studienbedingungen an der Frankfurt UAS durch FGS

- Partizipative Entwicklung von Hypothesen über Zusammenhänge, Wirkungsprozesse und Bedingungen, die fördernd oder störend auf den biographischen Bildungsprozess eingewirkt haben

- Empfehlungen zur „Studienfähigkeit“ der einzelnen Fachbereiche der Frankfurt UAS

Qualitative Methods – Example: First Generation Students

Graduierte FGS im Berufsleben bzw. in Master-Studiengängen an der Frankfurt UAS

Studierende, die erfolgreich durchs Studium gehen	Studierende, die erhebliche Probleme im Studium haben	Studierende, die das Studium abgebrochen haben	Geschwister der FGS an der Frankfurt UAS

FGS der Frankfurt UAS

Alle Fachdisziplinen (FB1, FB2, FB3, FB4)

Qualitative Methods – Example: First Generation Students

Lehrende an Schulen und Hochschulen	Eltern, Geschwister, Verwandte, Peers, Nachbarn, etc.	Berater_innen z.B. Berufsberatung des Arbeitsamts, Studienberatung Frankfurt UAS

Key-Persons mit positivem/negativen entscheidenden Einfluss auf den Bildungsweg der FGS aus ihrer Sicht

Alle Fachdisziplinen (FB1, FB2, FB3, FB4)

Qualitative Methods – Example: First Generation Students

Vorgehensweise

- Durchführung von Gesprächsgruppen (von ca. 2 Stunden Dauer) mit Studierenden der Frankfurt UAS im 3. Semester oder höher
  8 bis 10 Studierende pro Gruppe
- Hohe Diversität der Gruppenzusammenstellung
  Moderation und teilnehmende Beobachtung durch jeweils 2 Studierende/Absolvent_innen des Masterstudiengangs Forschung in der Sozialen Arbeit

- Audioaufnahme der Gespräche
- Transkriptionen der Gespräche
- Beobachtungsprotokolle der Gruppengespräche und Einzelinterviews

IV. Case studies

Case studies

- The case being the focus of the research
- Case study is defined by **what** you collect data about, rather than how you collect your data
- Watch out: it is not teaching case studies!
- Unit of analyses defines your case study: a person, group, organization, project etc.

Case studies

- Case studies often start with a description of the “entity” and its “boundaries”
- Case study to explain, describe, illustrate, explore or evaluate the phenomenon you are interested in
- You can combine both qualitative and quantitative methods (multiple method research)

Case studies – Generalizability

Case studies and its linkage to a body of theory and practice in the literature

- Case studies concentrate on special cases.
- Generalization from case studies must be handled with care.
- To serve as a foundation for generalizations, case studies should be related to a theoretical framework, which in turn may be adjusted as case study results provide new evidence.
- The generalizability of case studies can be increased by strategic selection of critical cases.

Case studies – Number of cases

Focus on one or perhaps two or three examples

- Single case studies are low on generalizability but high on realism

- Multiple cases studies
  - are powerful because of what is called replication , can identify similarities and differences across cases ,
  - develop a more complete picture
  - but your investigation is less depth than a single case study

Case studies – Grounded case study design

Grounded case study research approach to build theories from case study research

1. Getting started – problem definition
2. Selecting cases
3. Crafting instruments and protocols – preparing multiple data collection methods
4. Entering the field – collecting data
5. Analyzing data – within case analysis followed by cross-case analysis
6. Shaping hypotheses – building evidence and explanation
7. Enfolding literature – comparing findings within the literature
8. Research close – knowing when to stop

Case study - Strategies

Replication Strategy

select cases that are similar to each other and look for differences and what causes those differences

Variation Strategy

e.g. the best practice vs worst practice company

Remember … “at the end, you are telling a story”

Case study - Example

Success stories in the airline industry through case studies

Any idea?

Star Alliance passengers at London Heathrow

https://www.youtube.com/watch?v=ZqKzsiV6gbE

Mixed Methodologies

Qualitative and quantitative research approach may be productively combined

V. Data

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Primary data

Data observed or collected directly from first-hand experience.

Secondary data

Published data and the data collected in the past or other parties is called secondary data.

Types of data

Cross-sectional data:

Data is available for many individual units (banks, firms, households, states) but only for one point in time

Time series data:

Data is available for only one individual unit (banks, firms, market index,. . .)
but for many points in time (days, quarters, years, seconds,. . .)

Panel data:

Panel data combine the cross-sectional with the time series dimension.
Thereby you have observations for many individuals over time

Preparing for data analysis

VI. (Statistical) Data Analysis

Statistical Data Analysis slides: Dr. Orcun Kaya

Basic statistical analysis

Descriptive statistics: Summarizing and representing raw data

- Frequency counts
- Measurement of central tendency
- (graphs, etc.)

Bivariate statistics and simple hypothesis testing

- Correlation
- Simple linear regressions
- T-tests and ANOVAs
- Chi-squared test

Interpreting your quantitative results

Advanced statistical analysis

Analyzing multivariate relationships

- Have I included all the right variables?
- Have I included too many variables?
- Are my data appropriate for multivariate analysis?

Brief overview of multivariate data analysis methods

Population versus Sample

Population
The collection of ***all*** outcomes, responses, measurements, or counts that are of interest.

Sample
The collection of data from a subset of the population.

Example:

The *population* is all eligible voters in an election.

A *sample* is any subset of that population. For example, 2000 individuals that reveal their election preferences.

We collect data from a sample to make inferences about the population.

Frequency Distributions

- After constructing a SAMPLE of data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.
- This is the goal of descriptive/graphical statistical techniques.
- One method for simplifying and organizing data is to construct a **frequency distribution**.
- A frequency distribution is an organized tabulation showing exactly how many observations are located in each category on the scale of measurement.
- A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution.

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

USE COMMON SENSE!!!

90% of the information in a sample can be driven by

- A Table or Chart
- Central tendency measures

Frequency Distribution Tables

- A **frequency distribution table** consists of at least two columns - one listing categories on the scale of measurement (X) and another for frequency (f).
- In the X column, values are listed from the highest to lowest, without skipping any.
- For the frequency column, tallies are determined for each value (how often each X value occurs in the data set). These tallies are the frequencies for each X value. The sum of the frequencies should equal N.
- When a frequency distribution table lists all of the individual categories (X values) it is called a regular frequency distribution.

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Frequency Distribution Graphs

- In a **frequency distribution graph**, the score categories (X values) are listed on the X axis and the frequencies are listed on the Y axis.
- When the score categories consist of numerical scores from an interval or ratio scale, the graph should be either a histogram or a polygon.

Discrete Frequency Table: Qualitative data

- Suppose there are 2000 families in a small town and the distribution of children among them is as follows:
  - 300 families have no children
  - 400 families have 1 child
  - 700 families have 2 children
  - 300 families have 3 children
  - 100 families have 4 children
  - 100 families have 5 children
  - 100 families have 6 children

Class	Frequency	Relative Frequency
0 children	300	15%
1 children	400	20%
2 children	700	35%
3 children	300	15%
4 children	100	5%
5 children	100	5%
6 children	100	5%
TOTAL	2000	100%

Discrete Frequency Table: Quantitative data

- Suppose there are 80 different market prices (in USD thousands) in a second hand car shop:
  - 15 up to 18 , 8 cars
  - 18 up to 21 , 23 cars
  - 21 up to 24 , 17 cars
  - 24 up to 27 , 18 cars
  - 27 up to 30 , 8 cars
  - 30 up to 33 , 4cars
  - 33 up to 36 , 2 cars

Class	Frequency	Relative Frequency
15≤ price < 18	8	10%
18≤ price < 21	23	29%
21≤ price < 24	17	21%
24≤ price < 27	18	23%
27≤ price < 30	8	10%
30≤ price < 33	4	5%
33≤ price < 36	2	2%
TOTAL	80	100%

Frequency distribution graph

The four commonly used graphic forms are:

— Pie charts

In a **pie chart** a circular chart is divided into sectors and each sector shows the relative size of each value.

— Histograms

In a **histogram**, a bar is centered above each score (or class interval) so that the height of the bar corresponds to the frequency and the width extends to the real limits, so that adjacent bars touch.

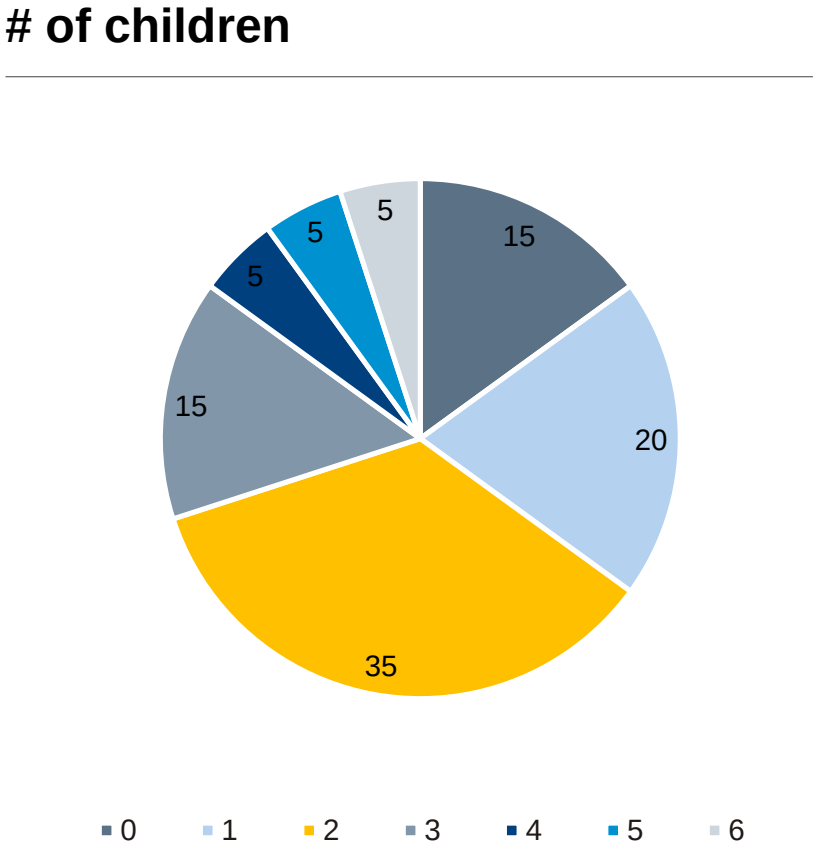
— Frequency polygons

In a **polygon**, a dot is centered above each score so that the height of the dot corresponds to the frequency. The dots are then connected by straight lines. An additional line is drawn at each end to bring the graph back to a zero frequency.

— Cumulative frequency distributions

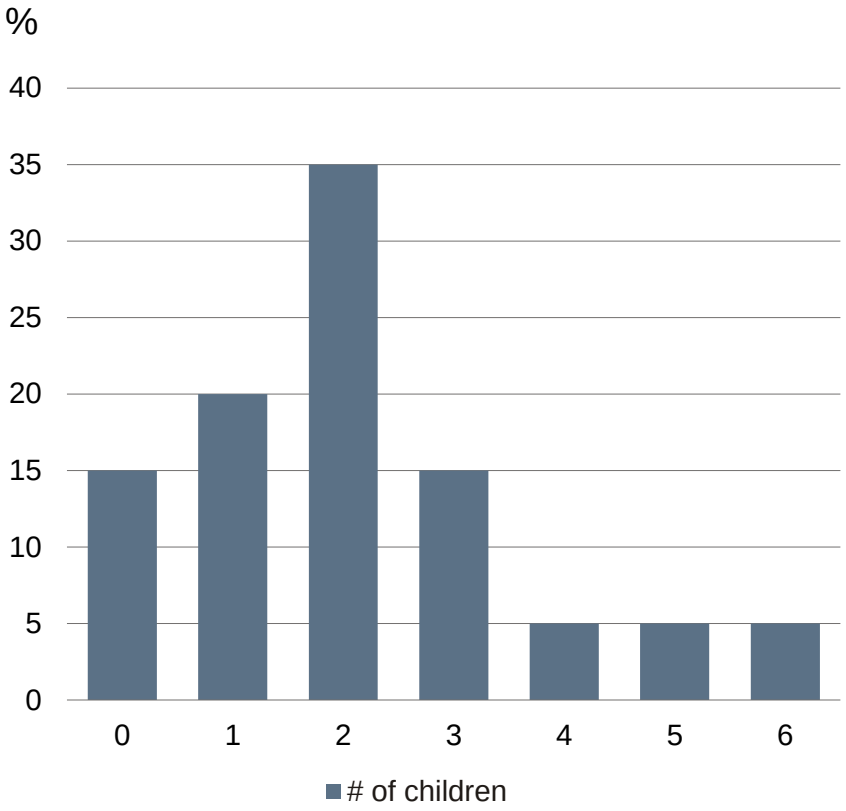
Pie chart

- Used usually to display a set of values each of which is associated with a single category of a factor or ordered factor.
- Most commonly the values are counts or proportions.

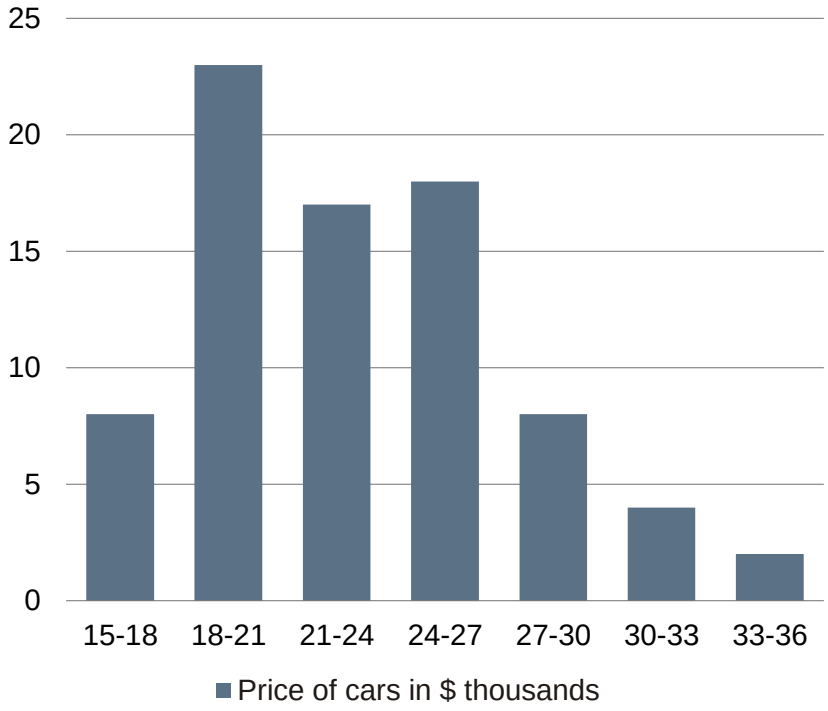


Histogram

Histogram for Qualitative data



Histogram for quantitative data with intervals



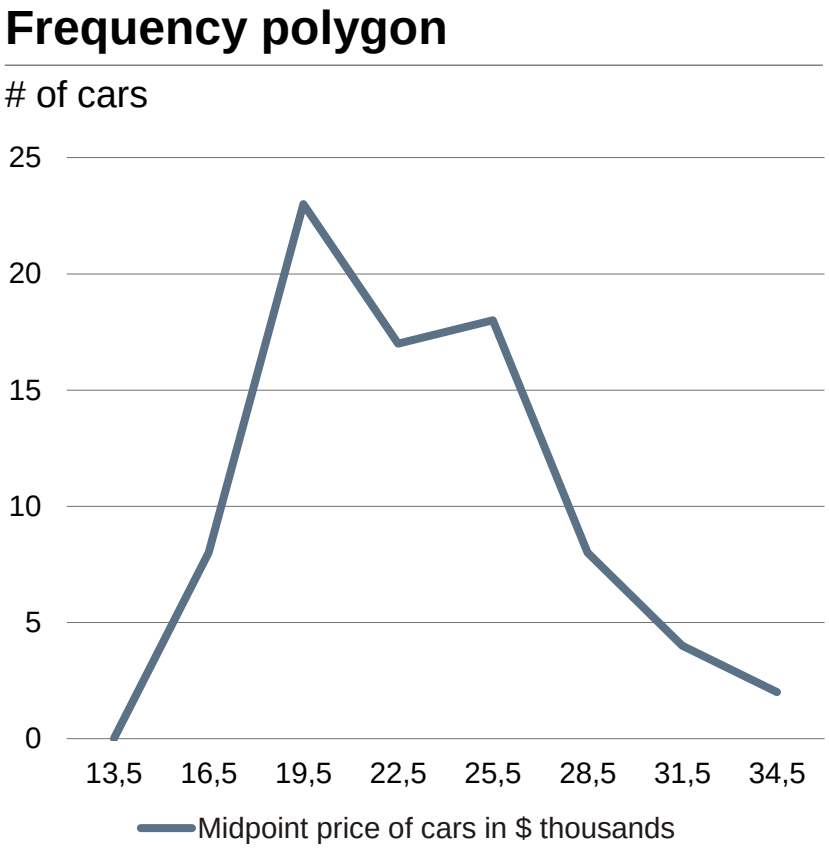
$$Width = \frac{Highest\ value - Lowest\ value}{Number\ of\ intervals}$$

Frequency polygon

- A frequency polygon also shows the shape of a distribution and is similar to a histogram.

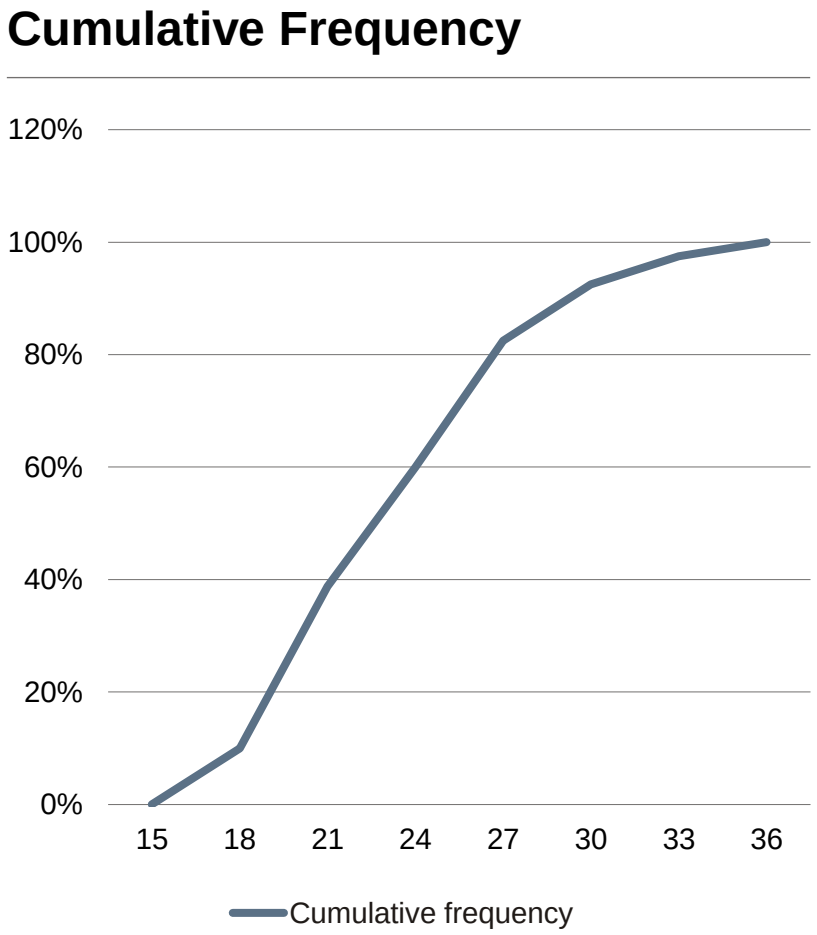
- It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies.

$$\text{Midpoint} = \frac{Higher\ boundary + lower\ boundary}{2}$$



Cumulative Frequency Distribution

Class	Relative Frequency	Cumulative Frequency
15≤ price < 18	10%	10%
18≤ price < 21	29%	39%
21≤ price < 24	21%	60%
24≤ price < 27	23%	83%
27≤ price < 30	10%	93%
30≤ price < 33	5%	98%
33≤ price < 36	2%	100%
TOTAL	100%	



Measures of Central Tendency

- In general terms, **central tendency** is a statistical measure that determines a single value that accurately describes the center of the distribution and represents the entire distribution of scores.
- The goal of central tendency is to identify the single value that is the best representative for the entire set of data
- By identifying the "average score," central tendency allows researchers to summarize or condense a large set of data into a single value.

central tendency serves as a descriptive statistic because it allows researchers to describe or present a set of data in a very simplified, concise form.

it is possible to compare two (or more) sets of data by simply comparing the average score (central tendency) for one set versus the average score for another set.

The Mean, the Median and the Mode

- A single number to serve as a representative value around which all the numbers in the set tend to cluster.
- No single procedure always produces a good, representative value.
- Therefore, researchers have developed three commonly used techniques for measuring central tendency:

1. **the mean (average)**
2. **the median (middle)**
3. **the mode (most)**

The Mean

- The mean is the most commonly used measure of central tendency.
- The **mean** (**arithmetic mean** or **average**) of a set of data is found by adding up all the items and then dividing by the sum of the number of items.
- The mean of a sample is denoted by $\overline{X}$ (read "x bar").
- The mean of n data items $x_1, x_2, \ldots, x_n$ is given by the formula:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$\overline{X} = \frac{\sum x}{n}$$

Mean computation

- Ten students were polled as to the number of siblings in their individual families.
- The raw data is the following set: {3, 2, 2, 1, 3, 6, 3, 3, 4, 2}.
- mean number of siblings for the ten students:

$$\overline{X} = \frac{3+2+2+1+3+6+3+3+4+2}{10} = \frac{29}{10} = 2.9$$

- The **weighted mean** of n numbers $x_1, x_2, \ldots, x_n$, that are weighted by the respective factors $f_1, f_2, \ldots, f_n$ is given by the formula:

$$\overline{w} = \frac{\sum(x \cdot f)}{\sum f}.$$

When the Mean Won't Work

- When a distribution contains a few extreme scores (or is very skewed), the mean will be pulled toward the extremes (displaced toward the tail).

  In this case, the mean will not provide a "central" value.

- With data from a nominal scale it is impossible to compute a mean
- When data are measured on an ordinal scale (ranks), it is usually inappropriate to compute a mean.

Thus, the mean does not always work as a measure of central tendency and it is necessary to have alternative procedures available.

The Median

- If the scores in a distribution are listed in order from smallest to largest, the median is defined as the midpoint of the list.
- The median divides the scores so that 50% of the scores in the distribution have values that are equal to or less than the median.
- Computation of the median requires scores that can be placed in rank order (smallest to largest) and are measured on an ordinal, interval, or ratio scale.

Usually, the median can be found by a simple counting procedure:

1. With an odd number of scores, list the values in order, and the median is the middle score in the list.
2. With an even number of scores, list the values in order, and the median is half-way between the middle two scores.

!!!! One advantage of the median is that it is relatively unaffected by extreme scores.

Example

Ten students in a math class were polled to find out the number of siblings in their individual families and the results were:

3, 2, 2, 1, 1, 6, 3, 3, 4, 2.

Find the median number of siblings

Position of the median: 10/2=5

Between the 5th and 6th values

Data in order: 1, 1, 2, 2, 2, 3, 3, 3, 4, 6

Median=(2+3)/2=2.5 siblings

Nine students in a math class were polled to find out number of siblings in their individual families and the results were:

3, 2, 2, 1, 6, 3, 3, 4, 2.

Find the median number of siblings

Position of the median: 9/2=4.5

the 5th and 6th value

Data in order: 1, 2, 2, 2, 3, 3, 3, 4, 6

Median=3 siblings

Median in a Frequency Distribution

Find the median for the distribution.

Value (x)	**1**	**2**	**3**	**4**	**5**
Frequency (f)	**4**	**3**	**2**	**6**	**8**

Position of the median is the sum of the frequencies divided by 2.

Position of the median $= \frac{\sum f}{2} = \frac{23}{2} = 11.5 = 12th\ term$

Add the frequencies from either side until the sum is 12.

The 12th term is the median and its value is 4.

The mode

- The mode is defined as the most frequently occurring category or score in the distribution.
- In a frequency distribution graph, the mode is the category or score corresponding to the peak or high point of the distribution.
- The mode can be determined for data measured on any scale of measurement: nominal, ordinal, interval, or ratio.

Example:

Ten students in a math class were polled as to the number of siblings in their individual families and the results were: 3, 2, 2, 1, 3, 6, 3, 3, 4, 2.

Find the mode for the number of siblings.

③, 2, 2, 1, ③, 6, ③, ③, 4, 2

The mode for the number of siblings is 3.

Mode in a Frequency Distribution

Find the median for the distribution.

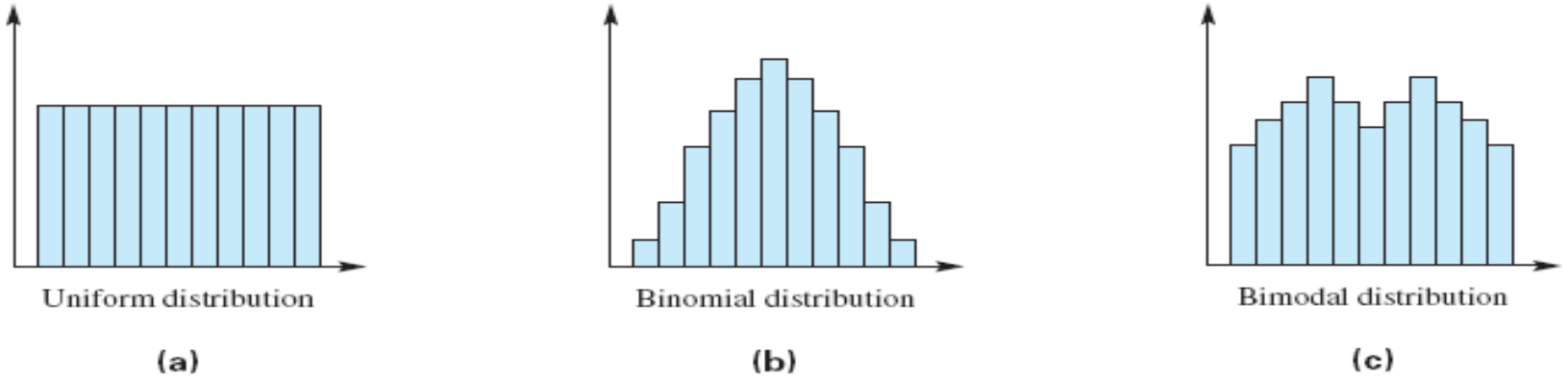
Value (x)	**1**	**2**	**3**	**4**	**5**
Frequency (f)	**4**	**3**	**2**	**6**	**8**

The mode in a frequency distribution is the value that has the largest frequency.

The mode for this frequency distribution is 5 as it occurs eight times.

Symmetry in data sets

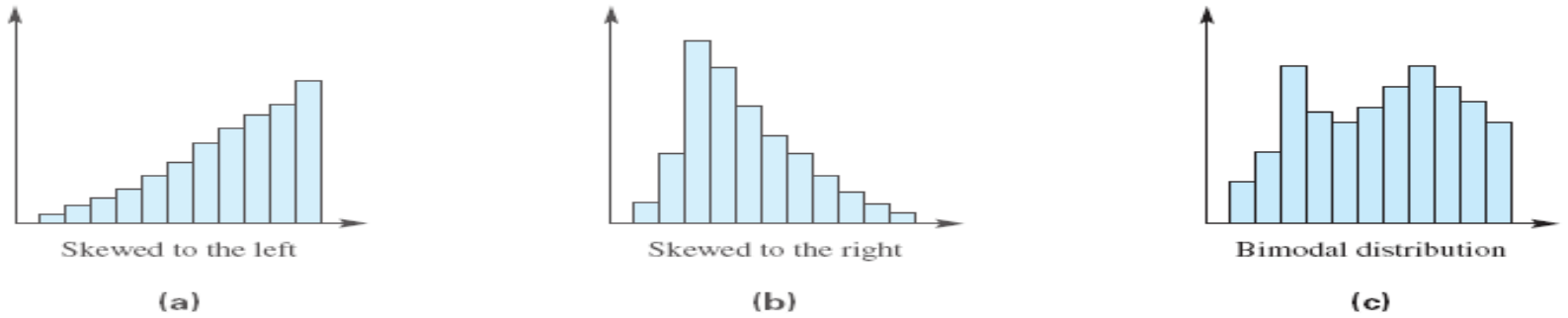
- The analysis of a data set often depends on whether the distribution is **symmetric** or **non-symmetric**.
- A distribution is **symmetrical** if the left side of the graph is (roughly) a mirror image of the right side.
- Because the mean, the median, and the mode are all measuring central tendency, the three measures are often systematically related to each other.
- In a symmetrical distribution, the mean and median will always be equal.



Some symmetric distributions

Non-symmetry in data sets

- **Non-symmetric distribution**: the patterns from a central point from the left and right are different.
- Distributions are **skewed** when scores pile up on one side of the distribution, leaving a "tail" of a few extreme values on the other side.
- In a **positively skewed** distribution, the scores tend to pile up on the left side of the distribution with the tail tapering off to the right.
- In a **negatively skewed** distribution, the scores tend to pile up on the right side and the tail points to the left.

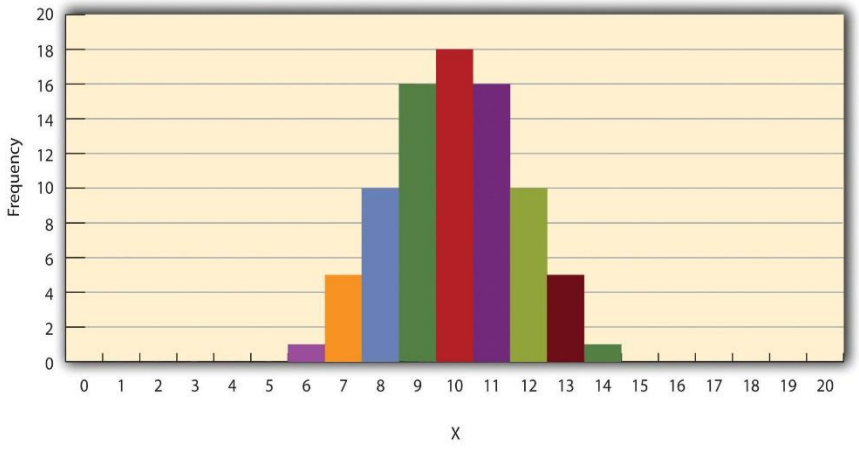


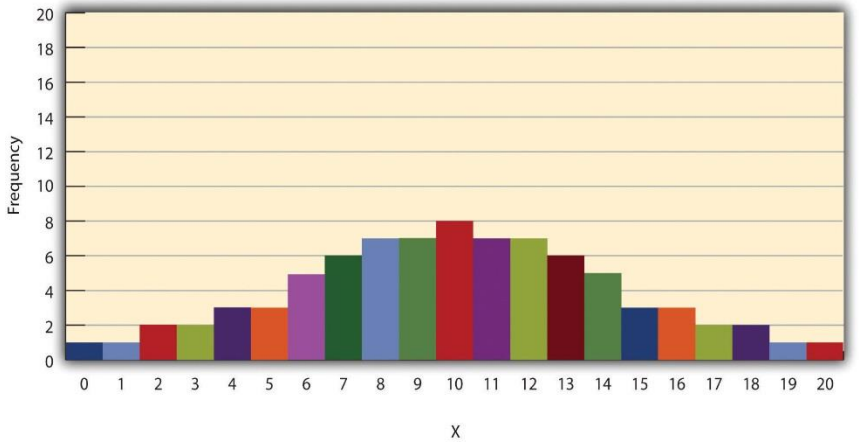
Some non-symmetric distributions

Non-symmetry in data sets (cont.)

- In a skewed distribution, the mode will be located at the peak on one side and the mean usually will be displaced toward the tail on the other side.
- The median is usually located between the mean and the mode.

Variability





- The goal for variability is to obtain a measure of how spread out the scores are in a distribution.
- Central tendency describes the central point of the distribution, and variability describes how the scores are scattered around that central point.
- Together, central tendency and variability are the two primary values that are used to describe a distribution of scores.

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Measuring Variability

Variability can be measured with

- the range
- the interquartile range
- the variance (standard deviation).

The Range

- The **range** is the total distance covered by the distribution, from the highest score to the lowest score (using the upper and lower real limits of the range).
- Easiest measure of variability to calculate

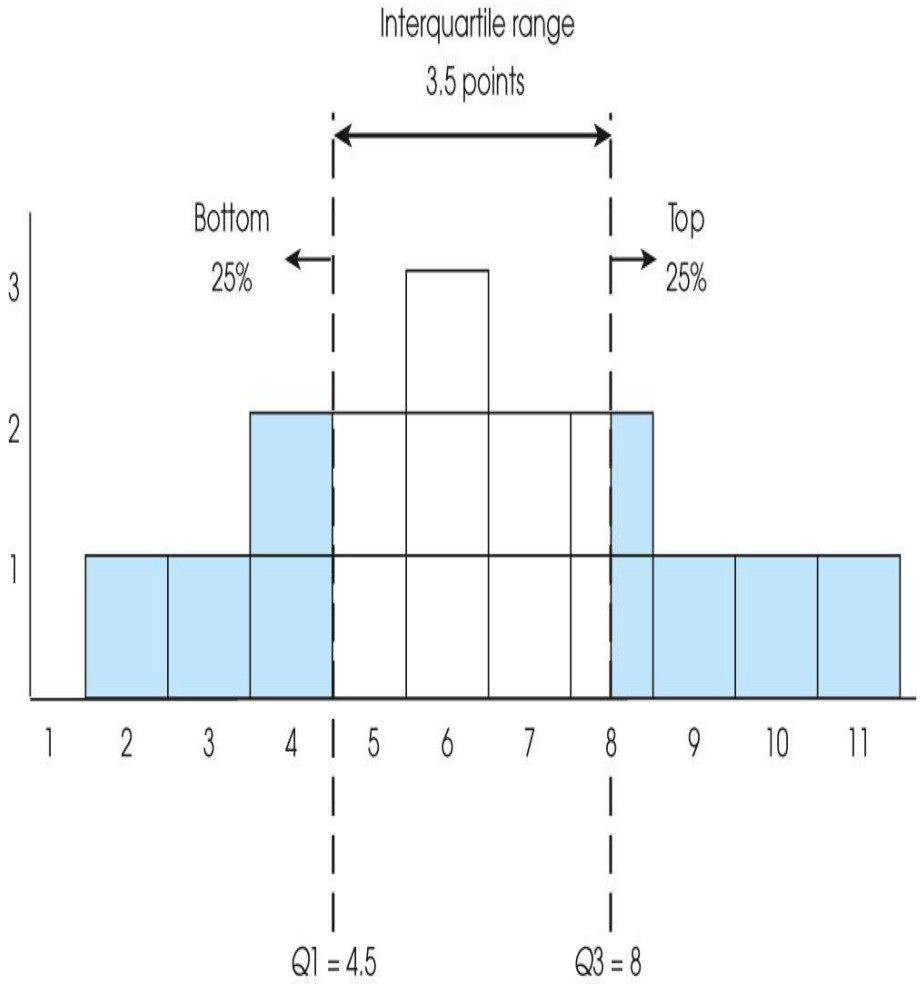
Example: Set of Scores

7, 2, 7, 6, 5, 6, 2

Range= highest score minus lowest score=7-2=5

The Interquartile Range

- The **interquartile range** is the distance covered by the middle 50% of the distribution (the difference between Q1 and Q3).
- The difference between the "**third quartile**" (75th percentile) and the "**first quartile**" (25th percentile).
- IQR = Q3-Q1
- Robust to outliers or extreme observations.
- Works well for skewed data.



The Variance

Variance measures the distance between a score and the mean.

The calculation of variance can be summarized as a three-step process:

1. Find difference between each data point and mean.
2. Square the differences, and add them up.
3. Divide by one less than the number of data points

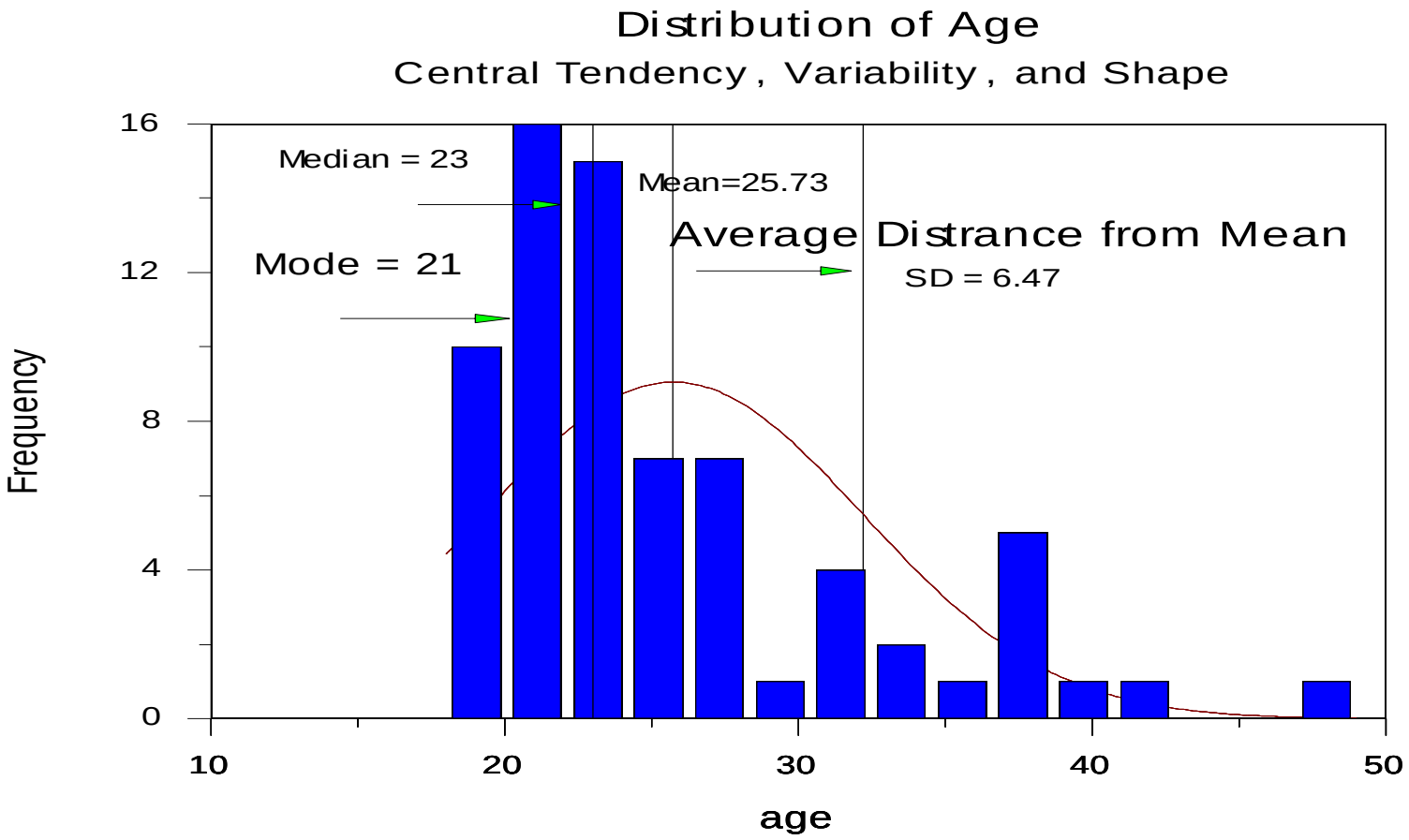
$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Sample standard deviation is square root of sample variance, and so is denoted by S.

Computing the Variance

(N=5) X	$\overline{X}$	$X - \overline{X}$	$(X - \overline{X})^2$
5	15	-10	100
10	15	-5	25
15	15	0	0
20	15	5	25
25	15	10	100
Total:	75	0	250
Mean:	Variance	Is →	62.5

Example


Distribution of Age
Central Tendency, Variability, and Shape
Median = 23
Mean=25.73
Mode = 21
Average Distrance from Mean
SD = 6.47
Frequency
16
12
8
4
0
10
20
30
40
50
age

How to make inferences about population mean?

- The t statistic allows researchers to use sample data to test hypotheses about an unknown population mean.
- The t-statistic can be used to test hypotheses about a *completely unknown* population; that is, both Population mean and σ (population standard deviation) are unknown
- The only available information about the population comes from the sample.

Significance Test for Mean

- **Null Hypothesis**: H_0: $\mu = \mu_0$ where μ_0 is particular value for population mean
- ***Alternative Hypothesis***: H_a: $\mu \neq \mu_0$
- *2-sided* alternative includes both > and < H_0 value
- *T-Test Statistic*: The number of standard errors that the sample mean falls from the H_0 value

$$t = \frac{\bar{y} - \mu_0}{se} \quad \text{where} \quad se = s / \sqrt{n}$$

Significance Test for Mean (cont.)

- When H_0 is true, the sampling distribution of the *t* test statistic is the *t* distribution with *df = n - 1.*

- *P-value*: Under presumption that H_0 true, probability the *t* test statistic equals observed value or even more extreme (i.e., larger in absolute value), providing stronger evidence against H_0

- *Conclusion*: Report and interpret *P*-value. If needed, make decision about H_0

Example

Year-end Portfolio returns (in %) for a sample of n=17 households

y = 11.4, 11.0, 5.5, 9.4, 13.6, -2.9, -0.1, 7.4, 21.5, -5.3,-3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7

Let μ = population mean weight change

Test H_0: μ = 0 (zero return) against H_a: μ ≠ 0.

Data have

Variable	N	Mean	Std.Dev.	Std. Error Mean
weight_change	17	7.265	7.157	1.736

Remember: $se = s/\sqrt{n} = 7.157/\sqrt{17} = 1.736$

Example cont.

Test Statistic *(df = 16):*

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{7.265 - 0}{1.736} = 4.18$$

***P*-Value**: $P = 2P(t > 4.18) = 0.0007$ (from software)

Interpretation: If H_0 were true, prob. would = 0.0007 of getting sample mean at least 4.18 standard errors from null value of 0.

Conclusion: Very strong evidence that the household portfolio returns differ from 0.

2 sample case

- How to compare the **mean** between 2 samples?

- if 2 samples are taken from the same population, then they should have fairly similar means

⇨ if **2 means are statistically different**, then the samples are likely to be drawn from 2 different populations

T-test for 2 independent samples

Difference between the means divided by the pooled standard error of the mean

Null hypothesis:

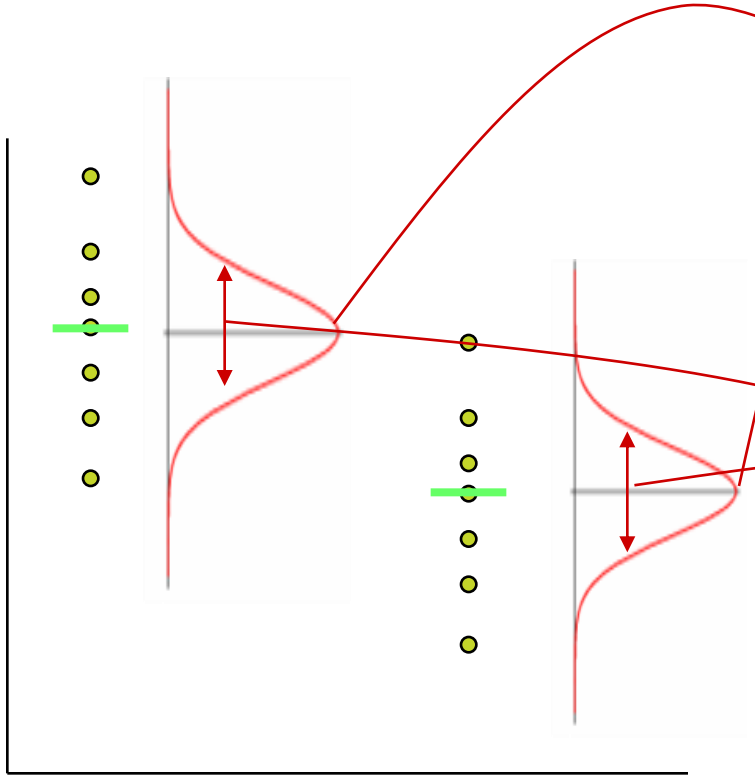
$$H_0: \overline{X_1} - \overline{X_2} = 0$$

Usually zero

Test statistic value:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{s_{\bar{x}_1 - \bar{x}_2}}$$

Formula cont.



$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

Weekly expenses of households:

First group: $\overline{X_1} = USD\ 191, s_1 = USD\ 38, n_1 = 8$

Second group: $\overline{X_2} = USD\ 199, s_2 = USD\ 12, n_2 = 10$

$$H_0: \overline{X_1} - \overline{X_2} = 0$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{191 - 199_2}{\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}} \approx .57$$

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Comparison of more than 2 samples: ANOVA

- ANalysis Of VAriance (ANOVA)
- Compares the differences in means between groups but it uses the variance of data to “decide” if means are different
- ANOVA can tell you if there is an effect but not where

Hypotheses of One-Way ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

— All population means are equal

i.e., no treatment effect (no variation in means among groups)

$$H_1 : \text{Not all of the population means are the same}$$

— At least one population mean is different i.e., there is a treatment effect

— Does not mean that all population means are different (some pairs may be the same)

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

ANOVA: a somewhat example

The statistics classroom is divided into three rows: front, middle, and back with exam grades below (left)

Data Summary Statistics			
Row	**Front**	**Middle**	**Back**
Sample Size	**7**	**9**	**8**
Mean	**75.71**	**67.11**	**53.50**
St. Dev	**17.63**	**10.95**	**8.96**
Variance	**310.90**	**119.86**	**80.29**

ANOVA TABLE					
Source	**SS**	**df**	**MS**	**F**	**P**
Between Variation	**1902**	**2**	**951.0**	**5.9**	**0.009**
Within Variation	**3386**	**21**	**161.2**		
Total Variation	**5288**	**23**	**229.9**		

Moving on…

How much linear is the relationship of two variables?
(descriptive)

CORRELATION

How good is a linear model to explain my data?

(inferential)

REGRESSION

Are two variables related?

How much depend the value of one variable on the value of the other one?

- Does one increase as the other increases?
  e. g. skills and income

- Does one decrease as the other increases?
  e. g. health problems and nutrition

- How can we get a numerical measure of the degree of relationship?

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Example: Smoking and Lung Capacity

Investigate relationship between *cigarette smoking* and *lung capacity*

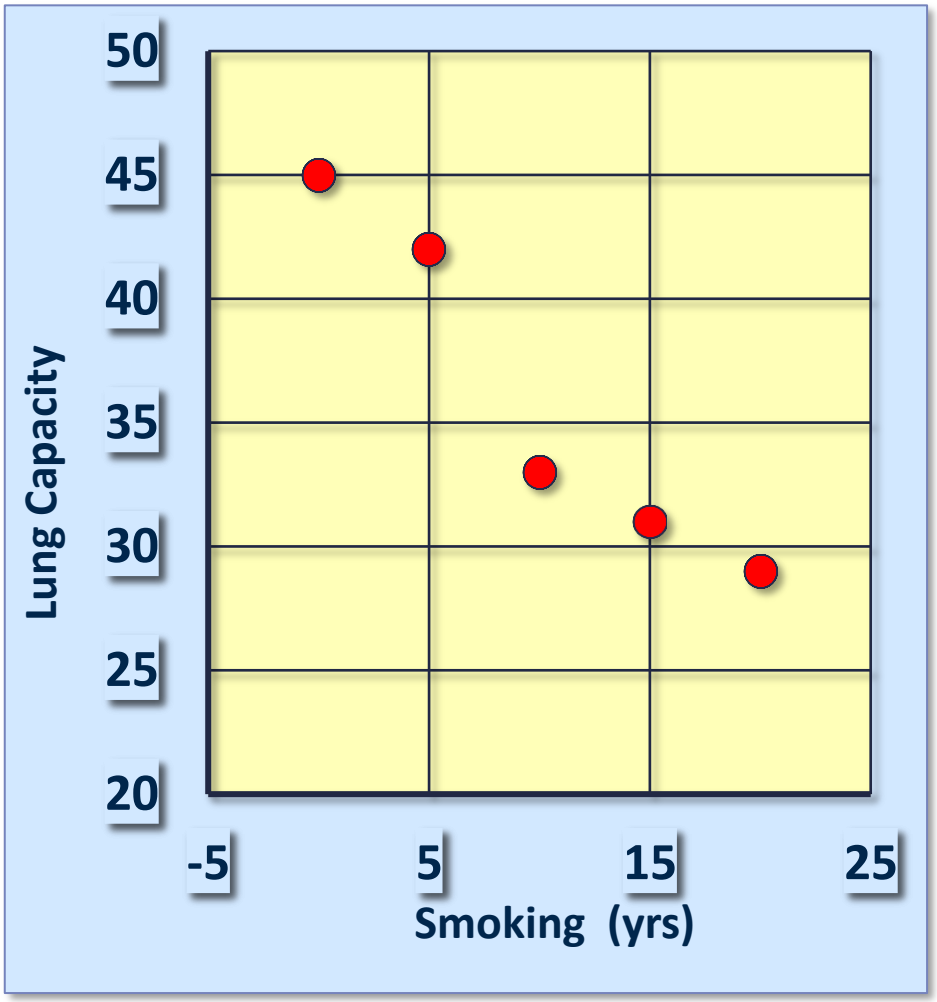
Data: sample group response data on smoking habits, *and* measured lung capacities, respectively

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

Smoking and Lung Capacity

Observe that as smoking exposure goes up, corresponding lung capacity goes down

Variables *covary* inversely!!!



The Sample Covariance

The covariance is a statistic representing the degree to which 2 variables vary together

cov(x,y) = mean of products of each point deviation from mean values

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Calculating Covariance

Cigs (X)	Lung Cap (Y)
0	45
5	42
10	33
15	31
20	29
$\bar{X} = 10$	$\bar{Y} = 36$

Calculating Covariance

Cigs (X)	$(X-\bar{X})$	$(X-\bar{X})(Y-\bar{Y})$	$(Y-\bar{Y})$	Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		Σ= -215		

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Pearson correlation coefficient (r)

The value obtained by covariance is dependent on the size of the data's standard deviations: if large, the value will be greater than if small

To overcome this problem use Pearson correlation coefficient

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y} \qquad \text{(S = st dev of sample)}$$

- r is normalised (dimensionless) covariance

Correlation

- Positive correlation
  - High values of X tend to be associated with high values of Y.
  - As X increases, Y increases

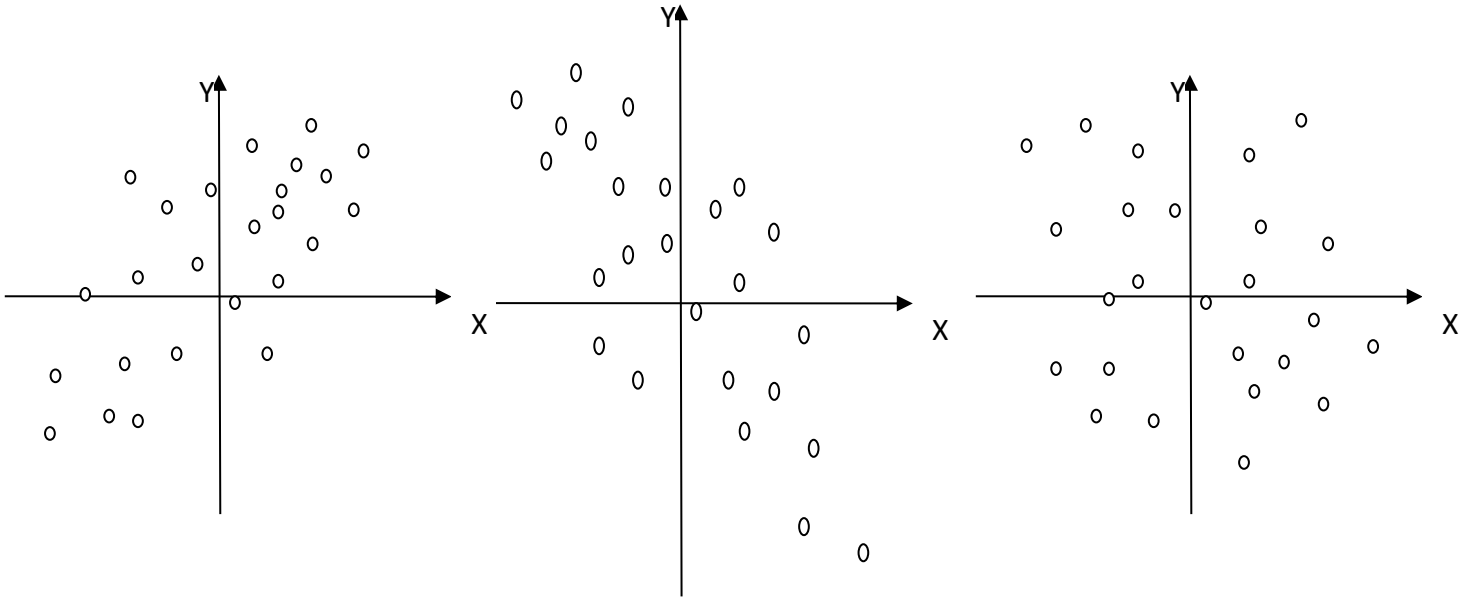
- Negative correlation
  - High values of X tend to be associated with low values of Y.
  - As X increases, Y decreases

- No correlation
  - No consistent tendency for values on Y to increase or decrease as X increases

- r takes values from -1 (perfect negative correlation) to 1 (perfect positive correlation). r=0 means no correlation

- Sign refers to direction.

sign of covariance = sign of correlation



Positive correlation: cov > 0

Negative correlation: cov < 0

No correlation. cov ≈ 0

Regression

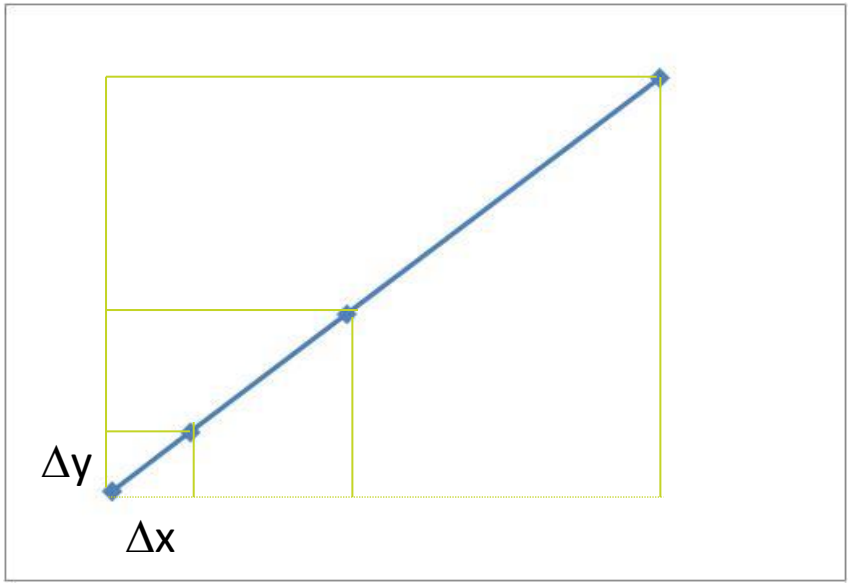
- Linear Regression: Prediction of one variable from knowledge of one or more other variables

- How good is a linear model ($y=ax+b$) to explain the relationship of two variables?

- If there is such a relationship, we can ‘predict’ the value y for a given x.

Lineal dependence between 2 variables

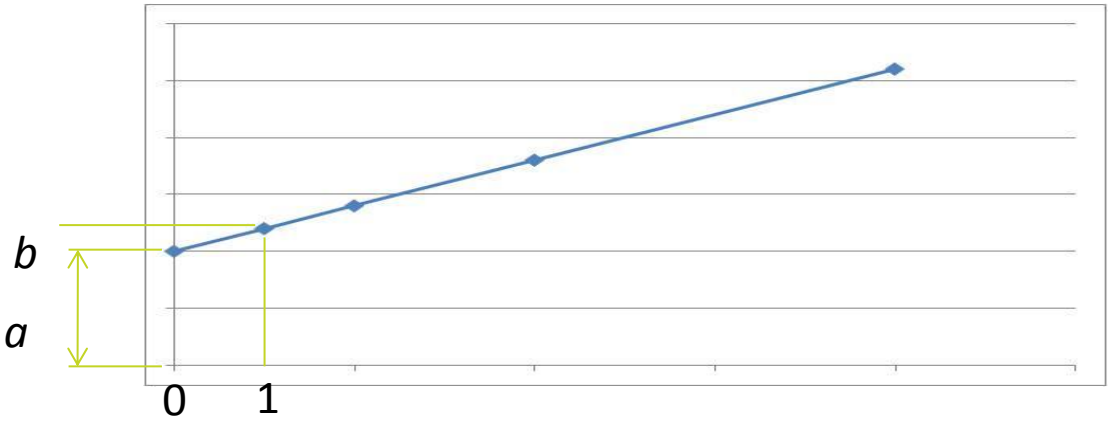
Two variables are linearly dependent when the increase of one variable is proportional to the increase of the other one



Examples: Energy needed to boil water, Money needed to buy goods

Lineal dependence (Cont.)

- The equation *y= mx+n* that connects both variables has two parameters:
- '*m*' is the unitary increase/decerease of y (how much increases or decreases y when x increases one unity)
- '*n*' the value of y when x is zero (usually zero)



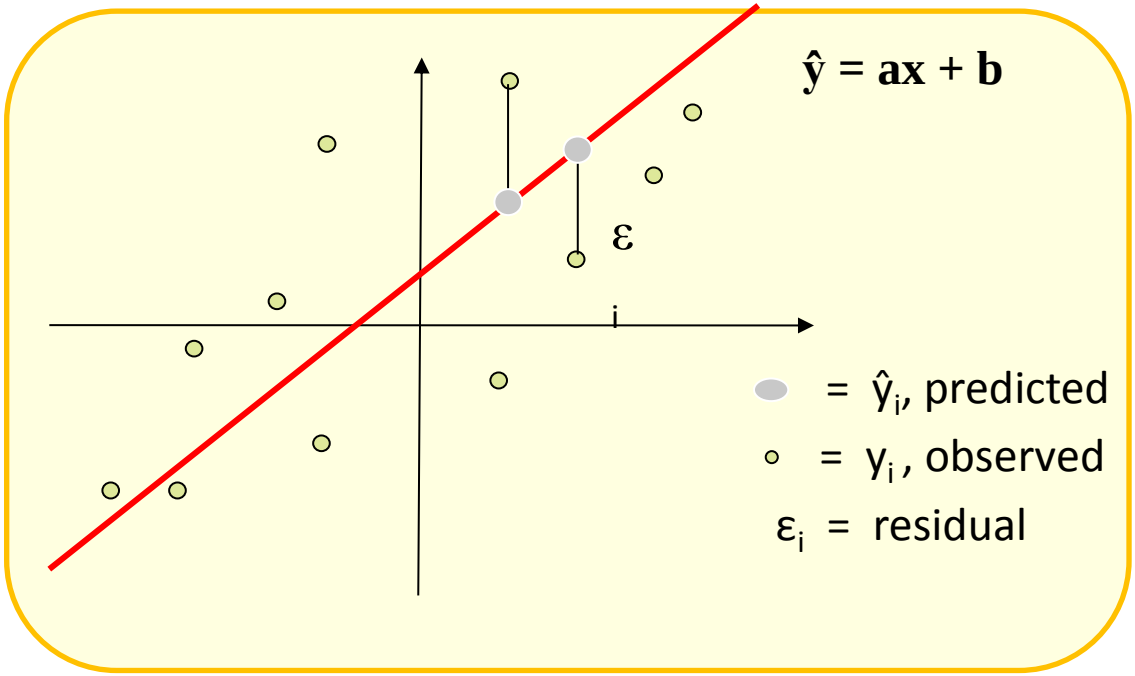
Examples: '*m*'= Energy needed to boil one litter of water , '*n*'=0

'*m*' = prize of an apple, '*n*'= fixed tax/commission to add

Fitting data to a straight line

Here, $\hat{y} = ax + b$

- $\hat{y}$: predicted value of y
- a: slope of regression line
- b: intercept



- Residual error (ε_i): Difference between obtained and predicted values of y (i.e. $y_i - \hat{y}_i$)
- Best fit line (values of *b* and *a*) is the one that minimises the sum of squared errors (SS_{error}) $\Sigma(y_i - \hat{y}_i)^2$

Fitting data to a straight line (Cont.)

- Minimise $\Sigma(y_i - \hat{y}_i)^2$, which is $\Sigma(y_i - ax_i + b)^2$
- Minimum SS_{error} is at the bottom of the curve where the gradient is zero – and this can found with calculus
- Take partial derivatives of $\Sigma(y_i - ax_i - b)^2$ respect parameters *a* and *b* and solve for 0 as simultaneous equations, giving:

$$a = \frac{rs_y}{s_x} \qquad b = y - ax$$

- This calculus can always be done, whatever is the data!!

How good is the model?

- We can calculate the regression line for any data, but how well does it fit the data?
- Total variance = predicted variance + error variance: $S_y^2 = S_{\hat{y}}^2 + S_{er}^2$
- Also, it can be shown that r^2 is the proportion of the variance in *y* that is explained by our regression model
- $r^2 = S_{\hat{y}}^2 / S_y^2$
- Insert $r^2 S_y^2$ into $S_y^2 = S_{\hat{y}}^2 + S_{er}^2$ and rearrange to get: $\mathbf{S_{er}^2 = S_y^2 (1 - r^2)}$
- From this we can see that **the greater the correlation the smaller the error variance, so the better our prediction**

Is the model significant?

i.e. do we get a significantly better prediction of *y* from our regression equation than by just predicting the mean?

F-statistic:

$$F_{(df_{\hat{y}}, df_{er})} = \frac{s_{\hat{y}}^2}{s_{er}^2} \quad \overset{\text{complicated rearranging}}{=\ldots\ldots=} \quad \frac{r^2 (n-2)^2}{1 - r^2}$$

And it follows that:

$$t_{(n-2)} = \frac{r (n-2)}{\sqrt{1 - r^2}}$$

So all we need to know are *r* and *n* !!!

Generalization to multiple variables

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc., on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:
  - $y = b_0 + b_1x_1 + b_2x_2 + \ldots.. + b_nx_n + \varepsilon$
- The a parameters reflect the independent contribution of each independent variable, x , to the value of the dependent variable, y
- i.e. the amount of variance in y that is accounted for by each x variable after all the other x variables have been accounted for

Last remarks:

- **Correlated doesn't mean related.**

e.g, any two variables increasing or decreasing over time would show a nice correlation: $C0_2$ air concentration in Antartica and lodging rental cost in London. **Beware in longitudinal studies!!!**

- **Relationship between two variables doesn't mean causality**

(e.g leaves on the forest floor and hours of sun)

- **Cov(x,y)=0 doesn't mean x,y being independents**

(yes for linear relationship but it could be quadratic,…)

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

MBA Aviation and Tourism Management

Faculty of Business and Law
Frankfurt University of Applied Sciences

September 2018