Bus 701: Advanced Statistics

Harald Schmidbauer

About this course.

Goal 1: Go through modern statistical terminology, concepts, and methodology.

- Based on slides.
- Four parts:
  - Part I: Introduction to Statistics; Basic Concepts; Descriptive Statistics
  - Part II: Probability and Stochastic Models
  - Part III: Statistical Inference
  - Part IV: Further Topics

About this course.

Goal 2: Learn about the following topics:

- Simulation of social systems. — Mainly based on:

  Gilbert, N. & Troitzsch, K.: *Simulation for the Social Scientist.* Mcgraw-Hill, 2005.

- Interpreting probability. — Mainly based on:

  Howie, D.: *Interpreting Probability. Controversies and Developments in the Early Twentieth Century.* Cambridge University Press, 2004.

About this course.

Goal 3: Learn how to use a powerful statistical software package.

- We recommend: R. Please visit:

  www.R-project.org

- R is a language and environment for statistical computing and graphics.

- R is a GNU project with contributors from all over the world.

About this course.

Your contribution to Bus 701; grading.

Your contribution:

- presentation of a topic (according to Goal 2 above)
- participation in a statistical research project

Grading:

presence (including presentation):	50%
research project:	50%

About using statistical methodology.

Advice.

- **If you have a statistics problem in your own research, please consult a statistician.**
- **Begin consulting the statistician before you start collecting your data set.**

About These Slides

- The present slides are not self-contained; they need to be explained and discussed.

- Even though being a "work in progress" and subject to revision, the slides constitute copyrighted material.
  If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer	**harald** at **hs-stat** dot **com**
Angi Rösch	**angi** at **angi-stat** dot **com**

- The slides were produced using LaTeX and R (the R project; website: www.R-project.org) on a GNU/Linux system.

- R files used for this course are available upon request.

PART I:

Introduction to Statistics;

Basic Concepts;

Descriptive Statistics

Chapter 1:

Introduction

1.1 Some Examples

1. Scratching down numbers.

Students in a first course of statistics were asked to enter their gender and body-height into a list. The result was:

(m,167)(m,181)(m,178)(m,180) (f,160) (f,174) (f,170) (f,173) (f,168)
(f,165) (m,195)(m,185)(m,183)(m,192)(m,195)(m,188) (f,168) (f,160)
(f,169) (f,175) (f,157) (f,162) (f,172) (f,160) (m,173) (f,157) (m,170)
(f,170) (f,168) (m,182)(m,175) (f,165) (m,185) (f,170) (m,182)(m,190)
(m,188) (f,160) (m,188)(m,182) (f,177) (f,165) (f,163) (f,160) (f,160)
(f,170) (m,175)(m,176)(m,183) (f,170) (f,170) (f,169) (m,181)(m,177)

Here, f=female and m=male, and the body-height is in centimeters. What can you tell from these data?

1.1 Some Examples

female:

15	77
16	0000002355588899
17	00000023457

male:

16	7
17	0355678
18	0112223355888
19	0255

Here, **15** | 7 = 157 cm. Such a diagram is called a **stem-and-leaf display**, or simply **stemplot**

1.1 Some Examples

2. Averages. Averages are very important. They beware us from getting lost in information. But we have to be careful with averages.

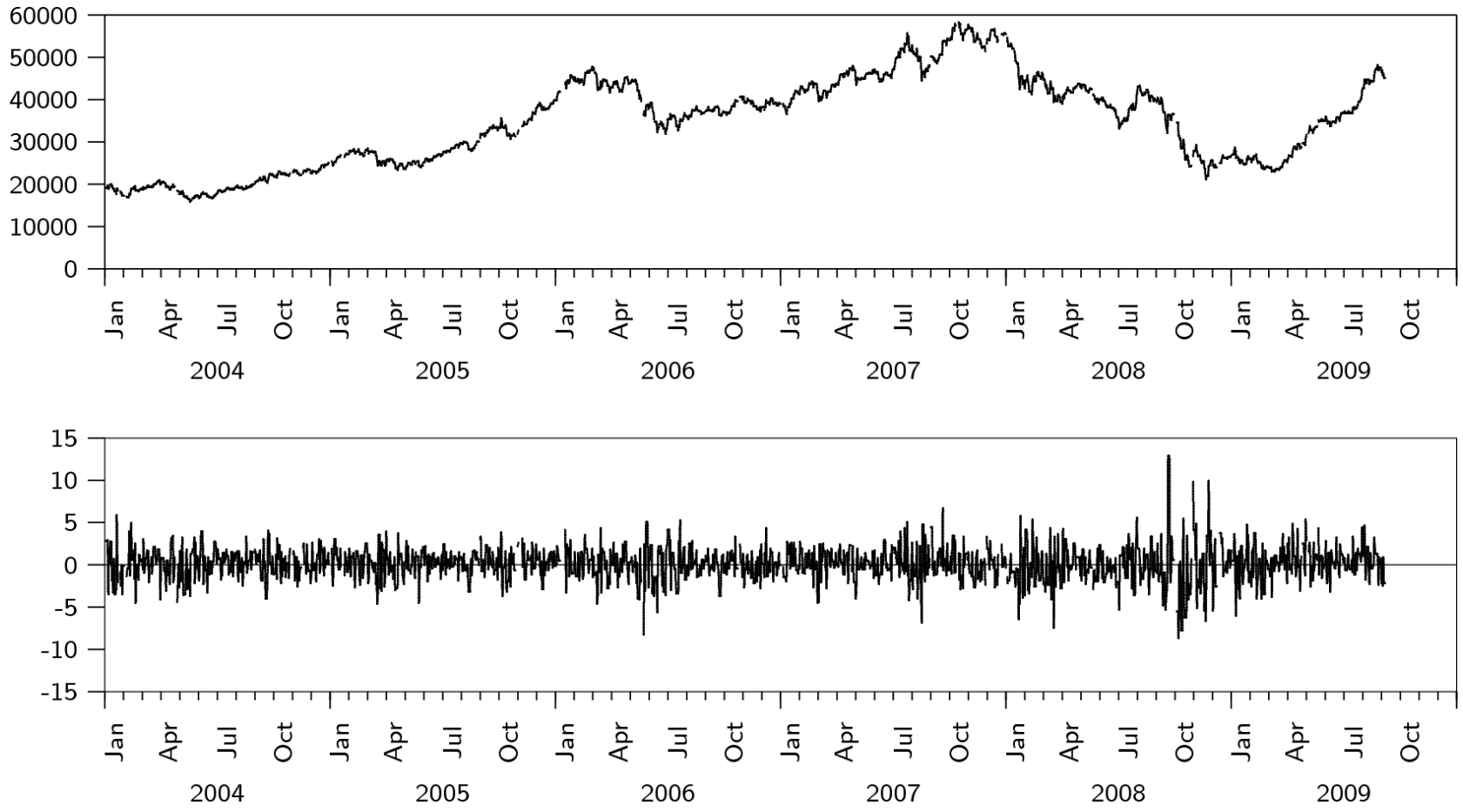
Consider a share of stock. Its price. . .

. . . 2 years ago:	$ 100
. . . 1 year ago:	$ 150
. . . today:	$ 90

What is the average annual gain or loss of the stock in percent?

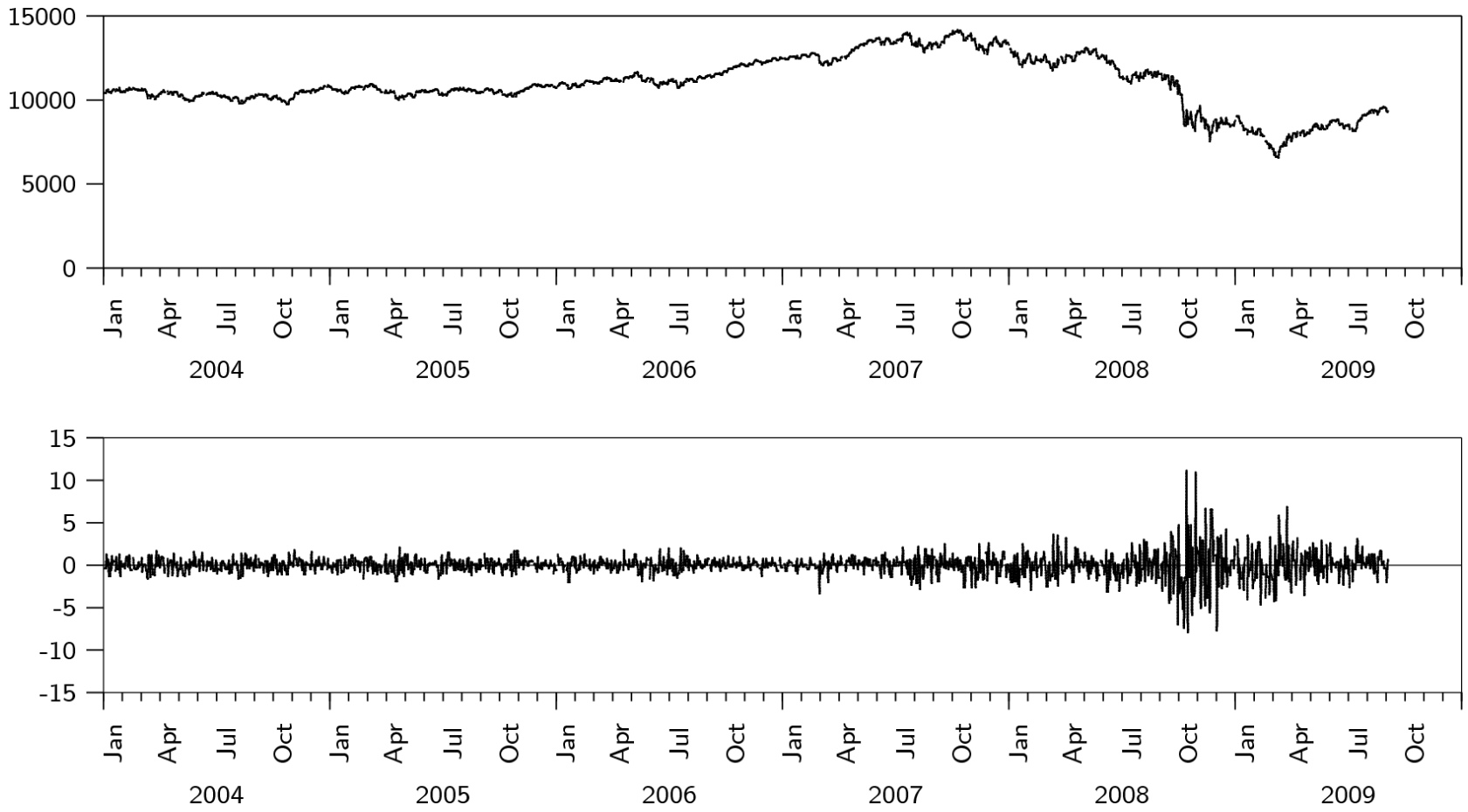
1.1 Some Examples

3. A stock index and its daily returns — İMKB 100.



1.1 Some Examples

3. A stock index and its daily returns — Dow-Jones.



1.1 Some Examples

4. Credit card management: Is a person credit-worthy?

A bank cannot see the person in detail. . . But there are clues:

- age, education, professional environment
- previous payment behaviour
- stability of residential area
- number of cellular phone contracts

How can we exploit these clues? Data protection issues?!?

1.1 Some Examples

5. Website visitor satisfaction.

- Customers making bad experience with website handling won't visit the website again.
- Common complaints: slow response time; difficult navigation
  (response time: the time it takes to answer a query)
- Effect on Google's website score and thus on ad fees!
- Quality management requirements:
  - monitor website response times
  - sample customers' click streams
- Data needs to be analyzed, conclusions drawn.

1.1 Some Examples

6. Television audience rating.

- Ayşe hanım is the program manager of a television channel.
- Her goal is that, in the future, the rating of "Çiçek Taksi" should be at least 10%.
- One evening, 350 televisions in 4000 randomly selected households were tuned into this program.
- Can we conclude that Ayşe hanım has not reached her goal?

1.2 Statistics as a Science

The term "Statistics".

The word "statistics" can refer to:

- a science ("statistics" is singular)
- results of this science ("statistics" is plural)

1.2 Statistics as a Science

Statistics is the science of reasoning with numbers.

Statistics is concerned with

- detecting the structure in data sets.
- facilitating the communication between people.
- making well-founded decisions.
- forecasting the future.
- determining what (numerical) information is needed to solve a given problem.

1.2 Statistics as a Science

On being misled by numbers.

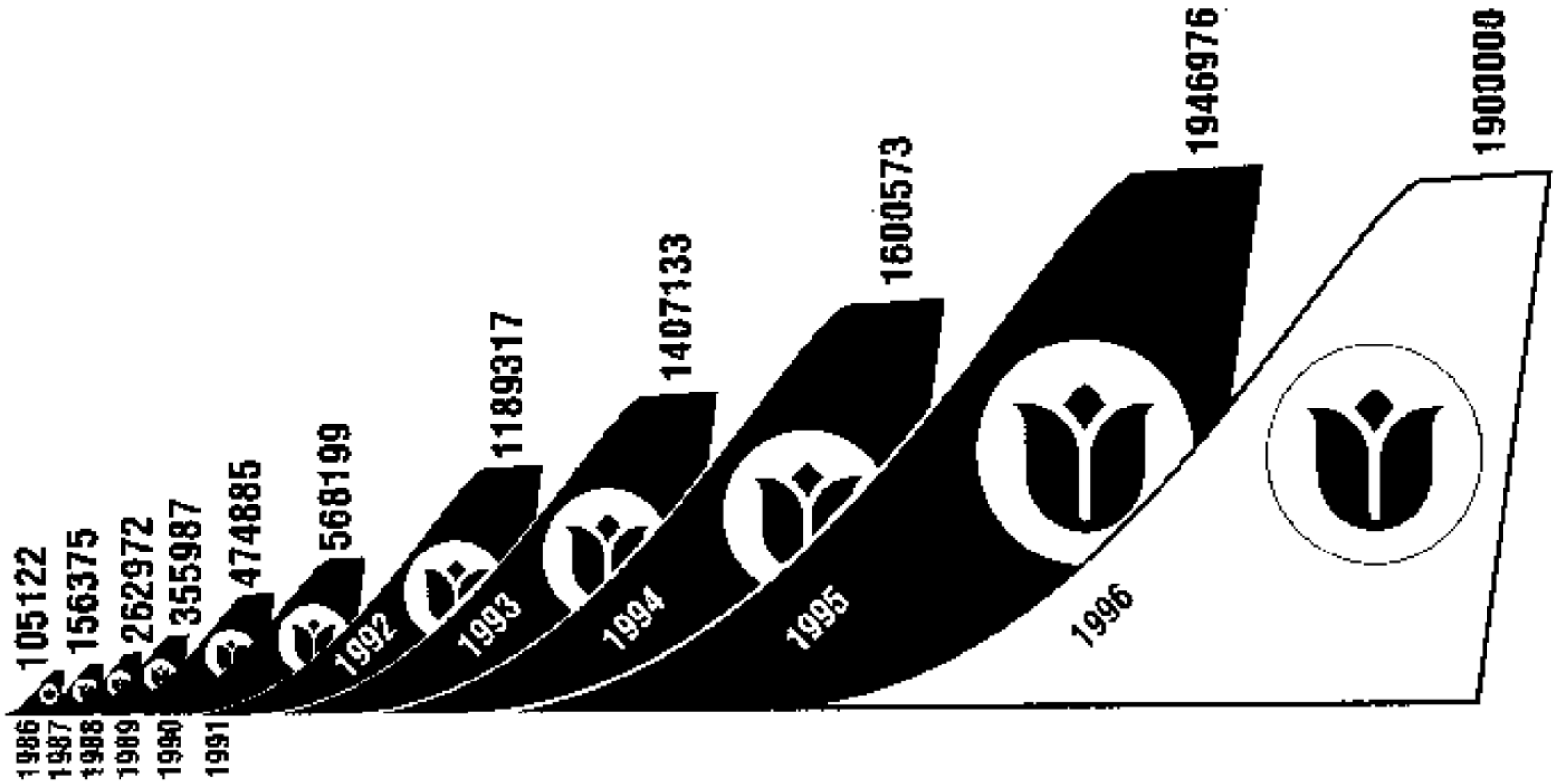
İstanbul'un nüfusu (2000):

$$\boxed{9\,822\,210}$$

(Source: Türkiye İstatistik Yıllığı 2005, TÜİK)

1.2 Statistics as a Science

Images can also be misleading:



1.2 Statistics as a Science

Statistics and the Computer.

- Although some techniques can be done using paper and pencil, statistics is a hi-tech science: It needs powerful software to be effective.
- We recommend: R. Please visit:

www.R-project.org

- R is a language and environment for statistical computing and graphics. It is a GNU project with contributors from all over the world.

1.3 Descriptive and Inductive Statistics

The goals of descriptive and inductive statistics.

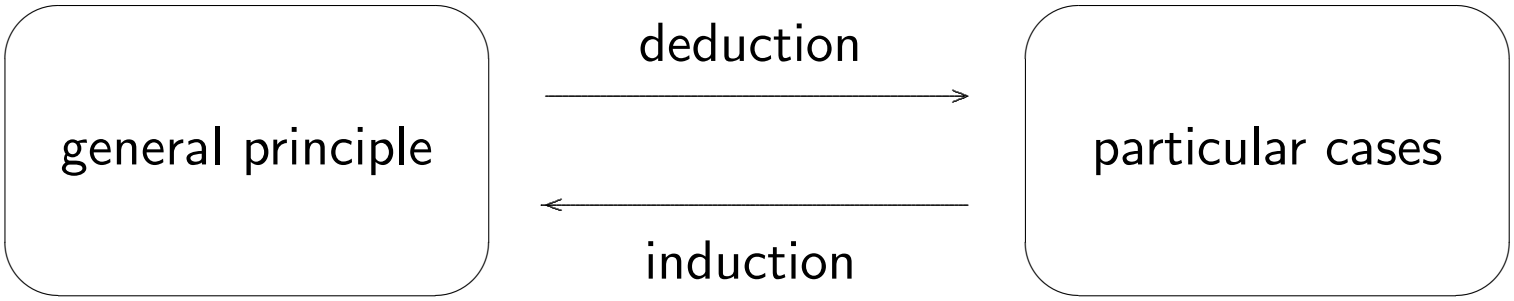
The goal of. . .

- . . . descriptive statistics: Describe, summarize, display given data (data reduction!).
- . . . inductive statistics: Draw conclusions from data (observations) to more general principles.
  This process is called statistical inference.

1.3 Descriptive and Inductive Statistics

Conclusions.

There are two kinds of conclusions:



In the context of inductive statistics:

- The "particular cases" are observed data (sample data).
- The "general principle" is a probability distribution, characterizing the entire population.

1.4 Some Historical Remarks

1. The Origins

- 2600 BC: construction of pyramids in Egypt; censuses to determine the power of the state
- population censuses for recruitment and taxation
- USA: population census to represent the states according to population (1787 Constitution)

1.4 Some Historical Remarks

2. "University Statistics"

- Hermann Conring (1606–1681): systematic description of state affairs

- Gottfried Achenwall (1719–1772):
  "Staatsmerkwürdigkeiten"
  ("phenomena of particular interest of a country or a people")
  $\longrightarrow$ "Statistik"

1.4 Some Historical Remarks

3. Political Arithmetics

- John Graunt (1620–1674):
  "Bills of Mortality of the City of London" (1662)

- William Petty (1623–1687):
  "Essays in Political Arithmetics" (1672)

- Thomas Robert Malthus (1766–1834): "An Essay on the Principle of Population, as it Affects the Future Improvement of Society" (1798)

1.4 Some Historical Remarks

4. Games of Chance

- Gerolamo Cardano (1501–1576):
  "De ludo aleae" (1560s; published 1663)
- Antoine Chevalier de Méré (1607–1685):
  questions concerning gambling
- Blaise Pascal (1623–1662)
- Pierre de Fermat (1601–1665)
- Christiaan Huygens (1629–1695)

1.4 Some Historical Remarks

5. Probability

- Jacob Bernoulli (1655–1705):
  "Ars Conjectandi" (published 1713)
- Abraham de Moivre (1667-1754):
  "The Doctrine of Chances" (1718)
- Pierre-Simon Laplace (1749–1827)
- Thomas Bayes (1702–1761): "Essay Towards Solving a Problem in the Doctrine of Chances" (1764)

1.4 Some Historical Remarks

6. Discovery of Statistical Regularity in Society

- Adolphe Quetelet (1796–1874):
  "l'homme moyen" (the average man) (1835)
- Henry Thomas Buckle (1821–1862)
- John Stuart Mill (1806–1873)
- Karl Marx (1818–1883)
- Émile Durkheim (1858–1917)

1.4 Some Historical Remarks

7. The "Statistical Revolution"

- Charles Darwin (1809–1882): "On the Origin of Species by Means of Natural Selection" (1859)
- Francis Galton (1822–1911)
- Karl Pearson (1857–1936)
- James Clerk Maxwell (1831–1879)
- Ludwig Boltzmann (1844–1906)
- Charles Sanders Peirce (1839–1914)

1.4 Some Historical Remarks

8. The Beginning of Statistical Inference

- Ronald Aylmer Fisher (1890–1961)
- Jerzy Neyman (1894–1981)
- Egon S. Pearson (1895–1980)