Bus 701: Advanced Statistics

Harald Schmidbauer

About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.

- 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 (www.latex-project.org) and R (www.R-project.org) on a GNU/Linux system, all of which are free and open source software (FOSS).

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

Chapter 2:

Populations and Observations

2.1 Defining Populations and Variables

The scope of this chapter.

- We are concerned with analyzing data.
- Data are the result of observations:
  - counting,
  - measurement,
  - more complex operations.
- Now, we'll have a closer look at
  - what data are,
  - where they come from,
  - how they are structured,
  - some problems when collecting data.

2.1 Defining Populations and Variables

Populations, variables, values.

- **population**: a set of objects of interest ("target population")
- **variable**: a property (or attribute) of interest of an object
- for each object, a variable takes on one among several **values**

2.1 Defining Populations and Variables

Elements of populations, variables, and their values.

- Example 1:
  - element: person
  - variable: gender
  - values: female, male
- Example 2:
  - element: person
  - variable: literacy
  - values: literate, illiterate
- Example 3:
  - element: person
  - variable: educational attainment
  - values: mezun olmayan, ilkokul, . . . , doktora

2.1 Defining Populations and Variables

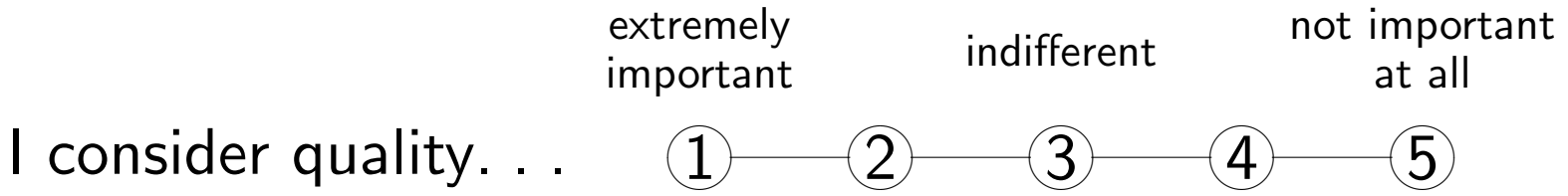
Elements of populations, variables, and their values.

- Example 4:
  - element: purchase from a supermarket
  - variable: total expenditure (YTL)
  - values: . . . , 10.00 YTL, . . .
- Example 5:
  - element: day in 2008
  - variable: closing quotation of the Dow-Jones Industrial Average
  - values: . . . , 10000.00, . . .

2.1 Defining Populations and Variables

Elements of populations, variables, and their values.

- Example 6:
  - element: potential customer of a clothing shop
  - variable 1: gender
  - values 1: female, male
  - variable 2: "I consider quality ..... when buying clothes."
  - values 2:



2.1 Defining Populations and Variables

Difficulties in defining populations and variables.

- How should the population be defined precisely?
  (Which elements belong to it?)
- How should the value taken on be measured precisely?
- How should the population be defined "logically"?
  (What **is** an element?)

2.1 Defining Populations and Variables

The scaling of variables. A variable is called. . .

- **categorical** if all we can say is if two observations are the same or not.
- a **rank** variable if there is a ranking (an ordering) among the observations.
- a **metric** variable if the observations are real numbers.
  - **ratio** variable: differences, as well as ratios, of two values are meaningful
  - **interval** variable: only differences are meaningful

2.1 Defining Populations and Variables

Example: A metric variable.

Population: TV sets;
variable: a TV set's price in €

For example,

TV set A: €200;
TV set B: €210.

Then, we can compute:

$$\frac{210 - 200}{200} \cdot 100\% = 5\%,$$

that is: B is 5% more expensive than A.

2.1 Defining Populations and Variables

Example: Another metric variable.

Population: historic events
variable: the year (AD) an event happened

For example,

Turkish conquest of Constantinople: 1453;
foundation of the Turkish Republic: 1923.

We compute:

$$\frac{1923 - 1453}{1453} \cdot 100\% = 32.35\%$$

What does this mean?

2.2 Data Collection

Techniques of data collection.

There are three basic techniques:

- questioning
- observation
- experiment

2.2 Data Collection

Example:

The clothing shop ByeStyle commissioned an inquiry. Goals:

- Gain insight into the motivation of potential customers.
- Investigate customer behaviour.

Method: face-to-face interviews, using a questionnaire.

- What is a suitable first question?
- What about a "Why"-question?
- Should we ask interviewees to list items?

2.3 Sampling from Populations

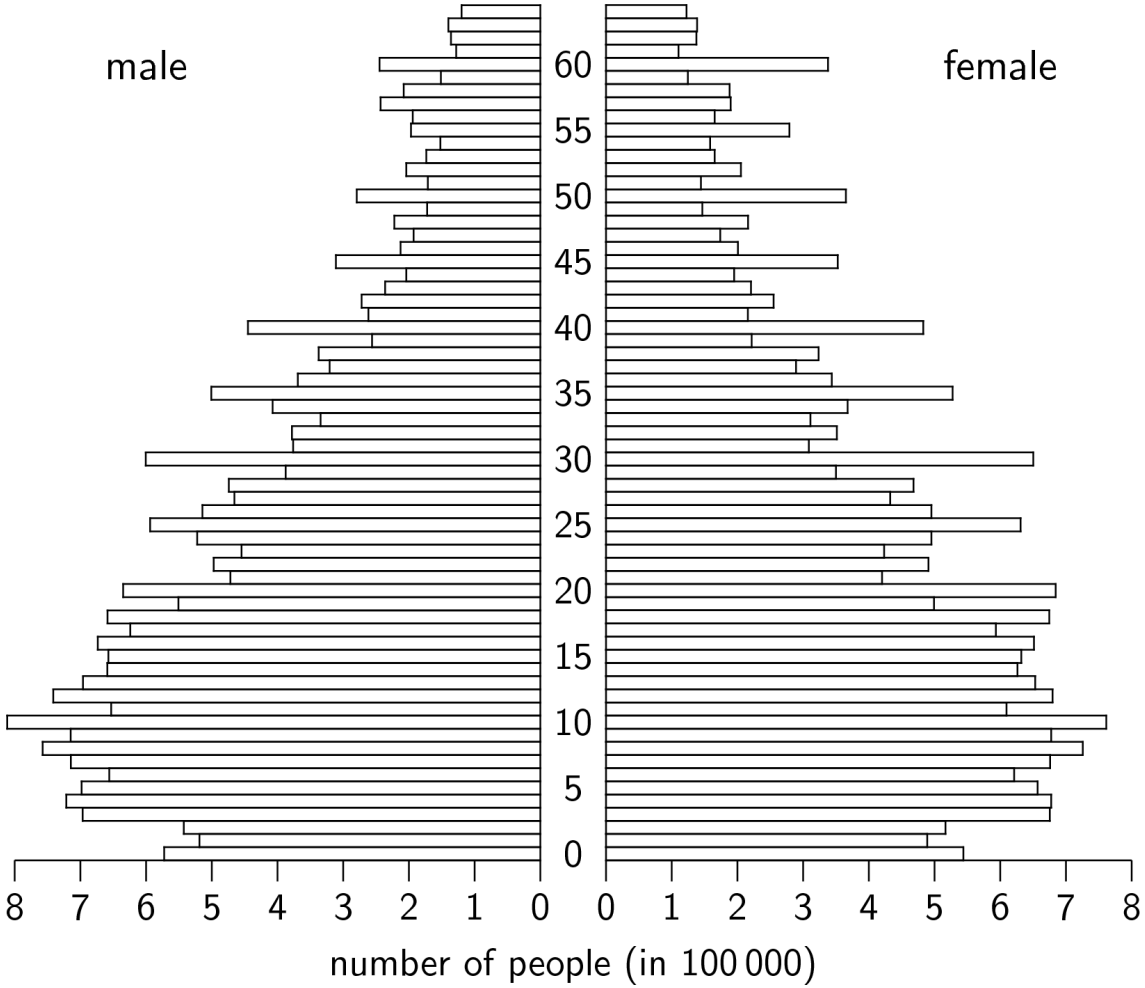
Census versus sampling.

- **Population census**: the process of collecting data from the whole population
- **Sampling**: the process of collecting data from a subset

A carefully drawn sample can be better than a population census!

2.3 Sampling from Populations

Is this the population of Turkey in 1990?



2.3 Sampling from Populations

Complete random sampling:

The process of drawing a sample of size n from a population of size N is called **complete random sampling** if each subset of n elements has the same chance of being selected.

Why do we prefer this way of sampling?

- It ensures representativity.
- It permits the use of inductive statistics.

2.3 Sampling from Populations

Other sampling methods:

- stratified sampling
- convenience sampling
- quota sampling

There are many pitfalls and difficulties!

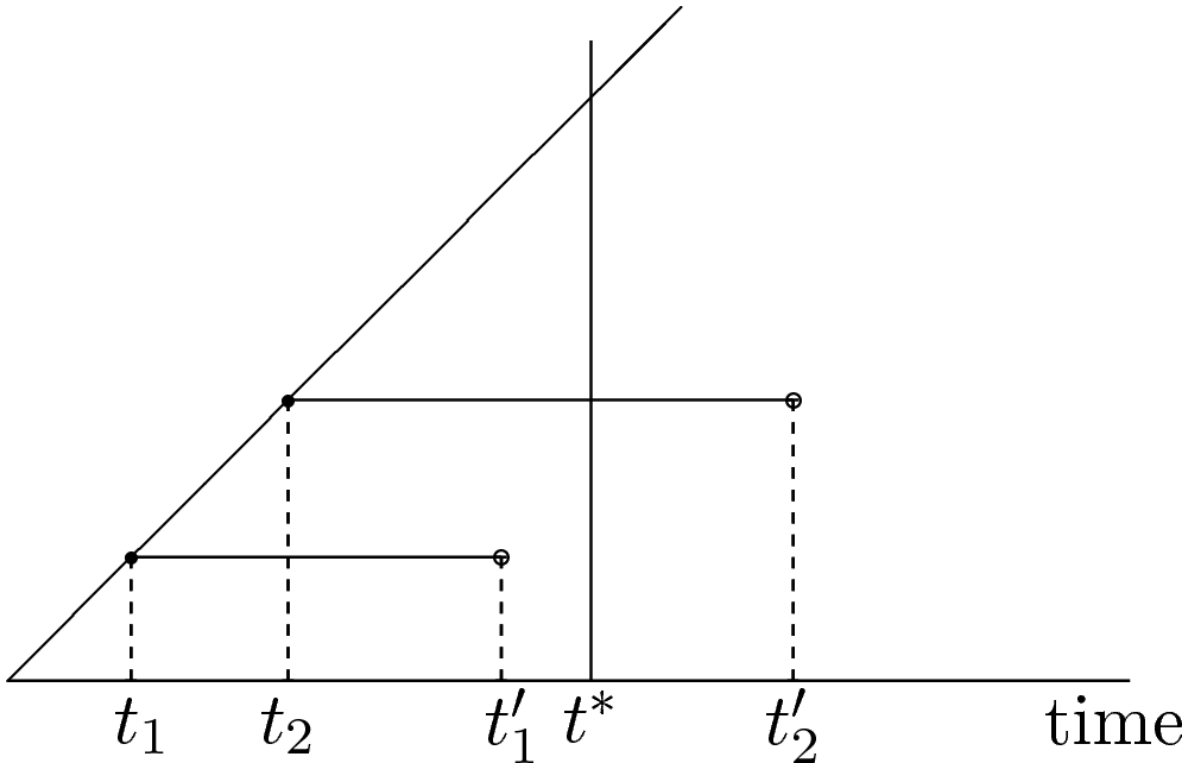
2.4 Difficulties and Pitfalls

Examples of problem sources. . .

- telephone sampling
- nonresponse error
- asking an embarrassing question
- the length sampling bias

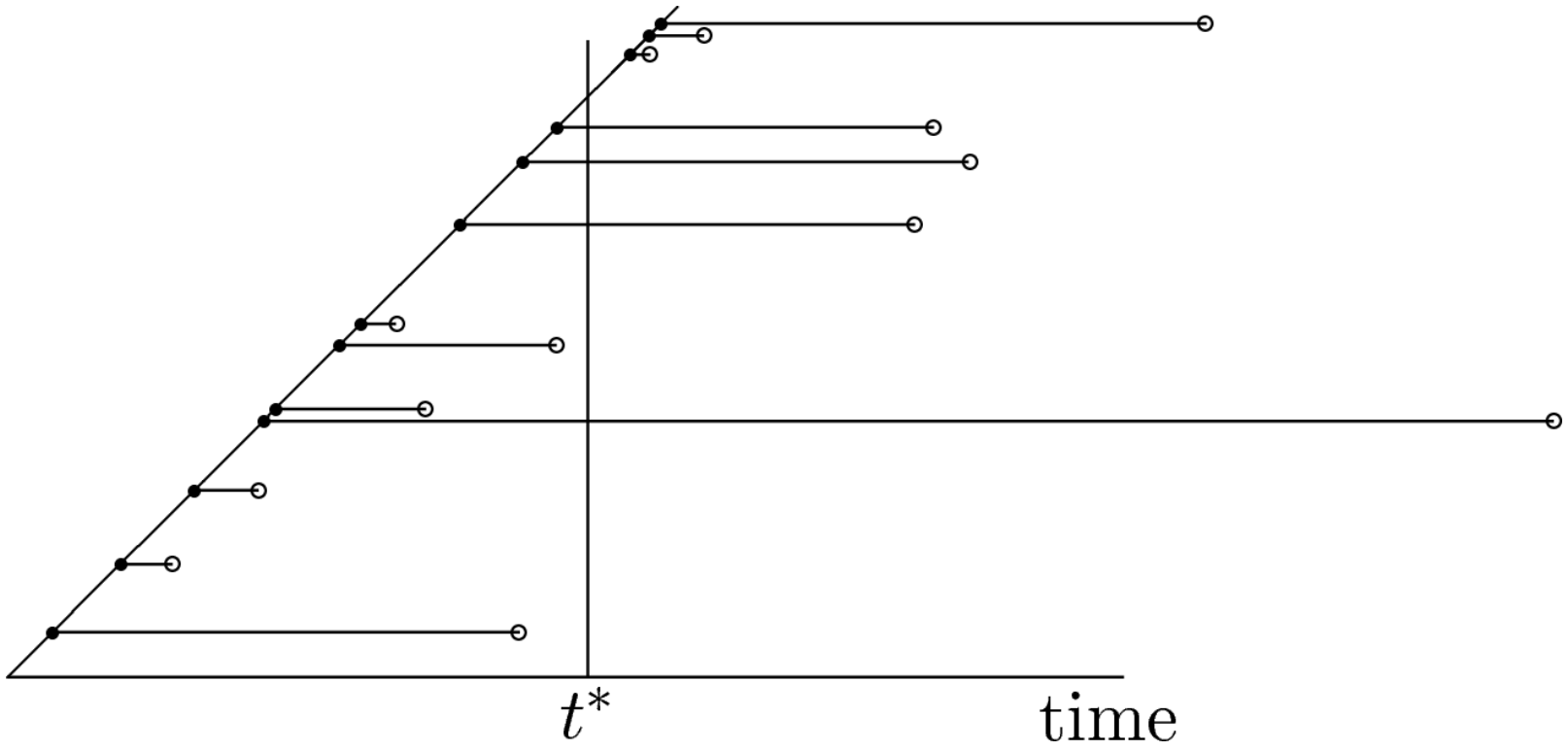
2.4 Difficulties and Pitfalls

The Becker diagram.



2.4 Difficulties and Pitfalls

The length sampling bias.



2.4 Difficulties and Pitfalls

The question of representativity.

- Goal: Data should be collected such that the sample is representative of the entire population.
- Length sampling bias: The sampled intervals are not representative. (They are too long.)
- A problem may occur whenever
  - the probability of being selected into the sample and
  - the variable of interest

  are somehow connected.

2.4 Difficulties and Pitfalls

The length sampling bias.

- Length sampling bias: The sampled intervals are not typical.
- The length sampling bias can serve as an illustration for certain cosmological studies. Some keywords:
  - anthropic bias / anthropic principle
  - Copernican principle / principle of mediocrity

  (Mankind as a "sample in universe.")

2.5 Data: Further Aspects and Problems

Data are needed for empirical research.

Some aspects:

- Check definitions and concepts. Do they match your aims?
- Look into your data. In particular:
  - Data entry mistakes?
  - Treatment of missing values?
  - Date and time format?
- Display your data!
- Consult and compare different data sources, if available.

2.5 Data: Further Aspects and Problems

Example 1: Hypothetical, but realistic.

- In the following table:
  - q_t = quantity in period t,
  - change over the previous period:

$$r_t = \frac{q_t - q_{t-1}}{q_{t-1}} \cdot 100\%$$

- There is a slight inconsistency in the series.
- How can the series (q_t) be re-constructed?

t	1	2	3	4	5	6
q_t	100	120	100	120		
r_t		+20%	−25%	+20%	+25%	+20%

2.5 Data: Further Aspects and Problems

Example 2: Also hypothetical, again realistic.

- In the following table:
  - q_t = quantity in quarter t,
  - change over the same quarter of the previous year:

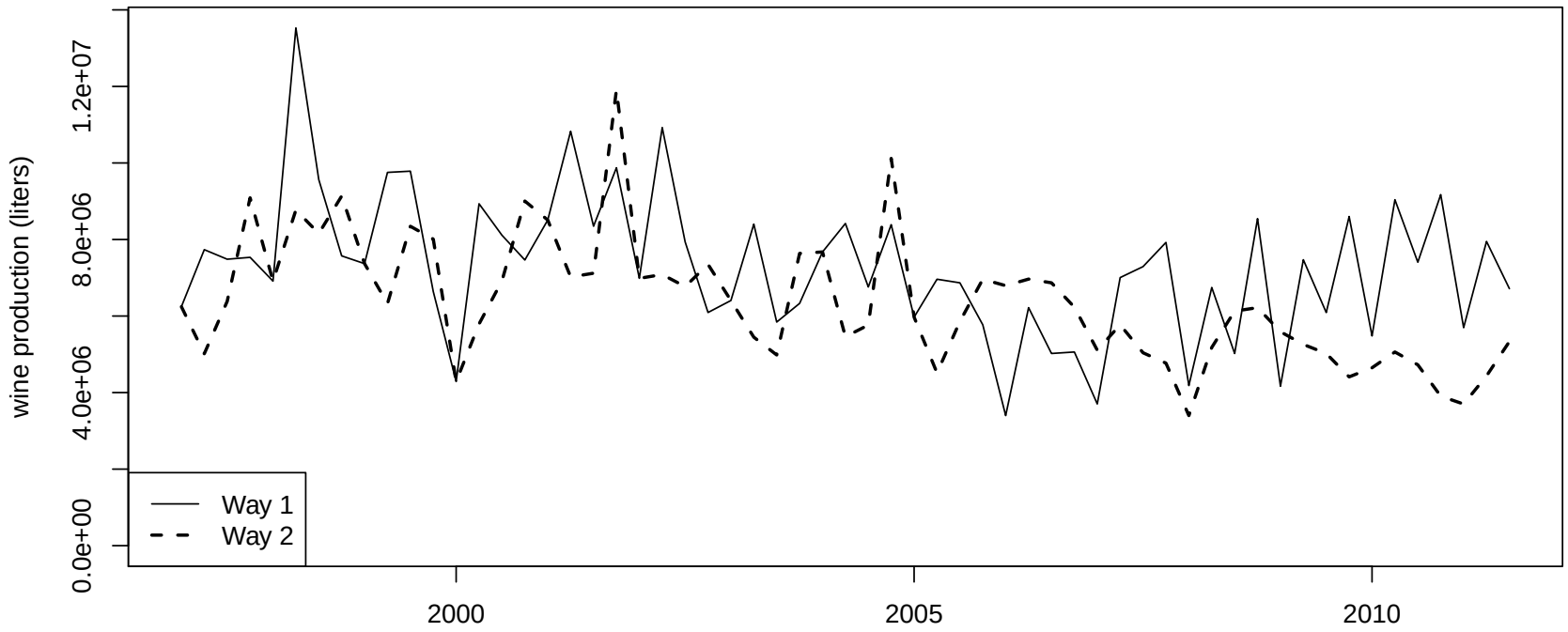
$$r_t = \frac{q_t - q_{t-4}}{q_{t-4}} \cdot 100\%$$

- There is again a slight inconsistency in the series.

t	1	2	3	4	5	6	7	8	9
q_t	100	100	100	100	100	80	80	80	
r_t					-20%	-20%	-20%	-20%	-10%

2.5 Data: Further Aspects and Problems

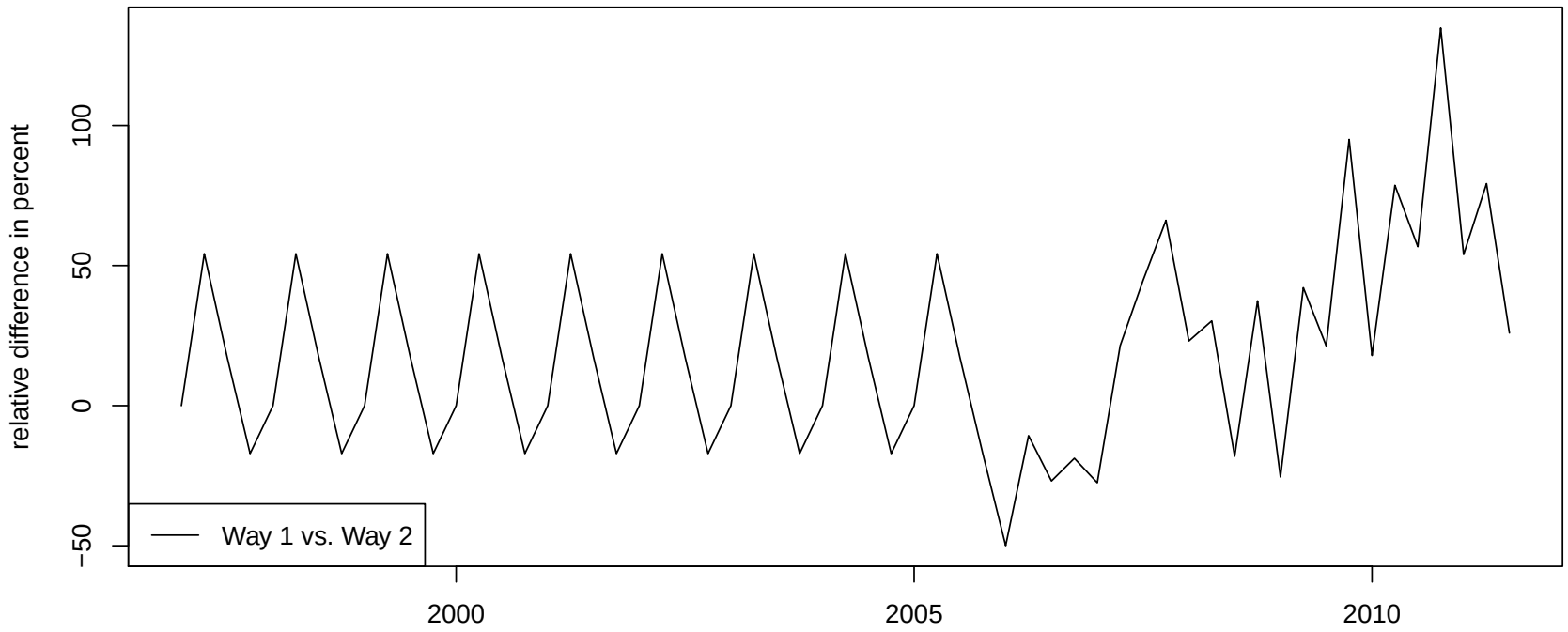
Example 3: Wine production in Turkey.



Result of putting together partial level and change series (source: TÜİK) in different ways.

2.5 Data: Further Aspects and Problems

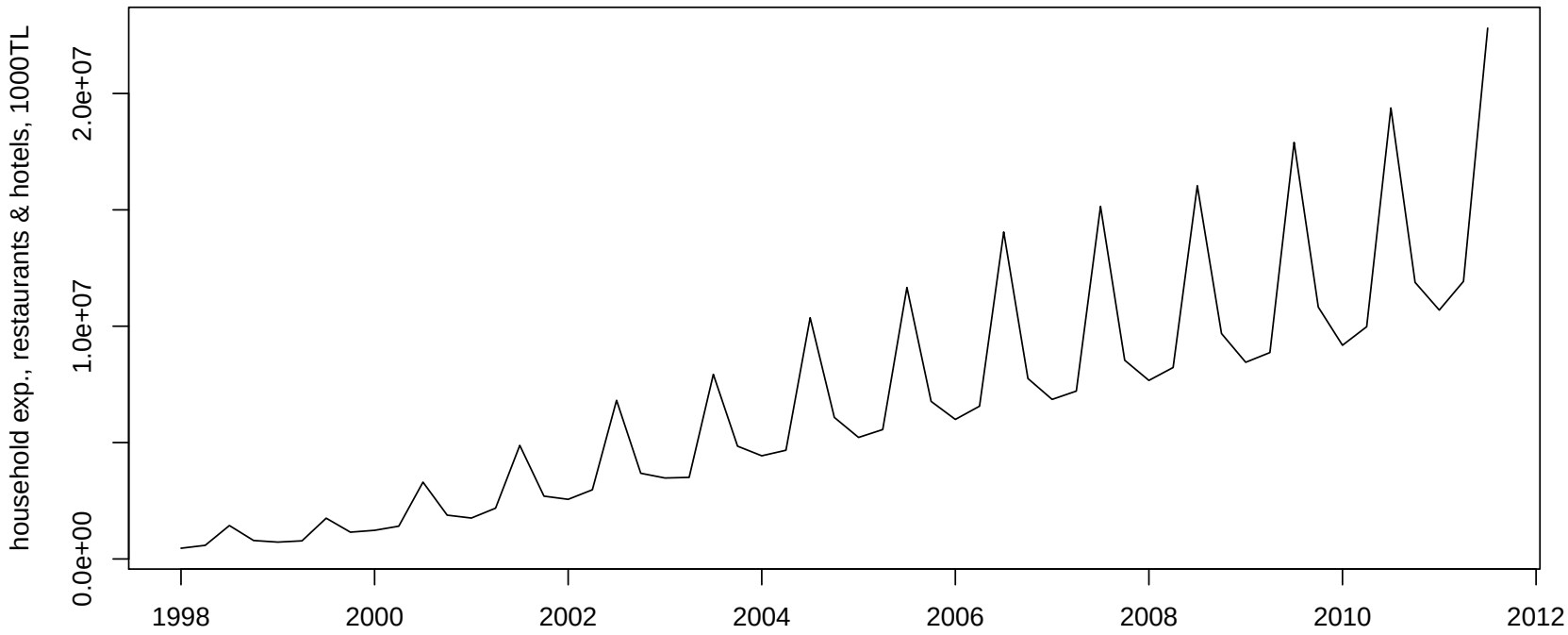
Example 3: Wine production in Turkey.



Relative differences between the two ways to reconstruct the level series.

2.5 Data: Further Aspects and Problems

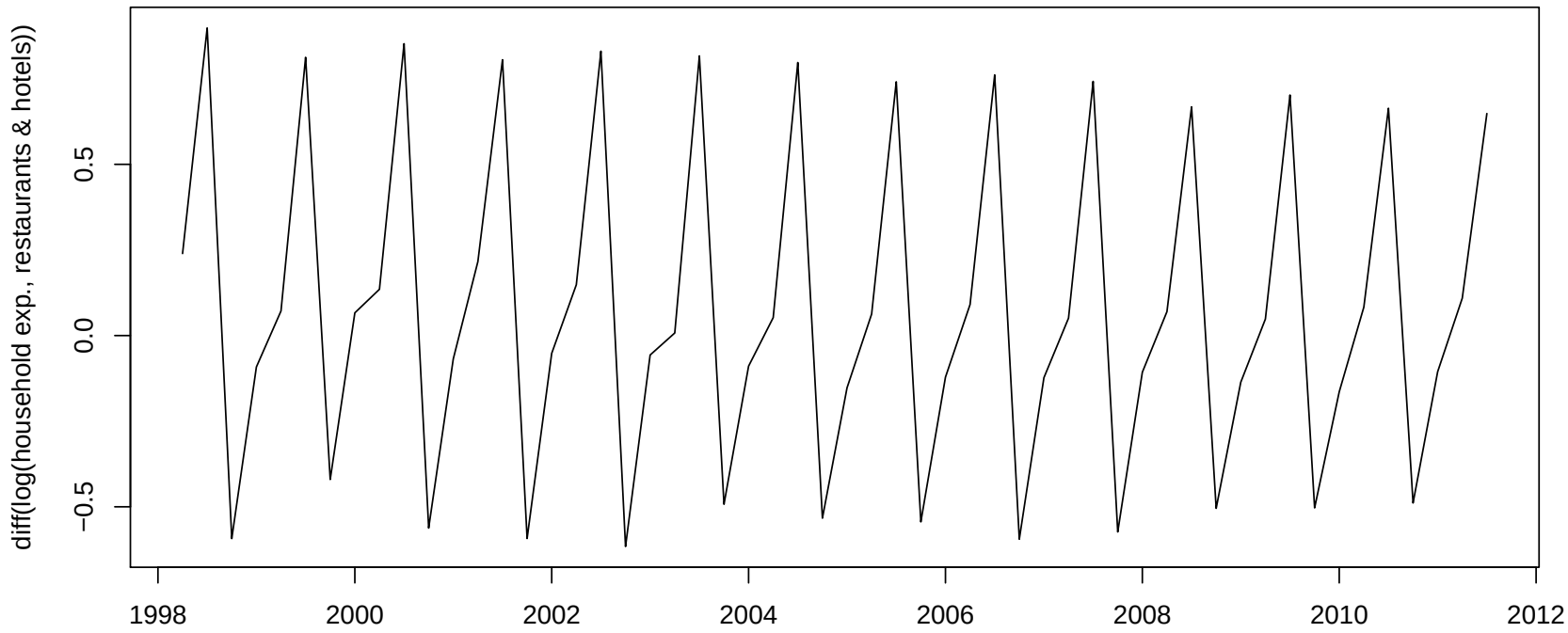
Example 4: Expenditure on hotels and restaurants.



"Final Consumption Expenditure of Resident and Non-Resident Households on the Economic Territory (at current prices), 1998–2011" (TÜİK); here: Restaurants and Hotels.

2.5 Data: Further Aspects and Problems

Example 4: Expenditure on hotels and restaurants.



Series as before; here: differences of logarithms.

2.5 Data: Further Aspects and Problems

"Data manipulation" — an example.

- Cronbach's α: measure of internal consistency of a data set.
- Some researchers say: "A good data set has $\alpha > 0.75$."
- What if your data set has $\alpha < 0.6$?
- Optimize your data set. (Exclude cases to increase α.)
  Simple exercise: Write a program in R to do that.
- Is this acceptable?
- Candidates for exclusion may contain valuable information.

2.5 Data: Further Aspects and Problems

Conclusion.

download/collect data

$\Downarrow$

analyze them!

2.5 Data: Further Aspects and Problems

Conclusion.

download/collect data

$\Downarrow$

analyze them!

This is an **illusion**.