

Elements of Statistics

Chapter 1: Introduction into Statistics

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier
Faculty IV
Economic and Social Statistics Department

Winter term 2021/22

© WiSoStat

The slides are provided as supplementary material by the Economic and Social Statistics Department (WiSoStat) of Faculty IV of the University of Trier for the lecture "Elements of Statistics" in WiSe 2021/22 and the participants are allowed to use them for preparation and reworking. The lecture materials are protected by intellectual property rights. They must not be multiplied, distributed, provided or made publicly accessible - neither fully, nor partially, nor in modified form - without prior written permission. Especially every commercial use is forbidden.

What is Statistics? (1)

status (lat.): State, Status

statisticum [Latin]: Concerning the state

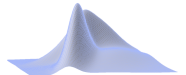
statista [Italian]: Statesman

→ **Statistik** [German]: Science of state affairs

Oxford English Dictionary:

Material sense (statistic): A statistical fact, statement, or piece of data

Instrumental sense (statistics): The systematic collection and arrangement of numerical facts or data of any kind; (also) the branch of science or mathematics concerned with the analysis and interpretation of numerical data and appropriate ways of gathering such data



What is Statistics? (2)

Statistics can be simple counting and measuring.

→ Presentation of the results in tables and graphics

Statistics is a scientific discipline which uses methods for description and analysis of aggregates through numbers.

→ Application of adequate methods in the applications

Statistics does not examine a specific field of human experience, but it is a methodical instrument for all sciences like economics, social sciences, medicine, natural sciences, etc.

The German Census

- ▶ Census of a population (inhabitants of Germany)
- ▶ Last *Volkszählung* in West Germany: 25.05.1987 (East: 31.12.1981)
- ▶ A census should be carried out approximately every 10 years.
- ▶ Census program:
 - ▶ Size of the population (State, Federal States, Districts, Communities)
 - ▶ Distribution of gender and age
 - ▶ Professions
 - ▶ Household structure
 - ▶ Housing

→ Current population statistics

→ Communal fiscal adjustment

Register based census in Germany 2021:

Methodological research by Trier University (Münnich) and
GESIS Leibniz-Institut für Sozialwissenschaften in Mannheim (Gabler);
see: <https://www.uni-trier.de/index.php?id=57986>

Further Examples for Statistics

- ▶ Inventory
 - ▶ Aim: Counting of items in, e.g., a warehouse
 - ▶ Methods: Complete inventory count vs. inventory sampling procedure
- ▶ National accounts
- ▶ Price indices and price changes
- ▶ Demographic development
 - ▶ Birth rate
 - ▶ Death rate
 - ▶ Migration
- ▶ Measurement of the progress in attaining EU targets (e.g. Lisbon Process) \Rightarrow need for *adequate* indicators

Stocks vs. flows

The German Microcensus

- ▶ Classification of sample units
 - ▶ 214 regional strata (RS)
 - ▶ 5 house size classes (GGK)
 - ▶ Clusters of approx. 20 persons by RS x GGK:
Sampling District (AWB)
- ▶ Sampling design
 - ▶ Pooling of 100 clusters each to one zone
 - ▶ Selection of one cluster per zone
 - ▶ $\approx 1\%$ of persons/households
- ▶ Since 2005 during the year, before annually
- ▶ Basis for the Access-Panel
 - ▶ European Union Statistics on Income and Living Conditions (EU-SILC)
 - ▶ Survey about the use of information technologies (IKT)

Unemployed in the Microcensus campus file

```
H <- with(subset(Y, Unemployed==1),
table(Gender, AgeGroup)
)
H <- cbind(H, total=apply(H,1,sum))
H <- rbind(H, total=apply(H,2,sum))
```

	[0;15)	[15;25)	[25;45)	[45;65)	at least 65	total
male	0	124	288	253	1	666
female	0	66	235	222	1	524
total	0	190	523	475	2	1190

- ▶ Data taken from [Microcensus](http://www.forschungsdatenzentrum.de) 2002 campus file
(Source: <http://www.forschungsdatenzentrum.de>)
- ▶ Subsequently: Projection of sample to federal level
- ▶ How high is the quality of the projections?

Data quality: Eurostat definition (1)

Relevance of statistical concepts

User needs, concepts and granularity

Timeliness and punctuality

Time lag between data release and phenomenon and
time lag between actual and targeted data release

Accessibility and clarity

Data (format) availability and documentation, visualisation etc.

Completeness

Coherence

- ▶ Preliminary and final statistics
- ▶ Yearly data and higher frequency data
- ▶ Definitions

Data quality: Eurostat definition (2)

Comparability

Geographical areas, domains, time

Accuracy of estimates

- ▶ **Sampling error:**
Standard error, confidence interval
- ▶ **Non-sampling error:**
Non-response, coverage error, measurement error, model error

Statistical production process

- ▶ **Input — Transformation — Output (and meta data)**
- ▶ Problem of statistical adequation
- ▶ Burden of respondents and costs incurred by data producer

Population vs. sample

Population

The set of all elements relevant to a specific research question is called population. Therefore, a population is a set of elements which share a certain system of defining characteristics relevant to the research question.

Delimitation of a population

- ▶ Time
(populace at point in time vs. births during period of time)
- ▶ Geography
(GDP vs. GNP)
- ▶ Fact
(resident population vs. expatriates)

Sample

A subset of or selection from the population is called sample, where the sample can be **random** or **non-random**.

Scaling of variables (1)

Aim: Specification of a research question

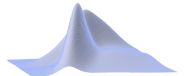
We have to distinguish

- ▶ **statistical units** from
- ▶ their **attributes** or **statistical variables** and those in turn from their
- ▶ **values** or **realisations**.

The definition of the set of possible values of a variable is called its **scaling**.

Statistical variables can be scaled differently and at different levels of intensity. Some examples:

- ▶ Marital status
- ▶ Social stratum
- ▶ Age
- ▶ Income



Scaling of variables (2)

Nominal variables

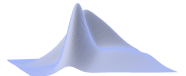
A variable is called nominal if its values only express equality or inequality but no order, no distance and no ratio.

Examples: Marital status, citizenship, soil colour

Ordinal variables

A variable is called ordinal if its values express inequality as well as a certain order, but neither a distance nor a ratio. Assigned numbers are supposed to express the order of the values.

Examples: Ranks, risk preference, water quality



Scaling of variables (3)

Interval-scaled variables

A variable is called interval-scaled if its values express inequality, an order as well as a distance or difference, but no ratio.

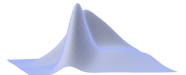
Examples: Speeding, *historical* year, temperature ($^{\circ}\text{C}$)

Ratio-scaled variables

A variable is called ratio-scaled if its values express an inequality, an order, a distance or difference as well as a ratio.

Examples: Income, duration, global radiation

Interval-scaled and ratio-scaled variables are both called **metric variables**.



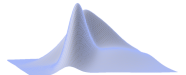
Intensity and measurement problems (metric variables)

Discrete: At most a countable number of values

- ▶ Finite
- ▶ Countably infinite

Continuous: An uncountable number of values

Problem of measurement accuracy: Age, weight, income, etc.



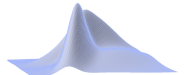
Descriptive and inferential statistics

Descriptive statistics

Collected data are typically processed (for better interpretation) or *condensed* (e.g. figures, tables, statement of *typical* values) in an appropriate manner. In statistics, this work is known as description; the methods used here are summarised under the term descriptive statistics.

Inferential statistics

If data are sampled, descriptions only pertain to the sampled elements of the population at first. Nevertheless, typically conclusions about the population as a whole are desired. The transfer of sample results to the population is called statistical inference or inferential statistics. When dealing with sampled data, inferential statistics have to complement the description of the sampled population subset.



Statistical institutions and types of statistics

Non-official statistics

Businesses, research institutes, organisations

Official statistics

National statistical institutes (NSIs), other public-law institutions,
Eurostat, European Central Bank (ECB), United Nations (UN)

Primary statistics

Statistics based on data collected for the specific task at hand

Secondary statistics

Statistics based on data originally collected for other purposes
(e.g. tax statistics)