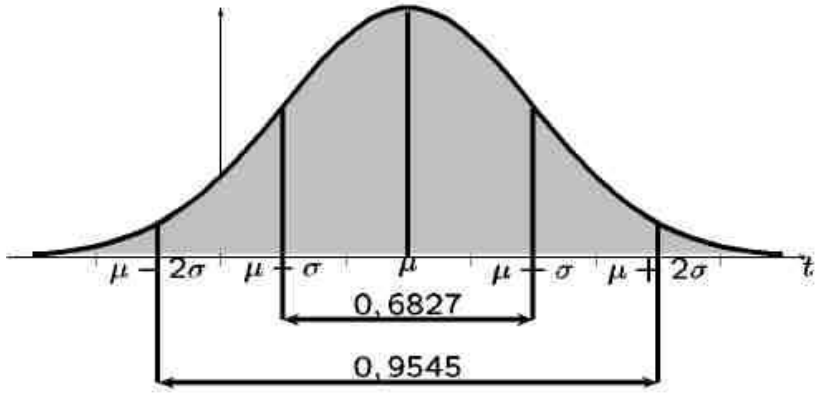
Statistics

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$$



Bachelor Studiengang Informatik

Prof. Dr. Egbert Falkenberg

Fachbereich Informatik & Ingenieurwissenschaften

Wintersemester 22/23

Section 1

Overview

Overview - Remarks I

- Lectures and exercises
  - dates according to the time schedule
  - lectures: tuesday, 8:15-10:45 am
  - exercises: 4 groups Schrader and 2 groups Falkenberg
  - please enroll in one of the groups
- Written examination: february
- Course material: exercises, slides, etc. $\rightarrow$ moodle plattform
  - course: Falkenberg: Statistics Bachelor Informatik
  - access-key: statistics_ws_22
- consultation hour: tuesday, 2:00-3:00 pm
- mail adress: falken@fb2.fra-uas.de

Overview - Remarks II

- Course based on
  Online Statistics: An Interactive Multimedia Course of Study
  - introductory-level statistics book
  - standard textbook and multimedia presentation available
  - interactive demonstrations and simulations, case studies, and an analysis lab
  - public domain $\rightarrow$ copies without limitiation
- Christian Heumann, Michael Schomaker, Shalabh: Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R, Springer 2016
- Sebastian Sauer; Moderne Datenanalyse mit R, Springer Gabler, 2019

Overview - Remarks III

- Statistical analyses are based on data.
- Data - usually extensive in practice - must be processed and prepared for statistical analysis before statistical analysis can be carried out.
- Performance of statistical analyses is unconceivable without appropriate tools.
- Here the open source software R will be used.
- In addition to a very extensive collection of statistical procedures, this provides efficient and powerful tools for handling and transforming data.
- R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.

Overview - Remarks IV

- Most of the exercises and parts of the exam are performed on the computer using R.
- I recommend to deal intensively with R and to try out the examples in R yourself as well as to solve the exercises with R.
- You will become familiar with R and its practical application.
- By applying statistical methods to examples, i.e. to (larger) datasets, you will gain a deeper understanding of the methods and their problems.
- A series of tutorials using learnr and swirl as well as Geogebra apps help them become more familiar with R as well as the content of the lectures.

Overview - RStudio

- The treatment of extensive data with statistical methods is not feasible without the use of computers.
- Here we will use R: a free software environment for statistical computing and graphics.
- A user friendly environment to run R is RStudio.
- Open-source, available at http://www.rstudio.com/
- Brief introduction in the use of RStudio: https://dss.princeton.edu/training/RStudio101.pdf

Please install RStudio on your computer!

Content

1. Introduction
2. Descriptive Statistics: one and two variables
3. Introduction to the Calculus of Probability:
   basic concepts, some important distribution, normal distribution, central limit theorem
4. Inferential Statistics:
   estimation, confidence intervals, logic of hypothesis testing, some parameter tests

Main Literature

- Christian Heumann, Michael Schomaker, Shalabh: Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R, Springer 2016
- Sebastian Sauer: Moderne Datenanalyse mit R, Springer Gabler, FOM-Edition, 2019
- Online Statistics: http://onlinestatbook.com/2/index.html

Additional Literature

- Virtual Laboratories in Probability and Statistics: http://www.math.uah.edu/stat/index.html an old version is on http://local.disia.unifi.it/VL/VL_EN/index.html
- David M. Lane, ...: Introduction to Statistics: http://onlinestatbook.com/Online_Statistics_Education.pdf
- Java-unterstützte Münsteraner Biometrie-Oberfläche: W. Köpcke, Institut für Medizinische Informatik und Biomathematikm 1998-2003
  http://campus.uni-muenster.de/fileadmin/einrichtung/imib/lehre/skripte/biomathe/bio/bio.html
- Pitman: Probability, Springer Text in Statistics, 1993
- Lehn/Wegmann: Einführung in die Statistik, Teubner Verlag 2006

Glossary of statistical terms

Glossar statistischer Fachbegriffe: Deutsch-Englisch, Englisch-Deutsch, Statistisches Bundesamt, Wiesbaden:
https://www.destatis.de/GPStatistik/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000180/GlossarStatFachbegriffeDE.pdf

Additional Literature R

- Hadley Wickham, Garret Grolemund: R for Data Science, O'Reilly 2018
- Nicholas J. Horton, Ken Kleinman: Using R and RStudio for Data Management, Statistical Analysis, and Graphics, Second Edition, 2015

Section 2

Introduction

Introduction

The aim of statistics is to provide methods to describe, summarise, interpret and analyse extensive data. Drawing conclusions from existing data is of great interest in many areas.

- Statistics are not only facts and figures.
- Statistics refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.
- Statistics provides tools that you need in order to react intelligently to information you hear or read.

Usually, a distinction is made between two statistical areas.

Descriptive Statistics

- **Objective:** Organizing, summarizing, and describing data but not generalizing beyond the data at hand
- **Generalizing from a sample to the population it was sampled from is the domain of inferential statistics.**
- Numbers and graphs are used to summarize and describe data.
- **Example birth certificates:**
  - descriptive statistics: birthrate in Germany, the average age of the mother, ...
  - Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed.
- Several descriptive statistics are often used at one time, to give a full picture of the data.

Inferential Statistics

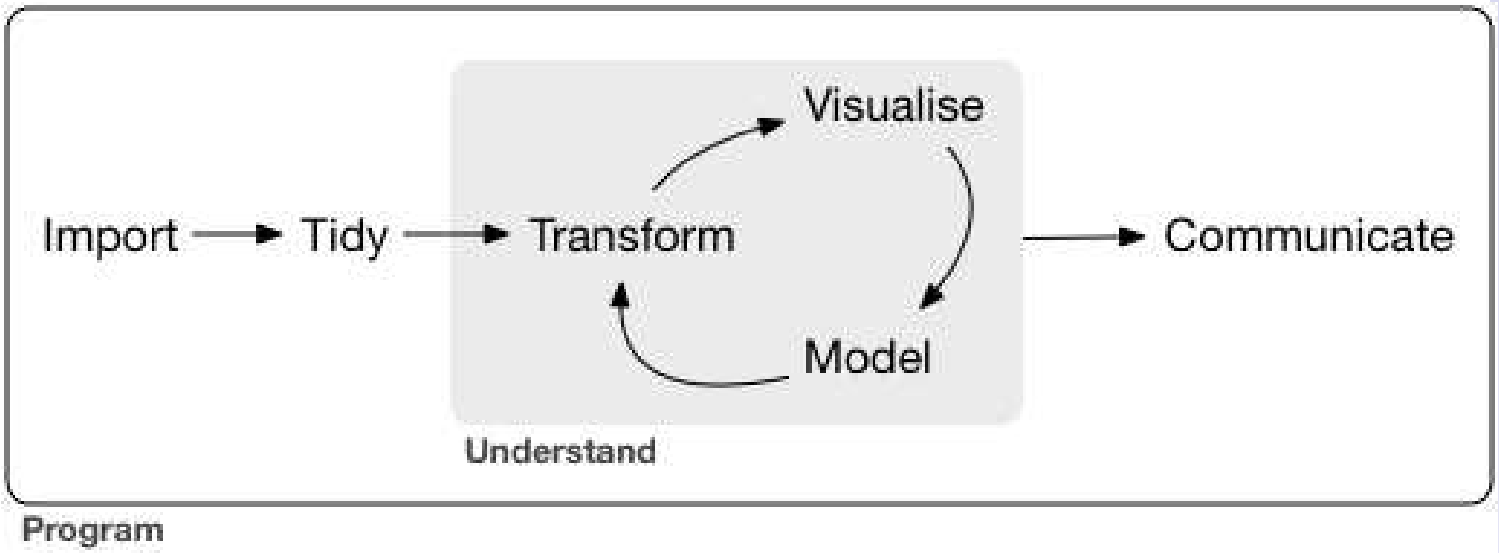
- Branch of statistics concerned with drawing conclusions about a population from a sample.
- Generally done through random sampling, followed by inferences made about characteristic numbers of a distribution like mean, standard deviation, etc..

Questions

- Which are of the following are part of statistics? Select all that apply.
  - ☐ numerical calculations
  - ☐ graphs
  - ☐ interpretations and decisions based on the numbers and graphs
- A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared.
  Is this an example of descriptive or inferential statistics?

Steps in Data Analysis I

- The application of statistics, i.e. the statistical analysis of data - usually extensive data - is not possible without the use of computers.
- Grolemund and Wickham [1] divide the data analysis process into several steps.



[1] https://r4ds.had.co.nz/introduction.html, chapter 1.1

Steps in Data Analysis II

Compare https://r4ds.had.co.nz/introduction.html, chapter 1.1

- Importing the data, i.e. take data stored in a file, etc. and load it into a appropriate software.
- Tidying the data, i.e. store it in a consistent form that matches the semantics of the dataset with the way it is stored.
- Transforming is a common first step. Transformation includes narrowing in on observations of interest, creating new variables that are functions of existing variables, and calculating a set of summary statistics.
- Visualising the data: A good visualisation will show things that you did not expect, or raise new questions about the data.
- Modeling or testing hypotheses: How can the data be explained?
- Communicating the results not the data.

Steps in Data Analysis III

- We will use the programming language R as a tool.
- With the help of R all these steps of a data analysis can be realized.
- Programming with R is therefore a cross-section tool that you can use in every part of a data analysis.

Population, Sample I

- population: large set of objects of a similar nature which is of interest as a whole
- sample: portion of the elements of a population

Example: The National Election Commission (USA) wants to examine how the American people feel about the fairness of the voting procedures in the U.S.

Problems:

- It is not practical to ask every single American.

$\rightarrow$ a relatively small number (**sample**) would be asked to draw inferences about the entire country (**population**) from their responses.

- It is crucial that the sample is representative.

Population, Sample II

Example: A teacher wants to know how students in the class did on their last test. He asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well.

Questions:

- What is the sample?
- What is the population?
- Can you identify any problems with choosing the sample in the way that the teacher did?

Population, Sample III

Basic assumption: sampling is random
Random sampling:

- most straightforward sample strategy
- Characteristics:
  - Every member of the population has an equal chance of being selected into the sample
  - The selection of one member must be independent of the selection of every other member.

$\Rightarrow$ Selection by pure chance

Problems can result, if selection is not by pure chance!

Population, Sample IV

Example: A random sample of 20 subjects were taken from a population with an equal number of males and females.
Problem: There would be a nontrivial probability (0.06) that 70% or more of the sample would be female.
$\Rightarrow$ The sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population.
Random samples, especially if the sample size is small, are not necessarily representative of the entire population.

Population, Sample V

Summary:

- Usually a sample is drawn.
  - $\Rightarrow$ Because of chance factors, the composition of the sample will differ somewhat from the composition of the population.
  - $\Rightarrow$ Statistical inferences based on a sample may be wrong.
- Wrong inferences can not be avoided but the risk of wrong inferences could be calculated.

$\Rightarrow$ Mathematical models describing random influences are necessary.

Probability and Probability Models

Section 3

Basic terms and definitions

Basic terms and definitions

A prerequisite for understanding statistics is to know how to correctly collect, manage, evaluate and analyze data. Basic terms and definitions for the correct handling of data are to be introduced here.

Population, Sample and Observation

- **Observations:** units on which data is measured, usual notation ω
- **Population:** collection of all possible observations, usual notation Ω
- **Sample:** selection of observations $\omega_1, \omega_2, ...\omega_n$, a sample is allways a subset of the population Ω

Example: I am interested in informations about those students participating this statistics course.

- Ω : all participants of the course
- ω : one participant of the course
- sample: for example all students sitting in the current lecture

Variables I

After defining the population of interest for a specific questions we should think of what is of interest about our observation.

- **Variable:** A specific feature of an observation is captured in a statistical variable X.
- If we are interested in several features, each of them is captured in a variable $X_i, i = 1, ..., p$. Each observation ω takes a particular value for X_i collected as $X = (X_1, X_2, ..., X_p)$. The concrete values of an obsveration ω are given by $(x_1, x_2, ..., x_p)$

$$X : \Omega \mapsto S, \omega \to x$$

where S denotes the set of all possible values of the variables.

Variables II

Example: Observations are students of FRA-UAS.
Following properties are described by variables:

- age
- gender
- religion
- weight
- size
- course
- semester
- marks in mathematics at school
- body temperature
- Credit Points
- date of birth

Qualitative and Quantitative Variables

- Qualitative variables: express a qualitative attribute
  Values of qualitative variable cannot be ordered in a logical or a natural way. They only denote a certain category.
- Quantitative variables: variables that are measured in terms of numbers
  Values of quantitative variable can be ordered in a logical or a natural way.

Discrete and Continuous Variables

- Discrete variables take on only a finite number of values
- Continuous variables can take on any value in a certain range
- All qualitative variables are discrete. Quantitative variables are either continuous or discrete.

Levels of measurement (Scales) I

- Variables differ in "how well" they can be measured, i.e., in how much measurable information their measurement scale can provide.
- A factor that determines the amount of information that can be provided by a variable is its "type of measurement scale".
- The variable's level of measurement determines the statistics that can be meaningfully computed with that variable.

Levels of measurement (Scales) II

1) nominal scale

- Nominal variables can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even order those categories.
- Examples: gender, colour,

2) ordinal scale

- ordering the measured items in terms of which has less and which has more of the quality represented by the variable is possible, but still it is do not allow us to say "how much more."
- For example, marks in mathematics, as measured in numbers from 1 to 5. We can not say that the differences between adjacent marks are equal.
- Examples: military ranking, consumer satisfaction rankings

Levels of measurement (Scales) III

3) continous scale:

The values can be ordered. Furthermore the differences between the values can be interpreted in a meaningful way. Subscales are

- **interval scale**
  - not only ordering the measured items, but also quantifying and comparing the sizes of differences between them is possible
  - Interval scales do not have a zero point even if one possible value is zero
  - Example: temperature measured in degrees Celsius; we can say that a temperature of 40 degrees is higher than a temperature of 20 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees. But a temperature of 40 degrees is not twice as warm as on 20 degress.
  - Example: dates of events

Levels of measurement (Scales) IV

- **ratio scale**
  - in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y.
  - For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, we can correctly state that it is twice as high.
  - Examples: amount of money in your pocket, length
- **absolute scale**
  - Same as ratio scale but additionaly the values are measured in natural units.
  - Example: number of semester studied but not weight where we have artificial units like kg, pound, etc..

Levels of measurement (Scales) V

variable	type	scale
age	quantitative, discrete	absolute
gender	qualitative, discrete	nominal
religion	qualitative, discrete	nominal
weight	quantitative, continous	ratio
size	quantitative, continous	ratio
course	qualitative, discrete	nominal
semester	quantitative, discrete	ratio
marks in mathematics	qualitative, discrete	ordinal
body temperature	quantitative, continuous	interval (Celsius) resp. ratio (Kelvin)
Credit Points	quantitative, discrete	absolute
date of birth	quantitative, discrete	interval

Grouped Data

In many situations the actual values are not considered, instead they are grouped and you only know to which group the value belongs.

Example:

- In reviews of the incomes of citizens of a country, incomes are often grouped into

  [0-20000) Euro
  [20000-30000) Euro
  ...
  > 100000 Euro

- In a survey for the next upcoming election, the consents are summarized to parties with small approval values or to “other parties”.

Data Set

Data is usually stored in a data matrix where the rows represent the observations and the columns represent the variables.

$$\begin{pmatrix} \omega & \text{Variable 1} & \text{Variable 2} & \dots & \text{Variable p} \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 2 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ n & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Usually the given raw data is not in this but a messy form and some work must spent in data cleaning and data preparation.

Questions I

- Describe the population and the observation for the following questions:
  1. Evaluation of the satisfaction of employees in a company.
  2. Description of the grades of students from an exam.
  3. Comparison of two drugs which deal with a certain sickness.
- An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.
  What is the independent variable? What is the dependent variable?

Questions II

- Categorize the following variables as being qualitative or quantitative and specify the level of measurement used for the items:
  1. Rating of the quality of a movie on a 7-point scale
  2. Age
  3. Country you were born in
  4. Favorite Color
  5. Time to respond to a question
  6. Intelliquence quotient
  7. Immatrikulation number
- Which of the operators
  $=$ $\neq$ $<$ $>$ $+$ $-$ x $/$ can be used for a variable, which is measured on a
  1. nominal scale?
  2. ordinal scale?
  3. interval scale?
  4. ratio scale?
  5. absolute scale?

Content