L1

Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

NCT 1.1-2 + (JB 1.1-3, 2.1-3)

Introduction to BSFE

To learn something from ***observations***

- Summarize and describe experiences (description)
- To draw conclusions (inference)
- To predict the future (prediction, forecasts)
- To decide on which action to take (prescription)

- Typically ***incomplete information***
  - We can't ask everybody, we don't have the time to test every combination, infinite number of possibilities …

$\Longrightarrow$ Statistical methodology!

Why do we observe?

Type of study / the purpose of the study:

- Descriptive *(”this is how it is”)*
- Explanatory, causality *(”it’s like this because …”)*
- Prediction, forecasting (*”what will happen tomorrow?”*)
- Normative, prescriptive *(”do like this and it’ll be like this”)*

- Explorative: search for (new) knowledge
- Confirmative: confirmation rejection of hypotheses

Main purpose is to enhance our knowledge of our world

Statistical surveys, investigations

Collection of data, i.e. observations that we study and analyze in order to get answers to the questions we find important.

Statistical Methods: collect, process, model, analyze, draw conclusions, predict

Note that the word **statistics** in this context refers to the methods, but often the word is used for the data and the numbers themselves; (“the statistics show that...")

Some important concepts

- **Population**
  - A set of well-defined ***objects*** (or *cases*) that possess properties
  - People, corporations, groups, stock, events, …
  - May be ***finite*** in size or ***infinite*** (examples?)
- **Sample (sv. *stickprov*)**
  - A ***subset*** of the population that we observe
  - Always finite in size
- **Variables**
  - The ***properties*** that the objects in the population possess and that we have collected or intend to collect information about

Statistical surveys

- ***Census* (sv. *totalundersökning*)**
  - Collect info on all objects in the population
  - Registers (Swedish Tax Agency, vehicle register, RTB)
  - Actual surveys/collections (FoB, HoB)

- ***Survey, sampling of a subset***
  - There are many ways to sample …
  - E.g. **NCT** p. 23: simple random sampling and systematic sampling; even more ways described in **JB**
  - These are ***random sampling*** designs (sv. *slumpurval*)
  - Other ways of doing it? **L11**

Types of empirical studies

- **Experimental (randomized controlled trials)**
  - Test for ***causation*** (sv. orsakssamband, *kausalitet*) e.g. medication dosage and health effects
  - Allocation to treatment and control groups is ***randomized***
  - You can till ***control*** for covariates – randomize within age groups, gender, …, etc.

- **Quasi-experiments**
  - Allocation to treatment or control groups isn't random
  - Create comparable "twins" (same age, gender, …)

- **Non-experimental, surveys**
  - Allocation to groups not possible, we get what we get

Stockholm University

Some more important concepts

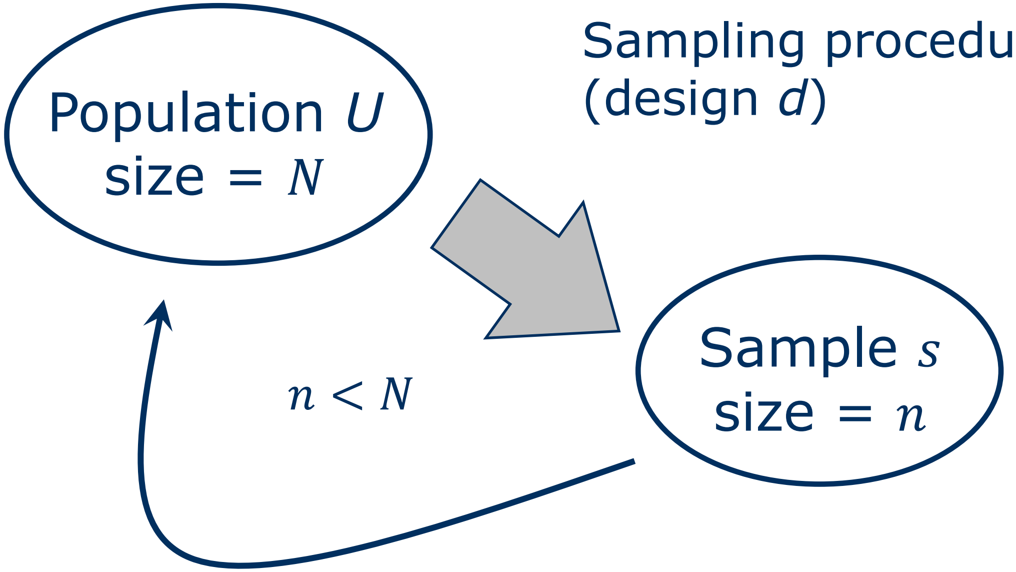
- **Parameter**
  - A ***numerical characteristic*** that defines a specific property of a ***population*** or a ***model***
  - E.g. proportion of smokers in Sweden, probability to be born with a heart condition

- **Statistic (sv. *statistika*)**
  - A ***numerical characteristic*** that defines a specific property of a ***sample***
  - E.g. proportion of smokers among 100 sampled residents of Sweden

A **statistic** is used to **estimate** (sv. *skatta*) a population or model parameter

Conclusions from incomplete information

Parameter:
proportion of smokers



Population U
size = N

Sampling procedure
(design d)

Sample s
size = n

$n < N$

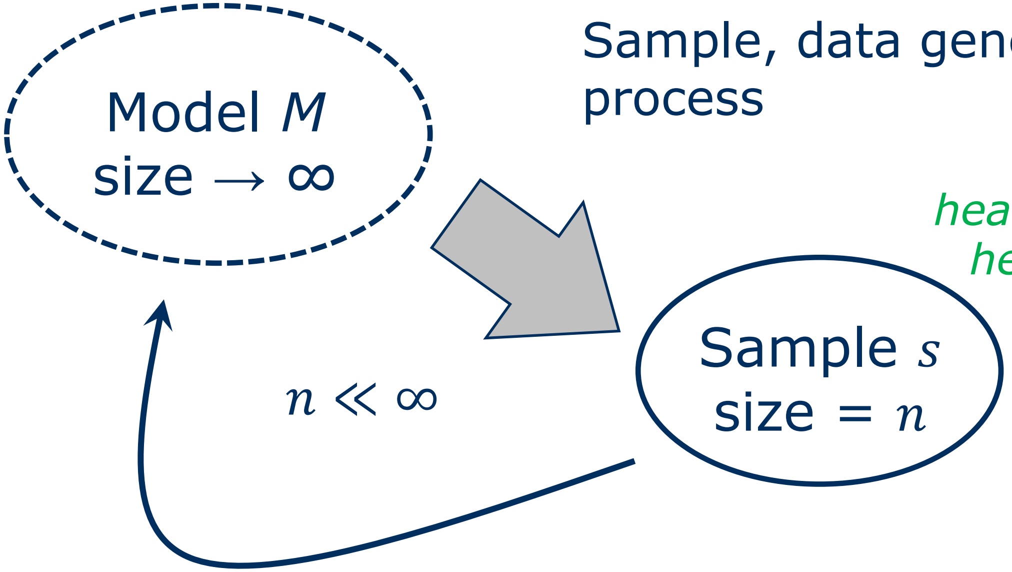
Variable:
smoker/non-smoker

Statistic:
Proportion of smokers

Inference: to say something about the ***finite population*** based on the information contained in the sample

Conclusions from incomplete information

Parameter:
risk of being born with a heart condition



Model M
size → ∞

Sample, data generating process

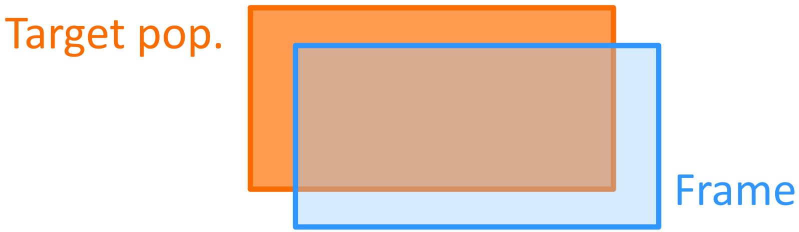
$n \ll \infty$

Sample s
size = n

Variable:
heart problem/no heart problem

Statistic:
proportion born with heart condition

Inference: to say something about a universal property of the objects of study, a data generating procedure or “super population” or ***infinite population***, that can be described with a ***model***, based on the information contained in the sample



More basic concepts

- **Sampling frame** or just **frame (sv. *ram*)**
  - A list, data file, register or similar, that lists all objects in the **finite** population
  - The sample is drawn from the frame
  - Should ideally match the population you are interested in, the **target population**

- **Coverage error (sv. *täckningsfel*)**
  - **Over coverage** – there are objects in the frame that do not belong to the target population
  - **Under coverage** – there are objects in the target population that are missing in the frame

Survey errors

- **Sampling error (sv. *urvalsfel*)**
  - One does not observe all, only a sample of the population
  - We are going to learn how to calculate the sampling error!
  - Named ***standard error***, or ***statistical margin of error***

- **Non-sampling errors (sv. *icke-urvalsfel*)**
  - Errors not due to sampling
  - ***Coverage error*** – mismatches between frame and target
  - ***Measurement error*** – uncertain or incorrect answers
  - ***Missing data*** – non-response
  - Other sources – editing, processing errors etc. …

Why surveys? Why not a census?

- ***Costs*** for data collection
  - Interviewing cost, other types of field collections (CPI)
  - Destroying measurement procedure

- Data collection and processing takes ***time***
  - e.g. 1880 US census, took 8 years to complete
  - Big Data – lots of (often un-structured) data
  - Time restraints, deadlines – decisions have to be taken now!

- ***Reliable*** measurements
  - Non-response follow-ups
  - Measurement error – better methods, interviewer training
  - Measurement mode affects results (interviewing, web, …)

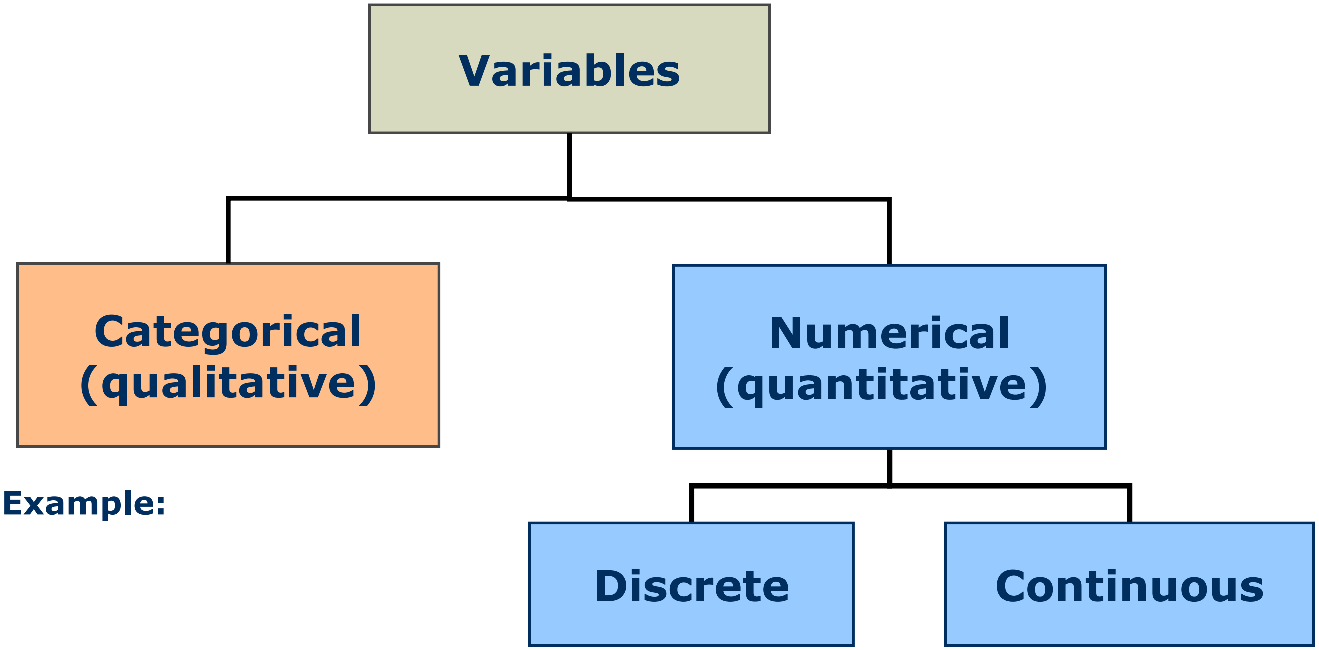
Allocate costs to quality rather than quantity

Variables – the stuff we observe

- **Categorical or qualitative variables**
  - Assumes non-numerical values, categories, types, labels

- **Numerical or quantitative variables**
  - Assumes numerical values, numbers
  - **Discrete variables**
    - Can assume ***only certain*** isolated number values
    - Countable, can be listed; e.g. integers
  - **Continuous variables**
    - Can assume ***any*** number from an ***interval***
    - The interval can be closed or open; limited or unlimited

Classifying variables


Variables

Categorical (qualitative)

Numerical (quantitative)

Discrete

Continuous

Example:

Example:

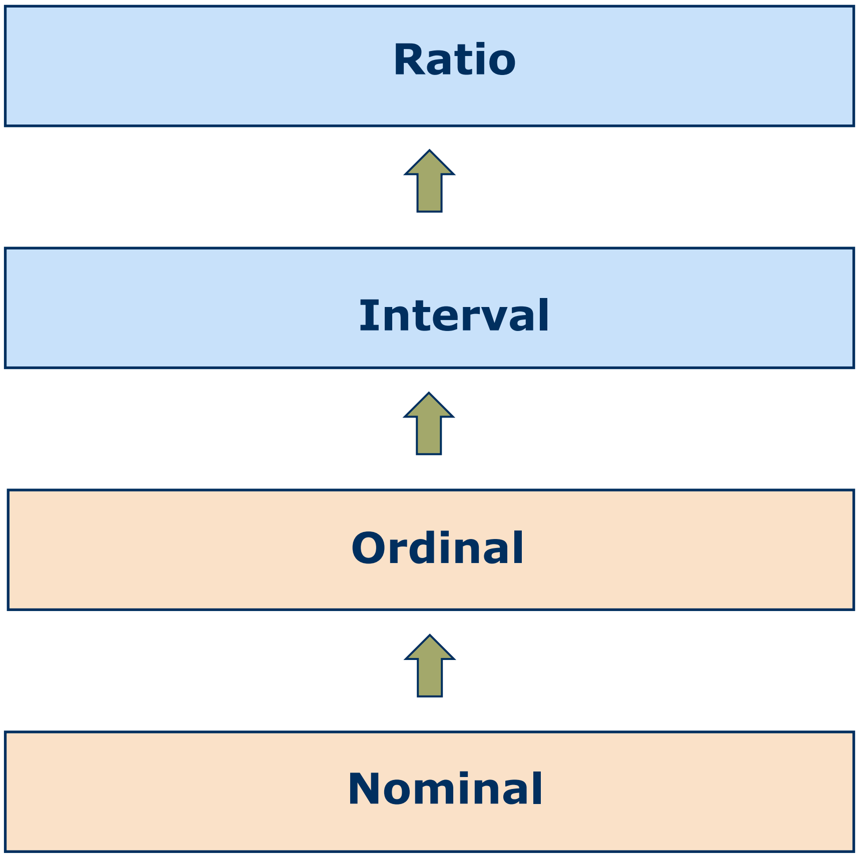
Example:

Scale types

The values that a variable can assume, classifies the variable:

- **Nominal scale**
  - Non-numerical (Latin: nomen = name)
  - E.g. brand names, occupation, countries etc.
- **Ordinal scale**
  - Non-numerical but can be ordered by ”size”/”level”
  - E.g. ”good, better, best”, management hierarchy
- **Interval scale**
  - Numerical values where the distance (differences) are well-defined but not ratios (zero is not properly defined)
  - E.g. temperature scales (”twice as hot”?), clock time
- **Ratio scale (sv. *kvotskala*)**
  - ”$20 is twice as much as $10”; time, distance, …

Scale level



Differences and ratios are well-defined, true zero exists

Ratio

Numerical, quantitative data

Differences are well-defined but not ratios, true zero does not exist

Interval

Ordered categories (ranking order)

Ordinal

Categorical, qualitative data

Categories but no natural ordering exists

Nominal

Data types

Data

- Measurements, observations, data
- e.g. 22 M 45 62 84

Metadata

- information about the data, explains the meaning of the symbols above, possible values (value sets) etc.

Paradata (production data, SCB and others)

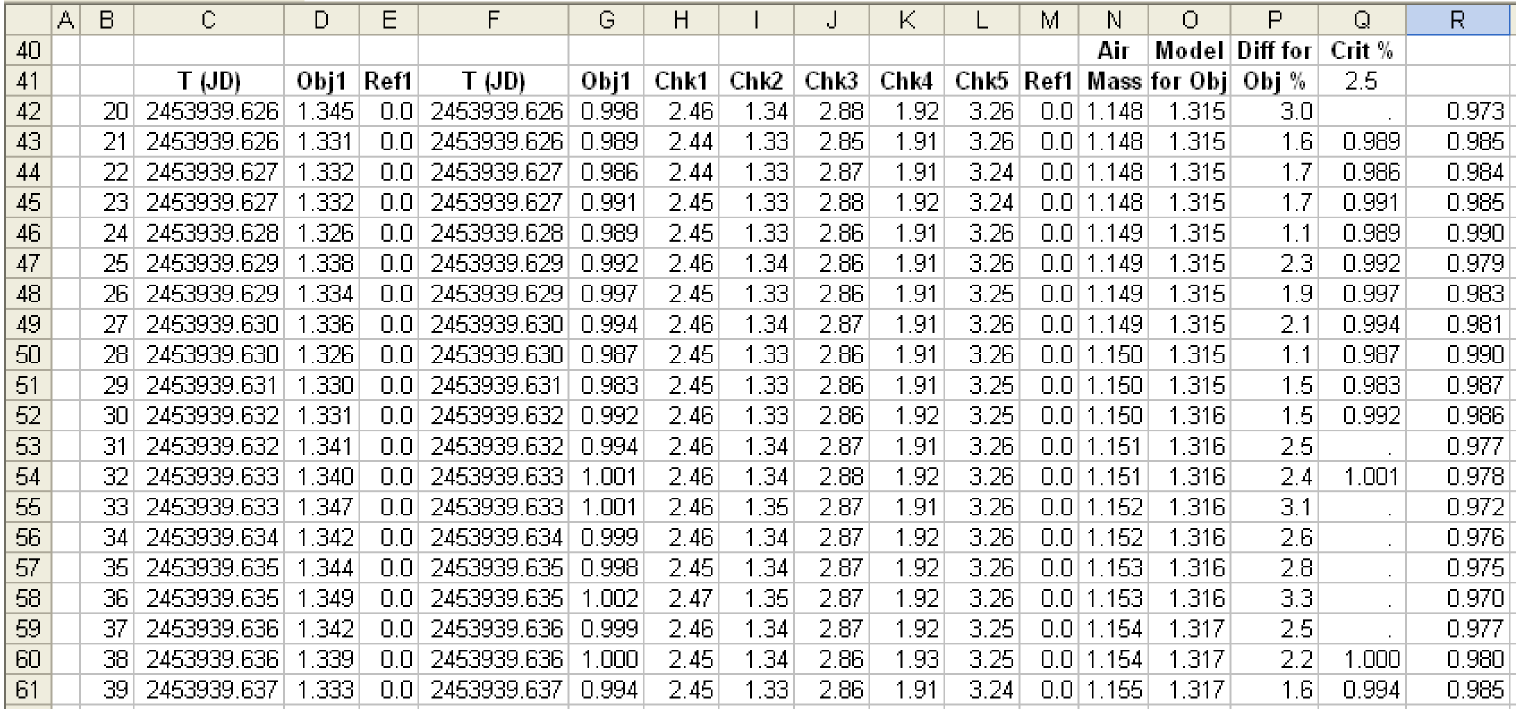
- Data about the data collection, processing etc.
- e.g. 4 T 11 62b

”**Metaparadata**”

- information about the paradata …

Statistics

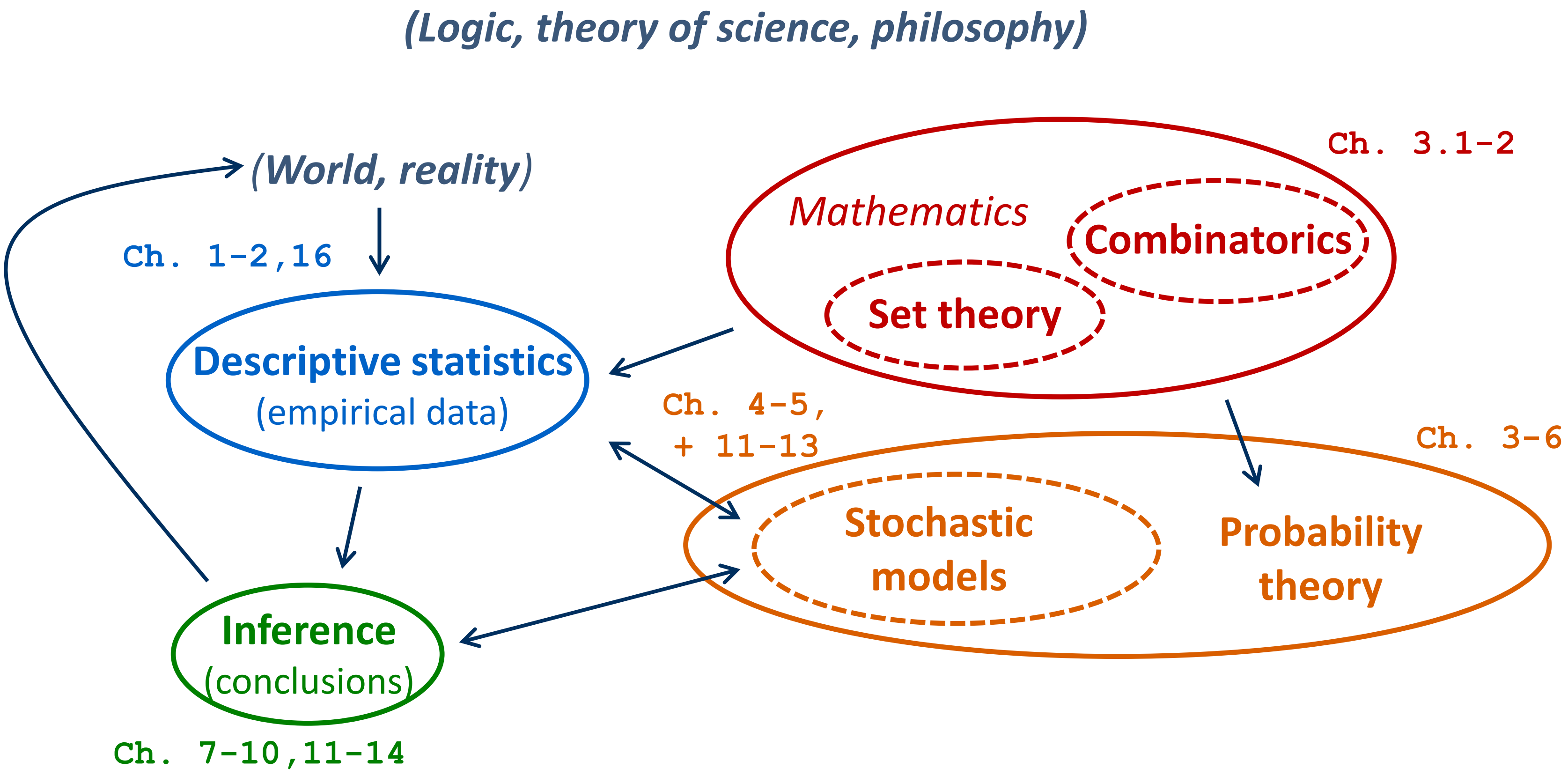
- To make this…



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
40														**Air**	**Model**	**Diff for**	**Crit %**	
41			**T (JD)**	**Obj1**	**Ref1**	**T (JD)**	**Obj1**	**Chk1**	**Chk2**	**Chk3**	**Chk4**	**Chk5**	**Ref1**	**Mass**	**for Obj**	**Obj %**	2.5	
42		20	2453939.626	1.345	0.0	2453939.626	0.998	2.46	1.34	2.88	1.92	3.26	0.0	1.148	1.315	3.0	.	0.973
43		21	2453939.626	1.331	0.0	2453939.626	0.989	2.44	1.33	2.85	1.91	3.26	0.0	1.148	1.315	1.6	0.989	0.985
44		22	2453939.627	1.332	0.0	2453939.627	0.986	2.44	1.33	2.87	1.91	3.24	0.0	1.148	1.315	1.7	0.986	0.984
45		23	2453939.627	1.332	0.0	2453939.627	0.991	2.45	1.33	2.88	1.92	3.24	0.0	1.148	1.315	1.7	0.991	0.985
46		24	2453939.628	1.326	0.0	2453939.628	0.989	2.45	1.33	2.86	1.91	3.26	0.0	1.149	1.315	1.1	0.989	0.990
47		25	2453939.629	1.338	0.0	2453939.629	0.992	2.46	1.34	2.86	1.91	3.26	0.0	1.149	1.315	2.3	0.992	0.979
48		26	2453939.629	1.334	0.0	2453939.629	0.997	2.45	1.33	2.86	1.91	3.25	0.0	1.149	1.315	1.9	0.997	0.983
49		27	2453939.630	1.336	0.0	2453939.630	0.994	2.46	1.34	2.87	1.91	3.26	0.0	1.149	1.315	2.1	0.994	0.981
50		28	2453939.630	1.326	0.0	2453939.630	0.987	2.45	1.33	2.86	1.91	3.26	0.0	1.150	1.315	1.1	0.987	0.990
51		29	2453939.631	1.330	0.0	2453939.631	0.983	2.45	1.33	2.86	1.91	3.25	0.0	1.150	1.315	1.5	0.983	0.987
52		30	2453939.632	1.331	0.0	2453939.632	0.992	2.46	1.33	2.86	1.92	3.25	0.0	1.150	1.316	1.5	0.992	0.986
53		31	2453939.632	1.341	0.0	2453939.632	0.994	2.46	1.34	2.87	1.91	3.26	0.0	1.151	1.316	2.5	.	0.977
54		32	2453939.633	1.340	0.0	2453939.633	1.001	2.46	1.34	2.88	1.92	3.26	0.0	1.151	1.316	2.4	1.001	0.978
55		33	2453939.633	1.347	0.0	2453939.633	1.001	2.46	1.35	2.87	1.91	3.26	0.0	1.152	1.316	3.1	.	0.972
56		34	2453939.634	1.342	0.0	2453939.634	0.999	2.46	1.34	2.87	1.92	3.26	0.0	1.152	1.316	2.6	.	0.976
57		35	2453939.635	1.344	0.0	2453939.635	0.998	2.45	1.34	2.87	1.92	3.26	0.0	1.153	1.316	2.8	.	0.975
58		36	2453939.635	1.349	0.0	2453939.635	1.002	2.47	1.35	2.87	1.92	3.26	0.0	1.153	1.316	3.3	.	0.970
59		37	2453939.636	1.342	0.0	2453939.636	0.999	2.46	1.34	2.87	1.92	3.25	0.0	1.154	1.317	2.5	.	0.977
60		38	2453939.636	1.339	0.0	2453939.636	1.000	2.45	1.34	2.86	1.93	3.25	0.0	1.154	1.317	2.2	1.000	0.980
61		39	2453939.637	1.333	0.0	2453939.637	0.994	2.45	1.33	2.86	1.91	3.24	0.0	1.155	1.317	1.6	0.994	0.985

… a little more comprehensible, understandable and useful

Structuring the course contents


(Logic, theory of science, philosophy)
(World, reality)
Ch. 1-2,16
Descriptive statistics
(empirical data)
Mathematics
Combinatorics
Set theory
Ch. 3.1-2
Ch. 4-5,
+ 11-13
Ch. 3-6
Stochastic
models
Probability
theory
Inference
(conclusions)
Ch. 7-10,11-14