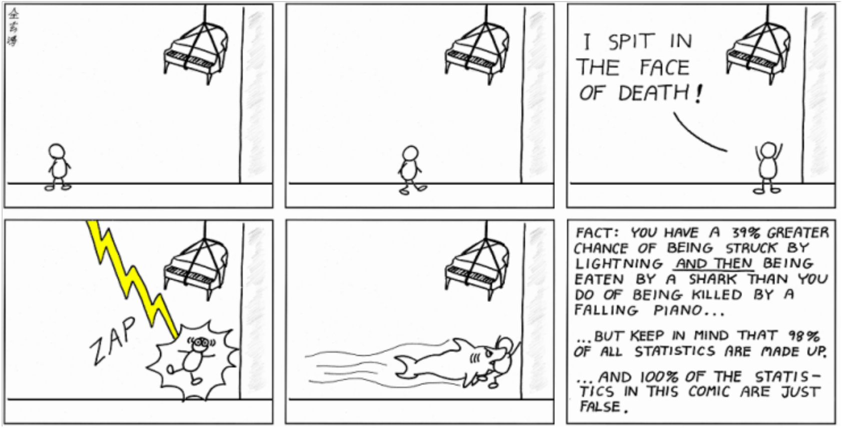
Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Applied Statistics

Introduction, Defining the Data

Release FS24

Introduction


I SPIT IN THE FACE OF DEATH!
ZAP
FACT: YOU HAVE A 39% GREATER CHANCE OF BEING STRUCK BY LIGHTNING AND THEN BEING EATEN BY A SHARK THAN YOU DO OF BEING KILLED BY A FALLING PIANO...
...BUT KEEP IN MIND THAT 98% OF ALL STATISTICS ARE MADE UP.
...AND 100% OF THE STATISTICS IN THIS COMIC ARE JUST FALSE.

(xkcd.com)

What is statistics?

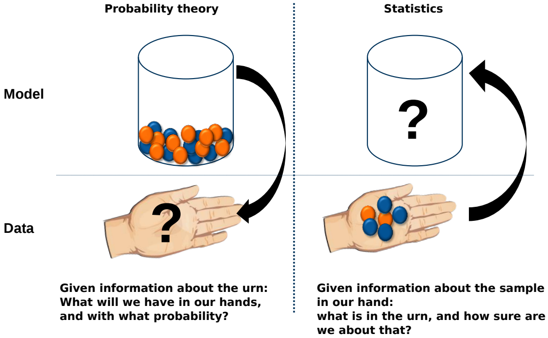
“Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. (. . .) Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments.” — Wikipedia

Statistics plays a key role in every scientific study from the very beginning (study plan) to the very end (interpretation).

What is statistics?

- Statistics is closely linked to probability theory
- Aim of probability theory: **modeling phenomena with uncertainty**
- Aim of statistics: performing **inference for probabilistic models**
- Sources of uncertainty:
  - variety of “samples” (e.g., individuals)
  - missing control of variables influencing outcome
  - missing knowledge

Statistics and probability theory

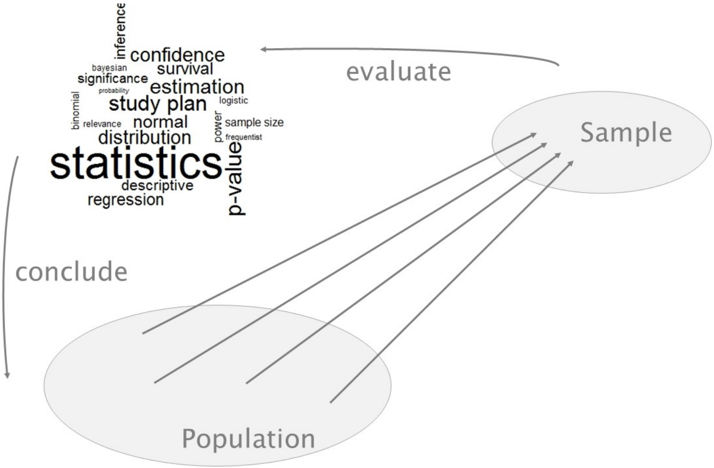


Given information about the urn:
What will we have in our hands,
and with what probability?

Given information about the sample
in our hand:
what is in the urn, and how sure are
we about that?

(Source: Meier (2014))

What is statistics?



evaluate

Sample

conclude

Population

Learning objectives

A data engineer should be able to . . .

- . . . actively join the planning phase of a scientific study
- . . . perform simple statistical analyses
- . . . apply appropriate software, such as `python`
- . . . understand and interpret the most common statistical methods in the scientific literature

To reach this goal:
Common sense is more important than a strong mathematical background!

What do you expect from the course?

- Let us ask **ChatGPT**
- What is the actual plan? **BFH**

Help shaping the course! What topics are you interested in?

Defining the Data

Learning objectives

Get to know important statistics vocabulary!

- Population and sample
- Different types of variables: numerical, categorical, binary
- Outcome and explanatory variables

Population and Sample

Population

The population is the totality of all individuals for which conclusions should be made.

An accurately defined group, e.g.

- all data engineer students at the BFH
- all data engineer students in Switzerland
- all data engineer students in the world
- all data engineer students in the world at present and in the future

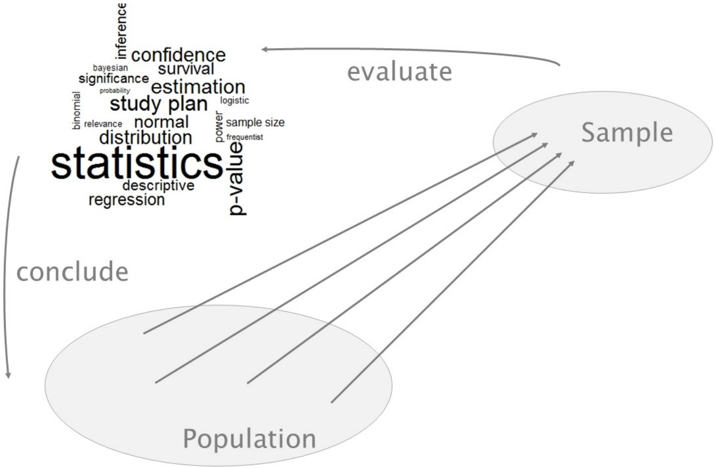
Population and Sample

Sample

A sample of a population is the set of individuals that are actually observed.

Ideally a random subset of the population
$\Longrightarrow$ **sampling variation**

Population and Sample



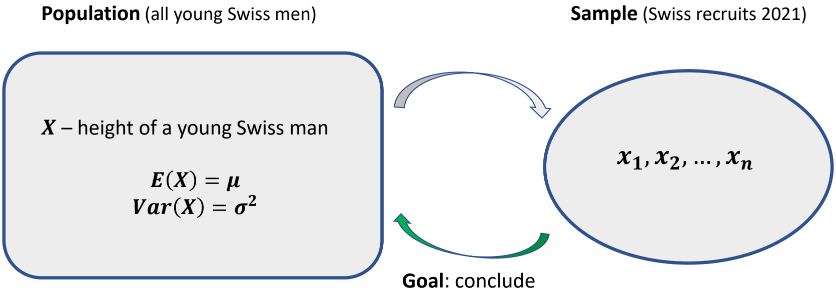
evaluate

Sample

conclude

Population

Example: height of a young Swiss man



Population (all young Swiss men)

Sample (Swiss recruits 2021)

$\boldsymbol{X}$ – height of a young Swiss man

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

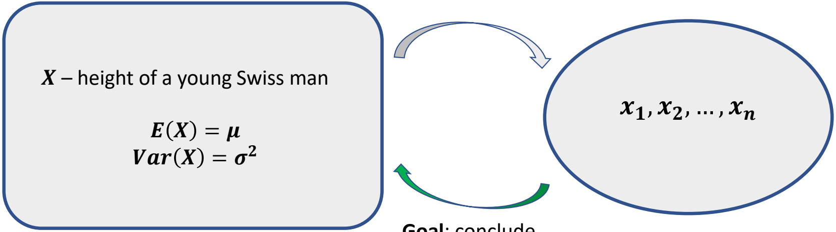
$x_1, x_2, \dots, x_n$

Goal: conclude

Example: height of a young Swiss man

Population (all young Swiss men)

Sample (Swiss recruits 2021)



$\boldsymbol{X}$ – height of a young Swiss man

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

$$x_1, x_2, \dots, x_n$$

Goal: conclude

estimate

μ - population mean ← **$\bar{x}$ - sample mean**

σ^2 - population variance ← **s_x^2 - sample variance**

Types of variable

A **variable** is an aspect of an individual in the sample that is measured or recorded.

E.g. data from a survival study after diagnosis of tuberculosis:

Id	Hospital	Age	Sex	Test result	6 months survival
001	1	57	M	Positive	yes
002	1	42	M	Positive	yes
003	1	51	F	Positive	no
004	2	64	F	Uncertain	yes
005	2	28	M	Negative	yes
006	3	37	M	Positive	yes

Types of variable

Types of variable

A first step in choosing how best to display and analyse data is to classify the variables into their different types – different types of variables ask for different methods.

Types of variable

We differentiate the following types of variable:

- **Categorical**:
  - No numerical interpretation
  - Subtypes:
    - **binary variable** (only two groups)
    - **nominal categorical variable** (no natural ordering)
    - **ordered categorical variable** (natural ordering)
  - E.g. sex, place of birth, pain scale (low – medium – high)
- **Numerical**:
  - Figures with numerical interpretation
  - Subtypes:
    - **continuous numeric variable**
    - **discrete numeric variable**
  - E.g. weight, number of adverse events

Tuberculosis example

Id	Hospital	Age	Sex	Test result	6 months survival
001	1	57	M	Positive	yes
002	1	42	M	Positive	yes
003	1	51	F	Positive	no
004	2	64	F	Uncertain	yes
005	2	28	M	Negative	yes
006	3	37	M	Positive	yes

Variable	Categories	Type of variable
Hospital	1, 2, 3	nominal categorical
Age	all natural numbers	numerical (continuous)
Sex	male, female	binary
Test result	Neg., Unc., Pos.	ordered categorical
6 months survival	yes, no	binary

Outcome and explanatory variables

Outcome variable

- Focus of attention
- The variable whose variation and occurrence we are seeking to understand
- The variable we try to model
- **Its type defines the appropriate statistical method**

Explanatory variables

- Variables that (may) explain occurrence of the outcome variable
- Goal: Try to quantify their influence on the outcome variable
- Common assumption: Explanatory variables have *no* uncertainty

Outcome and explanatory variables: Different vocabulary

Outcome variable	Explanatory variable
Response variable	Exposure variable
Dependent variable	Independent variable
y-variable	x-variable
Case-control group	Treatment group

Outcome and explanatory variables: Examples

Outcome variable	Explanatory variable
Baby born with low birth weight (yes/no)	Mother smoked during pregnancy (yes/no)
Anthropometric status at 1 year of age (numeric score)	Duration of exclusive breastfeeding (weeks)
Number of diarrhoea episodes experienced in a year	Access to clean water supply (yes/no)
Child develops leukaemia (yes/no)	Proximity to nuclear power station (km)
Survival time following diagnosis of lung cancer (months)	Socio-economic status (6 groups)

References

Lukas Meier. Statistik und Wahrscheinlichkeitsrechnung. *Lecture notes*, 2014.