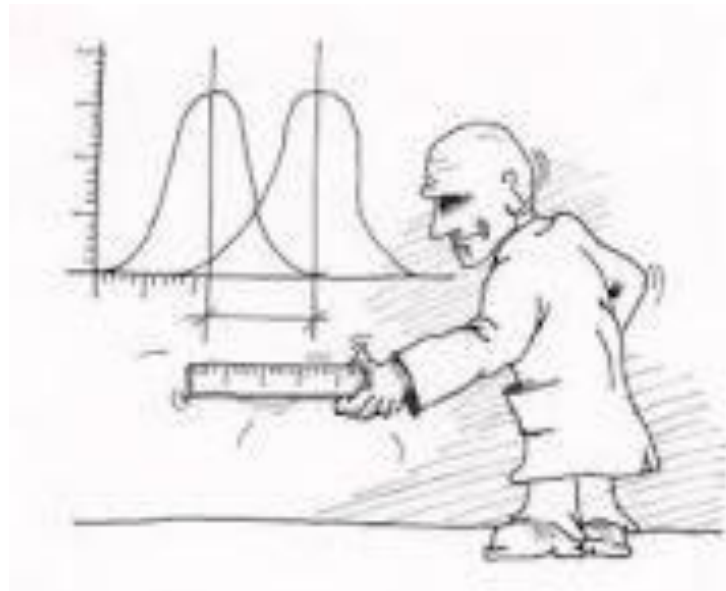


Biostatistics

Formalia

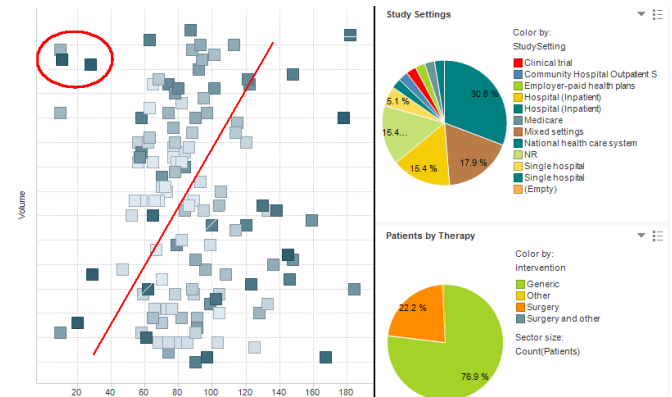
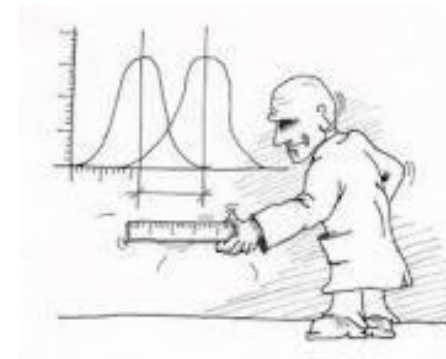
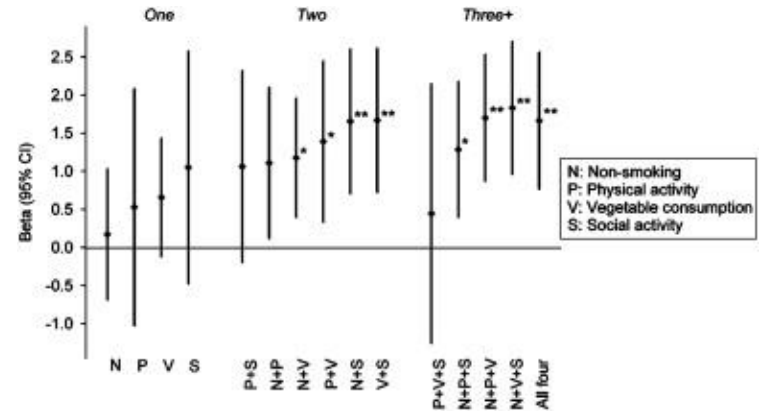
- Lectures: Tuesdays , 10:10-11:45 , HG E21, Beate Sick
- Exercises: Tuesdays, 16:15-17:00, HG D7.1, Lisa Herzog
- Material: <https://bsick.github.io/Biostatistics-Fall-2018/>



Goal of the module "Biostatistics"

■ Goal is to get more confident...

- in the most widely used statistical methods
- in reading data analysis sections in in scientific articles, especially in medical or biological journals
- Visualizing and analyzing own data



Biostatistics

Topics

Exam is on these topics, written, 45 minutes

- data visualization
 - basic terms and summary statistics
 - study types, confounding
 - diagnostic tests
 - models/distribution-types
 - parameter estimation
 - testing, confidence intervals, p-values
 - linear regression
-
- reliability analysis
 - logistic regression
 - outlook on more advanced or modern regression methods

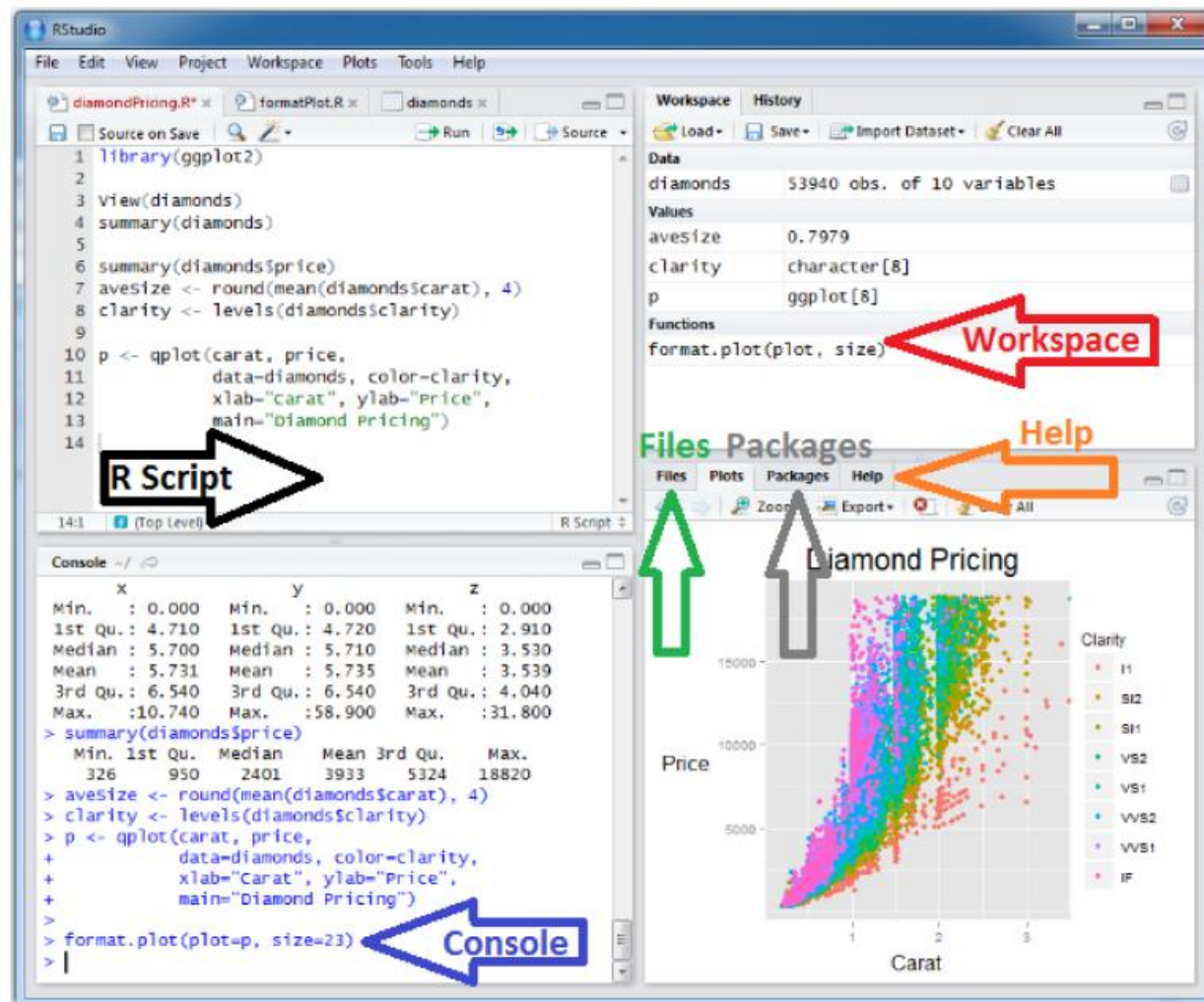
We use R for performing statistical data analysis

Recommended environment: RStudio

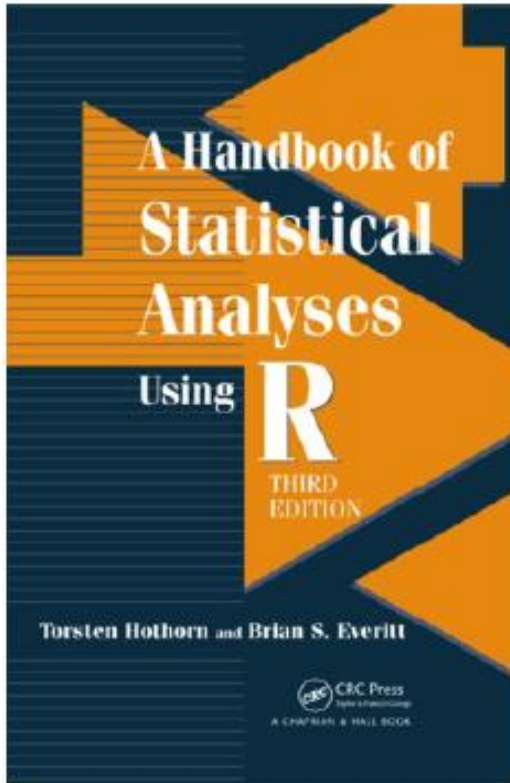


Main reasons:

- open source
- powerful
- wide spread
- reproducible
- transparent



Literature: no book needs to be purchased



Selected chapters from “A Handbook of Statistical Analysis Using R” by Torsten Hothorn, UZH, serve as recommended readings for this course and are provided on the course website. In addition some selected method articles will be recommended.

The R package HSAUR3 provides the selected chapters as PDF besides all data sets, examples and R code

<http://CRAN.R-project.org/package=HSAUR3>

Torsten.Hothorn@R-project.org

My background: Important stations

Heidelberg



Study of physics
& mathematics

Lausanne: UNIL, DAFL



Head of bioinformatics &
biostatistics at DNA
Array Facility UNIL

Zürich: ETH



PhD and Postdoc
Contract lecturer

Teaching
in Berlin



Winterthur: ZHAW, SoE



Prof. for applied
statistics & Scientist

Providing trainings for life
science researcher

Basel: OncoScore



Biomarker detection
CAS pharmaceutical Med.

UZH: EBPI, Biostatistics



Scientific collaborator,
consultant, and lecturer
in the field of biostatistics
and medical research

Biostatistics for Medical Physicists

Week 1

Topics this week:

- **Goals of this module**
- **Data types**
- **methods for uni-variate data visualization**
- **terms and key numbers**

Research Article Example: Paper on hyperactivity form McCann et al.

① Food additives and hyperactive behaviour in 3-year-old and 8/9-year-old children in the community: a randomised, double-blinded, placebo-controlled trial

Donna McCann, Angelina Barrett, Alison Cooper, Debbie Crumpler, Lindy Dalen, Kate Grimshaw, Elizabeth Kitchin, Kris Lok, Lucy Porteous, Emily Prince, Edmund Sonuga-Barke, John O Warner, Jim Stevenson

Summary

Lancet 2007; 370: 1560-67

Published Online

September 6, 2007

DOI:10.1016/S0140-

6736(07)61306-3

See [Comment](#) page 1524

See [Department of Error](#)
page 1542

School of Psychology

Background We undertook a randomised, double-blinded, placebo-controlled, crossover trial to test whether intake of artificial food colour and additives (AFCA) affected childhood behaviour.

Methods 153 3-year-old and 144 8/9-year-old children were included in the study. The challenge drink contained sodium benzoate and one of two AFCA mixes (A or B) or a placebo mix. The main outcome measure was a global hyperactivity aggregate (GHA), based on aggregated z-scores of observed behaviours and ratings by teachers and parents, plus, for 8/9-year-old children, a computerised test of attention. This clinical trial is registered with Current Controlled Trials (registration number ISRCTN74481308). Analysis was per protocol.

What does it mean?

Typical “table 1” in a medical research article

Table. Patient Clinical Characteristics

	All, n=68 (%)	No Recurrent Event Observed, n=54 (%)	Recurrent Event Observed, n=14 (%)	<i>P</i> Value
Demographic data				
Age, y (range)	65 (30–90)	64.7(30–88)	65.5 (47–90)	0.96
Men	47 (69)	38 (70.4)	9 (64.3)	0.75
Type of event				
TIA	5 (7.4)	4 (7.4)	1 (7.1)	0.6
Retinal ischemia	5 (7.4)	3 (5.6)	2 (14.3)	0.2
Stroke	58 (85.3)	47 (87)	11 (78.6)	0.2
Medical history, n (%)				
Smoking	29 (43)	20 (37)	9 (64.3)	0.21
Hypertension	49 (72)	37 (68.5)	12 (85.7)	0.32
Diabetes mellitus	13 (19)	7 (13.0)	6 (42.9)	0.02*
...				

Clinical characteristics of all 68 patients (all) and patient groups without or with ipsilateral recurrent ischemic event. Number (n) and percentage or median and IQR are shown. CAD indicates coronary artery disease; IQR, interquartile range; mRS, modified Rankin Scale; NIHSS, National Institute of Health Stroke Scale; pAOD, peripheral artery occlusive disease; TIA, transient ischemic attack; and TOAST, Trial of ORG 10172 in Acute Stroke Treatment classification scheme for stroke etiology.

**P* values <0.05 in Mann–Whitney *U* test or Fisher exact test.

Research Article Example: Paper on hyperactivity form McCann et al.

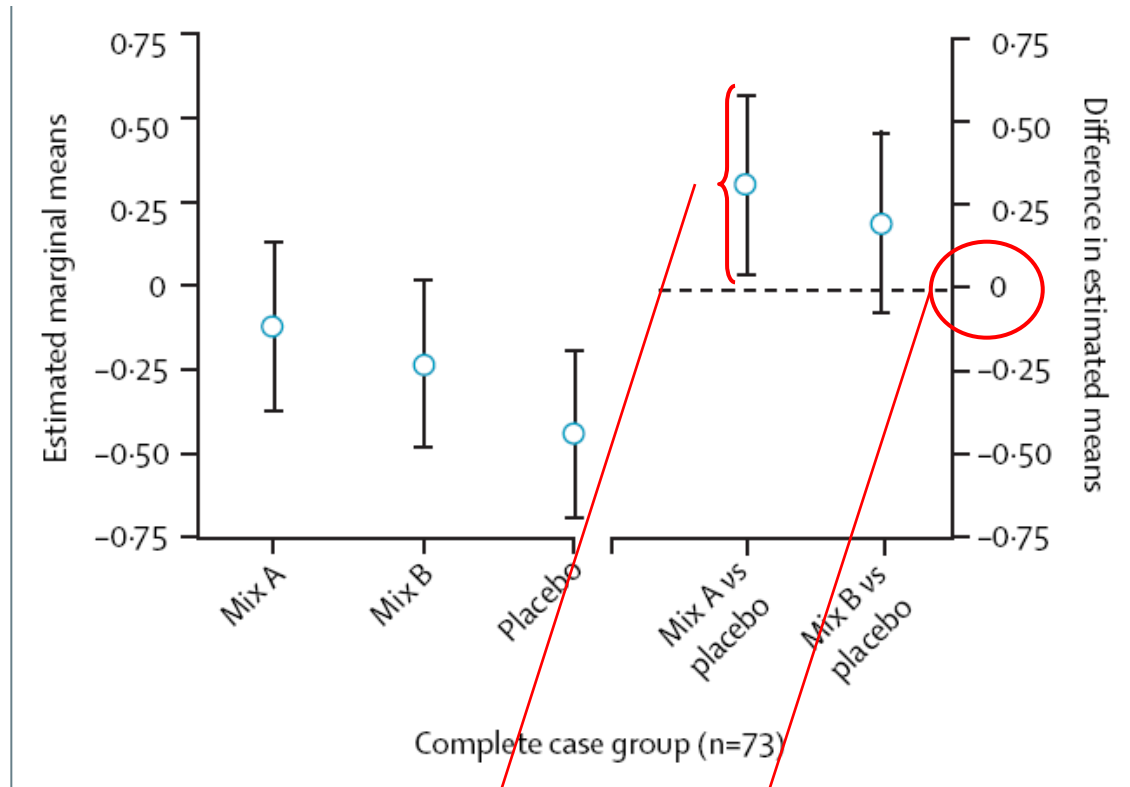


Figure 3: Estimated marginal means by challenge type and difference in estimated means in GHA under model 2 for 3-year-old children

What do bars indicate?
What is special with the 0-line?

Research Article Example: Paper on hyperactivity form McCann et al.

	Entire sample (n=140)	Group with $\geq 85\%$ consumption (n=130)
Model 1 Unadjusted		
Intercept	-0.31 (-0.49 to -0.13)*	-0.33 (-0.53 to -0.13)†
Challenge type		
Mix A vs placebo	0.20 (0.01 to 0.40)‡	0.24 (0.02 to 0.47)‡
Mix B vs placebo	0.16 (-0.04 to 0.35)	0.16 (-0.07 to 0.38)
Model 2 Adjusted		
Intercept	-0.54 (-0.89 to -0.18)*	-0.51 (-0.92 to -0.11)
Challenge type		
Mix A vs placebo	0.20 (0.01 to 0.39)‡	0.28 (0.05 to 0.51)‡
Mix B vs placebo	0.17 (-0.03 to 0.36)	0.19 (-0.04 to 0.41)

What does
“adjusted”
mean?

How is it
done?

In model 2, in

addition to challenge type, the effects of the following factors were adjusted for: week during study, sex, GHA in baseline week, number of additives in pretrial diet, maternal educational level, and social class.

What does *
mean?

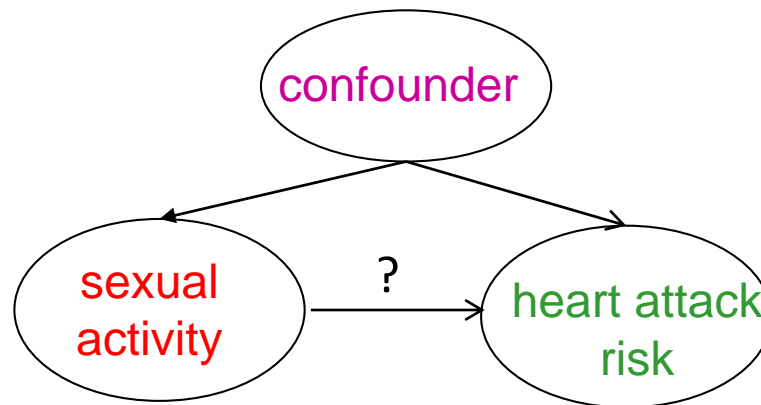
Example: Study in Caerphilly (Wales), 1979-2003

914 healthy men, between 45 and 95 years old, were chosen at random and followed over 10 years where they were interviewed e.g. about their sexual life. Moreover it was followed who suffered a heart attack in this period.

Result:

group	# men	# sexual active men	# sexual inactive men
all men	914	231	197
men suffering heart attack	11% 105	8% 19	17% 33

What can we conclude?



Why do we need statistics?

Data vary!

Samples are random!

„We need statistics to draw intelligent decisions in the presence of uncertainty.“

Numbers and Data

Numbers in mathematics	Number in data
exact	imprecise : random errors
certain	uncertain : “biased”, faked, accidental coarse error
Just a number	Need for interpretation
Two number are either equal or different	Two observations are normally not exactly equal, but are they significantly different?

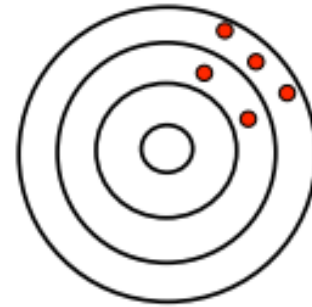
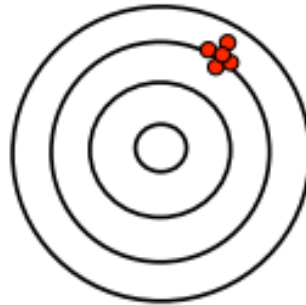
What kind of errors do we usually see?

Variance: by random errors

Small variance
Precise

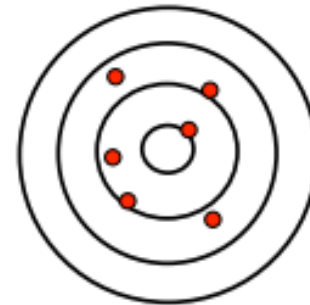
Large variance
Imprecise

Biased



Bias

Unbiased



Systematic error or bias, accidental coarse error

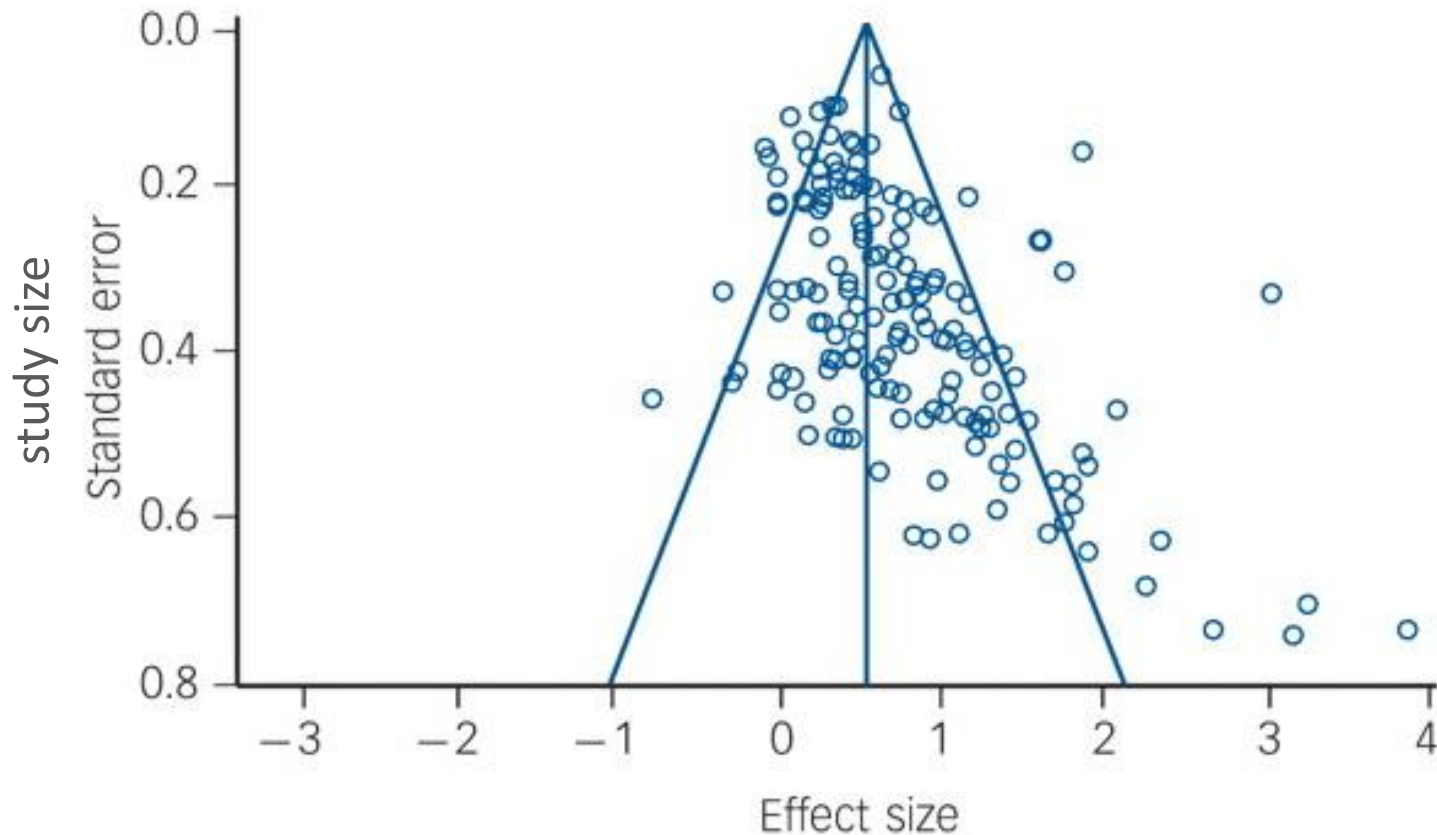


Wrong basic calibration



Accidentally shifted comma suggested wrongly high iron content in spinach

What is a publication bias?



Publication bias is the tendency to more likely publish results that are positive (i.e. showing a significant finding) than negative results, leading to a misleading bias in the overall published literature. This is often visible in a funnel plot, where smaller studies with higher standard error tend to report more often positive results than large studies which are published irrespectively of the result.

Willful error or fake: the case of Henrik Schön

- 2001 each week a paper, many in Nature or Science
• Proposed as youngest Max Plank direktor in Stuttgart
• Candidate for Nobelpreis
• ...
- 2002 Beasley Kommission...
- Withdrawl of PhD title because of dishonorable behaviour, but he did not fake during his PhD time in Konstanz



Der internationale Braunschweig-Preis geht an die Forschergruppe mit (von links) Jan Hendrik Schön, Christian Kloc (beide Bell Labs, Lucent Technologies, New Jersey) und Professor Bertram Batlogg (ETH Zürich) als Gruppenleiter.

picture:
ETH-Zürich

Fake: same noise implies same data -> data substitution

II. Data Substitution: Ambipolar triode characteristics

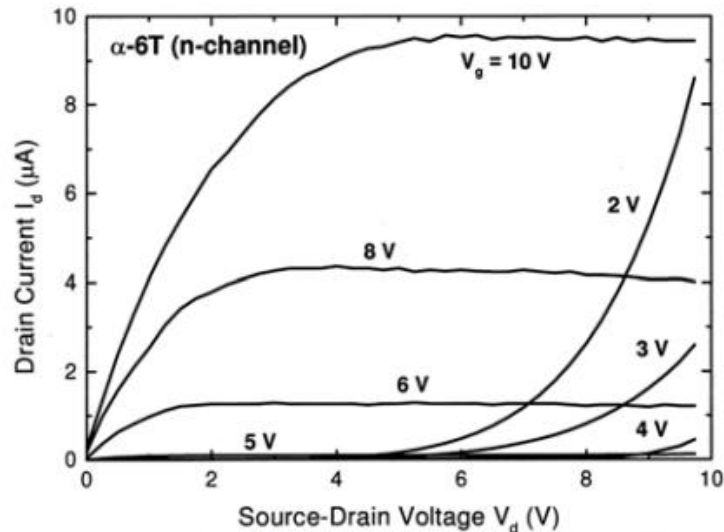


Figure 7. Triode characteristic from “LightEmitting” Paper (V), Fig. 1.: “alpha-sexithiophene (α -6T)”

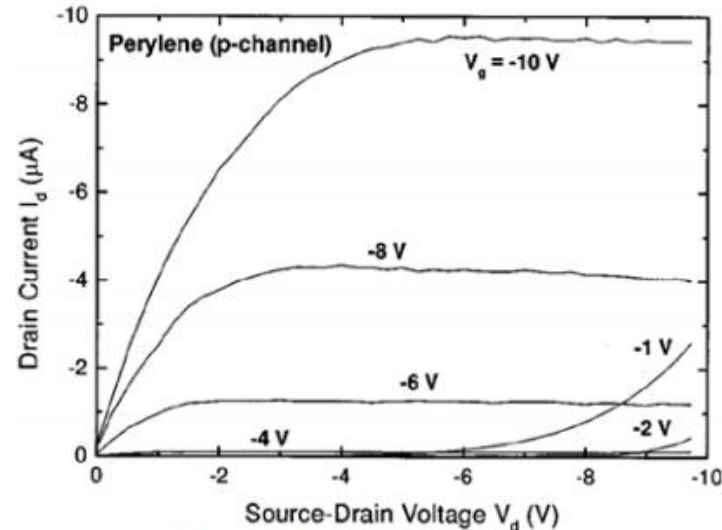


Figure 8. Triode characteristic from “Perylene” (VI), Fig. 2: “perylene”. Note the sign change from Figure 7. One curve is missing.

Same data, different effects, different publications

(V) “A Light-Emitting Field-Effect Transistor,” J. H. Schön, A. Dodabalapur, Ch. Kloc, and B. Batlogg, Science 290, 963 (November 3, 2000)

(VI) “Perylene: A promising organic field-effect transistor material,” J. H. Schön Ch. Kloc, and B. Batlogg, Appl. Phys. Lett. 77, 3776 (December 4, 2000)

Pitfall when faking: precision

Presented data

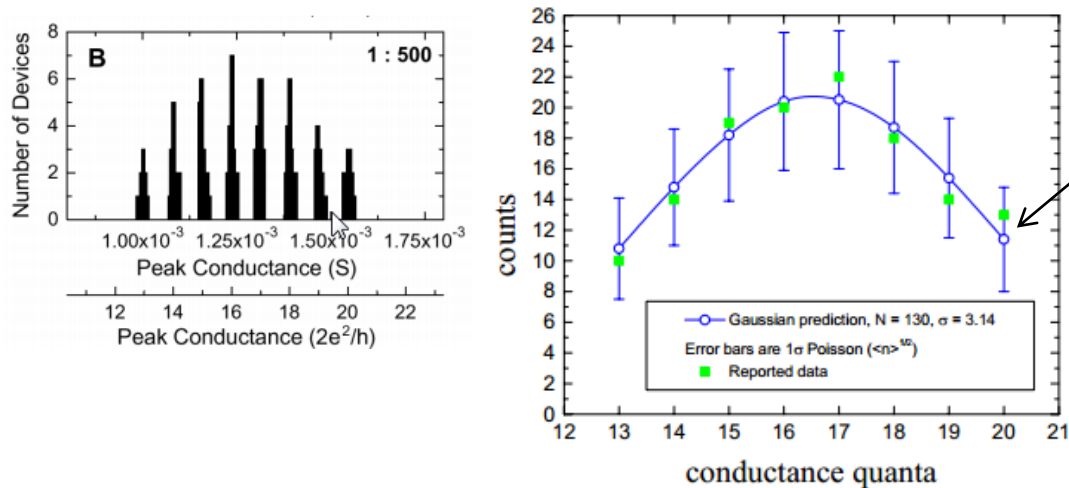


Figure 47. Comparison of reported number of devices in each conductance quantum bin with a fit to a Gaussian distribution. The reported data are extracted from original plotting data for Figure 46 from an electronic draft. The agreement exceeds the expected variation for such a small number of devices.

Blue: Theory

green: «observations»

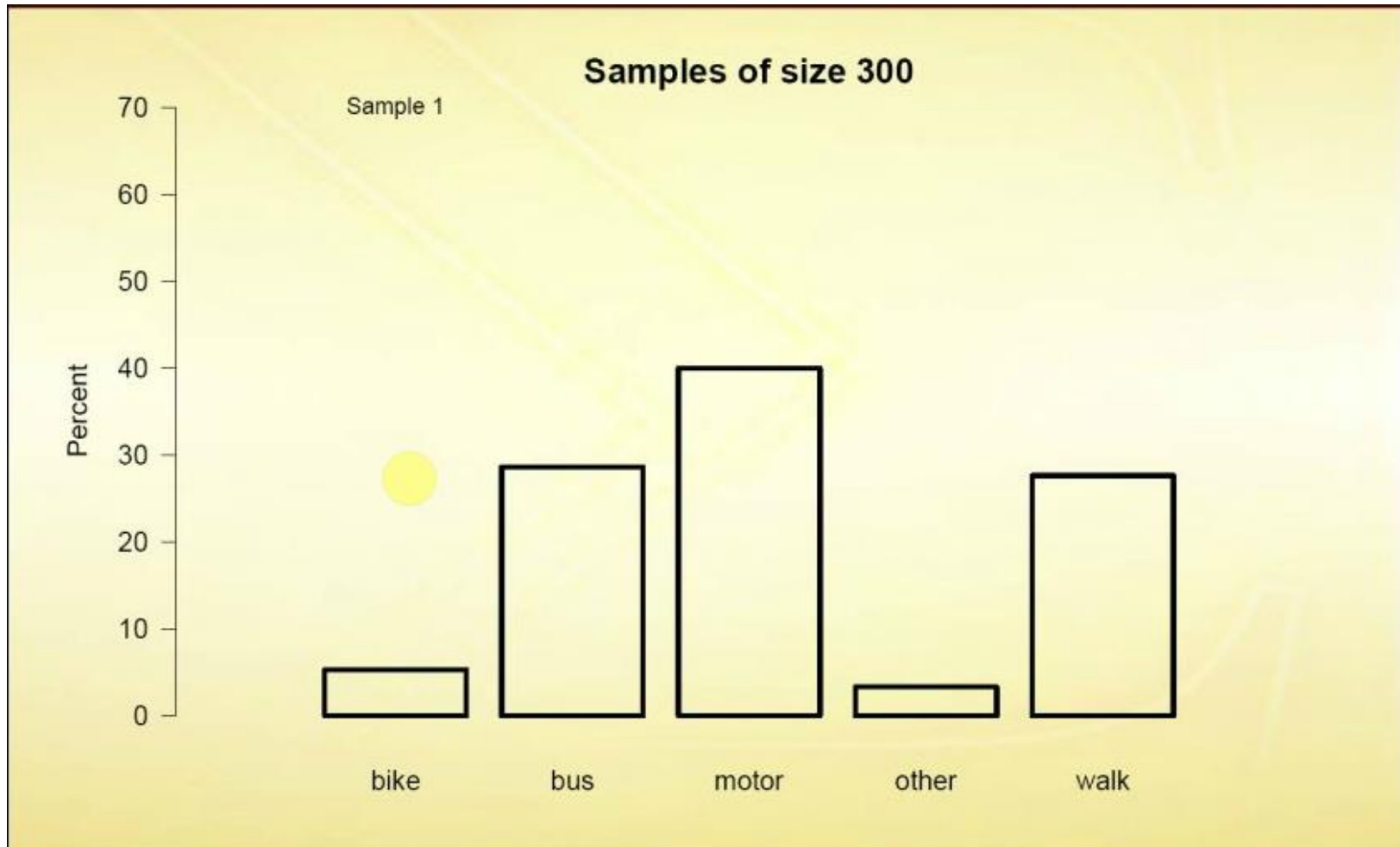
Blue standard error bar:
1/3 of all observations should
outside of the bars.

Probability to see such data by
chance: 0.0012:

“Field-Effect Modulation of the Conductance of Single Molecules,” Jan Hendrik Schön, Hong Meng, Zhenan Bao, Science 294, 2140 (December 7, 2001).

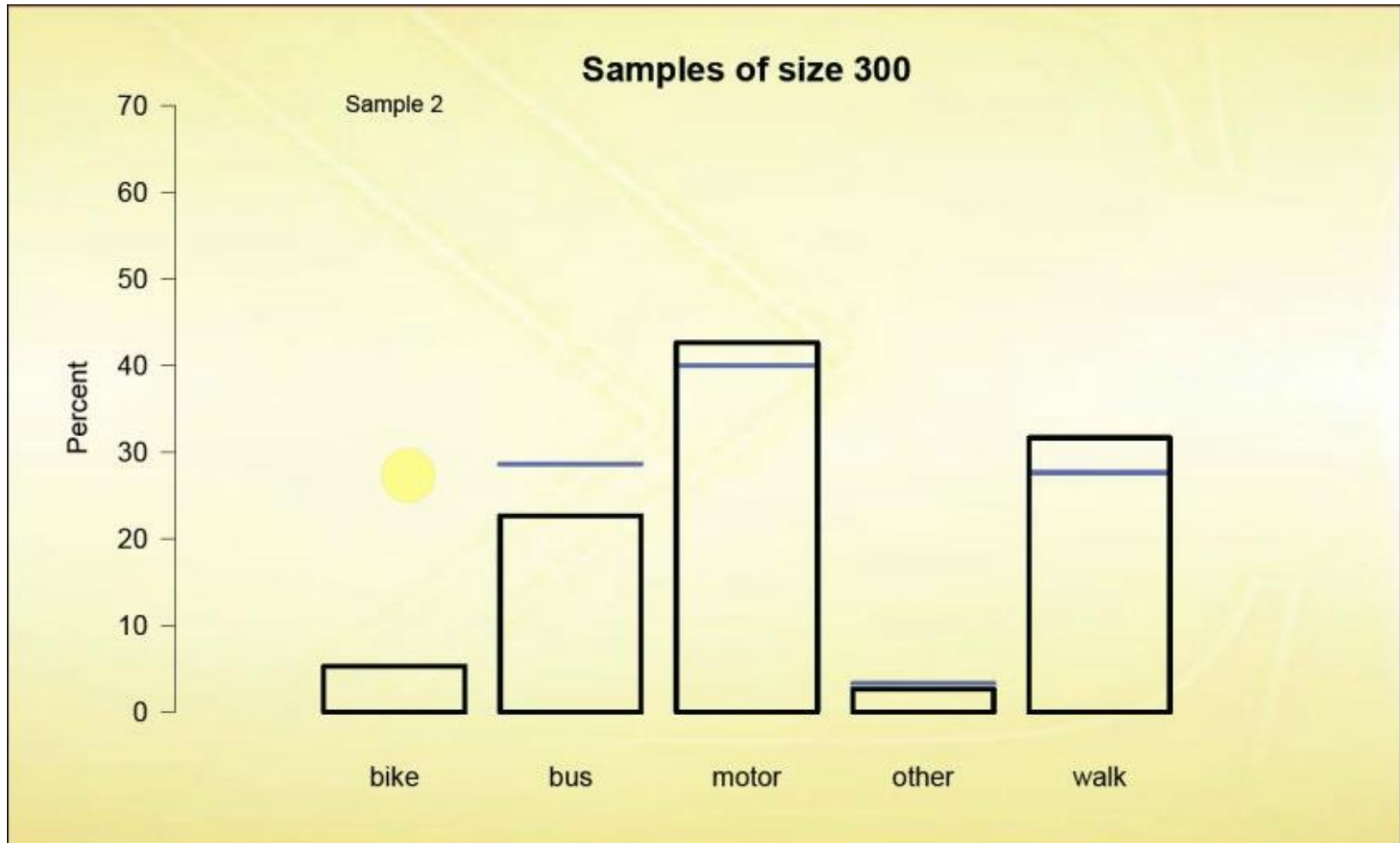
How reliable are the bar-heights in a barplot?

How do pupils get to school?



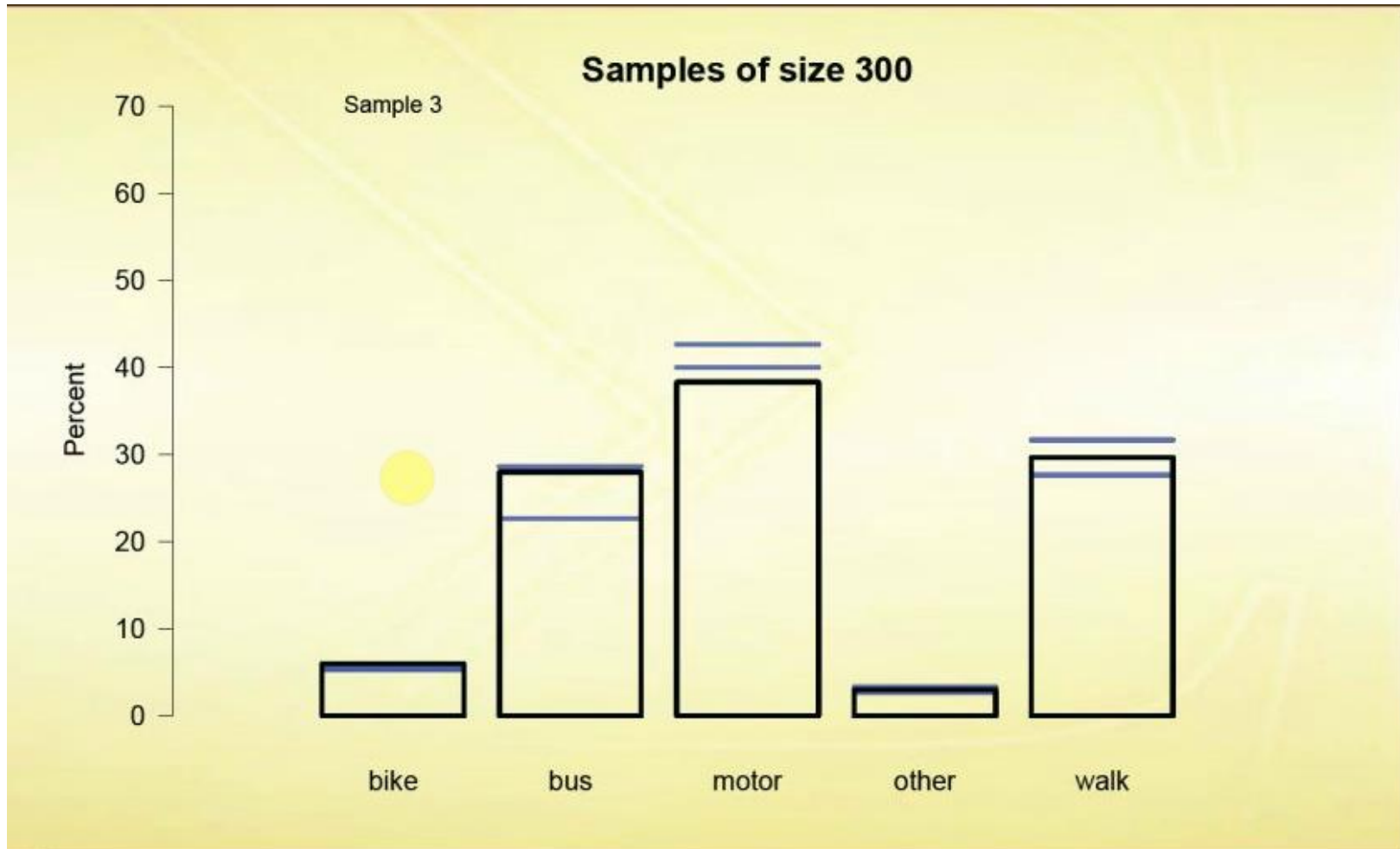
How reliable are the bar-heights in a barplot?

How do pupils get to school?



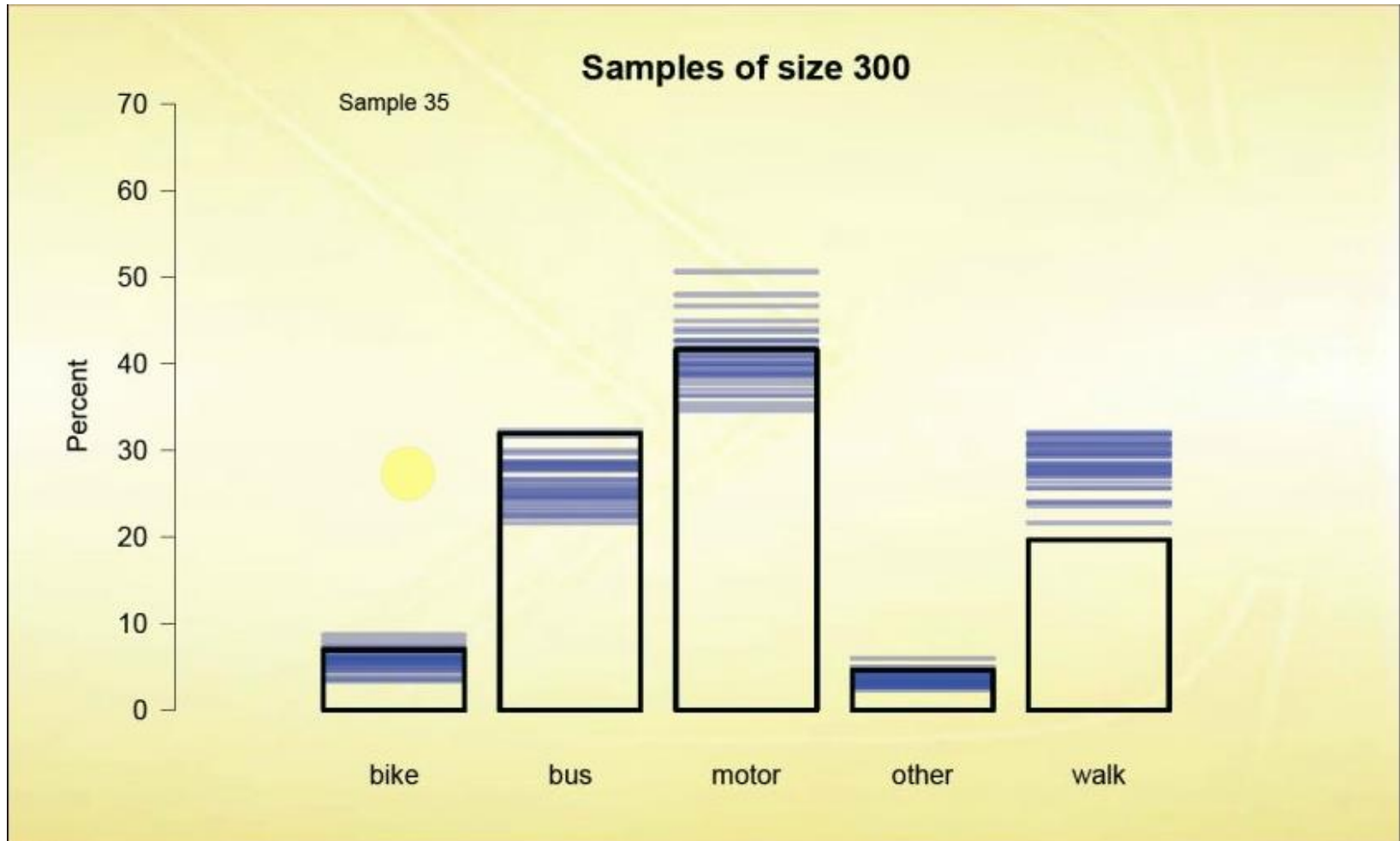
How reliable are the bar-heights in a barplot?

How do pupils get to school?



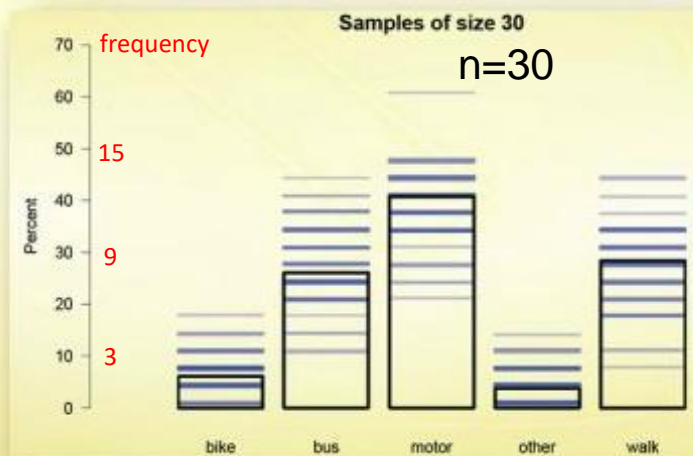
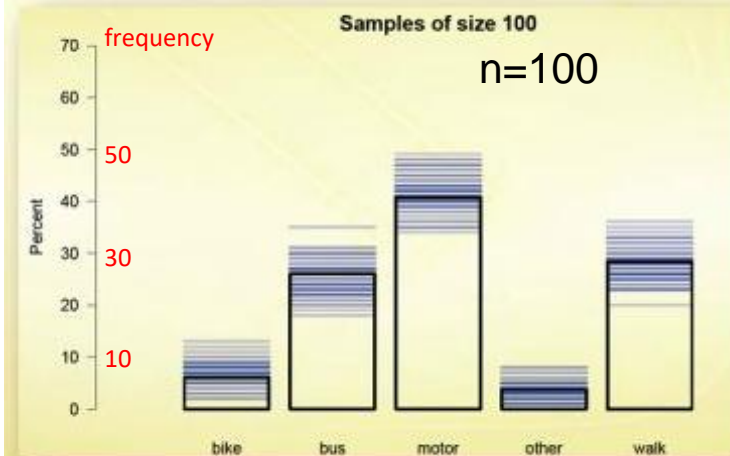
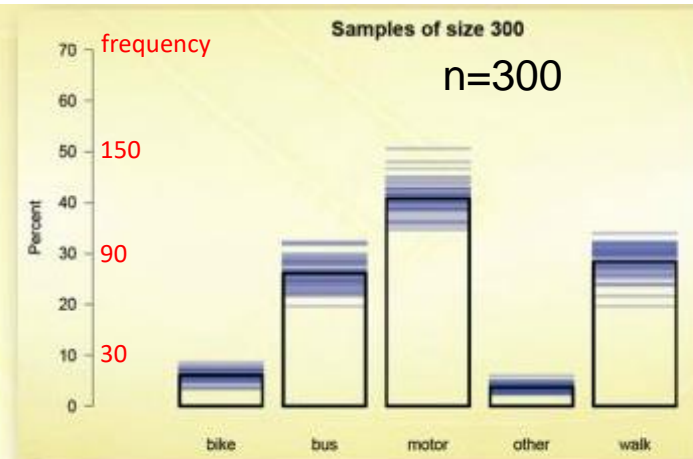
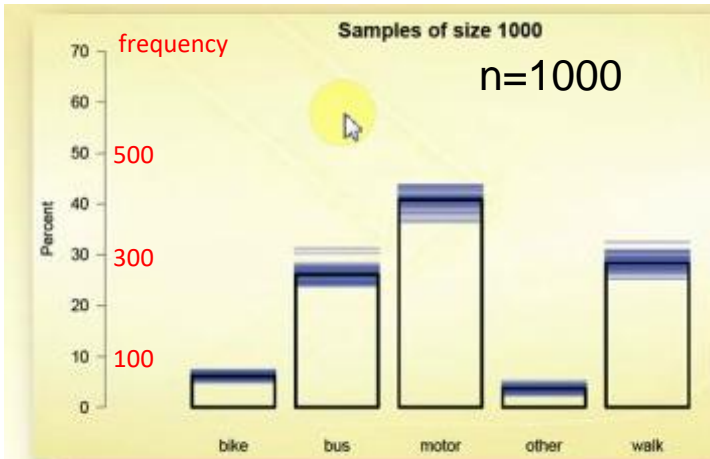
How reliable are the bar-heights in a barplot?

How do pupils get to school?



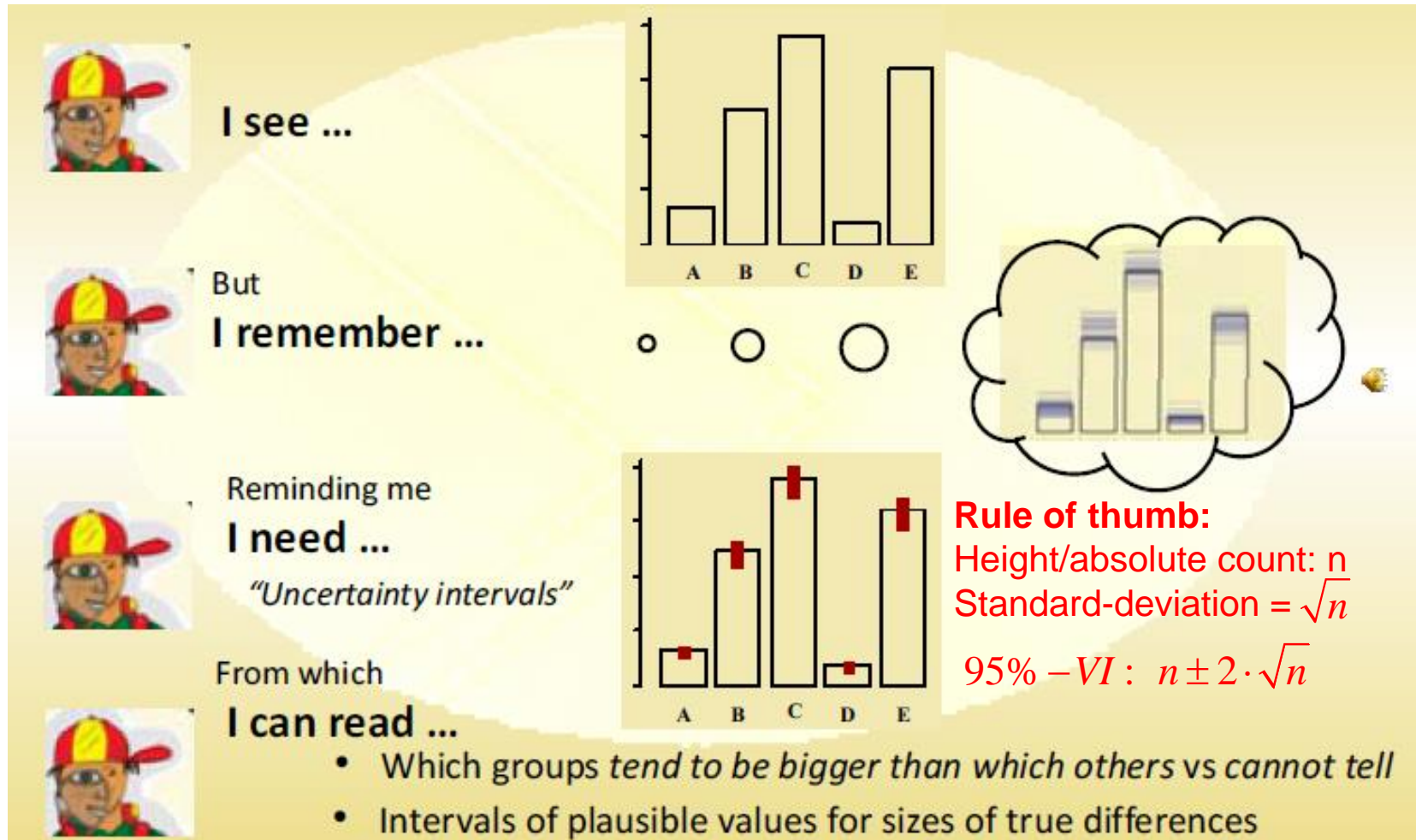
How reliable are the bar-heights in a barplot?

How do pupils get to school?



How reliable are the bar-heights in a barplot?

How do pupils get to school?



Fundamental terms

- **Feature or variable:**

Properties which can take different values

- Age of a patient
- Sex of a student
- braking distance from 100 km/h for different car types

- **Population:**

The complete set of all items that are relevant for the investigation

- All persons suffering a heart attack
- All at ETH inscribed students
- All uranium atoms

- **Sample:**

A subgroup of the entire population which have been selected in a certain manner (systematically, arbitrary, at random, stratified)

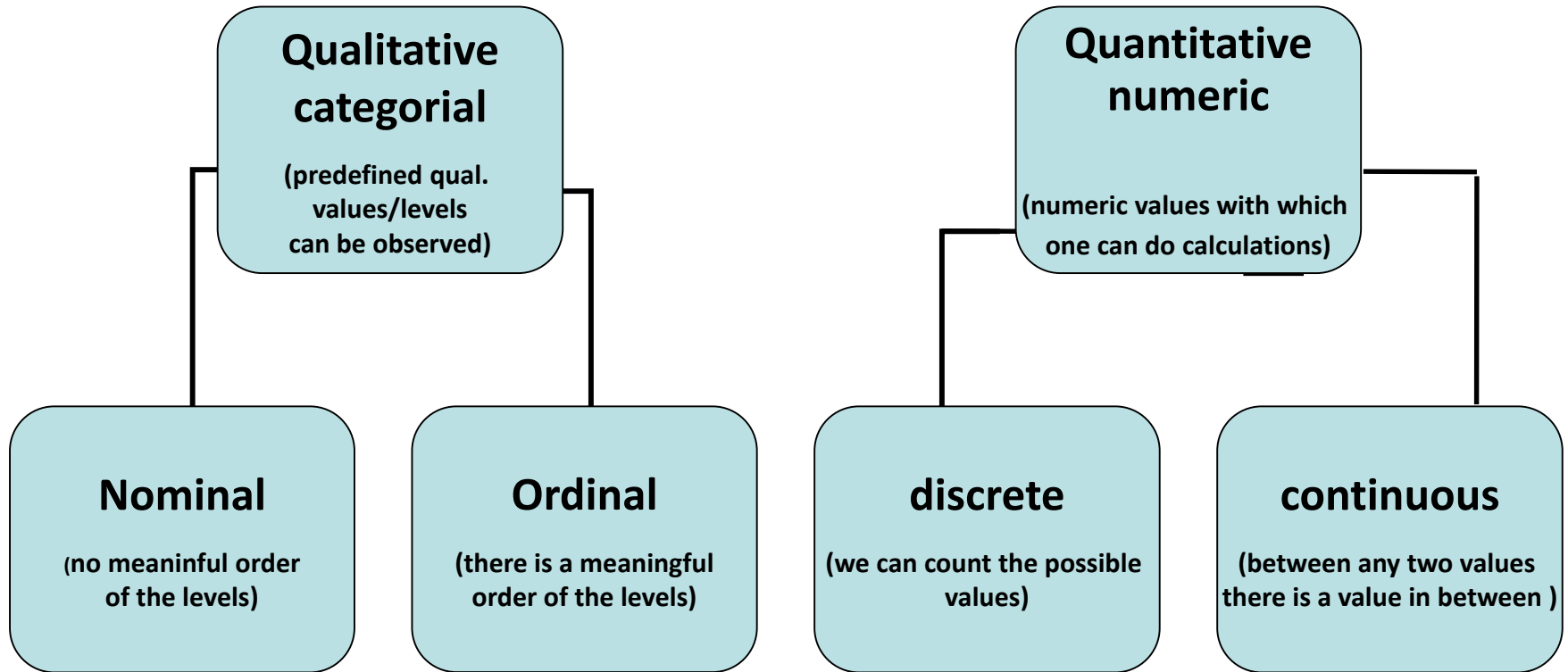
We can collect some data from each student

- 1) Age
- 2) Sex
- 3) Nationality
- 4) Height
- 5) Arm span
- 6) Number of siblings
- 7) Handedness
- 8) Rate the quality of the ETH Mensa food (0, 1, ..., 9)

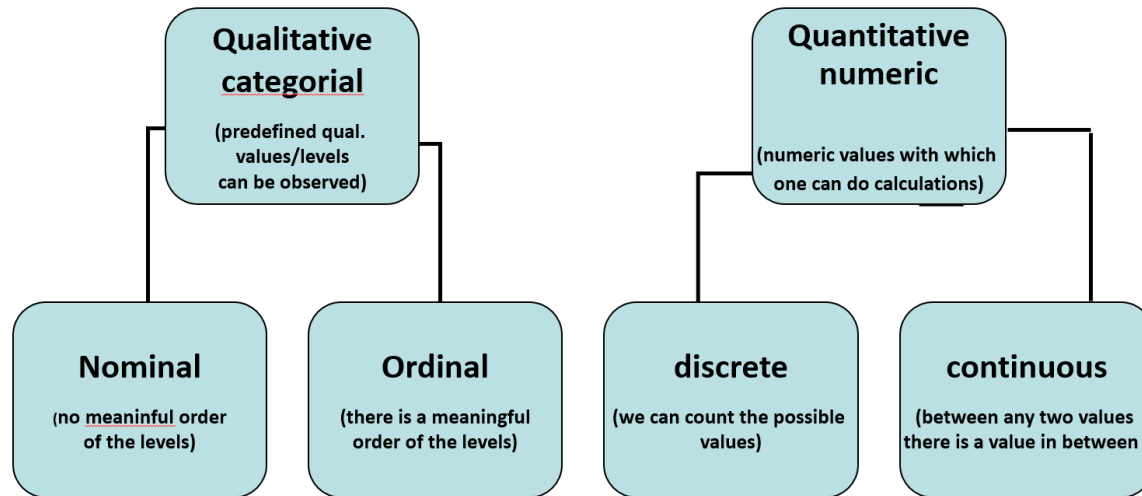
How do the type of collected data differ?

How could we visualize the answers from the class for each question?

There are different types of data



There are different types of data



- 1) Age
- 2) Sex
- 3) Nationality
- 4) Height
- 5) Arm span
- 6) Number of siblings
- 7) Handedness
- 8) Rate the quality of the ETH Mensa food (0, 1, ..., 9)