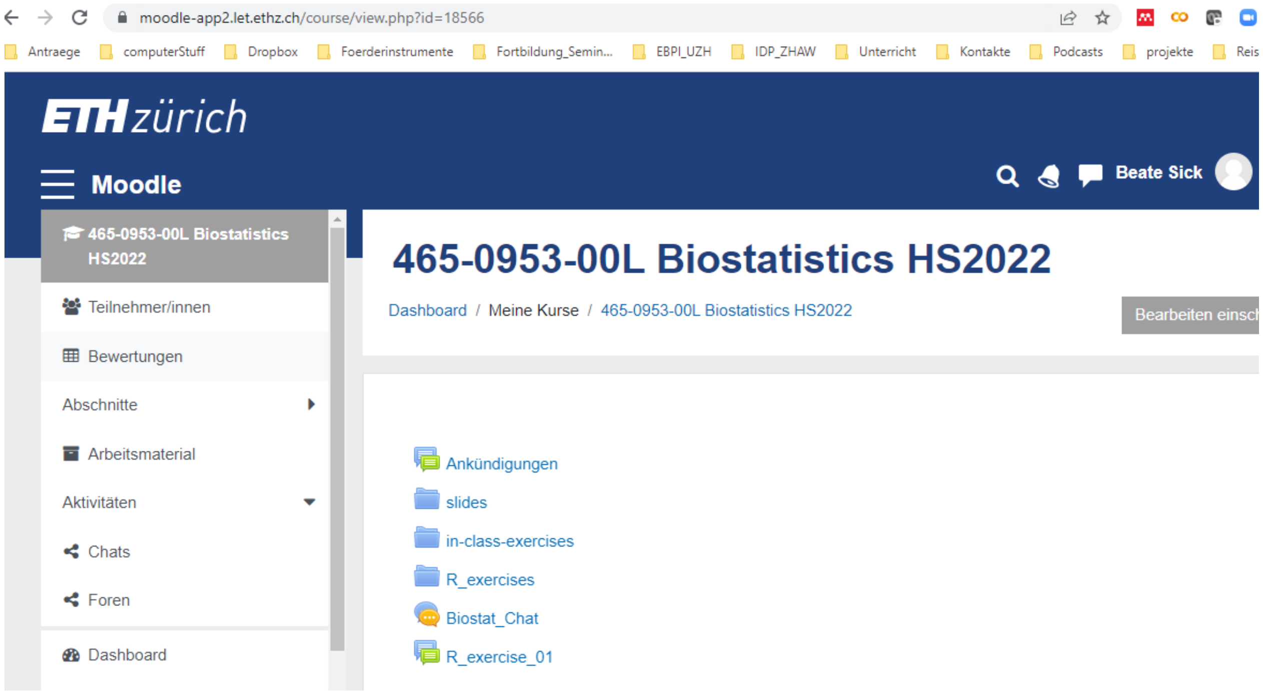
Biostatistics

- **Lectures: Tuesdays , 10:15-12:00 , ML F 36, Beate Sick**
- **Exercises: Tuesdays, 16:15-17:15 , online, Lisa Herzog**
  https://zhaw.zoom.us/j/68326457254?pwd=L0hsQjBtdHVrWXFTR21BUExRMDNGdz09
  Meeting-ID: 683 2645 7254
  Passwort: biobio

- **Material and Chats (no e-mail) via Moodle (Lectures and R-exercises):**
  https://moodle-app2.let.ethz.ch/course/view.php?id=18566


moodle-app2.let.ethz.ch/course/view.php?id=18566

465-0953-00L Biostatistics HS2022

Dashboard / Meine Kurse / 465-0953-00L Biostatistics HS2022

- Ankündigungen
- slides
- in-class-exercises
- R_exercises
- Biostat_Chat
- R_exercise_01

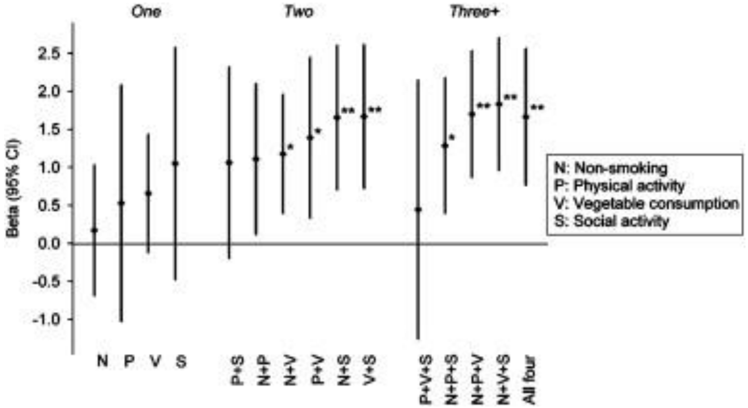
Goal of the module “Biostatistics”

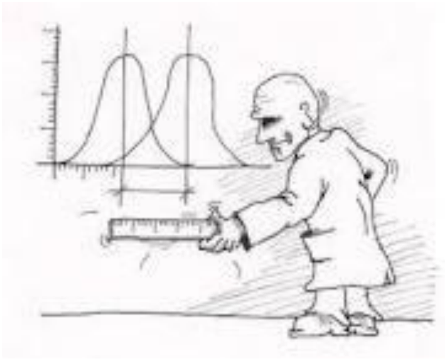
- **Goal is to get more confident…**

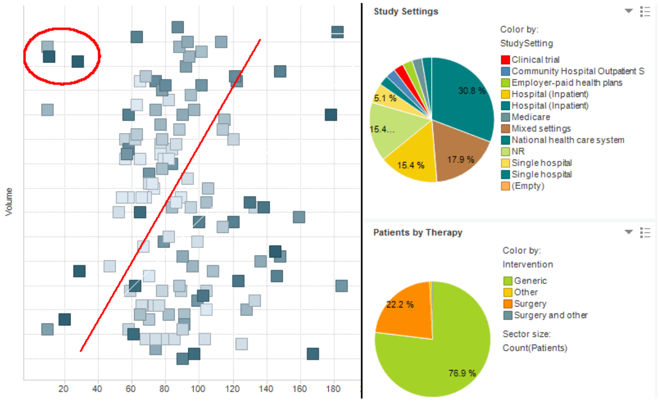
  - **in the most widely used statistical methods**

  - **in reading data analysis sections in in scientific articles, especially in medical or biological journals**

  - **Visualizing and analyzing own data**







Biostatistics

Topics

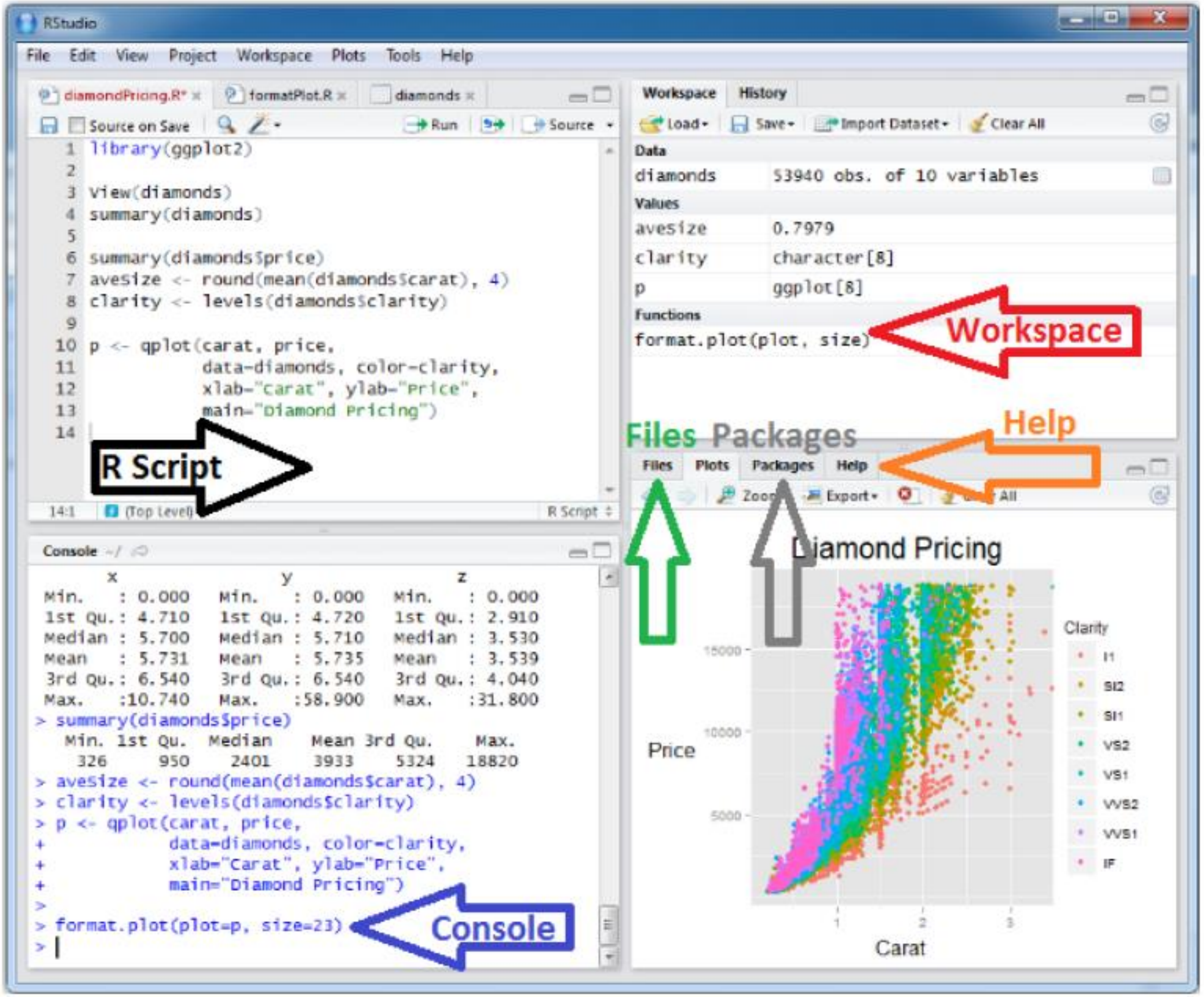
MC Exam is on these topics

- data visualization
- basic terms and summary statistics
- study types, risk measure
- models/distribution-types, parameter estimation
- testing, confidence intervals, p-values
- linear regression, adjusting
- diagnostic tests, classification
- logistic regression
- reliability analysis
- Causality
- 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



```r
library(ggplot2)

View(diamonds)
summary(diamonds)

summary(diamonds$price)
avesize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")

```

```
      x                 y                 z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> avesize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(plot=p, size=23)
> |
```

Data

diamonds 53940 obs. of 10 variables

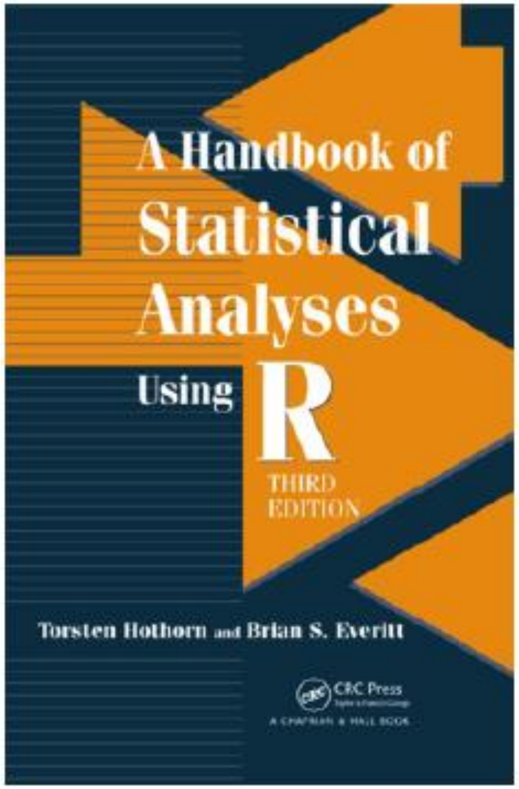
Values

avesize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

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

Zürich: ETH

PhD and Postdoc

Contract lecturer

Basel: OncoScore

Biomarker detection
CAS pharmaceutical Med.

Lausanne: UNIL, DAFL

Head of bioinformatics & biostatistics at DNA Array Facility UNIL

Winterthur: ZHAW, SoE

Prof. for applied statistics & Scientist

Focus: Deep Learning

UZH: EBPI, Biostatistics

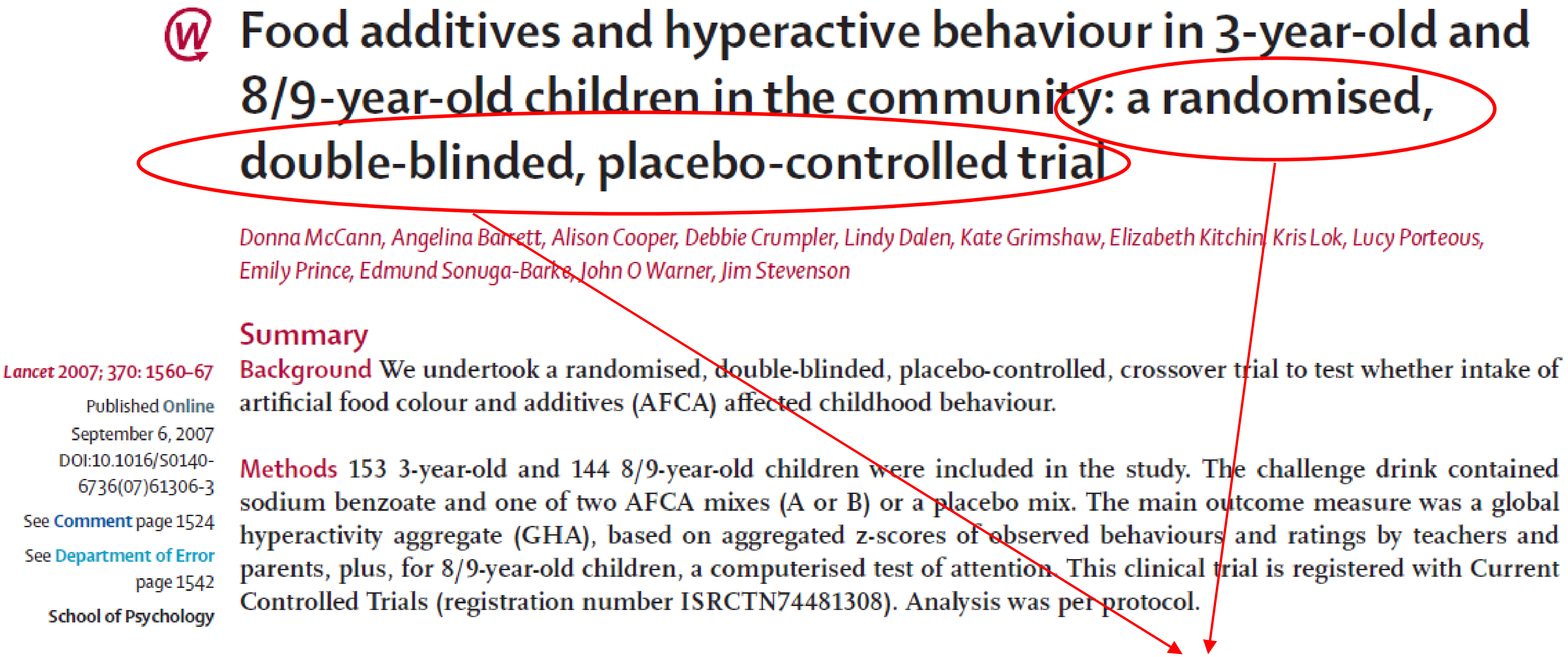
Scientific collaborator, consultant, and lecturer in the field of biostatistics and medical research

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.



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

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

Summary

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

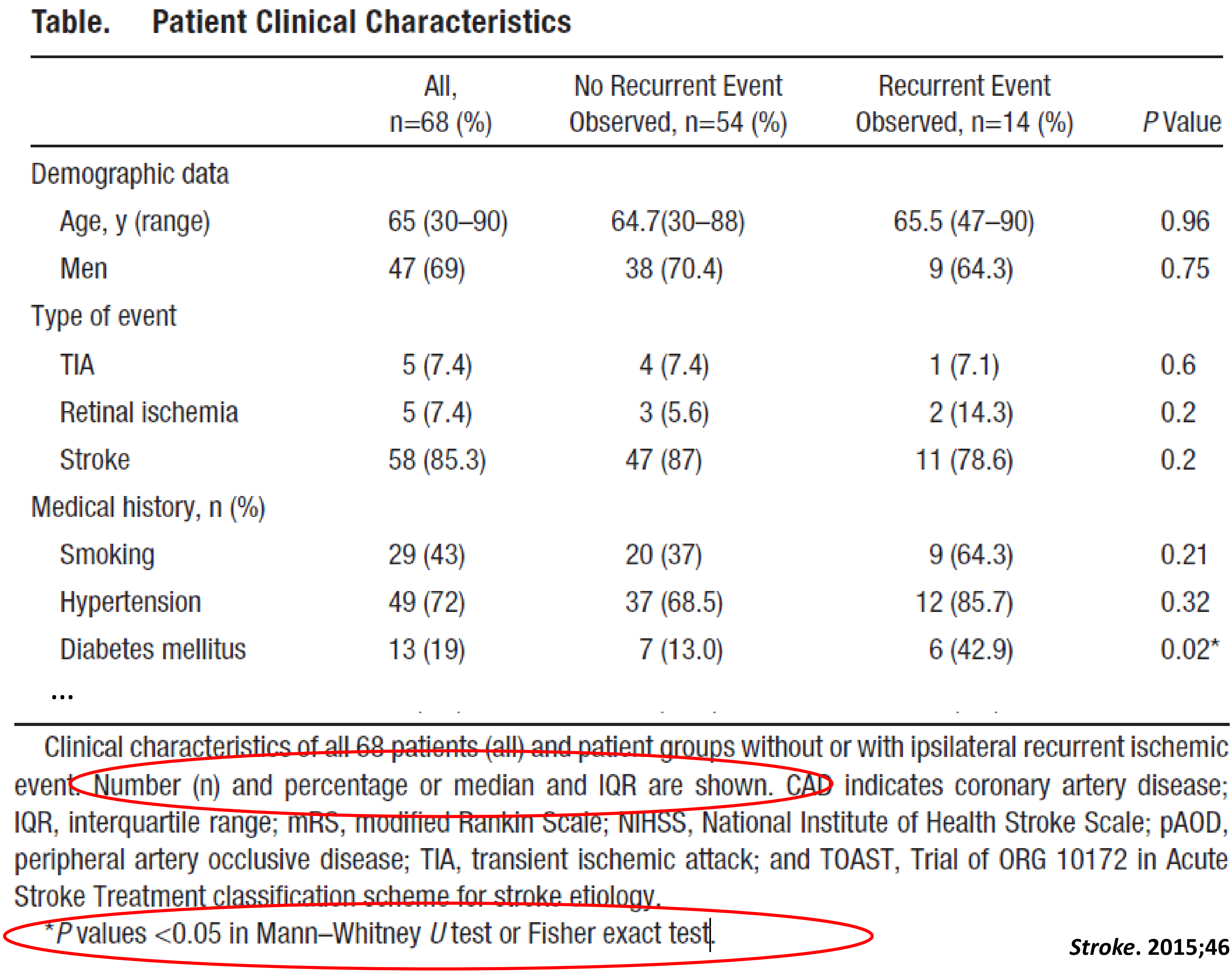


Table. Patient Clinical Characteristics

	All, n=68 (%)	No Recurrent Event Observed, n=54 (%)	Recurrent Event Observed, n=14 (%)	*P* 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.

Stroke. 2015;46:3274-3276.

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

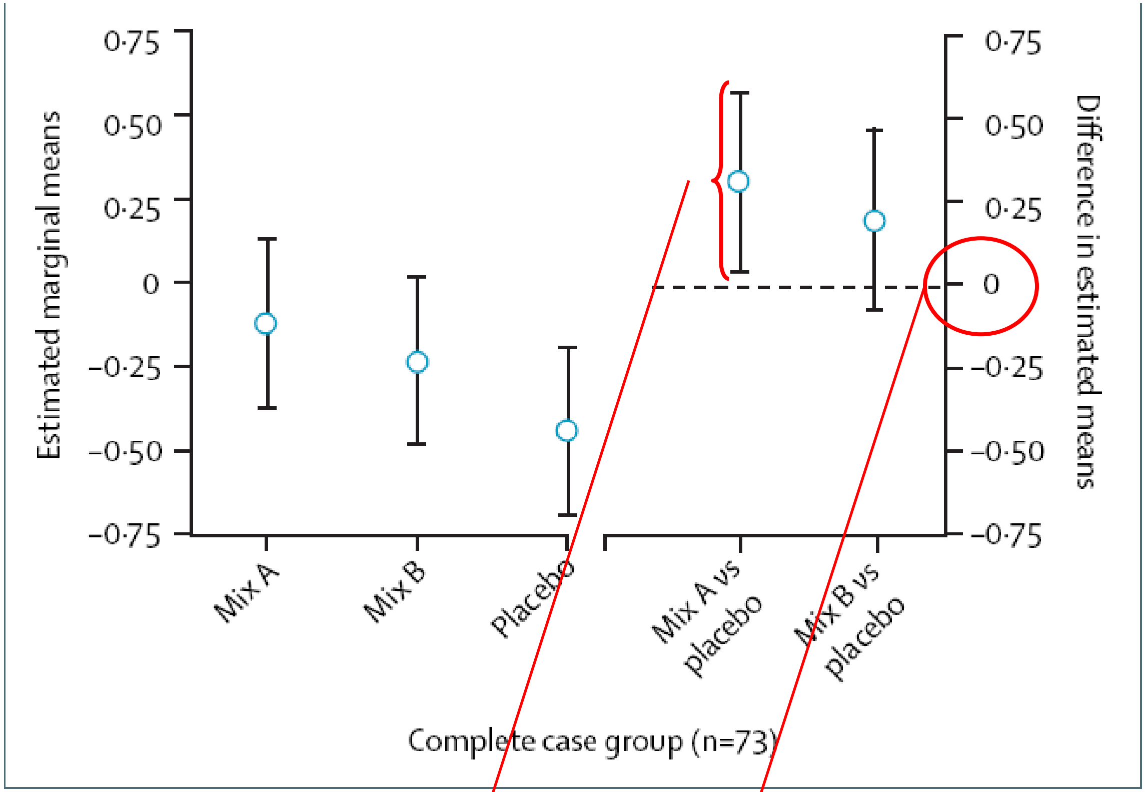
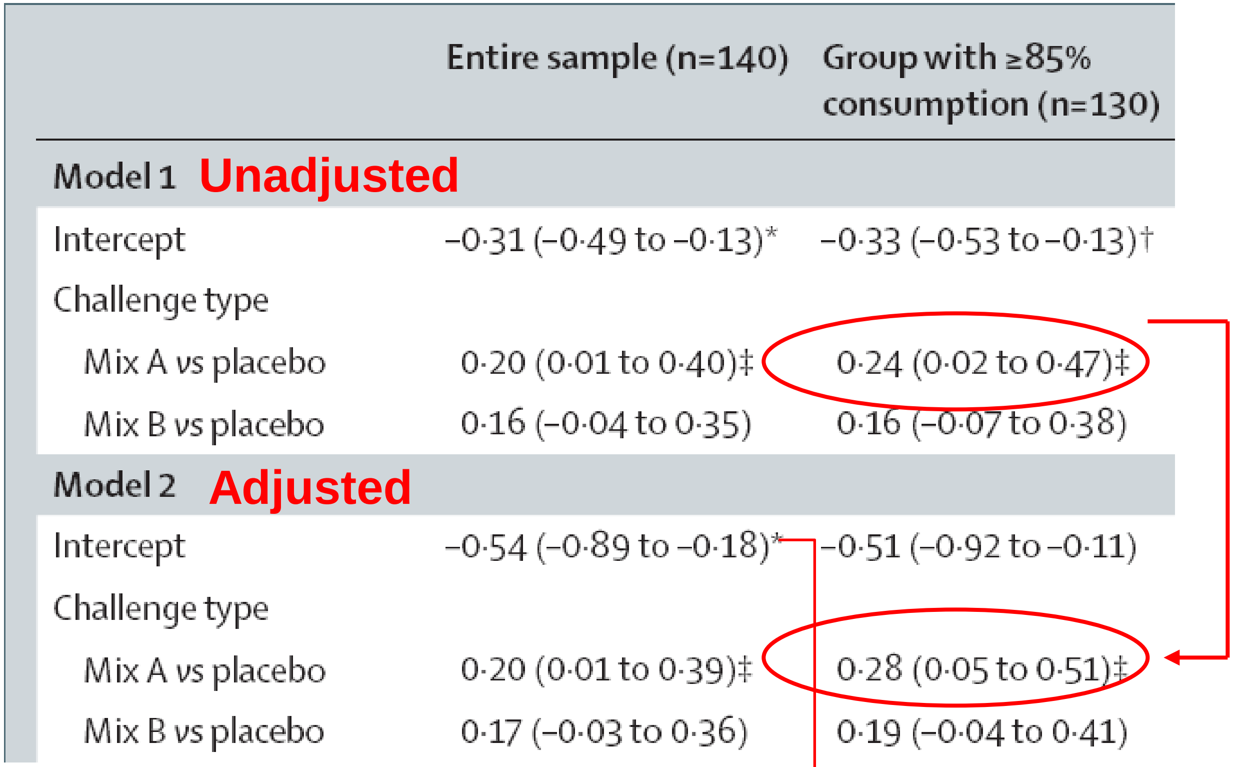


Figure 3: Estimated marginal means by challenge type and difference in estimated means in GHA under model 2 for 3-year-old children

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

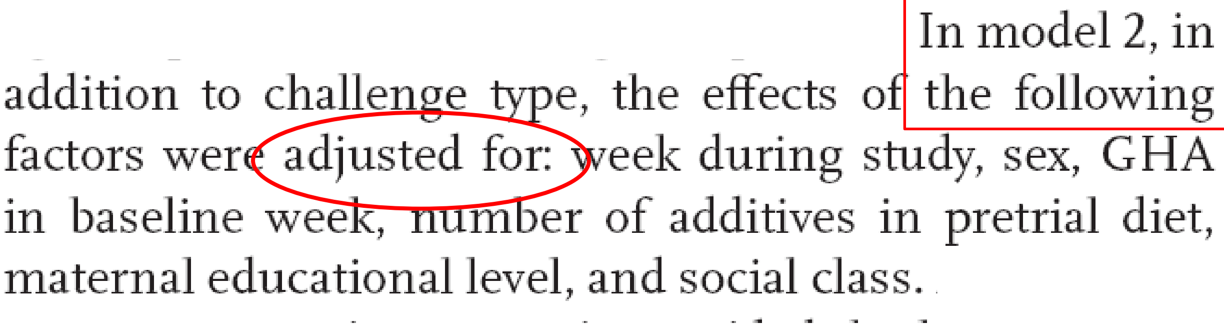
Research Article Example:
Paper on hyperactivity form McCann et al.



	Entire sample (n=140)	Group with ≥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?



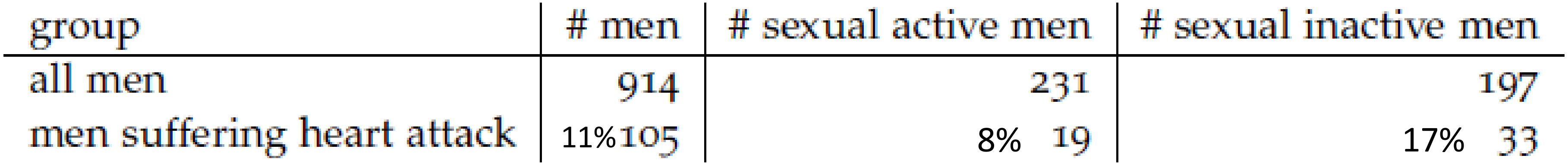
In model 2, in addition to challenge type, the effects of the following factors were adjusted for: week during study, sex, GHA in baseline week, number of additives in pretrial diet, maternal educational level, and social class.

What does * mean?

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

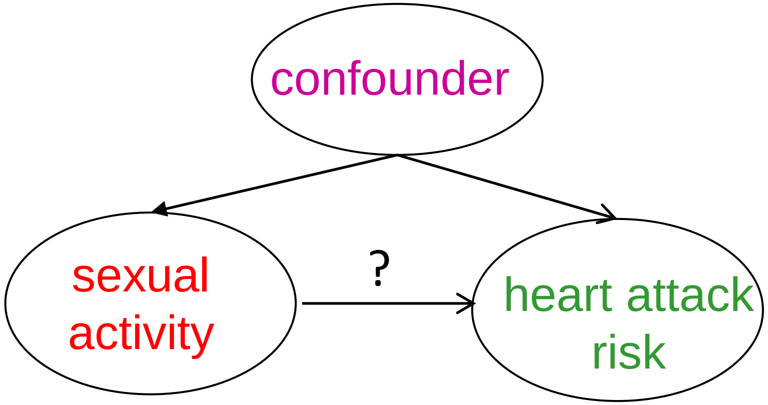
914 healthy men, between 45 and 95 years old, were chosen at random and followed over 10 years where they were interviewed e.g. about their sexual live. Moreover it was followed who suffered a heart attack in this period.

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.“

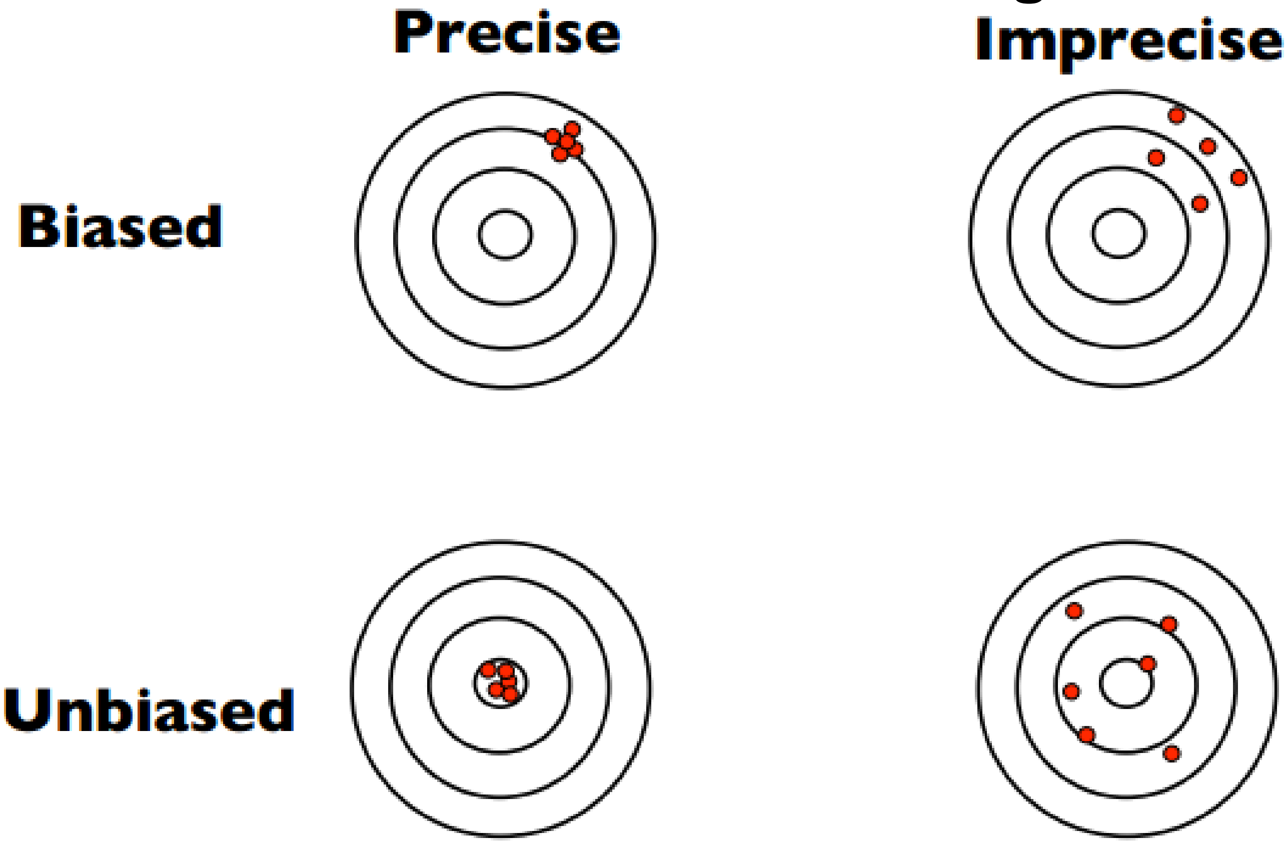
Numbers and Data

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 **Large variance**



Precise

Imprecise

Biased

Unbiased

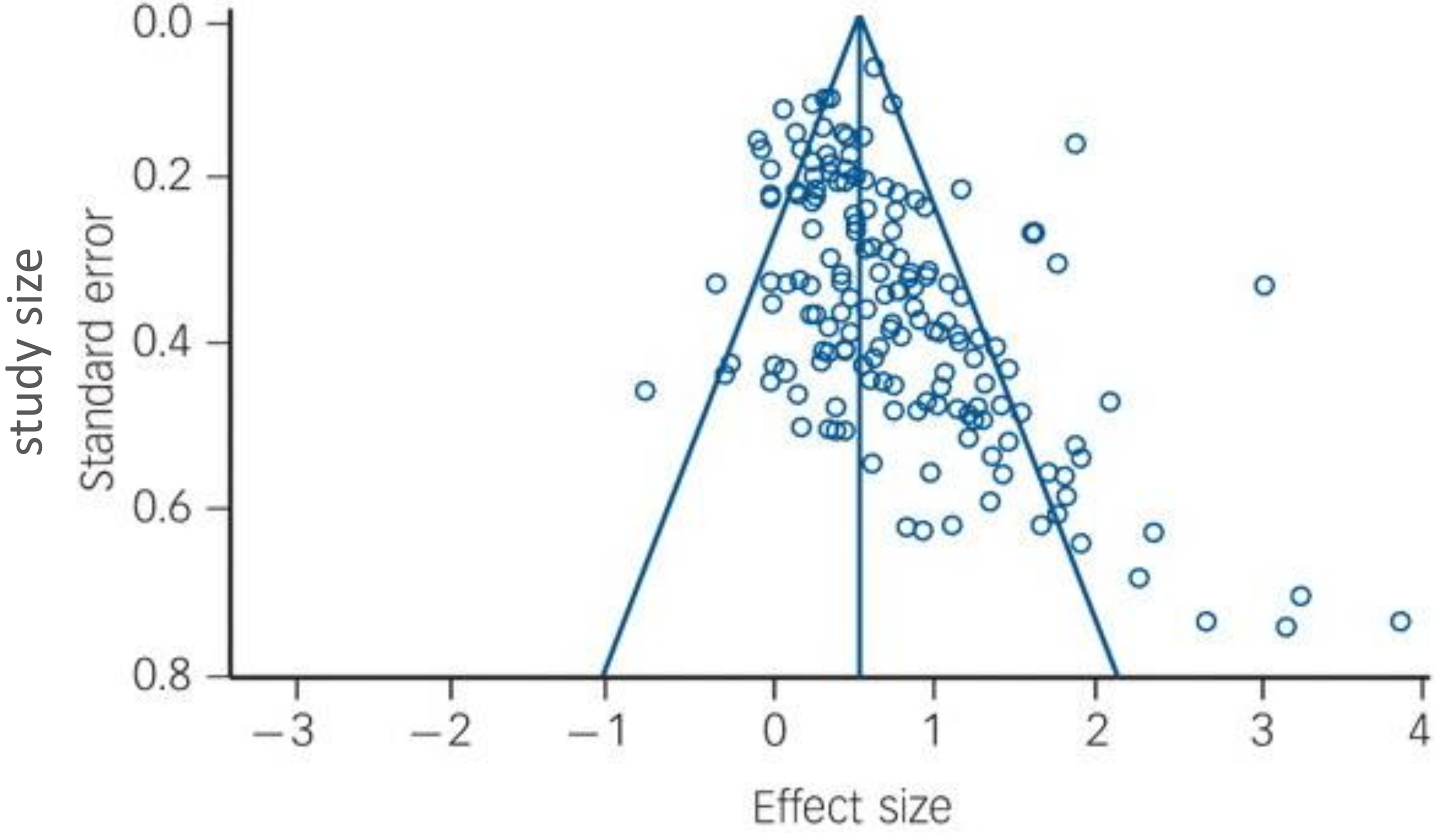
Bias

Systematic error or bias, accidental coarse error

Wrong basic calibration

Accidently 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.

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**:

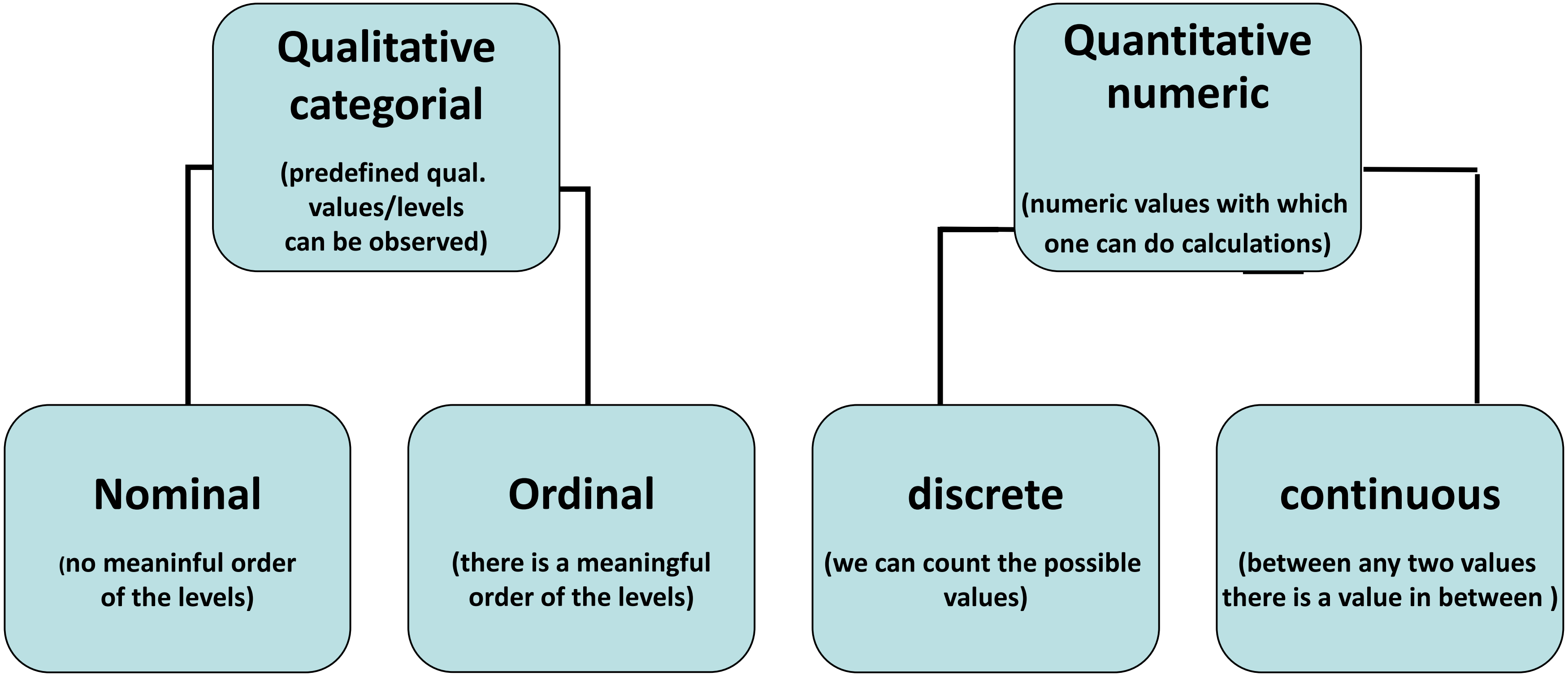
  A subgroup of the entire population which have been selected in a certain manner (systematically, arbitrary, at random, stratified)

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



Qualitative categorial

(predefined qual. values/levels can be observed)

Quantitative numeric

(numeric values with which one can do calculations)

Nominal

(no meaninful order of the levels)

Ordinal

(there is a meaningful order of the levels)

discrete

(we can count the possible values)

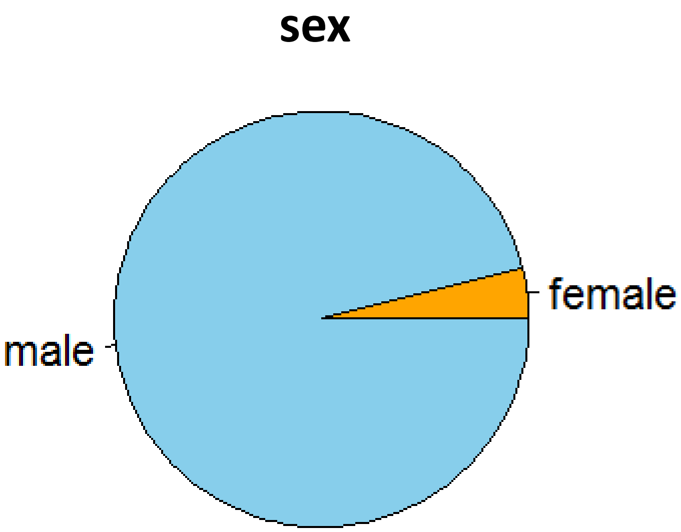
continuous

(between any two values there is a value in between)

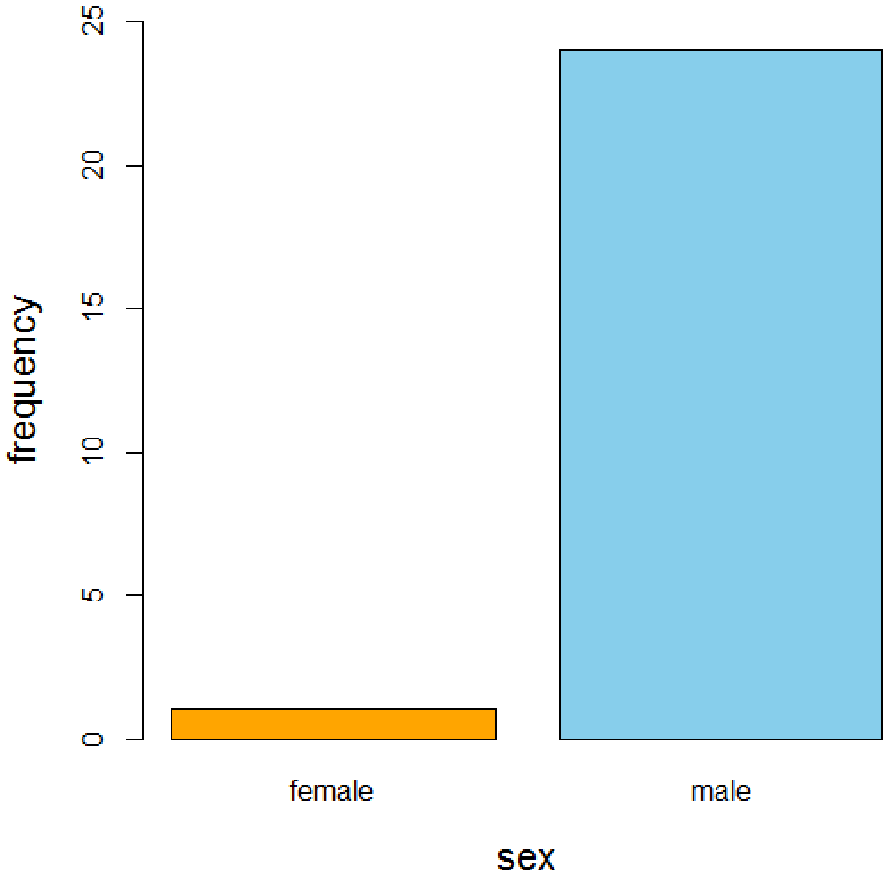
Visualization
of
categorical variables

How to summarize categorical data?

pie chart



bar chart

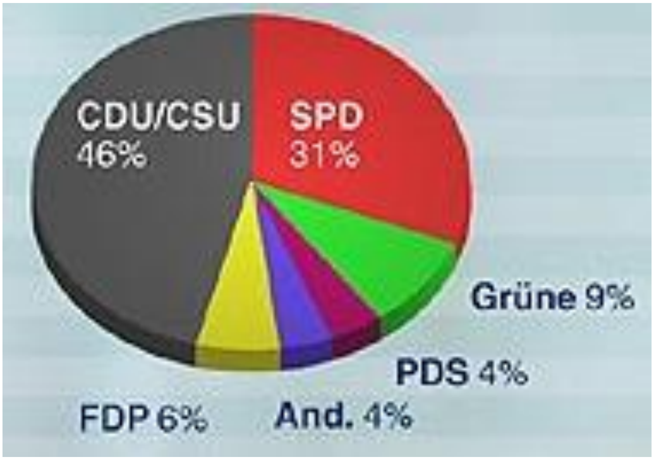


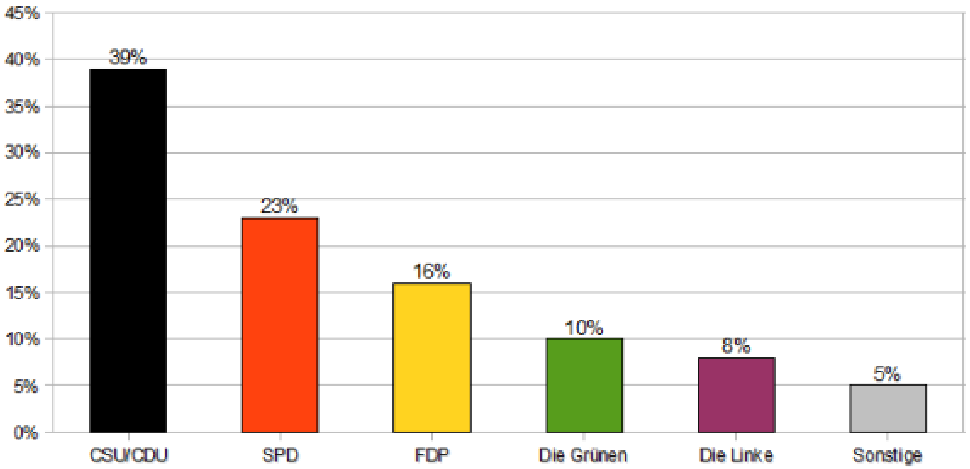
```
pie(table(dat$sex))
```

```
barplot(table(dat$sex))
```

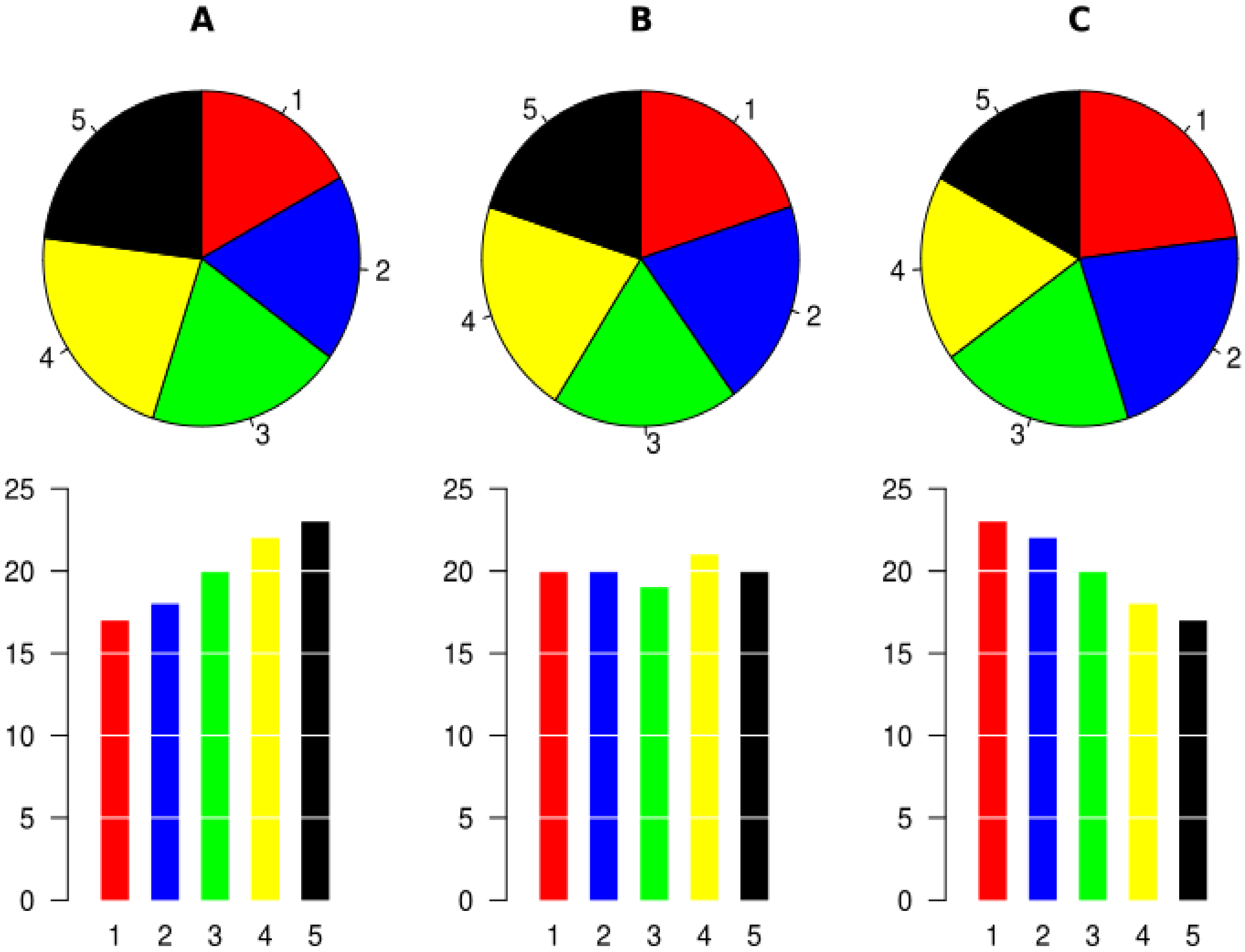
Visualizing categorical data by Bar-Chart or Pie-Chart

These charts are simple - is there room form manipulation?



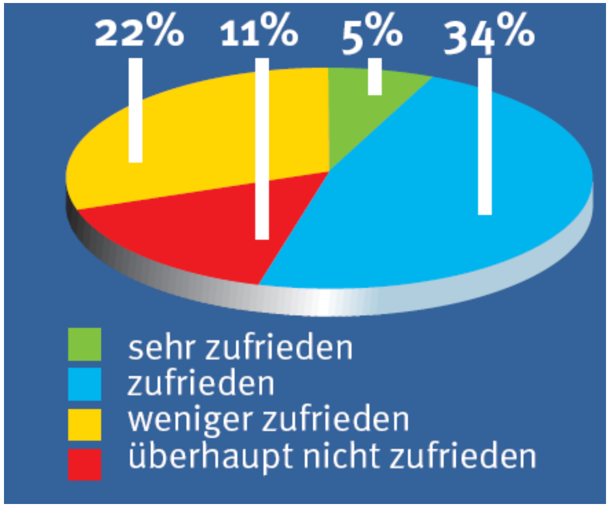


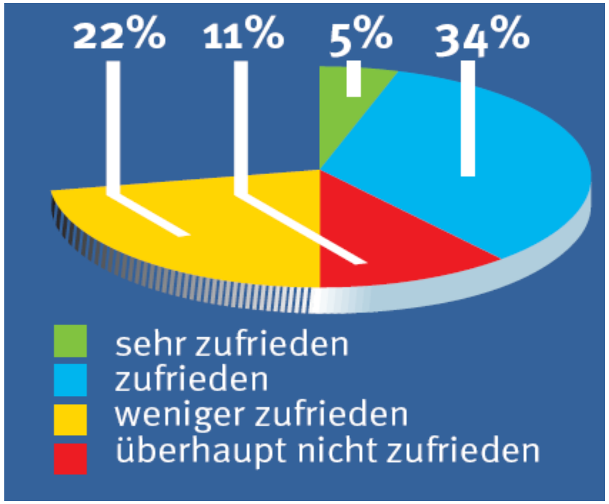
Barplots are often to prefere over pie-charts



Humans are much better in comparing heights than comparing areas.

Half of all reader are satisfied with Klinsmann - true?





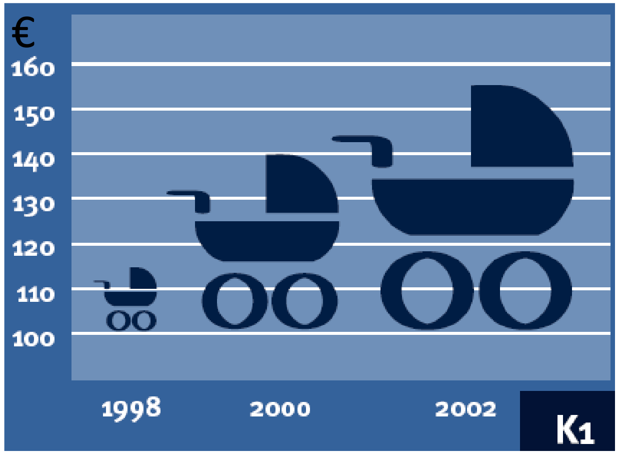
Pie-Chart from the German newspaper „Bild“ . Reader were asked how satisfied they are with soccer trainer Klinsmann.

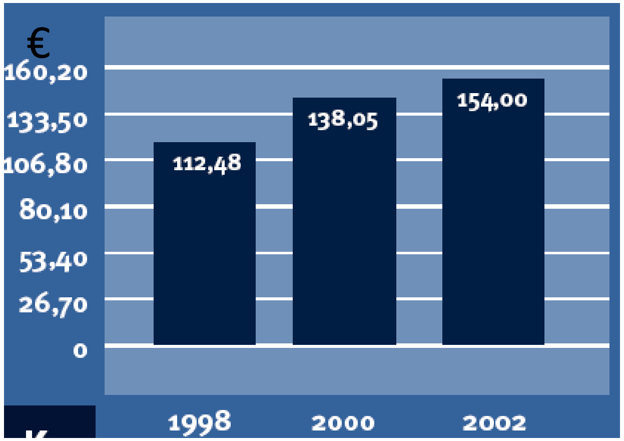
Manipulation:
pie segment does not correspond to percentage.

trick:
Percentages do not add up to 100%

Quelle: http://www.wdr.de.varnish.gl-systemhaus.de/tv/quarks/sendungsbeitraege/2006/1017/002_zahlen.jsp?smonat=2009-10

Generous increase of child allowance - true?



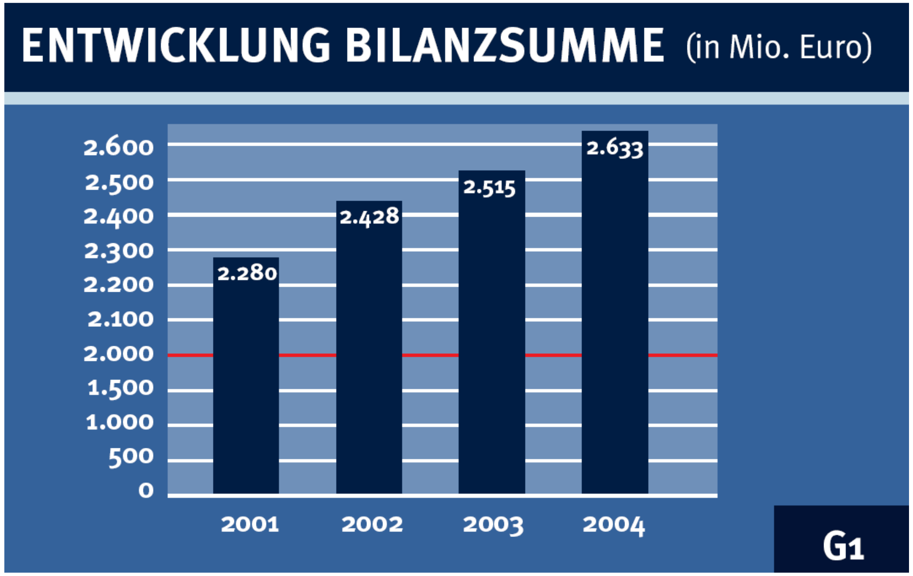


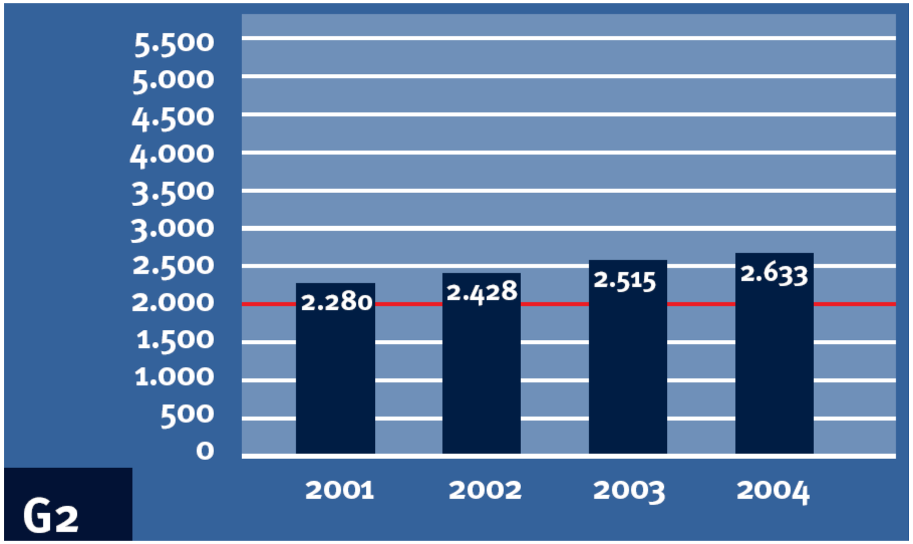
Graph from government statement in the German red-green agenda 2010.

Manipulation: area of the buggies are not proportional to the child allowance.

trick: scale starts not at 0 and the shape of the buggies are almost circles where the area increases proportional to the square of the height.

Good business development - true?



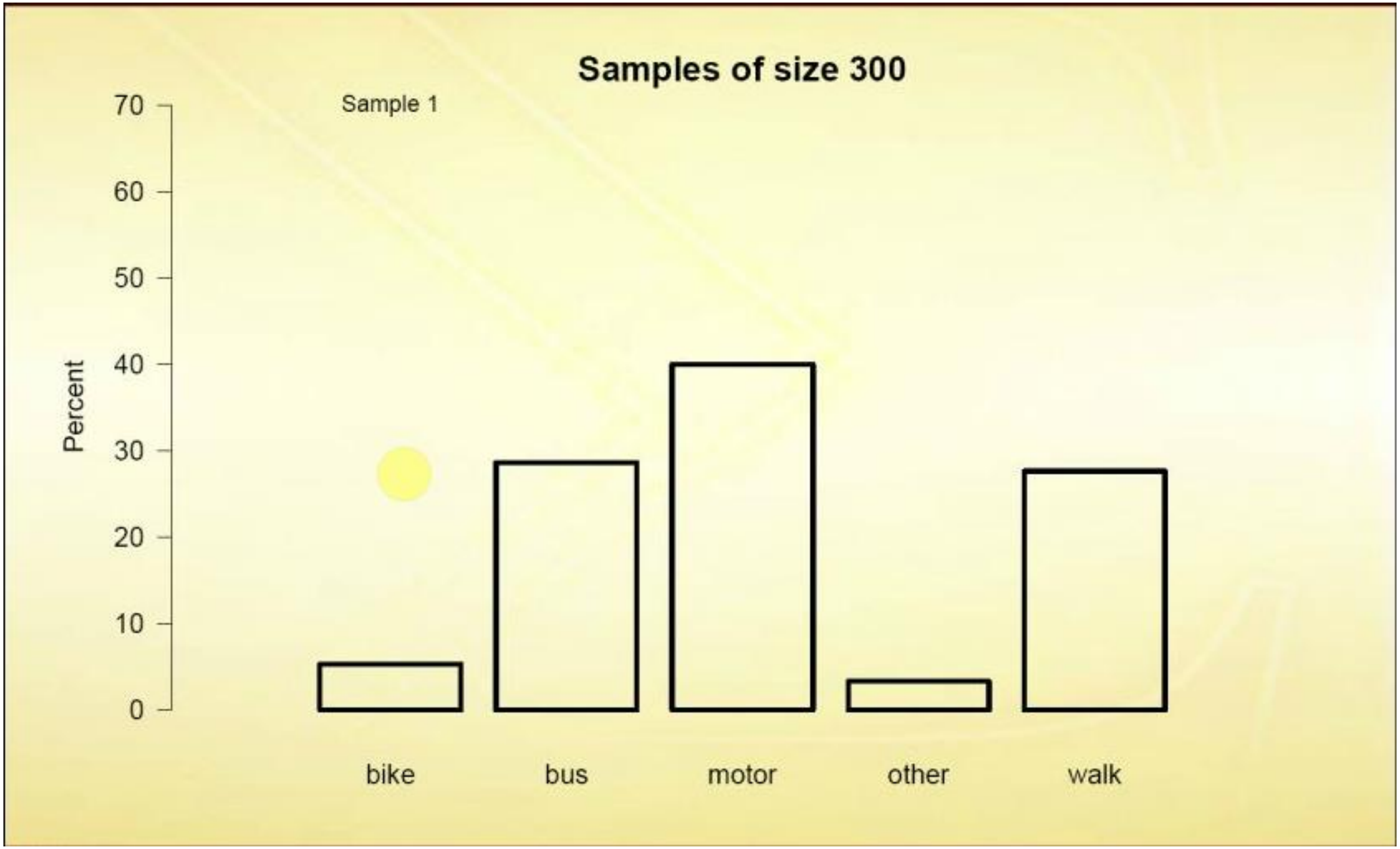


Bar Chart in the business report of a german bank
(psd-Bank Rhein-Ruhr 2004)

Manipulation: bar heights are not proportional to actual numbers.
trick: scale is changing above 2000.

How reliable are the bar-heights in a barplot?

How do pupils get to school?

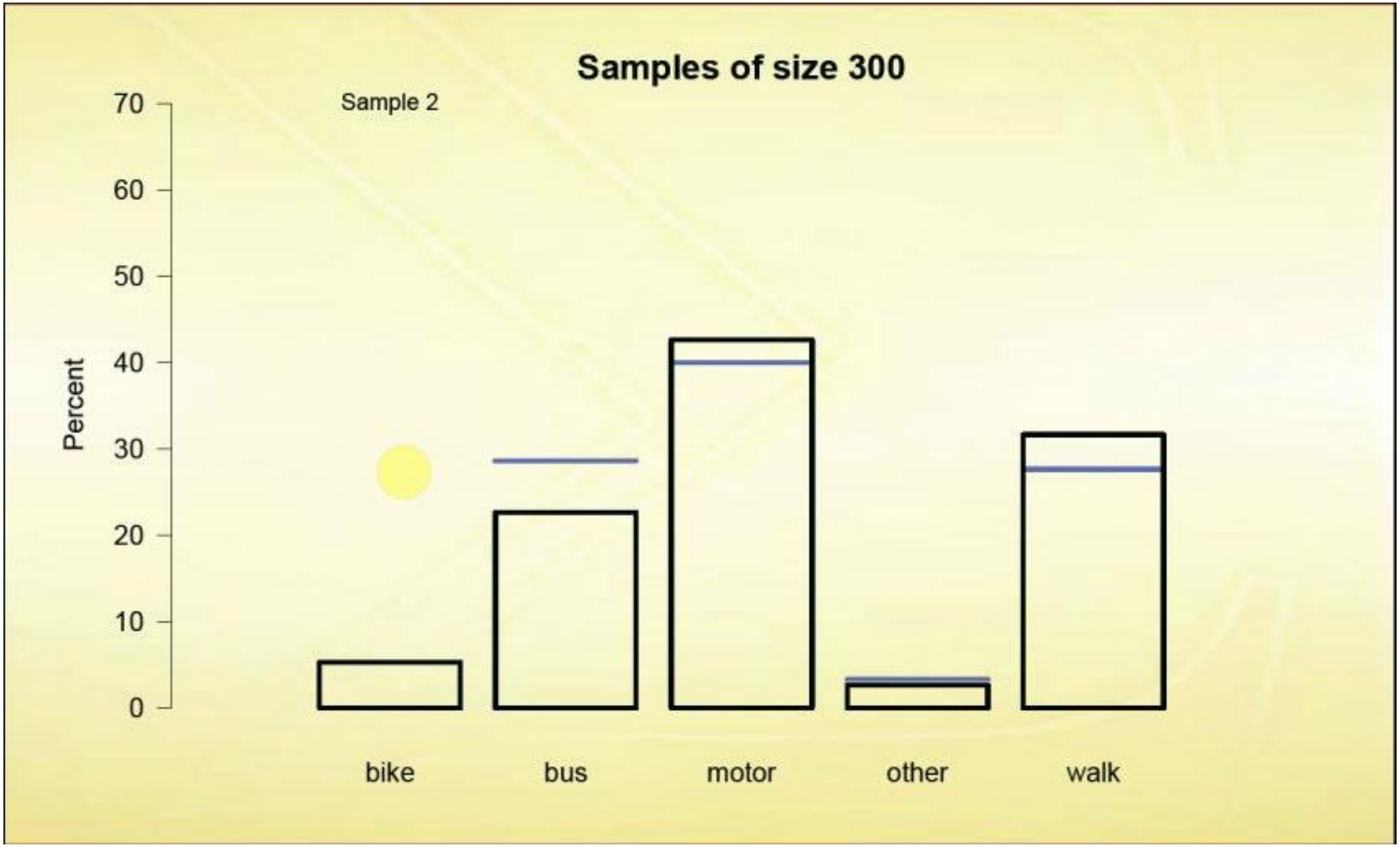


Balkendiagramm mit Gedächtnis

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

How reliable are the bar-heights in a barplot?

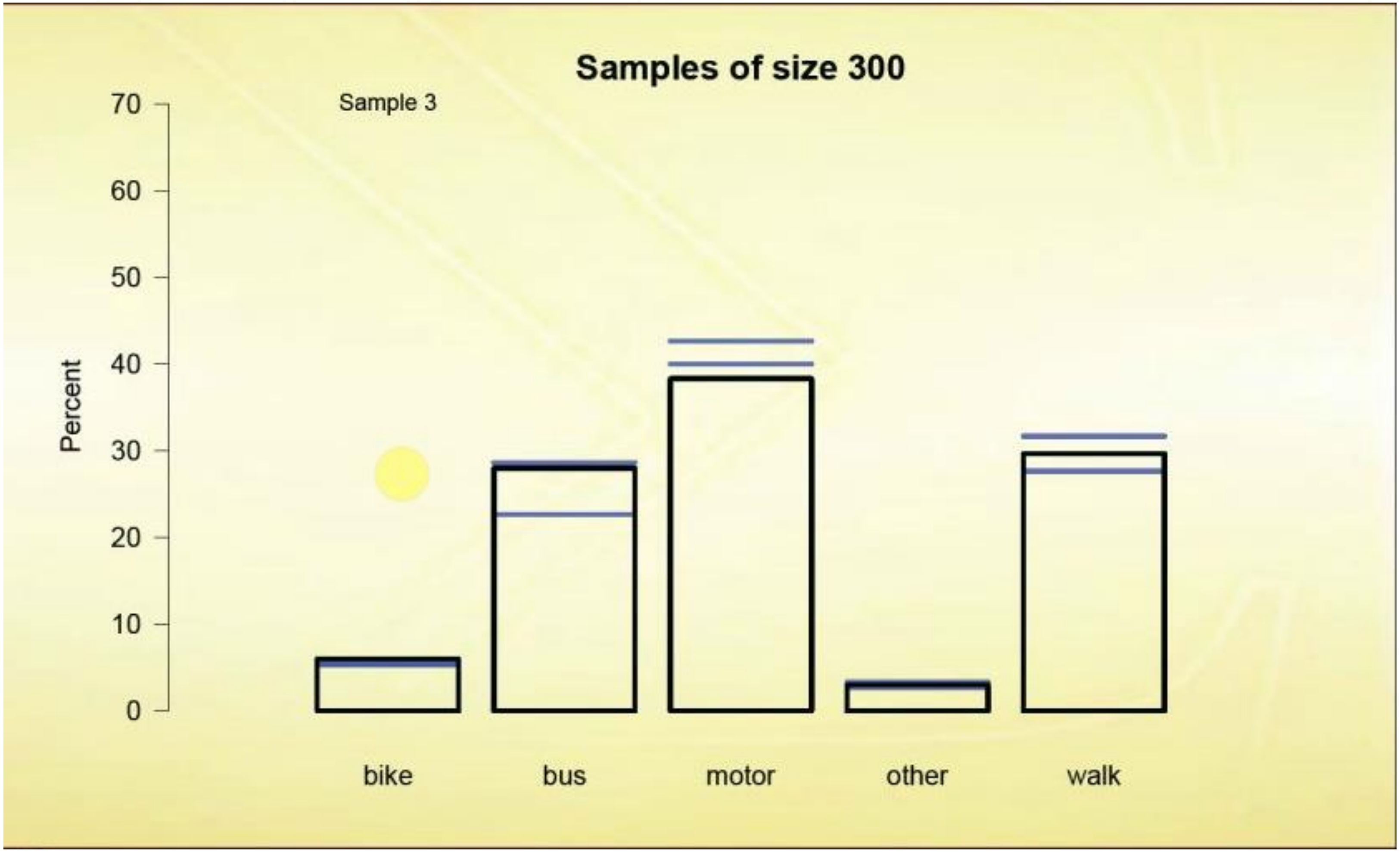
How do pupils get to school?



Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

How reliable are the bar-heights in a barplot?

How do pupils get to school?

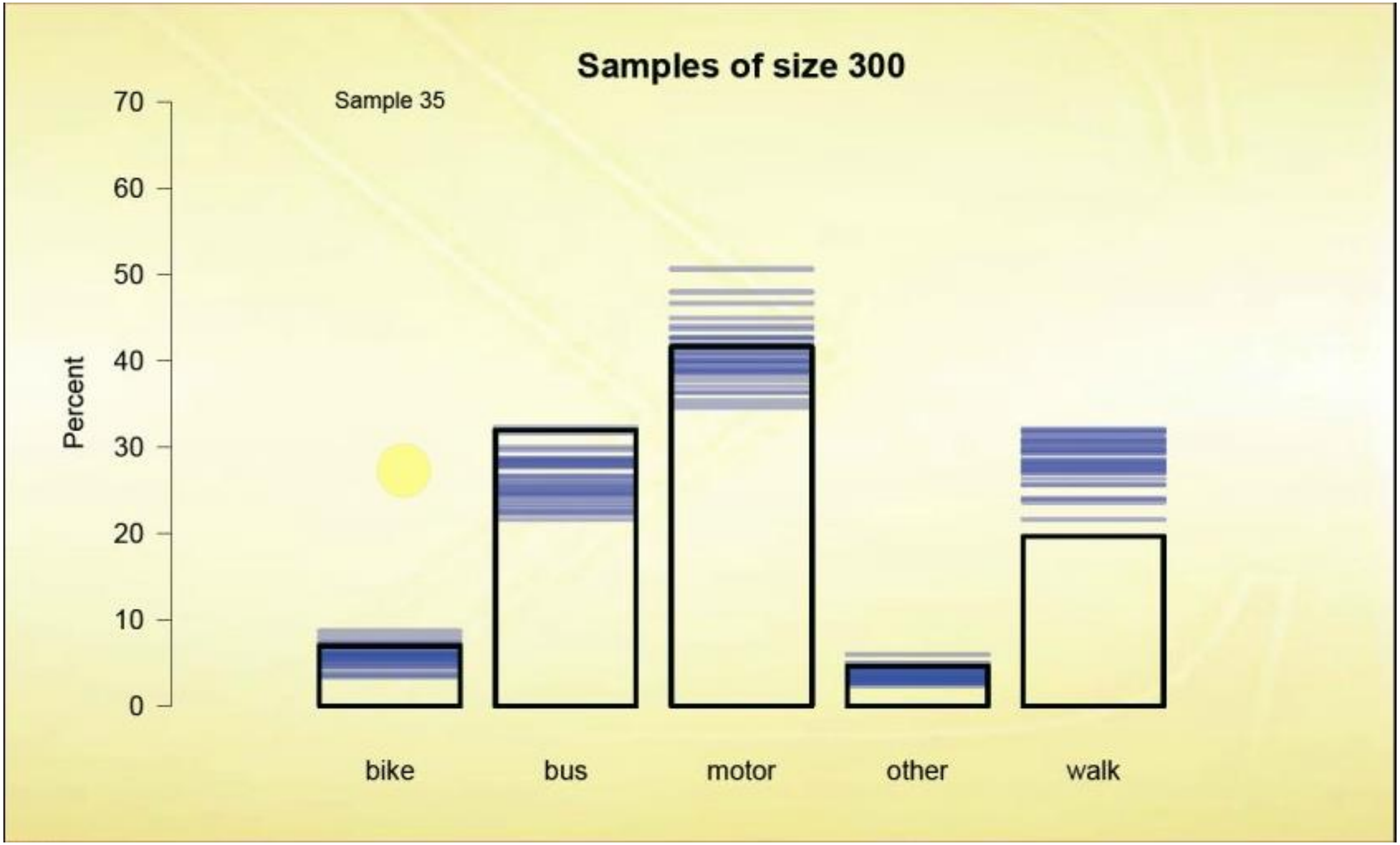


Balkendiagramm mit Gedächtnis

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

How reliable are the bar-heights in a barplot?

How do pupils get to school?

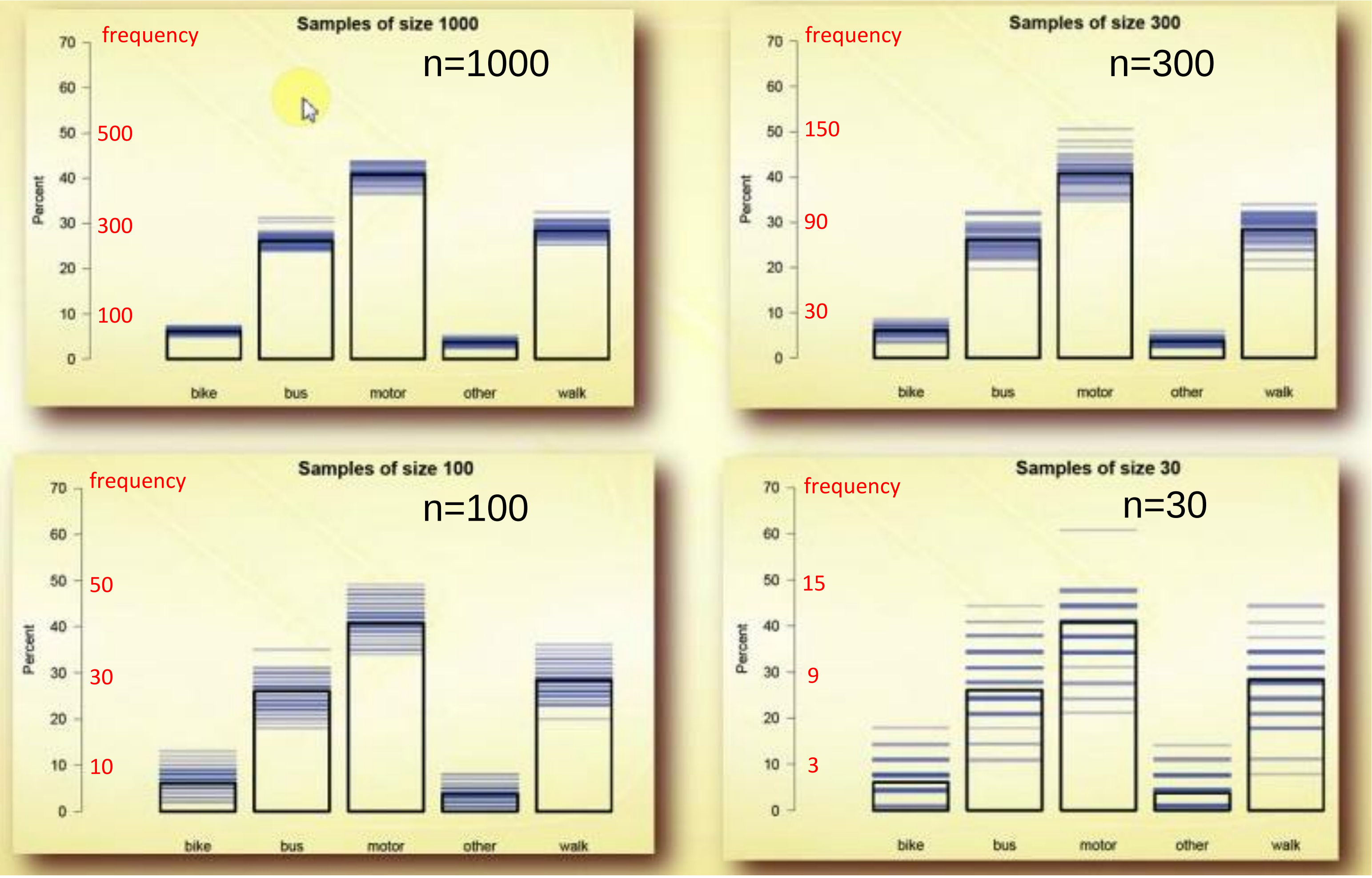


Balkendiagramm mit Gedächtnis

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

How reliable are the bar-heights in a barplot?

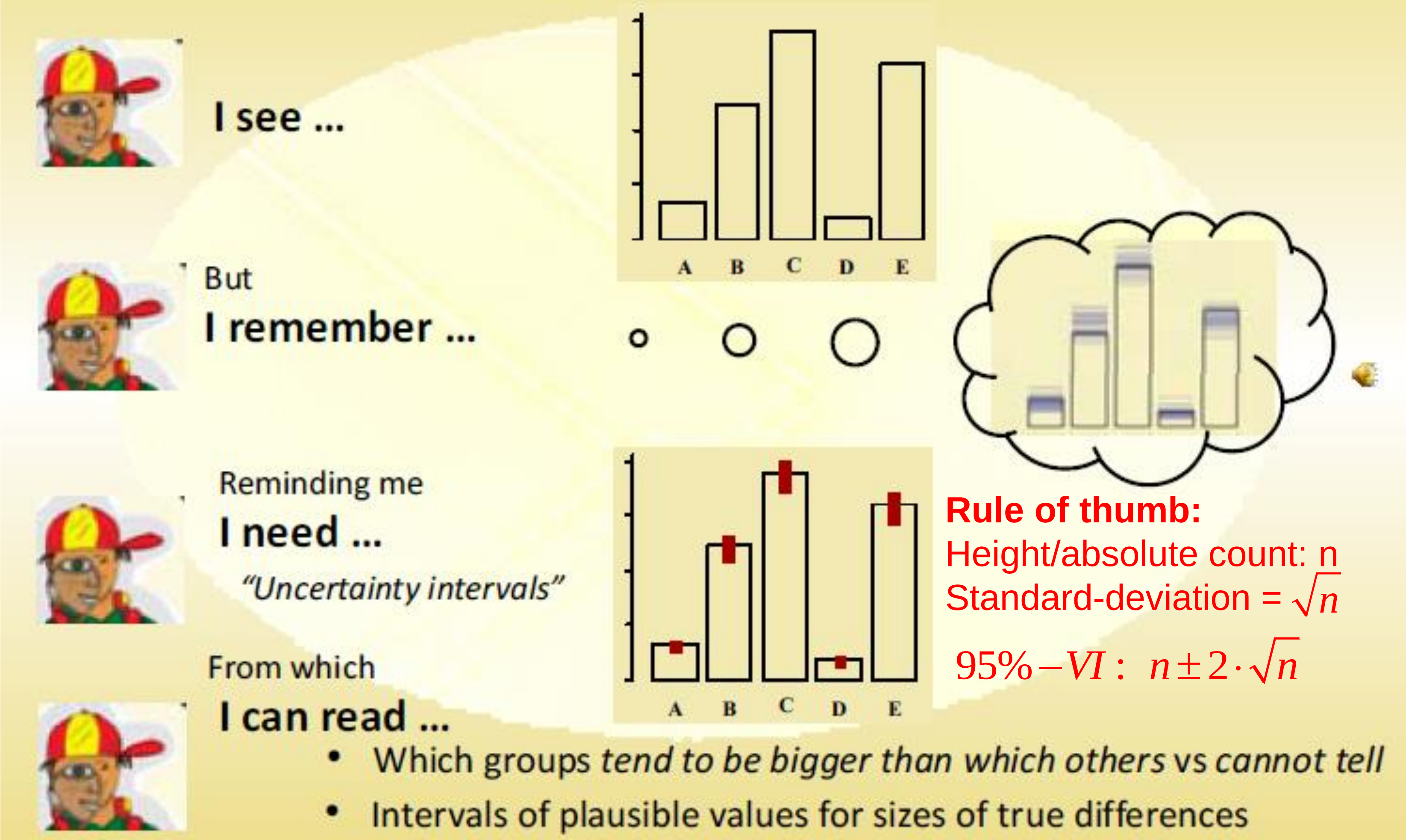
How do pupils get to school?



Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

How reliable are the bar-heights in a barplot?

How do pupils get to school?



I see ...

But

I remember ...

Reminding me

I need ...

“Uncertainty intervals”

From which

I can read ...

- Which groups *tend to be bigger than which others* vs *cannot tell*
- Intervals of plausible values for sizes of true differences

Rule of thumb:

Height/absolute count: n

Standard-deviation = $\sqrt{n}$

$$95\%-VI:\ n \pm 2\cdot\sqrt{n}$$

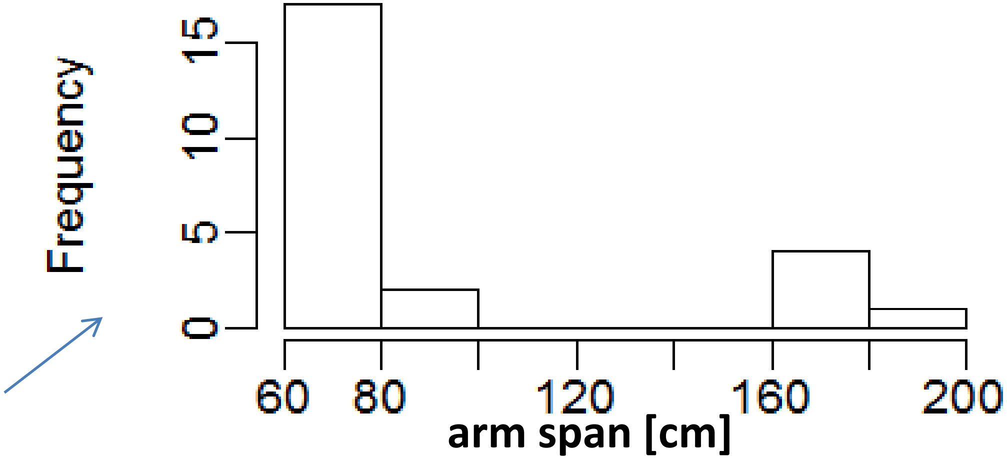
Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

Visualization of quantitative continuous variables

How to summarize continuous data – e.g. arm span?

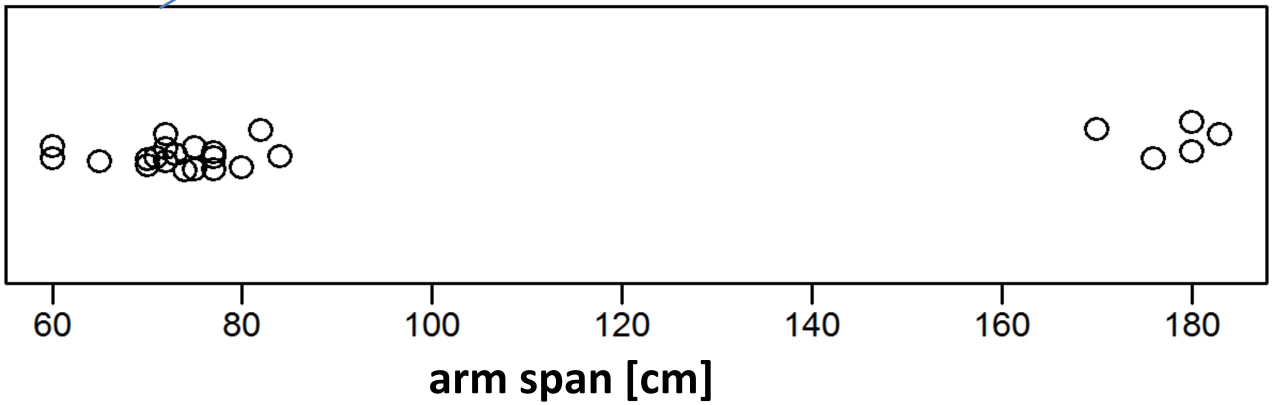
X: arm span	frequency
[60, 80)	17
[80,100)	2
[100,120)	0
[120,140)	0
[140,160)	0
[160,180)	4
[180,200)	1

- define non-overlapping classes/bins
- count number of observation per class
- draw histogram (no gaps between bars)



```
hist(x)
```

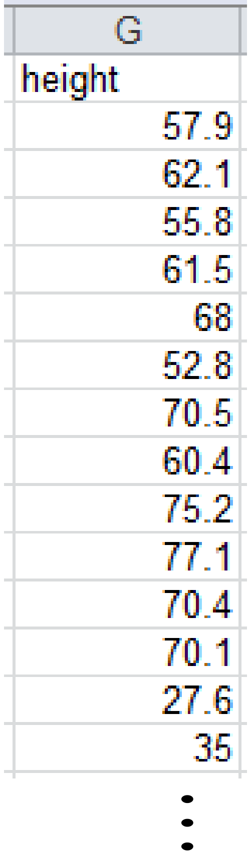
```
stripchart(x,method="jitter")
```



How to visualize continuous data?

The height (cm) of 376 plants were measured.

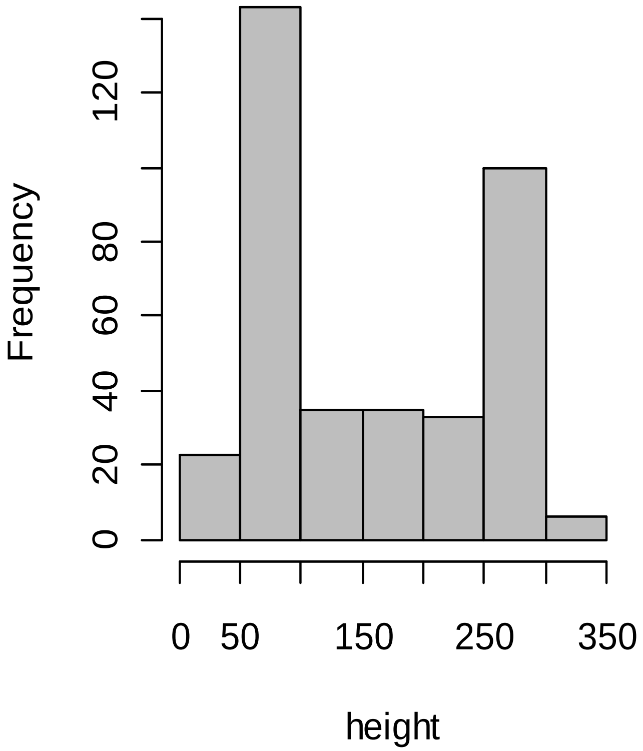
```
head(dat$height)
```

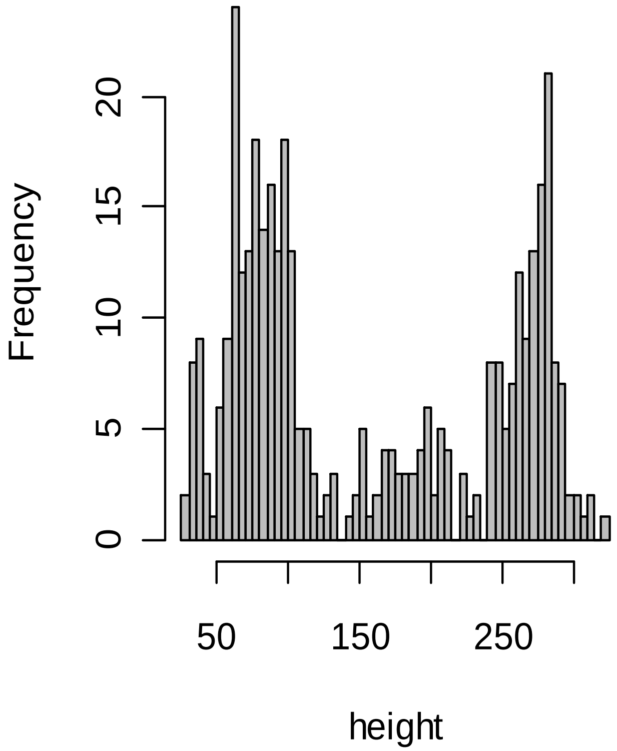


G
height
57.9
62.1
55.8
61.5
68
52.8
70.5
60.4
75.2
77.1
70.4
70.1
27.6
35

⋮

```
# hist, few classes, big bin width
hist(dat$height, nclass=7)
# hist, many classes, small bin width
hist(dat$height, nclass=100)
```

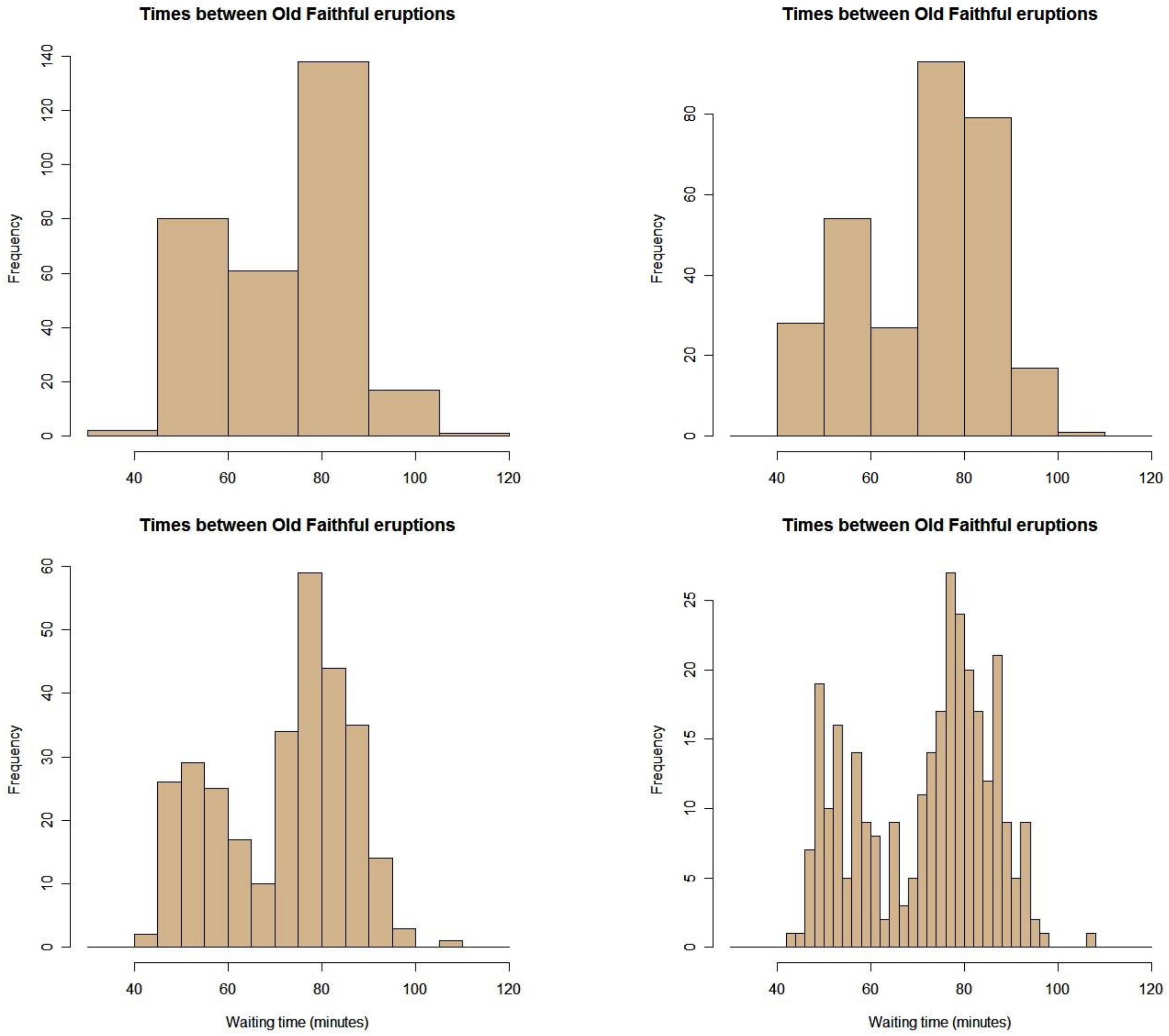




Are there subgroups? If yes – how many?

How reliable is the height of a bar?

How many classes do we need?



http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html

299 eruption intervals were observed

Shape of the histogram may depend on the class choices