What Is This Module All About?
Introduction to Explanatory Data Analysis (EDA)

Peter Büchel

HSLU W

SA: Week 1

Introduction

- What is “Statistics”?
- What is “Applied Statistics”?
- What is “Classical Statistics”?
- What is “Bayesian Statistics”?
- What are “Models”?
- What are “Simulations”?

What is Statistics?

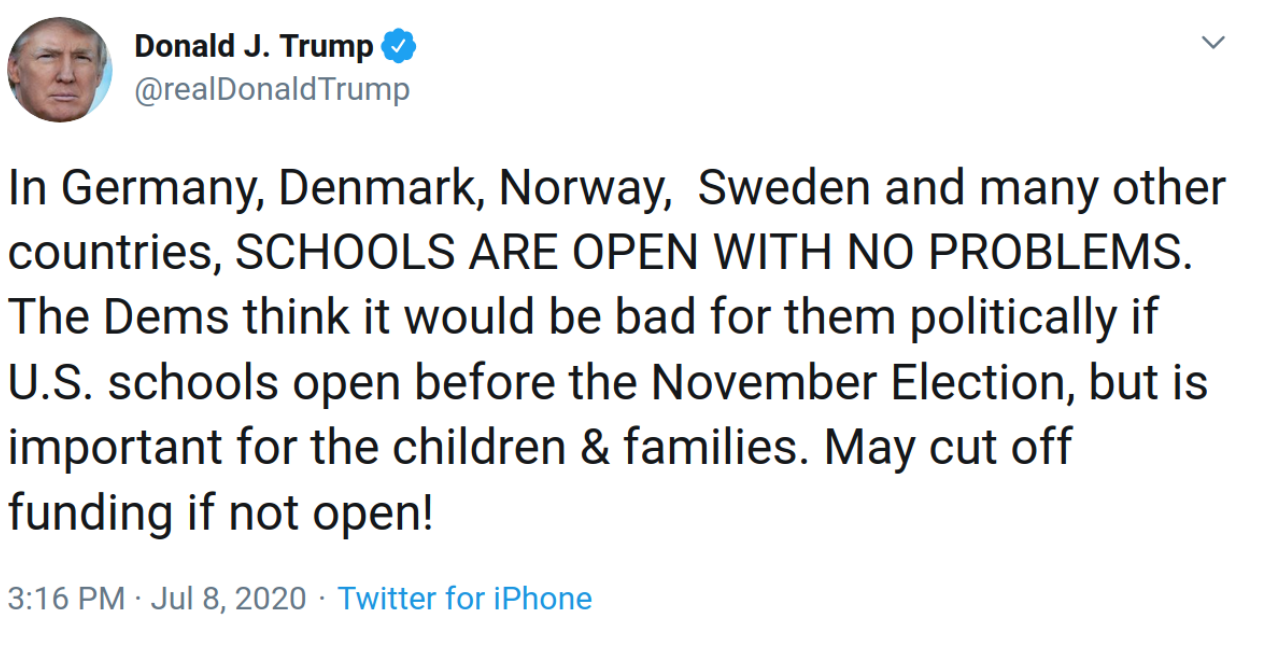
- Statistics: Discipline that concerns collection, organisation, analysis, interpretation, and presentation of data
- Very roughly: Statistics is the art of data handling
- Collecting data used to be a hassle, but that is history
- With computer and cheap sophisticated devices: Data are everywhere
- Example: Gazillion of photos on your smart phone
- Statistics has become very important in every ones life
- Google search: Uses a lot of statistics

What is "Applied Statistics"?

- Applied statistics: Applying statistics to real everyday problems
- Illustration with an example: *Critical thinking*
- Advantage of this example: No previous statistical knowledge is used (well, almost)

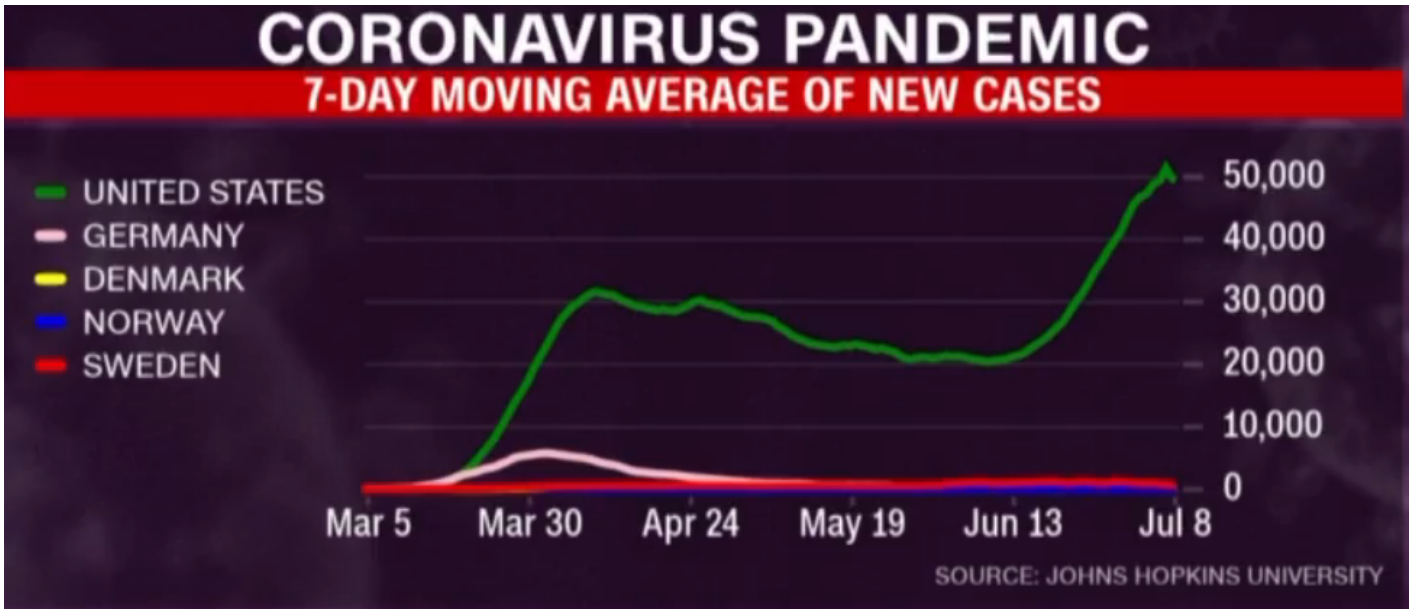
Example

- Following data: From 8 and 9 July 2020
- It began, like so much else then, with a tweet from Donald Trump:



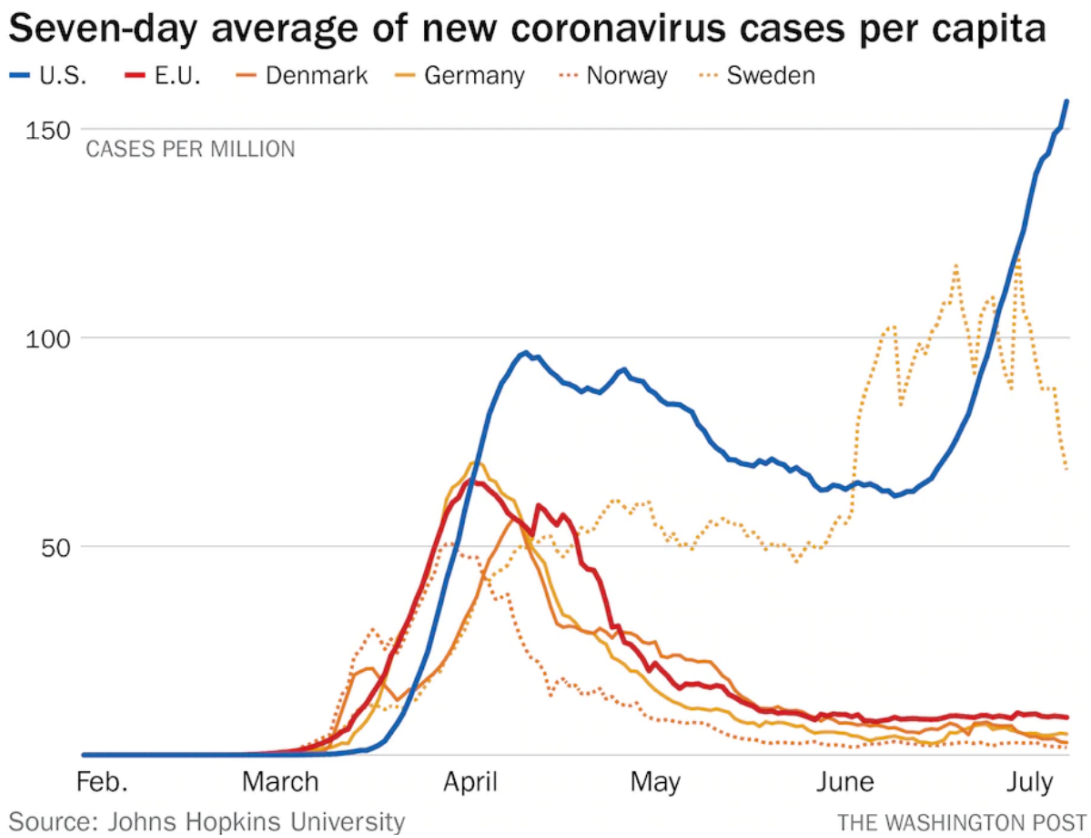
- Trump determined to reopen schools closed due to Covid-19
- Decision wasn't within his authority as president (but state governors)

- Argument for reopening: Germany, Denmark, Norway and Sweden had also done so without any problems
- CNN: Wanted to show this comparison is misleading (it is)
- But: CNN's reasoning was rather nonsensical
- Used following graph for its argument:



- Reasoning of CNN:
  - European case numbers are much lower than in US
  - Makes sense for European countries to reopen schools (case numbers low)
  - Does not make sense for US, because case numbers *very* high
- Plausible, but absurd: Comparing apples with pears
- Numbers in Figure above: *Absolute* numbers
- But: Norway population of about 5.5 million; US 330 million
- Curve of Norway *necessarily* much lower and flatter than of US
- Even if *relative* curve of Norway *had* a similar shape to curve of US, curve of absolute cases would not look much different than that in Figure above

- Based on same data: Graph in newspaper *Washington Post*:



- Case numbers: Per million inhabitants

- Makes more sense: Curves *can* be compared

- *Now*: Germany, Norway and Denmark in much better shape in terms of relative case numbers than US
- Sweden is not: Similar curve like US
- Latter observation not obvious at all in Figure slide 6
- Reopening schools: Make sense in Germany, Norway and Denmark
- Maybe not so in US and Sweden

- But: Also have to be careful with relative case numbers
- Table: Case numbers of 9 September:

All | Europe | North America | Asia | South America | Africa | Oceania

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	12,172,181	+16,579	552,164	+972	7,071,467	4,548,550	58,288	1,562	70.8			
1	USA	3,158,932		134,862		1,392,679	1,631,391	15,457	9,542	407	39,479,437	119,257	331,044,624
2	Brazil	1,716,196		68,055		1,152,467	495,674	8,318	8,073	320	4,468,829	21,021	212,591,154
3	India	769,150	+98	21,151	+7	476,565	271,434	8,944	557	15	10,740,832	7,782	1,380,270,828
4	Russia	700,792		10,667		472,511	217,614	2,300	4,802	73	21,790,705	149,317	145,935,982
5	Peru	312,911		11,133		204,748	97,030	1,265	9,488	338	1,842,316	55,862	32,979,917
6	Chile	303,083		6,573		271,703	24,807	2,053	15,852	344	1,220,790	63,850	19,119,526
7	Spain	299,593		28,396		N/A	N/A	617	6,408	607	5,734,599	122,652	46,755,218
8	UK	286,979		44,517		N/A	N/A	197	4,227	656	11,041,203	162,625	67,893,830
9	Mexico	275,003	+6,995	32,796	+782	167,795	74,412	378	2,132	254	684,804	5,310	128,958,893
10	Iran	248,379		12,084		209,463	26,832	3,309	2,956	144	1,872,391	22,287	84,012,442

- Countries first column: Ordered by absolute numbers in second column
- On top: Usual suspects at that time

- Table: Countries ordered by *relative* numbers of deaths (fourth to last column)

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
1	San Marino	716		42		660	14	1	21,093	1,237	6,865	202,239	33,945
2	Peru	691,575		29,976		522,251	139,348	1,488	20,921	907	3,386,625	102,450	33,056,487
3	Belgium	88,769	+402	9,909	+2	18,576	60,284	52	7,653	854	2,449,055	211,141	11,599,139
4	Andorra	1,261		53		934	274	3	16,316	686	137,457	1,778,504	77,288
5	Spain	525,549		29,516		N/A	N/A	1,034	11,240	631	9,987,326	213,595	46,758,226
6	UK	350,100		41,554		N/A	N/A	69	5,152	612	17,619,897	259,295	67,953,144
7	Chile	424,274		11,652		395,717	16,905	930	22,159	609	2,641,589	137,965	19,146,844
8	Bolivia	121,604	+835	7,054	+46	73,150	41,400	71	10,391	603	253,647	21,675	11,702,383
9	Ecuador	110,092		10,576		91,242	8,274	424	6,223	598	330,998	18,709	17,692,261
10	Brazil	4,147,794		127,001		3,355,564	665,229	8,318	19,488	597	14,408,116	67,694	212,842,596
11	Italy	278,784		35,553		210,238	32,993	142	4,612	588	9,271,810	153,393	60,444,825
12	USA	6,486,426	+851	193,586	+52	3,758,629	2,534,211	14,589	19,575	584	88,067,850	265,771	331,367,517
13	Sweden	85,707		5,838	+4	N/A	N/A	13	8,477	577	1,124,269	111,192	10,111,092
14	Mexico	637,509	+3,486	67,781	+223	446,715	123,013	2,836	4,935	525	1,435,703	11,114	129,184,522
15	Panama	97,578		2,099		70,247	25,232	149	22,550	485	369,420	85,371	4,327,225
16	France	328,980		30,726		87,836	210,418	537	5,038	471	8,500,000	130,167	65,300,897
17	Sint Maarten	516		19		321	176	7	12,009	442	2,450	57,022	42,966
18	Colombia	671,848		21,615		529,279	120,954	863	13,178	424	2,964,722	58,150	50,983,652
19	Netherlands	76,548	+964	6,244	+1	N/A	N/A	45	4,466	364	1,648,103	96,144	17,142,061
20	Ireland	29,774		1,777		23,364	4,633	7	6,017	359	906,432	183,191	4,948,020

- The two rankings look completely different
- Never heard: San Marino, country worst affected by Covid-19
- What's going on?

- Last column: Countries listed small or very small in terms of population
- Few absolute cases have big impact on relative case numbers
- First table: Large or very large countries in terms of population

- Observations just made occur very often in applied problems:
  - Relative case numbers clarify argument of CNN
  - But: Relative case numbers *not always* better or more meaningful than absolute ones
- For applied problems: *There is no cooking recipes how to solve problems*

Problem Solving in Applied Statistics

- Example above very simple
- But contains many aspects relevant for solving problems in applied statistics
- First: Not clear what question or problem is
  - Started with Trump tweet
  - And now?
  - CNN, generally anti-Trump, thought it was worth noting
- Second: It's not clear what solution is going to look like
  - *How* to respond to a tweet like that?
  - CNN: Arguing on basis of case numbers

- Third: It's not clear what elements to use for solution
  - CNN decided to use absolute case numbers instead of relative ones
- Fourth: It is not clear how to interpret result
  - Often most difficult part in solving problems in applied statistics
  - CNN went wrong direction: Compared things which cannot be compared

Example from School Mathematics

- Solve following equation for x:

$$2x + 1 = 5$$

  - Problem clear
  - Solution is clear (often practiced)
  - Nothing to discuss or interpret about solution $x = 2$

- *These kind of problems do not exist in applied statistics or science*

Everyday Problems (from: Eric Mazur; Principles & Practice of Physics)

- You're in a hurry to get somewhere, but you can't find your car keys
- You run out of flour baking a birthday cake and supermarket is closed
- Your flight's canceled en route to a job interview
- You want to buy these great new shoes, but there's no money in bank

- For all these problems there is no solution instruction
- Cannot be solved by formulae either
- Four-step problem-solving strategy
- Most statistical problems are formulated in words

First: First Steps

- It is not clear which is most efficient way to respond to a given problem
- First step of our problem-solving strategy, the beginning, is the most difficult one
- It therefore makes sense to start with something you *can* do:
  - Organize information given and make sure that you are clear what exactly is required in the problem
  - Make sure you are clear what information is included in the problem
  - Formulate problem in your own words
  - Finally, determine whether you have all information or not which are necessary to solve problem

Second: Prepare a Plan

- Next step: Work out a plan to solve your problem, i.e. figure out what you need to do to solve problem
- A good plan outlines steps you need to take to get a solution

Third: Execute Plan

- Execute your plan by following steps you outlined

Fourth: Interpret result

- You may think you are done
- But there is one last – and very important – step: interpret your answer
  - Check to see if your result is even possible
  - For example, if a probability becomes negative, then something must have gone wrong
  - Interpret result in words of problem

- School mathematics: 3rd point is often most important one
- In applied statistics: Not important, calculations are done by R
- 4th point *is* most important one
- Example: R-output (what R-command does is not important now)

```
t.test(x)

One Sample t-test

data:  x
t = 3.1814, df = 3, p-value = 0.05004
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.00151774  9.50151774
sample estimates:
mean of x
     4.75
```

- This output *has* to be interpreted
- All 4 points must be done correctly
- Even if 4th point is most important one, it is no use to do it correctly if a mistake was made in first 3 steps
- Application of these four points seems superfluous for very simple tasks
- But: As problems become more complex, they provide very good reference points for solving these problems
- We know from experience that it is difficult for students to read from problem definition what is actually being asked

What is Statistics that is Not Applied?

- Applied statistics: Procedures and methods are used and described
- Can often be explained in a simple way
- What we are *not* doing: *Why* these procedures and methods *exactly* do what they should
- While principle is often simple, details are difficult

- Details are proven in *mathematical statistics* and that looks like

$$S^2 := \frac{1}{m+n-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right\}.$$

Note that $\bar{X}$ has expectation μ and variance σ^2/n, and $\bar{Y}$ has expectation $\mu+\gamma$ and variance σ^2/m. So $\bar{Y} - \bar{X}$ has expectation γ and variance

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{n+m}{nm}\right).$$

The normality assumption implies that

$$\bar{Y} - \bar{X} \text{ is } \mathcal{N}\left(\gamma, \sigma^2\left(\frac{n+m}{nm}\right)\right)\text{−distributed.}$$

Hence

$$\sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{\sigma}\right) \text{ is } \mathcal{N}(0,1)\text{−distributed.}$$

To arrive at a pivot, we now plug in the estimate S for the unknown σ:

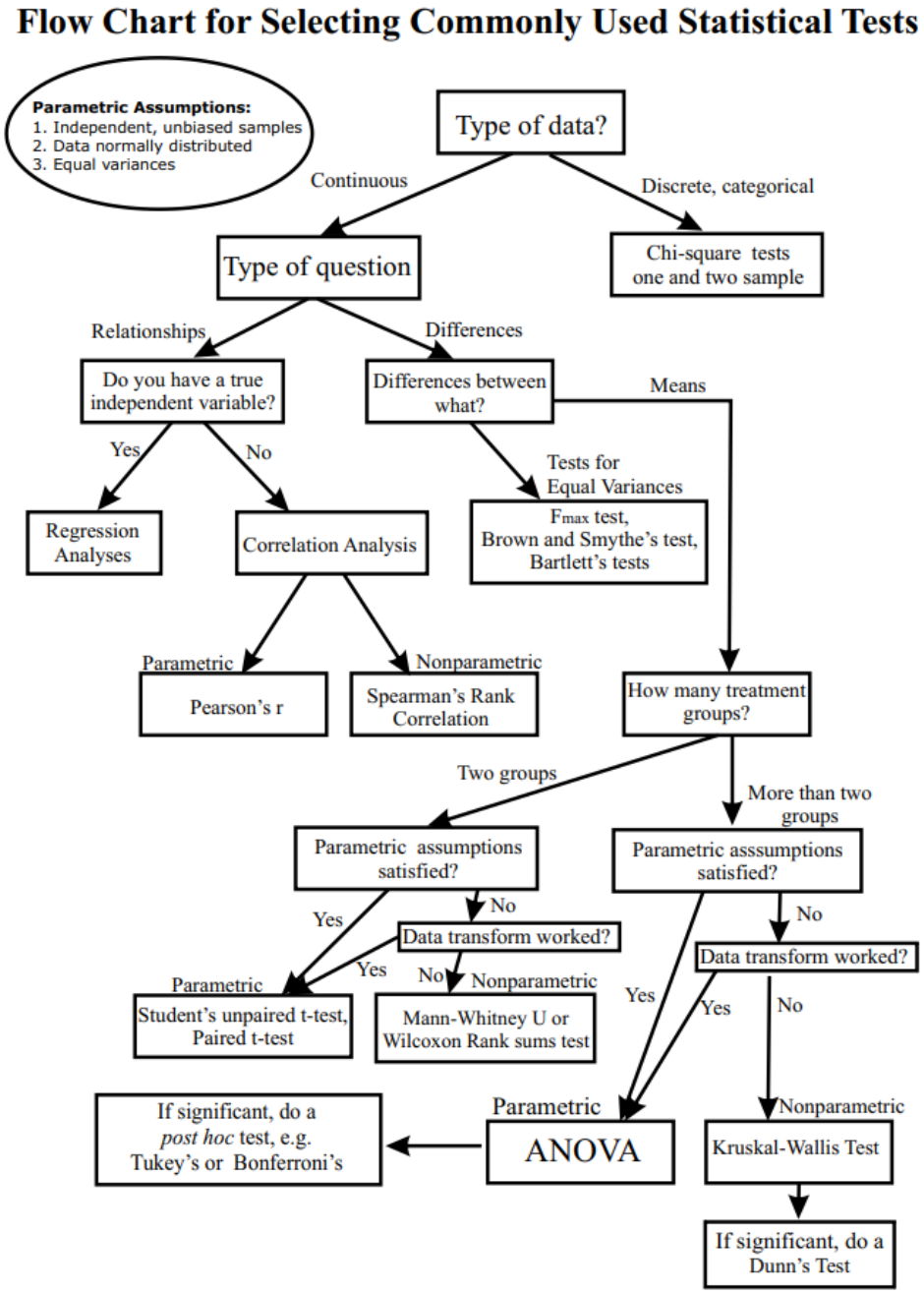
$$Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{S}\right).$$

- Don't despair, you won't see stuff like that

Classical Statistics

- Classical statistics: Set of tools for decision making using hypothesis (Null hypothesis significance testing (NHST))
- Invented around 1910-1935
- Some of these tests you might have heard of: t-test or χ^2-test
- Although around 100 years: Problematic
- One disadvantage: For different types of problems different tests are being used

- Find your test:



Example: Hypothesis test

- `Free newspaper:` Zürich consumes more drugs than Basel!
- Statement based on the following table (EMPA):

Weekdays	Wed	Thu	Fri	Sat	Sun	Mon	Tue
Zürich	16.3	12.7	14.0	53.3	117	62.6	27.6
Basel	10.4	8.91	11.7	29.9	46.3	25.0	29.4

- Values: Percentage of ecstacy in waste water per 1000 inhabitants per day in mg
- Statement seems to be clear Population of Zürich consumes more drugs than the one Basel

- But what does “obvious” mean?
- How can one decide *mathematically* whether the statement is correct?
- Fictitious table:

Weekdays	Wed	Thu	Fri	Sat	Sun	Mon	Tue
Zürich	16.3	12.7	12.0	33.3	117	62.6	27.6
Basel	15.4	13.91	15.7	29.9	86.3	55.0	29.4

- Statement no longer clear: „Guys form Zürich rather consume more drugs than the ones from Basel”
- How can we mathematically describe the boundary between „definitely” clear and „not so” clear?
- Remark: There is no such thing as „definitely” clear, only „very likely” clear

- A lot of hidden assumption
- But: Introduce a few tests (t-test, Wilcoxon-test)
- Need to know what a p-value or a confidence interval is

Bayesian Statistics

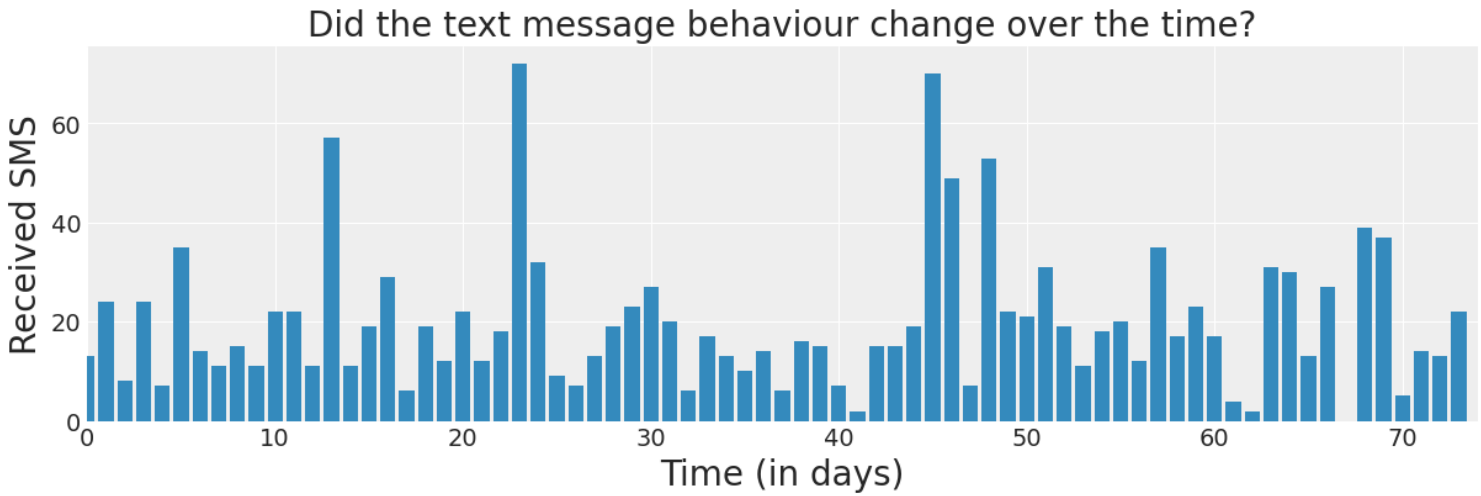
- One focus of this module: Bayesian statistics

> *When the Facts Change, I Change My Mind. What Do You Do, Sir?*
>
> John Maynard Keynes, Economist

- Bayesian statistics: Unified approach to statistics
- Much more natural approach of statistics for machine learning
- Have a model, collect data and "learn" from them using so-called *Bayesian inference*
- Very intuitive and is mathematical version of how we think (usually)

Example: SMS

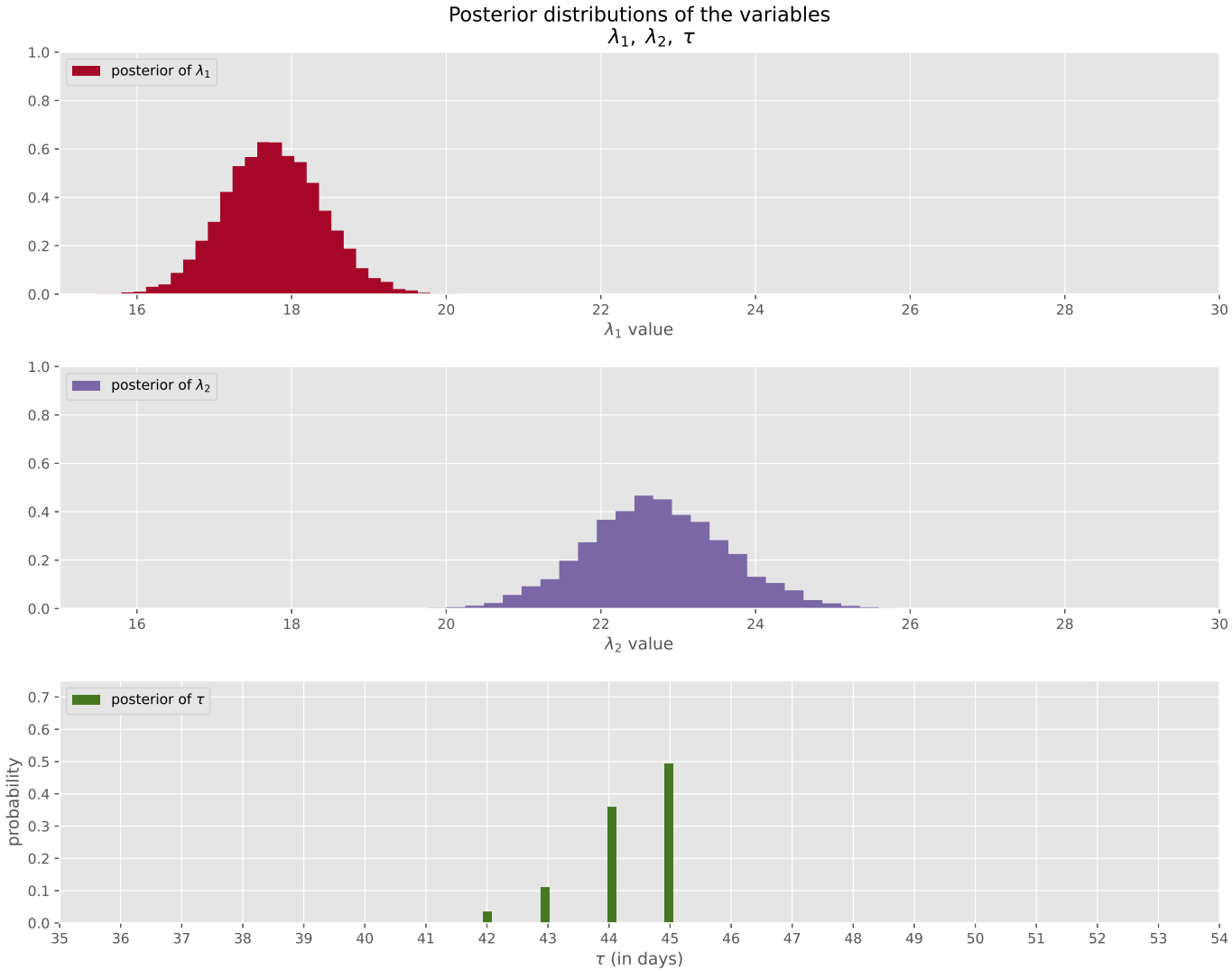
- Receive a series of daily text messages (SMS)
- Figure:



- Questions about the average number of SMS received
- Figure: Seems that more SMS arrive in the 2nd part from about day 40 onwards than in the 1st part

- Questions about the two averages and the day when it changes from one to the other average

- Using Bayes inference (all details later):



- Denote with λ_1 and λ_2 the mean values for number of received SMS in the first and second part of the observed period
- With τ: day where behaviour of received SMS changes, i.e. changes from λ_1 to λ_2
- Important: Bayesian inference does not simply return values for the λ's and τ, but *distributions*
- These contain much more information than simple numerical values
- What have we gained?
- Immediately see uncertainty in estimates: The wider the distribution is, the more uncertain our posterior belief should be

- Also see which values for parameters are plausible: λ_1 is at 18 and λ_2 is at 23
- Posterior distributions of the two λs are clearly different .
- Indicates that it is indeed likely that the user's text messaging behaviour has changed
- Distribution for τ: Posterior distribution looks a little different from the other two because discrete random variable (no w'kings for intervals)
- near day 45 has a 50 % chance of changing behaviour of user

- If no change had occurred or if the change had been gradual over time, posterior distribution of τ would have been more widely dispersed
- Reflects that many days would be plausible candidates for τ
- But in fact only three or four days make sense as potential transition points

Models

- Will use models *a lot*

All models are wrong, but some are useful.

George Box, Statistician

- Real world seems messy at times
- Models: Used to simplify nature of things
- Use them all the time without realising it
- Models: Essential part of this module (and statistics)

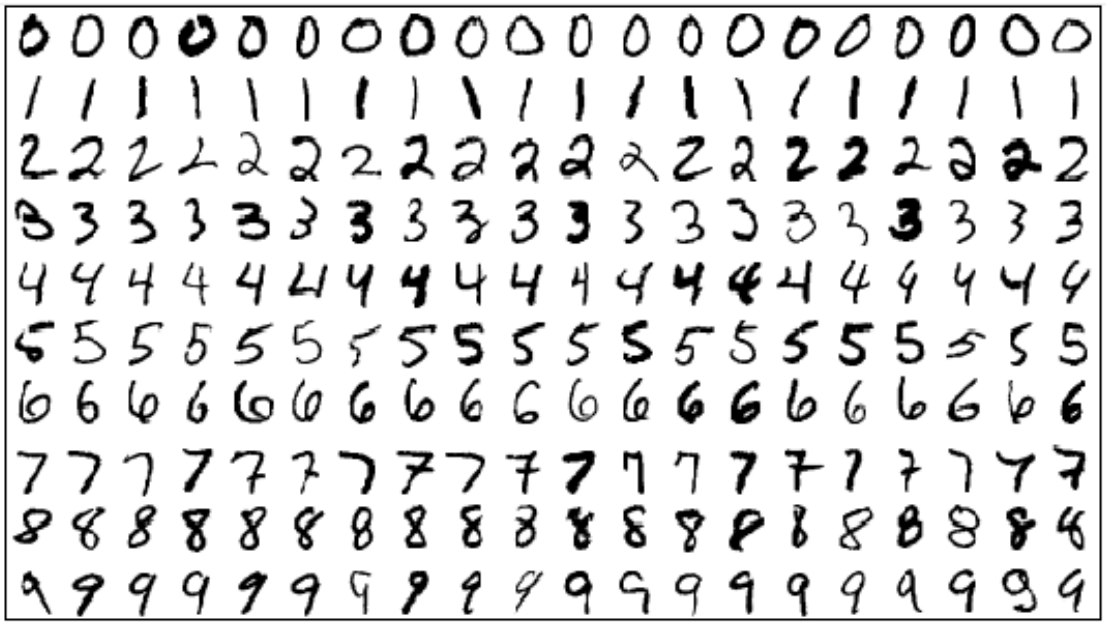
Non-Statistical Example

- Part of MRT (Mass Rapid Transit) map of Singapore

- Models some features of Singapore: Train lines and stations
- But omits almost all other geographical features
- Nobody can pretend that this a is proper map (also a model) of Singapore
- For example: Circle Line (yellow) in reality certainly not perfect circle
- Distances between stations are not displayed correctly
- However very useful map if you want to get with train from, say, Changi Airport to city center, say, Orchard
- Map is completely useless if you want to do same trip by car
- Models are not simply useful but only useful in a certain context

Example: Handwritten Digits

- Figure: Some handwritten digits



- Consider digit 8
- What makes an 8 an 8? What is essence of "eightness"?
- No problem for us humans: Recognise an 8 when we see one

- But how can computers recognise an 8?
- Define a model of an 8
- Maybe simplest model: Two circles on top of each other

- Unfortunately: Real 8's are never as simple: 8 of this font on the right
- Includes more features to enhance readability
- But model a good starting point to tell a computer what an 8 should look like

Example: Coin

- We all "know": Probability of tossing head with a fair coin is 0.5
- But there is catch: There is *no* fair coin in real life
- A perfect fair coin is absolutely symmetrical and has zero thickness (to avoid coin ending up on edge)
- Such a coin does *not* exist in real life
- A fair coin is a mental model of imperfect but almost fair coins
- If flipping a real coin which is almost symmetrical, probability of head showing up, will be, say, 0.497 or 0.500000132, but *never* exactly $\frac{1}{2}$
- Will see: Model *fair coin* is very useful

Statistical Example

- Dataset **`birthweight.csv`**: Information of newborn babies and their parents

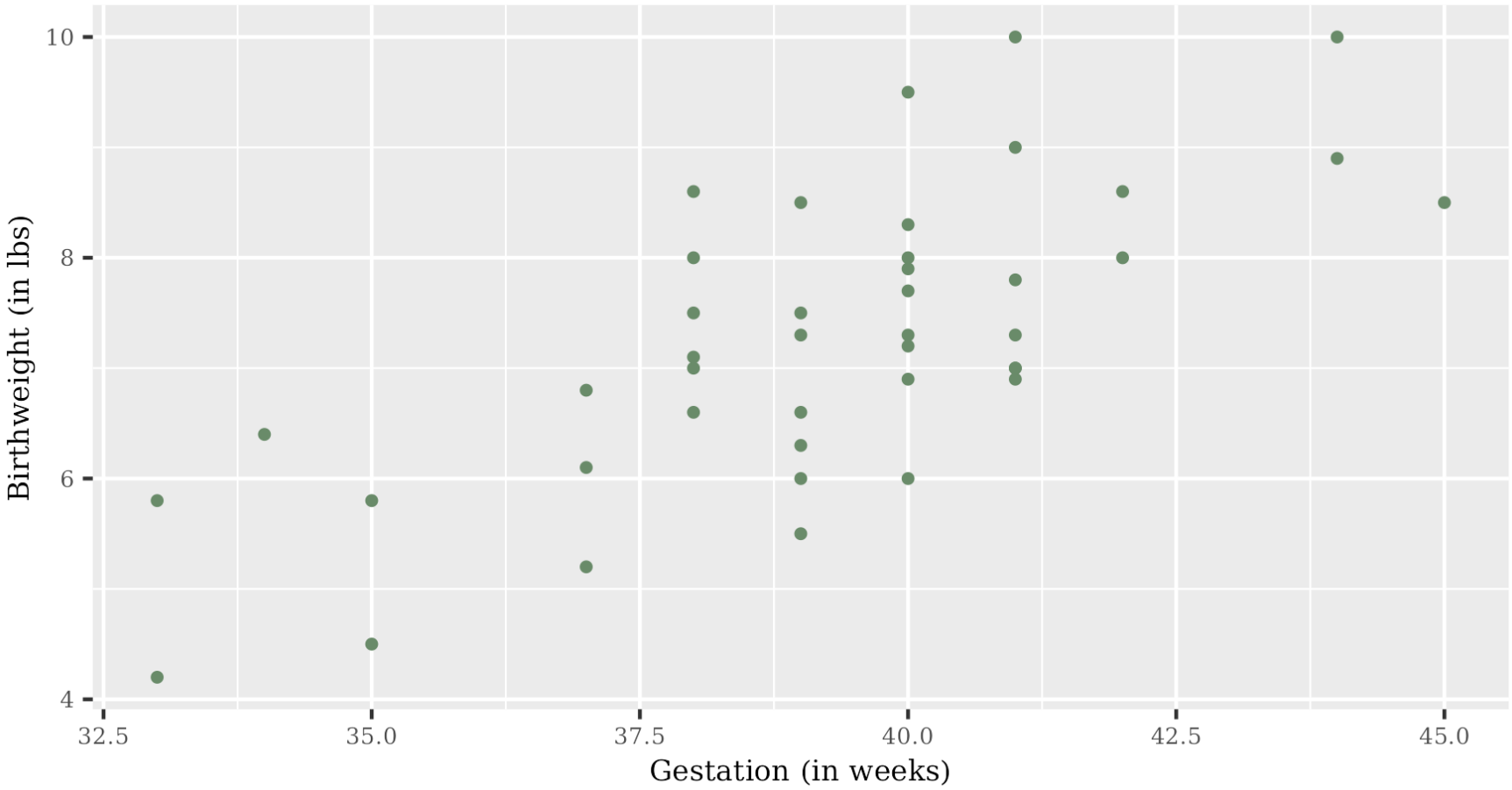
```
birthweight <- read.csv("birthweight.csv")
head(birthweight)
```

```
    id headcirumference length Birthweight Gestation smoker motherage mnocig
1 1313              12     17         5.8        33      0        24      0
2  431              12     19         4.2        33      1        20      7
3  808              13     19         6.4        34      0        26      0
4  300              12     18         4.5        35      1        41      7
5  516              13     18         5.8        35      1        20     35
6  321              13     19         6.8        37      0        28      0
  mheight mppwt fage fedyrs fnocig fheight lowbwt mage35 LowBirthWeight
1      58    99   26     16      0      66      1      0            Low
2      63   109   20     10     35      71      1      0            Low
3      65   140   25     12     25      69      0      0         Normal
4      65   125   37     14     25      68      1      1            Low
5      67   125   23     12     50      73      1      0            Low
6      62   118   39     10      0      67      0      0         Normal
```

- Question: Influence these variables birth weight (in lbs $= 0.454\,\text{kg}$) of babies?

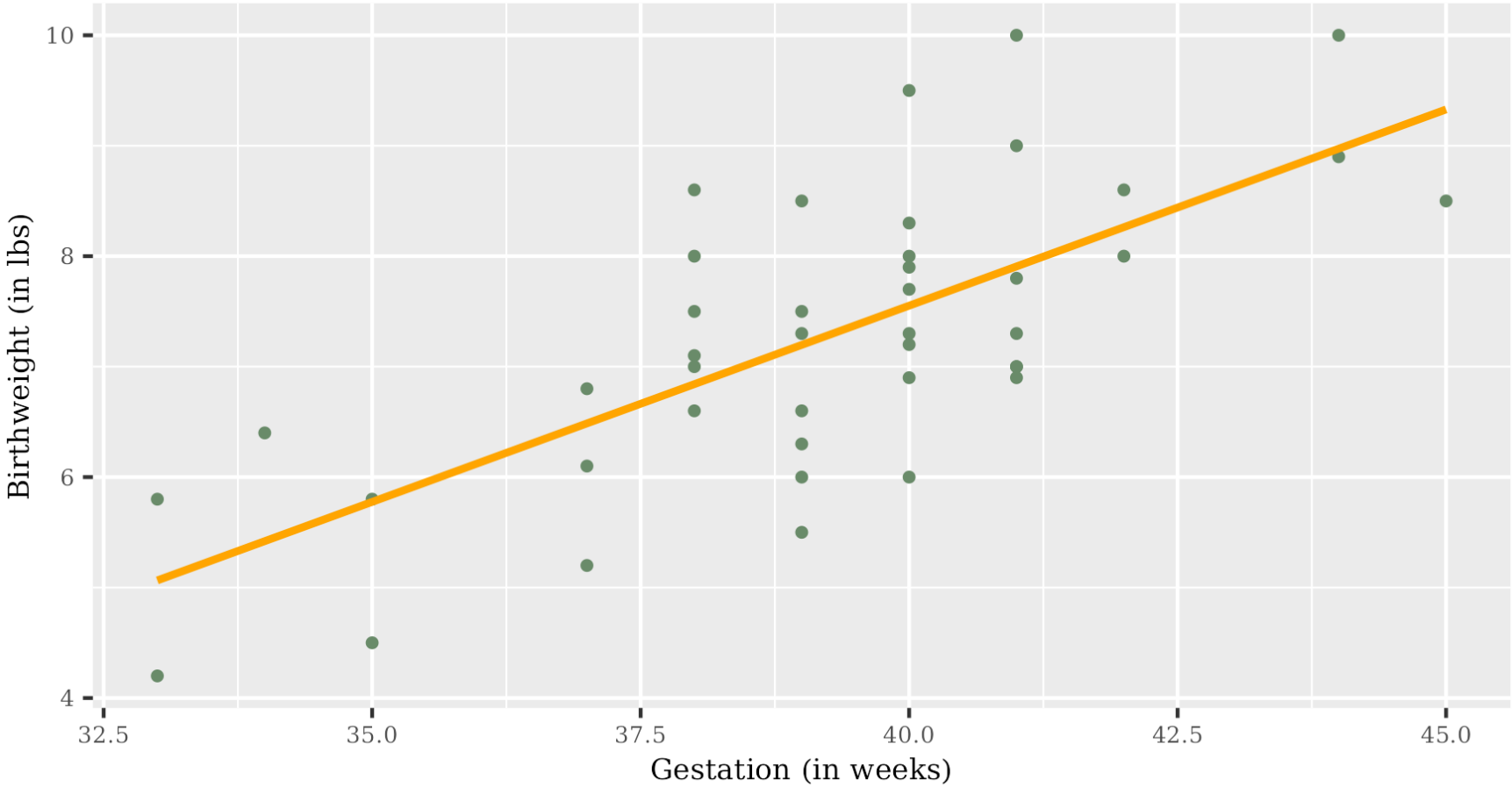
- Look at obvious relationship between `Gestation` and `Birthweight`
- It can be assumed that for longer pregnancies babies are heavier
- But how can we describe this relationship?

- Plot:


Birthweight (in lbs)
10
8
6
4
32.5
35.0
37.5
40.0
42.5
45.0
Gestation (in weeks)

- Plot: Confirms belief that a longer pregnancy results in heavier babies
- Because dots have tendency to ascend from left to right
- To get a quantitative result: Determine regression line (if you never heard about that one, do not worry, you will)

- Regression line:



- Equation of blue regression:

$$\texttt{Birthweight} = -6.660 + 0.355 \cdot \texttt{Gestation}$$

```
coef(lm(birthweight$Birthweight ~ birthweight$Gestation))

          (Intercept) birthweight$Gestation
           -6.6601895              0.3553025
```

- *No* data point lies exactly on line but are close to it
- Regression line: Does not describe individual points *exactly but approximately*
- Regression line is a model for data points
- Equation: If pregnancy lasts, say, 41 weeks, we can *expect* that baby has a weight of about 7.895 lbs:

$$\texttt{Birthweight} = -6.660 + 0.355 \cdot 41 = 7.895$$

- Of course, other factors have also an influence on the birth weight
- A first and certainly not last note of caution concerning models
- Interpretation of intercept -6.66: If pregnancy lasts 0 weeks, baby has weight of -6.66 lbs
- Hmmmm...: What's going on?
- Models have their limitations

Summary

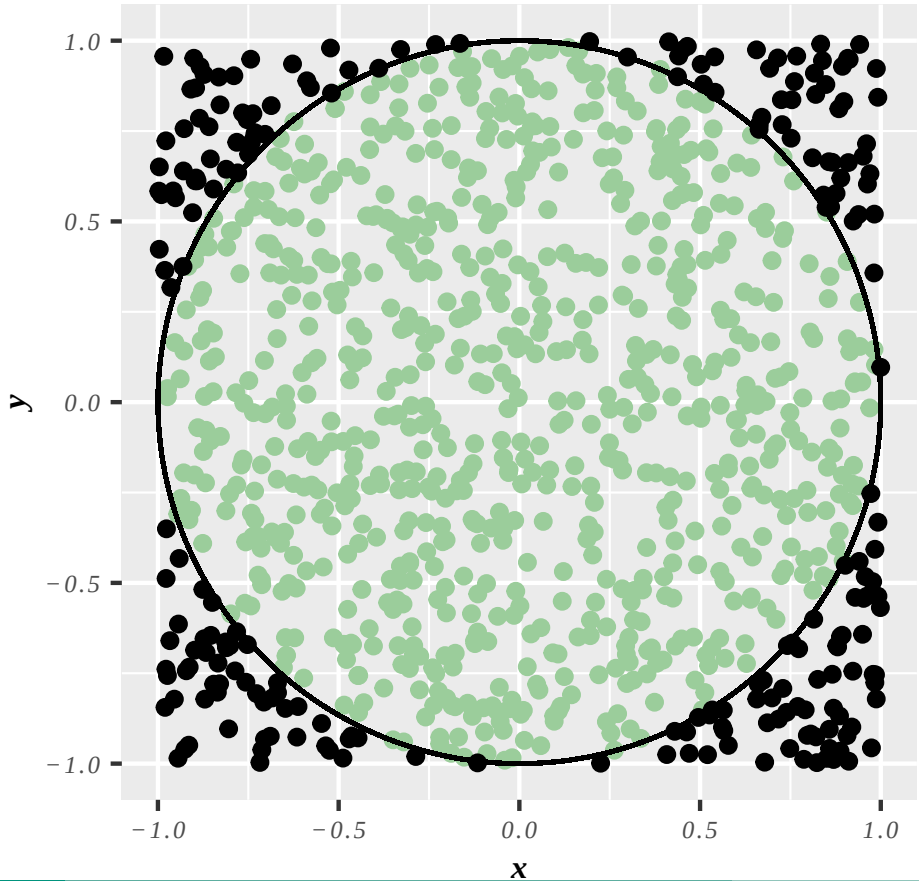
- Models are statements about the operation of nature that purposefully omit many details and thus achieve insight that would otherwise be discursively obscured
- They provide unambiguous statements of what we believe is important

Simulations

- *Very* useful tools in applied statistics are simulations and sampling
- These are used to approximate quantities for which exact solution is very difficult or even impossible to determine
- Simulations rely heavily on computer power

Example: Approximation of π

- Want to *approximate* $\pi = 3.14159\ldots$ using a simulation
- Principle is very simple
- Take a circle with radius 1, circumscribed by a square with side 2



- Throw dart 1000 times randomly on square:
  - Mark points where dart landed with green dot if dart is in circle
  - With black dot otherwise
- It can be shown (see Example **??** in lecture notes):

$$\pi \approx 4 \cdot \frac{\text{Number of green dots}}{\text{Total number of dots}}$$

- Symbol $\approx$ stands for „approximately”
- Of course: Do not use real dart, but R instead
- Code is not important for now but shows R at work
- Code will be explained in Problems

- Code:

```
set.seed(2)
n <- 1000

x <- runif(n, min = -1, max = 1)
y <- runif(n, min = -1, max = 1)

r_squ <- x^2 + y^2

sum(r_squ < 1) / n * 4

[1] 3.004
```

- Approximation of π is 3.004 which is not a particularly good
- If dart is thrown 1000 times again Get different result, because dart is thrown randomly

- Code below: Simulates 10 times these 1000 random throws

```
n <- 1000

for (i in 1:10)
{
  x <- runif(n, -1, 1)
  y <- runif(n, -1, 1)
  r_squ <- x^2 + y^2
  pi <- sum(r_squ < 1)/n*4
  cat(pi," ")
}

3.196  3.156  3.104  3.132  3.18  3.052  3.224  3.104  3.096  3.12
```

- To improve approximation: Increase number of throws, say 1 000 000 (replace `1000` in first line above by `1000000`)

```
3.140788  3.140772  3.14478  3.14116  3.140268  3.140312  3.140868
3.141224  3.140916  3.137056
```

- Approximations closer to π than before, but still not very good
- This algorithm is not very effective but shows how simulations work quite nicely
- This kind algorithm: *Monte Carlo algorithm*, because random numbers are used

Practical Relevance

- Will see: Often very difficult or even impossible to get an *exact* result for a problem
- Aim: To *approximate* result
- Quite often: Not even interested in exact results, because approximation is good enough
- Speak of *practical relevance* or *significance*
- For example, to calculate area of a circle: Do not need exact value of π, which is unknown anyway, but a good enough approximation of π

Data

- Data and statistics: Becoming more and more influential
- Newspaper: Predictions for next election
  $\rightarrow$ Based on polls
- Google: How does Google come up with what we want to search?
  $\rightarrow$ Google evaluates search queries, gets better with each query
- Passport control at the airport: How does software "recognise" faces?
  $\rightarrow$ Faces are characterised using statistics
- Weather report: How do forecasts come about?
  $\rightarrow$ Model based on previous weather data (and theory)
- Stock prices: How to predict the price of a share for next few days based from stock market performance of last few days?
  $\rightarrow$ Modeling from old data

Datasets (One-Dimensional)

- *List*: Simplest kind of a dataset
- Example: Body heights of 5 people (in m)

$$1.75, \quad 1.80, \quad 1.72, \quad 1.65, \quad 1.54$$

- Such lists: *One-dimensional datasets* or *measurement series*

Datasets (Two-Dimensional)

- Most common form of datasets: *Tables* or *two-dimensional datasets*
- Example:

Person	Height	Weight	Gender	Nationality
A	1.82	72	m	CH
B	1.75	82	f	D
C	1.61	70	f	CH
D	1.80	83	m	A
E	1.89	95	f	FL

- Height and weight: *Quantitative* data, i.e. numbers (measurements)
- Can take, at least theoretically, any numerical value within an interval of number line
- Gender and nationality: *Qualitative* data
- Can only take certain number of values (don't have to be numbers)

Exploratory Data Analysis (EDA)

- Sometimes called *descriptive statistics*
- Subject of EDA: Representation of datasets
  - Characterise datasets by certain numbers (e.g. average)
  - *And* display data graphically
- First: *One dimensional* data: *One* measurement is determined on a test object (two-dimensional later)

Aim of EDA

- *Summarise* data by *numerical parameters*
- *Graphical representation* of the data

Dataset: Body Weight

- Measurements of body weight
- Experience: Stand on scale in the morning and record weight
- Stand on scale again and get a slightly different result
- Use 80 kilogram metal block, which is calibrated, i.e. it has with very high accuracy a weight of 80 kg
- Weight of metal block: Measured several times with scales A and B
- Two *datasets* (in kg, accurate to 10 g)

- Table:

Scale *A*	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Scale *A*	80.03	80.02	80.00	80.02					
Scale *B*	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

- First of all: Why do different measurements on the same object lead to different results?
- Measurements never take place under *exactly* the same conditions
- Temperature and/or humidity and so on can change
- Apparently exact data are only *approximate* data
  - Calorie declaration on a bar of chocolate
  - Content 500 ml of can: No two cans have *exactly* the same shape
  - Face recognition at the airport: You *never* have the *same* facial expression

Back to the Scales Example

- Measurements were carried out with the greatest care
- Nevertheless values vary within both scales
- We may ask:
  - Is there a *difference* between scale A and scale B?
  - If so, how to determine this difference?
- Take a closer look at Table slide 65:
  - Both scales: Values are around 80 (which it should be)
  - Scale A: Only 2 values of 13 *below* 80
  - Scale B: Only 2 of 8 values *over* 80
  - Values of scale A are therefore *rather* larger than those of scale B

- But what does “rather” mean?
- How can the two datasets be compared with each other?
- Aim: To somehow *summarise* the datasets in order to be able to compare the two scales with each other
- First idea (of course): Compare averages (means)
- *EDA* deals with ways to organise and summarise data
- Goal: Simplify interpretation and subsequent statistical analysis of these data

- Key figures: Summarise data numerically to characterise them roughly
- Statistical analysis: Very important not to blindly adapt a model or apply *one* statistical procedure
- Data should (if possible) *always* be graphically displayed *and* compared with corresponding key figures
- Only in this way can one discover (sometimes unexpected) structures and special features of the data
- But:

Be careful!!!

Whenever a dataset is "reduced" (by key figures or graphics), *information is lost*!

Example

- Fictitious grades of an exam with 24 students:

  4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

- Mean of grades: 4.51
- Average tells something about class as a whole, but nothing about individual students
- If only known mean 4.51: No information how each student performed
- Do not even know how many students attended exam
- *Gain information of the whole, but loose information of the parts*

Notation

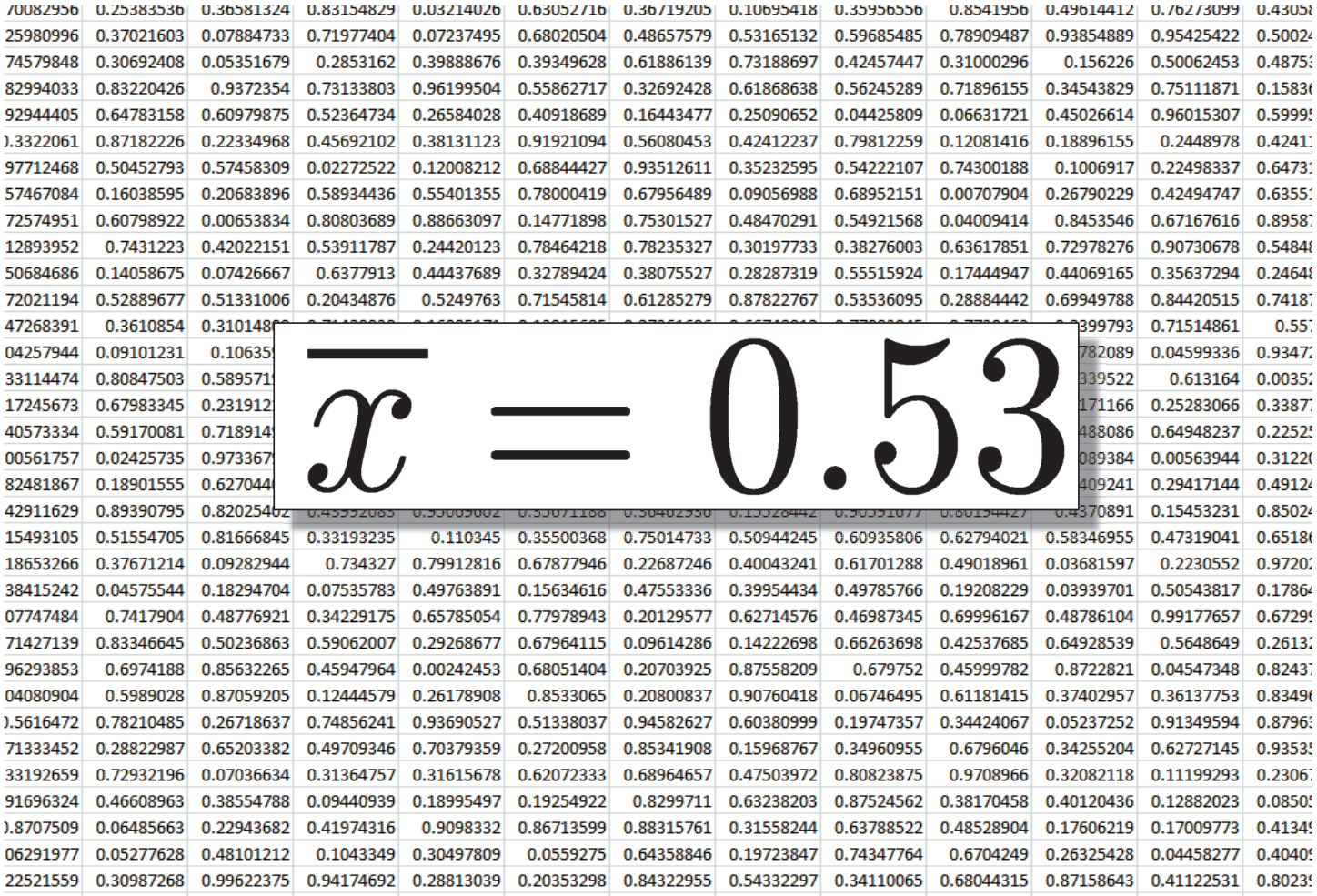
- Standard notation for individual data (points):

$$x_1,\ x_2,\ \ldots,\ x_n$$

- n: *Scope* of the (one-dimensional) dataset
- Example of dataset scale A: Scope $n = 13$:

$$x_1 = 79.98,\quad x_2 = 80.04,\quad \ldots,\quad x_{13} = 80.02$$

Key Figures



Overview of Key Figures

- Distinguish between location and spread (dispersion) parameters
- *Location parameters* ("*Where* are the observations on the measuring scale?")
  - Arithmetic mean ("average")
  - Median
  - Quantile
- *Spread parameters* ("*How* do the data scatter around their central location?")
  - Empirical variance / standard deviation
  - Interquartile range

Arithmetic Mean

- Colloquially: *Average* or simply *mean*
- Add up all data and divide it by number of data (scope)
- Definition:

Arithmetic mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

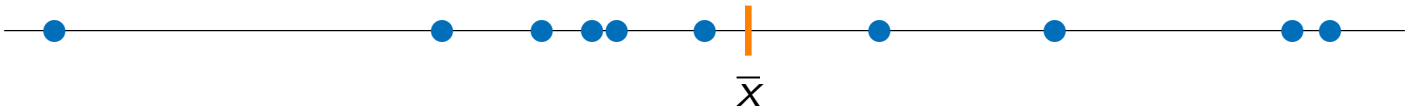
- Say: "x bar"
- Example scale A: Arithmetic mean of $n = 13$ observations

$$\bar{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.020\,77$$

- R command `mean(...)`:

```
scaleA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
            80.03, 80.02, 80.00, 80.02)

mean(scaleA)

[1] 80.02077
```

- Illustration of mean:



Comparison of Scales

- Arithmetic mean for scale *B*:

```
scaleB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)

mean(scaleB)

[1] 79.97875
```

- Scale *B* has thus *on average* lower values than scale *A*
- Comparison possible even though scopes are different

Example: Grades

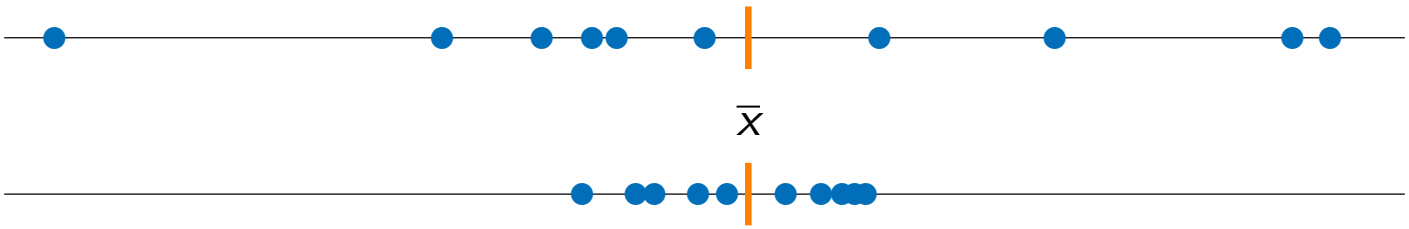
- Mean of the grades:

```
grades <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5,
            5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)

mean(grades)

[1] 4.5125
```

- Average score is 4.51 (rounded to two decimal places)

Spread Graphically

- Mean tells a lot about a dataset: "Center" of data
- But: Average does not tell whole story about (quantitative) datasets
- Graphical example:



- Both datasets have same mean
- Points of second dataset "on average" much closer to mean $\overline{x}$ than points of first dataset
- Say: "Different *spread* of data around mean"
- Second dataset has *less* spread than first

Spread Numerically

- Example of (fictitious) grades in school:

$$2,\ 6,\ 3,\ 5 \qquad \text{and} \qquad 4,\ 4,\ 4,\ 4$$

- Both mean 4, but spread of data around mean quite different
  - 1st case: Two good and two poor students
  - 2nd case: All students equally good
- Datasets have a different *spread* around mean
- Second dataset has spread 0, first dataset has (positive) spread
- How can this spread be described numerically?
- Want to describe "difference" to mean

- 1st idea: Take average of *differences to mean*
- 1st case:

$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = \frac{-2+2-1+1}{4} = \frac{0}{4} = 0$$

- 2nd case 0 as well:

$$\frac{(4-4)+(4-4)+(4-4)+(4-4)}{4} = \frac{0+0+0+0}{4} = \frac{0}{4} = 0$$

- No statement about *different* spread
- This idea is of no use
- Problem: Differences can become *negative*
- Numbers in numerator can cancel each other out

- Next idea: Replace differences with *absolute values* of differences
- 1st case:

$$\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = \frac{2+2+1+1}{4} = 1.5$$

- I.e.: Grades deviate on average 1.5 grades from mean 4
  r
- 2nd case:

$$\frac{|(4-4)|+|(4-4)|+|(4-4)|+|(4-4)|}{4} = \frac{0+0+0+0}{4} = 0$$

- 0, as it should be: Dataset has no spread
- The greater this value (always greater than or equal to 0), the more the data differ from each other with the same mean value
- This definition for spread: *Average absolute deviation*
- But: Theoretical and numerical disadvantages

Empirical Variance and Standard Deviation

- Better: *Empirical variance* and *empirical standard deviation*
- Measure of variability or spread of observations
- Definition:

Empirical variance *var*(x) and standard deviation s_x

$$\text{Var}(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_x = \sqrt{\text{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Properties of Variance

- *Squaring* differences $x_i - \bar{x}$: Differences can't cancel each other out
- Denominator $n - 1$, instead of n: Mathematical reasons
- In some books: n instead of $n - 1$
- But: Difference negligible for large n (see lecture noters)
- Equal in terms of practical relevance
- Standard deviation is root of variance
- Root: Standard deviation has same unit as data itself

- Example:
  - Data in cm
  - Unit of variance due to squaring of data is cm^2
  - Having same unit as original data: Take root of variance
- If empirical variance (and thus standard deviation) is large, spread of values around mean is large
- *Value* of empirical variance has no physical interpretation
- Only known: The larger the value the greater the spread
- Important: *Only standard deviation s_x can be interpreted in a natural way*
- For normally distributed data: Standard deviation has a nice geometric interpretation (see later)

Example: Scale A

- Mean of $n = 13$ observations is $\bar{x} = 80.02$ (see slide 73)
- Empirical Variance:

$$\text{Var}(x) = \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1}$$
$$= 0.000\,574$$

- Empirical standard deviation:

$$s_x = \sqrt{0.000574} = 0.024$$

- "Average" deviation from mean 80.02 kg is 0.024 kg

- Calculation by hand very tedious
- Easy with R:

```
var(scaleA)

[1] 0.000574359


sd(scaleA)

[1] 0.02396579
```

- sd: Standard deviation

Median

- Another measure for "central location": *Median*
- Very simplified: Value where half of observations are below or equal to this value
- Other half is equal to or greater than this value
- Example: Results of exam at school has median 4.6
  - I.e.: Half of class has *this grade or lower*
  - Conversely: Other half of class has *this grade or higher*
- Interpretation above: Median very simplified
- Exact definition now follows

Ordered Dataset

- Sort values in dataset in ascending *order*
- Notation of *ordered data* with $x_{(i)}$:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- Round brackets in index: Data in ascending order
- For calculation of *median*: Sort data in ascending order first

Example: Scale A

- Data scale A ordered:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Median is now very easy to determine
- Among these 13 observations: Value of central observation
- Median: Value of 7th observation:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Median of dataset scale A is 80.03
- Nearly half observations, namely 6, are smaller or equal to 80.03
- Conversely: 6 observations are greater than or equal to median
- Odd number of observations: *Exactly* one central value

Example: Scale B

- Before: For odd scope of data, central observation is unique
- Scope of data even: *No* single central observation
- *Define* median: Mean of the *two* central observations
- Example: Dataset of scale B has 8 observations
- Order dataset: Median is average of 4th and 5th observations

$$79.94,\ 79.95,\ 79.97,\ \underbrace{79.97,\ 79.97}\ ,\ 79.94,\ 80.02,\ 80.03$$

$$\frac{79.97 + 79.97}{2} = 79.97$$

- R–Command

```
median(scaleA)

[1] 80.03

median(scaleB)
[1] 79.97
```

- Median does not have to be in dataset
- Example:

```
grades <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7,
            5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)

median(grades)

[1] 4.65
```

- Grades are given to one decimal place
- Median is mean of 4.6 and 4.7

Median vs. Mean

- Two location parameters for "center" of a dataset
- Which one is "better"?
- Wrong question in applied statistics
- There is no "better":
  - Depends on specific problem
  - Best practice: Consider both location parameters simultaneously
- Important property of median: *Robustness*
- Median is much less influenced by extreme observations than mean

Example: Scale A

- Example: For largest observation ($x_9 = 80.05$) of scale A, a typo occurred and $x_9^* = 800.5$ was used
- New mean: Instead of 80.02:

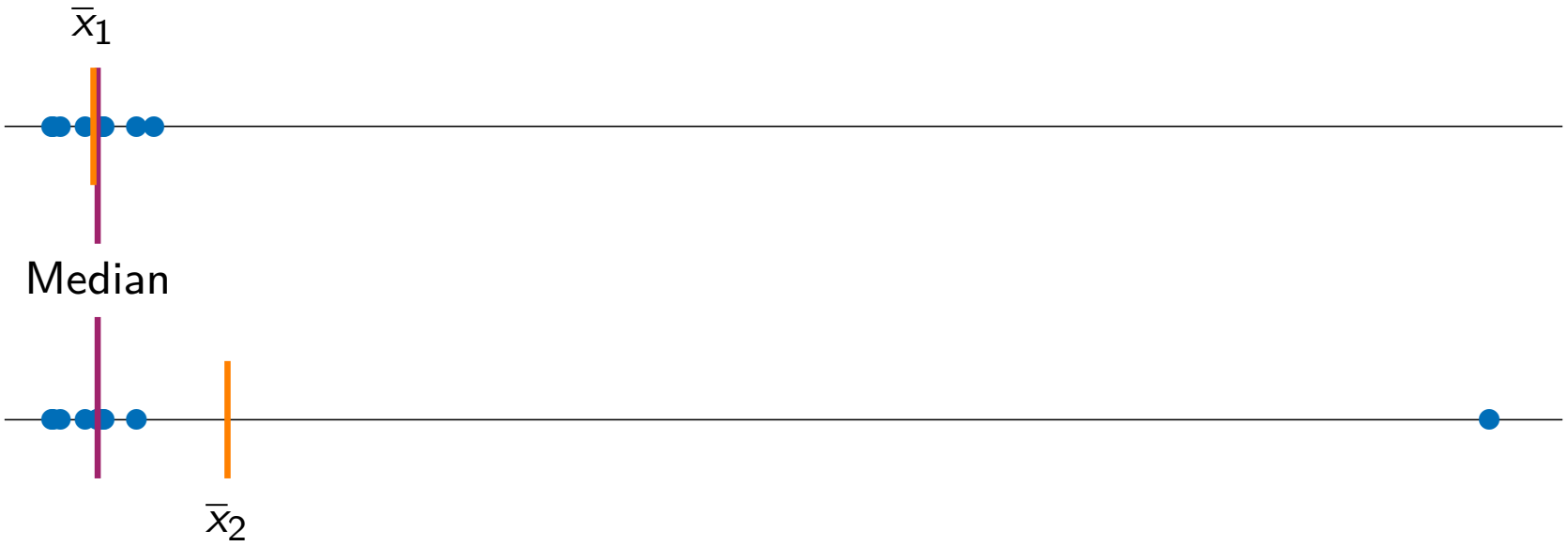
$$\bar{x}^* = 135.44$$

- But median is still

$$x^*_{(7)} = x_{(7)} = 80.03$$

- Mean: Very strongly influenced by one extreme observation
- Median remains the same in this case: Robust

Graphically



Example: Income

- Typical household income of Seattle suburbs around Lake Washington
- Average income of Medina and Windermere will be very different
- Reason: Bill Gates lives in Medina
- Generally: Median is used for "middle" income and not mean, as this is fairer
- If Bill Gates moves from Medina to Windermere, average income changes, but nobody else benefits in Windermere
- Conversely: People of Medina still have same income, although average income has fallen

Remark

- Median: Also called *central value* or *average value* (not to be confused with the mean value)
- Exact interpretation of median is surprisingly difficult
- Good enough for us: (About) half of values are less than or equal to median, other half is greater than or equal to median

Quartiles

- Median: Value where half of observations are less or equal to this value
- Similarly: Lower and upper quartile
- Lower quartile: Value where 25 % of all observations are less or equal and 75 % are greater or equal to this value
- Upper quartile: Value where 75 % of all observations are less or equal and 25 % are greater or equal to this value
- Also called: 1st and 3rd quartile
- *Caution*: Most of the time there is no *exact* 25 % of observations
- *Define* value for lower quartile or upper quartile

Example: Scales

- *One* way to define quartiles
- Scale A: $n = 13$ observations: $25\,\%$ of which is 3.25
- *Choose* next higher value $x_{(4)}$ as lower quartile:

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Lower quartile is 80.02
- Nearly a quarter of observations are equal to or less than 80.02
- Upper quartile: *Choose* $x_{(10)}$, because for $0.75 \cdot 13 = 9.75$ number 10 is next higher integer

79.97, 79.98, 80.00, 80.02, 80.02, 80.02, 80.03, 80.03, 80.03, 80.04, 80.04, 80.04, 80.05

- Almost three quarters of values are less than or equal to 80.04

Remarks

- There is no uniform definition for quartiles: R has 9 implemented
- Use default of R
- For large datasets: Practically irrelevant which definition is used
- Equal in terms of practical relevance
- Difference between respective quartiles negligible
- Important: Interpretation as seen above

- R does not have own commands for quartiles
- More general command `quantile` (soon)

```
# Syntax for lower quartile: p = 0.25
quantile(scaleA, p = 0.25)

  25%
80.02

# Syntax for upper quartile: p = 0.75
quantile(scaleA, p = 0.75)

  75%
80.04
```

Interquartile range

- *Interquartile range*: Measure for spread of data:

  $$\text{upper quartile} - \text{lower quartile}$$

- Measures length of interval that contains about half of the "central" observations

- The smaller this value, the closer half of central values are around median and the smaller the spread

- This spread parameter is robust

- Interquartile range of scale A:

  $$80.04 - 80.02 = 0.02$$

- R command IQR:

```
IQR(scaleA)

[1] 0.02
```

- (About) half of middle observations in a range of length 0.02
- Boxplot: Visual interpretation of interquartile range

Example

- Grades from 24 students:

$$4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1$$

- Calculate with R quartiles and interquartile range:

```
quantile(grades, p = c(0.25, 0.75))

  25%   75%
3.850 5.275

IQR(grades)

[1] 1.425
```

- (About) half of students are within 1.425 grades, between 3.85 and 5.275
- (About) 25 % of class has 3.85 or less, around 25 % of class 5.275 or more

Quantile

- Generalise quartiles to any other percentage: Quantile
- 10 %-quantile: Value where 10 % of observations are less than or equal to and 90 % of observations are greater than or equal to this value
- Definition analogous to quartiles
- Median: 50 %-quantile
- 25 %-quantile: Lower quartile
- 75 %-quantile: Upper quartile

- R: 10 %- and 70 %-quantile of scale A:

```
quantile(scaleA, p = .1)

   10%
79.984

quantile(scaleA, p = .7)

   70%
80.034
```

- About 10 % of observations are less than or equal to 79.97
- Accordingly: About 70 % of observations less than or equal to 80.04

Example

- Grades at exam in class with 24 students:

  4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

- Various quantiles with R:

```
quantile(grades, p = seq(from = .2, to = 1, by = .2))

 20%  40%  60%  80% 100%
3.66 4.26 4.98 5.54 6.00
```

- About 20 % of students are worse than or equal to 3.66
- Exactly 20 % of students not possible: 4.8 students
- 60 %-quantile: (About) 60 % of students are worse than or equal to 4.98