What can go wrong with statistics: Some typical errors & How to lie with statistics

Many slides borrowed from:

Lutz Prechelt

Daniel Huff

Jon Hasenbank

Motivation

"There are three kinds of lies:
Lies, Damned Lies, and Statistics."
– attributed to Benjamin Disraeli

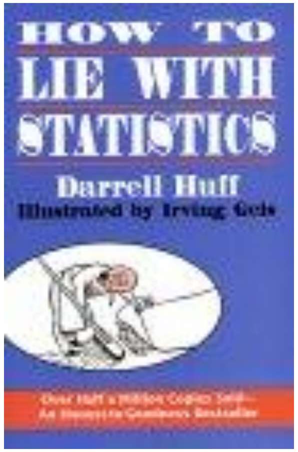
- Statistics are commonly used to make a point or back-up one's position
  - 82.5% of all statistics are made up on the spot.
- Three sources of errors:
  - If done in manipulative way, statistics can be deceiving
  - If not done carefully, statistics can be deceiving
    - Inadvertent methodological errors also will fool the person who is doing the statistics!
  - If not read carefully, statistics can be deceiving

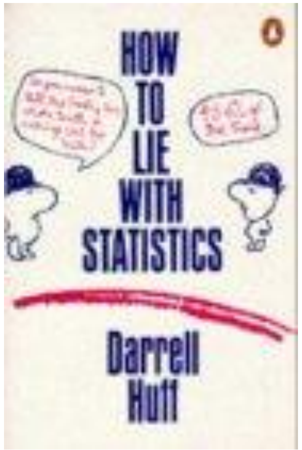
Purpose of this section

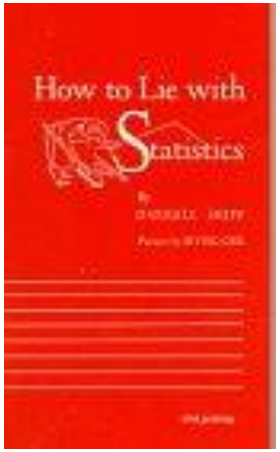
- Avoid common inadvertent errors
  - “Lessons for author”
- Be aware of the subtle tricks that others may play on you
  - (and that you should never play on others!)
  - “Lessons for reader”

Source

- Large parts of this slide set is based on ideas from Darrell Huff: "How to Lie With Statistics",
  (Victor Gollancz 1954, Pelican Books 1973, Penguin Books 1991)
  - but the slides use different examples
  - Most slides made by Lutz Prechelt
  - The book is short (120 p.), entertaining, and insightful
  - Many different editions available
  - Other, similar books exist as well







Example: Human Growth Hormone Spam (HGH)

GET HGH NOW!

Human Growth Hormone will add years to your life

Defy aging! As seen on CBS, NBC, The Today Show, and Oprah

Learn how now! click here for details

STOP THE AGING PROCESS WITH HGH!

*Body Fat Loss................ up to 82%

*Wrinkle Reduction......... up to 61%

*Energy Level.................. up to 84%

*Sexual Potency.............. up to 75%

*Memory......................... up to 62%

*Muscle Strength............. up to 88%

HUMAN GROWTH HORMONE WORKS!

Remark

- We use this real spam email as an arbitrary example
- and will make unwarranted assumptions about what is behind it
  - for illustrative purposes
  - I do not claim that HGH treatment is useful, useless, or harmful

Note:

- HGH is on the IOC doping list
  - http://www.dshs-koeln.de/biochemie/rubriken/01_doping/06.html
  - *"Für die therapeutische Anwendung von HGH kommen derzeit nur zwei wesentliche Krankheitsbilder in Frage: Zwergwuchs bei Kindern und HGH-Mangel beim Erwachsenen"*
  - *"Die Wirksamkeit von HGH bei Sportlern muss allerdings bisher stark in Frage gestellt werden, da bisher keine wissenschaftliche Studie zeigen konnte, dass eine zusätzliche HGH-Applikation bei Personen, die eine normale HGH-Produktion aufweisen, zu Leistungssteigerungen führen kann."*

Problem 1: What do they mean?

- "Body fat loss: up to 82%"
  - OK, can be measured
- "Wrinkle reduction: up to 61%"
  - Maybe they count the wrinkles and measure their depth?
- "Energy level: up to 84%"
  - What is this?
  - Also note they use language loosely:
    - Loss in percent: OK; reduction in percent: OK
    - Level in percent??? (should be 'increase')

Lesson for readers: What did they actually measure?

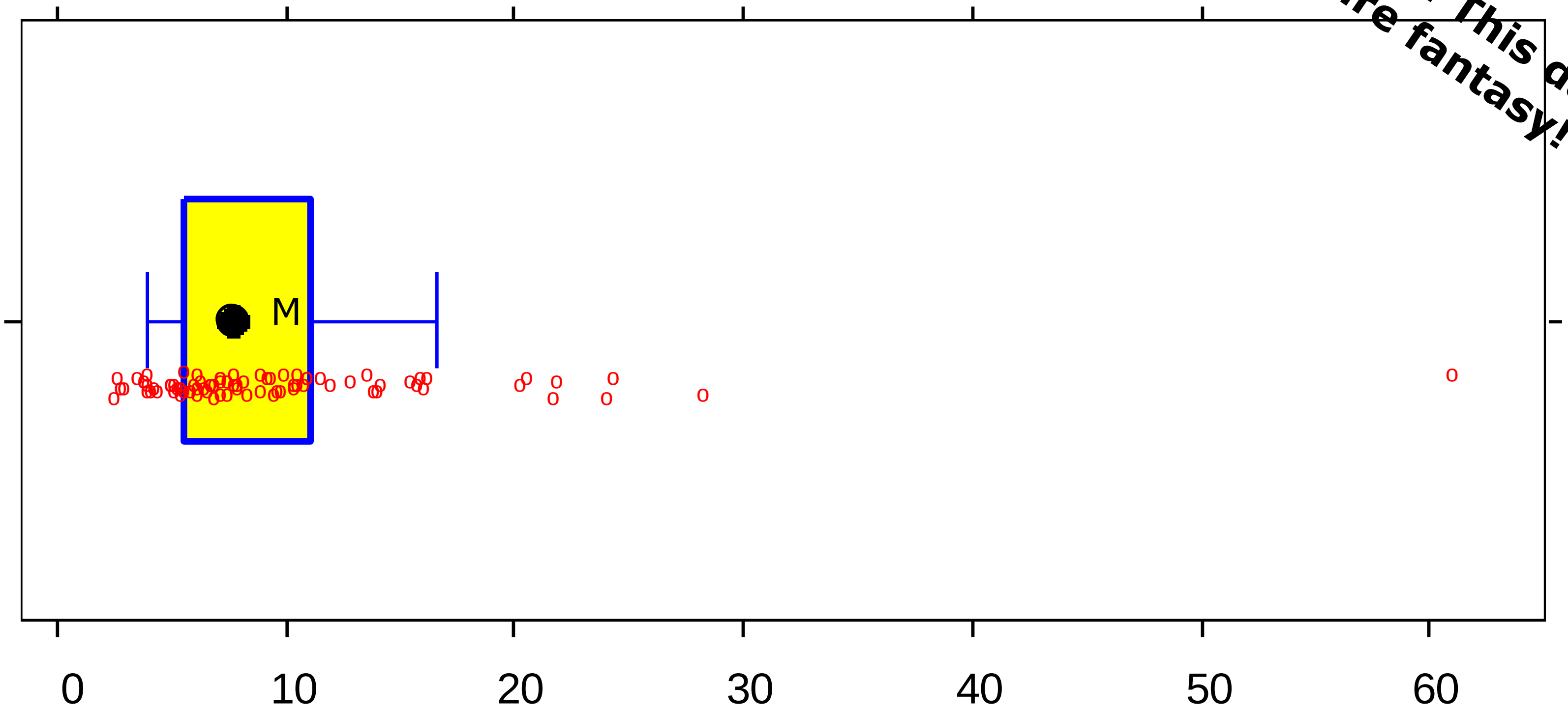
- Always question the definition of the measures for which somebody gives you statistics
  - Surprisingly often, there is no stringent definition at all
  - Or multiple different definitions are used
    - and incomparable data get mixed
  - Or the definition has dubious value
    - e.g. "Energy level" may be a subjective estimate of patients who knew they were treated with a "wonder drug"

Lesson for authors: Be clear about what you measure

- Before you start:
  - What effect do you want to analyze?
  - What could be good metrics to measure it?
  - Try out different metrics and compare them
- When writing things up:
  - Define your metrics clearly and understandable.
  - Bad example: “We analyzed the delays in our simulated network”.
    - One-way or RTT?
    - Total delays? But what if wire length is constant?
  - Good example: “We analyzed the one-way delays in our simulated network. Since propagation delays are constant in a wired network, we analyzed only the queueing delays and transmission delays.”

Problem 2: A maximum does not say much

- Wrinkle reduction: up to 61%
- So that was the best value. What about the rest?
- Maybe the distribution was like this:



Note: This data is pure fantasy!

Lesson for readers: Dare ask for unbiased measures

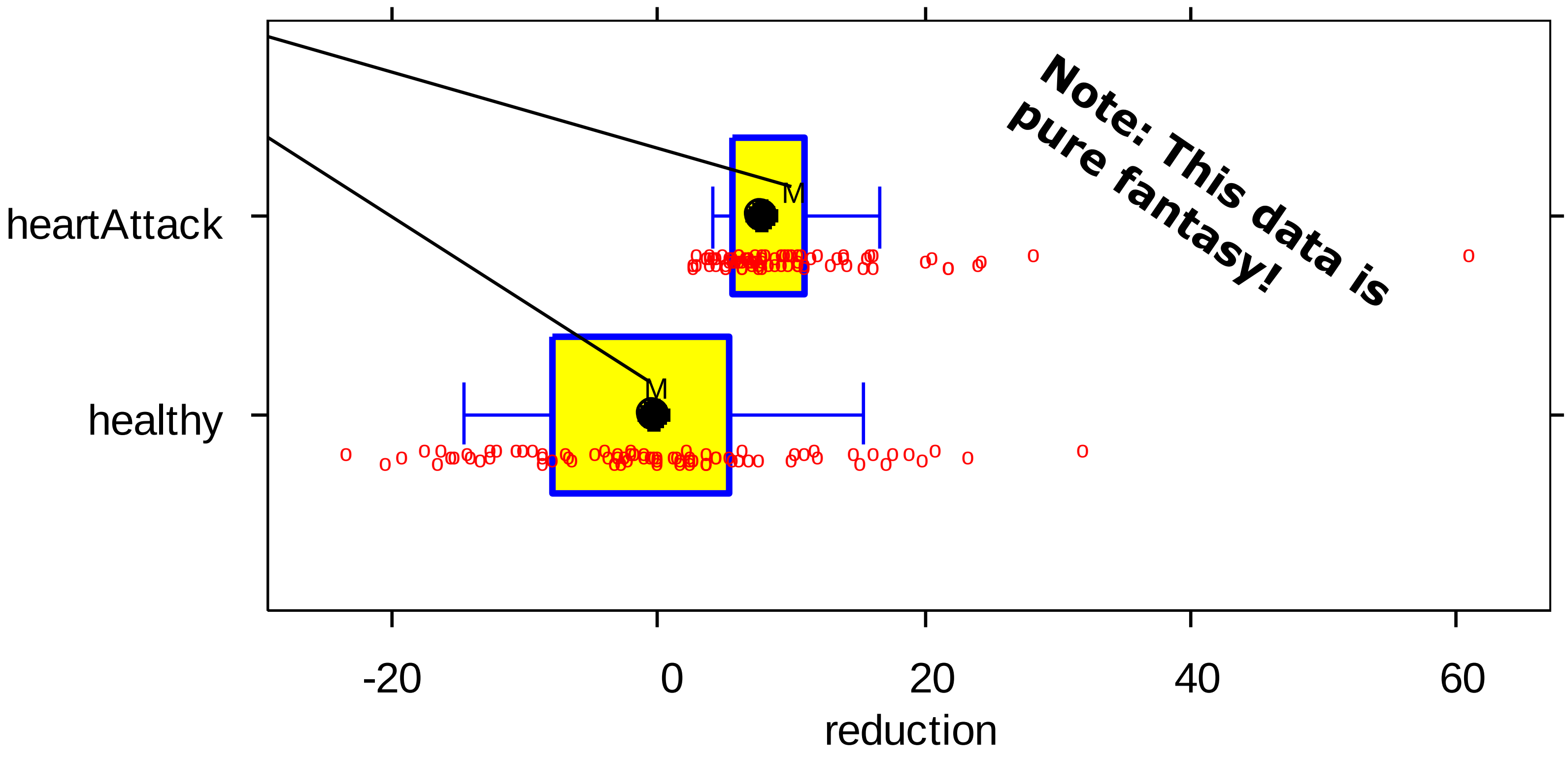
- Always ask for neutral, informative measures
  - in particular when talking to a party with vested interest
  - Extremes are rarely useful to show that someting is generally large (or small)
  - Averages are better
  - But even averages can be very misleading
    - see the following example later in this presentation
  - If the shape of the distribution is unknown, we need summary information about variability at the very least
    - e.g. the data from the plot in the previous slide has arithmetic mean 10 and standard deviation 8
  - Note: In different situations,
    rather different kinds of information
    might be required for judging something

Lesson for authors: Is it really significant?

- Are there many outliers?
- Do not use minimum or maximum values for comparison of, e.g., “before – after”
  - Compare the means
  - Think about what kind of mean to use:
    - Arithmetic mean?
    - Hyperbolic mean?
    - Geometric mean?
  - Better: compare the medians
- Or even better: Use statistical tests (e.g., Student’s t test) to prove that the change (before – after) is statistically significant

Problem 3: Underlying population

- Wrinkle reduction: up to 61%
- Maybe they measured a very special set of people?



Lesson: Insist on unbiased samples

- How and where from the data was collected can have a tremendous impact on the results
- It is important to understand whether there is a certain (possibly intended) tendency in this
- A fair statistic talks about possible *bias* it contains
- If it does not, ask.

Notes:

- A biased sample may be the best one can get
- Sometimes we can suspect that there is a bias, but cannot be sure

Lesson 4: 'Cum hoc ergo propter hoc' is wrong!

- Translation: "With this, therefore because of this"
- Meaning: Correlation does not mean causation
- Correlation may suggest causation (effect A causes effect B), but there also can be other reasons for a correlation between A and B

- Nitpicking: 'Post hoc ergo propter hoc' is almost the same thing:
  - After this, therefore because of this
  - Implies a temporal relation between A and B,
  - whereas 'cum hoc…' only implies some correlation

Correlation does not mean causation

- "If A is correlated with B, then A causes B"
  - Perhaps neither of these things has produced the other, but both are a product of some third factor C
  - It may be the other way round: B causes A
  - Correlation can actually be of any of several types and can be limited to a range
  - The correlation may be pure coincidence,
    e.g. #pirates vs. global temperature
  - Given a small sample, you are likely to find some substantial correlation between any pair of characters or events
- Ex: "Queueing delays increased, therefore throughput for individual TCP connections decreased"
  - Could be true
  - Could be due to an increased # of total TCP conections
  - Could be actually unrelated

Lesson: Question causality

- Sometimes the data is not just biased,
  it contains hardly anything else than bias

- If you see a presumably (=author) or assertedly (=reader) causal relationship ("A causes B"), ask yourself:
  - Does it really make sense?
  - Would A really have this much influence on B?
  - Couldn‘t it be just the other way round?
  - What other influences besides A may be important?
  - What is the relative weight of A compared to these?

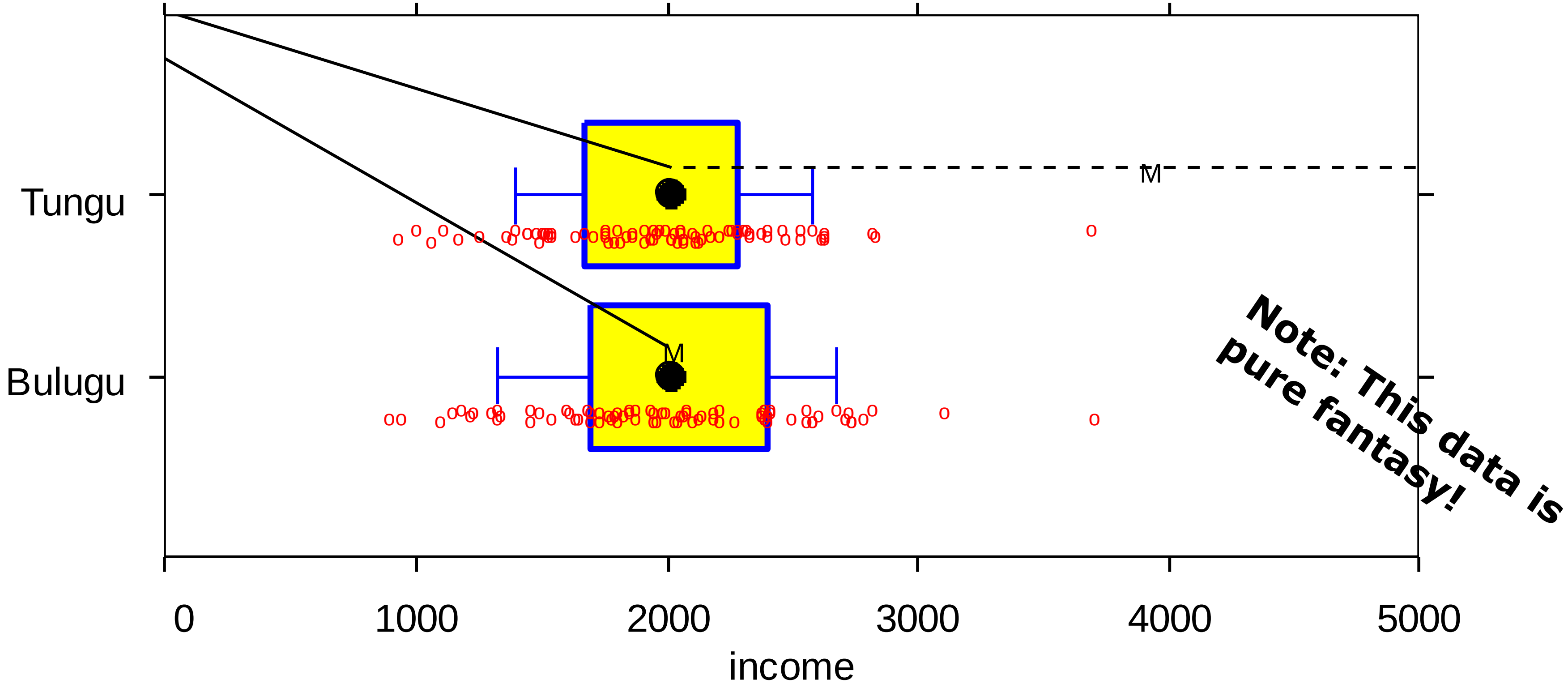
Example 2: Tungu and Bulugu

- We look at the yearly per-capita income in two small hypothetic island states:
  Tungu and Bulugu

- Statement:
  "The average yearly income in Tungu is 94.3% higher than in Bulugu."

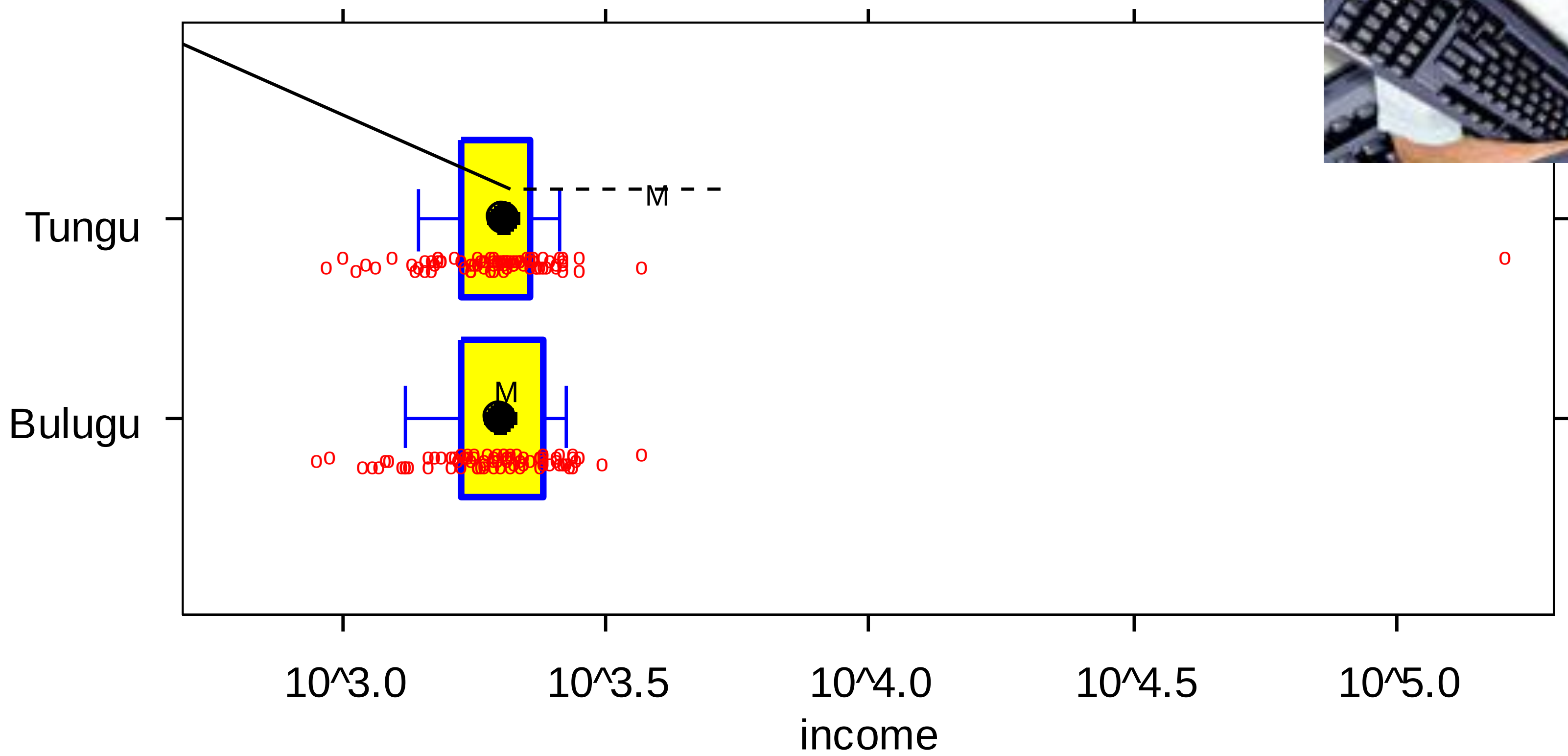
Problem 1: Misleading averages

- The island states are rather small:
  81 people in Tungu and **80** in Bulugu
- And the income distribution is not as even in Tungu:



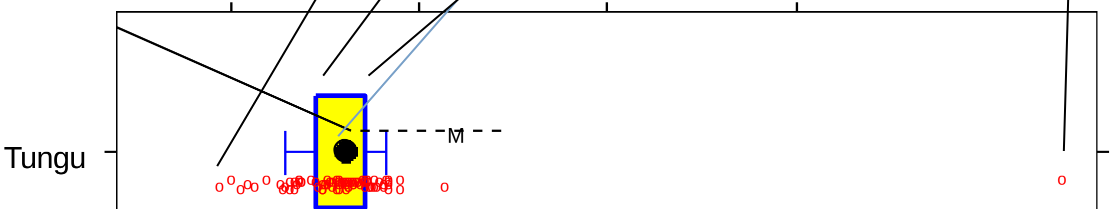
Misleading averages and outliers

- The only reason is Dr. Waldner, owner of a small software company in Berlin, who since last year is enjoying his retirement in Tungu



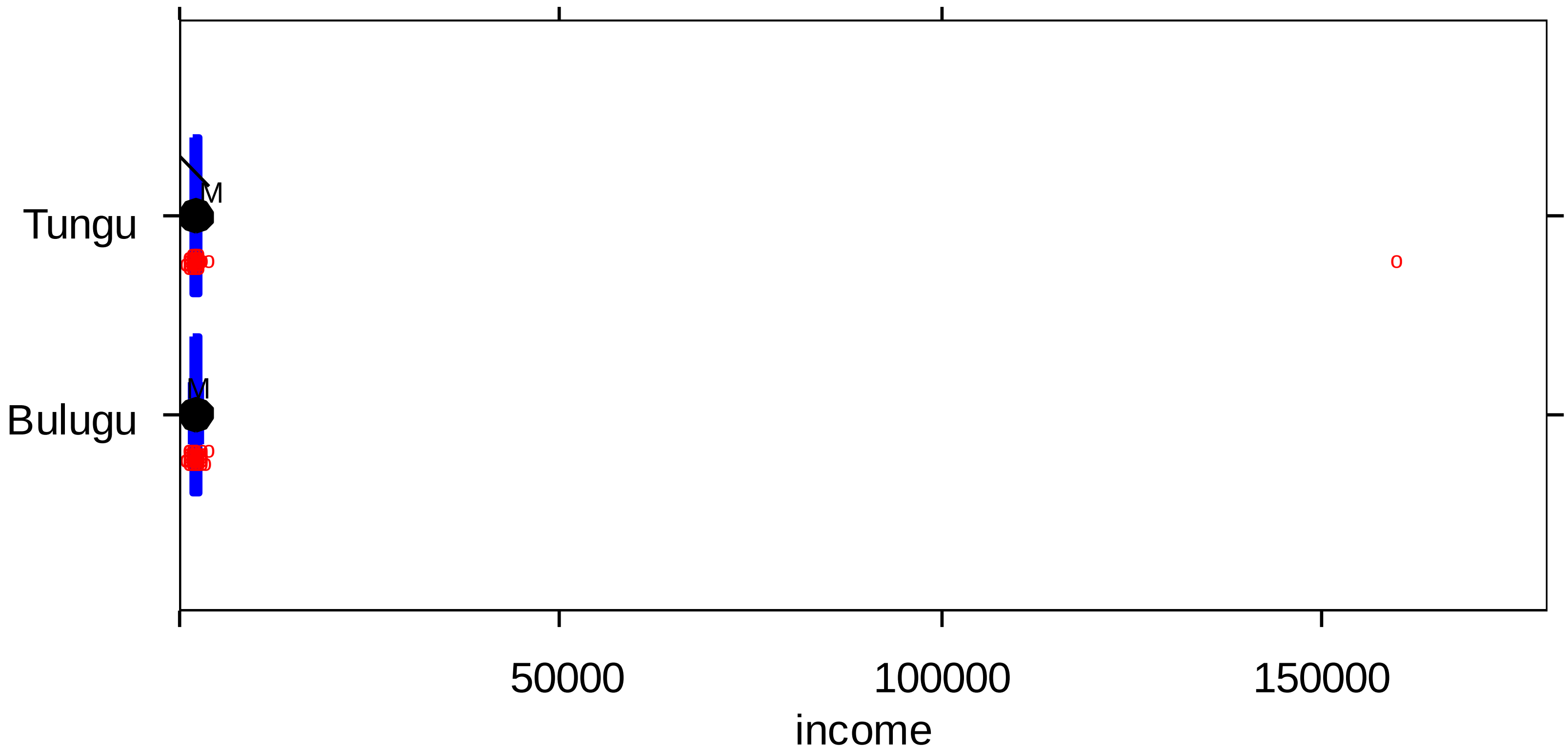
Lesson: Question appropriateness

- A certain statistic (very often the arithmetic average) may be inappropriate for characterizing a sample
- If there is any doubt, ask that additional information be provided
  - such as standard deviation
  - or some quantiles, e.g.: 0, 0.25, 0.5, 0.75, 1
    Note: 0.25 quantile
    is equivalent to
    25-percentile
    etc.



Logarithmic axes

- Waldner earns 160.000 per year.
  How much more that is than the other Tunguans have, is impossible to see on the logarithmic axis we just used



Lesson: Beware of inappropriate visualizations (#1)

- Lesson for reader: Always look at the axes. Are they linear or logarithmic?

- Lesson for author:
  - Logarithmic axes are very useful for reading hugely different values from a graph with some precision
  - But they totally defeat the imagination!
  - If you decide to use logarithmic axes, always state this fact in your text!

- There are many more kinds of inappropriate visualizations
  - see later in this presentation

Problem 3: Misleading precision

- "The average yearly income in Tungu is **94.3%** higher than in Bulugu"

- Assume that tomorrow Mrs. Alulu Nirudu from Tungu gives birth to her twins

- There are now 83 rather than 81 people on Tungu
- The average income drops from 3922 to 3827
- The difference to Bulugu drops from 94.3% to 89.7%

Lesson for reader: Do not be easily impressed

- The usual reason for presenting very precise numbers is the wish to impress people
  - *„Round numbers are always false“*
  - But round numbers are much easier to remember and compare

- Clearly tell people you will not be impressed by precision
  - in particular if the precision is purely imaginary

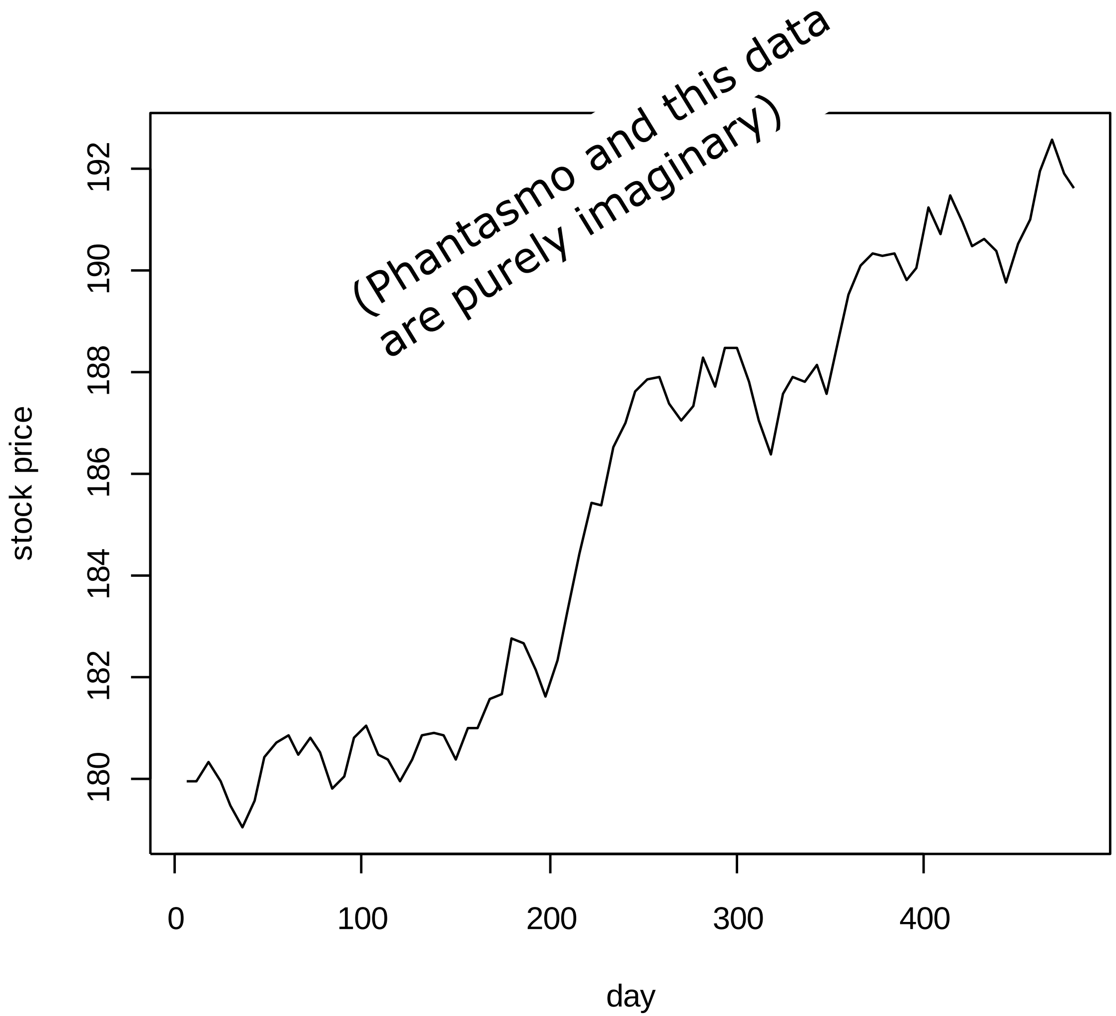
Lesson for author: Think about precision

- Do you really have enough data that would make sense to give out precise numbers?

- Compromise: Give exact number in tables/figures, but round them in text.
- Do not exaggerate: If you find your systems yields a 53,9% increase in throughput
  - Don’t say: “Our system increases throughput by more than 50%”
  - Do say: “Our experiments suggest that our system can achieve throughput increases of around 50%”

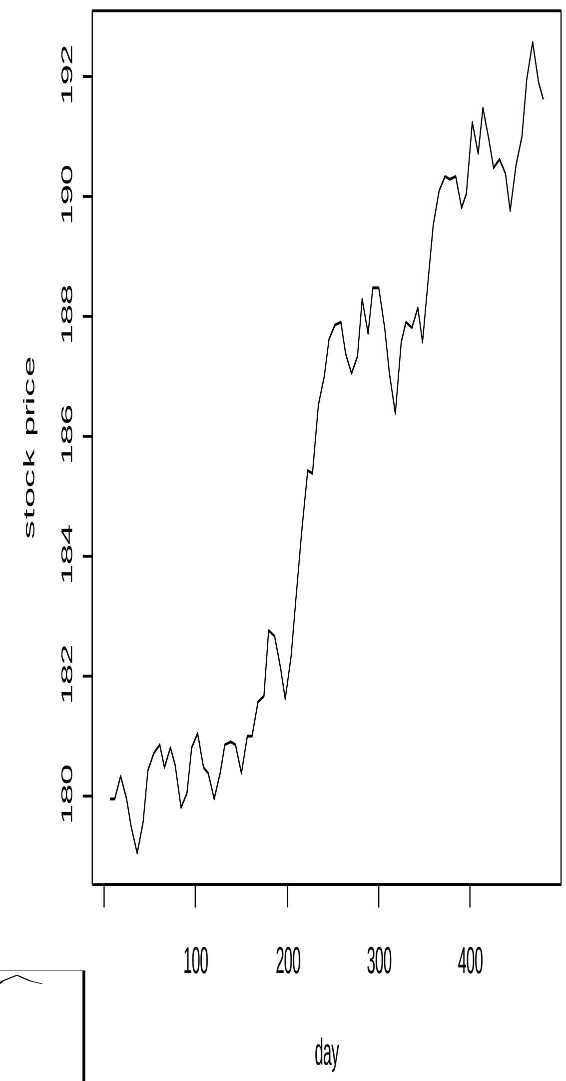
Example 3: Phantasmo Corporation stock price

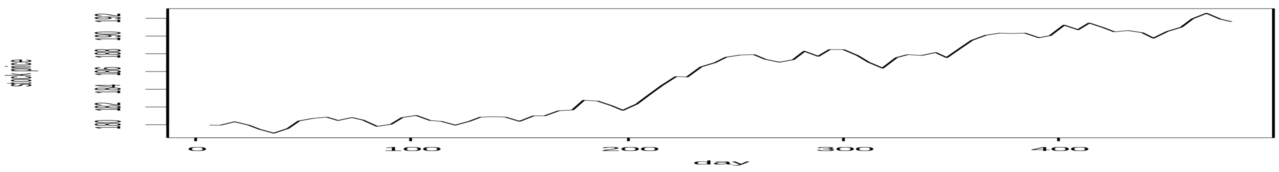
- We look at the recent development of the price of shares for Phantasmo Corporation
- *"Phantasmo shows a remarkably strong and consistent value growth and continues to be a top recommendation"*



Problem: Looks can be misleading

- The following two plots show exactly the same data!
  - and the same as the plot on the previous slide!

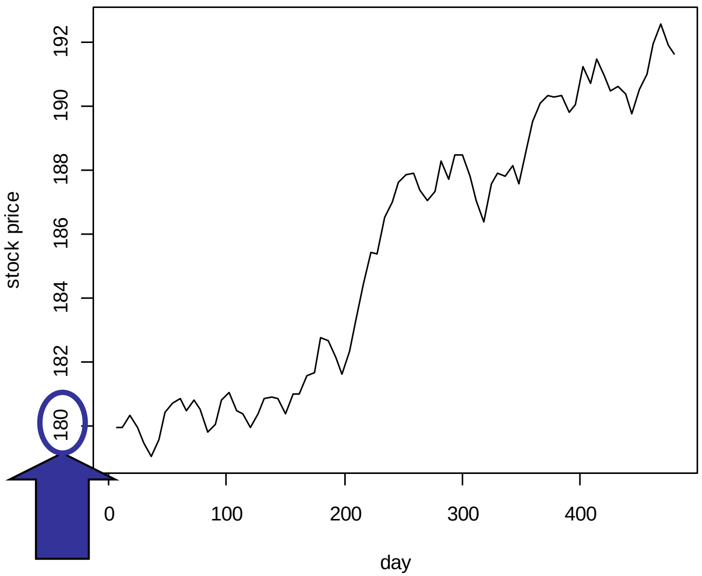


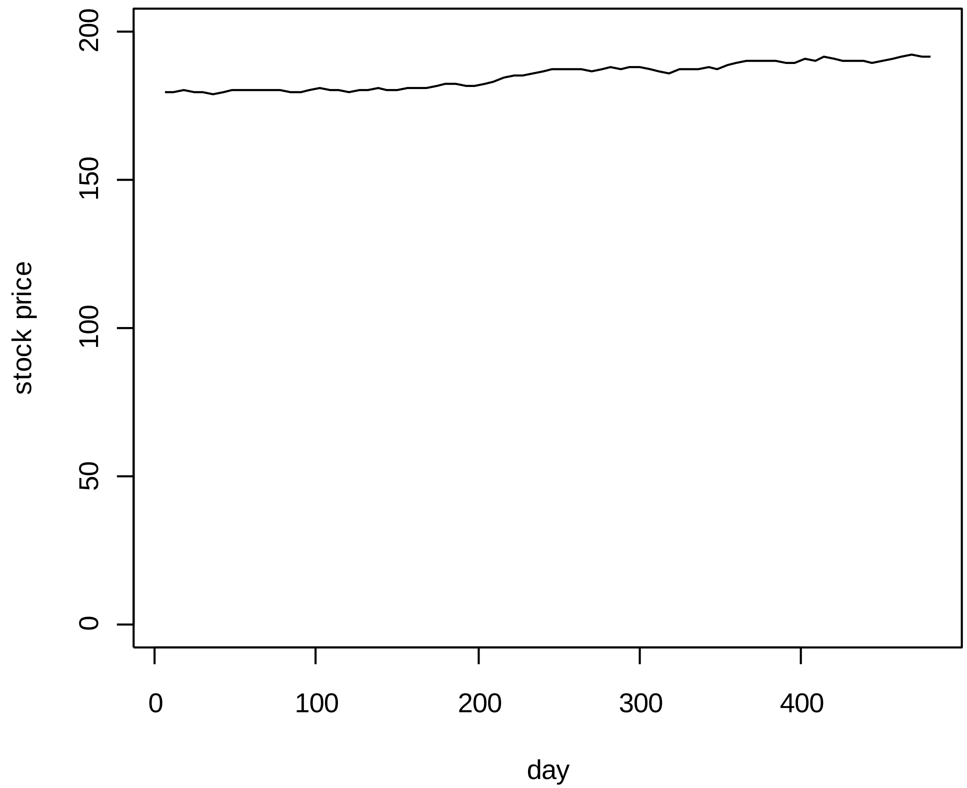


Problem: Scales can be misleading

- What really happened is shown here:

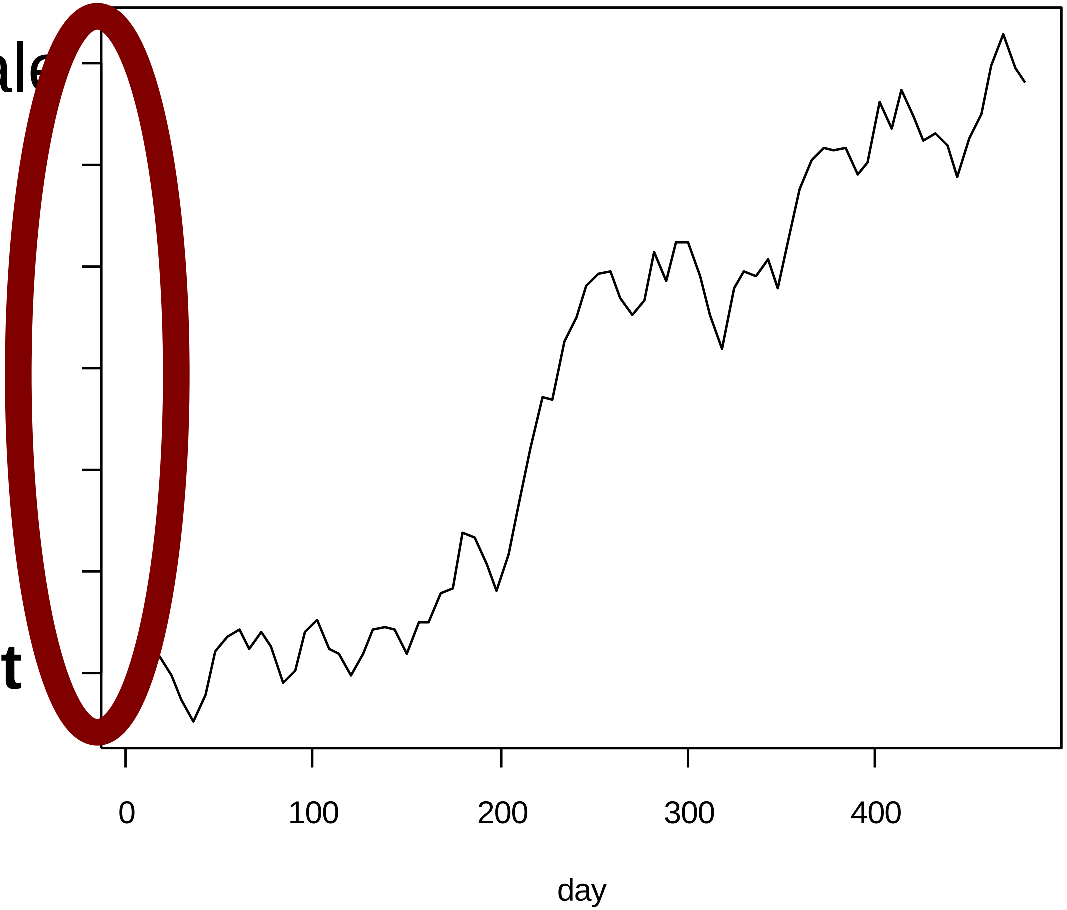
  We intuitively interpret a trend plot on a ratio scale





Problem: Scales can be missing

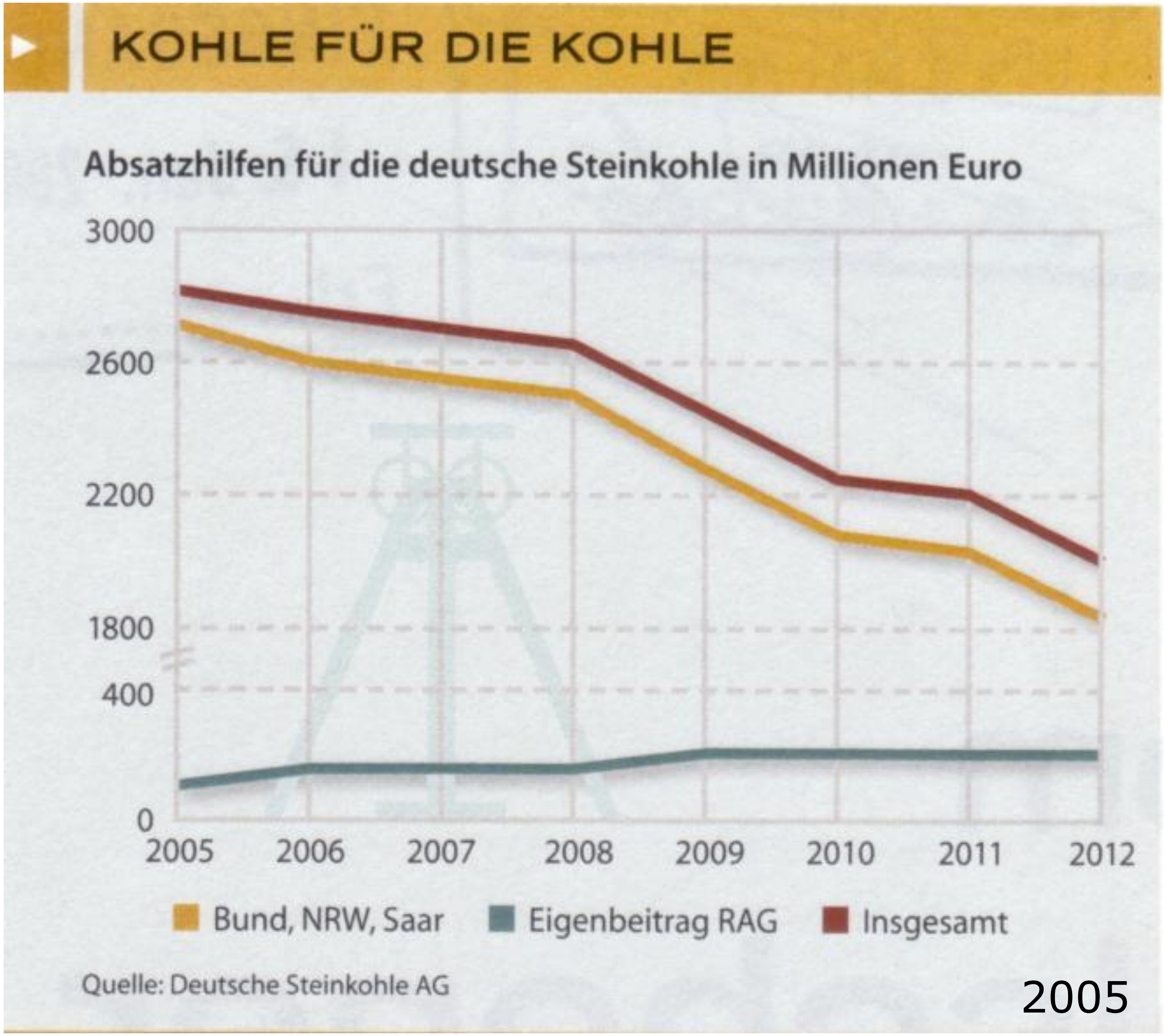
- The most insolent persuaders may even leave the scale out altogether!



- **Never forget to label your axes!**
- **Never forget to put a scale on your axes!**

Problem: Scales can be abused

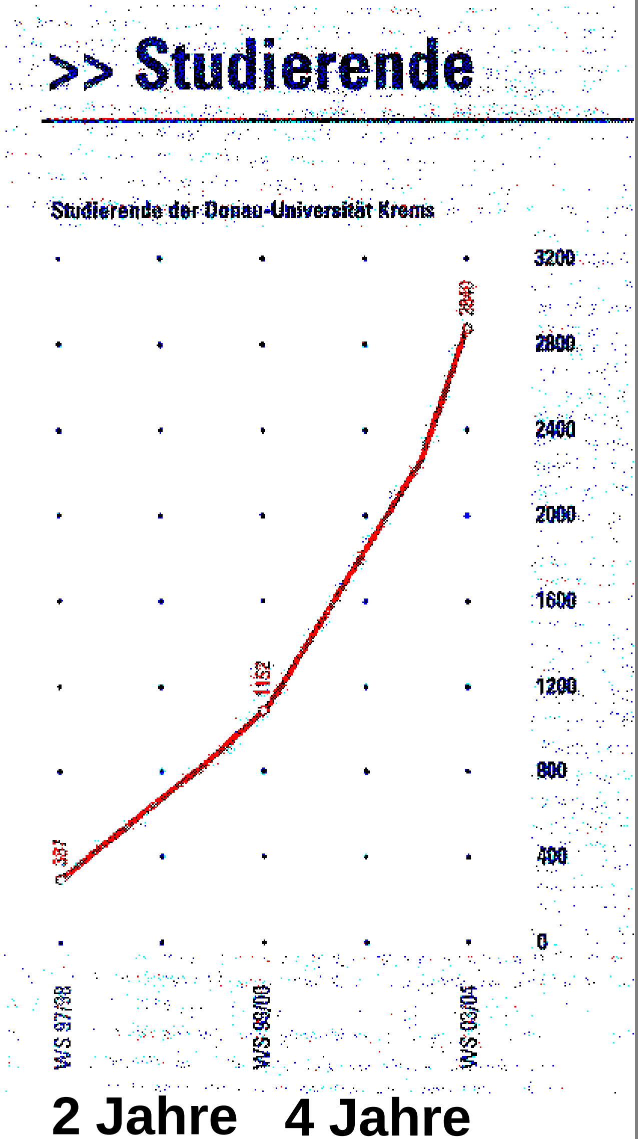
- Observe the global impression first



2005

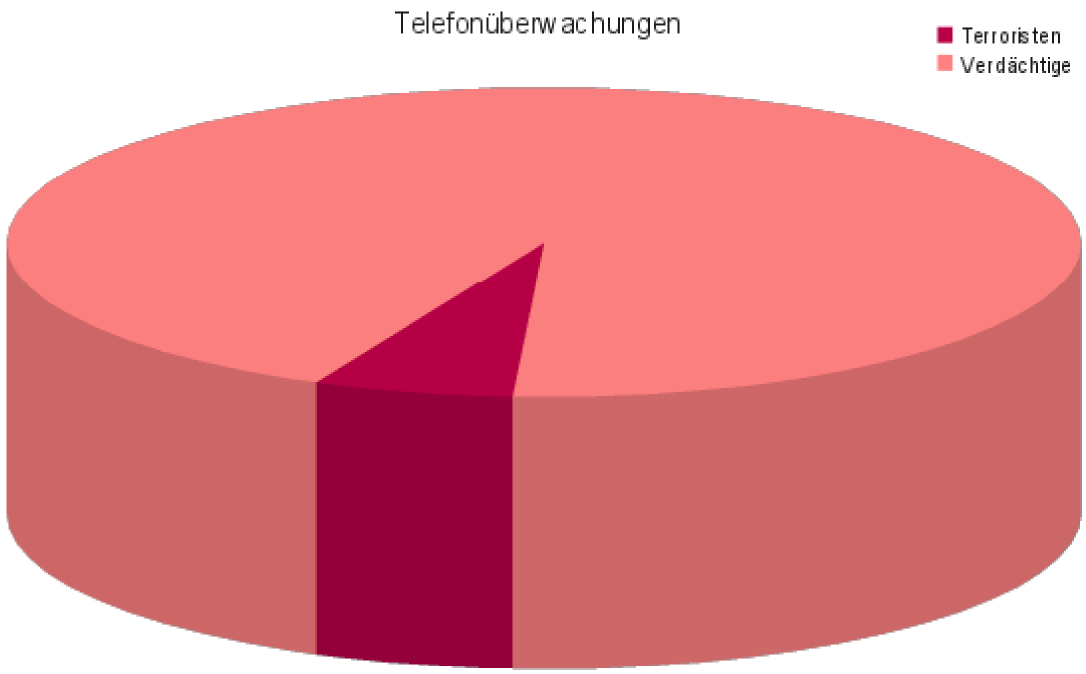
Problem: People may invent unexpected things

- Quelle: Werbeanzeige der Donau-Universität Krems
  - DIE ZEIT, 07.10.2004
  - What‘s wrong?



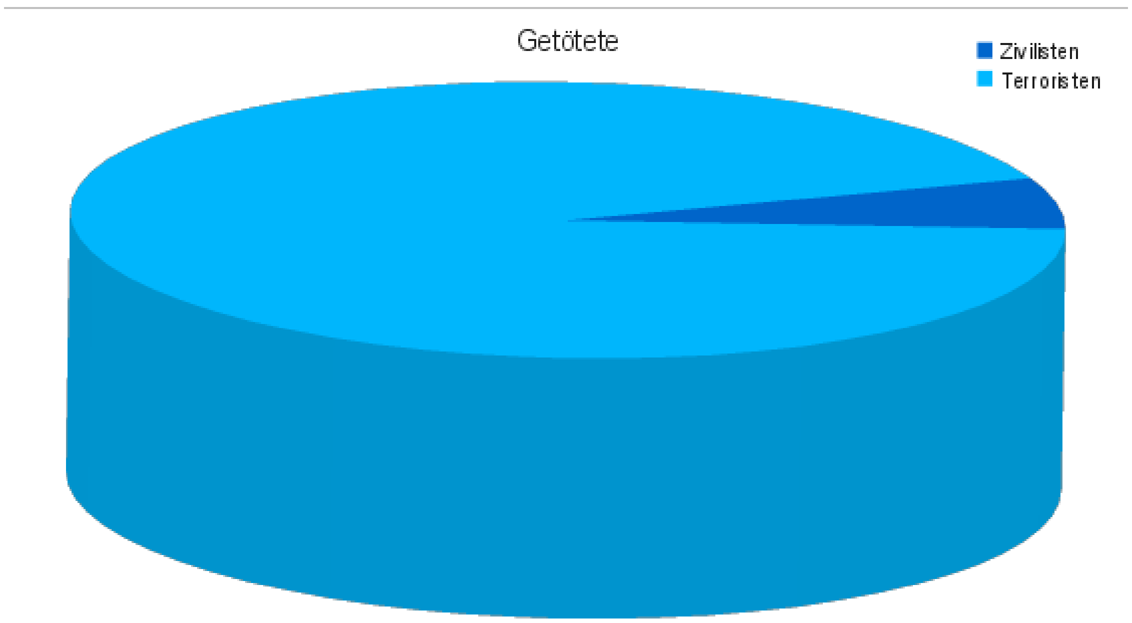
Pie charts (1/3)

Note: This data is pure fantasy!



Pie charts (2/3)

Note: This data is pure fantasy!



Pie charts (3/3)

- What percentages do the two graphs show? Guess!

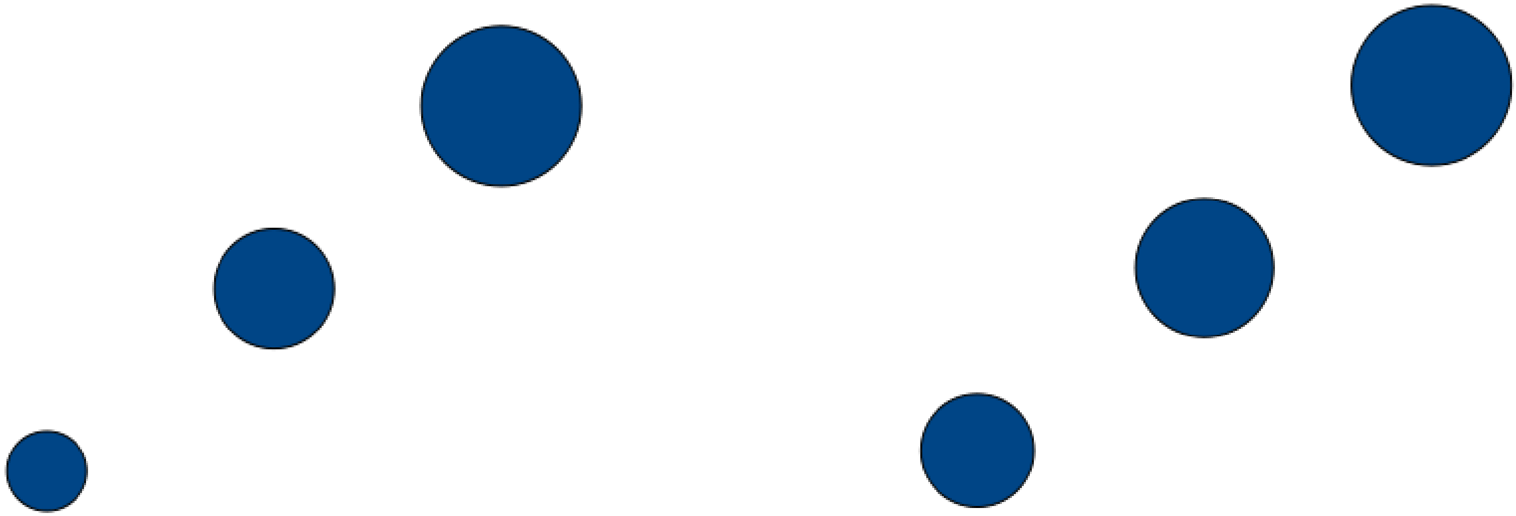
- Answer:
  - **Both** show the same data: A 94% : 6% ratio!
  - The difference only lies in the angle of the pies.

Lesson: Distrust pie charts!

- Pie charts should not be used
  - Perception dependent on the angle
  - Even worse with 3D pie charts:
    Parts at the front are artificially increased due to the pie's 3D height; they thus seem to be bigger
  - A very subtle way to visually tune your data
  - Unfortunately, still very common

- Distrust pie charts that do not give numbers as well
  - Think about the numbers, compare them
  - Think about the presentation: are they trying to beautify the impression?

Bubble charts



Note: This data is pure fantasy!

- ❑Which diagram shows the values 2, 3, 4?
- ❑Both do!
- ❑Left one: Radius is proportional to measurements
  - Exaggerates differences: 4 looks much larger than 2
- ❑Right one: Area is proportional to measurements
  - Underestimates differences: 4 looks only slightly larger than 2

Lesson: Bubble charts

- This lession is more or less similar to pie charts…:

- Bubble charts usually should not be used
  - Radius proportionality exaggerates differences, area proportionality lets underestimate differences
  - A very subtle way to visually tune your data
  - Of course, a bubble chart + pie chart may convey more information, but *please* try to visualize it differently…
  - If you really, really want to use a bubble chart, then use the area proportionality variant, and clearly explain this in your text

- Distrust bubble charts that do not give the numbers as well
  - Think about the numbers, compare them
  - Think about the presentation: Did they really need to use bubble charts? Or are they trying to beautify the impression?

Summary lesson for the reader: Seeing is believing…

- …but often, it shouldn't be!

- Always consider what it really is that you are seeing
- Do not believe anything purely intuitively
- Do not believe anything that does not have a well-defined meaning

Example 4: blend-a-med Night Effects

- What do they <u>not</u> say? Think about it…

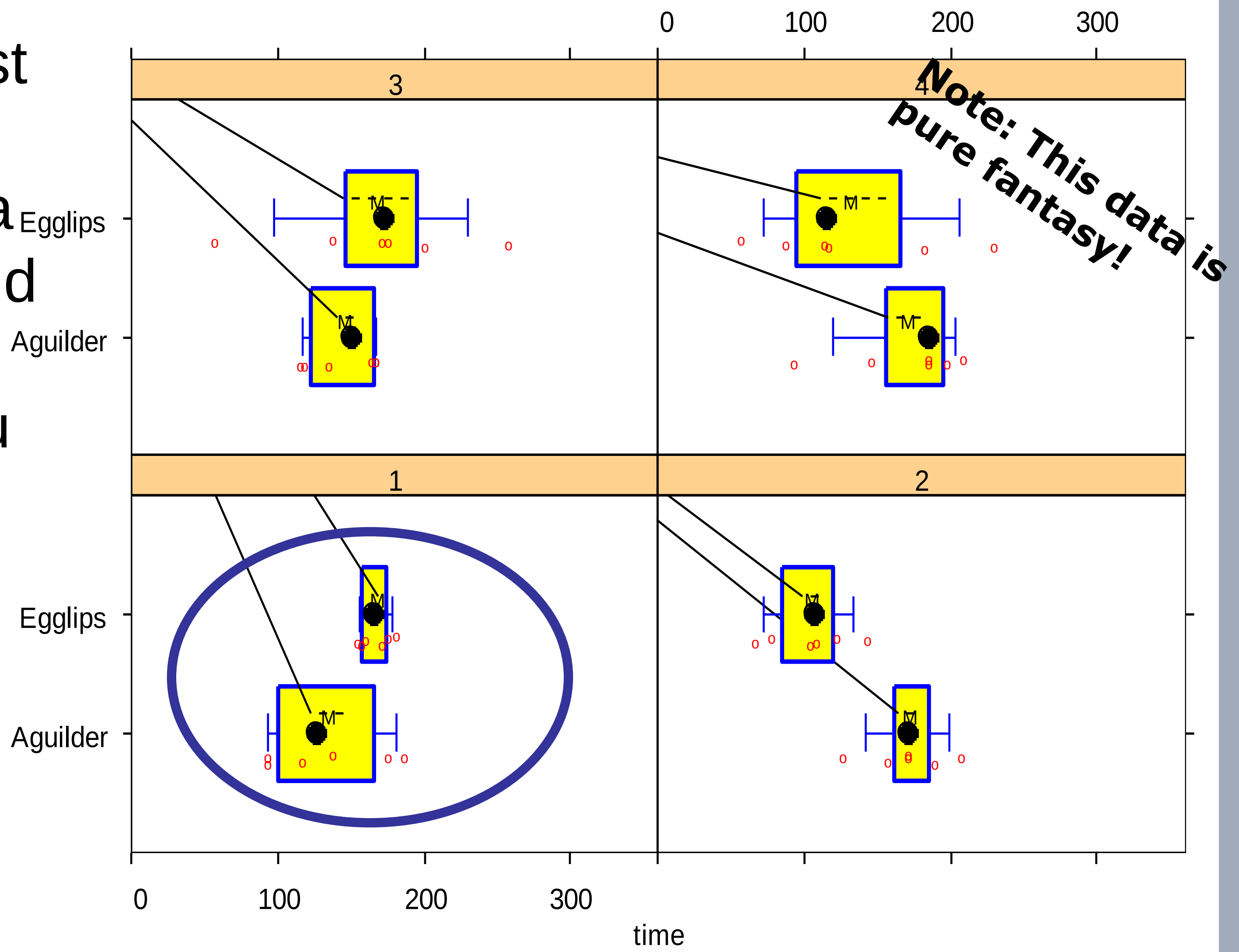
- What exactly does "sichtbar" mean?
  What exactly does „hell“ or „heller“ mean?
- What was the scope, what were the results of the clinical trials?
- What other effects does Night Effects have?

Example 5: The better tool?

- We consider the time it takes programmers to write a certain program using different IDEs:
  - *Aguilder* or
  - *Egglips*
- Statement (by the maker of Aguilder):
  *"In an experiment with 12 persons, the ones using Egglips required on average **24.6% more time** to finish the same task than those using Aguilder.*
  Both groups consisted of equally capable people and received the same amount and quality of training."
- Assume Egglips and Aguilder are in fact just as good. What may have gone wrong here?

Problem: Has anybody ignored any data?

- Solution: Just repeat the experiment a few times and pick the outcome you like best



Lesson for the reader: Demand complete information

- If somebody presents conclusions
  - based on only a subset of the available data
  - and has selected which subset to use
  - then everything is possible

- There is no direct way to detect such repetitions,

BUT for any one single execution . . .

Digression: Hypothesis testing

- …a *significance test* (or confidence intervals) can determine how likely it was to obtain this result if the conclusion is wrong:
  - Null hypothesis: Assume both tools produce equal worktimes overall
  - Then how often will we get a difference this large when we use samples of size 6 persons?
    - If the probability is small, the result is plausibly real
    - If the probability is large, the result is plausibly incidental

Statistical significance test: Example

- Our data:
  - Aguilder: 175, 186, 137, 117, 92.8, 93.7 (mean 133)
  - Egglips: 171, 155, 157, 181, 175, 160 (mean 166)
- Null hypothesis: We assume
  - the distributions underlying these data are both normal distributions with the same variance
  - the means of the actual distributions are in fact equal
- Then we can compute the probability for seeing this difference of 33 from two samples of size 6
- The procedure for doing this is called the *t-test* (recall the confidence intervals? – It‘s a very similar calculation)
- Results (10 degrees of freedom):
  - p value: 0.08
    - the probability of the above result if the null hypothesis is true (i.e., difference is indeed zero)
  - 95% confidence interval for true difference: -5…71

So? (Lessons for the author)

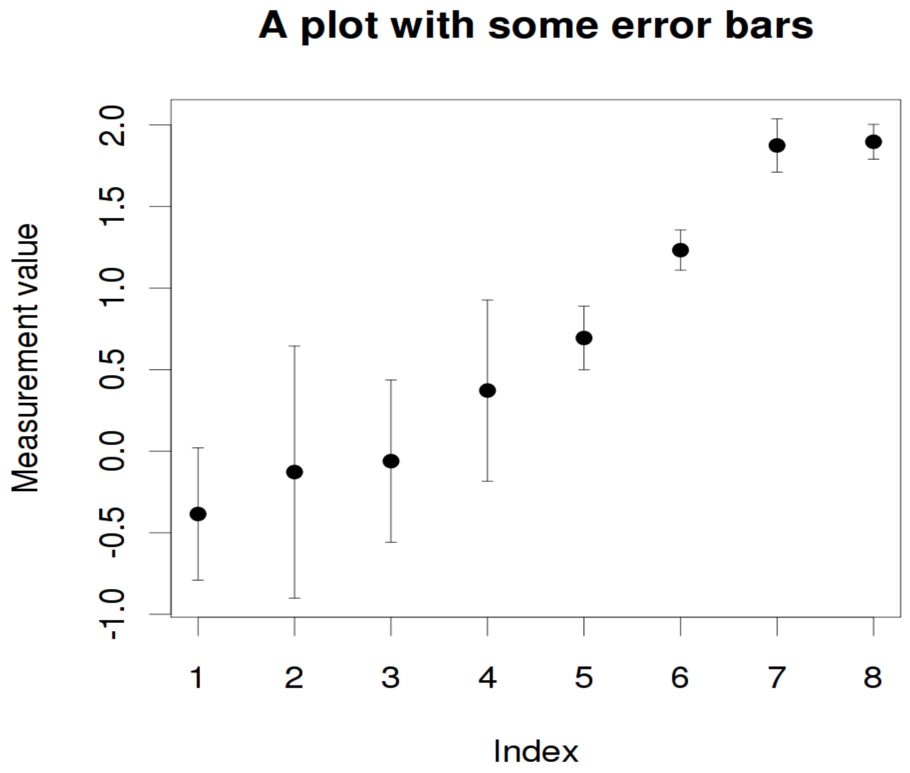
- So in our case we probably would believe the result and not find out that the experimenters had in fact cheated
  - (And indeed they were lucky to get the result they got)

Note:

- There are many different kinds of hypothesis tests and various things can be done wrong when using them
  - In particular, **watch out what the test assumes**
  - **and what the p-value means, namely:**
    - The probability of seeing this data *if the null hypothesis is true*
    - Note: **The p-value is not the probability that the null hypothesis is true!**
  - But unless the distribution of your samples is very strange or very different, using the t-test is usually OK.
    - Note: There are quite a number of different tests called “t test”.
    - They have subtle yet important differences…

Example: Error bars

- "Although a high variability in our measurements results in rather large error bars, our simulation results show a clear increase in [whatever]."
- What's wrong here?



Lesson: Error bars

- What are the error bars? How are they defined?
  - Minimum and maximum values?
  - Confidence intervals?
    - If so, at which level? 95%? 99%?
  - Mean ± two standard deviations?
  - First and third quartile? 10% and 90% quantile?
  - Chebyshov* or Chernoff bounds?
    *also: Tschebyscheff, Tschebyschow, Chebyshev, …

- Reader: Distrust error bars that are not explained
- Author:
  - Clearly state what kind of error bars you're using
  - Usually, the best choice is to use confidence intervals, but stddev is also quite common

Lesson for the author: Common errors for t tests and confidence intervals

- Recall: “But unless the distribution of your samples is very strange or very different, using the t-test is usually OK.”
- If you do not have many samples (less than ~30), then you must check that your input data looks more or less normally distributed
  - At least check that the distribution does not look terribly skewed
  - Better: do a QQ plot
  - Even better: use a normality test
- You might make many runs, group them together and exploit the Central Limit Theorem to get normally distributed data, but…:
  - Warning: Only defined if the variance of your samples is finite!
  - Therefore won’t work with, e.g., Pareto-distributed samples ($\alpha<2$)
- You must ensure that the samples are not correlated!
  - For example, a time series often is autocorrelated
  - Group samples and calculate their average (Central Limit Theorem); make groups large enough to let autocorrelation vanish
  - Check with ACF plot
or autocorrelation test
or stationarity test

Lesson for the author: Check your prerequisites and assumptions!

- Similar errors can be committed with other statistical methods
- Usual suspects:
  - Input has to be normally distributed, or follow some other distribution
  - Input must not be correlated
  - Input has to come from a stationary process
  - Input must be at least 30 samples (10; 50; 100; …)
  - The two inputs must have the same variances
  - The variance must be finite
  - The two inputs must have the same distribution types
  - …
  - of course, all this depends on the chosen method!

Summary

- When confronted with data or conclusions from data one should always ask:
  - Can they possibly know this? How?
  - What do they really mean?
  - Is the purported reason the real reason?
  - Are the samples and measures unbiased and appropriate?
  - Are the measures well-defined and valid?
  - Are measures or visualizations misleading?
  - Has something important been left out?
  - Are there any inconsistencies (contradictions)?

- When we collect and prepare data, we should
  - work thoroughly and carefully
  - check our assumptions and prerequisites
  - avoid distortions of any kind

Thank you!