Course "Empirical Evaluation in Informatics"

How to lie with statistics

Lutz Prechelt
Freie Universität Berlin, Institut für Informatik

- What do they mean?
- Biased measures
- Biased samples
- What is the real reason?
- Misleading averages
- Misleading visualizations
- Pseudo-precision
- Plain false statements
- What is not being said?
- "Just try again"
- Incomparable measures
- Invalid measures

"Empirische Bewertung in der Informatik"

Wie man mit Statistik lügt

Lutz Prechelt
Freie Universität Berlin, Institut für Informatik

- Was ist überhaupt gemeint?
- Verzerrt das benutzte Maß?
- Verzerrt die Stichprobenauswahl?
- Ist das wirklich der Grund?
- Irreführende Mittelwerte
- Irreführende Darstellungen
- Pseudopräzision
- Glatte Falschaussagen
- Was wird <u>nicht</u> gesagt?
- "Probier einfach noch mal"
- Unvergleichbare Daten
- Gültigkeit von Maßen

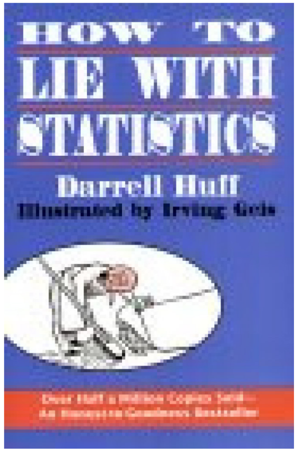
Source

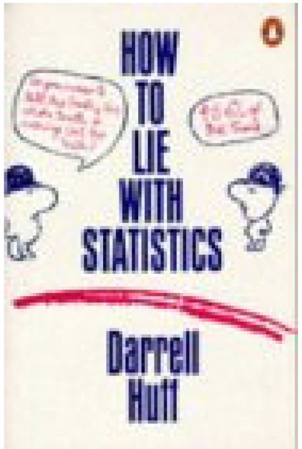
- This slide set is based on ideas from

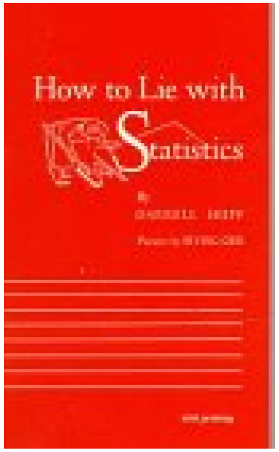
  Darrell Huff: "How to Lie With Statistics",
  (Victor Gollancz 1954, Pelican Books 1973, Penguin Books 1991)

  - but the slides use different examples

- I recommend everyone to read this book in full
  - It is short (120 p.), entertaining, and insightful
  - Many different editions available
  - Other, similar books exist as well







Example:
Human Growth Hormone (HGH)

Original spam email, received 2004-02

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 WADA doping list
- HGH has health risks when taken by healthy individuals

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 the "energy level"?
  - Also note they use language loosely:
    - Loss in percent: OK; reduction in percent: OK
    - Level in percent??? (should be 'increase')

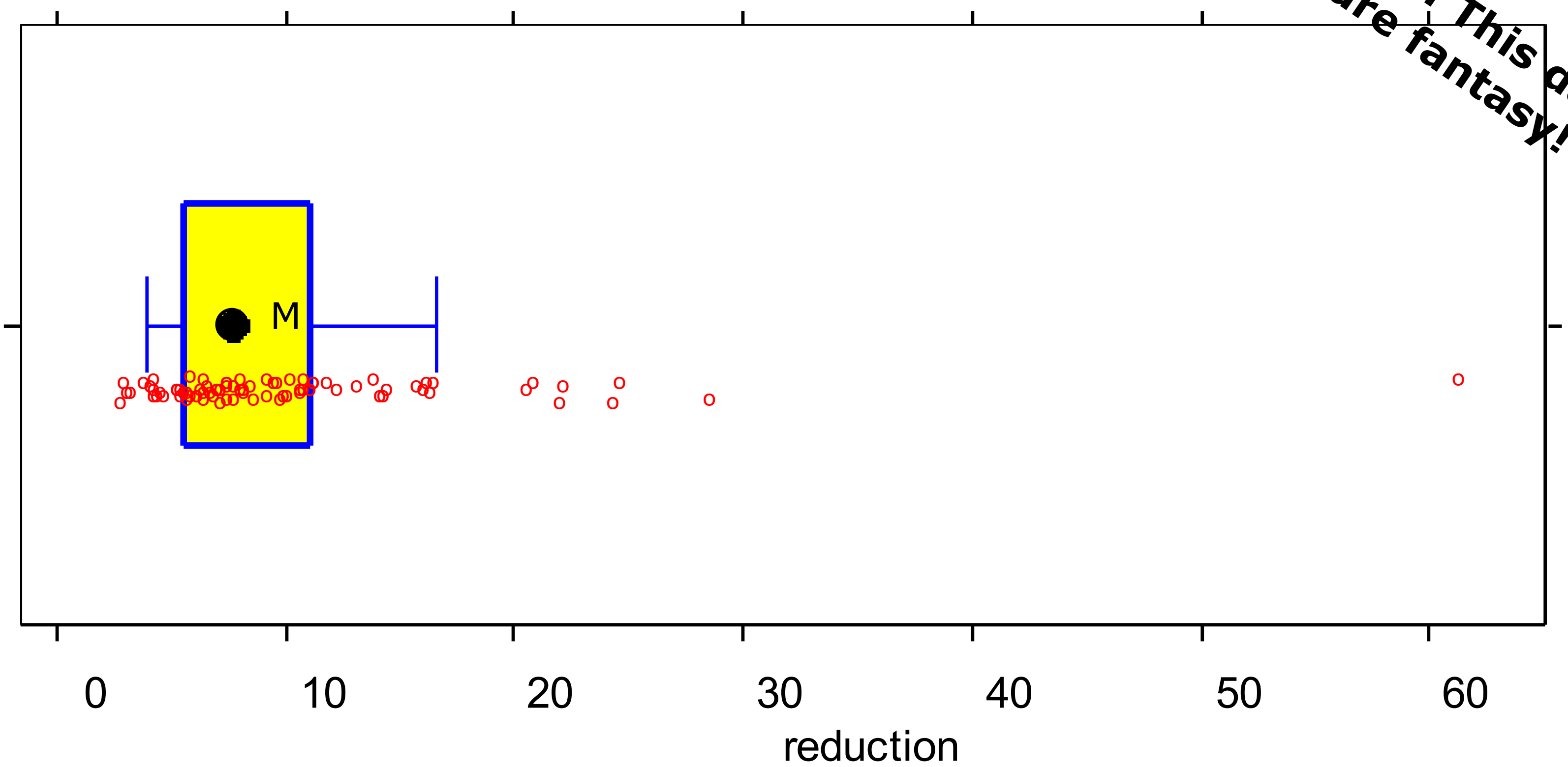
Lesson: Dare ask what

- 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"

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!



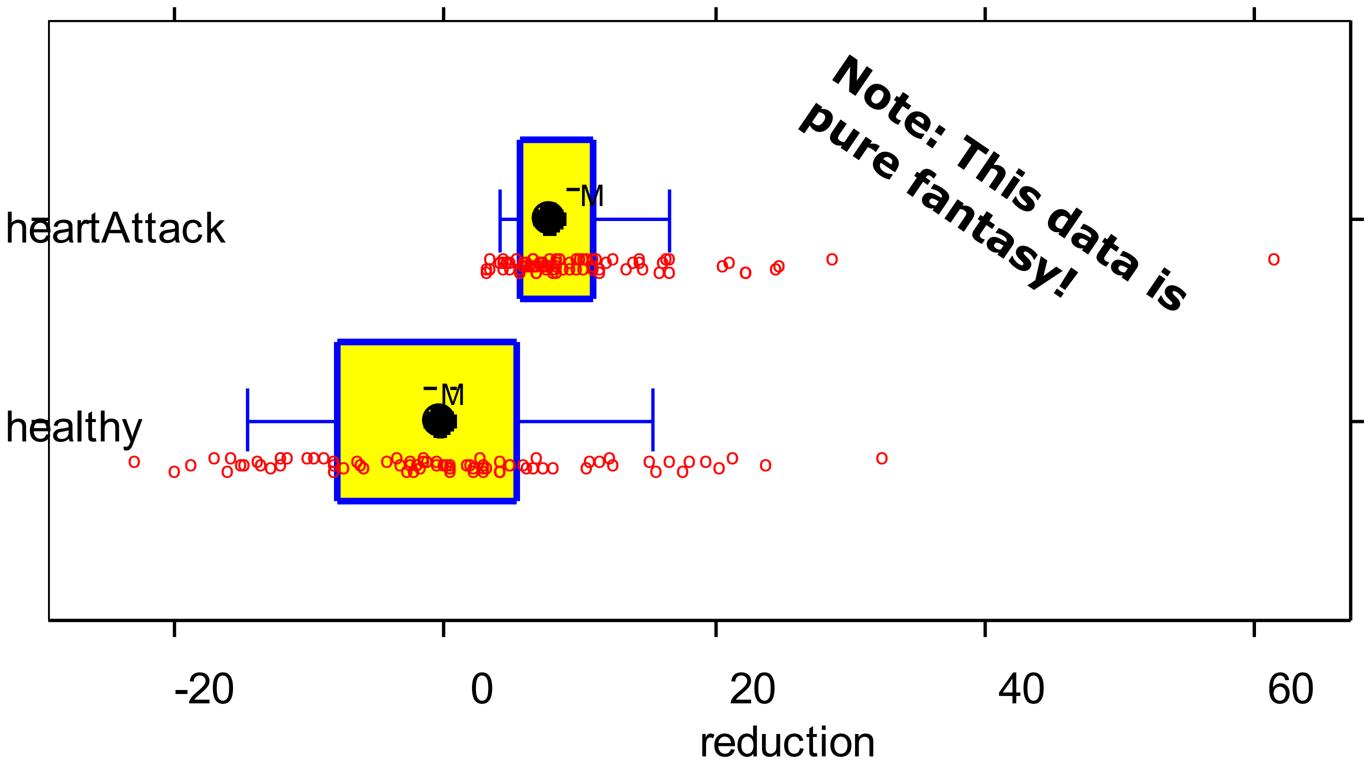
Lesson:
Dare ask for unbiased measures

- 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

Lutz Prechelt, prechelt@inf.fu-berlin.de

Problem 3: Underlying population

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



Lesson: Insist on unbiased samples

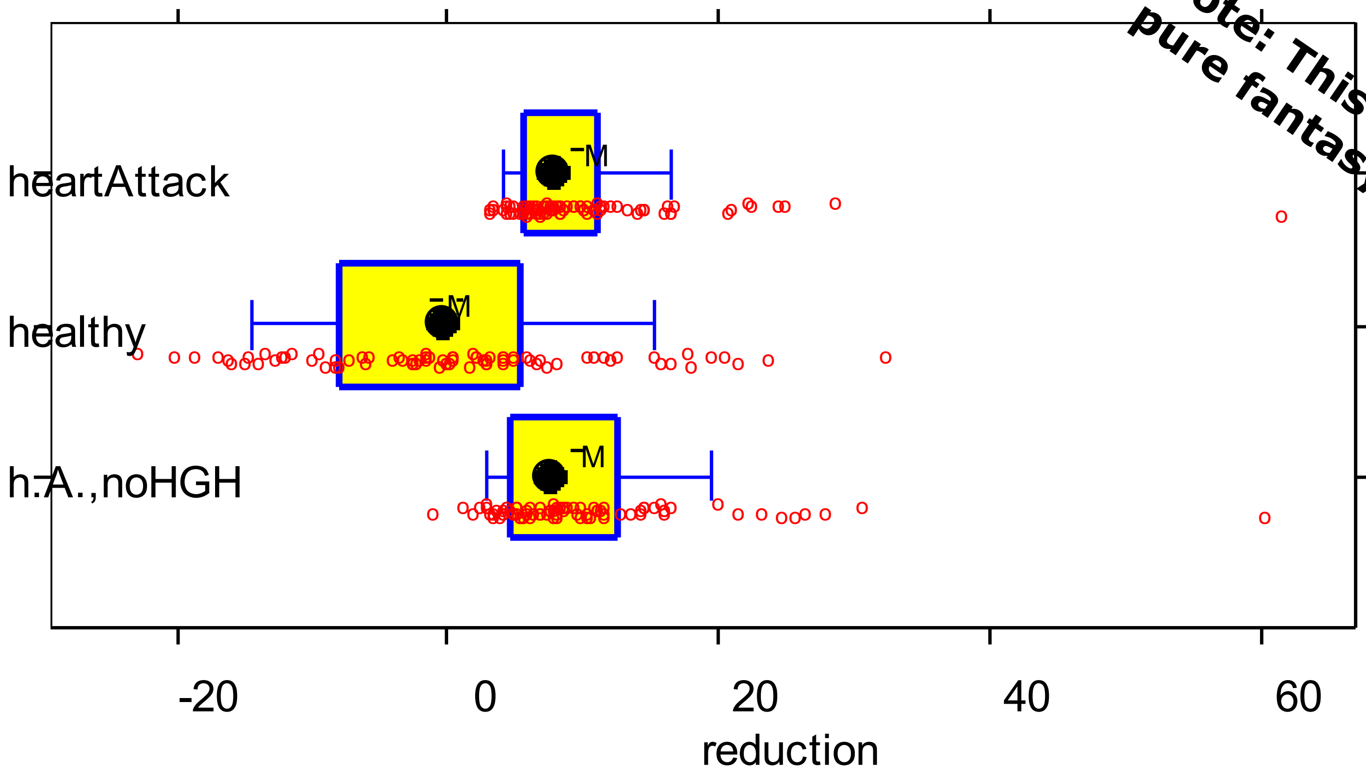
- How and where from the data was collected can have a tremendous impact on the results
- It is important to understand whether there is a certain (possibly intended) tendency in this
- A fair statistic talks about possible *bias* it contains
- If it does not, ask.

Notes:

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

Problem 4:
Is HGH even part of the cause?

- Wrinkle reduction: up to 61%
- Maybe that could happen even without HGH?



Note: This data is pure fantasy!

Lesson: Question causality

- Sometimes the data is not just biased,
  it contains hardly anything else than bias
- If somebody presents you with a presumably causal relationship ("A causes B"), ask yourself:
  - What other influences besides A may be important?
  - What is the relative weight of A compared to these?

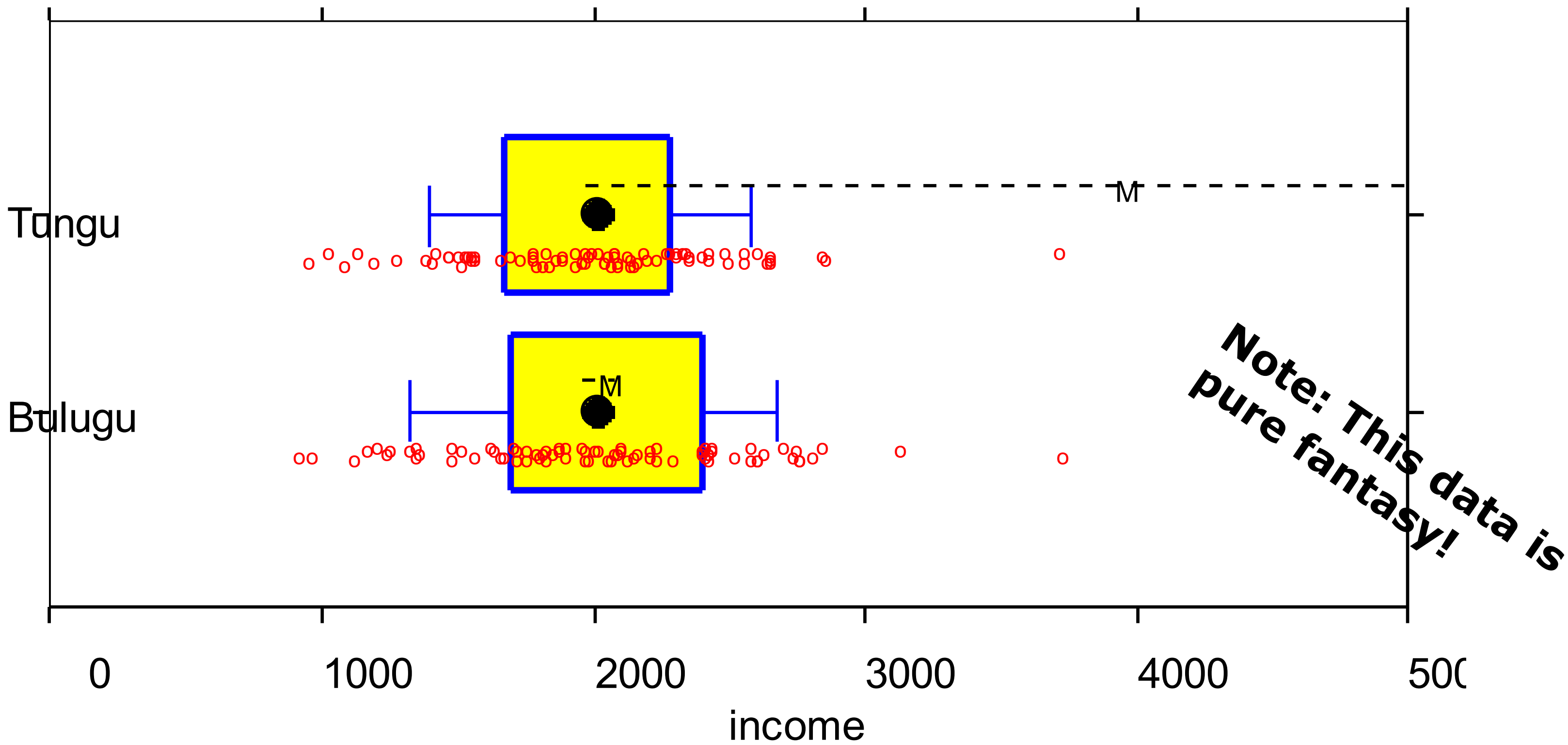
Example 2: Tungu and Bulugu

- 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

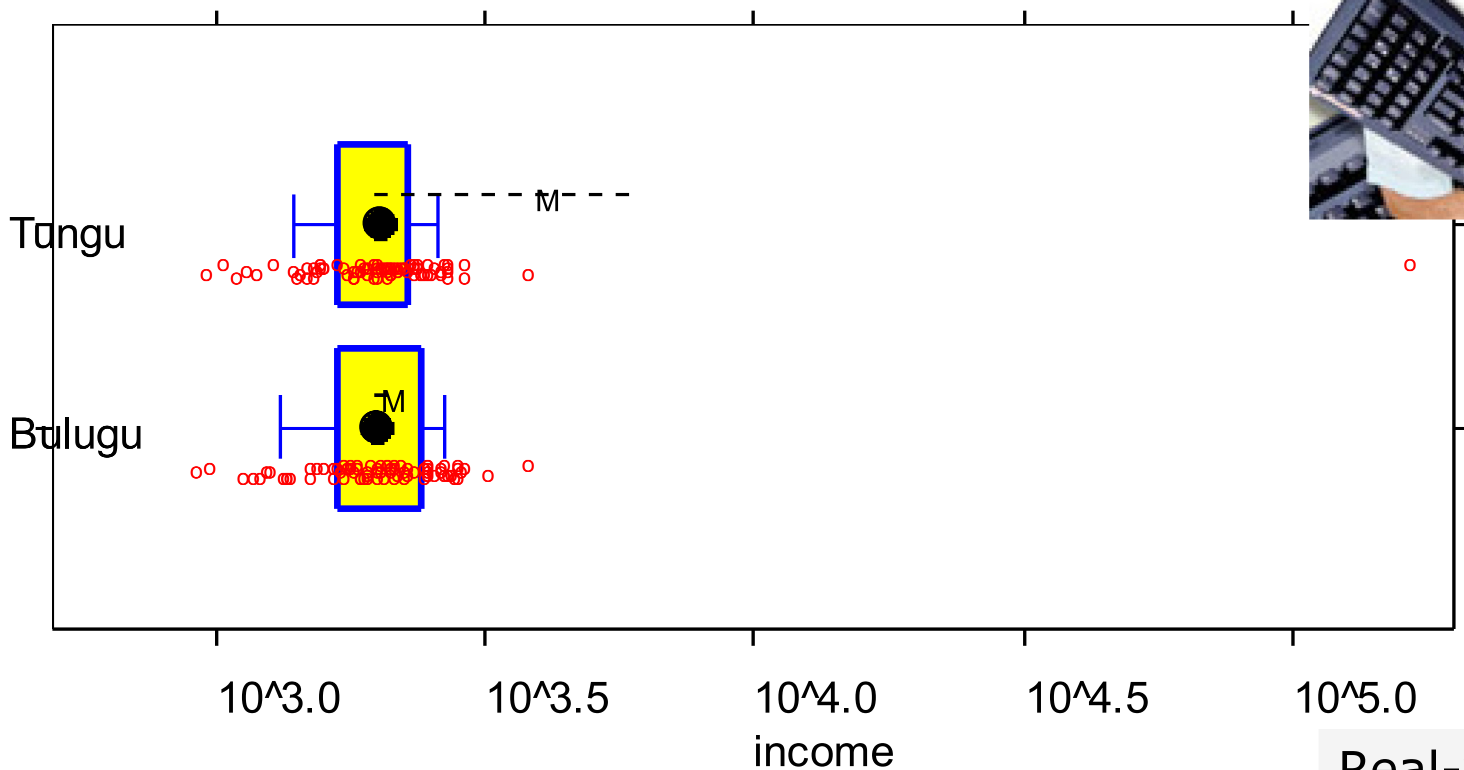
- The island states are rather small:
  81 people in Tungu and **80** in Bulugu
- And the income distribution is not as even in Tungu:



Lutz Prechelt, prechelt@inf.fu-berlin.de

Misleading averages and outliers

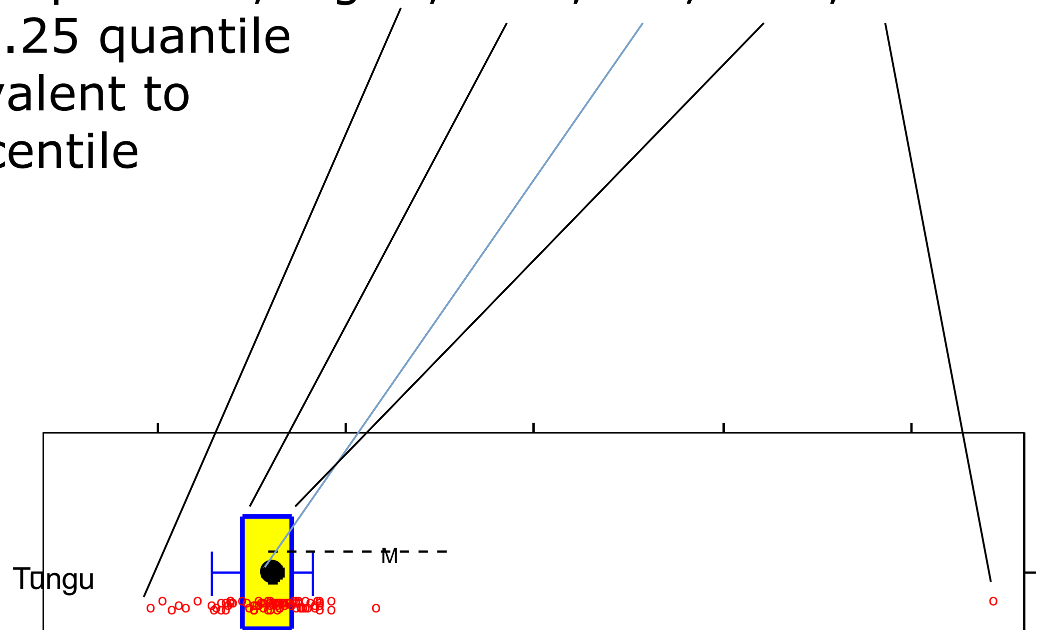
- The only reason is Dr. Waldner, owner of a small software company in Berlin, who since last year is enjoying his retirement in Tungu



Real-world example: [Dieter Schwarz]

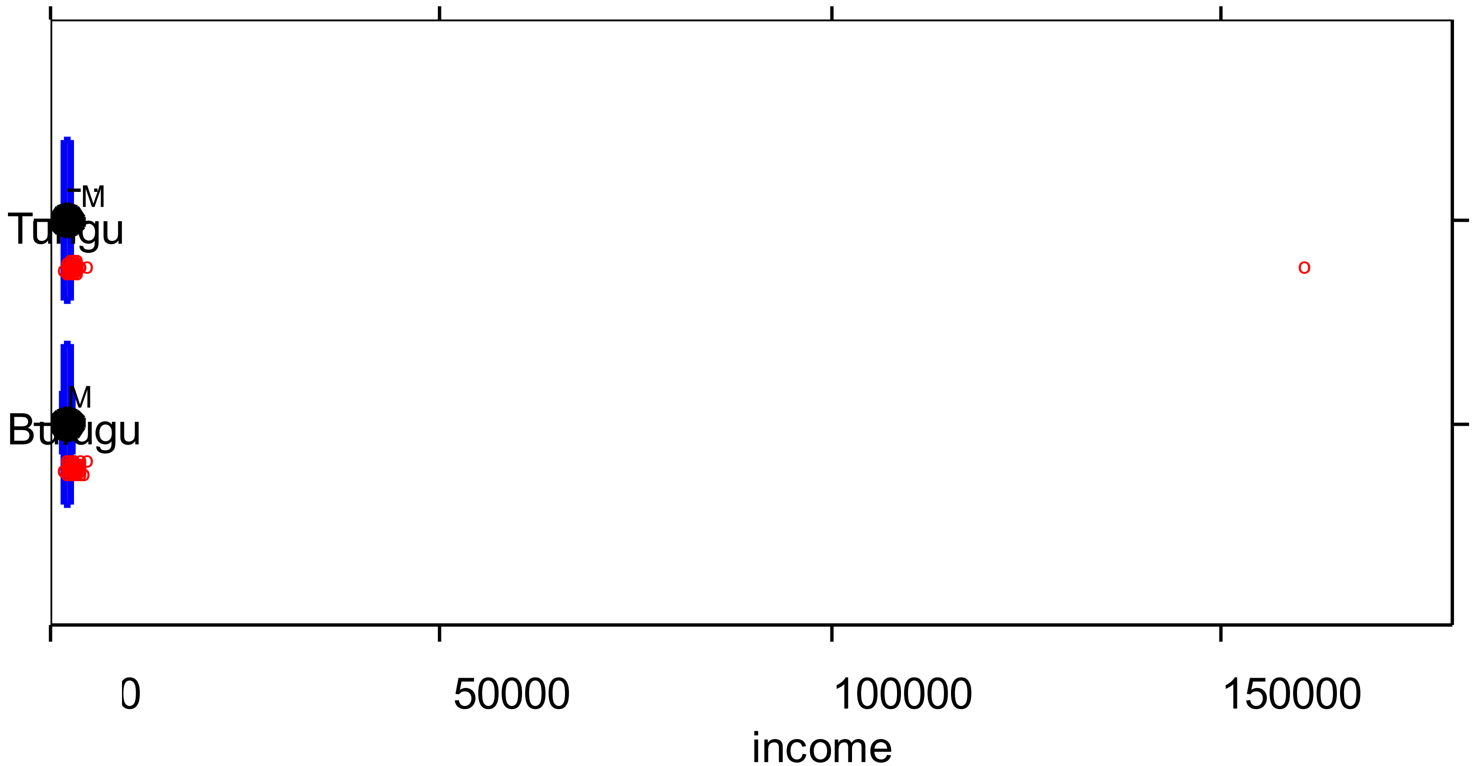
Lesson: Question appropriateness

- 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
  - So let's use a linear one for comparison:



Lesson:
Beware of inappropriate visualizations

Freie Universität Berlin

- Logarithmic axes are useful for reading hugely different values from a graph with some precision
- But they totally defeat the imagination

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

Problem 3: Misleading precision

Freie Universität Berlin

- "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%

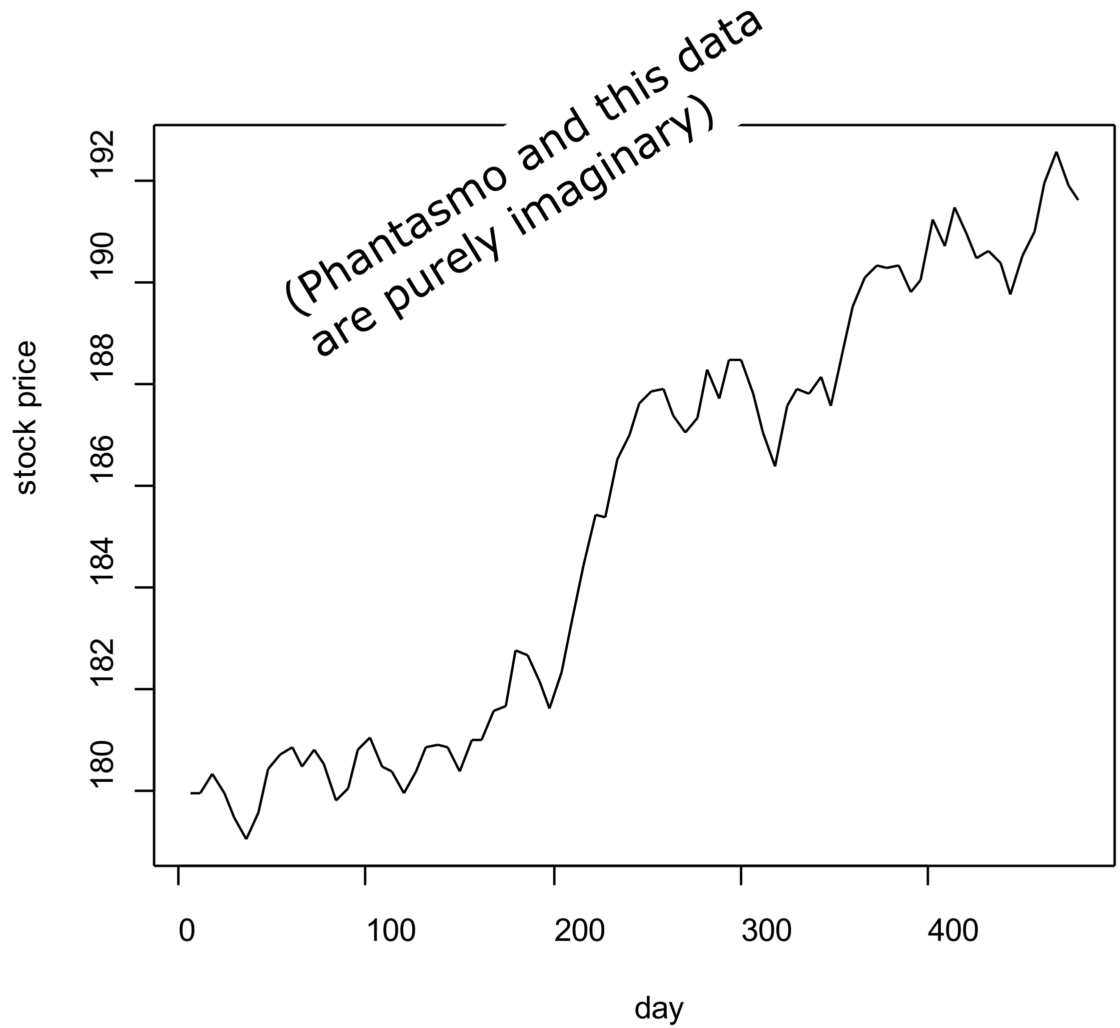
Lesson: Do not be easily impressed

- The usual reason for presenting very precise numbers is the wish to impress people
  - Attitude: *"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

Example 3: Phantasmo Corporation stock price

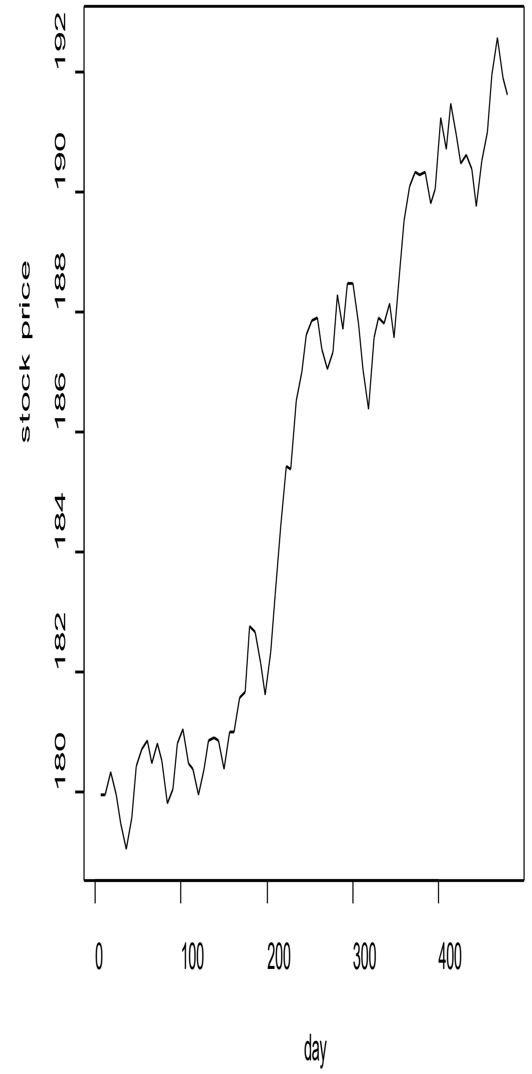
* We look at the recent development of the price of shares for Phantasmo Corporation

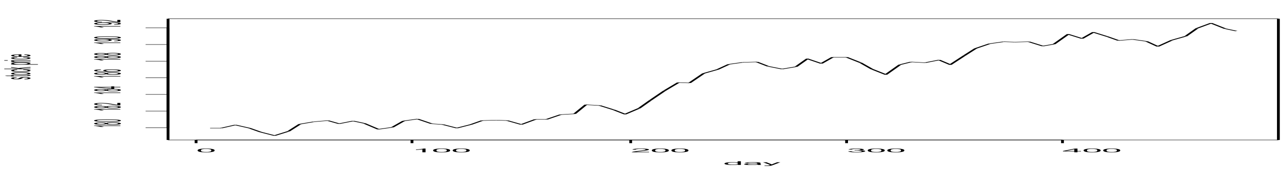
* *"Phantasmo shows a remarkably strong and consistent value growth and continues to be a top recommendation"*



Problem: Looks can be misleading

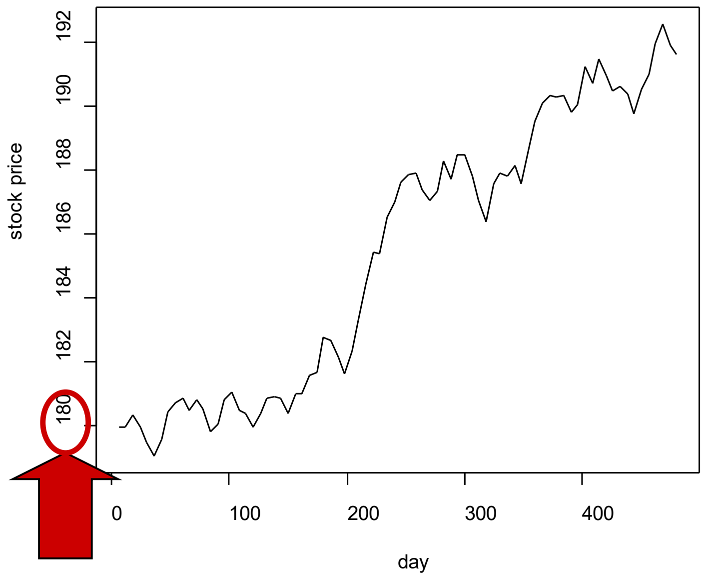
- The following two plots show exactly the same data!
  - and the same as the plot on the previous slide!

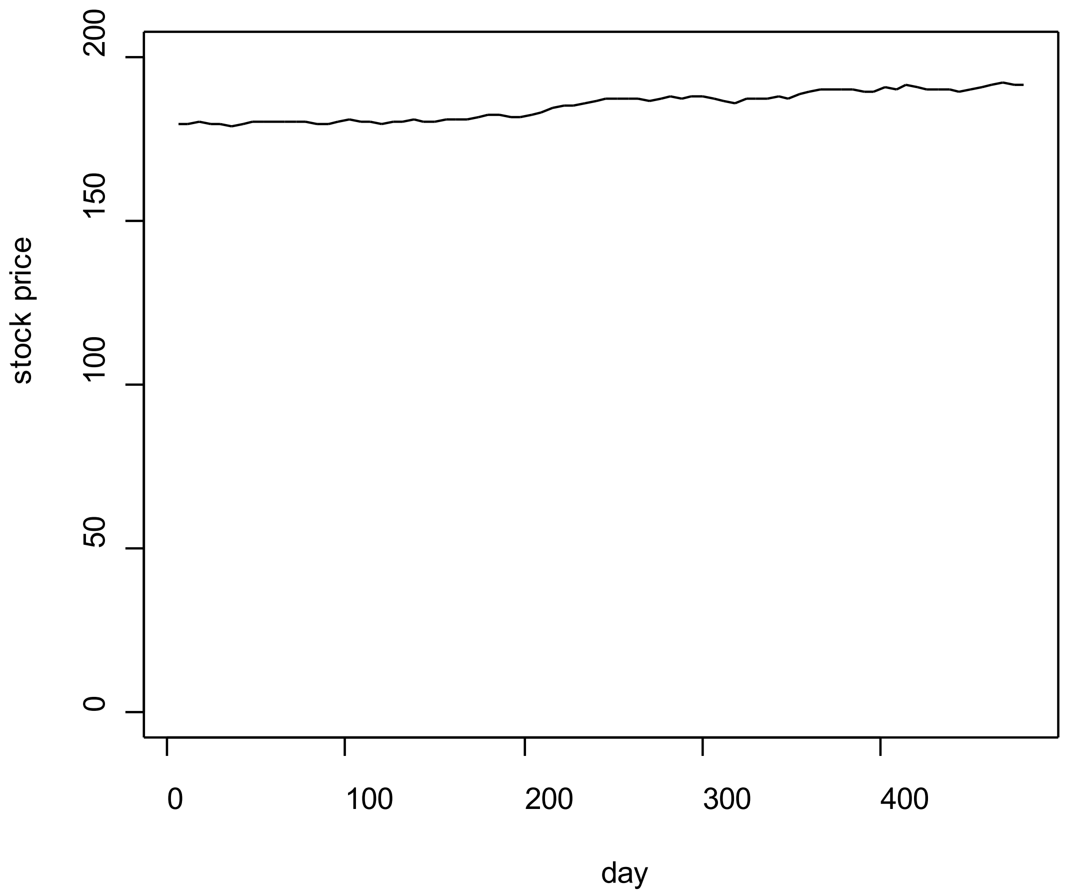




Problem: Scales can be misleading

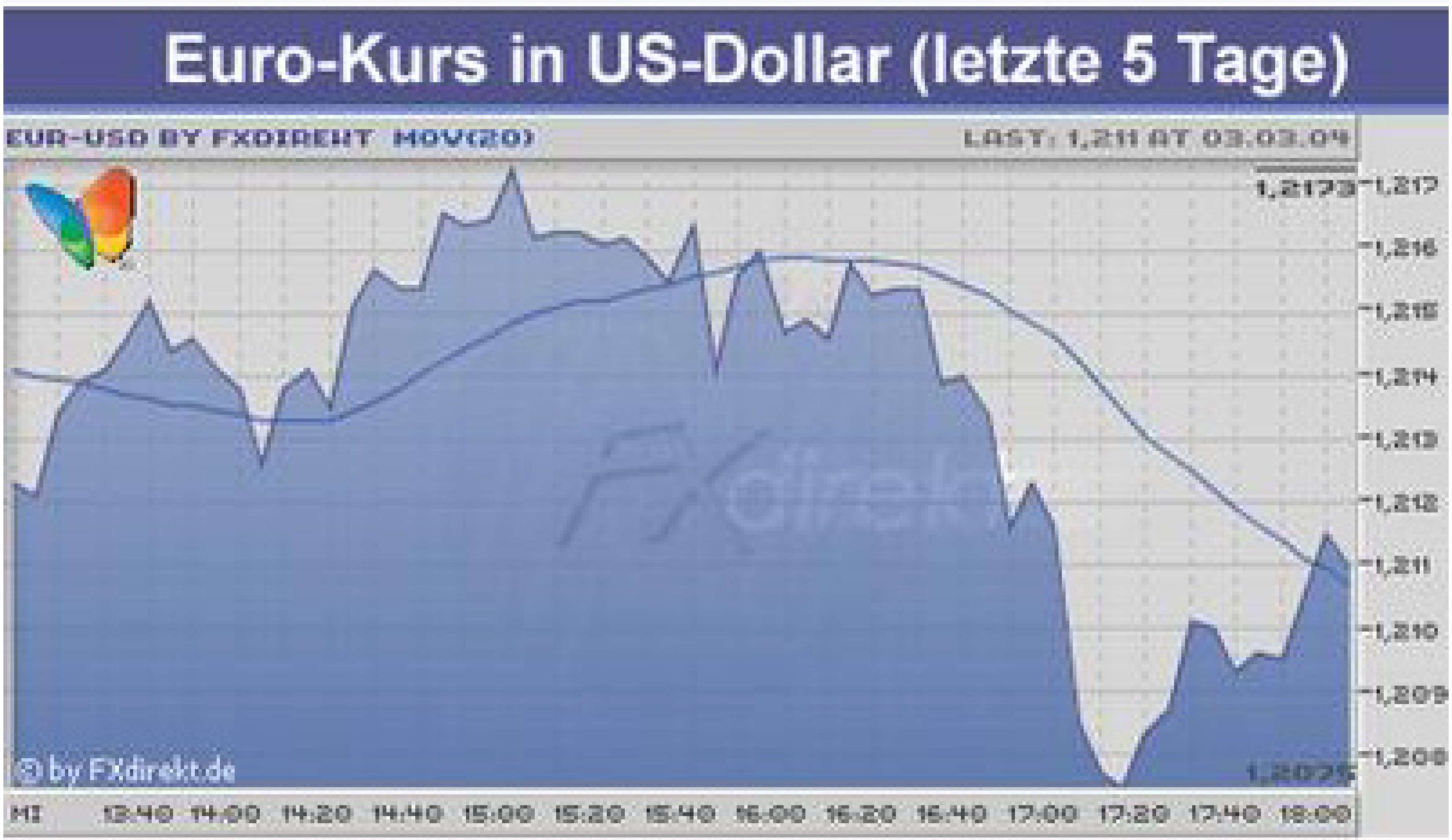
- What really happened is shown here
  - We intuitively interpret a trend plot on a ratio scale





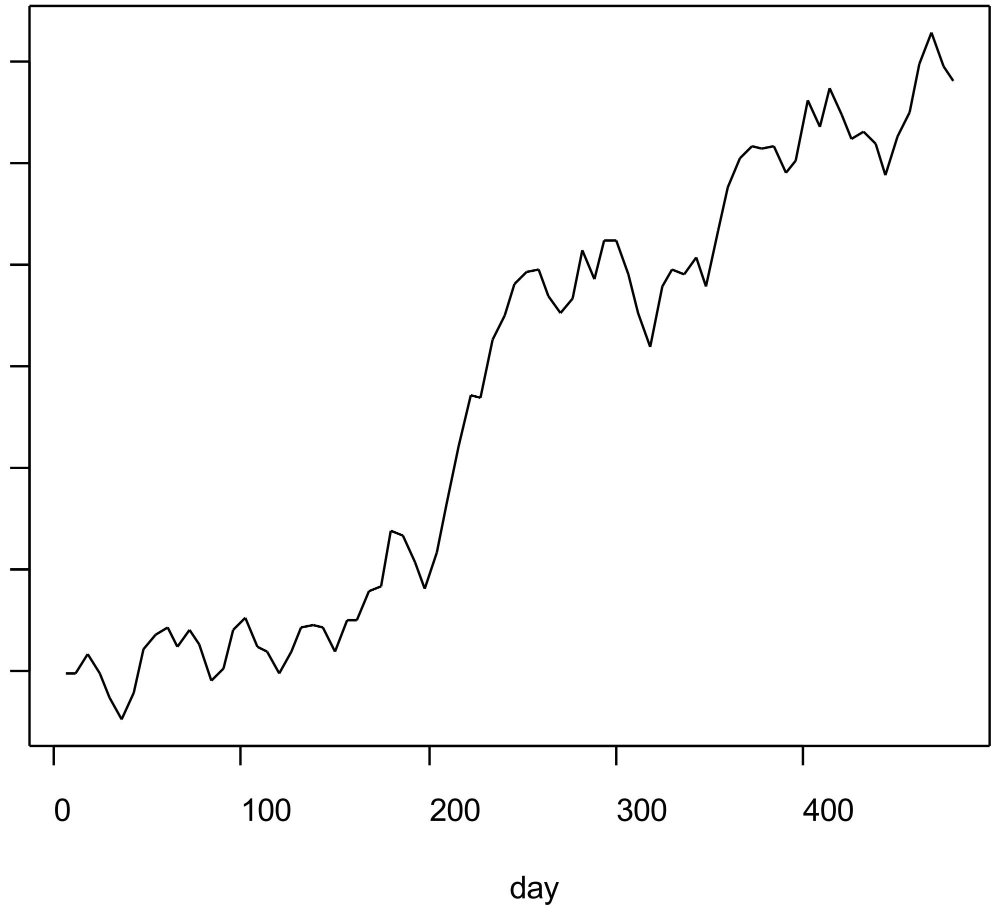
So look carefully!

found on focus.msn.de on 2004-03-04:



Problem: Scales can be missing

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

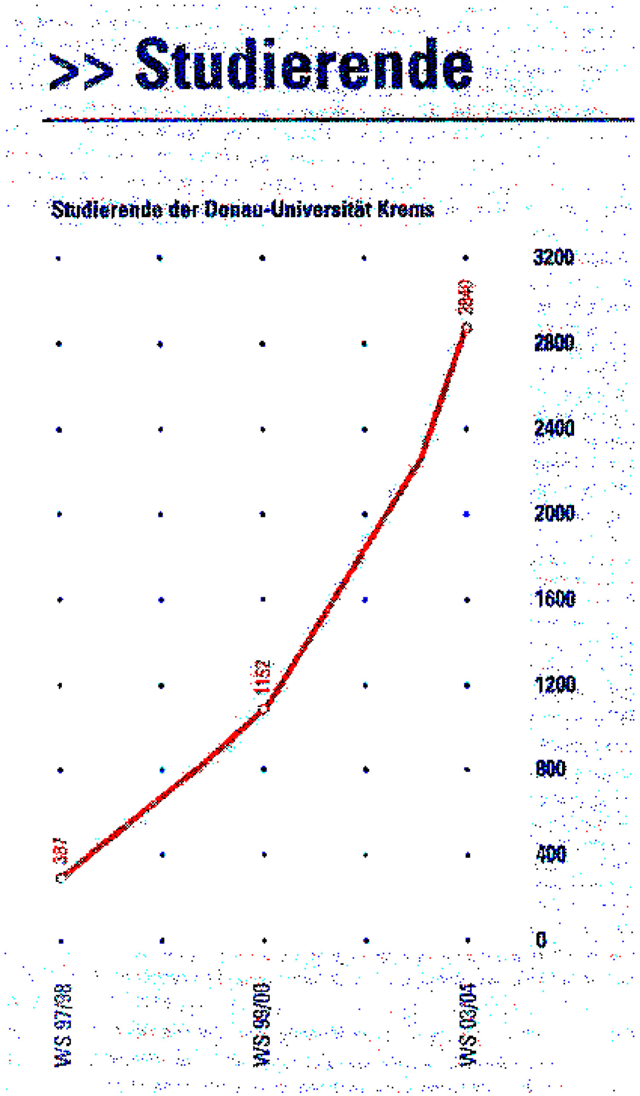


Problem: Scales can be abused

- Observe the global impression first

Problem:
People may invent unexpected things

- Quelle: Werbeanzeige der Donau-Universität Krems
  - DIE ZEIT, 07.10.2004



Lesson: 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?

blend-a-med Night Effects

Sichtbar hellere Zähne nach 14 Nächten – für mindestens 6 Monate.

- Zahnaufhellungsgel für die Nacht
- Klinisch getestet
- Einfach aufpinseln
- Mit patentierter LiquidStrip Technologie

- What exactly does "sichtbar" mean?
- What were the results of the clinical trials?
- What other effects does Night Effects have?

Example 5: the better tool?

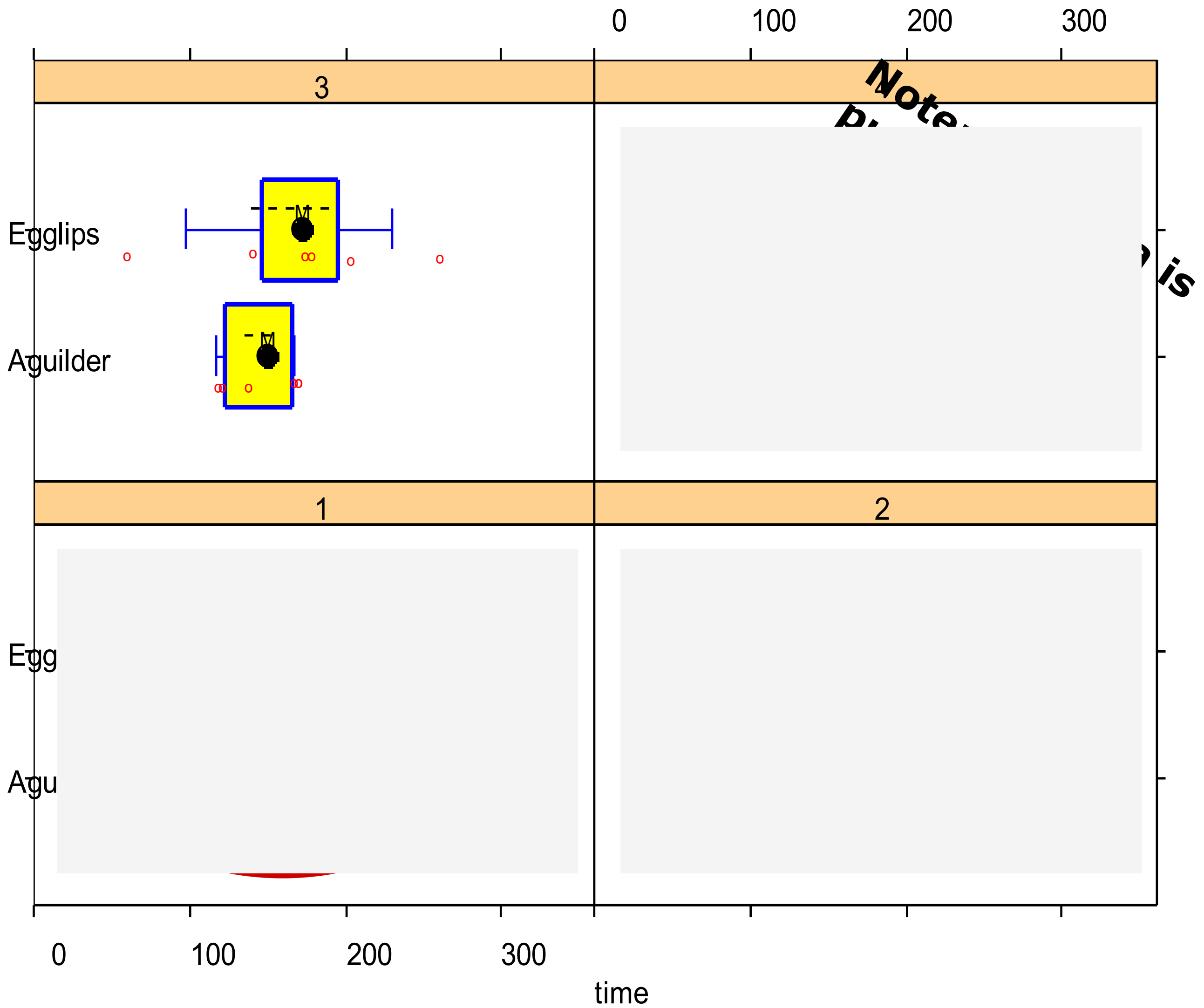
- We consider the time it takes programmers to write a certain program using different IDEs:
  - *Aguilder* or
  - *Egglips*

- 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: Demand complete information

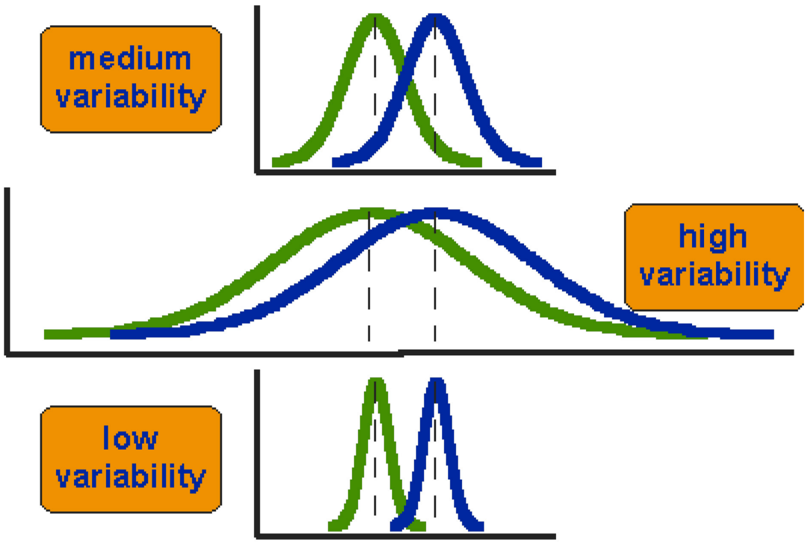
- If somebody presents conclusions
  - based on only a subset of the available data
  - and has selected which subset to use
  - then everything is possible

- There is no direct way to detect such repetitions,
  - which is why in some areas of clinical medicine, trials have to be registered before they start in order to become publishable results,

BUT for any one single execution . . .

Digression: Hypothesis testing

- …a so-called *significance test* can determine how likely it was to obtain such a result by chance:
  - assume both tools produce equal worktimes overall
    - as indeed they do in our case
    - this assumption is called the *null hypothesis*
      - the name means: the assumption that there is not really any difference (a null difference)
  - then how often will be get a difference at least this large when we use samples of size 6 persons?
  - If the probability is small, an effect is plausibly real
  - If the probability is large, the result is plausibly incidental



Statistical significance tests

- Our data:
  - Aguilder: 175, 186, 137, 117, 92.8, 93.7 (mean 133)
  - Egglips: 171, 155, 157, 181, 175, 160 (mean 166)
- 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 (or more) from two samples of size 6
- One procedure for doing this is called the *t-test* (`t.test` in R)
- Results (10 degrees of freedom):
  - p value: 0.08
    - the probability of getting at least the above difference of the means if the true difference is zero
  - 95% confidence interval for true difference: -5…71

So?

- So in our case we would probably believe the result and not find out that the experimenters had in fact cheated
  - because we do not know about the other 3 tries

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.

- (End of digression on hypothesis tests)

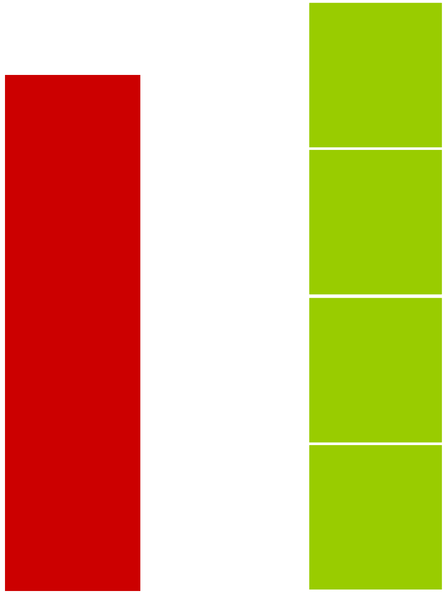
(Note: Significance testing on *correlations* is very nearly bullshit)

Example 6:
economic growth (D vs. USA)

- On 2003-10-30, the *US Buerau of Economic Analysis* (BEA) announced
  - USA economic growth in 3rd quarter: 7.2%
- Assume that same day the German *Statistisches Bundesamt* had announced
  - D economic growth in 3rd quarter: 2%
  - (Note: This value is fictitious)
- Note: Both values refer to gross domestic product (GDP, "Brutto-Inlandsprodukt", BIP)
- Which economy was growing faster?

Problem: Different definitions

- The US BEA <u>extrapolates</u> the growth for each quarter to a full year
  - Statistisches Bundesamt does not
- Thus, the actual US growth factor during (from start to end of) this quarter was only x, where $x^4 = 1.072$.
  - x = 1.0175
  - ➔ US growth was only 1.75% <u>in</u> this quarter



Example 7:
unemployment rate (D vs. USA)

Freie Universität Berlin

- (Source: DIE ZEIT 2004-02-05, p. 23: "Rot-weiß-blaues Zahlenwunder")

- 2003-11: USA: **5.9%** D: **10.5%**
- Which country had the higher unemployment rate?
- What does the number mean?:
  - D: registered as unemployed at the Arbeitsamt
  - USA: telephone-based micro-census by Bureau of Labor Statistics (BLS):
    - 1. Are you without work? (less than 1 hour last week)
    - 2. Are you *actively* searching for work?
    - 3. Could you start on a new job within 14 days?
    - <u>Only people with 3x "yes"</u> qualify as unemployed
  - A phone census is also performed by Statistisches Bundesamt
    - Result: **9.3%** unemployed (rather than 10.5%)
      - called "erwerbslos" (as opposed to "arbeitslos")
    - Because people are more honest on the telephone
    - But the rules are still not quite the same…

Unemployment rate (continued)

- USA: The census ignores
  - people who read job ads, but do not search actively
  - people who do not believe they can find a job
    - counting them would increase the rate by **0.5%**
  - 15-year-olds (who are unemployed very frequently)
- D: All these are included in the numbers
- Furthermore: People disappear from the statistic
  - USA: 760 of every 100000 people are in prison (as of 2003). That decreases the rate by **0.76%**
  - D: 80 of every 100000. Decreases rate by **0.08%**
  - D: Some people are "parked" on ABM
  - And more effects (in both countries)

- The overall result is hard to say

Lesson: Demand precise definitions

- Only because two numbers have the same name does not mean they are equivalent
  - in particular if they come from different contexts

- If no precise definitions of terms are available, only very large differences can be trusted

Example 8: productivity

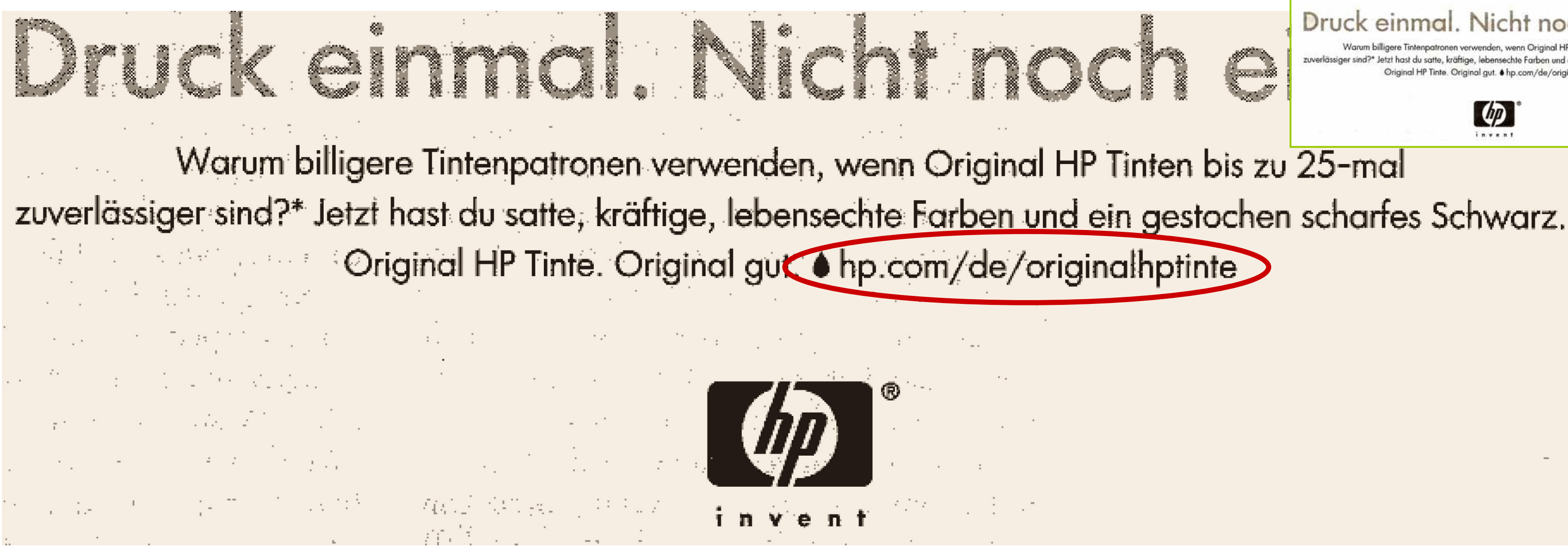
- Steve Walters on comp.software-eng (early 1990s):
  - "We just finished a software development project and discovered some curious metrics. This was a project in which we had good domain experience and about six years of metrics, both team productivity and other analogous software of similar scope and functionality.
  - The difference with this project was that we switched from a functional design methodology to OO.
  - First the good news: the overall team productivity (SLOC/personmonth) was almost three times our previous rate.
  - Now for the bad news: the delivered SLOC was almost three times greater than estimated, based on the metrics from our previous projects."

Lesson: Precise measurements can be invalid

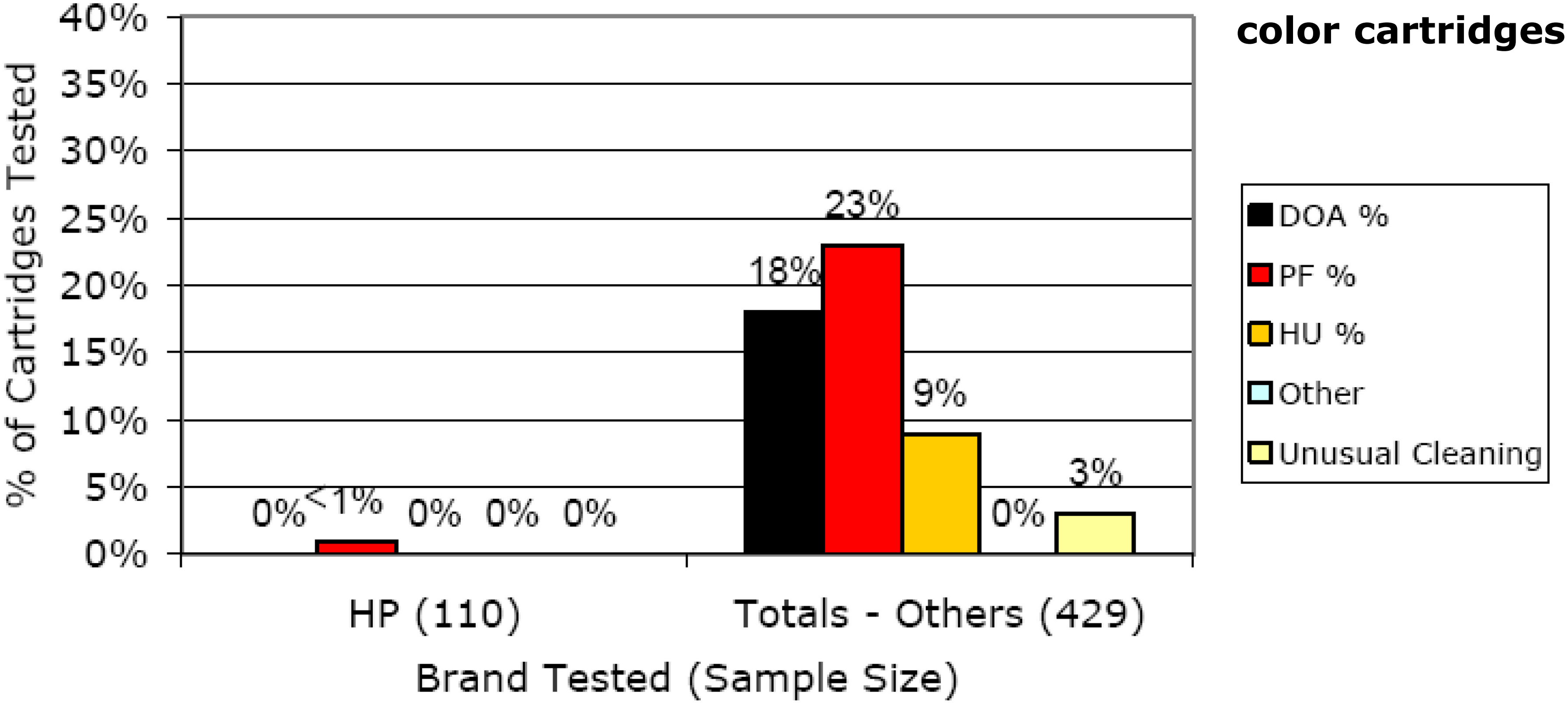
- Often a statistic is used for a purpose that it does not exactly fit to.
  - Perhaps nothing better is realistically possible
- But even if the numbers themselves are correct and precise, the conclusions may be totally wrong.
  - It is not sufficient that statistics are correct when at the same time they are inappropriate
    - Here: SLOC/personmonth has low construct validity for measuring productivity
- Such proxy measurements are very common.
  - Beware!

Real-world example:
25-fold reliability

- "Warum billigere Tintenpatronen verwenden, wenn Original HP Tinten bis zu 25-mal zuverlässiger sind?"
  - "Why use cheaper ink cartridges when genuine HP ink is up to 25 times more reliable?"



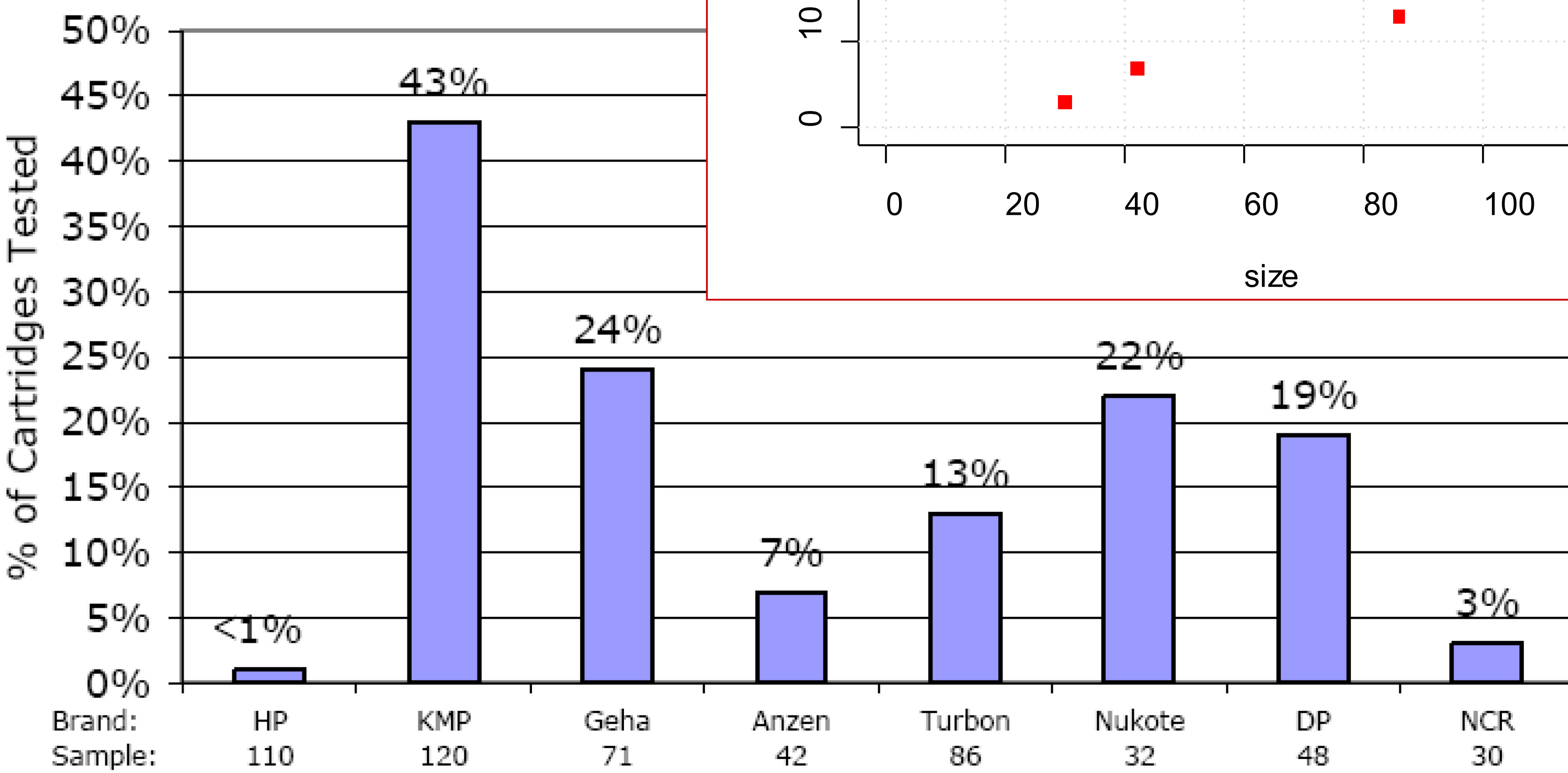
25-fold reliability explanation

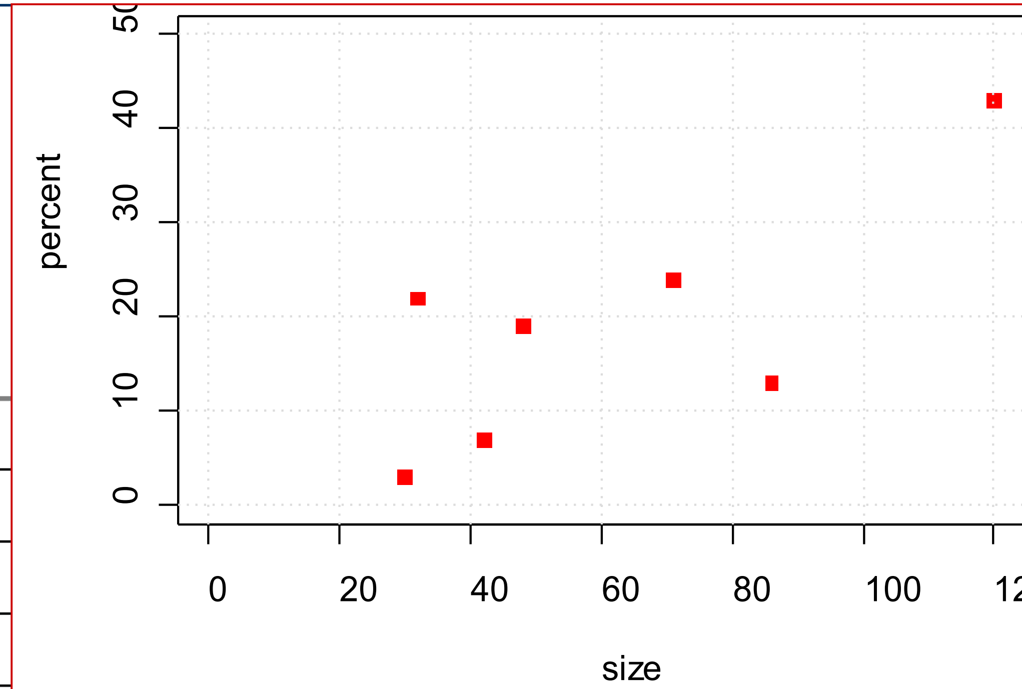


- DOA: Dead-on-arrival (<10 pages usable capacity)
- PF: premature failure (<75% of avg. non-DOA yield)
- HU: high unusable (>10% pages with low quality)

25-fold reliability explanation (2)

- Percentage of PF cartridges (less than 75% of the avg. capacity of all cart's.) per brand





25-fold reliability explanation (3)

Manufacturer	Model	Sample Size	Adjust Sample Size	DOA #	DOA %	Premature Failure #	Premature Failure %
KMP	656c	40	39	1	3%	2	5%
	990cxi	80	64	16	20%	50	63%
KMP Total		120	103	17	14%	52	43%

More problems with this data:

- 52/120 = 43% is what they used
- 52/103 = 50% is right if PF excludes DOA (as claimed)

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
  - and avoid distortions of any kind

Thank you!