

Course "Empirical Evaluation in Informatics"

How to lie with statistics

Christopher Oezbek, Prof. Dr. Lutz Prechelt
Freie Universität Berlin, Institut für Informatik

<http://www.inf.fu-berlin.de/inst/ag-se/>

- 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

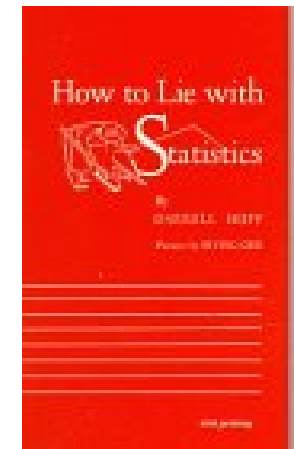
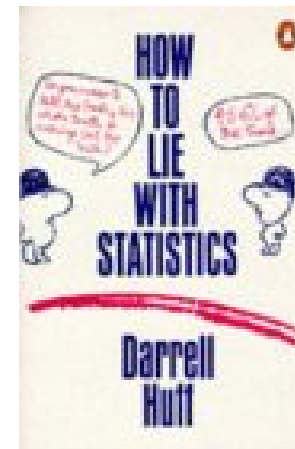
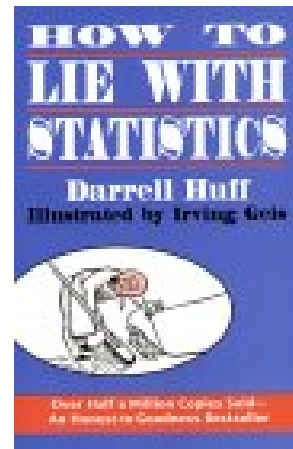
"Empirische Bewertung in der Informatik"

Wie man mit Statistik lügt

Christopher Oezbek, Prof. Dr. Lutz Prechelt
Freie Universität Berlin, Institut für Informatik
<http://www.inf.fu-berlin.de/inst/ag-se/>

- 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 nicht gesagt?
- "Probier einfach noch mal"
- Unvergleichbare Daten
- Gültigkeit von Maßen

- This slide set was created *roughly* along the lines of
Darrell Huff: "How to Lie With Statistics",
(Victor Gollancz 1954, Pelican Books 1973, Penguin Books 1991)
- I urge 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



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?

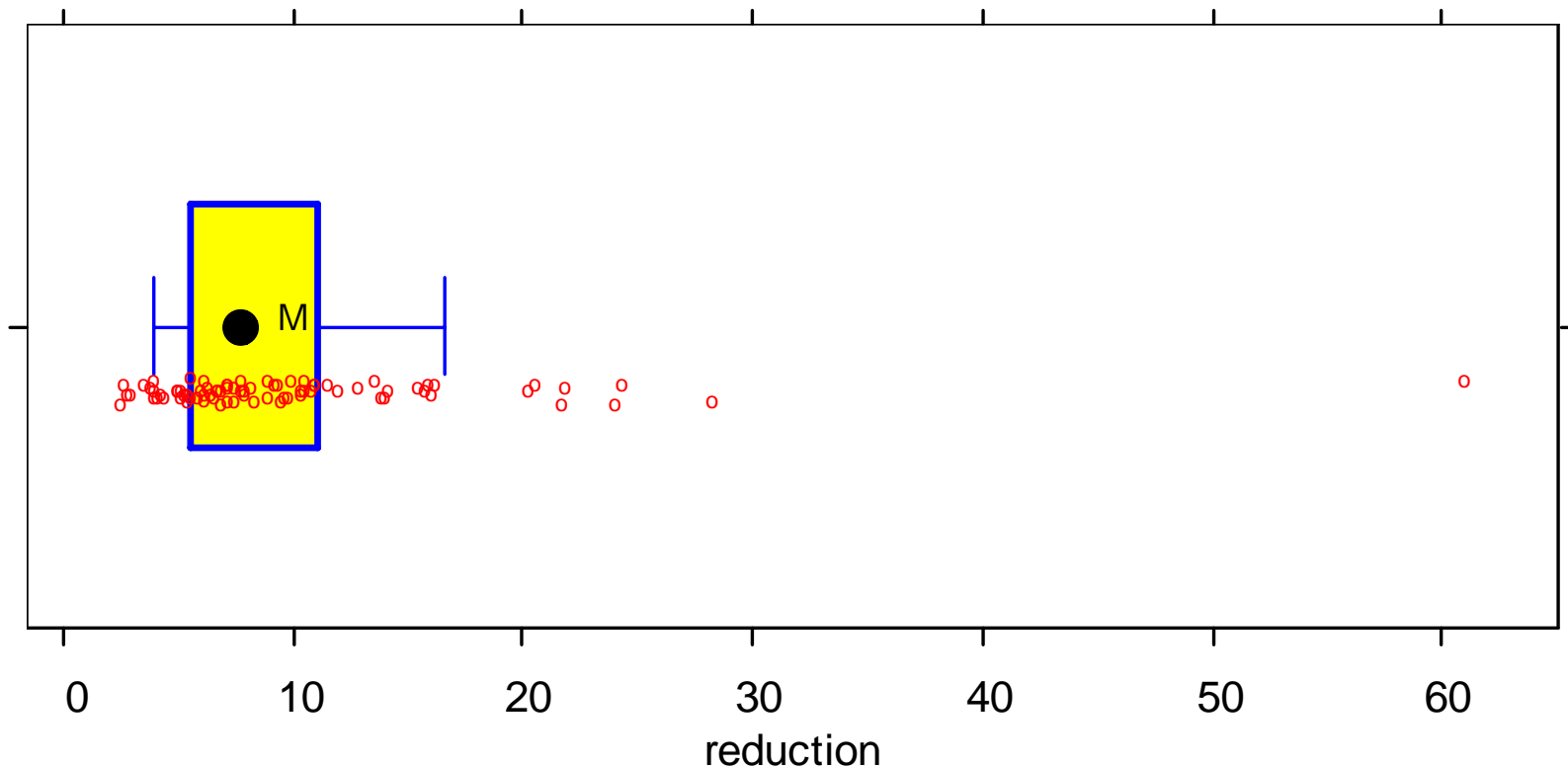
- "Body fat loss: up to 82%"
 - OK, can be measured
- "Wrinkle reduction: up to 61%"
 - Maybe they count the wrinkles and measure their depth?
- "Energy level: up to 84%"
 - What is this?
 - Also note they use language loosely:
 - Loss in percent: OK; reduction in percent: OK
 - Level in percent??? (should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

Lesson: 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:



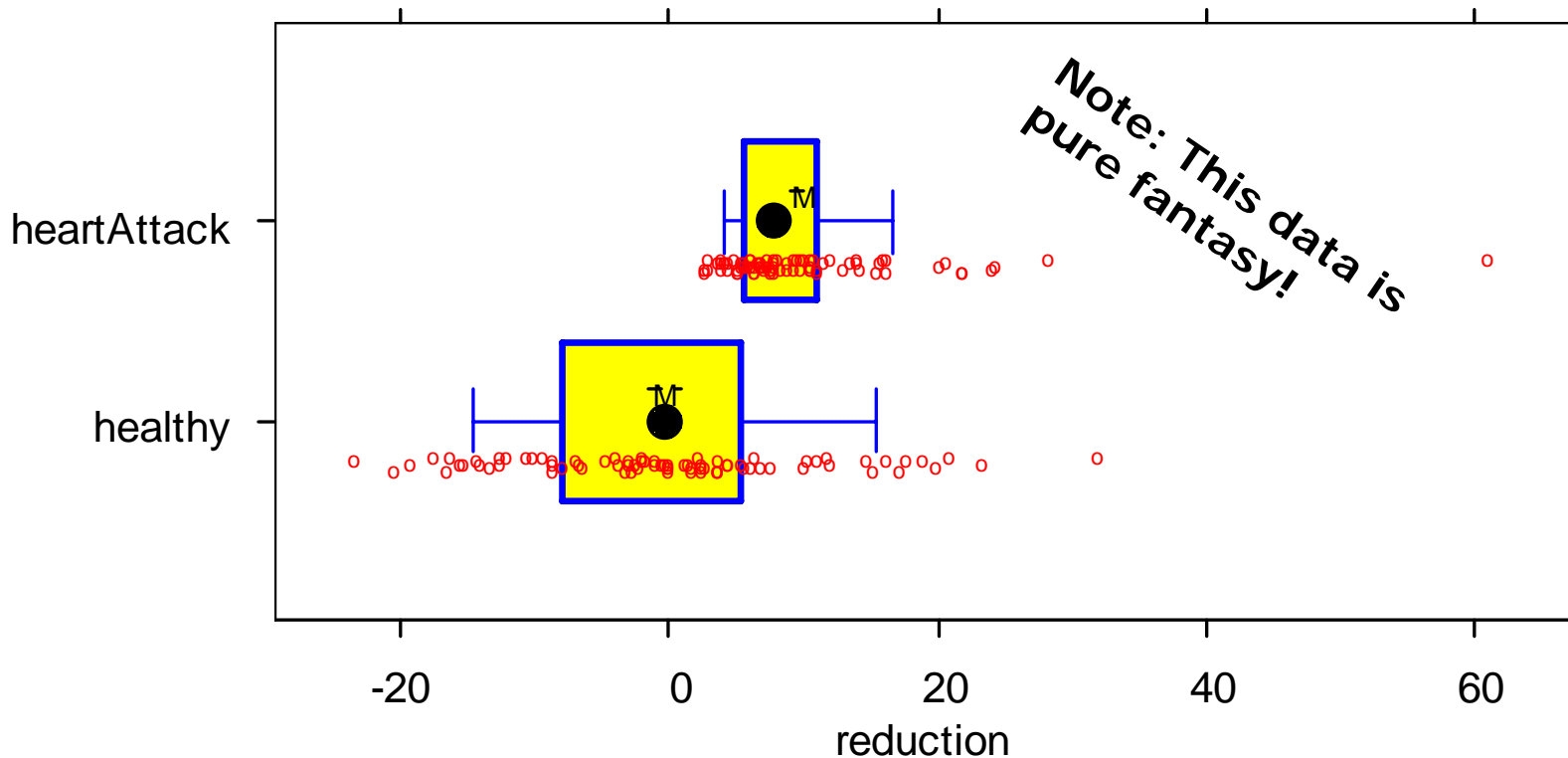
Lesson:

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 something 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

Problem 3: Underlying population

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



Lesson: Insist on unbiased samples

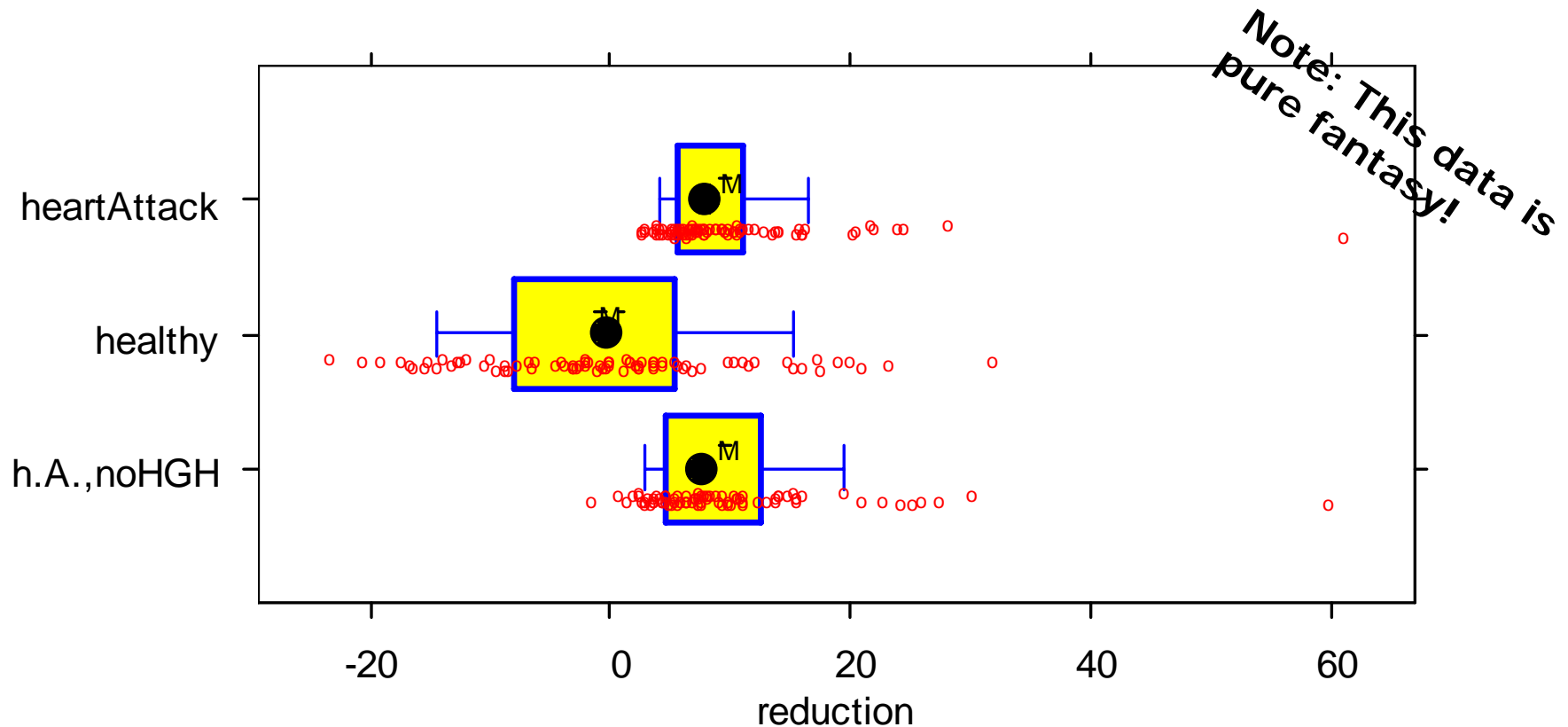
- How and where from the data was collected can have a tremendous impact on the results
- It is important to understand whether there is a certain (possibly intended) tendency in this
- A fair statistic talks about possible *bias* it contains
- If it does not, ask.

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?



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?

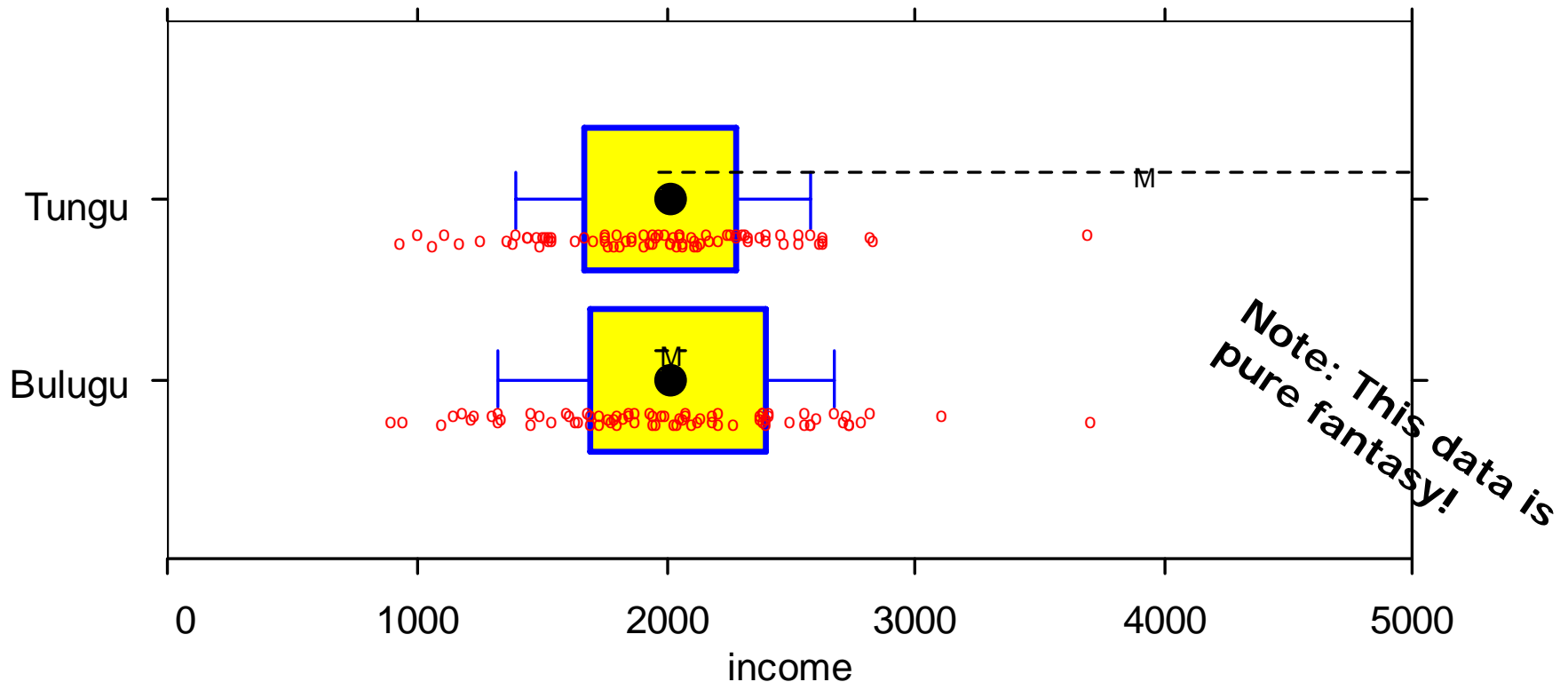
Example 2: Tungu and Bulugu

- We look at the yearly per-capita income in two small hypothetical island states:
Tungu and Bulugu
- Statement:
"The average yearly income in Tungu is 94.3% higher than in Bulugu"



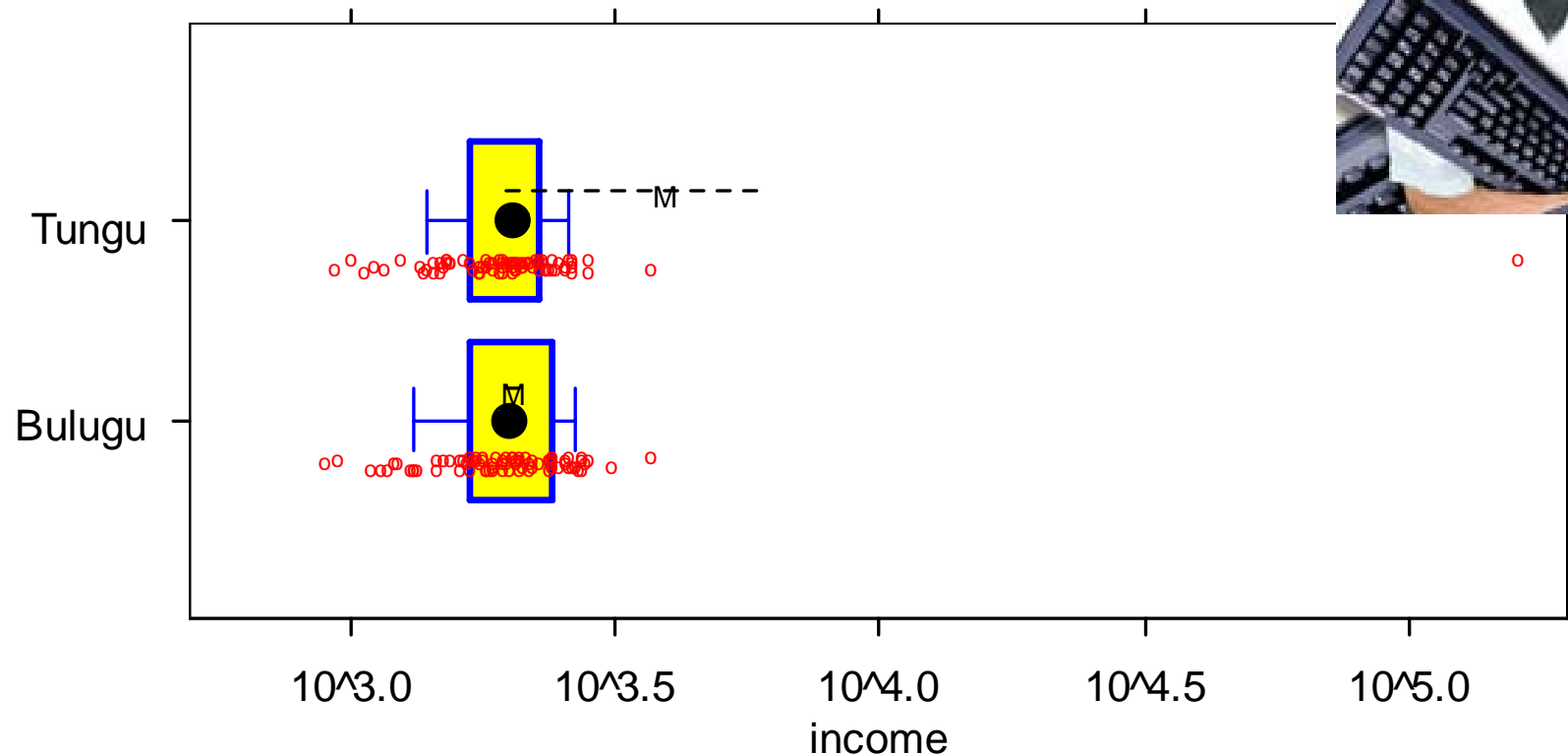
Problem 1: Misleading averages

- 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

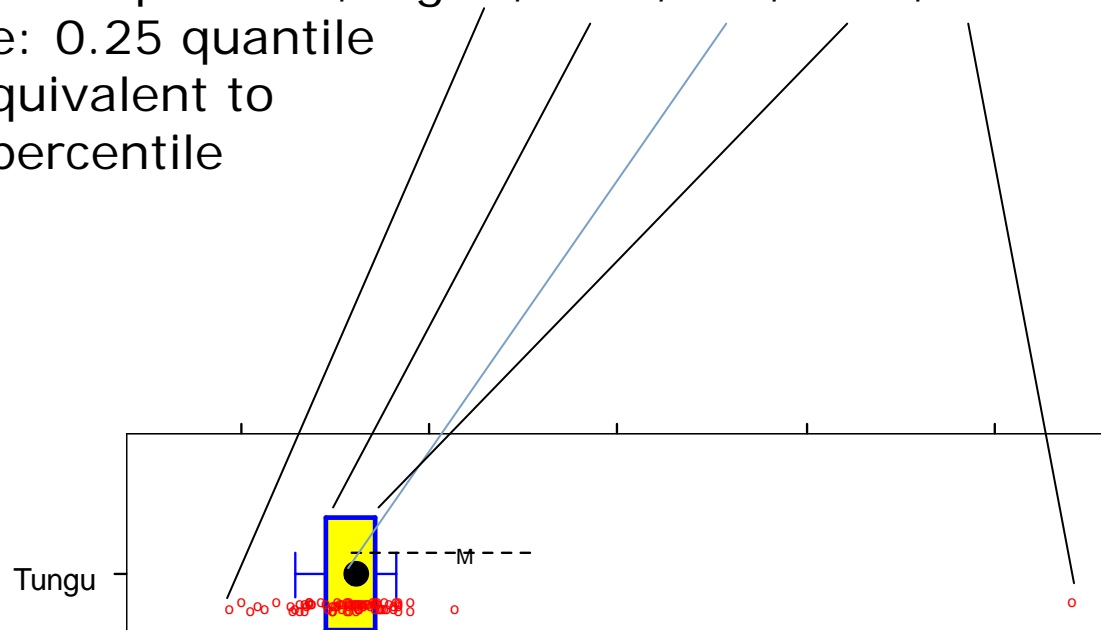
- The only reason is Dr. Waldner, owner of a small software company in Berlin, who since last year is enjoying his retirement in Tungu



Lesson: Question appropriateness

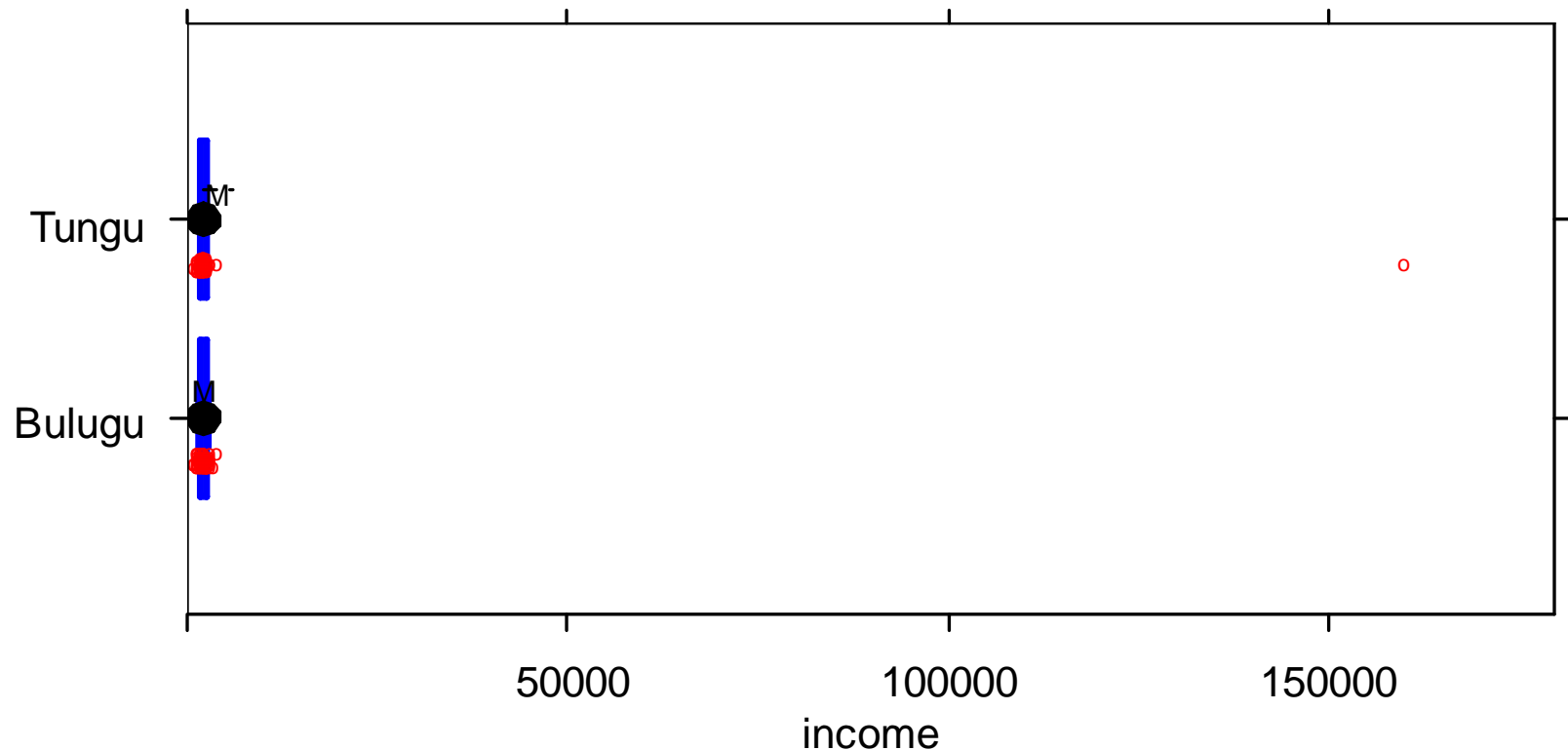
- A certain statistic (very often the arithmetic average) may be inappropriate for characterizing a sample
- If there is any doubt, ask that additional information be provided
 - such as standard deviation
 - or some quantiles, e.g. 0, 0.25, 0.5, 0.75, 1

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



- 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

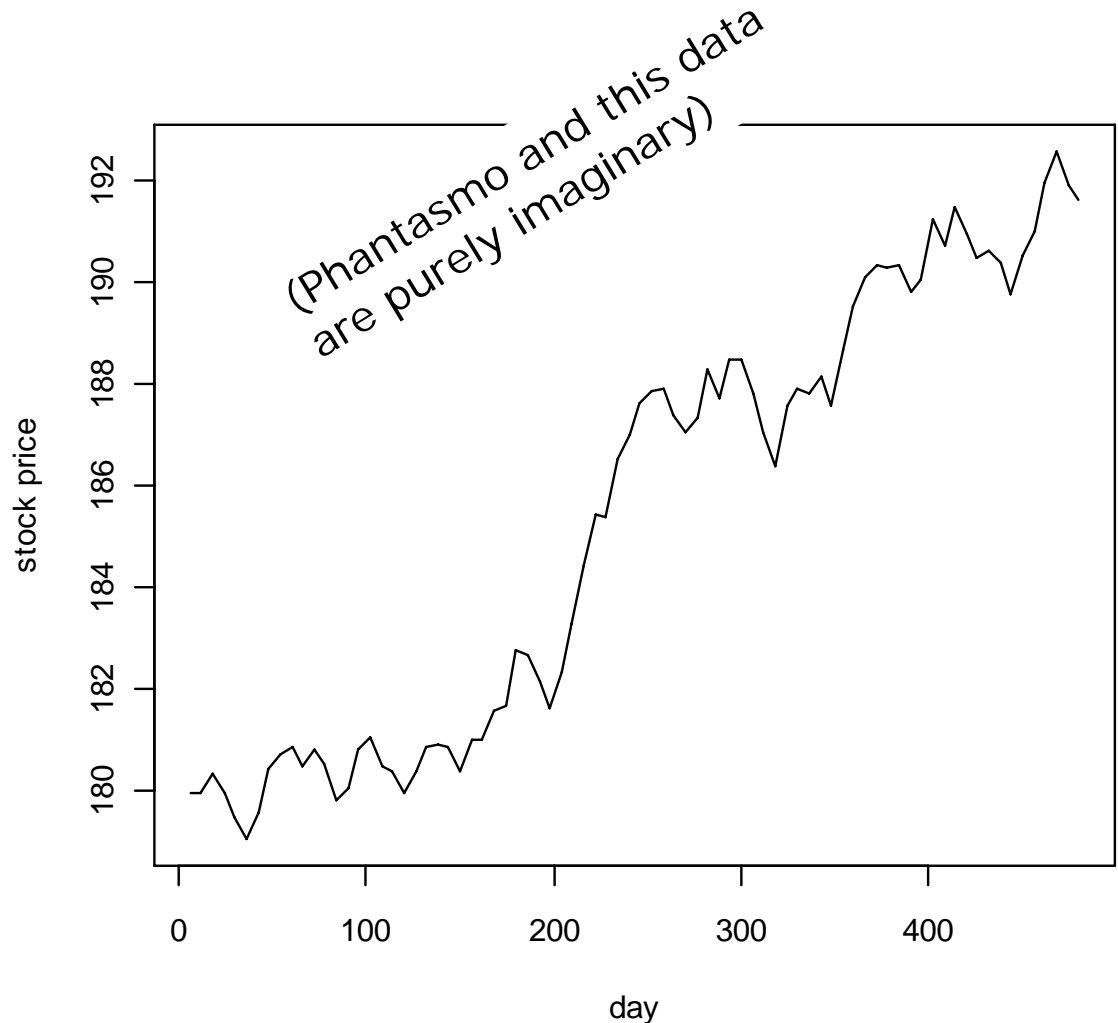
- "The average yearly income in Tungu is **94.3%** higher than in Bulugu"
- Assume that tomorrow Mrs. Alulu Nirudu from Tungu gives birth to her twins
- There are now 83 rather than 81 people on Tungu
- The average income drops from 3922 to 3827
- The difference to Bulugu drops from 94.3% to 89.7%

Lesson: Do not be easily impressed

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

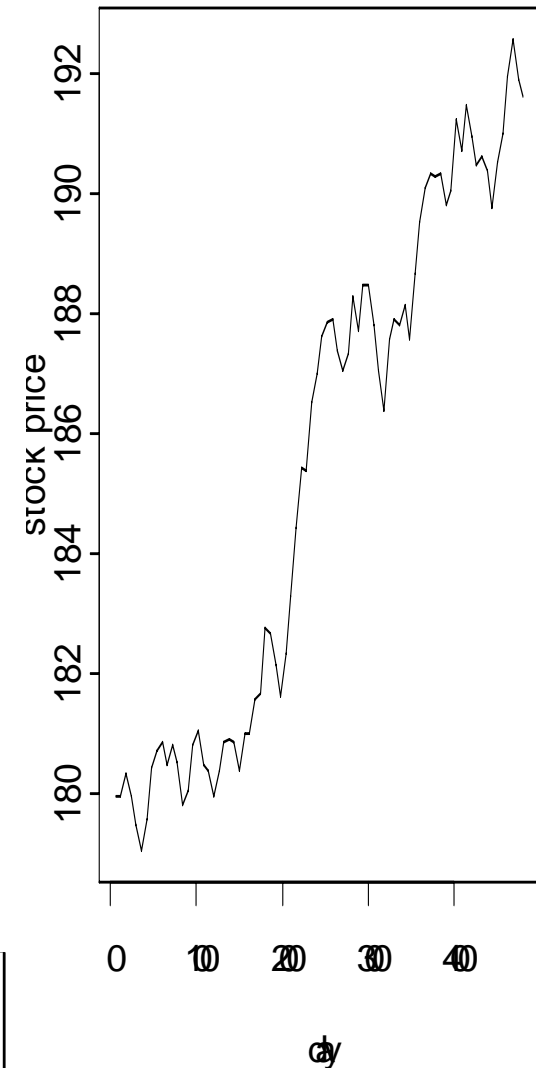
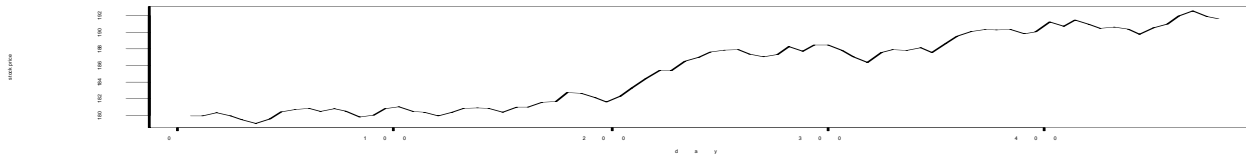
Example 3: Phantasma Corporation stock price

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



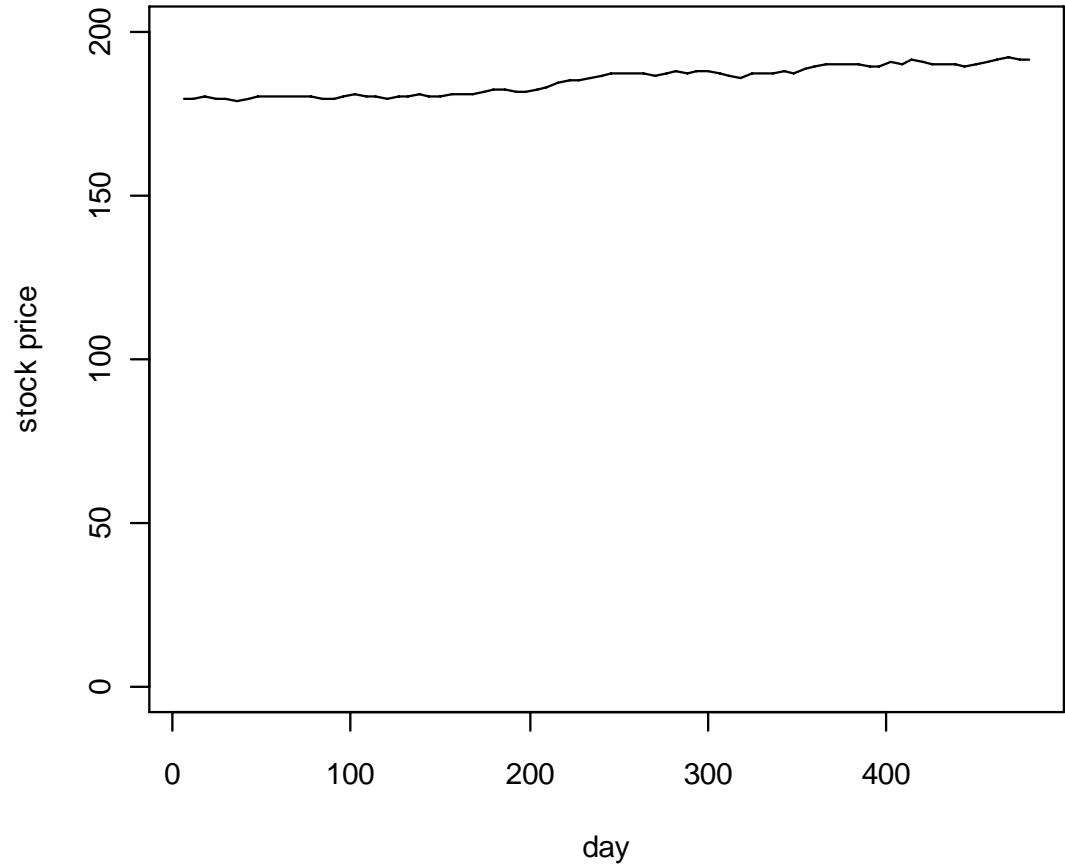
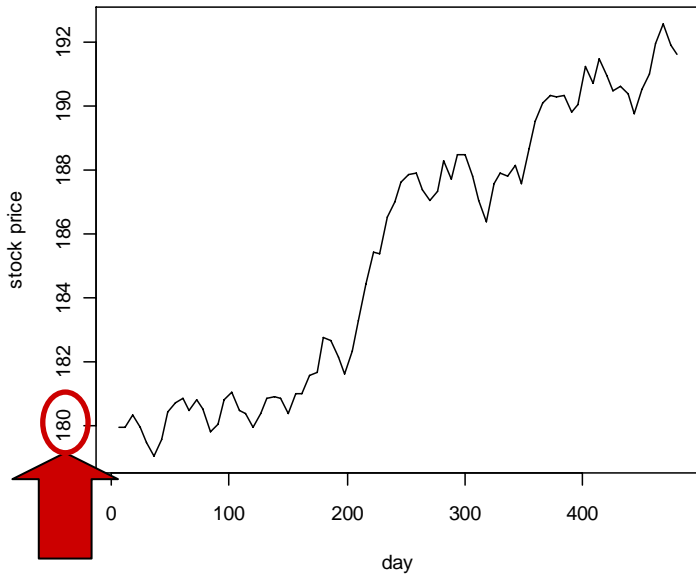
Problem: Looks can be misleading

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



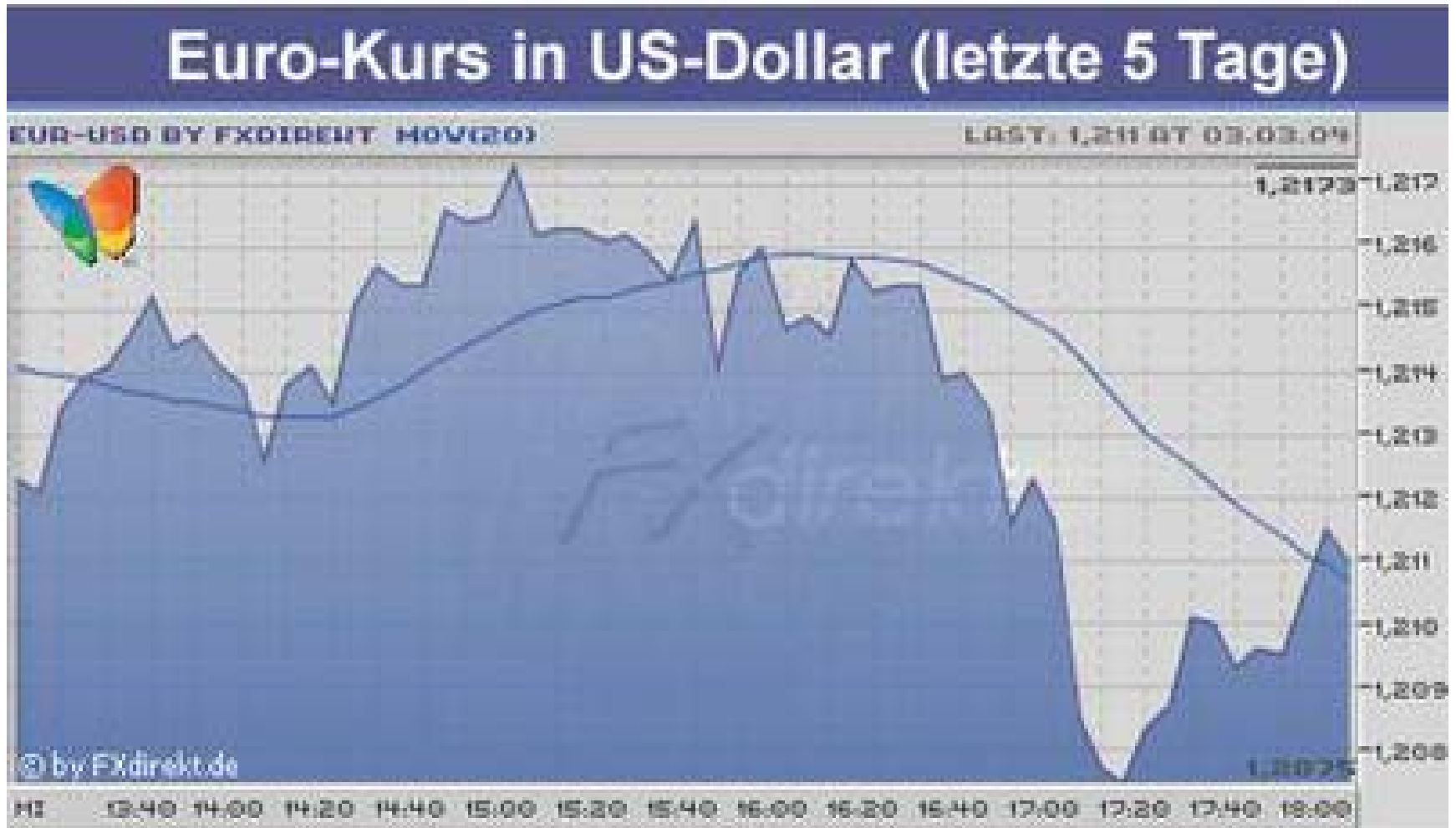
Problem: Scales can be misleading

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



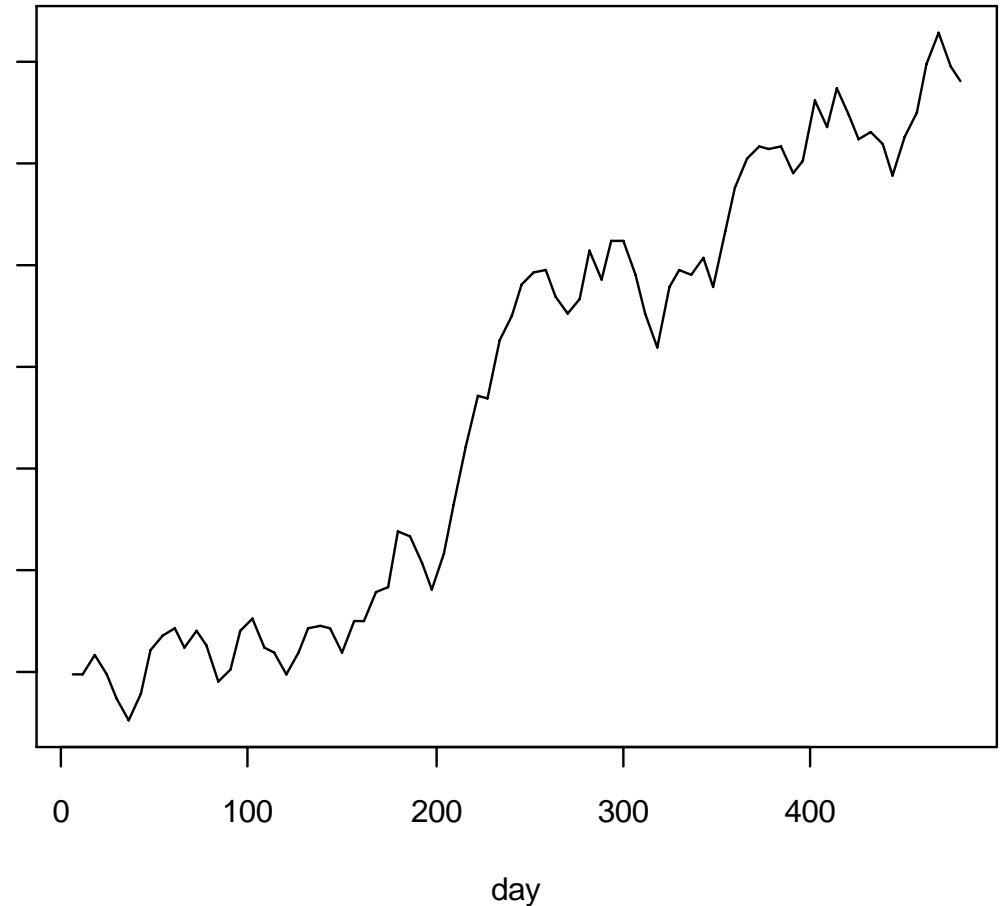
So look carefully!

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



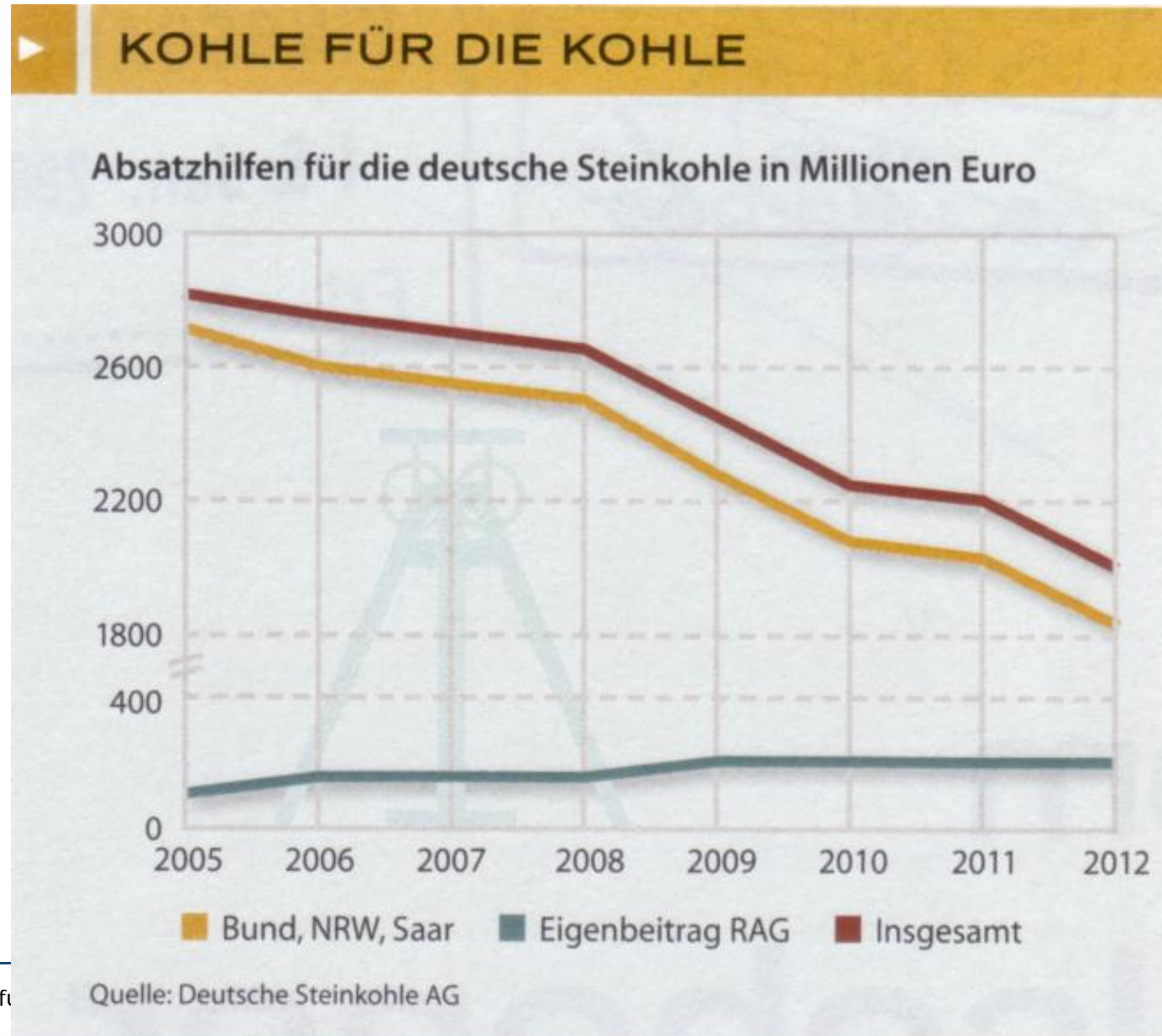
Problem: Scales can be missing

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



Problem: Scales can be abused

- Observe the global impression first

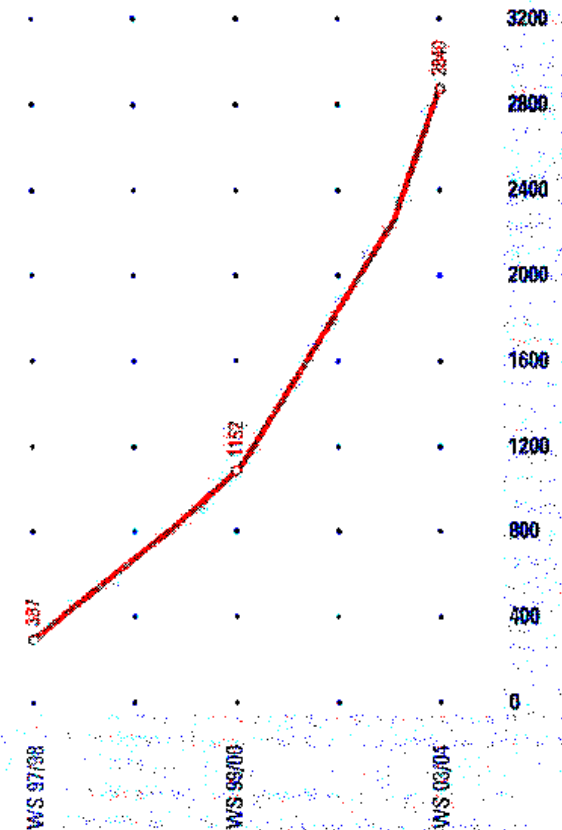


Problem: People may invent unexpected things

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

>> Studierende

Studierende der Donau-Universität Krems



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 not 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

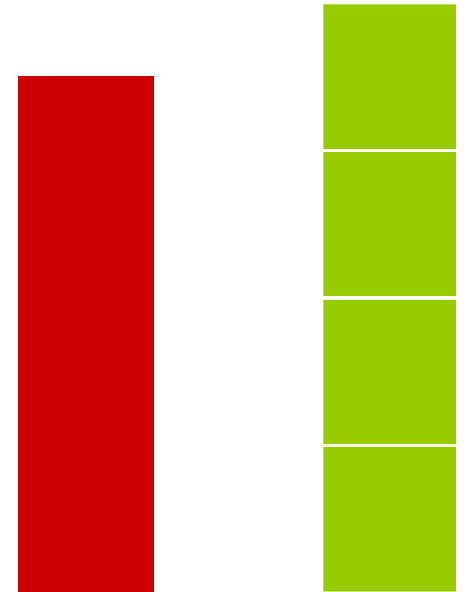
- What exactly does "sichtbar" mean?
- What were the results of the clinical trials?
- What other effects does Night Effects have?

Example 6: economic growth (D vs. USA)

- On 2003-10-30, the *US Bureau 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 extrapolates 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% in this quarter



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?"



Druck einmal. Nicht noch e

Druck einmal. Nicht noch einmal.

Warum billigere Tintenpatronen verwenden, wenn Original HP Tinten bis zu 25-mal zuverlässiger sind?* Jetzt hast du satte, kräftige, lebensechte Farben und ein gestochen scharfes Schwarz. Original HP Tinte. Original gut. • hp.com/de/originaltinte

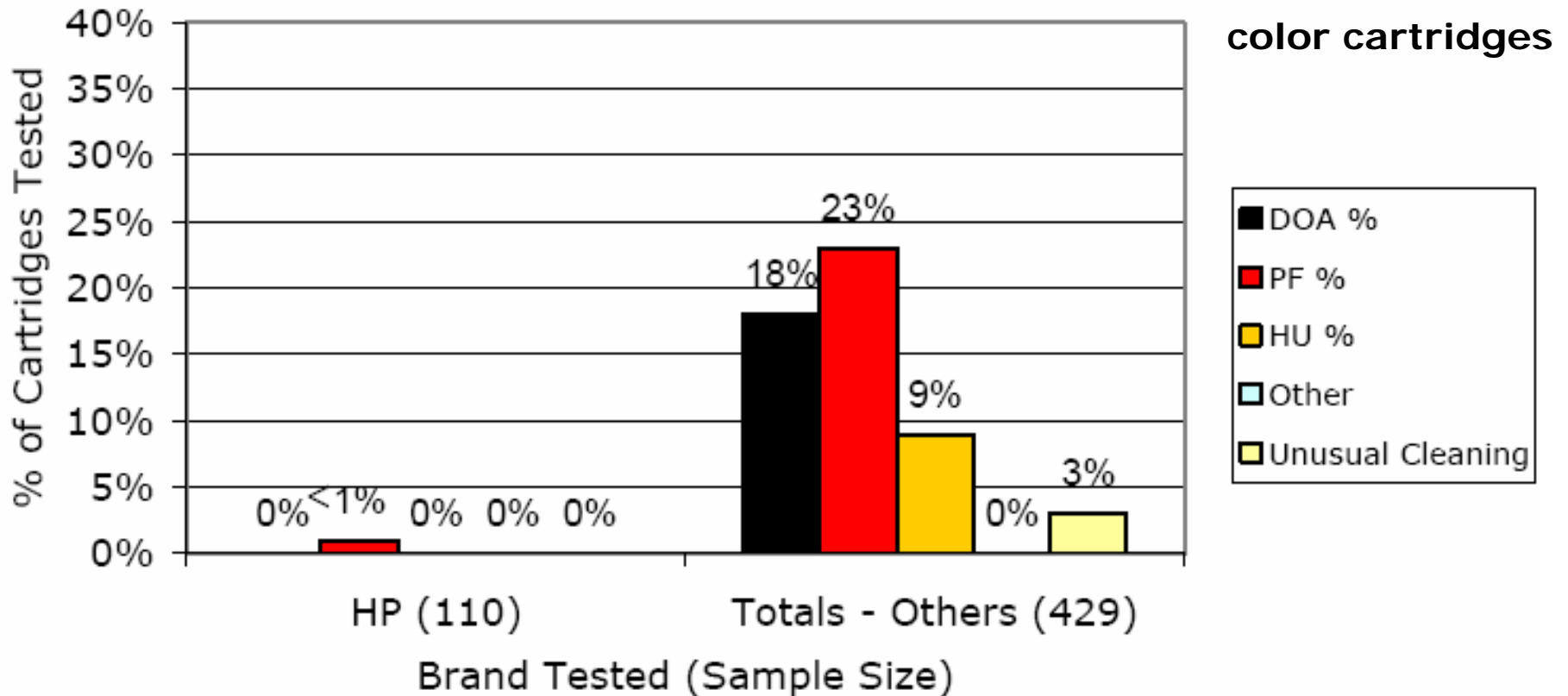


Warum billigere Tintenpatronen verwenden, wenn Original HP Tinten bis zu 25-mal zuverlässiger sind?* Jetzt hast du satte, kräftige, lebensechte Farben und ein gestochen scharfes Schwarz.

Original HP Tinte. Original gut. hp.com/de/originalhptinte



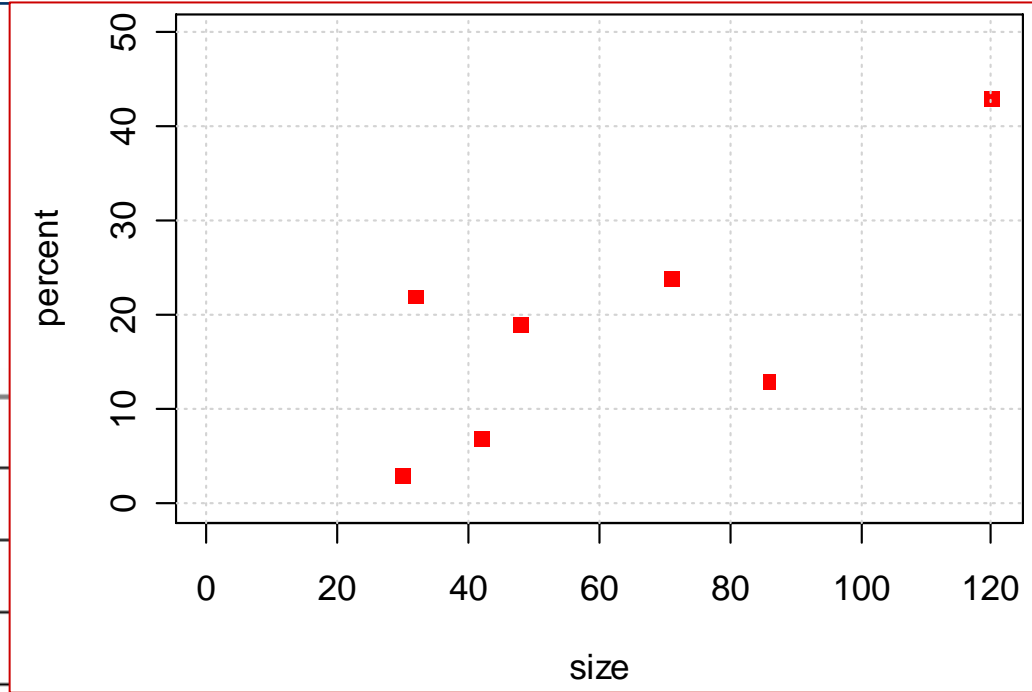
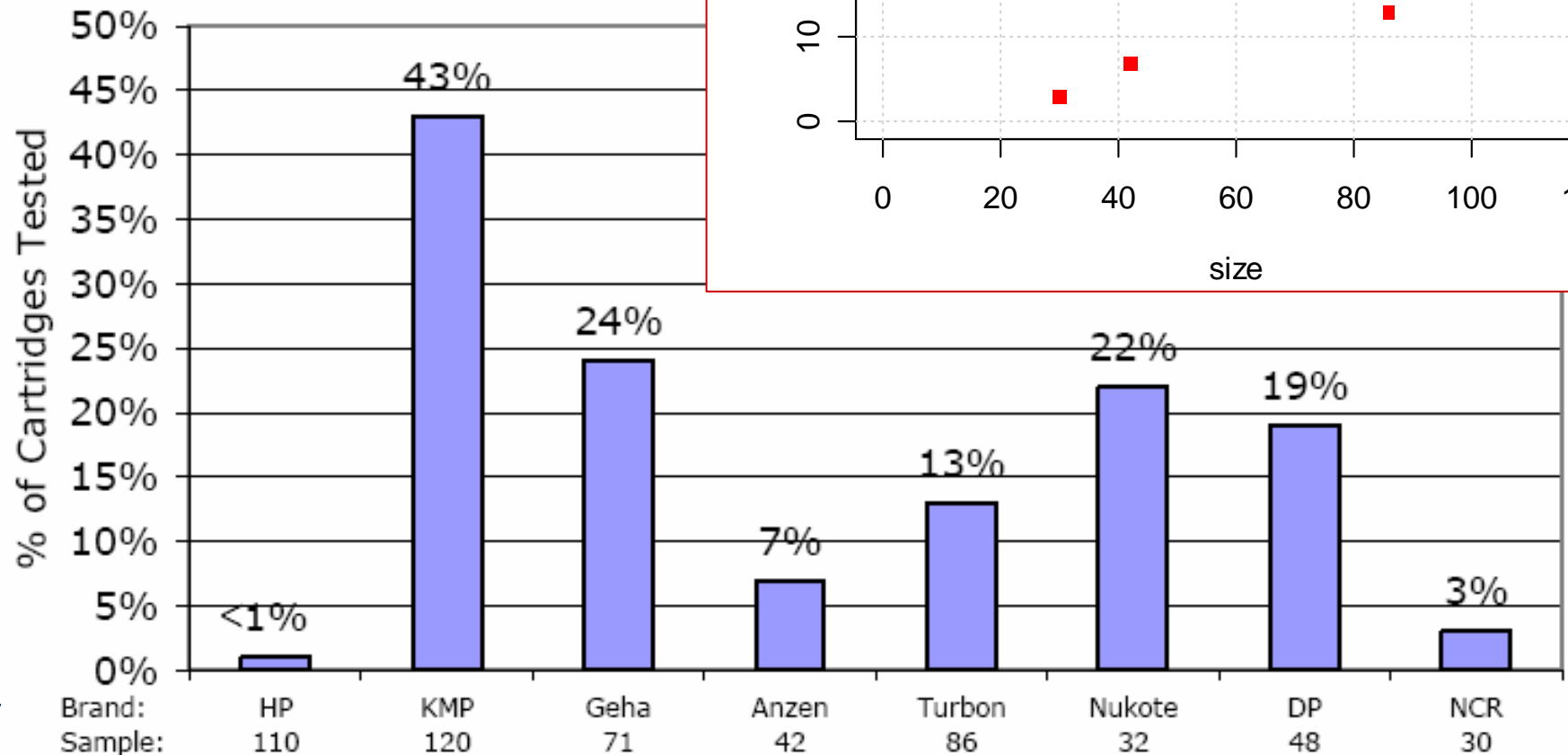
25-fold reliability explanation



- 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		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)
- $(52-17)/103 = 34\%$ is right if PF includes DOA

Thank you!