Basic Principles of Excellent Data Graphs

Frederic Malter, PhD

Evaluation, Research & Development Unit, University of Arizona

November 11, 2009

520-318-7259 fmalter@email.arizona.edu

EVALUATION, RESEARCH AND DEVELOPMENT UNIT
THE UNIVERSITY OF ARIZONA

Funded by the Arizona Department of Health Services

Why data graphs?

"The graphical method has considerable superiority for the exposition of statistical facts over the tabular. A heavy bank of figures is grievously wearisome to the eye and the popular mind is as incapable of drawing any useful lessons from it as of **extracting sunbeams from cucumbers.**"

Farquhar & Farquhar, 1891

This presentation & workshop…

- general principles that constitute a reference for judgments about graphs: normative & authoritative
- Will anchor basic principles in cognitive science & information theory (Tufte)
- Focuses on COMMUNICATION, not data analysis (e.g. Tukey)
- Will debunk „infographics“

Disclaimer

- I’m not paid by Microsoft OR Tableau.
- I do not want to devaluate other people’s work or opinions.
- I‘m not the pope. Ed Tufte is neither.

Cognitive science of statistical reasoning

- Gigerenzer et al. (2007): „Collective statistical illiteracy“ (inability to understand numbers)
  - Lay people & „experts“ (MDs) don‘t understand %.

- Tufte (1997): „Visual presentation (...) should be governed by principles of reasoning about quantitative evidence. Clear and precise seeing becomes one with clear and precise thinking“. (Snow & Cholera epidemic, Challenger disaster)

Natural frquency graph (Gigerenzer et al. 2007, p. 55)

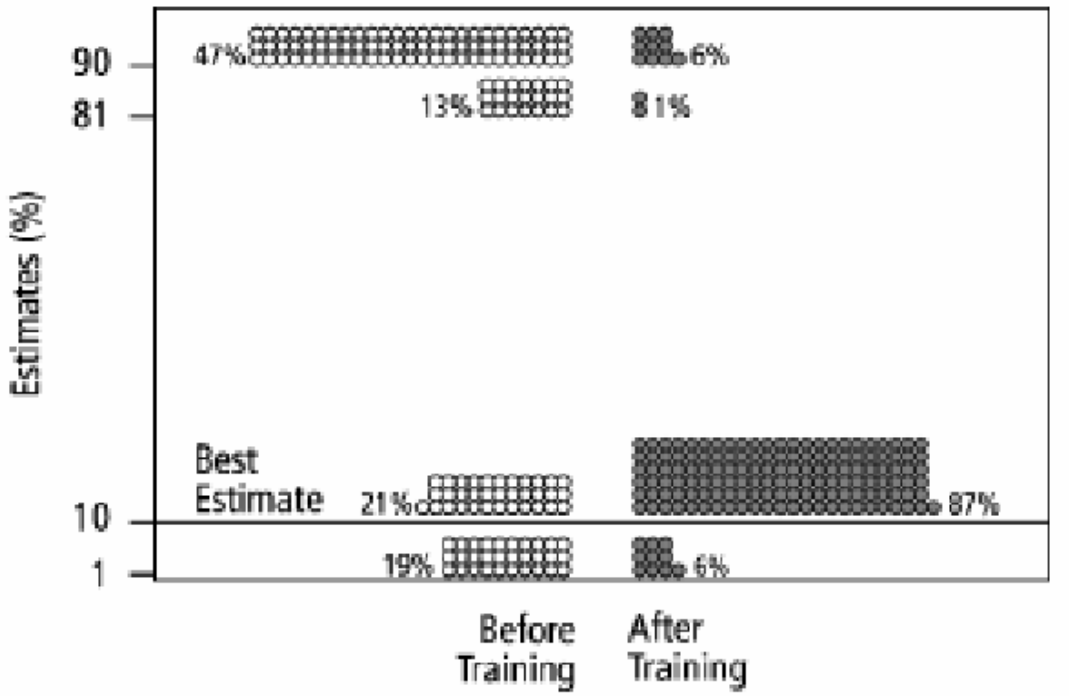


Fig. 2. Estimates by 160 gynecologists of the probability that a woman has breast cancer given a positive mammogram, before and after receiving training in how to translate conditional probabilities into natural frequencies.

Which format is most appropriate when?

- Common question: when text, when tables, when graphs?
- Start with a clear notion on what you want to communicate
  - Specific
  - Precise
  - Critical

- Tables work best when the data presentation:
  - Is used to look up individual values
  - Is used to compare individual values
  - Requires precise values
  - Values involve multiple units of measure.

- Graphs work best when the data presentation:
  - Is used to communicate a message that is contained in the shape of the data
  - Is used to reveal relationships among many values.

Graphical excellence (E. Tufte, 2001)

…is about communicating complex ideas with clarity, precision & efficiency

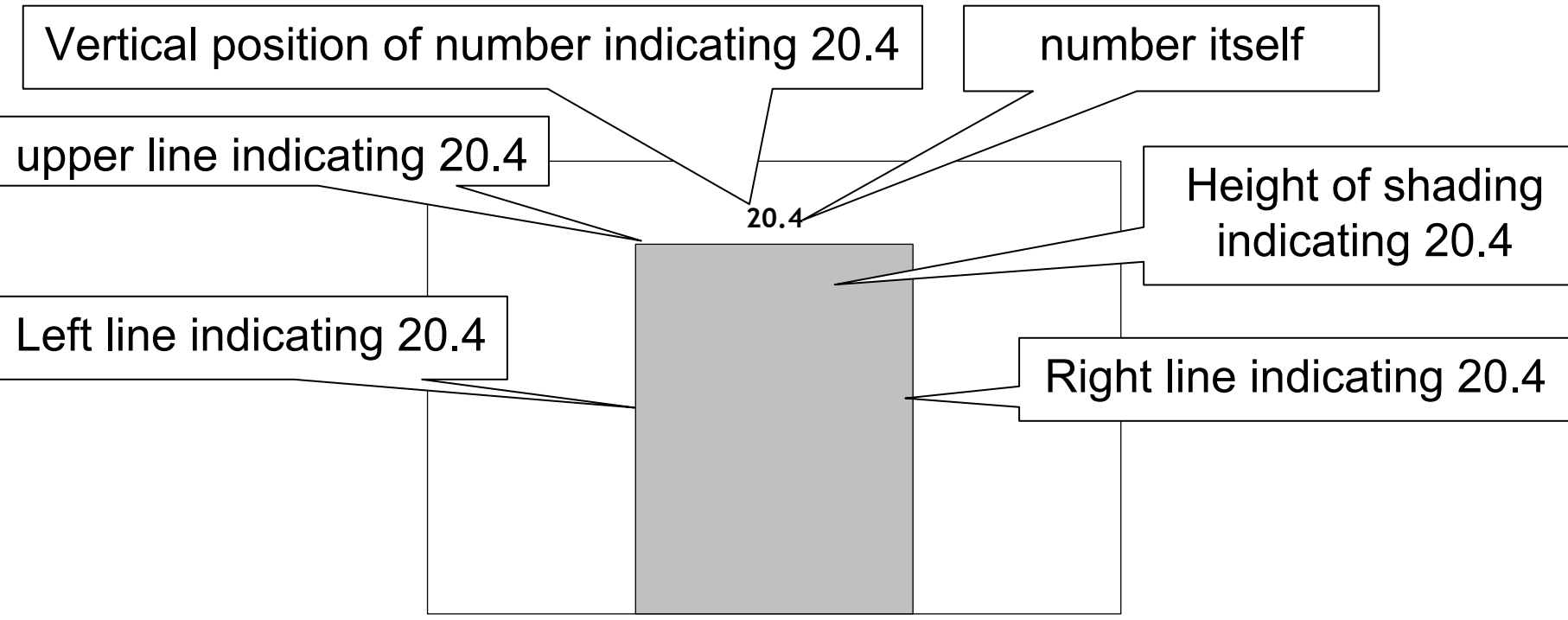
- Principles
  - Avoid distortion!
  - Graphs should present many numbers in a small space (data density)
  - Graphs should encourage the eye to compare different pieces of data
  - Data graphs do serve a **clear purpose**: **description, exploration, tabulation**.
  - Data graphs are NOT art (mostly auto-telic).

Theory of Data Graphics (Tufte)

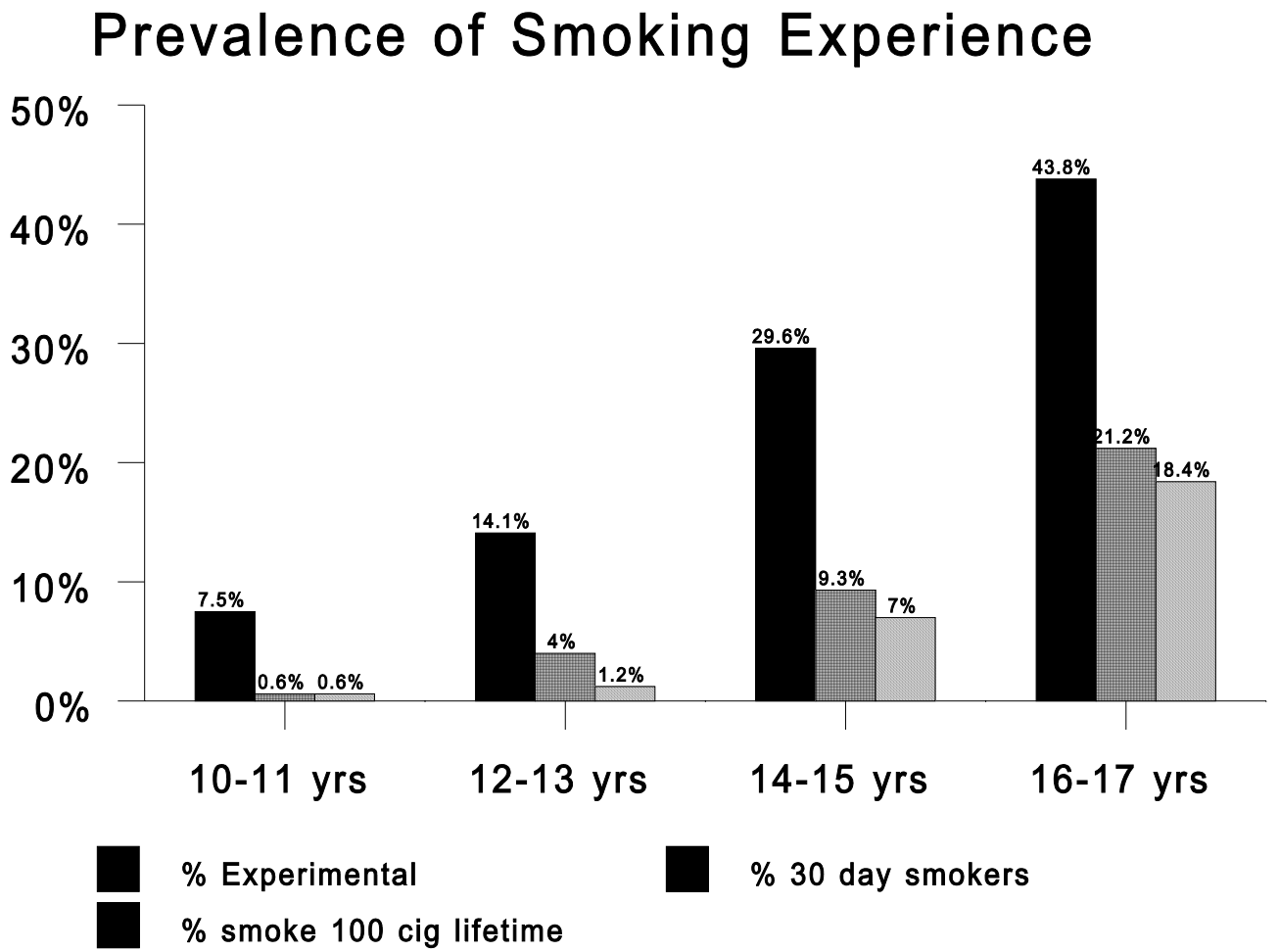
- Maximize data density.
  - Data density = $\frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$
  - Maximize the data-ink ratio.
  - Erase non-data-ink.
  - Erase redundant data-ink.
- Revise & edit, i.e. decide & delete.

Theory of Data Graphics

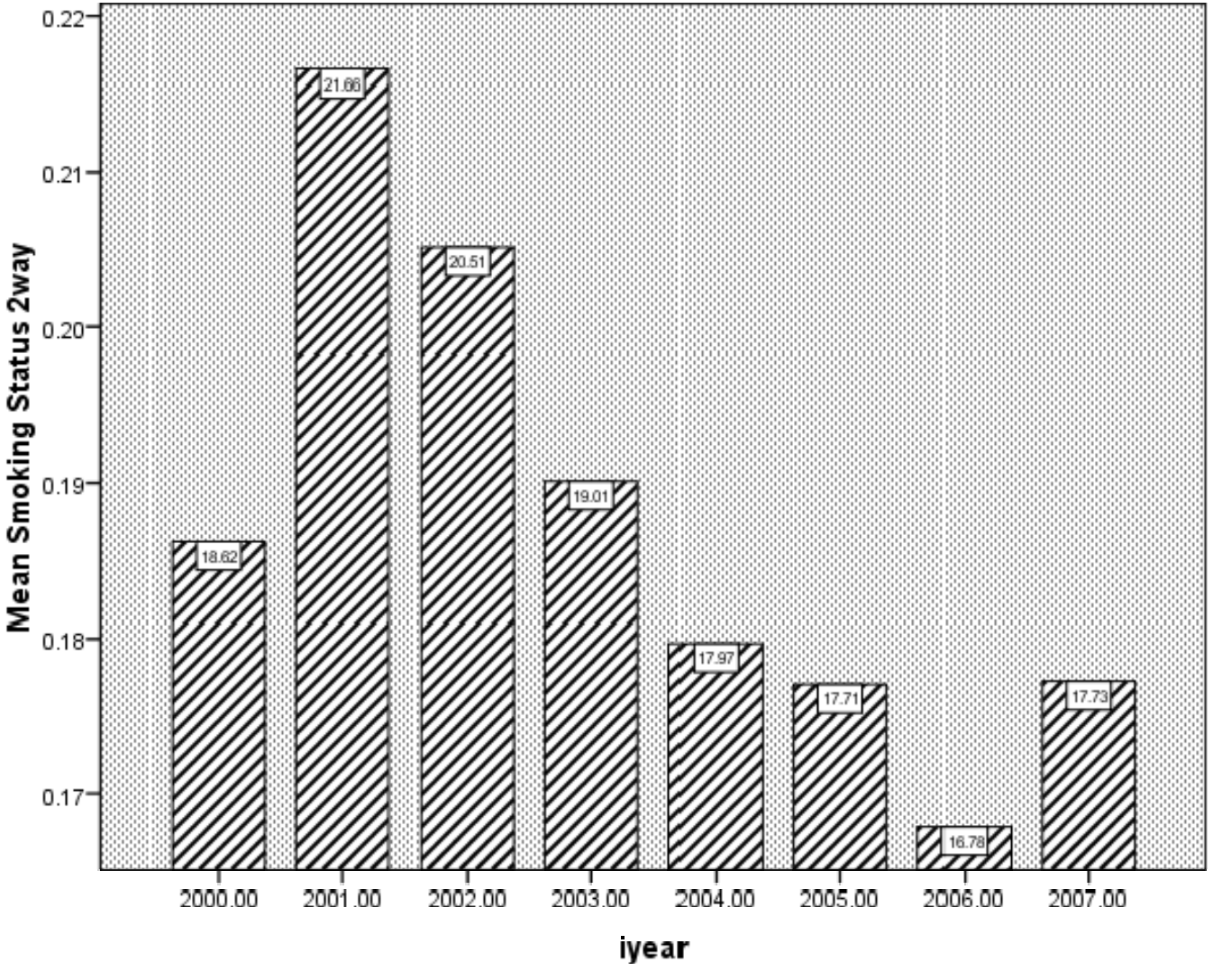
- The infamous bar chart – a prime example of maximizing redundancy: one single number gets multiplied 5 times!



Redundancy & Clutter: Moire effect

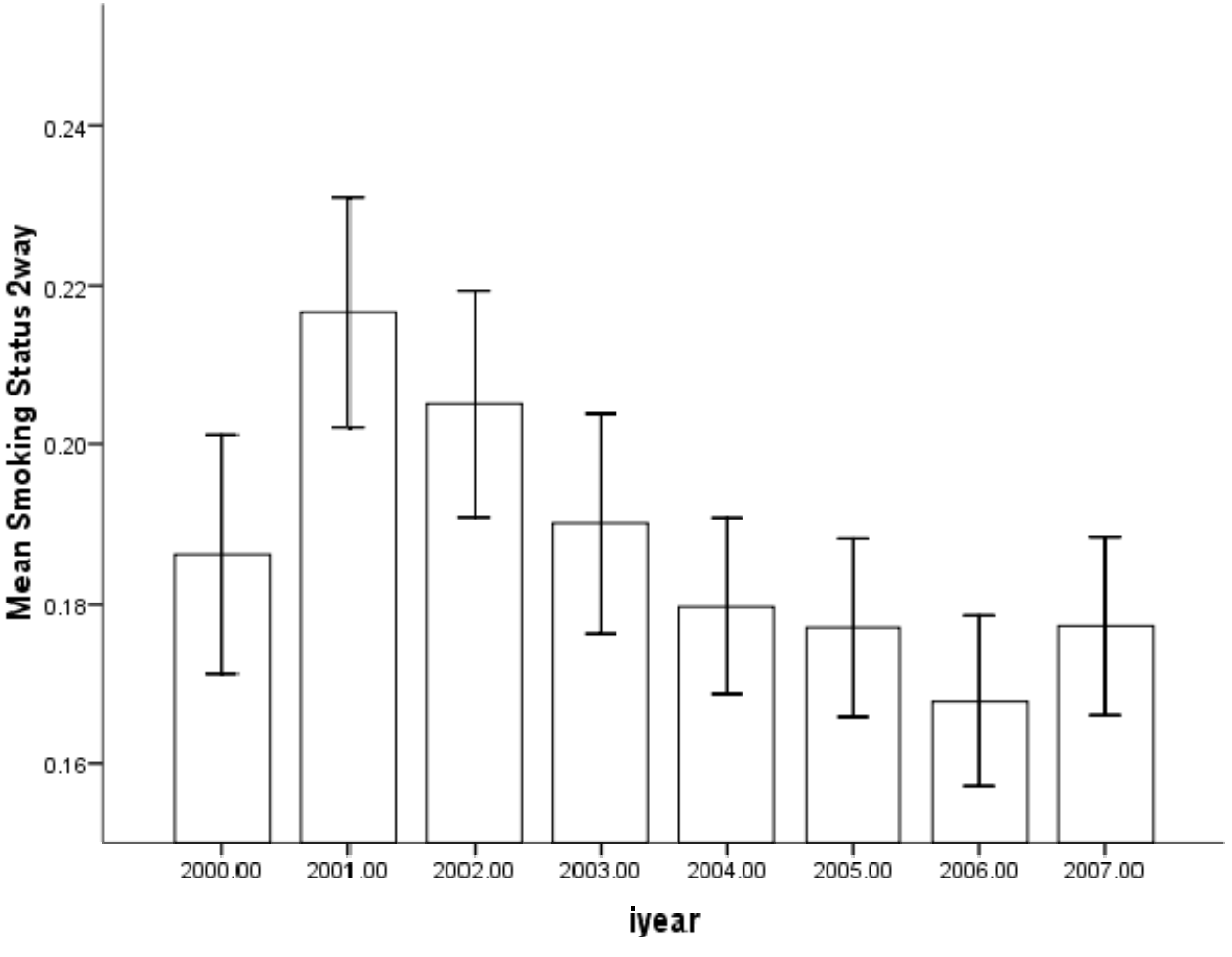


De-cluttering a bar chart: Worst case scenario



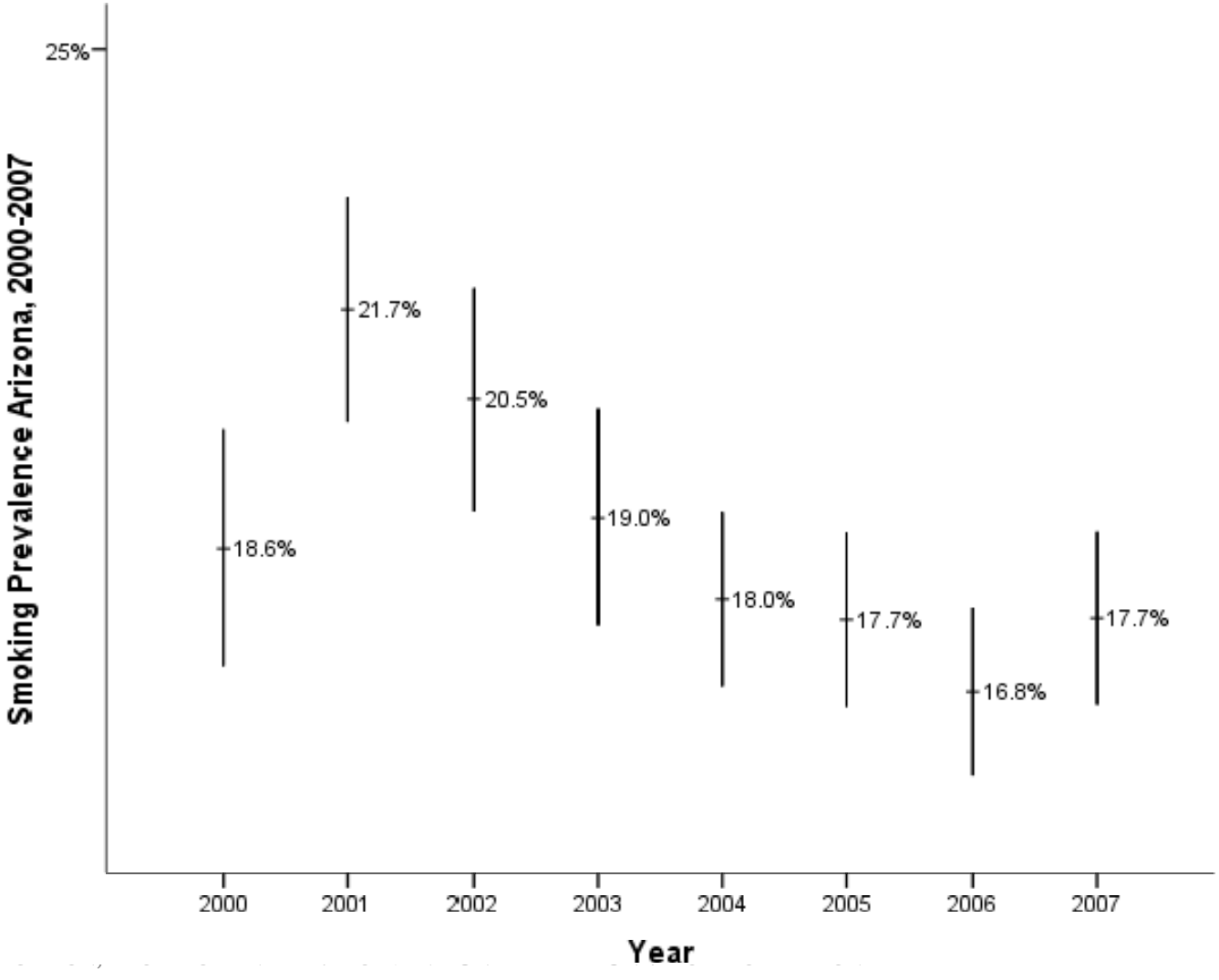
SPSS

Somewhat improved…



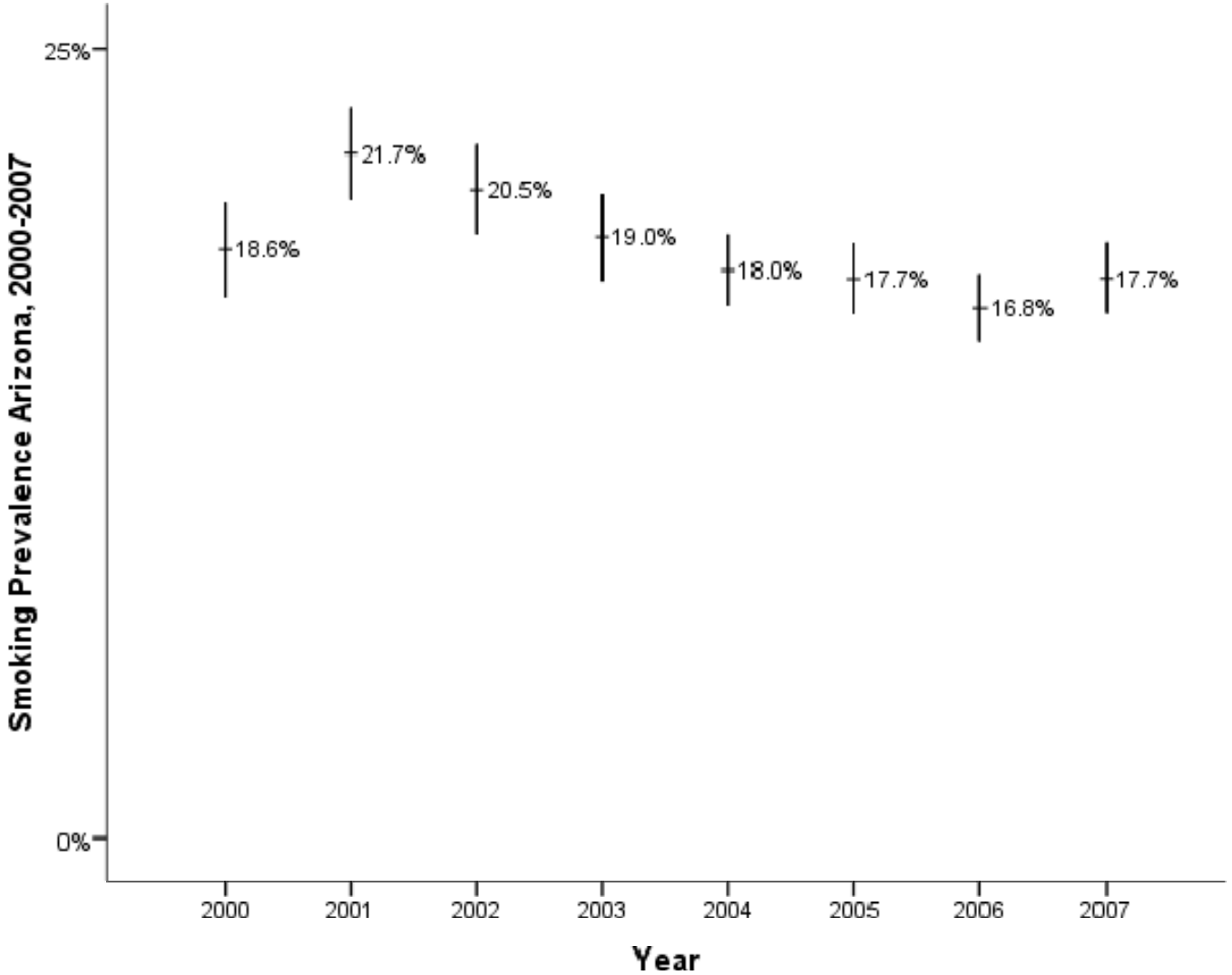
SPSS

Even better…



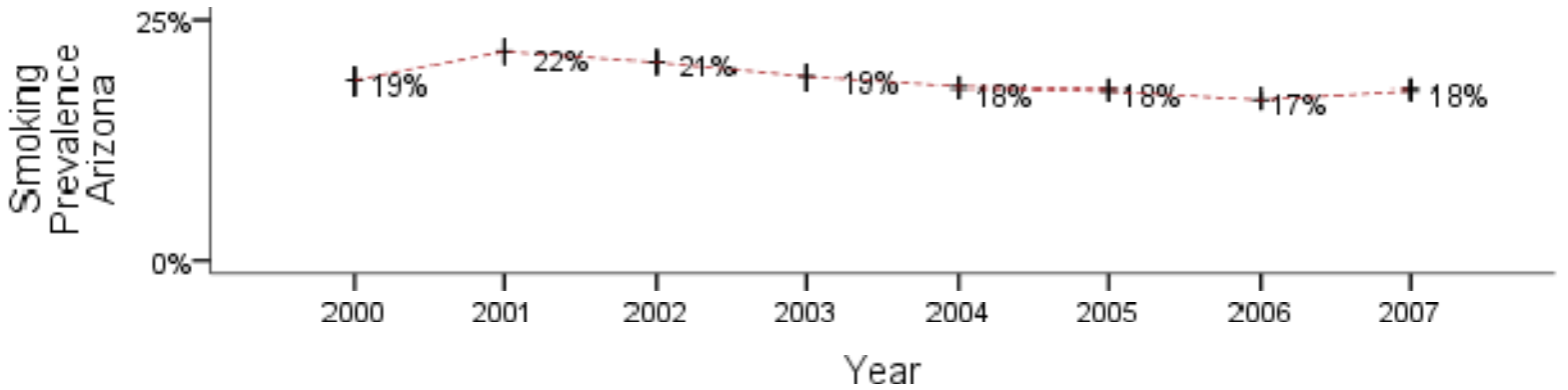
SPSS

Almost there…



SPSS

From junk to excellence in 4 steps



SPSS

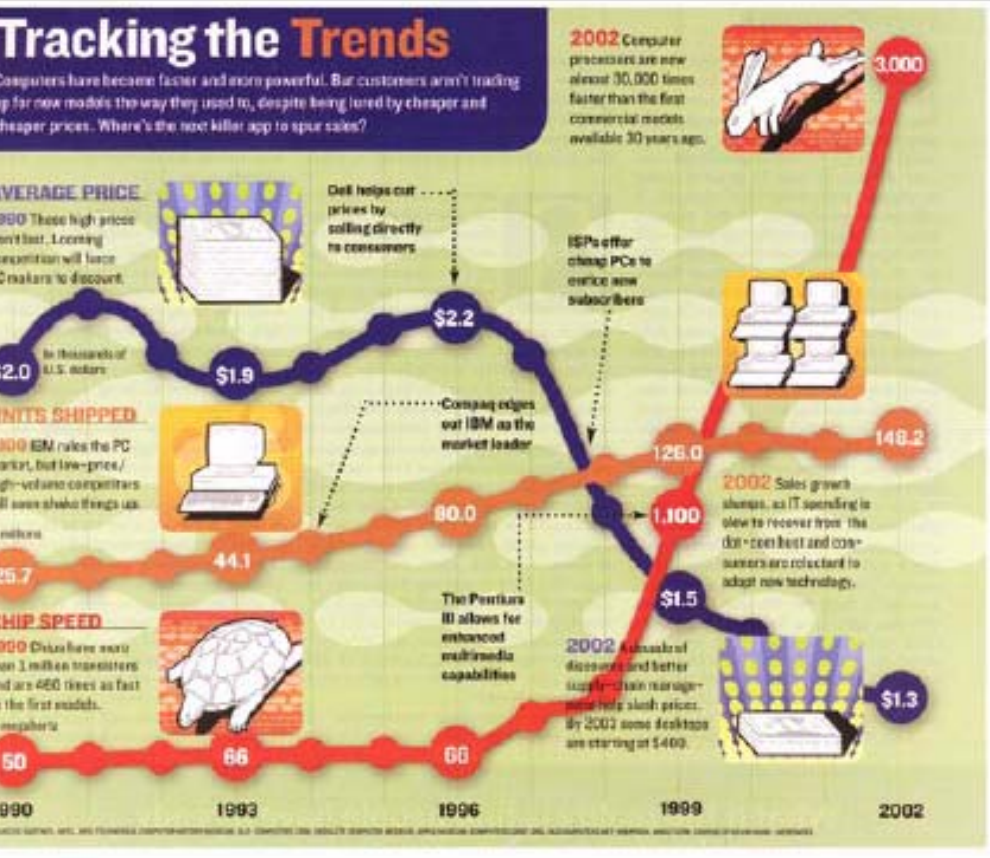
Quality benchmarks

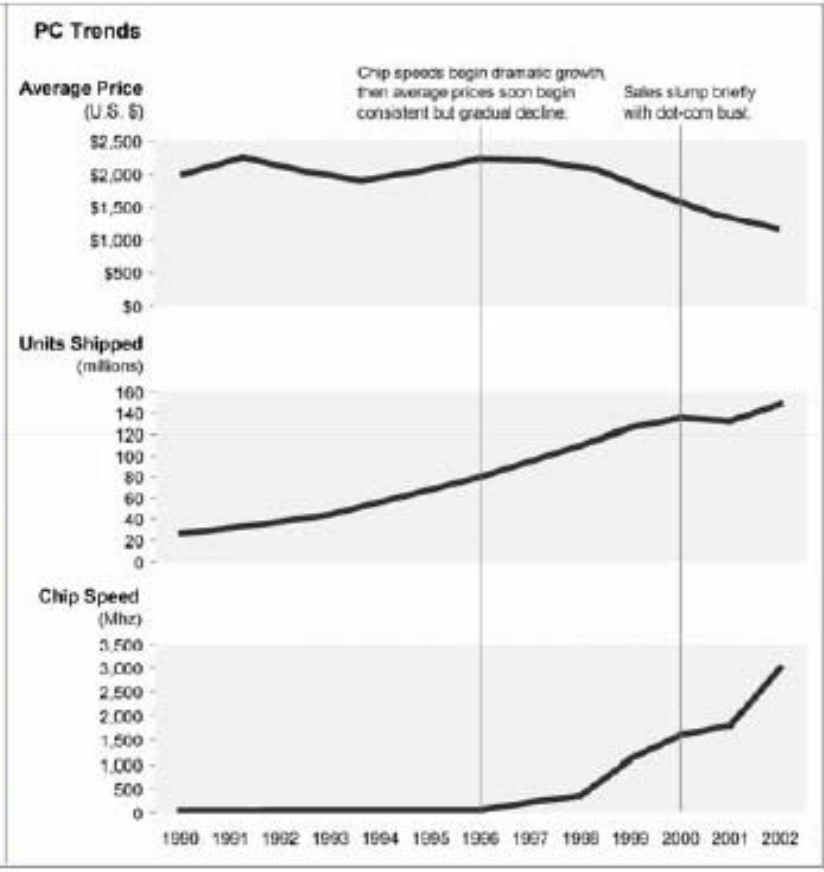
- Self-sufficiency
- Data density
  - vs FF (fanciful frippery, aka „infographics“)
- Clarity
  - Right balance between density and clarity (“loss aversion”)
- Honesty
- Properness (no, not trivial)
- Thoughtfulness

David vs. Goliath

Data graph	Infographic
Express	Impress
Data density	Flashiness
Clarity	Pompousness
Clear message	Visual appeal
Clear quality benchmarks	Anything goes
Limited number of formats	Infinite variety

Stephen Few's example





Graphical integrity

- History of data graphs
  - 1930s to ~1970s: “decoration for dullards”
  - Preconceived as always fraudulent

- Lie Factor = $\frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$

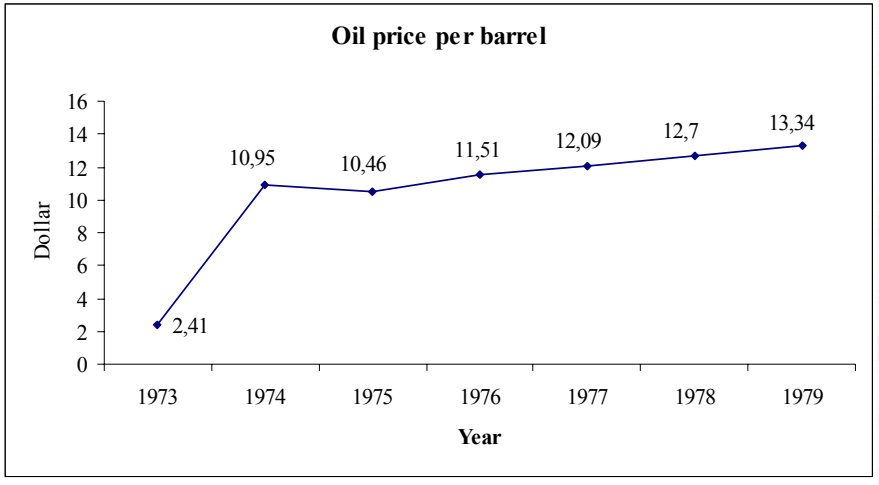
- Design variation ≠ data variation

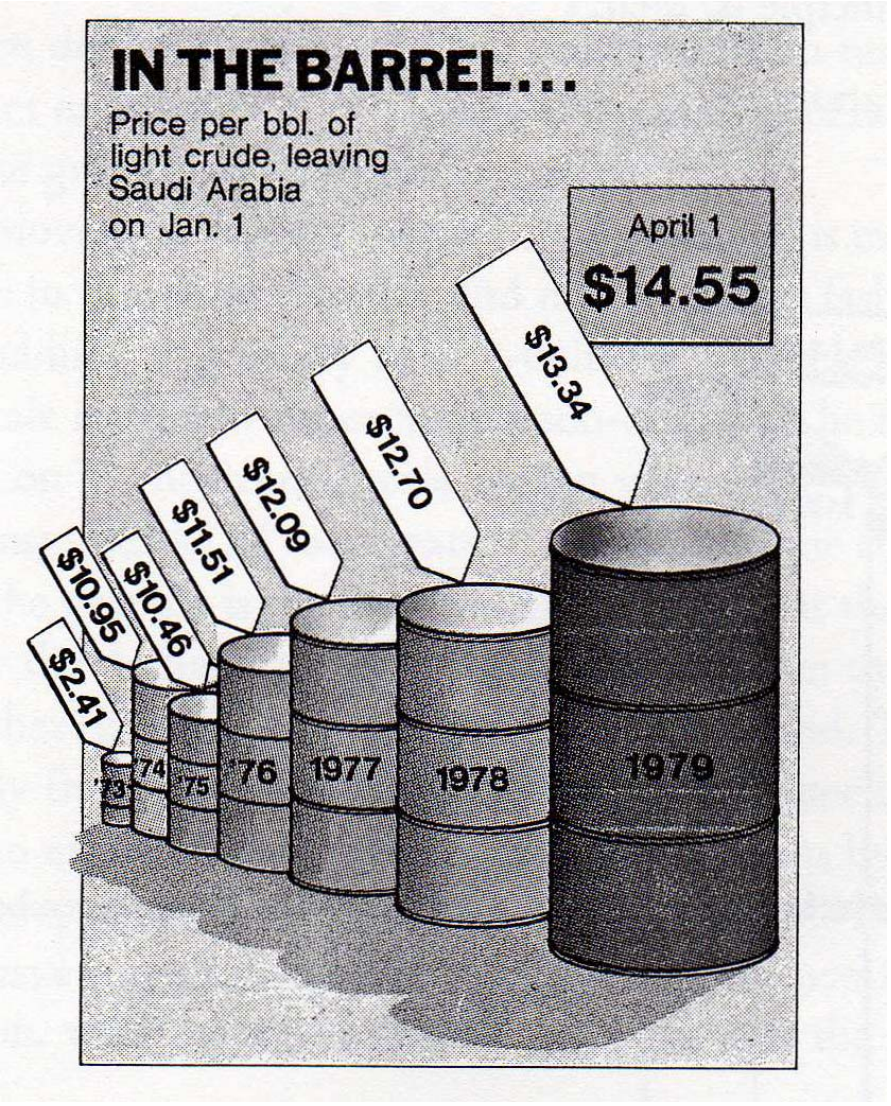
The deceptive barrels

Real increase of dollar value: 454%

Increase in barrel volume: 4280%

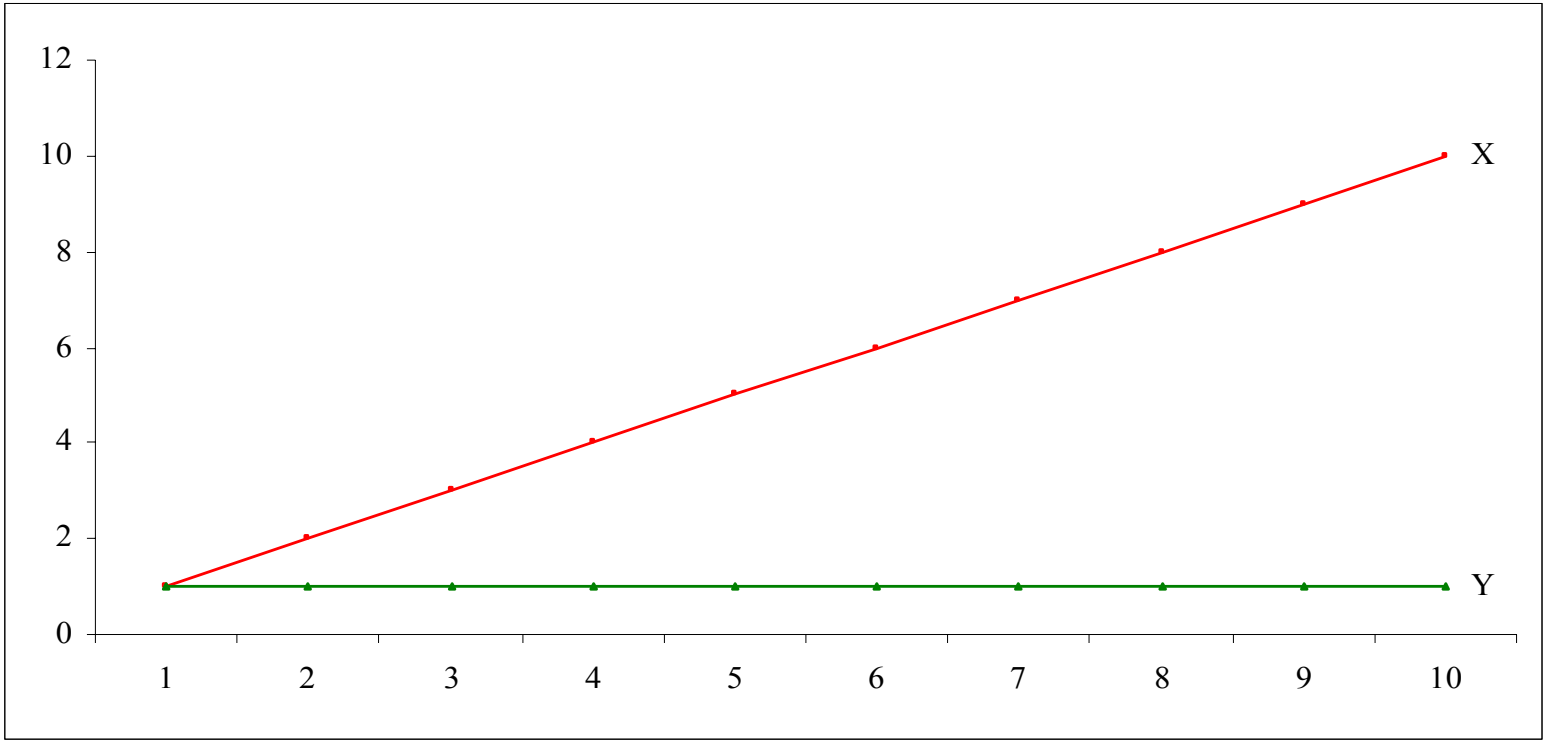
Lie factor = 4280% / 454% = 9.4





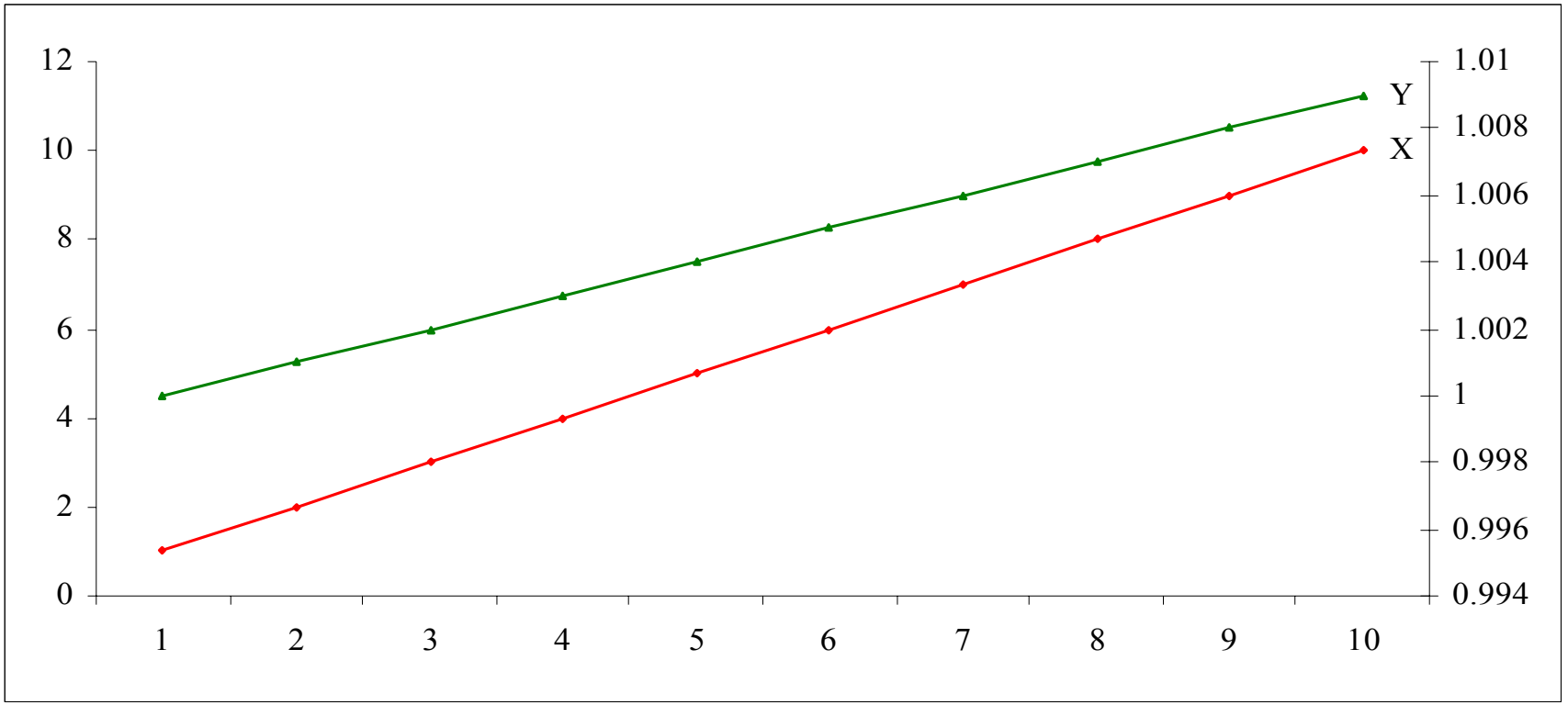
The Secondary Axis of Evil

- Two time series: X & Y
- Your guess: correlation X,Y?



The Secondary Axis of Evil

- Two time series: X & Y
- Your guess: correlation X,Y?



The Secondary Axis of Evil

X	Y
1	1
2	1.001
3	1.002
4	1.003
5	1.004
6	1.005
7	1.006
8	1.007
9	1.008
10	1.009

z-standardized

X	Y
-1.486301	-1.4863
-1.156012	-1.15601
-0.825723	-0.82572
-0.495434	-0.49543
-0.165145	-0.16514
0.165145	0.165145
0.495434	0.495434
0.825723	0.825723
1.156012	1.156012
1.486301	1.486301

$r_{x,y} = 1.0$

Never use secondary axis!

If two variables in time series, use standardized series!

X & Y standardized

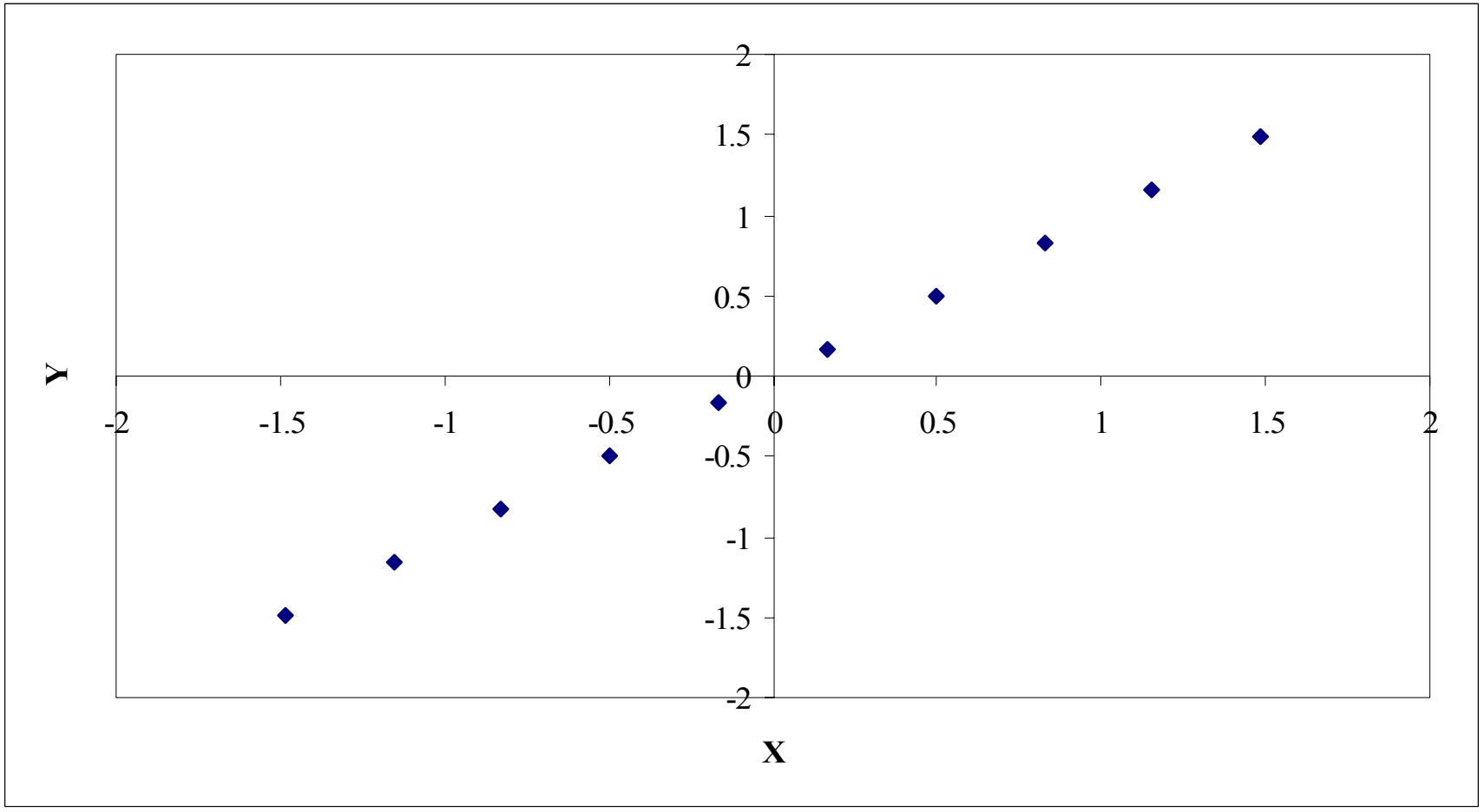
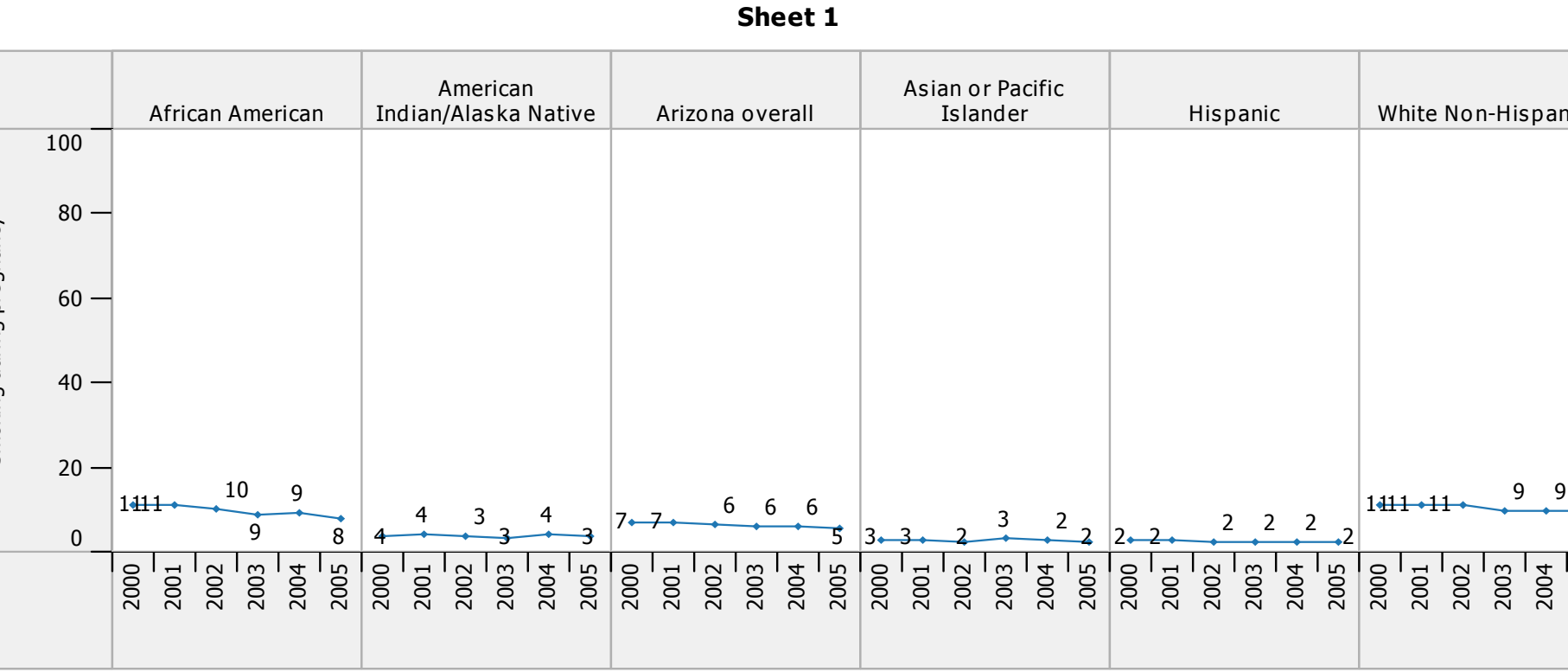


Tableau | Excel

X | **X**

Format of Y-Axis

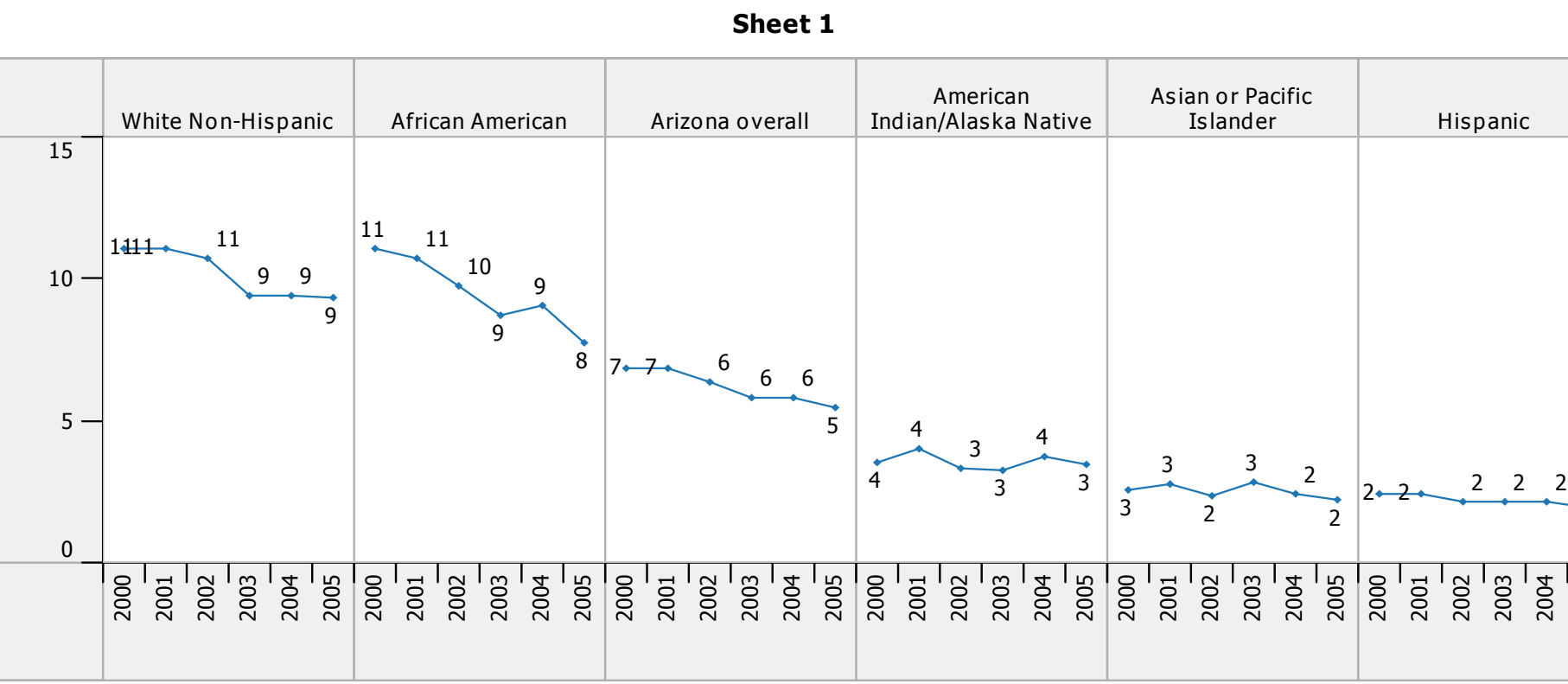


oking during pregnancy for each Year broken down by Race. The marks are labeled by sum of Smoking during pregnancy.

Tableau | Excel

X |

Format of Y-Axis

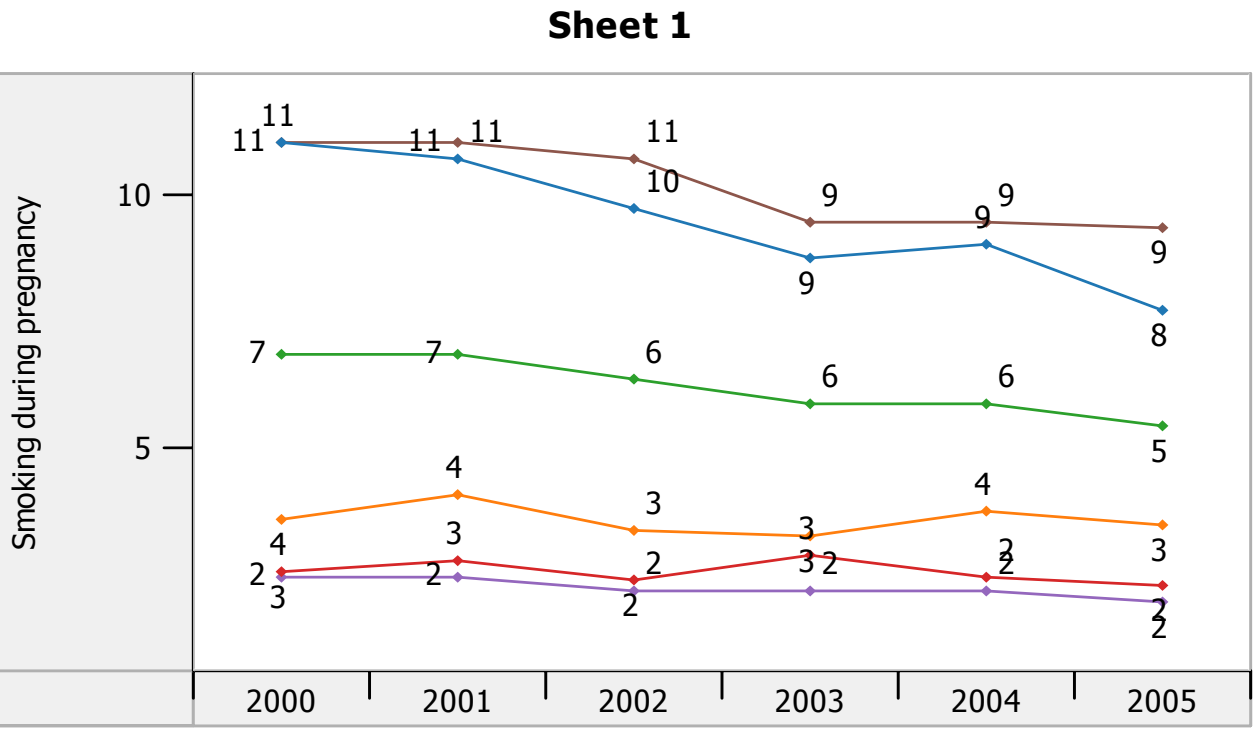


oking during pregnancy for each Year broken down by Race. The marks are labeled by sum of Smoking during pregnancy.

Tableau | Excel

X |

Format of Y-Axis



Smoking during pregnancy for each Year. Color shows details about Race. The marks are labeled by sum of Smoking during pregnancy.

Tableau | Excel

X |

Context is essential!

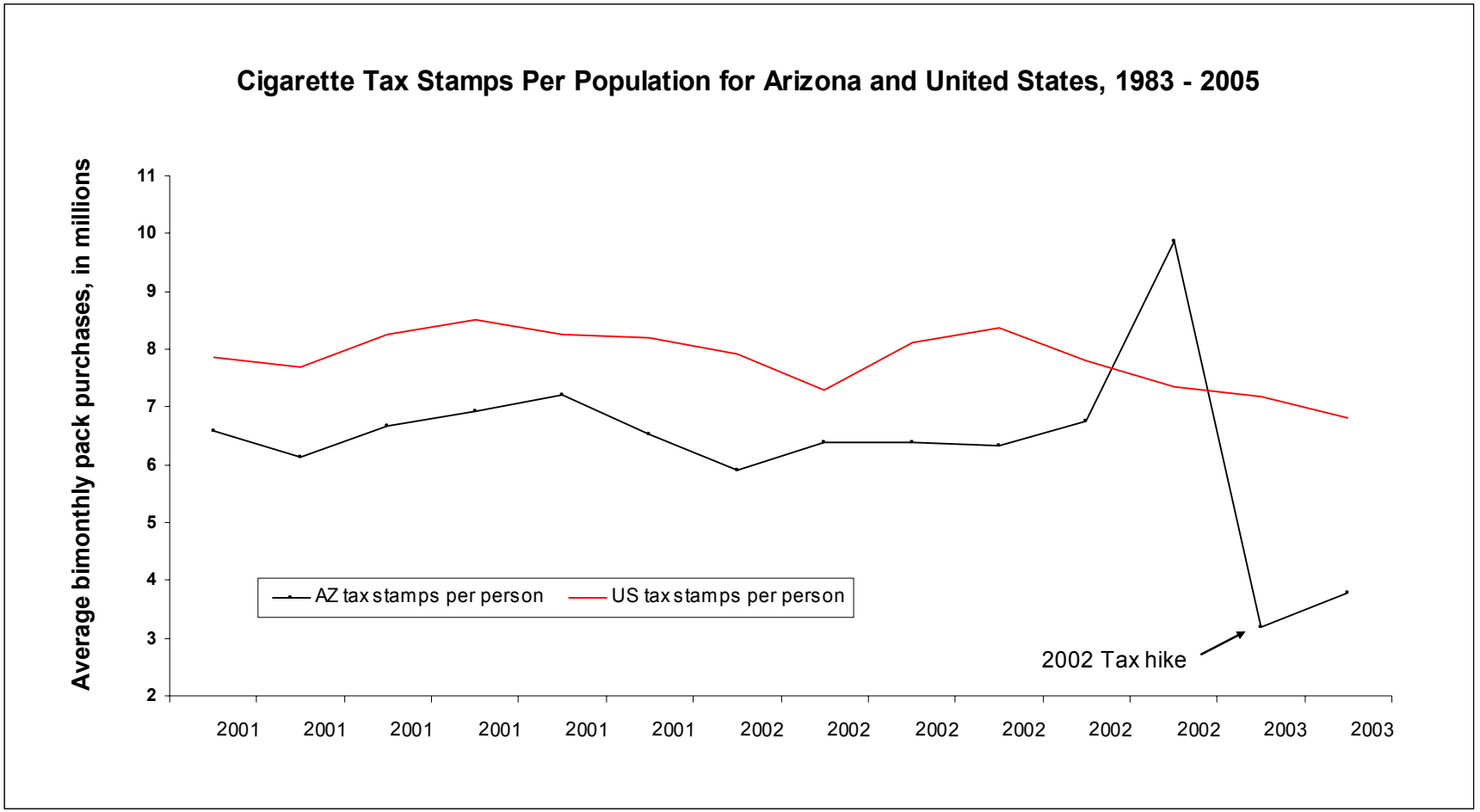


Tableau	Excel
X	X

Don’t present data out of context!

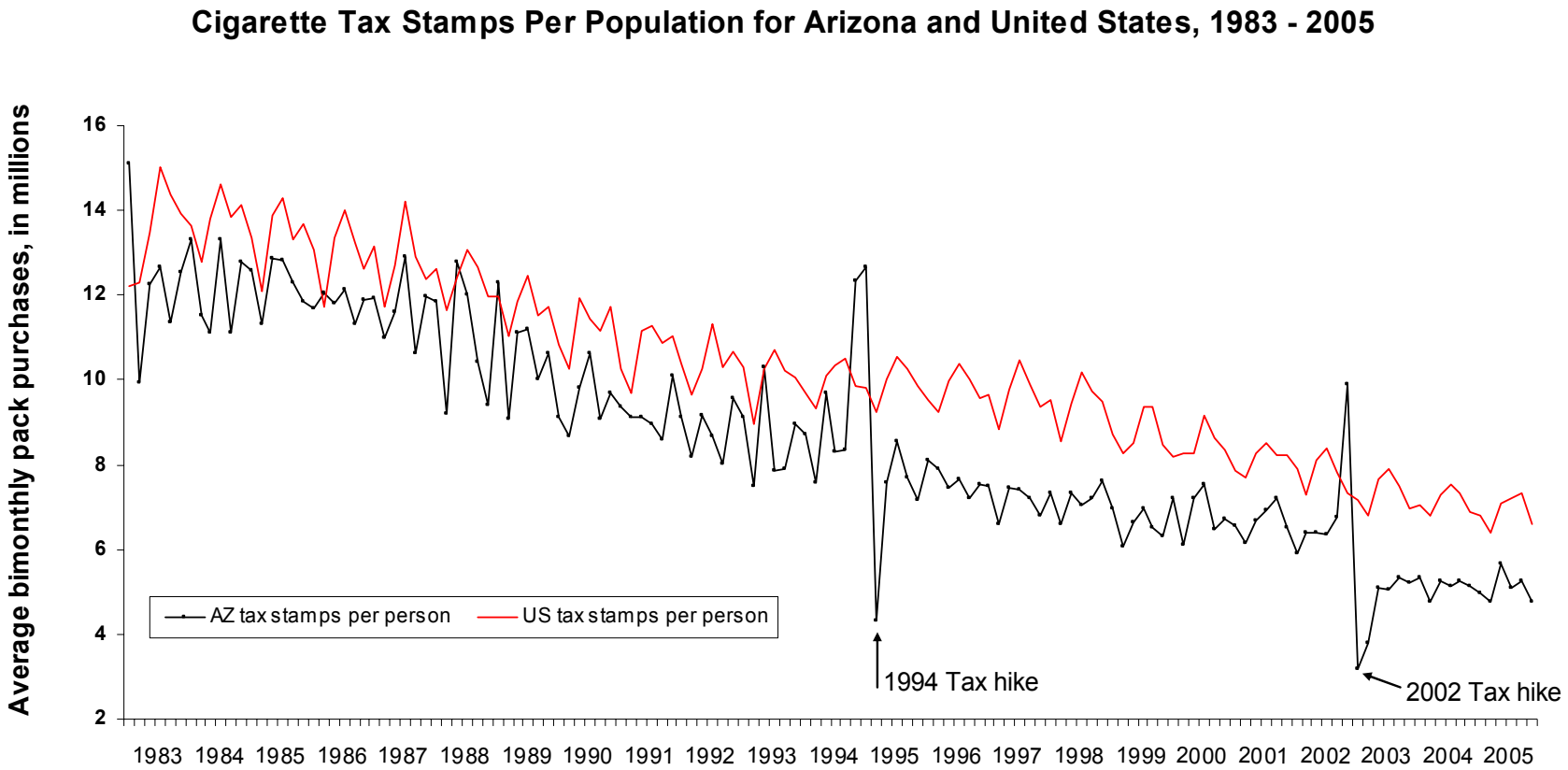


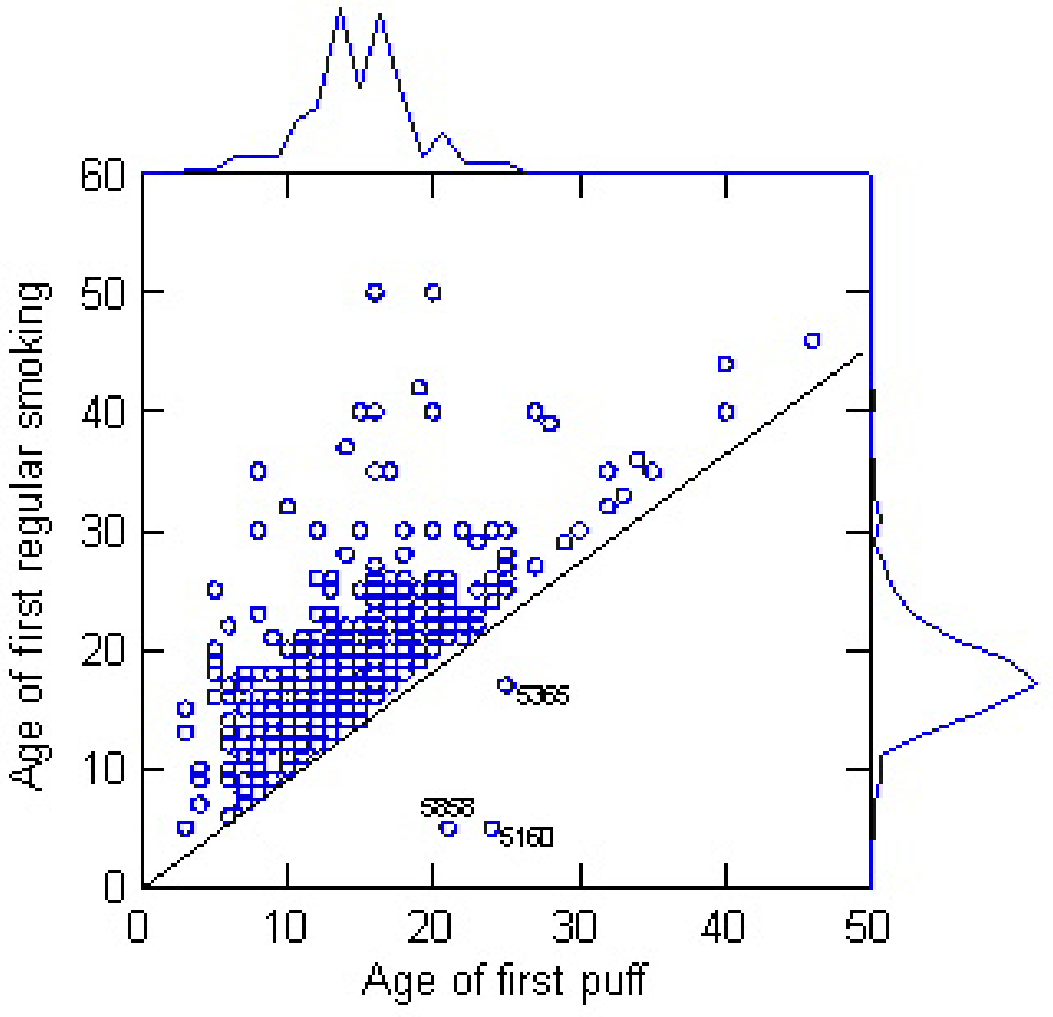
Tableau	Excel
X	X

Graphical integrity

- Clear, detailed, unambiguous labeling on the graph itself
- In time-series display of money, use deflated & standardized units
  - Time series: very sensitive to aggregations
- Spatial data
  - Consider natural frequencies
  - Very sensitive to aggregations

- Graphics must not quote data out of context

Visualization IS data analysis!



SYSTAT

Issues

- Personal preferences
- Breaking (bad) customs & habits
  - Bar charts are ubiquitous, everybody WANTS to see them
  - The notorious pie-chart…
- When tables, when graphs, when both?
  - There are rules for good tables, too.

Resources: Software

- Tableau ($999 „Desktop personal edition“)
- Mondrian (free) http://theusrus.de/Mondrian/index.html
- Panopticon ($$) http://www.panopticon.com/

- SPSS
- SYSTAT
- STATA
- R
- Maps: ArcGIS

Resources: Online

- http://junkcharts.typepad.com/
- http://graphs.gapminder.org (TED talk Hans Rosling)
- http://flowingdata.com/
- http://infosthetics.com/
- http://services.alphaworks.ibm.com/manyeyes/app
- http://statisticalgraphics.blog.com/
- http://www.informationisbeautiful.net/

- http://www.math.yorku.ca/SCS/Gallery/
- http://www.perceptualedge.com/library.php (Stephen Few)

Resources: Articles

- Wainer, H. 1984. How to Display Data Badly. Am Stat
- Wainer, H. 1992. Understanding Graphs and Tables. Educ Researcher
- Reese, 2008. Scatterplot revisited. Significance

Resources: Books

- Everything by Edward Tufte
- Everything by William S. Cleveland
- Everything by Herbert Wainer
- Gelman, A. (2008) Red State Blue State