Applied Mechanism Design and Big Data

→ sub-topic: Statistical Data Analysis

Max Baak

(with many thanks to Wouter Verkerke)

Roadmap for this course

1. Statistics basics
   - Probability theory
   - Probability distributions

2. Parameter estimation

3. Pitfalls in (big) data analysis
   - Spurious correlations
   - Data-quality assessment

4. Hypothesis Testing

Correlation & covariance in >2 variables

- Concept of covariance, correlation is easily extended to arbitrary number of variables

$$\mathrm{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)} x_{(j)}} - \bar{x}_{(i)} \bar{x}_{(j)}$$

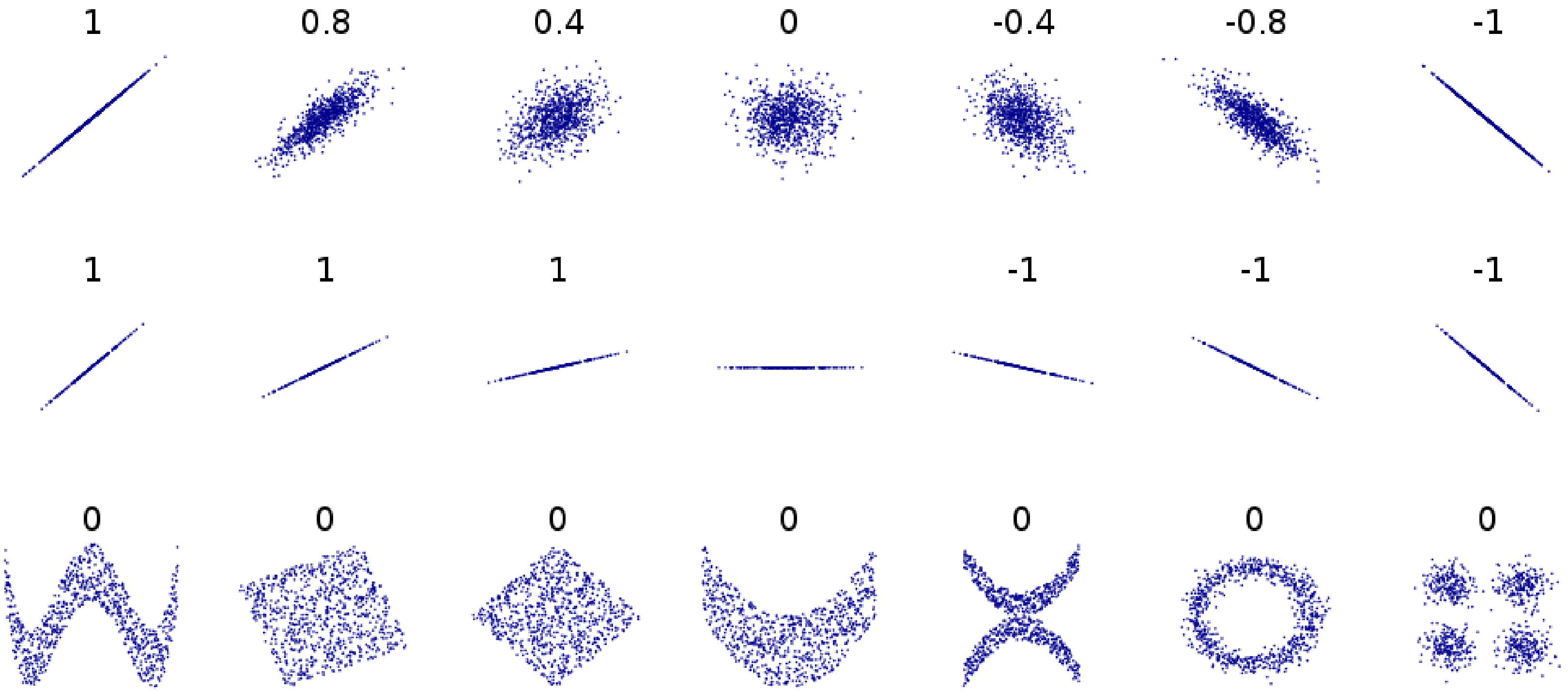
- so that $V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)})$ takes the form of a *n x n symmetric matrix*

- This is called the ***covariance matrix***, or ***error matrix***
- Similarly the correlation matrix becomes

$$\rho_{ij} = \frac{\mathrm{cov}(x_{(i)}, x_{(j)})}{\sigma_{(i)} \sigma_{(j)}} \longrightarrow V_{ij} = \rho_{ij} \sigma_i \sigma_j$$

Linear vs non-linear correlations

- Correlation coefficients used here are (linear) Pearson product-moment correlation coefficients
- Data can have more subtle (non-linear) correlations that contained in these coefficients



- Always check correlation by eye!

Wouter Verkerke, UCSB

Trivia Quiz!

Trivia quiz! (1/2)

- Which of the following coin-flip series is random, and which one made-up?

a. 1001101011101011101111010101010111110101

b. 0010110110011101011011010110001101001110

Trivia quiz! (2/2)

- What percentage of the course material have you understood so far?

a. More than 70%

b. Less than 70%

- Write down the number you think is right

Overconfidence bias!

- (From KPMG Audit test)

Group Tested	Information Type	(% Misses) Target/Observed	
Harvard MBAs	Trivia facts	2%	46%
Computer Co. Managers	General business Company-specific	5% 5%	80% 58%
Physicians	Probability of pneumonia	0-20%	82%

- E.g. when physicians were asked to assess the likelihood of pneumonia, they were highly confident that they would only be wrong between 0-20% of the time
- Instead, they were wrong more than 80% of the time

Analysis Pitfalls

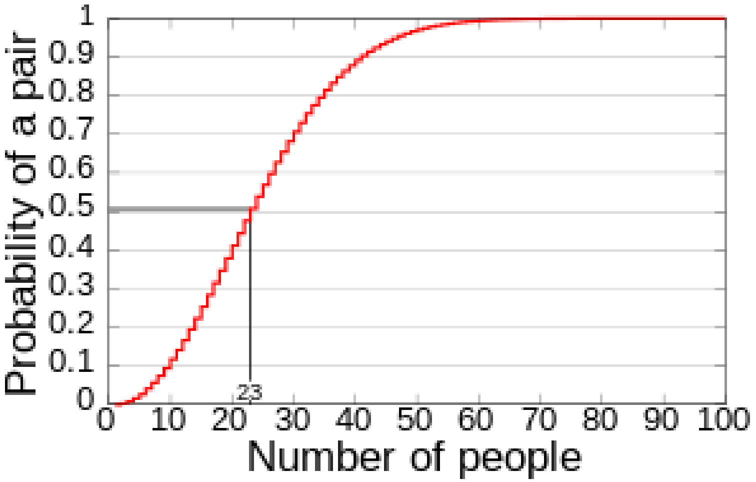
Just to show…

KPN NL
22:12
39%
True!
"If you torture data long enough, it will confess to anything you'd like."
R.H. Coase, British economist
News Feed
Messenger
Notifications
More

Look Elsewhere Effect – The Birthday Problem

- Humans have tendency to highlight unlikely events, but to ignore all likely events.
- The more experiments performed (on the same dataset), the less significant the result(s) found.
  - Bonferroni correction (conservative): multiply the observed *p*-value by the number of tests performed.

“the birthday problem” as illustration of the “look elsewhere effect.”



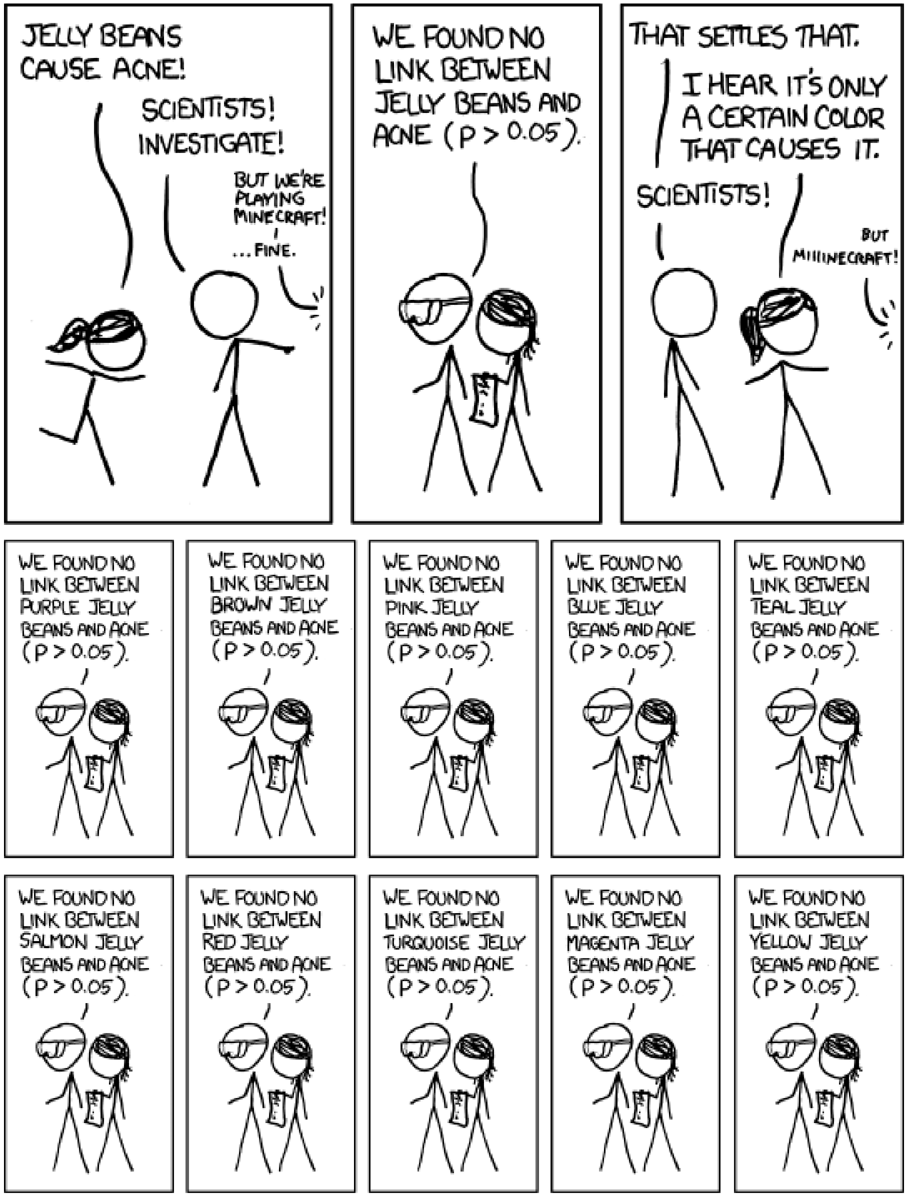
With 23 people the chance of a matching birthday is already 50%.

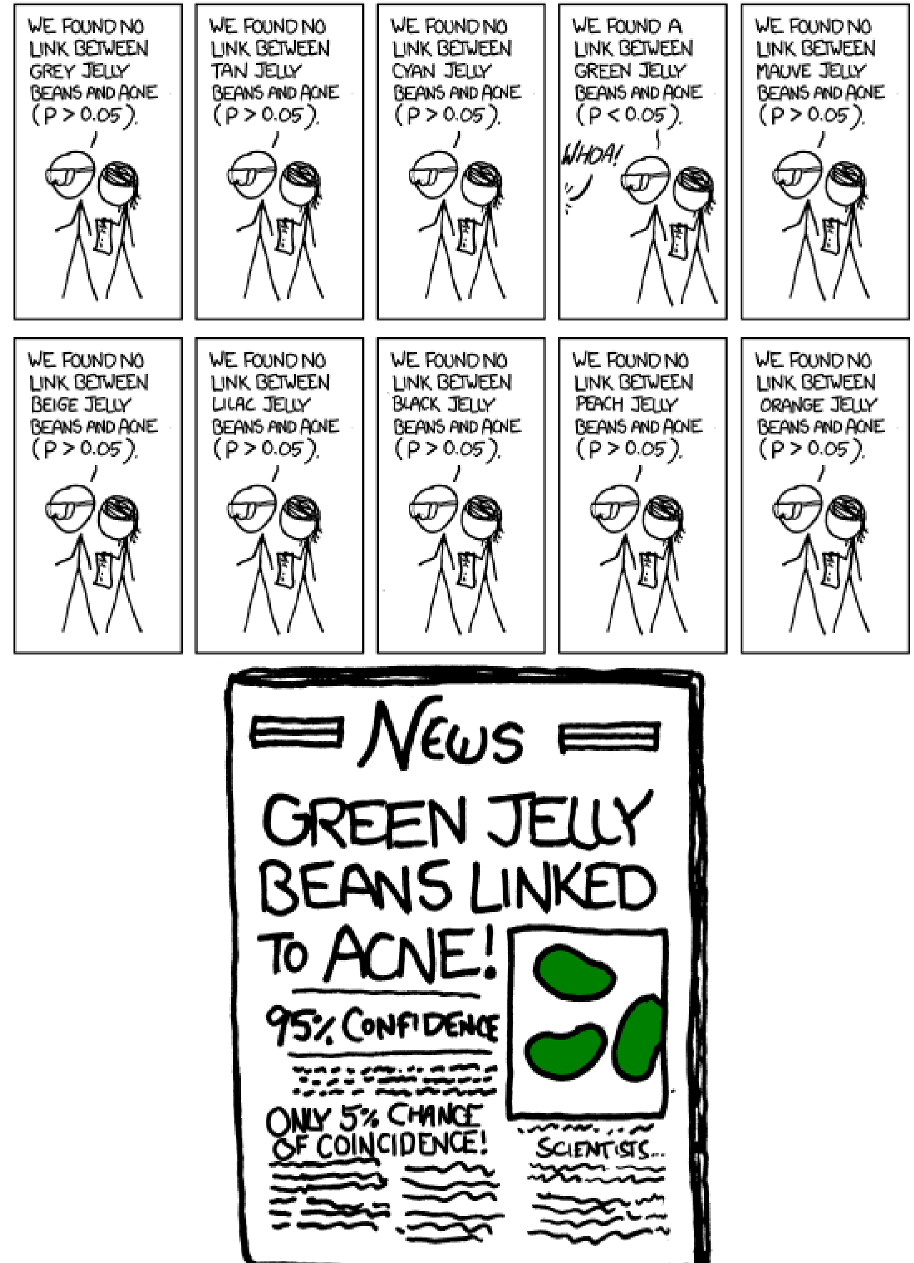
Example Look Elsewhere Effect

- A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects.
- The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 ailments.
- The study found that the incidence of childhood leukaemia was four times higher among those that lived closest to the power lines.
- → It spurred calls to action by the Swedish government.

(Wikipedia, LEE)

aka: Law of very large numbers
“With a sample size large enough,
any outrageous thing is likely to happen.”

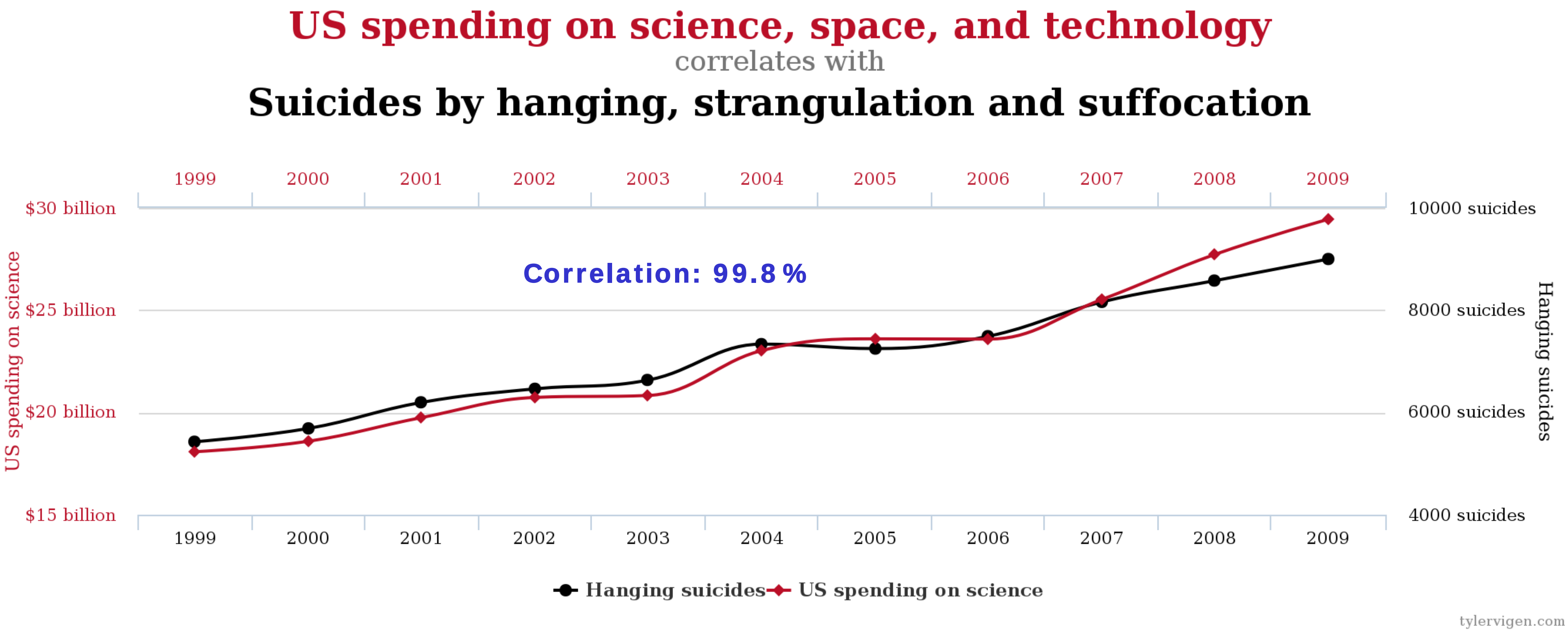




https://xkcd.com/882/

Curious correlation?

http://www.tylervigen.com/spurious-correlations



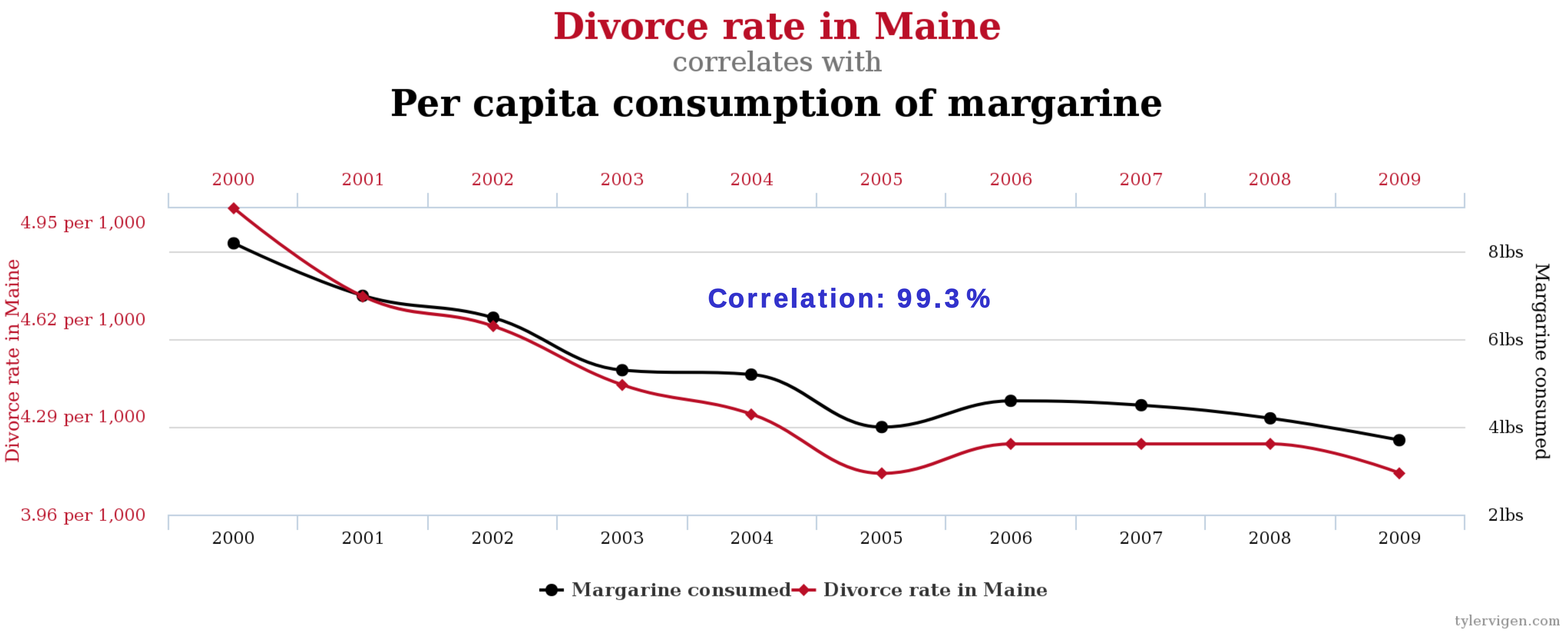
Correlation and causality

- Correlation does not necessarily imply causality
- Spurious correlations are caused by dependence on "hidden" confounding variables

Example: Increased rates of drownings in a city's swimming pools in summers with high ice cream sales do not necessarily imply that eating ice cream causes drowning. There might be a common, (not so) confounding dependence on the weather.

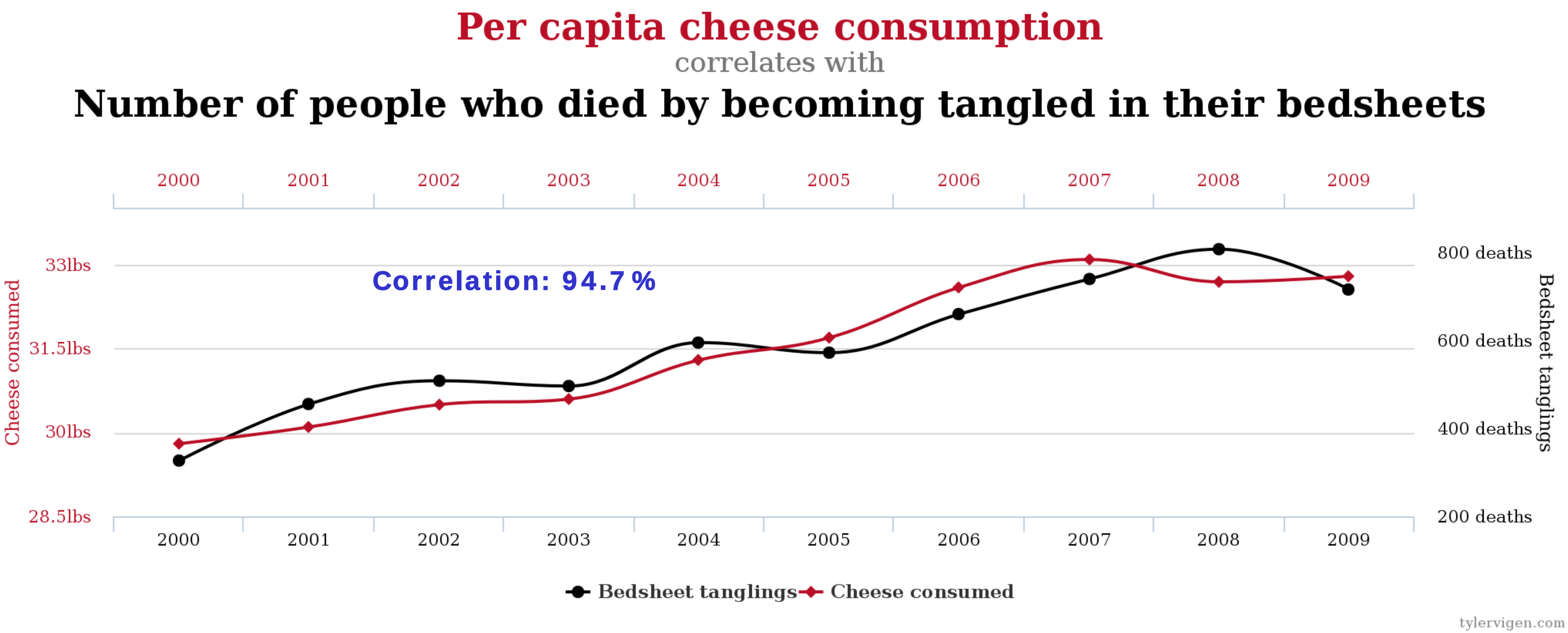
Spurious correlations - example 2

http://www.tylervigen.com/spurious-correlations



Spurious correlations - example 3

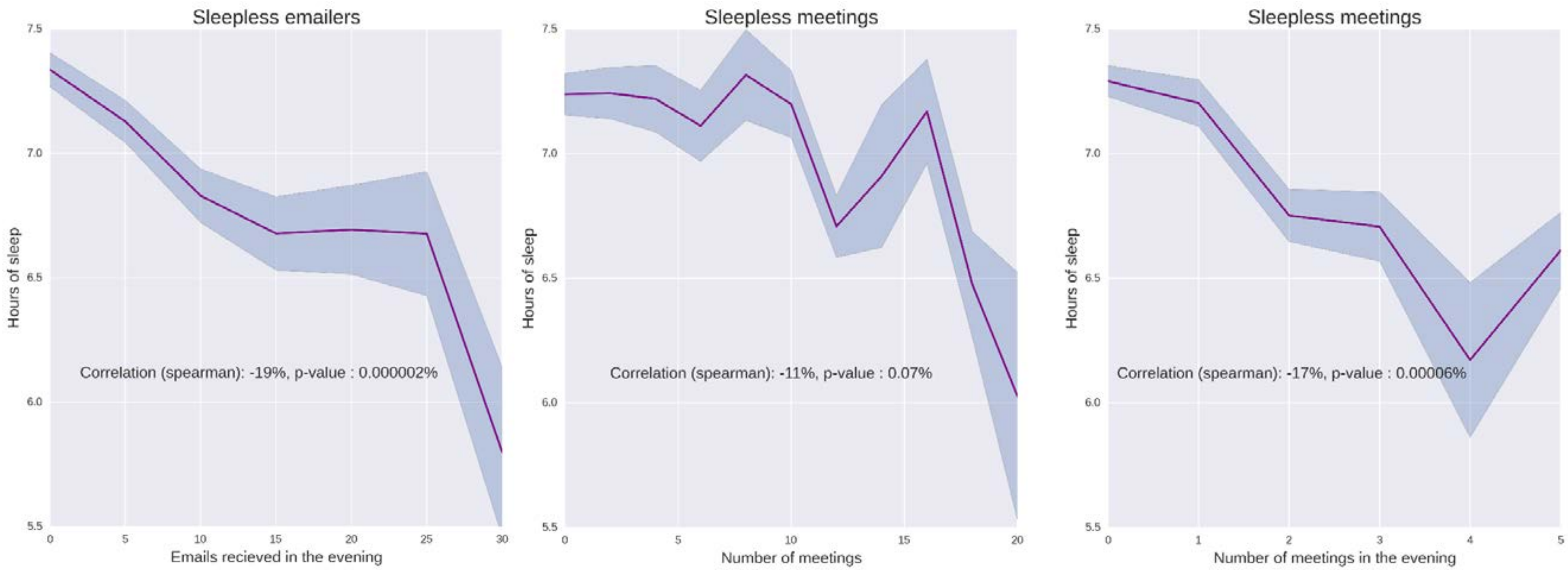
http://www.tylervigen.com/spurious-correlations



Spurious correlations - example 4

From internal KPMG study on
correlation between fitness and health

Hours of sleep



Emails in evening

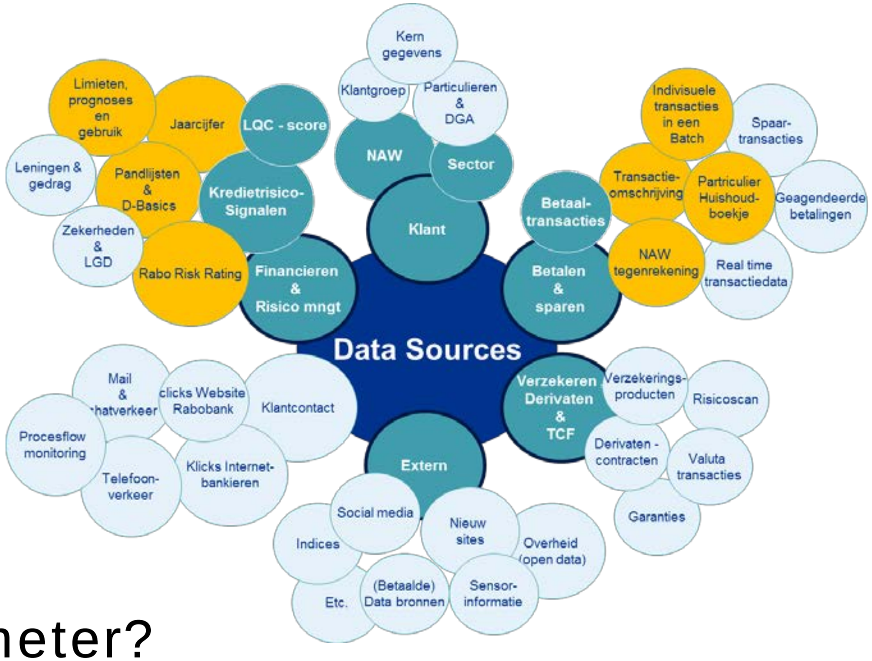
Meetings

Meetings in evening

Spurious correlations

- Typical client question:
  "We have multiple (large) datasets [with many variables]
  Please extract any relevant insights [...]"

- With N variables, number of (spurious) correlations proportional to N^2

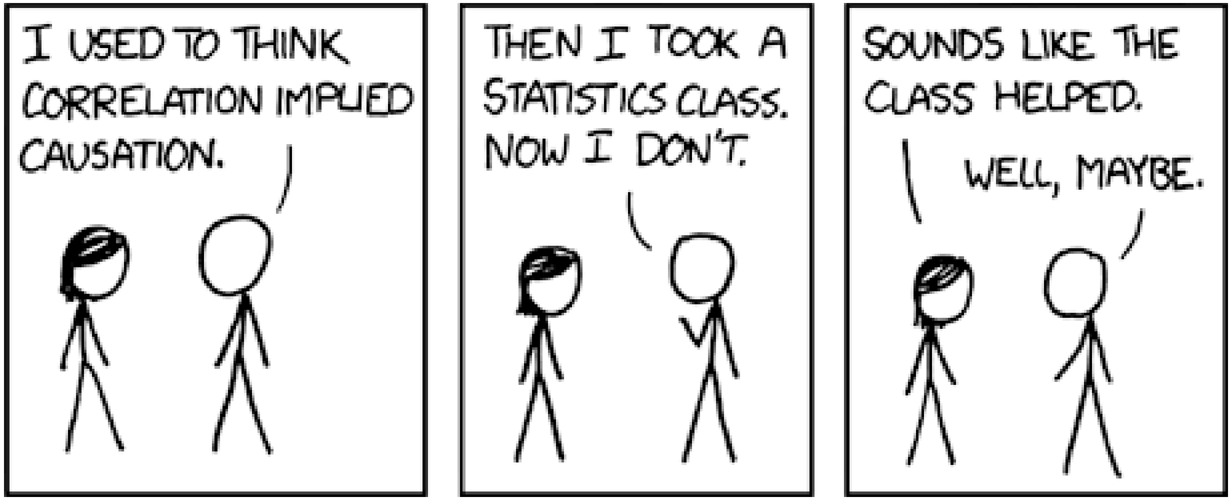


Debunking spurious correlations:

- Causal relationship?
- Correlation through other parameter?
- Awareness! Use your brains!

Correlation and causality

- https://xkcd.com/552/

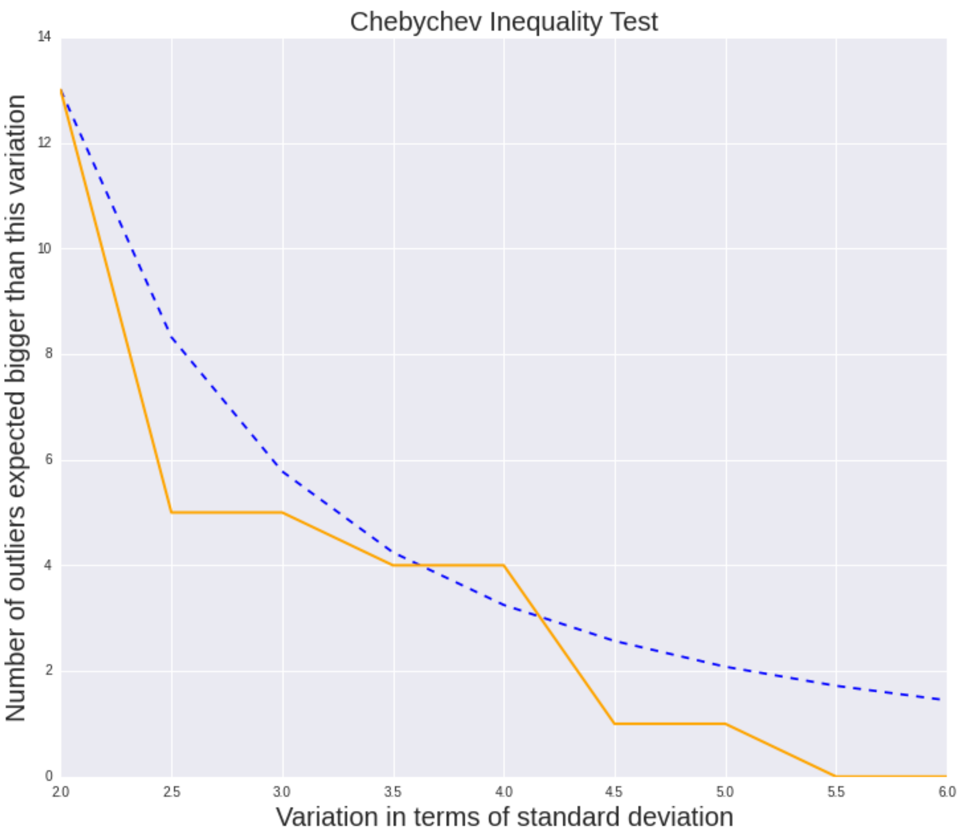


Chebyshev Inequality Test

- For *arbitrary* continuous distributions in x the Chebyshev inequality theorem holds:

$$P(|x - \mu| > k\sigma) \leq \frac{1}{k^2}$$

- → Gives upper bound on size & significance of random fluctuations
- Clearly more strict bounds apply for Gaussian distribution



blue curve = Chebyshev prediction:

If the orange line is on top (or below) of the blue curve, the findings are not inconsistent with random fluctuations

Simpson's paradox

- Other (extreme) example of dependence on confounding variables: Simpson's paradox
- Conclusion of a study is reversed when confounding variables are taken into account

Example: study of success rate in removing kidney stones

Two methods were compared:
- Open surgery: 78% of treatments successful
- Small puncture: 83% of treatments successful

Conclusion: "Small puncture method" more successful. Or is it? →

BMJ 1994;309:1480 - http://dx.doi.org/10.1136/bmj.309.6967.1480
BMJ 1986;292:879-82 - http://dx.doi.org/10.1136/bmj.292.6524.879

BMJ 1994;309:1480 - http://dx.doi.org/10.1136/bmj.309.6967.1480
BMJ 1986;292:879-82 - http://dx.doi.org/10.1136/bmj.292.6524.879
https://en.wikipedia.org/wiki/Simpson%27s_paradox

Simpson's paradox

Example: study of success rate in removing kidney stones

Split success rates into cases with small stones and cases with large stones:

	Open surgery	Small puncture
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Overall	78% (273/350)	**83% (289/350)**

lower rates

preference in treatment: dominates overall rates

- Surgery more often successful for both small and large stones
- Treatment large stones less often successful for both surgery and puncture
- Large stones usually treated by surgery, small stones by puncture
- As a result, surgery less often successful overall

Note:

- Success probabilities in table are also conditional on the initial choice of treatment. I.e., success rate is 87%, given small stones, choice for puncture, and treatment by puncture. For (hypothetical) cases treated by puncture, where surgery would have been preferred choice, the success rate is likely to be smaller.
- Statistical significance of differences in surgery and puncture rates (binomial): small stones: 1.9 (σ_Δ=3%); large stones: 0.7 (σ_Δ=6%); overall: 1.5 (σ_Δ=3%)

$$\sigma_r = \sqrt{\frac{r(1-r)}{N}}$$

Data Quality

What data did we get?

Where it goes wrong...

1999 - Mars Climate Orbiter:
Crash caused by incorrect data from thruster software

Nov. 10, 1999: Metric Math Mistake Muffed Mars Meteorology Mission

LISA GROSSMAN 11.10.10 7:00 AM

NOV. 10, 1999: METRIC MATH MISTAKE MUFFED MARS METEOROLOGY MISSION

hat NASA's
n atmosphere
n English to

Mystery of Orbiter Crash Solved

By Kathy Sawyer
Washington Post Staff Writer
Friday, October 1, 1999; Page A1

NASA's Mars Climate Orbiter was lost in space last week because engineers failed to make a simple conversion from English units to metric, an embarrassing lapse that sent the $125 million craft fatally close to the Martian surface, investigators said yesterday.

Scie
Mar

BBC ONLINE NETWORK HOMEPAGE | SITEMAP | SCHEDULES | BBC INFORMATION | BBC EDUCATION | BBC WORLD SERVICE

BBC NEWS

News in Audio | News in Video | Newyddion | Новости | Noticias | أخبار | 国际新闻 | 粤語廣播

Front Page
World
UK
UK Politics
Business
Sci/Tech
Health
Education
Sport
Entertainment
Talking Point
In Depth
On Air
Archive

Feedback
Low Graphics

Thursday, September 30, 1999 Published at 18:53 GMT 19:53 UK

Sci/Tech

Confusion leads to Mars failure

The Mars Climate Orbiter: Now in pieces on the planet's surface

The Mars Climate Orbiter Spacecraft was lost because one Nasa team used imperial units while another used metric units for a key spacecraft operation.

Sci/Tech Contents

Relevant Stories

24 Sep 99 | Sci/Tech
Scientist fights Mars setback

23 Sep 99 | Sci/Tech
Mars probe feared destroyed

23 Sep 99 | Sci/Tech
What the loss of Mars Climate Orbiter means

17 Jul 99 | Sci/Tech
Astronauts call for Mars mission

Internet Links

Mars Climate Orbiter

The BBC is not responsible for

Unit of thruster impulse in data was pound-seconds, where it should have been Newton-seconds according to specifications

Where it goes wrong...

Review of British hospital data:
Pregnant men?!

thebmj Research Education News & Views Campaigns Archive

Letters
Hospital episode statistics

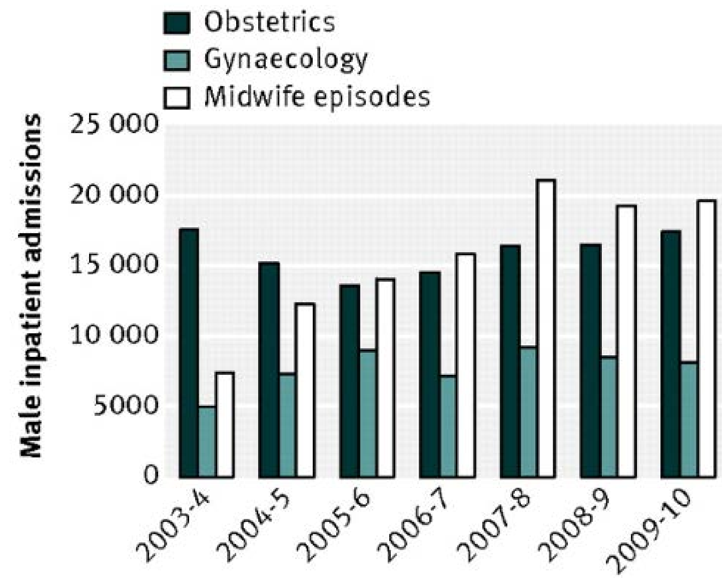
The importance of knowing context of hospital episode statistics when reconfiguring the NHS

BMJ 2012 ; 344 doi: http://dx.doi.org/10.1136/bmj.e2432 (Published 04 April 2012)
Cite this as: *BMJ* 2012;344:e2432

Article | Related content | Metrics | Responses

Lauren Brennan, specialty doctor and honorary clinical research fellow[1], *Mando Watson, consultant paedia*
Robert Klaber, consultant paediatrician[1], *Tagore Charles, consultant paediatrician*[1]

BMJ 2012;344:e2432



Britain's 17,000 Pregnant Men Aren't Really Pregnant

Sam Ro
Apr. 8, 2012, 1:47 AM 2,234 1

Call it the mother of all medical coding errors.

Sarah Kliff of Ezra Klein's WonkBlog recently wrote about an interesting nugget that appeared in a letter published in the British Medical Journal.

> Between 2009 and 2010, thousands of British men turned up at hospitals to be treated for many pregnancy-related services, things like obstetric exams and midwife services. All told, there were 17,000 of them.

"Junior" YouTube

Hospital data of men visiting gynaecologists and midwifes were probably entered with incorrect admission codes

Data quality (DQ)

Are the data in a dataset what one expects them to be?

- Missing files or records
  missing few hours in a daily log file (less activity that day or missing data?)
- Missing / inconsistent / incorrect data fields
  birth date in 2984, “John Dough” instead of “John Doe”, “7” instead of “007”,
  record for a car with a length of 1m
- Inconsistent / duplicate records
  mutations bank account don’t add up to total change in balance,
  daily log files with overlap between 23:00 and 01:00,
  company details for both “Anderson Enterprises” and “Anderson Enterpr.”
- Biased data sets
  manual preselection of “noteworthy” maintenance tasks on a train
- Distributions changing in time
  increasing temperature values over last month
  (indication of failing monitored machine or running calibration of sensor?)
- ...

Lead to systematic uncertainties
in results of data analysis

Data-quality requirements

Example: data-quality levels large insurance company

- Strict risk-management regulations ("Solvency II"), especially after financial crisis
- Regulations also affect data requirements

The required data-quality level depends on the purpose of the data. Data-quality guide lines must be designed for each purpose. The goal is to "extend the project *DQ in source systems* beyond Solvency-II".

Financial reporting (Solvency requirement)	minimal DQ: 100%
Risk and premium definition	minimal DQ : 90%
Advanced analytics	
Reporting, marketing & sales	minimal DQ: 85%
Identify/monitor trends in data	minimal DQ: 75%

?

The desired DQ definitions and levels are determined iteratively on the job and evolve in time

Central question

How to quantify level of data quality for any given dataset?

... and to design automated data-quality checks on our data?

Data-quality business rules

- DQ business rule: specific, well-defined test of data
  Results in a “traffic light”: red, yellow, green

- DQ rules are applied at various levels

1. Field level
   Is a certain data field filled? Does it have the right format? Does it have an allowed value?
2. Record level
   Are all related elements filled? And filled correctly (wrt each other)?
3. Cross-record level
   Are all records corresponding to one person consistent with each other?
4. Analysis level / Over time
   Detects anomalies or trends in data over time → **Focus of this lecture**

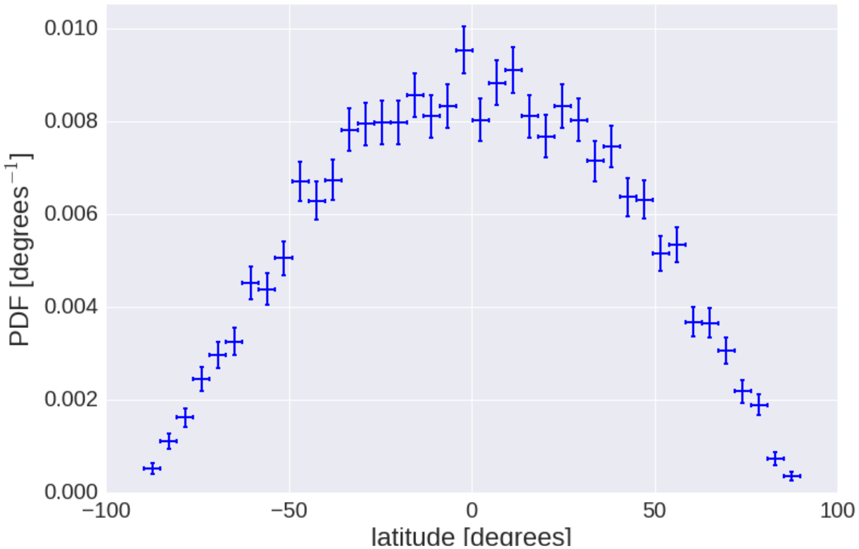
- Assess the DQ level:
  Apply DQ rules, count how often they work/fail

Data profiling

"Data profiling": collecting generic information on data

- About each data field (column) in a dataset
  length, #nans, #infs, #zeros, #unique, their fractions, mean, std, max, min, ...
- Histogram of distribution of each field
  For strings: count how often each string occurs
- Histograms of derived distributions
  - Distribution of most significant digit
    Check for Benford's law, e.g. for money values
  - Distribution of relative value counts
    Histogram of (normalized) value-histogram counts...
- Note: All quite easy evaluations!
  "Count how often something happens"

Data profiling



profile for single column
in input data

	length	frac. unique	frac. NaN	frac. 0	frac. positive	minimum	maximum	mean	std	median
datetimeCol										
2015-01-31	8390	0.984148	0.010250	0	0.605006	-999.99	1499.74	257.213800	721.762423	264.405
2015-02-28	7653	0.983928	0.008624	0	0.598981	-999.94	1499.78	247.177364	716.717271	243.140
2015-03-31	8508	0.981782	0.010696	0	0.605783	-999.88	1499.50	259.871051	717.811026	266.865
2015-04-30	8282	0.981043	0.010746	0	0.600217	-999.63	1500.00	245.900256	721.913009	255.275
2015-05-31	8405	0.982867	0.010708	0	0.590720	-999.85	1499.88	237.254653	723.026722	226.310
2015-06-30	8117	0.981767	0.009363	0	0.601084	-999.93	1499.88	252.777563	720.835069	245.820
2015-07-31	8566	0.978403	0.008639	0	0.595844	-998.95	1499.64	244.516981	723.872802	227.695
2015-08-31	8430	0.983274	0.010202	0	0.599644	-999.91	1499.77	250.079751	718.966038	247.775
2015-09-30	7894	0.983785	0.010894	0	0.609704	-999.50	1500.00	264.444297	730.142341	273.130
2015-10-31	8344	0.983341	0.010187	0	0.603787	-999.62	1499.90	261.942858	723.093769	254.820
2015-11-30	8175	0.982997	0.008073	0	0.598532	-999.93	1499.93	251.151890	716.849333	240.870
2015-12-31	8258	0.983410	0.010051	0	0.597239	-999.76	1499.29	242.961905	721.729556	243.145

Define generic DQ rules based on profiling

- def traffic_light(x):

  ```
  if x < red_lo or x >= red_hi: return 'red'
  if x < yellow_lo or x >= yellow_hi: return 'yellow'
  return 'green'
  ```

- def nan_traffic_light(num_nans)

  ```
  if num_nans >= 3: return 'red'
  if num_nans >= 1: return 'yellow'
  return 'green'
  ```

- Setting fixed thresholds would be kind of annoying ☹
  - Could get ranges automatically from reference data
  - This also enables further comparison of test data and reference

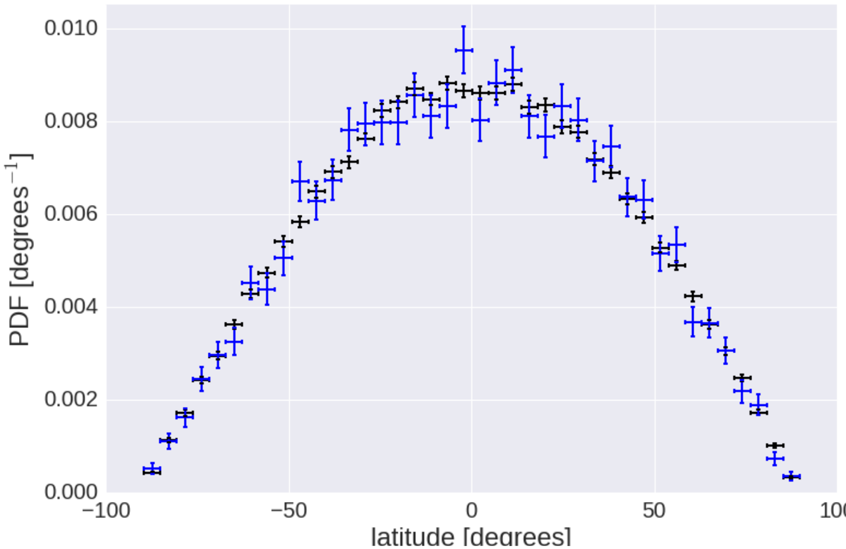
Compare with reference dataset

Key concept for generic DQ monitor

- Base DQ evaluation on comparison of test dataset with reference dataset
- For this example: records in datasets are time-stamped
  Useful, but not a necessary requirement
- DQ assessment done per (user-defined) unit time block e.g. per day/week/month
  Flag data if inconsistencies are found
- Possibly reject block of test data if DQ is insufficient (red)
- Apply data profiling to
  1. Entire reference dataset
     If possible to the blocks of the reference dataset
  2. All blocks of the test dataset
     If not possible, to entire test dataset only

Compare with reference

profile for single column
in input data



	length	frac. unique	frac. NaN	frac. 0	frac. positive	minimum	maximum	mean	std	median
period mean	8261.333333	0.983654	0.008721	0	0.599139	-999.769167	1499.694167	250.237557	722.372284	250.537500
period std.	297.886596	0.001208	0.000591	0	0.004021	0.251413	0.380513	6.570123	4.984437	11.410786
2015-02-28	7653	0.983928	0.008624	0	0.598981	-999.94	1499.78	247.177364	716.717271	243.140
2015-03-31	8508	0.981782	0.010696	0	0.605783	-999.88	1499.50	259.871051	717.811026	266.865
2015-04-30	8282	0.981043	0.010746	0	0.600217	-999.63	1500.00	245.900256	721.913009	255.275
2015-05-31	8405	0.982867	0.010708	0	0.590720	-999.85	1499.88	237.254653	723.026722	226.310
2015-06-30	8117	0.981767	0.009363	0	0.601084	-999.93	1499.88	252.777563	720.835069	245.820
2015-07-31	8566	0.978403	0.008639	0	0.595844	-998.95	1499.64	244.516981	723.872802	227.695
2015-08-31	8430	0.983274	0.010202	0	0.599644	-999.91	1499.77	250.079751	718.966038	247.775
2015-09-30	7894	0.983785	0.010894	0	0.609704	-999.50	1500.00	264.444297	730.142341	273.130
2015-10-31	8344	0.983341	0.010187	0	0.603787	-999.62	1499.90	261.942858	723.093769	254.820
2015-11-30	8175	0.982997	0.008073	0	0.598532	-999.93	1499.93	251.151890	716.849333	240.870
2015-12-31	8258	0.983410	0.010051	0	0.597239	-999.76	1499.29	242.961905	721.729556	243.145

Pulls with respect to reference data

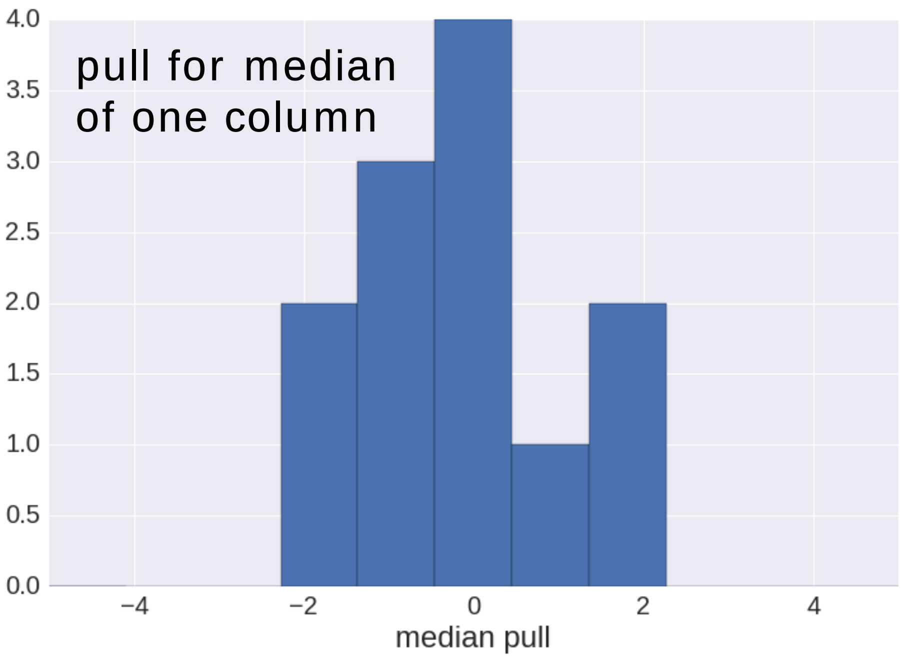
value of single block

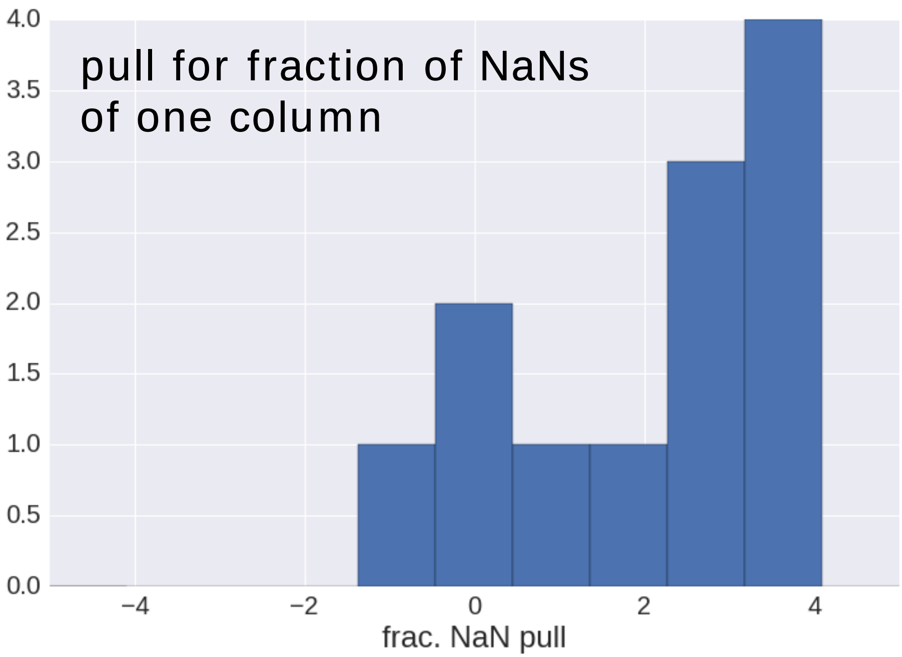
mean of block values

$$\text{pull} = \frac{\text{test value} - \text{reference mean}}{\text{reference std.}}$$

standard deviation
of block values

Expect normal distributions with
$\mu = 0$ and $\sigma = 1$



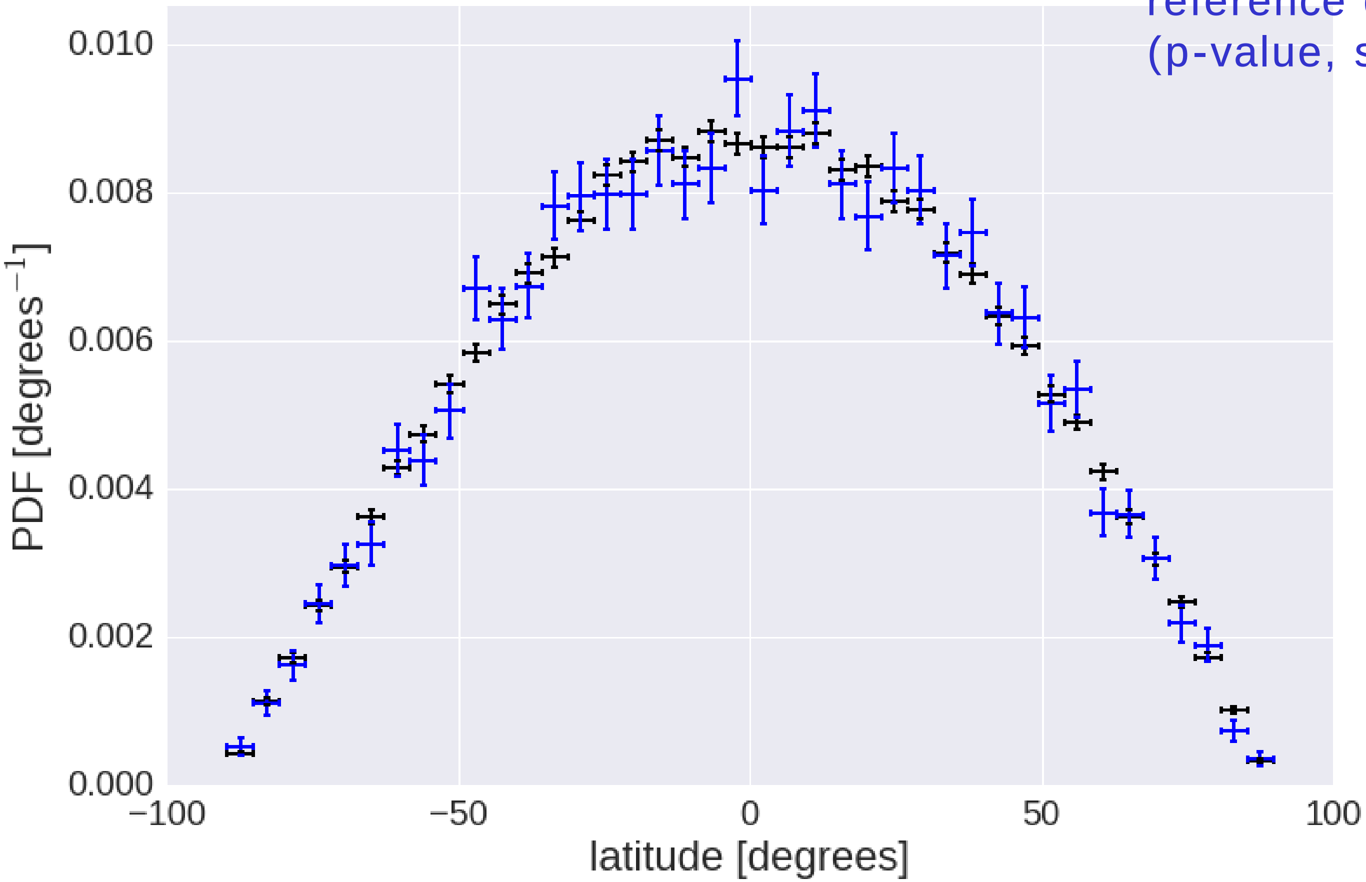


Further tests of distributions

Compare distribution of test-column values in each block with reference distribution

- χ^2 test
- Kolmogorov-Smirnov test

probability to find test distribution with equal or greater difference with reference distribution (p-value, see next lecture)



Define additional DQ rules

- Test all evaluated pulls for outliers

```
def pull_traffic_light(pull):
        if |pull| > 5: return 'red'
        if |pull| > 3: return 'yellow'
        return 'green'
```

- Test distribution p-values for outliers:

```
def distr_test_traffic_light(p_value):
        if p_value < 0.001: return 'red'
        if p_value < 0.01: return 'yellow'
        return 'green'
```

- Simply count how often the DQ rules work and fail...

Data-quality dashboard

Data Quality Summary

DQ% Overall

93,12%

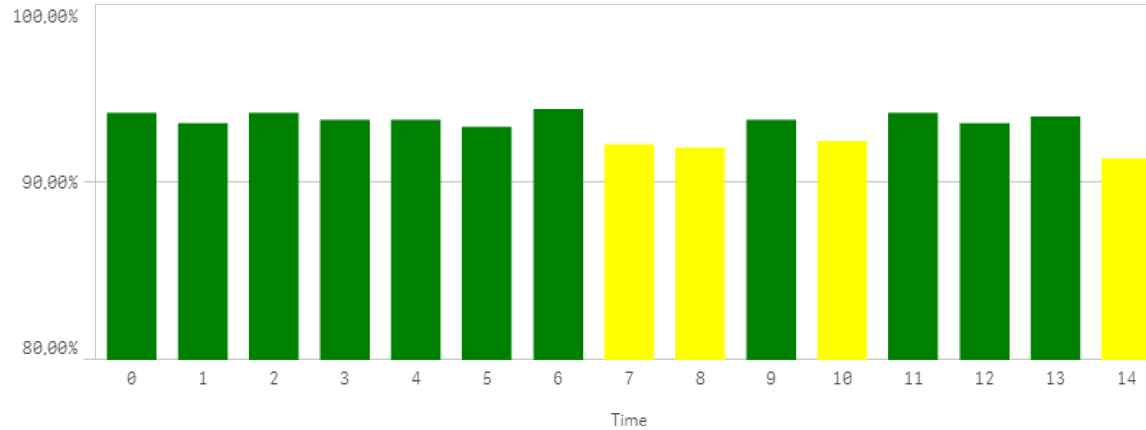
Last Period

91,29%

Change from last Period

-2,38%

DQ% over Time



Green

7,05k

Yellow

400

Red

121

Observable	DQ%
Totals	**93,12%**
cat	86,67%
amount	88,70%
f7	92,44%
f2	92,74%
f0	93,04%
f4	93,19%
f3	93,63%
f8	93,93%
f1	94,37%

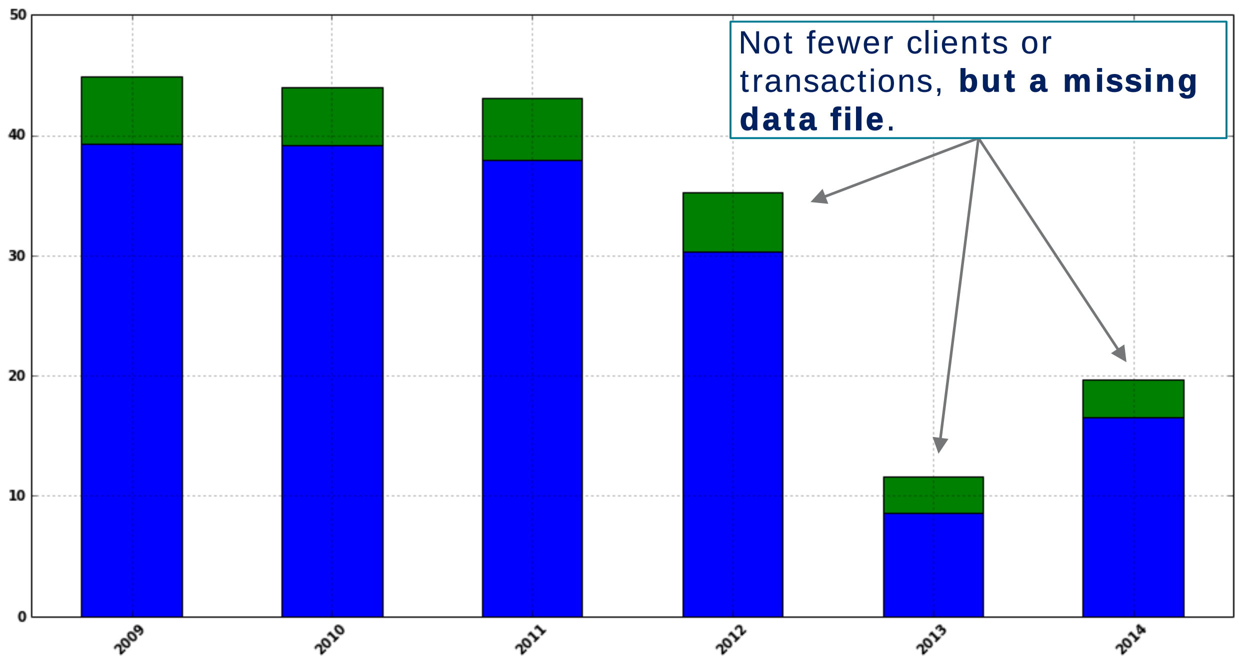
Backup

Data Errors

Types of data ‘errors’:

- Missing files or records
- Missing or incorrect fields

Follow-up number	Date	Amount	Transaction type
20150004	Jan 13 3015	-100.00 €	Pin withdrawal
20150005	Jan 15 2015	-50.00 €	Pin payment
20150006	Jan 20 2015	2500.00 €	
20150007	Jan 22 2015	-40.00 €	Pin payment
20150008	Jan 29 2015	-1500.00 €	failed
20150008	Jan 29 2015	-1500.00 €	Bank transfer
20150009	Feb 3 2015	-250.00	Pin payment
20150010	Jan 4 2015	-150.00	Pin payment



Data Errors

Types of data 'errors':

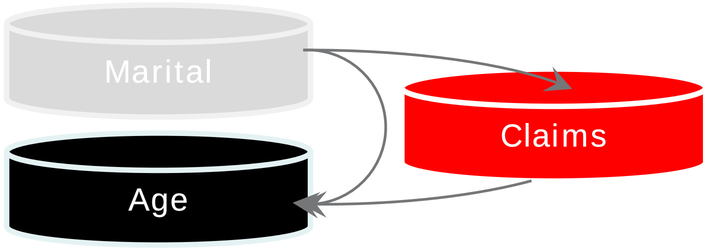
- Missing or ill understood relations between fields
- Missing or ill understood relations between records

Follow-up number	Date	Amount	Transaction type
20150004	Jan 13 3015	-100.00 €	Pin withdrawal
20150005	Jan 15 2015	-50.00 €	Pin payment
20150006	Jan 20 2015	2500.00 €	
20150007	Jan 22 2015	-40.00 €	Pin payment
20150008	Jan 29 2015	-1500.00 €	failed
20150008	Jan 29 2015	-1500.00 €	Bank transfer
20150009	Feb 3 2015	-250.00	Pin payment
20150010	Jan 4 2015	-150.00	Pin payment

Data Errors

Types of data 'errors':

- Biased data



Enriching Age data with Marital status:

1. If name given in Age data, then get Marital status directly with name.
2. Else get name from Claims data through address, and then get Marital status with name.
3. Else marital status remains unknown.

Now what is the probability that a married person makes a claim?

But people with claims tend to have better known marital status than others!