Data Fraud Detection

("Money makes the world go 'round...")

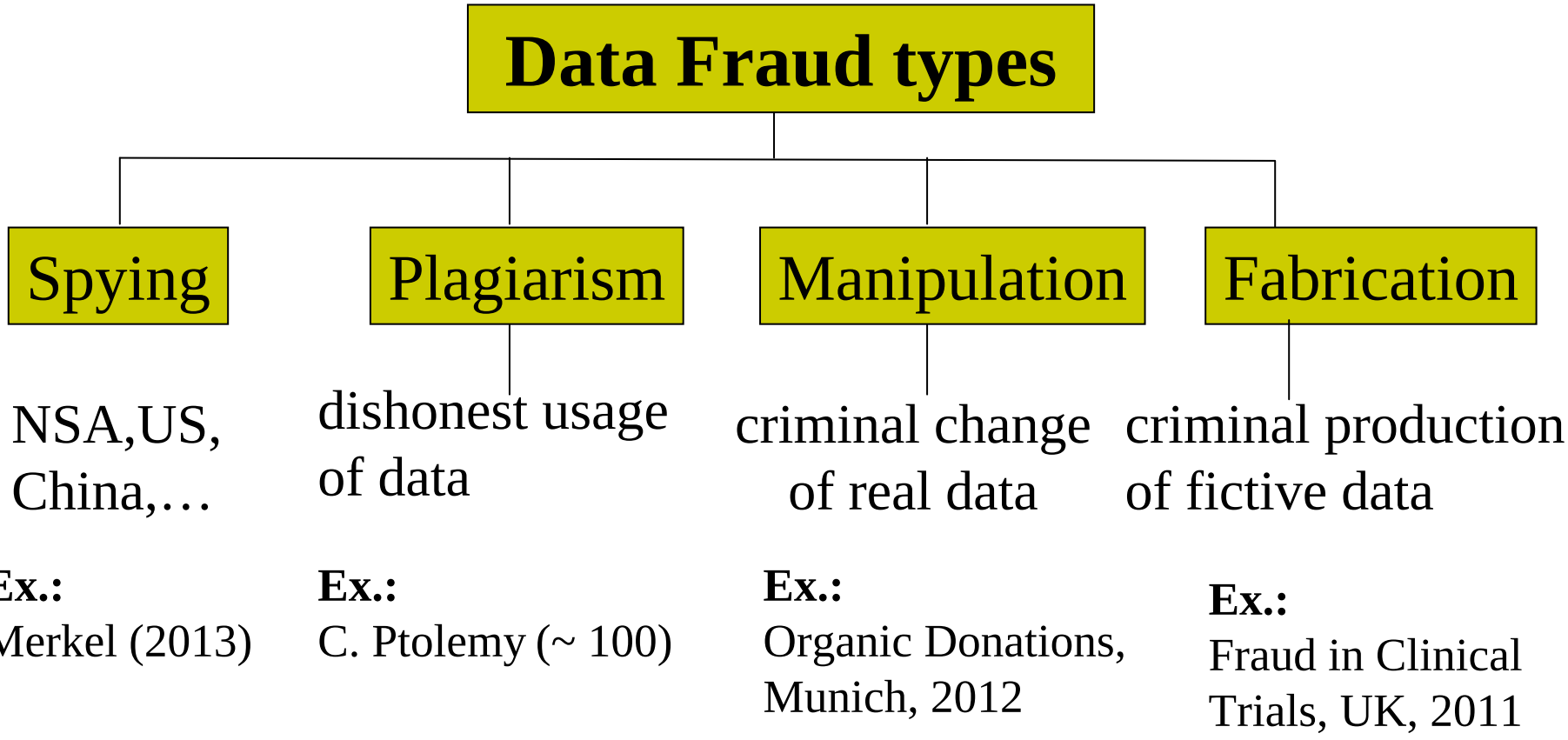
Hans - J. Lenz
Inst. f. Statistik und Ökonometrie
Freie Universität Berlin

hans-j.lenz @ fu-berlin.de

Overview

- **Short History** of Deception and Fraud
  from *Ptolemy* to …
- **Ontology** of Data Fraud
- **Data Spy out** (military, governmental & business misuse) (not my topic today)
- **Data Plagiarism** (not my topic today)
  - Swarm intelligence
  - String and pattern matching
  - Meta analysis by IR techniques
- **Data Manipulation**
  - Stochastic independence assumption, multiple tests on same data, manipulating p-values,…
  - Benford's Law
  - Data-Model Conforming (DMC) Tests
- **Data Fabrication**
  - Benford's Law
  - Inlier tests
  - Outlier tests
- **Next Steps**

Ontology
of Data Fraud



I A short History of Deception and Fraud

in Astronomy

Claim: Claudius Ptolemy (~100 p. C.) copied astronomical data from Hipparchos von Nicaea (~ 190-120 a. C.) ,
but he argued to have them collected himself.

☞ **Data Plagiarism**

Source: Di Trocchio (1994)

A short History (cont.)

in Science

Galilei's Experiments couldn't have been run in its way at that time.

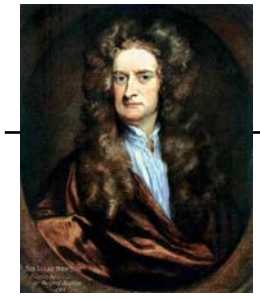
Claim: "The Genius was motivated by the objective of supporting the final break-through of his ideas."

☞ **Principle of Reproducibility**

Source: R. Sheldrake (1996, p 176), W. Broad & N. Wade (1985, p 27)

A short History (cont.)

in Physics



- **Newton's Principia** convinced due to the reported high precision of his observations far from being legitimate.
- **Claim:** "Nobody was so brilliant and effective in cheating than the master mathematician."

☞ **Principle of Reproducibility**

Source: R. Sheldrake (1996, p 176), R.S. Westfall (1973, p 1118)

A short History (cont.)

in Physics

Robert Millikan, US Nobel Price Winner

Claim: strikingly precise measurements of the charge of electrons in 1913

- quite opposite to his rival *Felix Ehrenfeld* who experienced large deviations.
- Milikan's lab protocols showed later:
  He published only the **'best 58 out of 140' experiments** having smallest variance.

☞ **Experiments Selection Bias;**

Principle of Reproducibility

Source: R. Sheldrake (1996, p 176-177), W. Broad and N. Wade (1985, p 34)

A short History (cont.)

in Banking

- **Deutsche Bank**, Germany's largest bank was a member of the Libor Cartel from 2006-2009 **.**
- **Claim:** EU, USA accused the cartel for secretly up/down-raising the Libor.
- **Note**: Libor is used by banks as the leading interest rate.

  It influences the interest rates of credit and saving accounts of all of us customers.

☞ **Data manipulation** (by Ackermann, Jain,…)

Source: Frankfurter Allgemeine Sonntagszeitung (10.02.2013, S. 21-22)

A short History (cont.)

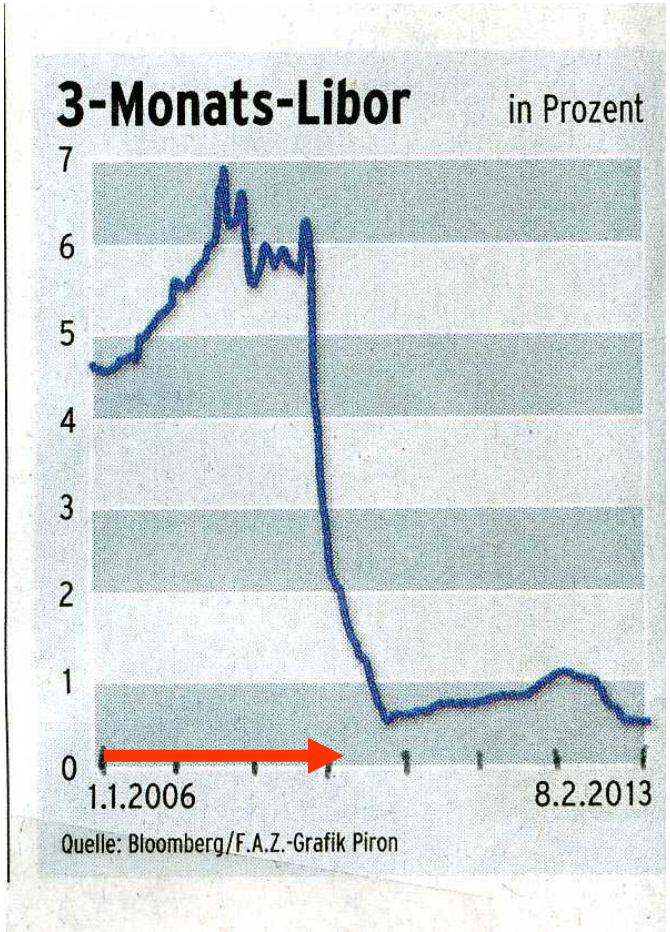
in Banking: 3m/6m-Libor Manipulation

Libor Cartel:

UBS

RBS
The Royal Bank of Scotland

BARCLAYS



Source: Ch. Siedenbiedel, Die Libor-Bande, Frankfurter Allgemeine Sonntagszeitung, Nr.6, 10.02.2013, S. 21-22

A short History (cont.)

in clinical research

Illegale Studie über Harninkontinenz bei Frauen von Urologe Dr. H. Strasser, MUI

Text
Editorial

Nature **454**, 917-918 (21 August 2008) | doi:10
20 August 2008

Scandalous behaviour

Top of page
Abstract

Austria's most serious report of scientific misconduct in recent memory must be handled properly.

[…]

According to a report from the Austrian Agency for Health and Food Safety, a urologist at the university, Hannes Strasser, has conducted a high-profile clinical trial so inappropriately that it must be considered entirely invalid (see page 922). Moreover, that trial represents just a fraction of the total number of patients who paid handsomely for the stem-cell treatment for urinary incontinence without knowing it was experimental.

"There is no official body in Austria responsible for addressing issues of scientific misconduct."

[…]

Austria is a small country, and networks between power-brokers are small and tight. But something, it seems, is rotten in the state of Austria, and it needs to be faced and dealt with openly.

End of page

☞ **Unsound statistical experiments & methods**

A short History (cont.)

in Health Care (Organic Transplantations)

Fachleute von Eurotransplant, der zentralen Vergabestelle für Organe mehrerer Länder in Mitteleuropa, begrüßen die Festnahme. Die Abschreckungswirkung für zukünftige Manipulationen, so ein leitender Arzt, „ist nun natürlich riesengroß".

Der Haftbefehl gibt auch den Prüfern der Bundesärztekammer größere Bedeutung. Sie wurden nach der Aufdeckung der ersten Organskandale in Göttingen und Regensburg eingesetzt und sollen das Vertrauen der Öffentlichkeit in die Organspende wieder herstellen. Doch nachdem die Kontrolleure ein Viertel der 49 Transplantationszentren abgearbeitet haben, sind sie auf mehr Auffälligkeiten gestoßen, als sie für möglich gehalten hätten.

Im ersten Bericht der Prüfkommission, der noch in diesem Monat veröffentlicht werden soll, geht es um das Universitätsklinikum

Beschuldigter Arzt O.
Allmacht und Eitelkeit?

schweiger Ermittler für eine abstraktere Argumentation: Irgendwo im Eurotransplant-Verbund wird auf jeden Fall ein schwerkranker Patient benachteiligt, wenn ein Arzt seinen Patienten unrechtmäßig bevorzugt. Er nimmt damit billigend in Kauf, dass dieser andere Patient stirbt.

Während in Göttingen der beschuldigte Mediziner O. in Untersuchungshaft sitzt, prüft die Staatsanwaltschaft München noch, wie sie mit den Vorfällen im Klinikum rechts der Isar umgehen soll. Dort hatten interne Prüfer im September bei neun Lebertransplantationen aus 2010 und 2011 Unregelmäßigkeiten entdeckt. Auch hier sollen Blutwerte manipuliert, Dialysen vorgetäuscht und Krebspatienten transplantiert worden sein, die möglicherweise kein Organ hätten bekommen dürfen.

Im Fokus steht der Transplantationschirurg Peter B. Er soll im Januar 2010 einer

DER SPIEGEL 3/2013

Der Spiegel 3/2013 S.42

☞ **Data Manipulation**

Source: Der Spiegel, Nr. 3/2013, S. 42-44

A short History (cont.)

in Politics: Greece's manipulated Euro Entry (2001)

Monday 18 February 2013

The INDEPENDENT

NEWS | VOICES | SPORT | TECH | LIFE | PROPERTY | ARTS & ENTS | TRAVEL | MONE

UK | World | Business | People | Science | Environment | Media | Technology | Education | Obituar

News > World > Europe

Greece admits deficit figures were fudged to secure euro entry

BY DANIEL HOWDEN AND STEPHEN CASTLE IN BRUSSELS | TUESDAY 16 NOVEMBER 2004

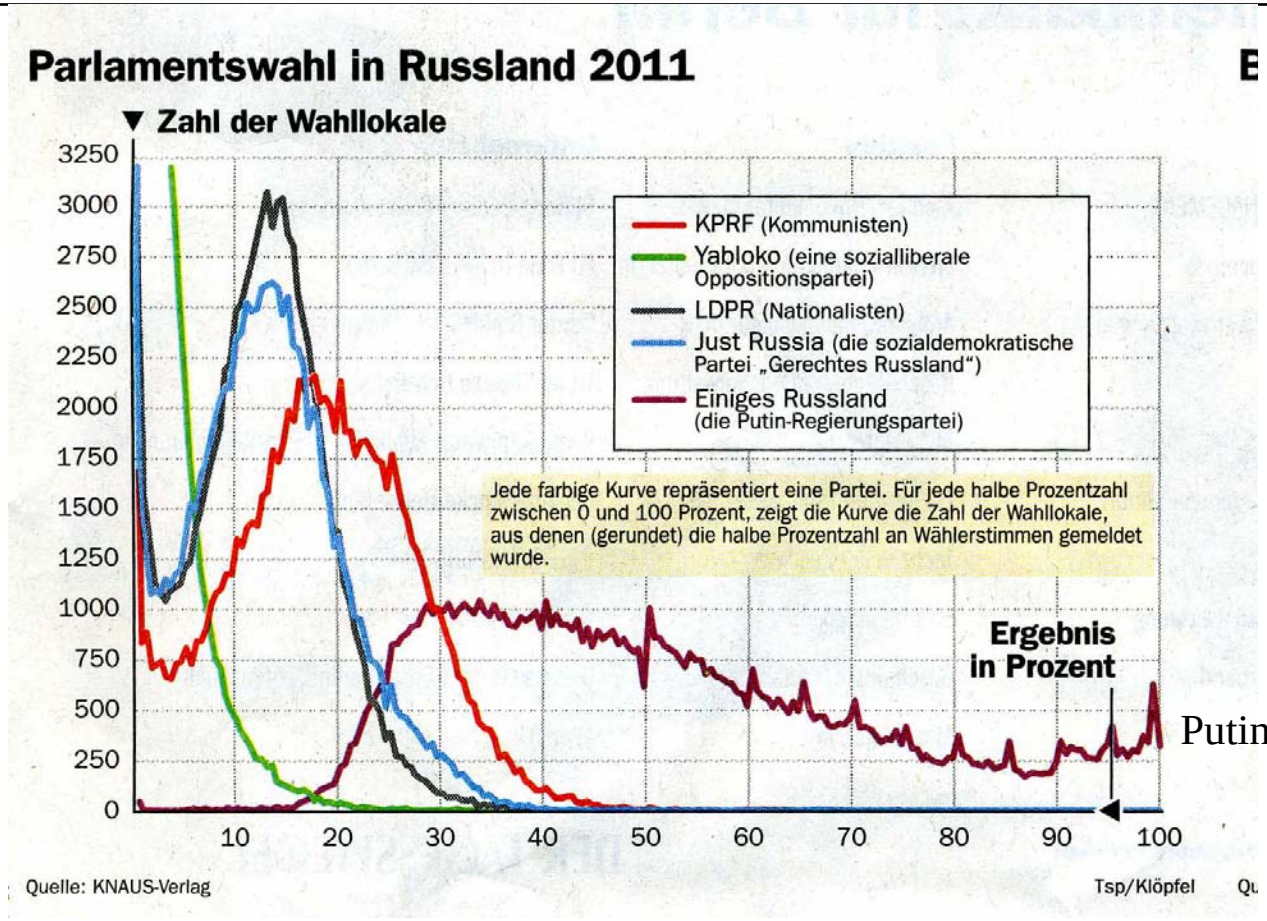
The Greek financial daily *Naftemboriki* said the corrected deficits for the crucial period from 1997 to 1999, when the country's economic data was scrutinized to decide on its eligibility for the Euro Zone, were 6.44 per cent, 4.13 per cent and 3.38 per cent respectively. The conservative government in Athens has placed the blame on its Socialist predecessors.

Greece former Finance Minister:

"It has been proven that the deficit had not fallen below 3 per cent in every year since 1999," Mr Alogoskoufis told reporters.

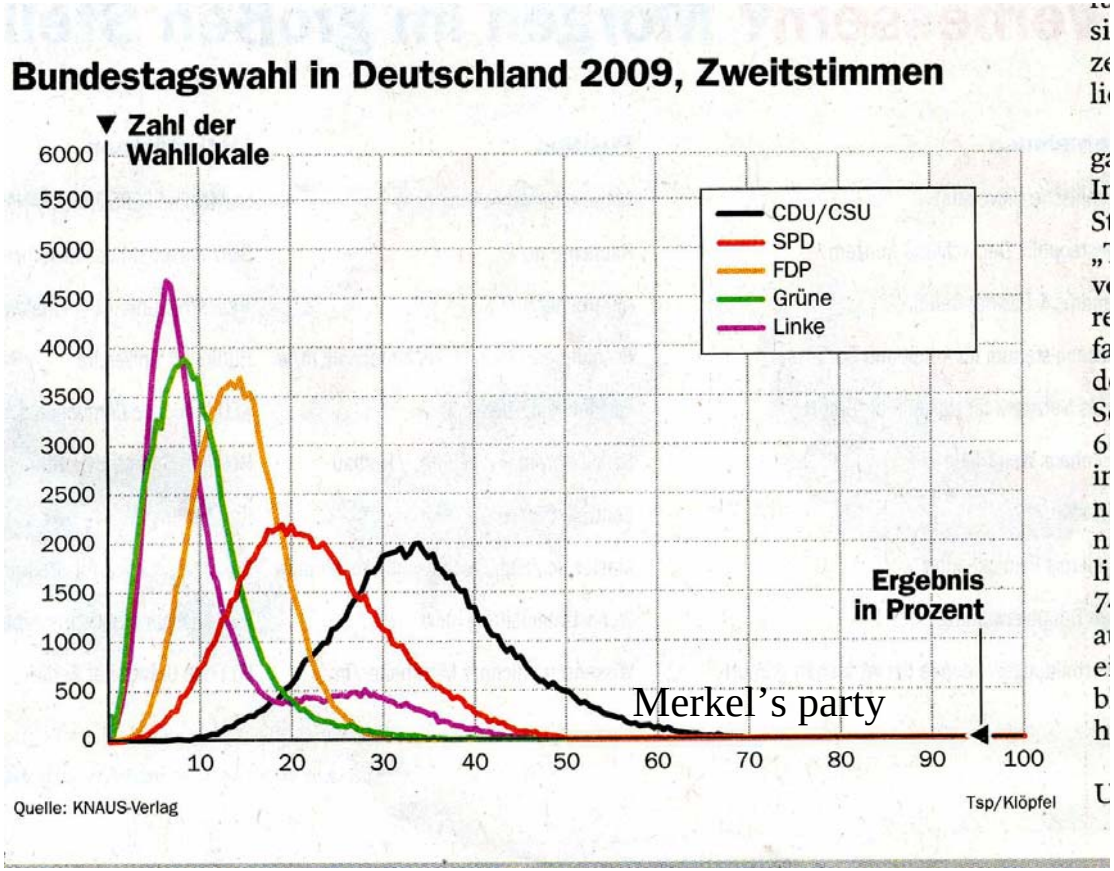
Source: The Independent, 18.02.2013

Distribution of votes on electoral districts Russia 2011



Putin's party

Distribution of votes on electoral districts Germany 2009



Field Study on US Science

Martinson, Anderson and de Vries (2005)

- sample size n= 3247 out of 7760 scientists who got a grant from National Institutes of Health
- anonymous self-report based on a standardized questionnaire with (10 + 6) items on (top + other) fraud types

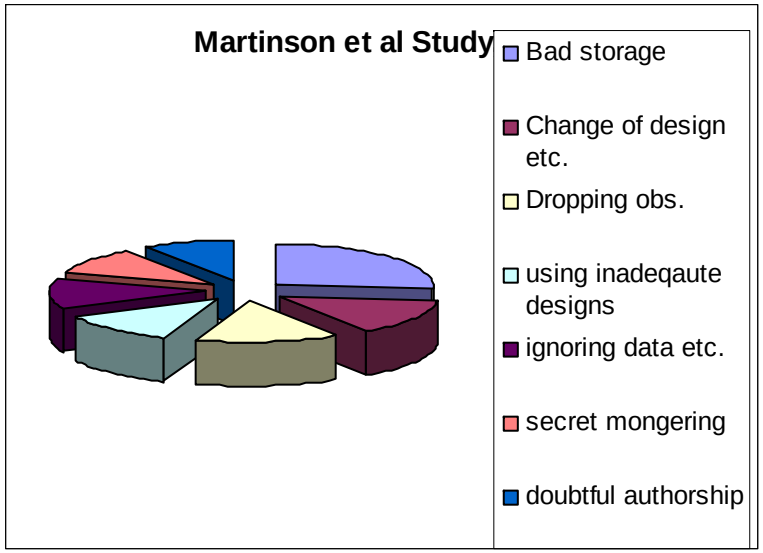


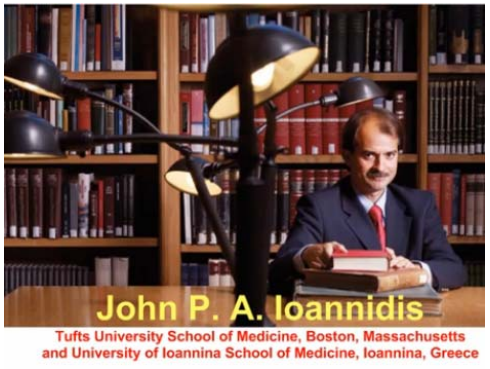
Table 1 | Percentage of scientists who say that they engaged in the behaviour listed within the previous three years ($n = 3,247$)

Top ten behaviours	All	Mid-career	Early-career
1. Falsifying or 'cooking' research data	0.3	0.2	0.5
2. Ignoring major aspects of human-subject requirements	0.3	0.3	0.4
3. Not properly disclosing involvement in firms whose products are based on one's own research	0.3	0.4	0.3
4. Relationships with students, research subjects or clients that may be interpreted as questionable	1.4	1.3	1.4
5. Using another's ideas without obtaining permission or giving due credit	1.4	1.7	1.0
6. Unauthorized use of confidential information in connection with one's own research	1.7	2.4	0.8 ***
7. Failing to present data that contradict one's own previous research	6.0	6.5	5.3
8. Circumventing certain minor aspects of human-subject requirements	7.6	9.0	6.0 **
9. Overlooking others' use of flawed data or questionable interpretation of data	12.5	12.2	12.8
10. Changing the design, methodology or results of a study in response to pressure from a funding source	15.5	20.6	9.5 ***
Other behaviours			
11. Publishing the same data or results in two or more publications	4.7	5.9	3.4 **
12. Inappropriately assigning authorship credit	10.0	12.3	7.4 ***
13. Withholding details of methodology or results in papers or proposals	10.8	12.4	8.9 **
14. Using inadequate or inappropriate research designs	13.5	14.6	12.2
15. Dropping observations or data points from analyses based on a gut feeling that they were inaccurate	15.3	14.3	16.5
16. Inadequate record keeping related to research projects	27.5	27.7	27.3

Note: significance of χ^2 tests of differences between mid- and early-career scientists are noted by ** ($P < 0.01$) and *** ($P < 0.001$).

Source: http://www.vub.ac.be/phd/doctoralschools/lsm/docs/435737a.pdf

Unsound Methods of US Psychologist



http://de.wissenschaftlichepraxis.wikia.com/wiki/Datei:Why_Most_Published_Research_Findings_Are_False

FAS, 29.09.13, S.59

Seven Shades of Grey

Im Fachblatt *Psychological Science* erschien im April 2012 eine Studie, für die Forscher um Leslie John von der Harvard Business School mehr als zweitausend amerikanische Psychologen anonym befragten, wwelche fragwürdigen Praktiken sie anwenden. Am häufigsten wurde zugegeben:

1. Das weitere Sammeln von Daten davon abhängig zu machen, ob die bisherigen die Signifikanzschwelle erreicht haben (58 Prozent gaben an, es schon einmal gemacht zu haben).

2. Teilexperimente unterschlagen, die nicht den erwarteten Effekt zeigten (27 Prozent).

3. Abrunden von Daten in den signifikanten Bereich hinein (23 Prozent).

4. Messpunkte als Ausreißer ausschließen, nachdem überprüft wurde, ob dies das Ergebnis in die gewünschte Richtung bringt (43 Prozent).

5. Überraschende Ergebnisse in der Publikation als von vornherein erwartet ausgeben (35 Prozent).

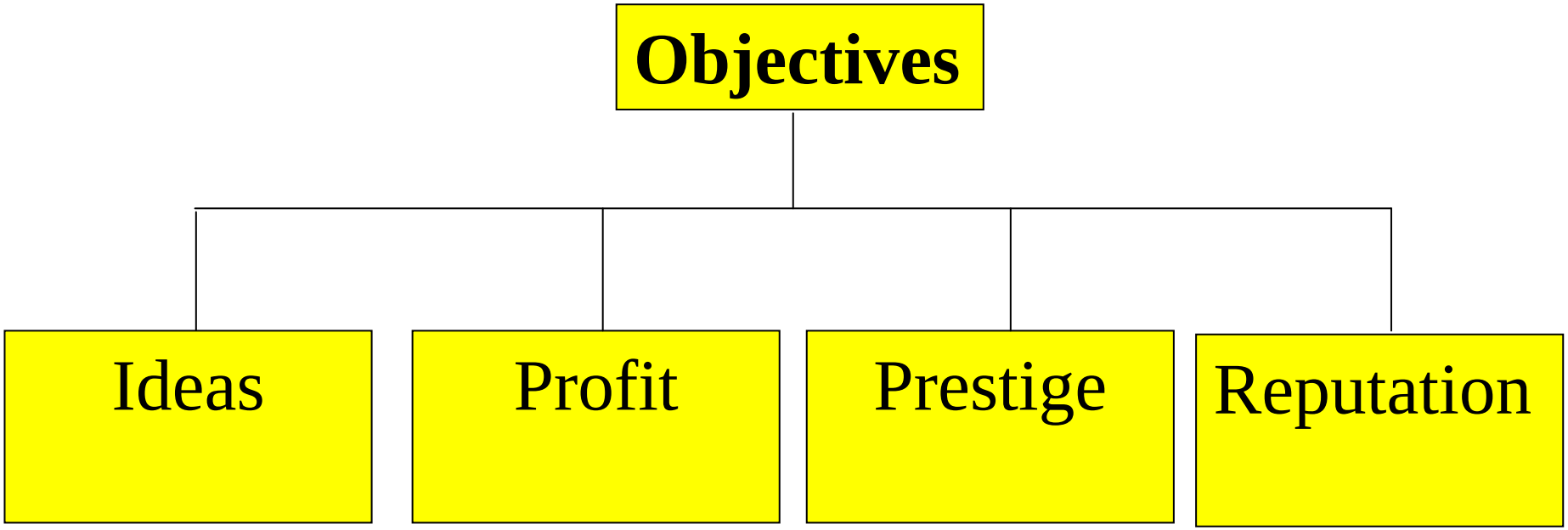
6. Ergebnisse für unabhängig von demographischen Variablen wie dem Geschlecht der Probanden erklären, wenn man sich in Wirklichkeit nicht sicher ist oder vielleicht sogar weiß, dass dies nicht stimmt (vier Prozent).

7. Daten fälschen (zwei Prozent).

Dass es gang und gebe ist, grenzwertige Ergebnisse mit solchen Tricks irgendwie über die Signifikanzschwelle zu retten (meist angegeben mit dem p-Wert, der die Wahrscheinlichkeit angibt, mit der ein gemessener Unterschied in Wirklichkeit ein reines Zufallsprodukt ist), lässt eine 2012 im *Quarterly Journal of Experimental Psychology* veröffentlichte Analyse annehmen. Bei Durchsicht der Studien, die innerhalb eines Jahres in drei anerkannten Fachzeitschriften erschienen waren, fiel eine Häufung von p-Werten auf, die gerade unterhalb von fünf Prozent lagen. Eine Lücke klaffte bei Werten, die knapp über der Signifikanz-Marke lagen. Ein Schelm, wer Böses dabei denkt. *geru*

Source: Psychological Science, April 20012

II Data Manipulation & Data Fabrication



Ptolemy, Galileo, Newton,…

Ackermann,…

Greece, Belgium, …

H. Strasser, MUI, …

Roles of People (stage holder)

Reality

Collective Self Deception
(‘Wunschdenken’)

Mainstream Opinion
(‘Schulmeinung‘)

Poor Wretch
(‘Sündenbock‘)

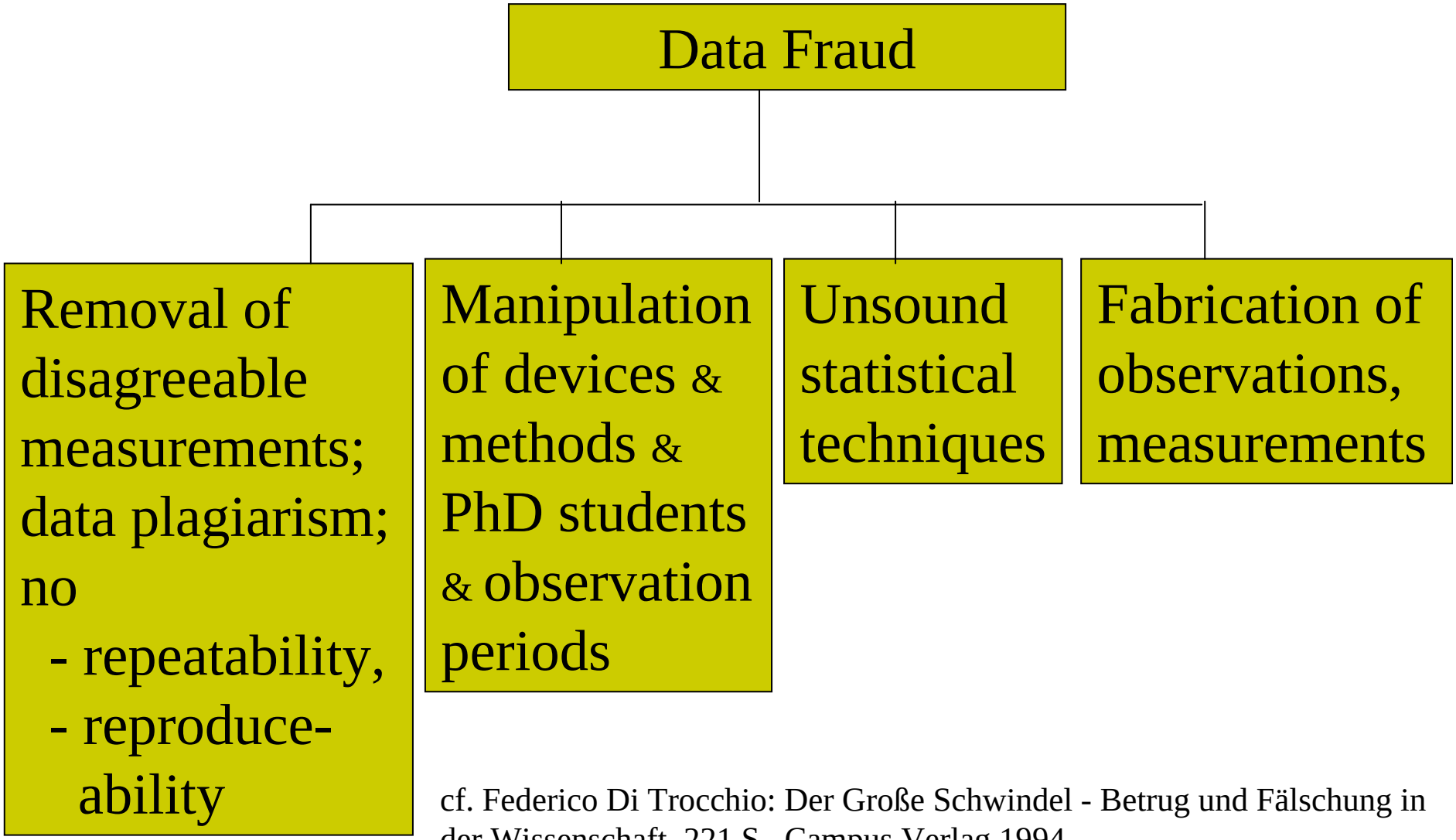
Cheater

Truth Finder

Denouncing
Informer

Ontology

of Data Fraud in science



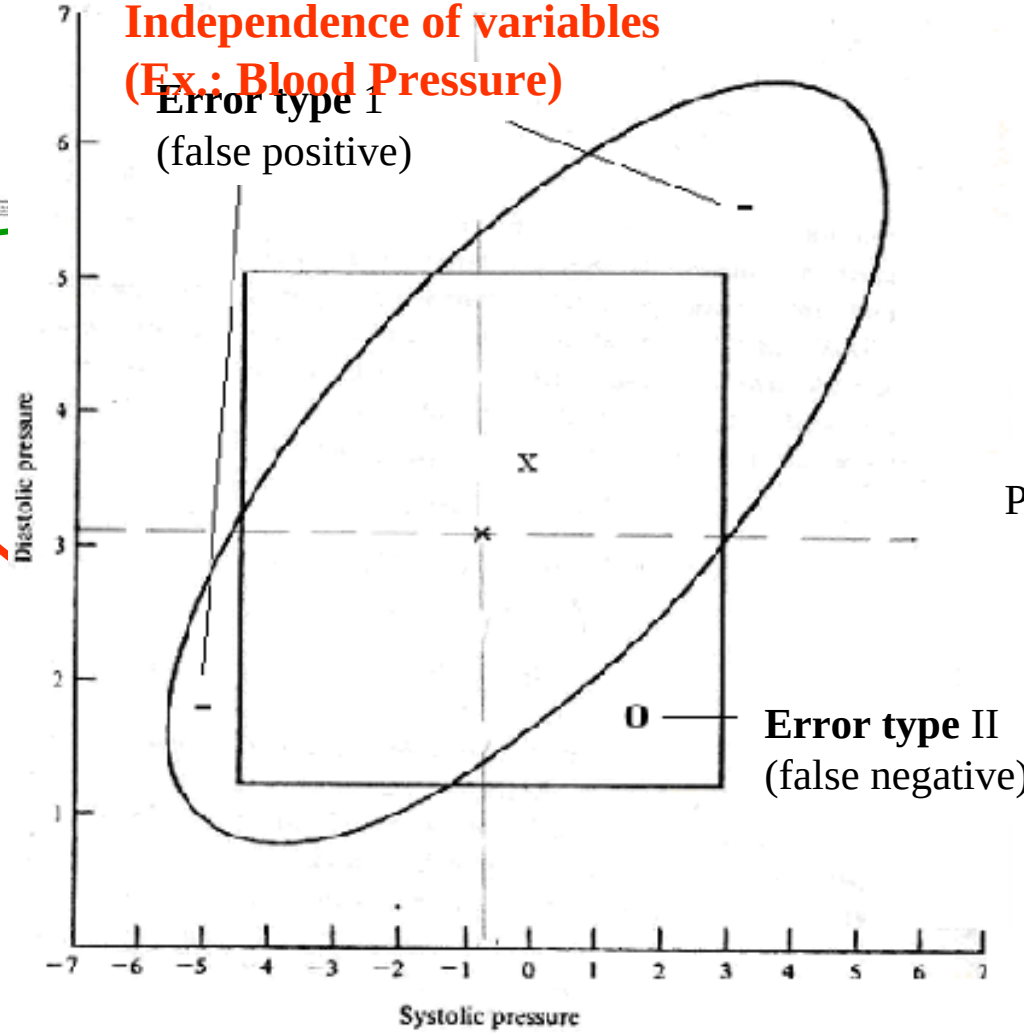
cf. Federico Di Trocchio: Der Große Schwindel - Betrug und Fälschung in der Wissenschaft, 221 S., Campus Verlag 1994

III Data Manipulation
(the little tricks you know…)

Statisticians prefer simultaneous 2-dim intervals (ellipses)

Labs and doctors use two 1-dim confidence intervals (rectangles)

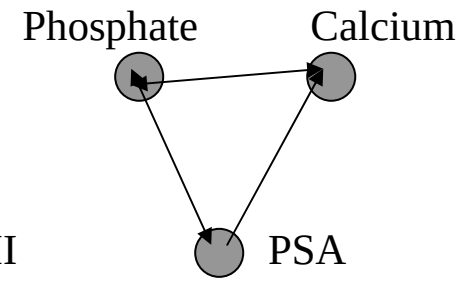
Independence of variables (Ex.: Blood Pressure)



Error type I
(false positive)

Error type II
(false negative)

- Boundaries of tumor marker tests in feedback systems of the human body



Dependency Graph

Misuse of statistical techniques (cont.)

(multiple tests on same data set)

Problem: A set ***H*** of hypotheses and one data set ***D***. Which $h \in \boldsymbol{H}$ is 'true'/ 'best'?

Ex.: n=20 hypotheses, (nominal) significance level α=5%

What is the probability of observing at least one significant result?

P(at least one significant result)=1-P(no significant results)=

$1-(1-0{,}05)^{20} \approx 0{,}64 = \alpha_{eff} > \alpha$

Bonferroni Correction (BC)

Set significance level $\alpha_n = \alpha/n$.

Ex.: n=20, α=5% $\Rightarrow$ α_n=0,0025

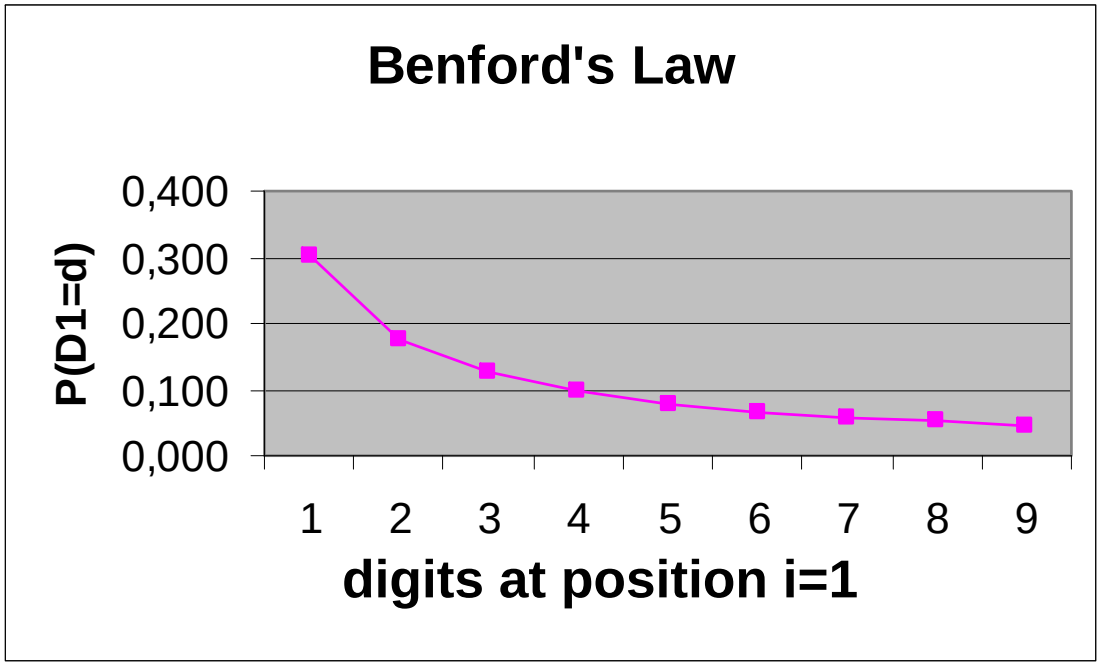
P(at least one significant result)=$1-(1-0{,}0025)^{20} \approx 0{,}0488 < \alpha$=5%

Note: BC leads to a conservative test ***but may lead*** to a high rate P(II) (>50%) of false negative!

Source: M. Goldman (2008)

Data Manipulation (cont.)

Benford‘s Law



Distribution $P_C(D_i=d)$ of Numeral D_i at Position i = 1,2,…
of figures from a well defined corpus C (numerical database)

Newcomb (1881), Benford (1938), Hill (1995), Göb (2007)

Data manipulation (cont.)

Benford's Law

Scientific notation of numbers:

Each $x \in \boldsymbol{R}_+$ can be written as $x = \langle x \rangle 10^n$ with $n \in \boldsymbol{Z}$ and (normalized) mantissa $1 \leq \langle x \rangle < 10$

Ex.: $0{,}00123 = 1{,}23\ 10^{-3}$

$\langle\ 0{,}00123\ \rangle = 1{,}23$

Theorem (Benford's Law)

Let $\log x = \log \langle x \rangle + n$ where $0 \leq \log \langle x \rangle < 1$.
If the random variable X is distributed according to a Benford df the logarithm of its mantissa is uniformly distributed over [0,1)

$P(\log \langle X \rangle < t) = t$.

Data manipulation
Benford's Law (cont.)

DEF.: i-th leading digit

Let D_i: $\boldsymbol{R}_+ \rightarrow \{0,1,2,\ldots,9\}$ be a random variable with $D_i(x) = d$ where $\langle x \rangle = d_1,d_2d_3\ldots d_k$

Ex.: $D_1(0,00123)=1$

Theorem (pdf of first digit)

If X ~ Benford pdf $\Rightarrow$ $P(D_1=d)=\log(1+1/d)$

Proof

$D_1=d \Leftrightarrow d \leq \langle X \rangle < d+1 \Rightarrow P(D_1=d)=P(d \leq \langle X \rangle < d+1)=$

$P(\log d \leq \log \langle X \rangle < \log (d+1)) = \log(d+1)-\log(d)=\log(1+1/d)$

Data manipulation
Benford's Law (cont.)

Theorem (pdf of first k digits)

If X ~ Benford pdf
then $P(D_1=d_1, D_2=d_2,\ldots,D_k=d_k))=\log(1+1/(d_1d_2\ldots d_k))$

where $d_1 \in \{1,2,\ldots,9\}$ und $d_j \in \{0,1,2,\ldots,9\}$ f.a. $j=2,3,\ldots,k$.

Theorem (Scale and basis Invariance of Benford's law)
Pinkham(1961)

Let f_X be a probability distribution.
f_X is scale and basis scale invariant $\Leftrightarrow$ f_X is a Benford pdf

Data Manipulation (cont.)

Fraud Detection by Benford‘s Law

Case: Intl. hotel company X runs its own health insurance department (HID)

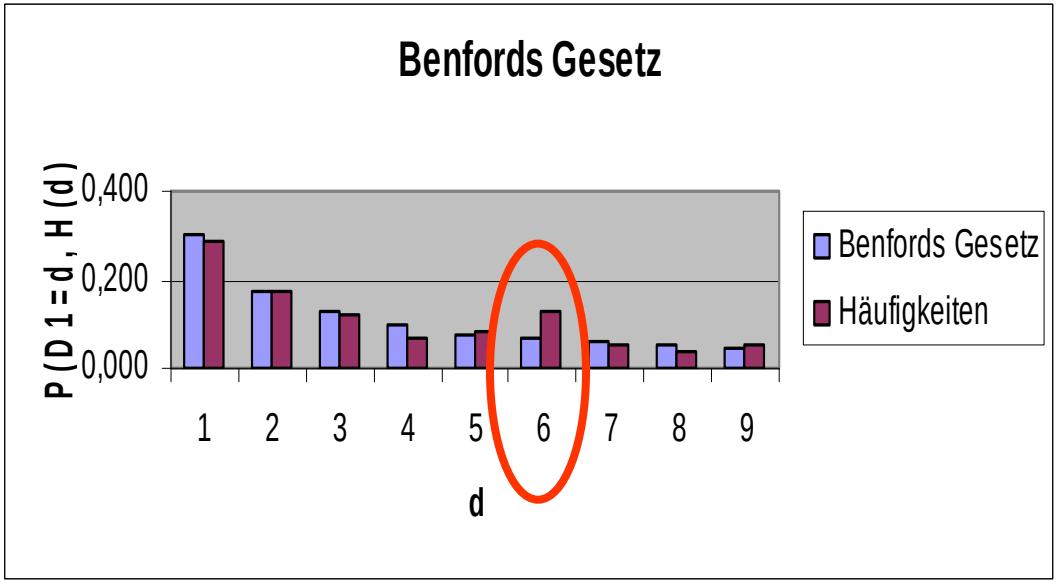
- **Claim** of the auditing group:
  Clerk C issued bad checks by *check manipulations* or fictitious open heart operations with an amount of US$ *6.500,00* per check
- Population: all checks issued by HID in the past

☞ **Data manipulation**

Source: Nigrini (2000)

Data Manipulation (cont.)

Conformance Test by Benford‘s Law



- **Goodness-of-Fit test**:

  $\chi^2_{emp} > \chi^2_{0,95;8} = 2,73$

- **Evidence by histogram inspection**:

  Frequencies $h_d > p_d$ for d =5,6 (cf. US$/check: 6500,00)

IV Data Fabrication

Criminal production of artificial figures driven by power, prestige or profit ('Gier')

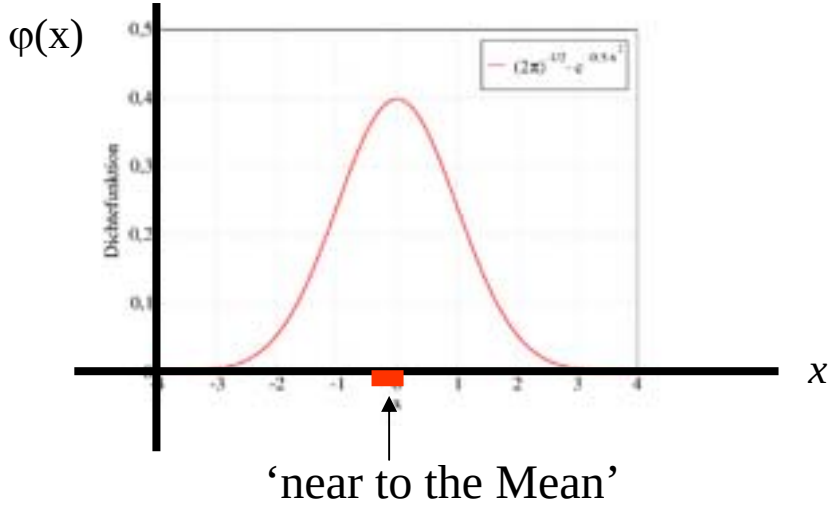
- **Prominent Cases**:
  - Medical Clinic, University of Ulm (1995)
  - Mobile phone human damage research project, Medical University Vienna (2007-2008)
  - 26h/day (!) X-Ray diagnostics of (dead) patients in Ruhrgebiet (~2000)
  - …

Source: http://www.arztwiki.de/wiki/Abrechnungsbetrug (retrieved: 15.02.13)
Adelkofer and Rüdiger (2009)

Data Fabrication (cont.)

Inlier Tests

- **Inliers** are values which are “too near to the mean”
- **Evidence**: ‘*Trickster* try to avoid outliers’
- Log-Score Approach due to Weir and Murray (2011) under i.i.d. assumption



Data Fabrication (cont.)

Inlier Score Test (Weir and Murray (p 167, 2011)

Algorithm 1 **Inlier Test**

Input: Problem size (*n*, *p)*, confidence level (1-α), data matrix $X_{(n \times p)}$

Output: Scores sz_i^2, $ln\ sz_i^2$ for $i=1,2,\ldots,n$

Standardize data by $z=(x- \mu)/\sigma$

1. Let $sz_i^2 = \sum_{d=1}^{p} z_{di}^2$ summed over all *p* variables

   for each test object $i=1,2,\ldots,n$

3. Compare $sz_i^2 >$ (approx.) $\chi^2_{p;(1-\alpha)}$

4. Plot Histogram of $ln\ sz_i^2 = ln \sum_d z^2_{di}$ against $i=1,2,\ldots,n$

Alternative: Mahalanobis Distance

Data Fabrication (cont.)

Inlier Score Test

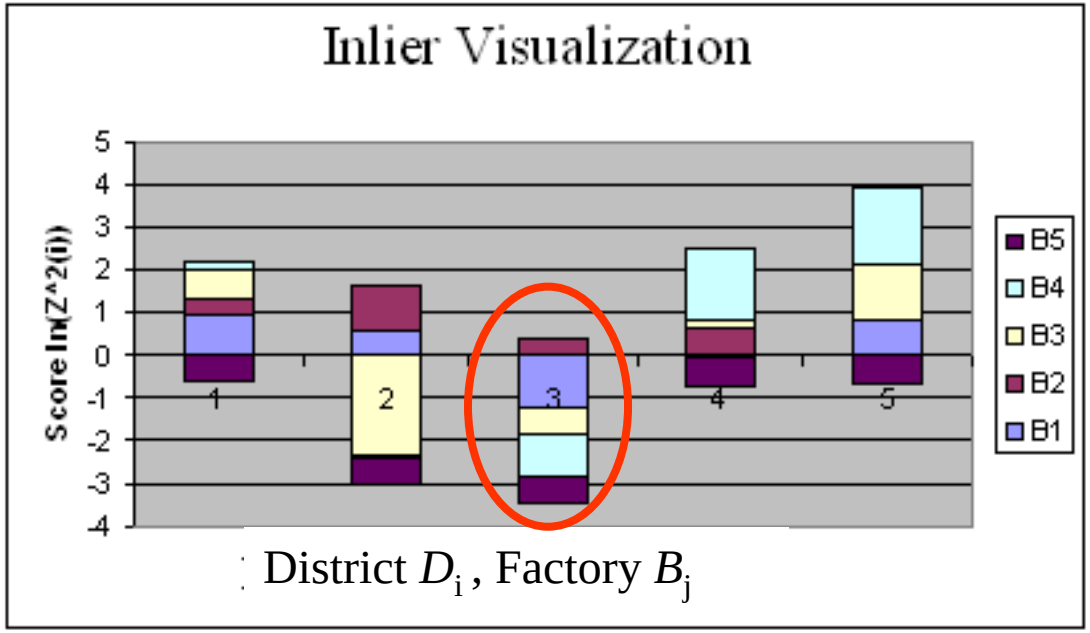
- **Case Study:** Energy consumption (electric power & gas of comp. X) ($p=2$ variables)

- 5 factories B, $j \in J$, within 5 districts D_i, $i \in I$ where $I,J = \{1,2,\ldots,5\}$ ($n_{all}=25$ objects)

- Compute sum of scores $sz^2_{ij} = z^2_1(ij) + z^2_2(ij)$, and plot $ln\ sz^2_{ij}$ for all pairs (i, j).

Data Fabrication (cont.)

Case: Inlier Analysis



Note: The scores *ln* sz^2 of district D_3 are "too small"

Overall mean: $\overline{\ln sz^2} = +0{,}29$

Mean of D_3: $\overline{\ln sz^2}_3 = -1{,}52$

Data Fabrication (cont.)

Outlier Tests

- **Naïve 3σ-Rule**, Cramer (2002)
  - Gaussian Assumption: $X \sim N(\mu,\sigma)$
  - Hypothesis H_0: Observation $x \in \boldsymbol{R}$ generated by $N(\mu,\sigma)$
  - Confidence $(1-\alpha) = 0{,}9973$
    Reject H_0 if $|x-\mu|/\sigma > 3$
  - ML-Estimates $\hat{\mu} = \bar{x}$ *and* $\hat{\sigma}^2 = s^2$

Problem: Masking effect leads to 'Masking' of outliers by distorting estimated mean and standard deviation.

Davies and Gather(2004)

Data Fabrication (cont.)

Weakness of Outlier Tests (Masking Effect)

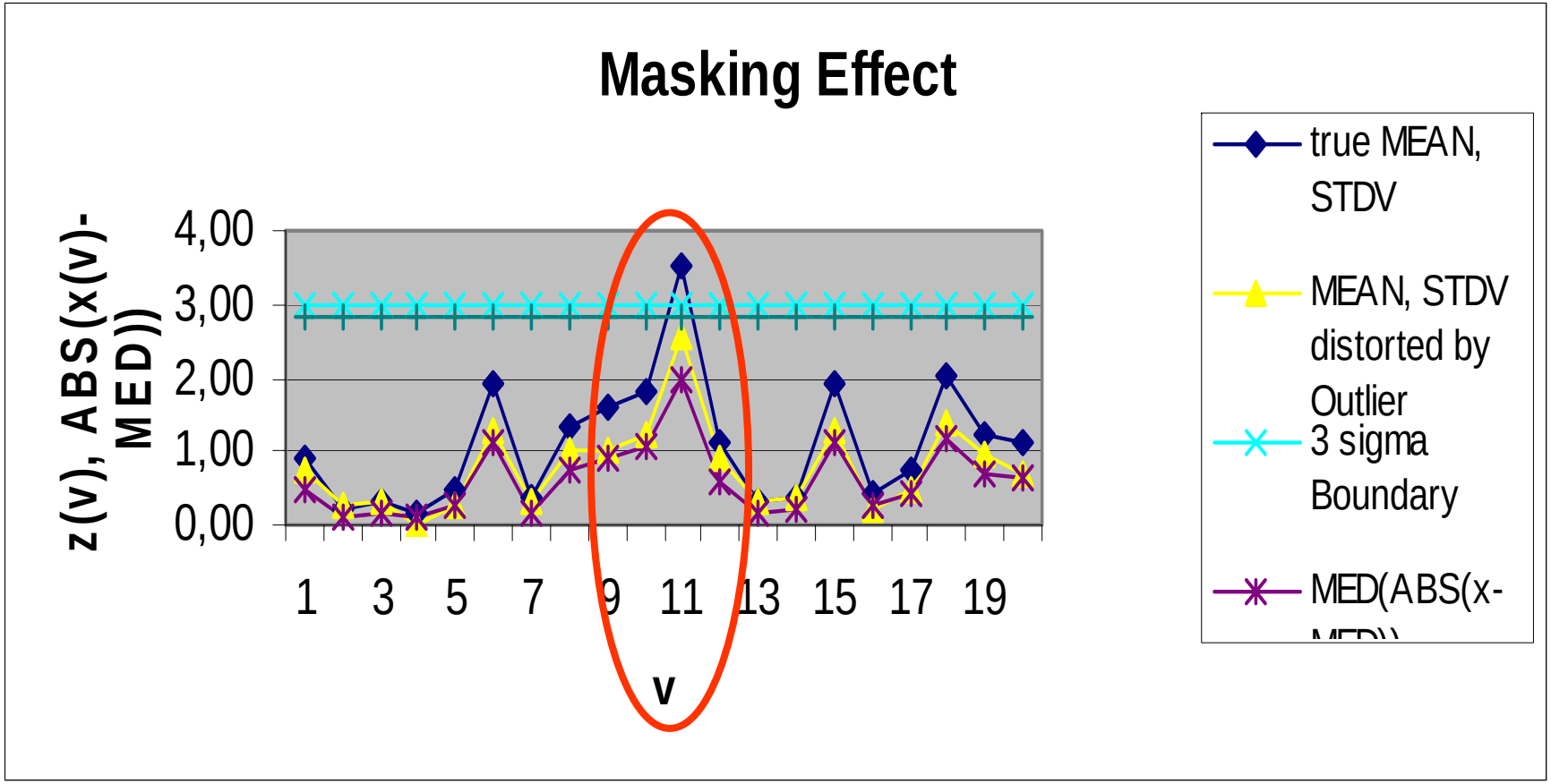
Masking Effect of 3σ-Rule (λ=3)

- 3σ-Rule is ineffective to locate outliers with the same sign at a rate of $1/(1+\lambda^2) = P(I)$ Davies&Gather(2004)
- **Ex.:** Let λ=3 => $P(I)$ = 10% !
- **Conjecture** (Rule of thumb) by Davies&Gather(1993)
  Let (1- α) = 95% and set $\lambda_n = \sqrt{2\ log\ (n)}$
- **Ex.:** n=90 => $\lambda_n \approx 3$.

Problem: Robustness w. r. t. to outliers + Non Normality

Masking Effect

Outliers distort MEAN, STDV by 3Sigma Rule



Source:

Data Fabrication (cont.)

Outlier Test based on *MED*, *MAD* Hampel (1985)

- **λ_H MAD-Rule:**
  Reject $x \in \boldsymbol{R}$ if $\left|x - \tilde{x}\right| \geq \lambda_H MAD((x_\upsilon)_{\upsilon=1,2,\ldots,n})$
  where
  $\tilde{x} = MED((x_\upsilon)_{\upsilon=1,2,\ldots,n})$ is the median

  $MAD((|x_\nu - \tilde{x}|)_{\nu=1,2,\ldots,n})$ is the median of absolute deviations from $\tilde{x}$
- Rule of Thumb of Hampel: λ_H=5,2

Data Fabrication (cont.)

Outlier test based on MAD & Simulation

- **c-MAD Rule** Davies and Gather (2004)
- Reject $x \in \boldsymbol{R}$ if $\left|x-\tilde{x}\right| > c_{n,\alpha_n} MAD((x_\upsilon)_{\upsilon=1,2,\ldots,n})$

  where $\alpha = \alpha_n = 1-(1-\tilde{\alpha})^n$ and

  c_{n,α_n} is estimated by MCMC simulation solving

$$P_{n,\alpha_n}(X \notin \{x \in R \,||\, x-\tilde{x}\,| > c_{n,\alpha_n} MAD(x_1,x_2,\ldots,x_n)\}) \geq 1-\tilde{\alpha}$$

WC-Studies for $\tilde{\alpha} = 0{,}05$: c_{20}=3,02; c_{50}=3,28, c_{100}=3,47

Davies, Gather(1993)

Data Fabrication (cont.)

multivariate outlier tests

- Let $(\mathbf{x}_\nu \in \boldsymbol{R}^p)_{\nu=1,2,\ldots,n}$ be a p-variate data set
- Assume $\mathbf{X} \sim N(\mu, \Sigma)$ with $\mu \in \boldsymbol{R}^p$ and $\Sigma \in \boldsymbol{R}^{p\times p}$ known
- Mahalanobis distance

  Reject $\boldsymbol{x} \in \boldsymbol{R}^p$ if $(\boldsymbol{x}-\mu)^T \Sigma^{-1} (\boldsymbol{x}-\mu) > \chi^2_{\boldsymbol{p};1-\alpha}$
- Problem: - (μ, Σ) unknown
  - Non-Normality
  - Outlier Robustness
- Estimate critical values $\chi^{*2}_{p;1-\alpha_n}$ for a given (p, α, n) by MC -Simulation

Lenz , Rödel (1991) Davies, Gather (2004)

V Next Steps & Perspectives

- **Hot topic** ~ 1/ethic
  - in industry, science, public sector / media
  - in clinical / pharmaceutical research, health care
- **State-of-the-art data analytics**
  - Robustness (inliers / outliers / non-normality)
  - Missing units and values
  - Sequential testing
- **Principle of transparency**:
  - Deposit of data & paper & experiments/observations protocol
  - Reproducibility & Repeatability of designed experiments / observations
  - Checks of scientific studies (devices, methods) by independent authorities
  - Transparency in science by *ResearchGate* (start-up, Berlin) etc.

Vague boundaries between terms

- data appraisal, cheat, deception, fraud, scouting

Enough is not enough!