MONTAN UNIVERSITÄT
WWW.UNILEOBEN.AC.AT

Fake Data – Rationale, Detection and Implications

Alessandra Rachetti & Wolfhard Wegscheider

Fraudulent „Fake“ Data - Facts

- About 2 % of reseachers have admitted to faking data at least once in their careers.
- Blurred boundaries between innocent error, misunderstandings, avoidable faults, intentional “bending” and massive falsification.
- **Fraud implies intention to cheat.**

Research Misconduct
Official Definition

"fabrication, falsification or plagiarism (FFP) in proposing, performing or reviewing research or in reporting research results"

US office of Science and Technology Policy (OSTP)

Fake Data - Characteristics

- **Falsified, manipulated data:** observations that do not fit the desired results are deleted or amended and the variability as a whole is reduced.
- **Fabricated, invented data:** very little variation, total absence of outliers, and because of human intervention, a pattern of digit preference. Invented distributions tend to be flat, evenly spread over a limited range.

Fake Data - Objective

- **Falsification / Bending / Data manipulation:** to achieve a desired result or increase the statistical significance of the findings and affect the overall scientific conclusions, to achieve publication, or to produce results confirming a particular theory.
- **The object of most falsifications is to demonstrate a "statistically significant" effect that the genuine data would not show.**

Fake Data - Objective

- **Fabrication / Invention of data** for non-existent or incomplete cases (in clinical studies, market research), usually for financial gain.

- **The most serious cases of fraud are those in which there is an expectation of gain in terms of prestige, advancement, or money.**

- Almost never occurs in fields like physics, astronomy and geology.

David Goldstein, 2005

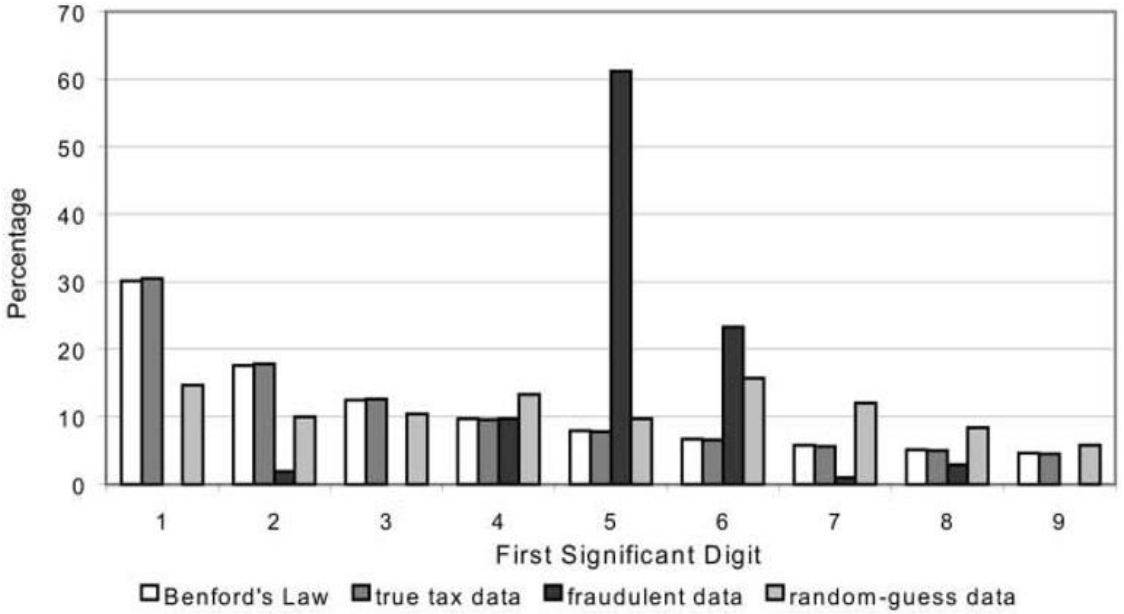
Statistical Methods for Detecting Fake Data

- Look at digit distribution and preferences
- Look at variances, standard deviations, percentile ranges, range, kurtosis
- Multivariate associations - Look for relationships that should exist

Digit Preference – First Digit Benford's Law

- Runs against intuition
- mainly for counting and measurement data
- not for assigned data or numbers influenced by human thought

Benford's Law -Examples



from Theodore P. Hill, 1998

Digit Preference – Terminal Digit

- Terminal Digits are supposed to be uniformly distributed as they are expected to contain mostly random measurenment error.
- Humans instinctively do exhibit digit preferences
- Well suited for graphic methods of detection, Histogram, Stem & Leaf plot

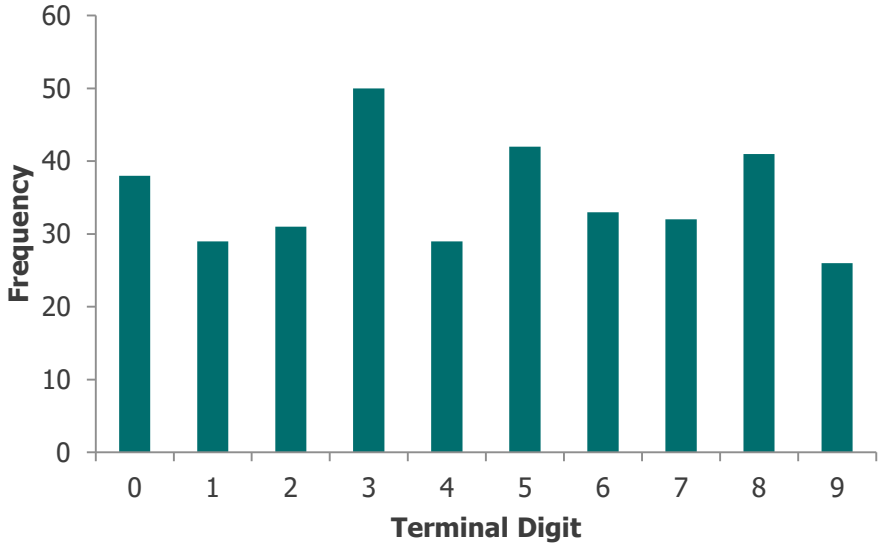
Digit Preference – Stem and Leaf Plot

```
14 : 2
14 : 555
14 : 67777
14 : 889
15 : 000000111111
15 : 22222222222233333333333333333
15 : 44444444445555555555555555555
15 : 66666666666666666667777777777777777777
15 : 88888888888888888888888888888899999999999999999
16 : 00000000000000000001111111111111111111
16 : 22222222222222222233333333333333333333333333333
16 : 44444444444444444555555555555555555
16 : 666666666667777777
16 : 88888899999999
17 : 00000000000111
17 : 333
17 : 4
17 : 67
17 : 88
```

heights of 351 (elderly) women.

Data source: http://what-when-how.com/statistics/skewness-to-systematic-review-statistics/

Digit Preference – Histogram



heights of 351 elderly women.

Fake Data / Possum Example

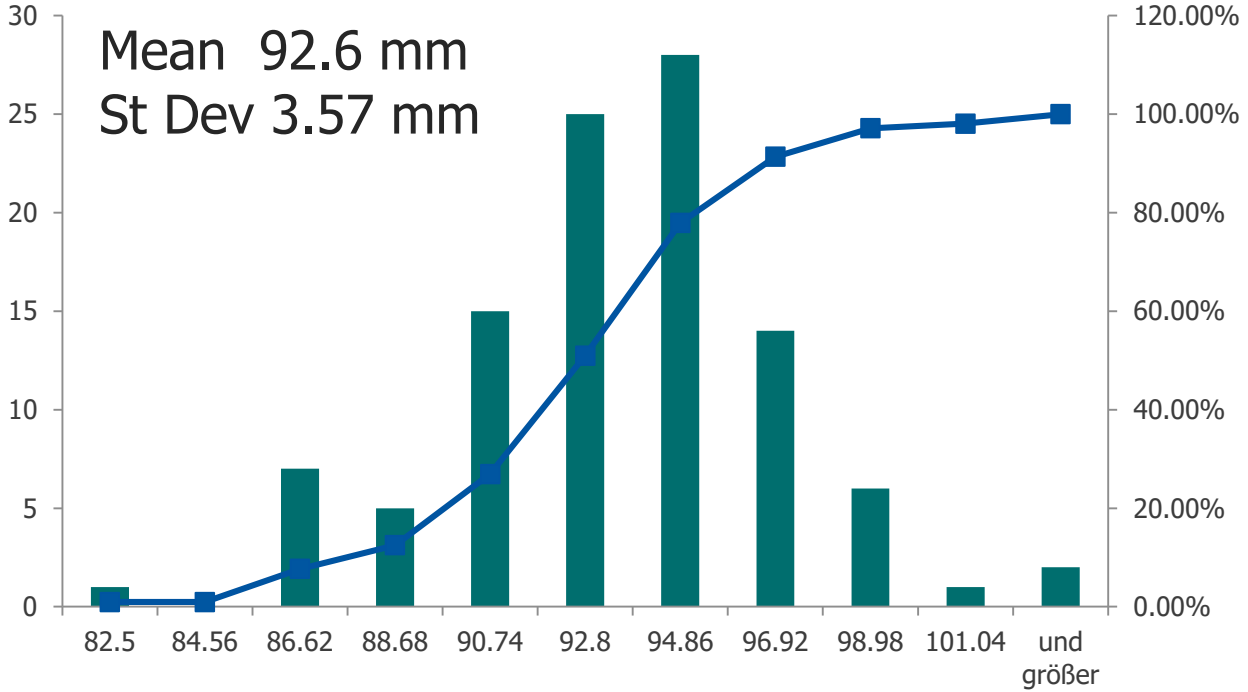
- 104 mountain brushtail possums
- 9 morphometric measurements

- Head length, skull width

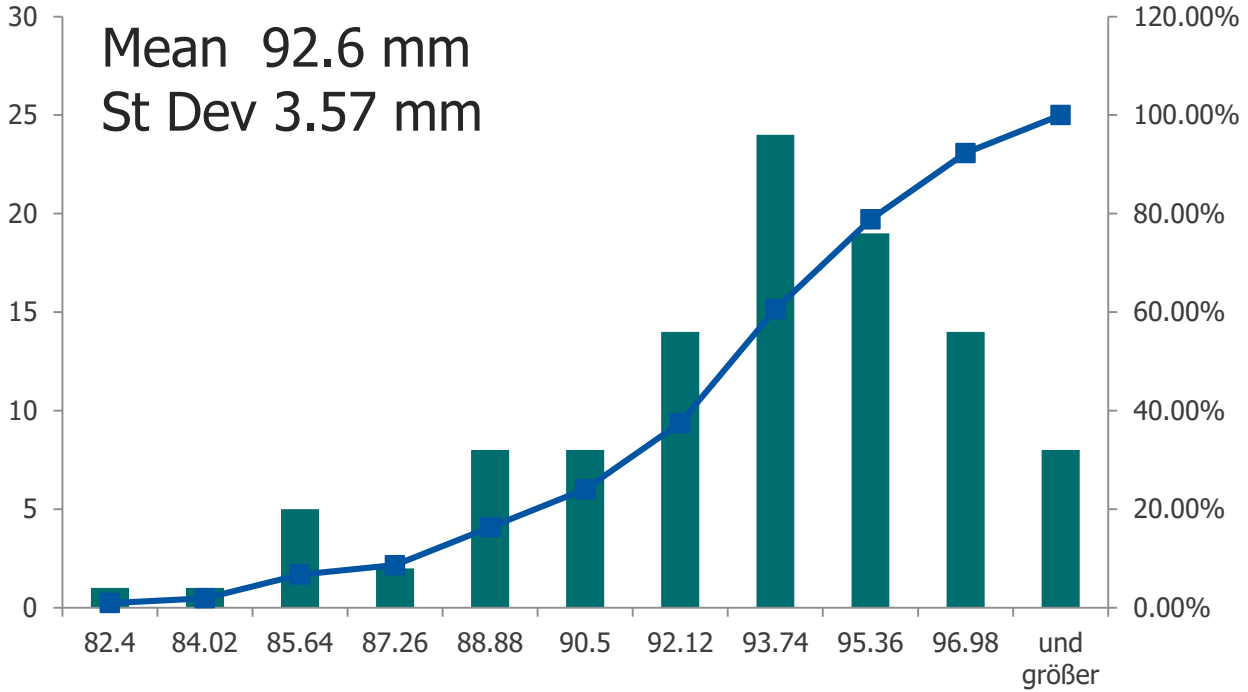
Picture source: http://www.environment.nsw.gov.au/topics/animals-and-plants/native-animals /native-animal-facts/brush-tailed-possum

Data source: Lindenmayer, D. B., Viggers, K. L., Cunningham, R. B., and Donnelly, C. F. 1995. Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala). Australian Journal of Zoology 43: 449-458.
https://vincentarelbundock.github.io/Rdatasets/datasets.html

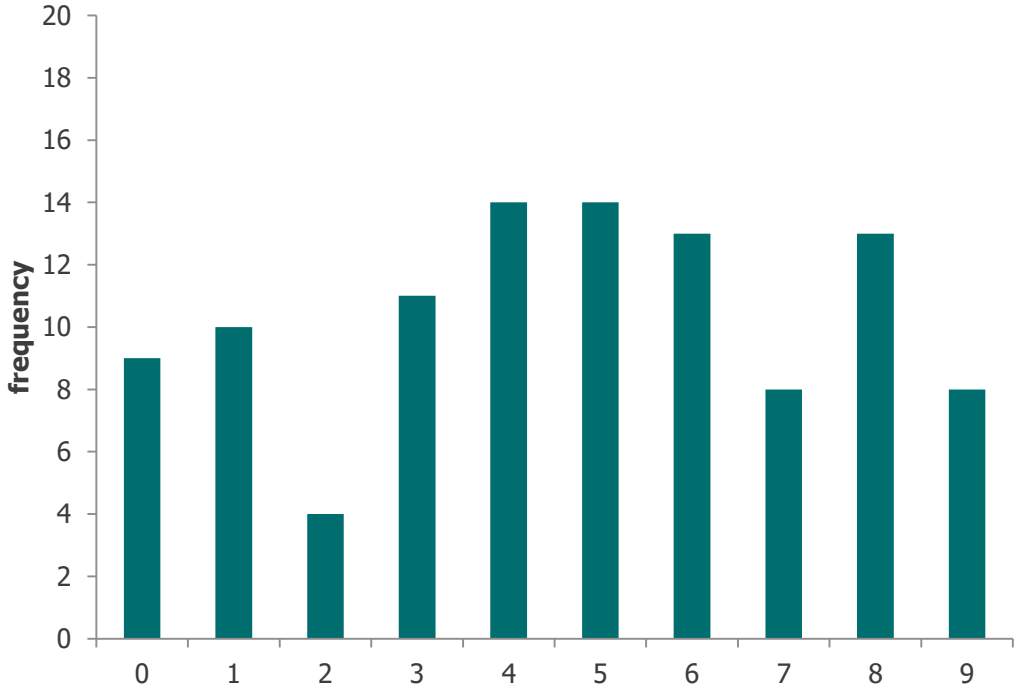
Possum head length / true data Histogram



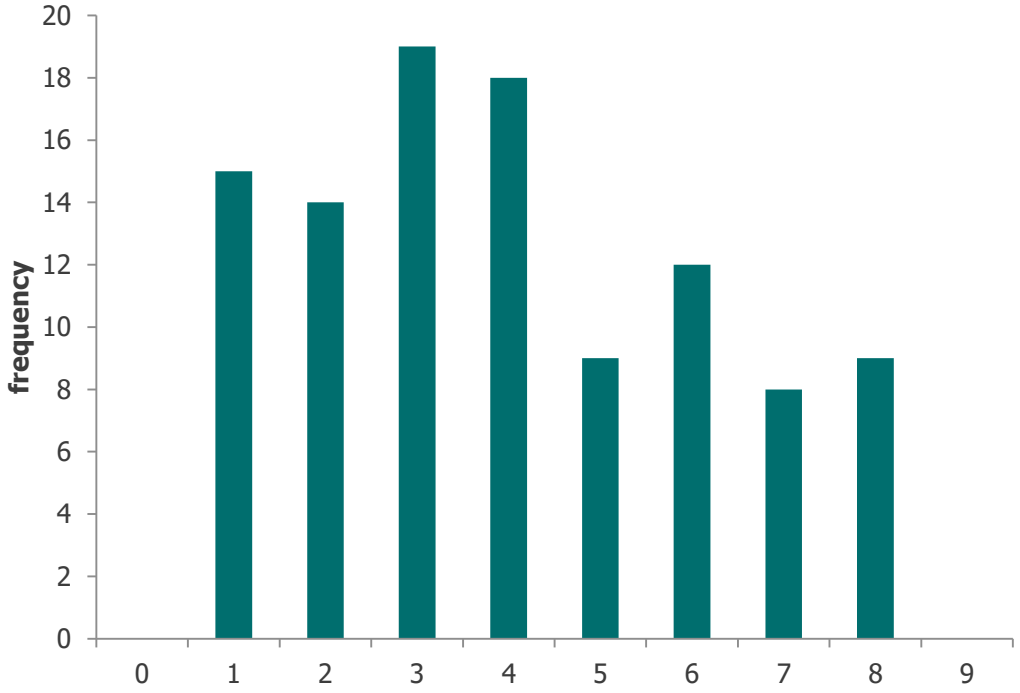
Possum head length / fake data Histogram



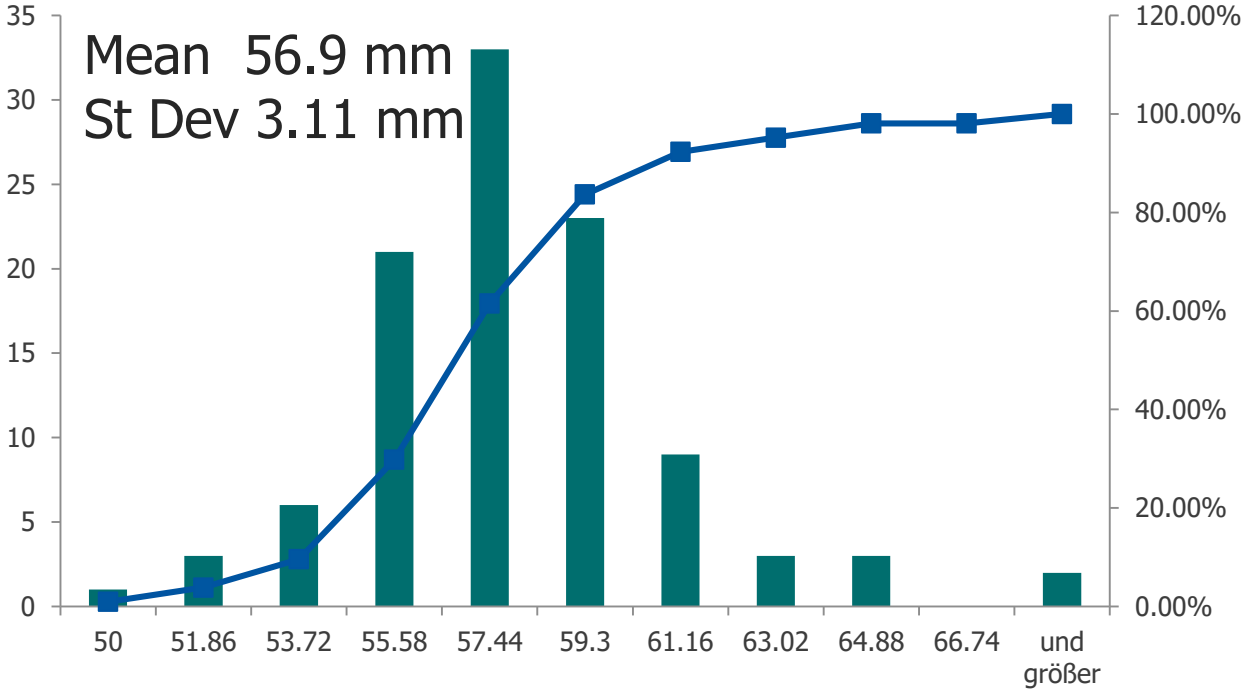
Possum head length / true data terminal digit distribution



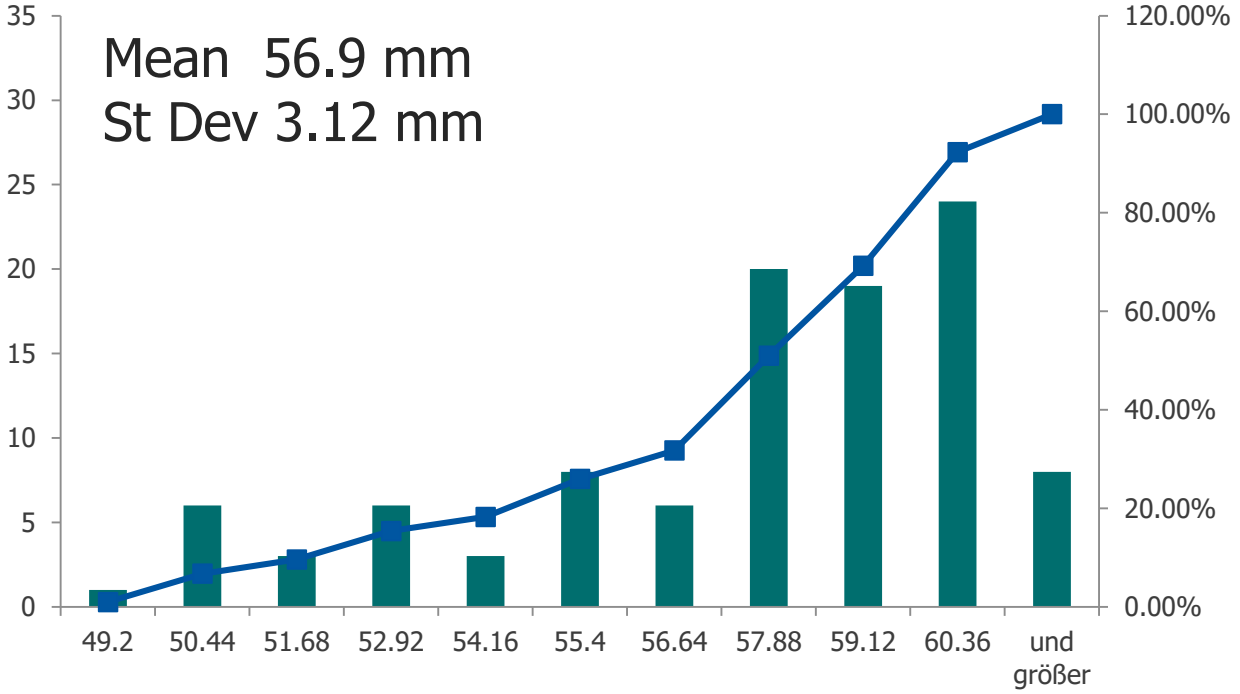
Possum head length / fake data terminal digit distribution



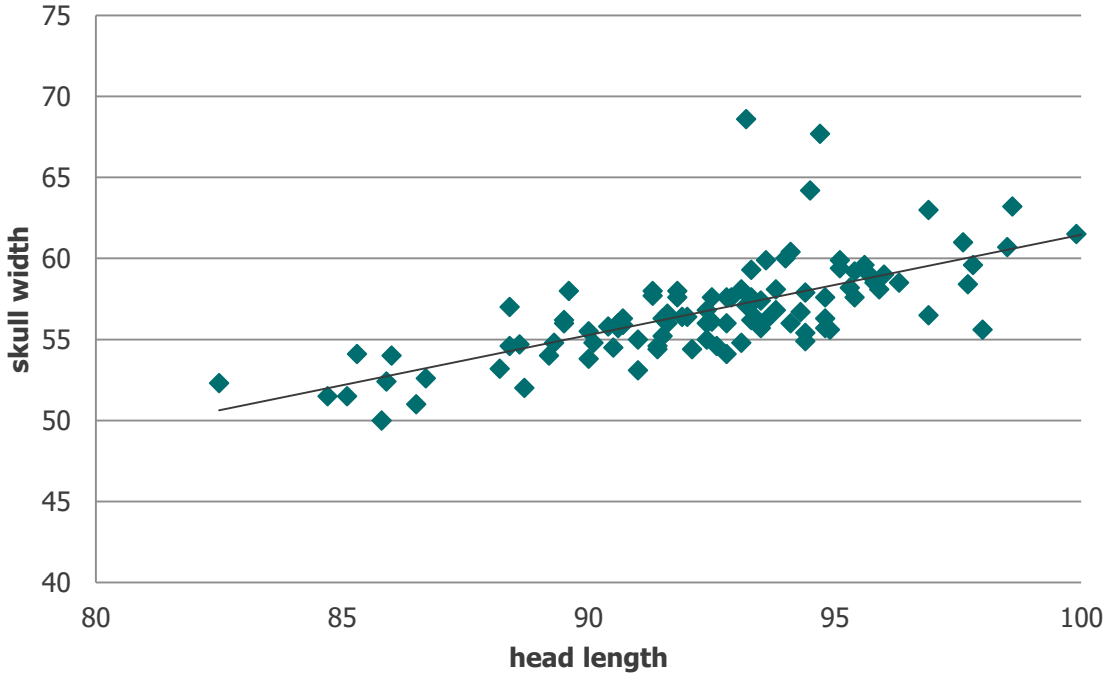
Possum skull width/ true data Histogram



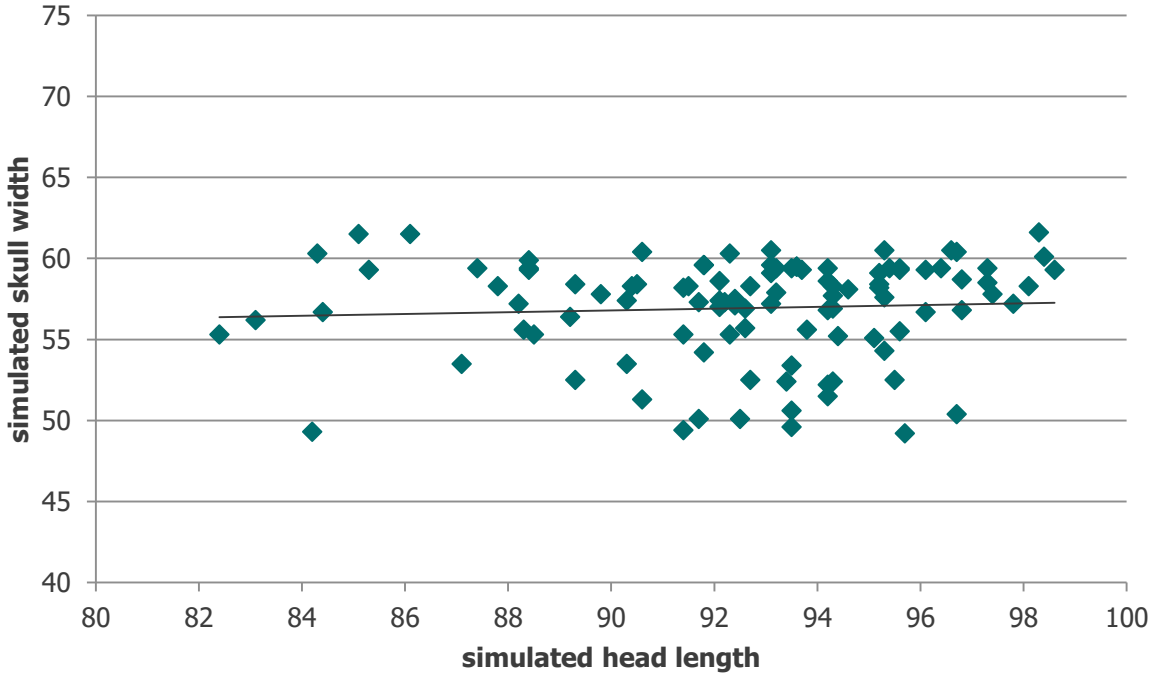
Possum skull width/ fake data Histogram



True data: Possum skull width / head length



Fake data: Possum skull width / head length



Fake Data – Risk Factors

- Career pressure
- „Knowing the answer"
- Working in a small team
- Working in a field where individual experiments are not expected to be precisely reproducible

David Goldstein, 2005

Fake Data - What to do ?

Increase risk of exposure

- Peer review
- Full access to original data
- Public data repositories
- Better education of statisticians
- Devote a significant amount of research funds for replications
- Automated scanning of publications

Antonakis' 5 scientific diseases

- **Significosis**, an inordinate focus on statistically significant results
- **Neophilia**, an excessive appreciation for novelty
- **Theorrea**, a mania for new theory
- **Arigorium**, a deficiency of rigor in theoretical and empirical work
- **Disjunctivitis**, a proclivity to produce large quantities of redundant, trivial and incoherent works

“If you copy from one author, it’s plagiarism, but if you copy from many, it’s research.”

Wilson Mizner

Resources

- Fanelli, D., How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. Plos One 2009, 4, e5738.
- OSTP Federal Policy on Research Misconduct, cited from Martinson, B. C., Anderson, M. S., de Vries, R. (2005). Scientists behaving badly. Nature, 435, 737-738 doi:10.1038/435737a
- Fanelli, D. Redefine misconduct as distorted reporting (2013). Nature 494, 149. doi:10.1038/494149a
- Hill, T. P. (1998). The first digit phenomenon. *Amer. Scientist* 86:358–363.
- Goldstein, D. (2005). Conduct and Misconduct in Science. Retrieved from http://www.physics.ohio-state.edu/~wilkins/onepage/conduct.html
- Evans, S. (2001). Statistical aspects of the detection of fraud. In Lock, S., Wells, F., Farthing, M. (Eds), Fraud and Misconduct in Biomedical Research (186-203). London, England: BMJ Books,
- Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. The Leadership Quarterly, 29, 5-21. http://dx.doi.org/10.1016/j.leaqua.2017.01.006