



Can Data speak for itself (themselves)?

----- basic statistical principles that are still valid in the era of big data

Prof W K Li

Department of Mathematics and Information Technology

Faculty of Liberal Arts and Social Sciences

The Education University of Hong Kong



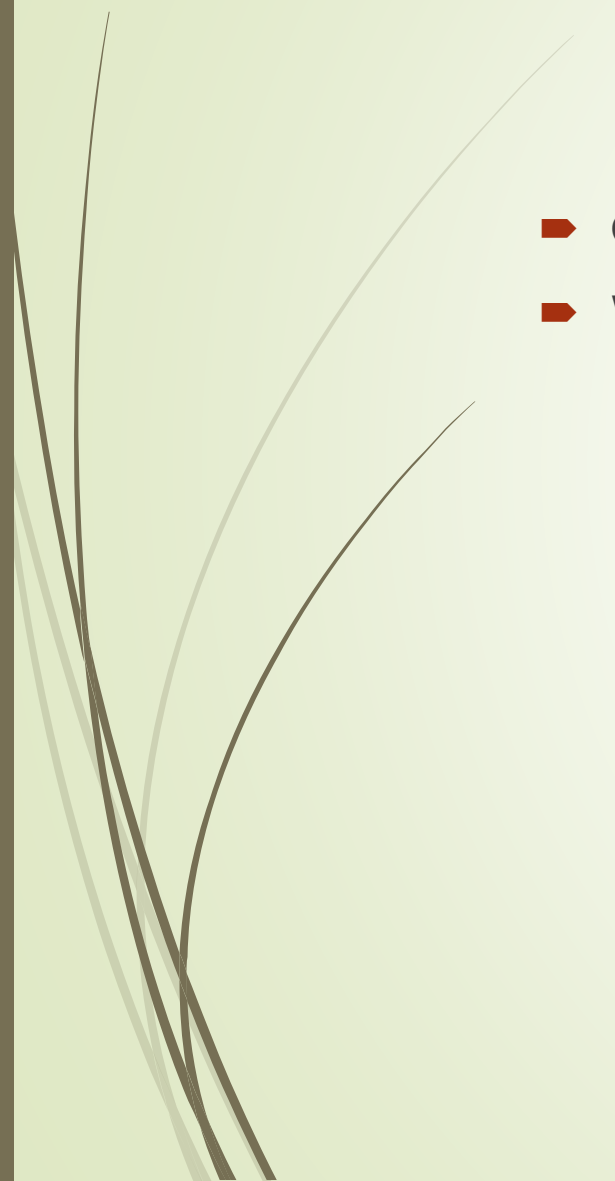


(I) Biased sampling remains biased --- the issue of self selection

- Ann Landers was a US advice columnist of Chicago Sun-Times. She once asked the readers of her advice column: “If you had it to do over again, would you have children?”
- Results – she got nearly 10,000 write-in responses and nearly 70% saying “NO!”
- But would you believe it?
- What type of people will response?
- Issue – for voluntary response sample only those who feel strongly about an issue, especially those with strong negative feelings, are more likely to respond. (Moore and Notz, 2009)



The above is an example of Self Selection Bias

- Can you think of other examples?
 - What is your re-action to most signature campaigns? Why?
- 



Selection bias is ubiquitous !

- Selection bias cannot be resolved by a larger sample (bigger data set).
- Some surveys conducted by *Playboy* on its readers in the 80's have some 100,000 responses.
- A LARGE SAMPLE!
- Yet this only represents 2% of the readership. (Asher 2012)
- Double problems:
 - 1. The readers may not be representative of the whole population
 - 2. Those who response may not represent the whole readership.



Healthy food and longevity

- Some TV programs on healthy living will make recommendations on what to eat for longevity. E.g. eating wasabi/okra.
- They will often interview some healthy local old folks who have consumed the recommended food for years to support their claim
- What kinds of issues are there?
 - 1. There may be other factors behind (genes, living style, environment, etc.)
 - 2. Only the survivors can be interviewed. (survivorship bias!)
 - 3. Reversed causality! They choose such food because they do not have other choices.



Selection bias

- ▶ What other examples can you think of?

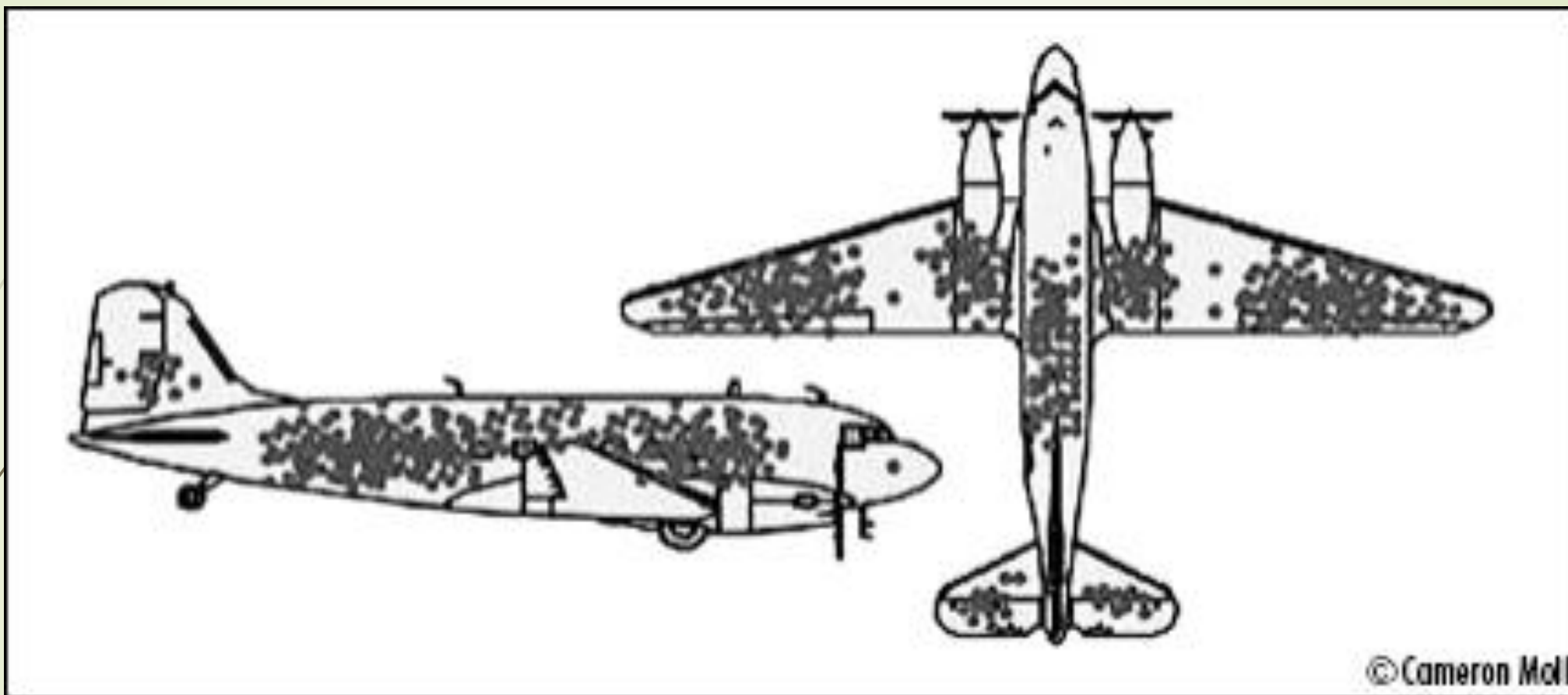


What does no data imply?

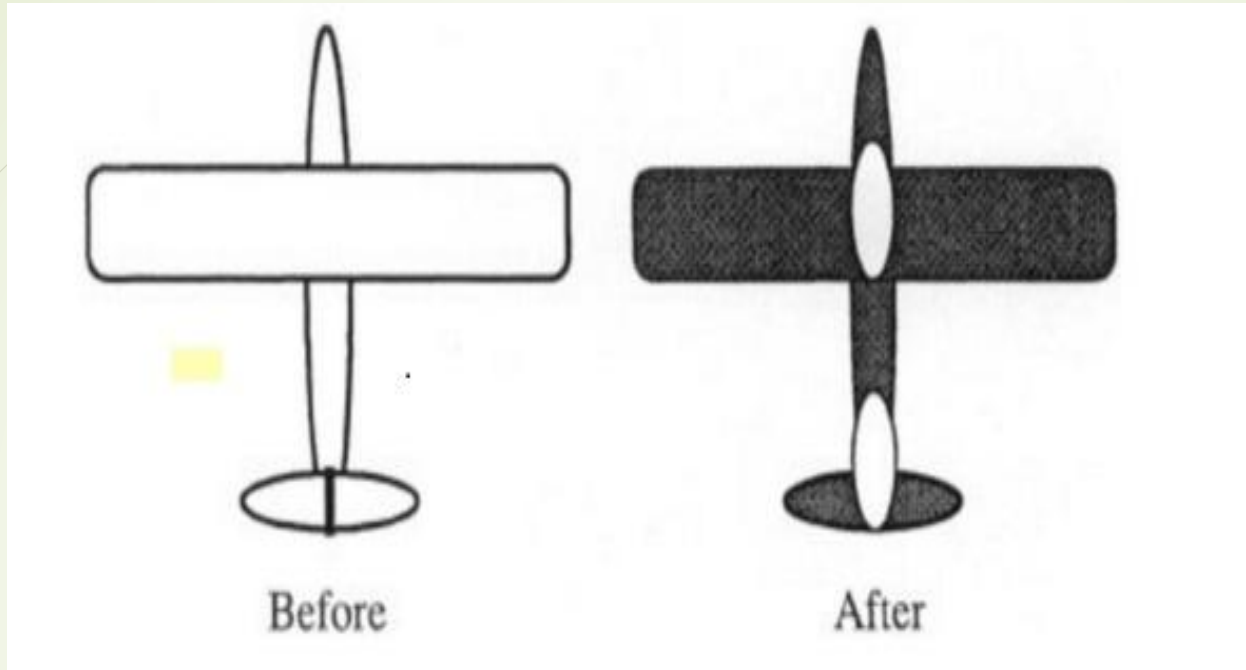
- During World War II the allies sent bombers to bomb strategic targets on the Nazi occupied Europe almost daily.
- However, losses of planes and pilots were heavy due to anti-aircraft flaks.
- To reduce losses it was suggested to put more armour plates on the fuselage of the bombers where it would receive the most fatal hits.
- Bullets holes on returning bombers were located and marked on a model of the plane
- Question: where to put more additional armour

<https://clearthinking.co/survivorship-bias/>






©Cameron Moll



Where should you put more armour?

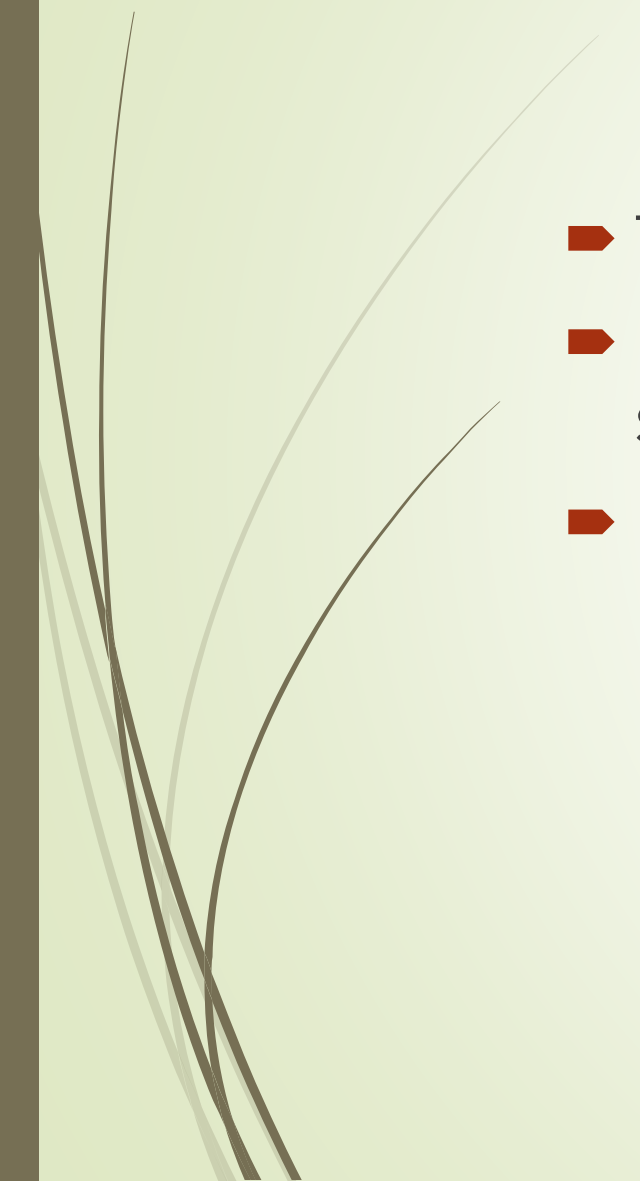


Should more armour be put on where you have the highest number of bullet counts?

- Observation: What are those planes that can provide information about the holes?
- Answer: Those that can return to base!
- Those planes that were hit at the vital locations never made it back and therefore there were no information available.
- Survivorship bias again!
- Abraham Wald, a Hungarian Jewish statistician was the first to note
- the problem and recommended reinforcement on locations without bullet holes.
- And this is the right thing to do!



Beware of selection biases!

- Think carefully on reading the results of a poll/survey.
 - Biased samples/surveys/polls remain biased regardless of sample size!
 - Read newspapers slowly!
- 

(II) Spurious correlation everywhere!

Example: Children's traffic accidents and the consumption of ice-creams (Inversen & Gergen, 1997)

- Observe:
- In some months a high consumption of ice-cream are associated
- with higher traffic accidents involving children and in months with
- lower ice-cream consumption, fewer accidents occur
- Question:
- Does ice-cream consumption cause more traffic accidents for
- children?
- (or vice versa?)

Higher Temperature

```
graph LR; A[Higher Temperature] --> B[more ice cream consumption]; A --> C[more accidents (summer holidays)];
```

**more ice cream
consumption**

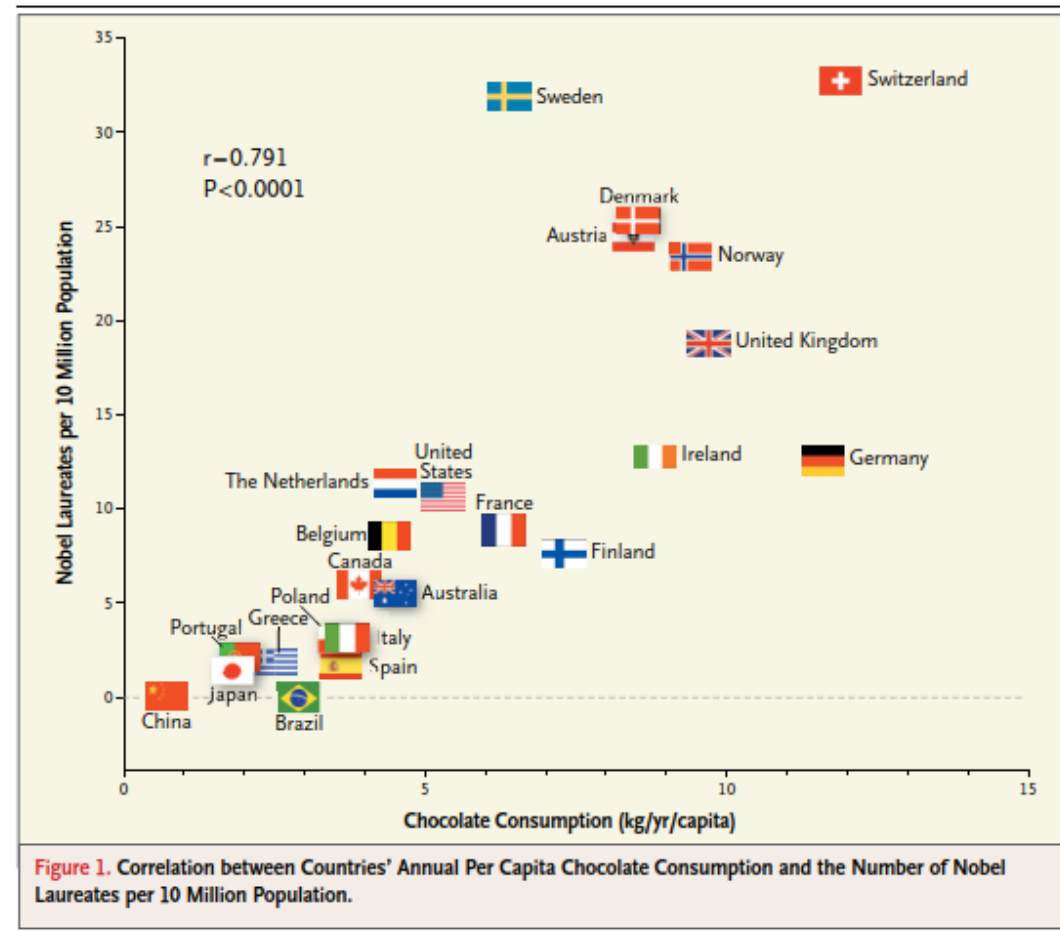
**more accidents
(summer holidays)**



Correlation does not imply causation!

- A strong correlation (association) between two variables does not always imply that changes in one variable cause changes in the other.
- Often, the relationship may be because of influences of another variable(s) lurking in the background.
- Spurious correlation refers to a high correlation that is actually due to some third factor.
- Causation is best investigated by carefully controlled randomized experiments. (beyond the scope of this talk)

Chocolate consumption and Nobel Laureates (New England Journal of Medicine, F H Messerli, 2012)





Evidence of Causation

- The association is strong.
 - The association is consistent.
 - Higher doses are associated with stronger responses.
 - The alleged cause precedes the effect in time.
 - The alleged cause is plausible. For example, there is experiment evidence.
-
- Example: smoking and lung cancer.
 - The evidence for causation is *overwhelming* but it is not as strong as evidence provided by well-designed experiments.



Does Big Data Imply The End of Theory ?

“The **Data Deluge** Makes the Scientific Method Obsolete with enough data, the numbers speak for themselves. Correlation supersedes causation, and science can advance even without coherent models, unified theories!”

According to C. Anderson (2008), former editor-in-chief of Wired Magazine

Results of 200 tosses of a fair coin: one of the data sets below is a fake. Can you tell? (Bennet, Briggs and Triola 2018)

Data Set 1 (H = heads; T = tails)

H T H T H H T T T T T H T H T T T T H
H H T T T T H T T H T T H H H T T H H T
H T H T H H H H T T H T H T H H H H T H
T T H H T T H H H T T T T T H H H T H T
T H T H T T H H T T H T H T H H T T H T
T T H T H T H H T T T H H T T H H H H T
H T H H T H T T T T H T T T H T H T H H
T H T H H H H T H T H H H T T T H T T H
T H T T T H T H H T H H H H T T T H H T
T H T H H T T H T H H H T H H T H T T H

Data Set 2 (H = heads; T = tails)

T H H T T H H T T H T H H T T H H T H T
H T T H T T H H T T H T T T H H T H T H
H H T T H T H H T H T T H T T H H T T H
H T H H T T H T H T H H T H T H T H H T
H T H T T H T T H H T T H H T H T H H T
T H T H H T T H T H T T H T T H T H H T
H T H T T H T H H T H T H T H T H H T H
T T H T H T H H T T H T T H T H H T H H
H T H H T T H T H T H T H H T H T T
T H H T H T H H T H H T T H H T H T H T



Some hints:

Features of data set 1:

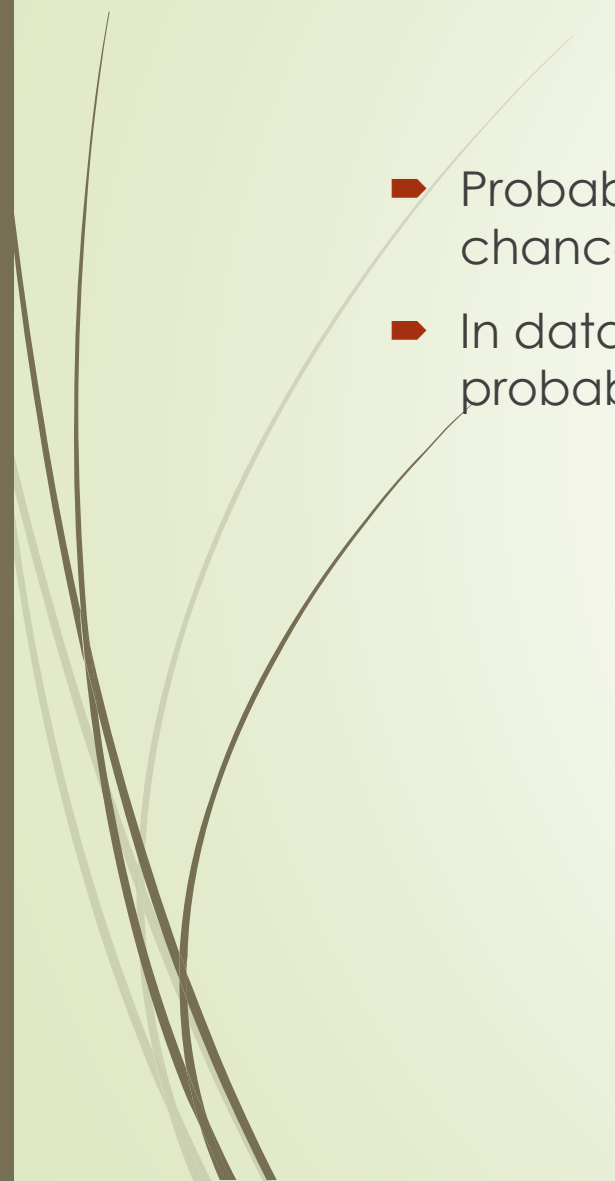
- 97 H 103 T
- Two cases of 6 T in a row
- Five cases of 4 H in a row
- 3 cases of 4 T in a row

Features of data set 2:

- 101 H 99 T
- No cases of more than 3 H or 3 T in a row




Answer: data set 2 is fake!

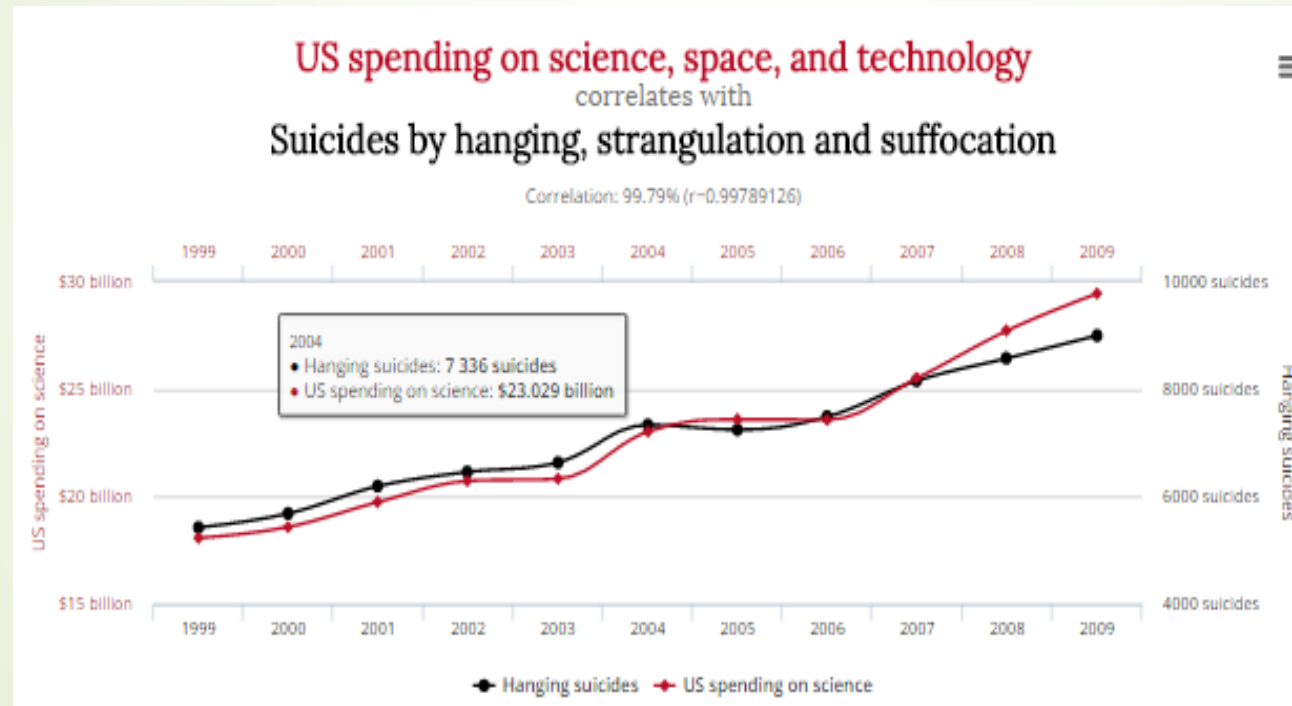
- Probability of getting 6 heads or tails in a row is 1 in 64. With 200 tosses the chance of observing one case of 6 heads or tails is quite high!
 - In data set 2 there is no case of 4 Heads or 4 Tails in a row which has a probability of 1 in 16.
- 



Lessons to be learnt:

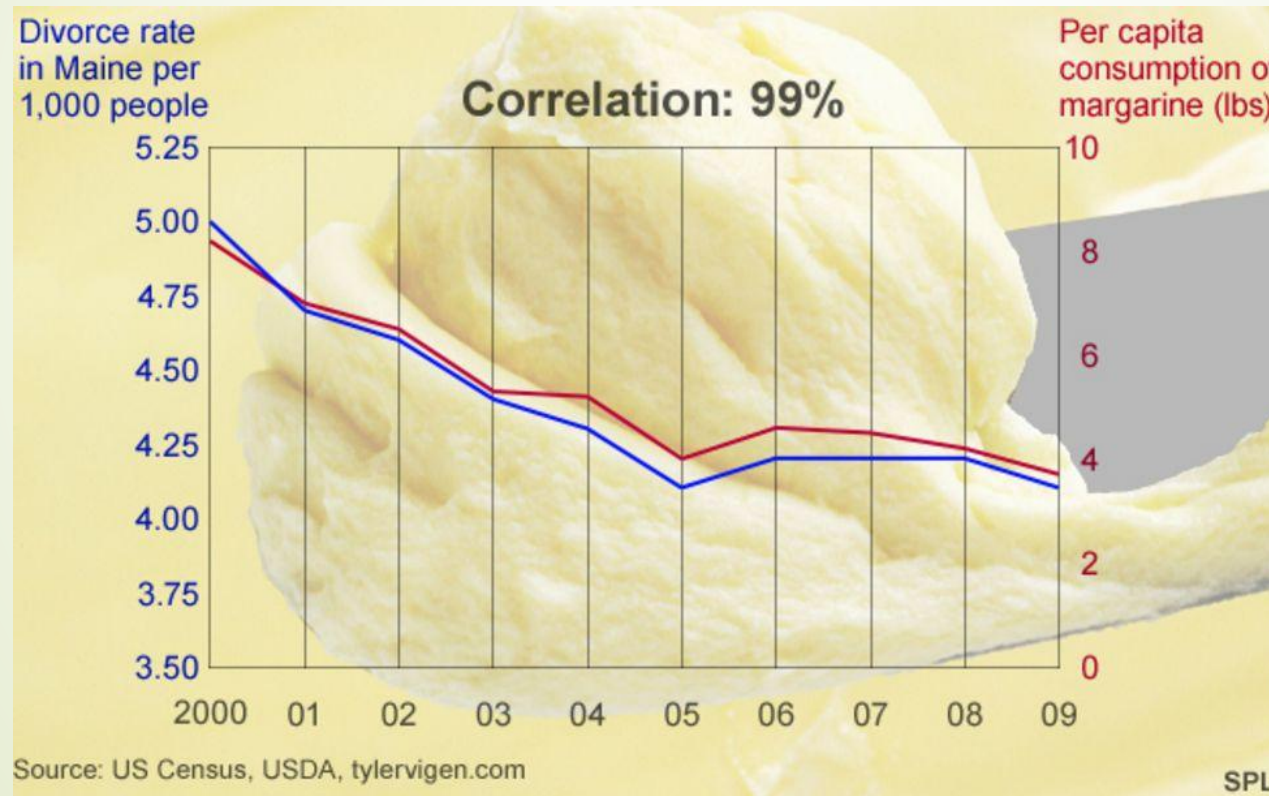
- ▶ Lesson 1: even in randomly generated data sets there will be “some (spurious) patterns” according to an untrained eye!
 - ▶ Lesson 2: the faked random sequence is ‘too good(regular/clean) to be true’!
- 

Many Examples of (absurd) Spurious Correlations
<https://www.statisticshowto.datasciencecentral.com/spurious-correlation/>



Absurd Correlations!

<https://www.bbc.com/news/magazine-27537142>






The Deluge of Spurious Correlations in Big Data

by Cristian S. Calude & Giuseppe Longo, 2016

Found Sci DOI 10.1007/s10699-016-9489-4

- They prove that very large databases have to contain arbitrary correlations.
- These correlations appear only due to the size, not the nature, of data.
- They can be found in randomly generated, large enough databases, which implies that
- Most correlations are spurious. Too much information tends to behave like very little information.
- The scientific method can be enriched by computer mining in immense databases,
- but not replaced by it.



With data, enough computing power and statistical algorithms patterns will be found. But are these patterns of any interest? Not many of them will be, as spurious patterns vastly outnumber the meaningful ones.

Data will never speak for itself, we give numbers their meaning, the Volume, Variety or Velocity of data cannot change that.

(Do Numbers really speak for themselves with big data? John Poppelaar)



References:

- <https://hbr.org/2015/06/beware-spurious-correlations>
- <https://www.bbc.com/news/magazine-27537142>
- <https://www.statisticshowto.datasciencecentral.com/spurious-correlation/>
- <https://blogs.dnvgl.com/software/2017/11/thinking-outside-of-the-box/>
- <https://clearthinking.co/survivorship-bias/>
- <http://john-poppelaars.blogspot.fr/2015/04/do-numbers-really-speak-for-themselves.html>
- <https://www.youtube.com/watch?v=Kol6pnhfkRQ>
- Asher, H. (1995). Polling and the Public: What Every Citizen Should Know (3rd ed.). CQ Press.
- Bennett J., Briggs W. L. and Triola M. F. (2018) Statistical Reasoning for Everyday Life. Boston: Pearson
- Calude C. S. and Longo G. (2016) The Deluge of Spurious Correlations in Big Data. Foundation of Science
- Moore, D. S., & Notz, W. (2009). Statistics: Concepts and Controversies (9th ed.). New York: W. H. Freeman.
- Smith G. (2014) Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics



Thank you for your attention!