How to Lie with Statistics

a book by Darrell Huff, 1954

Peter Hügel
Seminar “How do I lie with statistics?”
Supervisor: Prof. Dr. Ullrich Köthe
Heidelberg, 17.10.2019

Heidelberg Collaboratory
for Image Processing

Outline

- Introduction
- Simple Ways to Lie
    - Selection Bias
    - The Average
    - Missing Figures
    - Charts and Pictographs
    - Semi Attached Figures
    - Correlation and Causation
- Causes for Lies
- Identifying Lies

Introduction

- Statistics are all around us
  - Actual statistics
  - Assumptions we make based on available information
- There can be a lot more or a lot less to the statistics we are exposed to
- They can distort reality while technically being correct
  - Incompetence
  - Ill intent
  - Misinterpretation

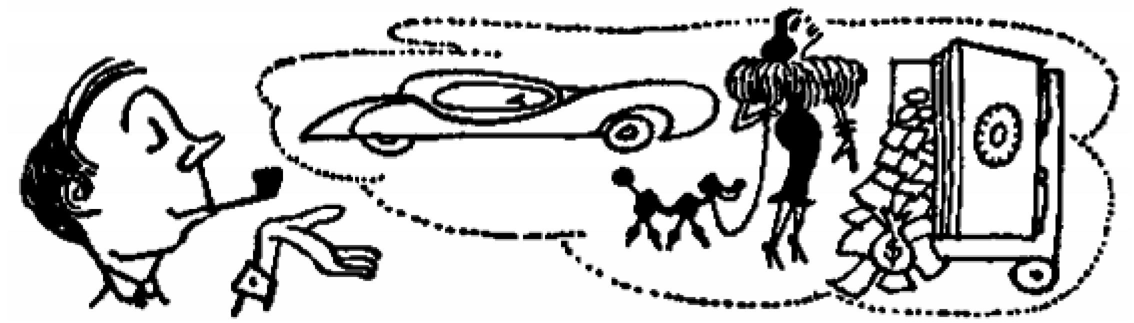
Introduction – Nitrate Levels in Groundwater

- Germany monitors nitrate levels in accordance to an EU conservation directive from 1991
- “More and more nitrate in groundwater”[1]
- “From 2013 to 2017 the average nitrate concentration in the top 15 polluted regions has increased by 40mg/l”
- What could be wrong with this statement?

[1] Rheinische Post, 8.8.2019, https://rp-online.de/wirtschaft/immer-mehr-nitrat-im-grundwasser-gefahr-fuer-mensch-und-natur_aid-44825553

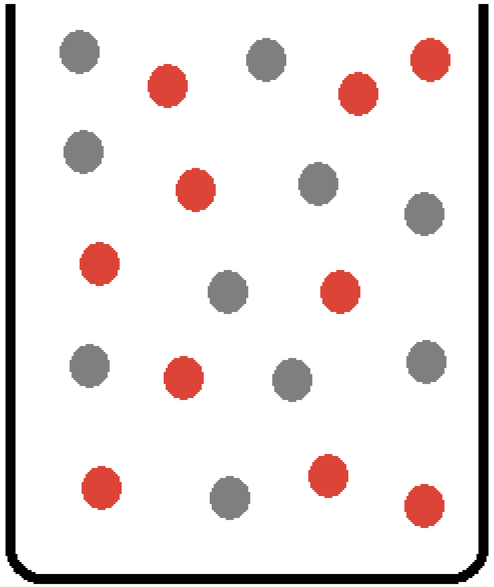
Simple Ways to Lie – Selection Bias

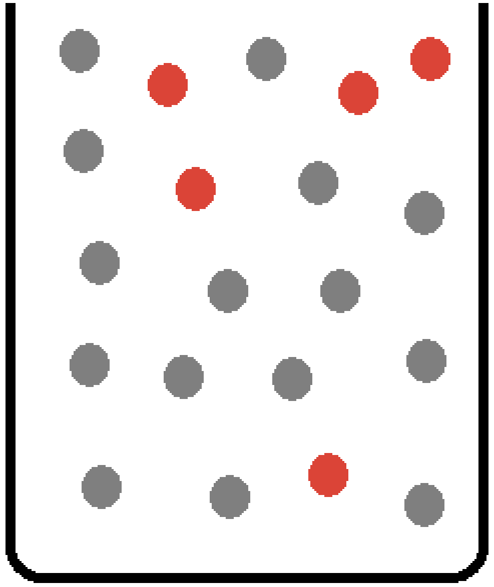
- “The average Yaleman, Class of 1924 makes $24,111 a year.” Time Magazine
- How was this number derived?
  - Graduates of that year had be asked:
    - How were they located and contacted?
      - Are all equally likely to be found?
      - Are all equally likely to respond?
    - Do they answer honestly?
      - Exaggerate?
      - Understate?
- → Selection bias



Simple Ways to Lie – Selection Bias

- Imagine a barrel of red and white beans
- How can we find out the ratio of red and white beans?
  - Count all of them
  - Take a sample from the top
- Taking a sample requires a uniform distribution
- Different densities may cause a different distribution

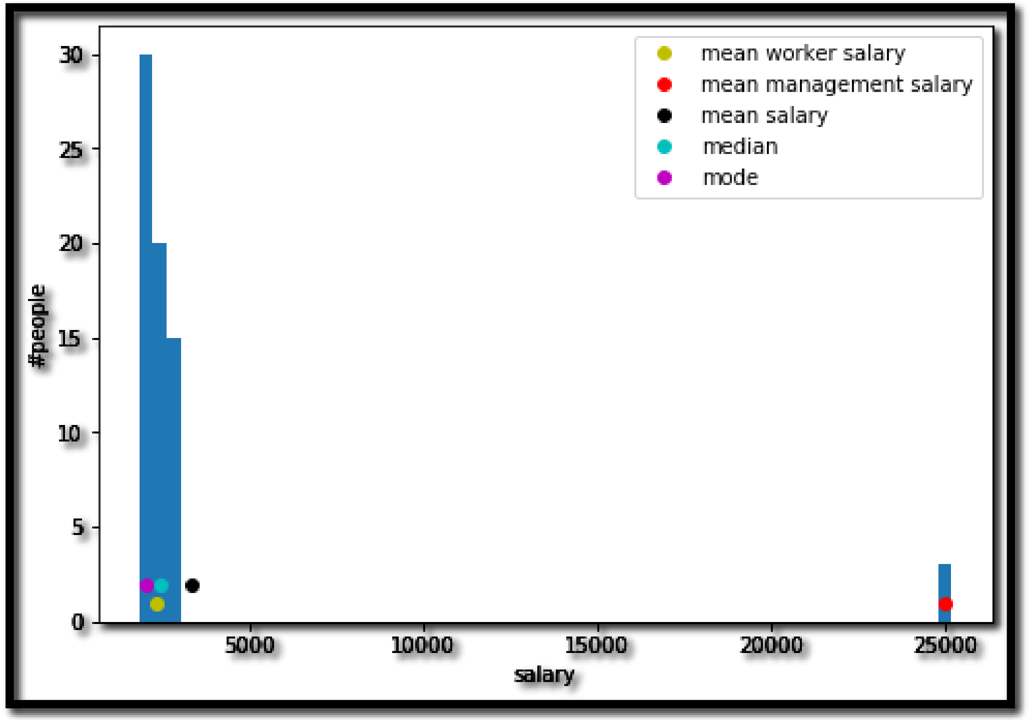




- A result of a sampling study is no better than the sample it is based on
- → Is a sample representative for the whole distribution?

Simple Ways to Lie – the Average

- “The average” is ambiguous:
  - Mean – arithmetic average
  - Median – middle value when sorted
  - Mode – value that appears most often
- Assume a company with workers and management:
  - Mean salary of workers: $2,308
  - Mean salary of management: $25,000
  - Mean salary: $3,309
  - Median: $2,400
  - Mode: $2,000



Labor Union:

Average salary of workers:	$2,000
Average salary of management:	$25,000

Management:

Average salary payed:	$3,309

Simple Ways to Lie – the Average

- Time magazine in “A Letter from the Publisher” about their readers:
- “Their **median** age is 34 years and their **average** family income is $7,270 a year”
- Was the mean used to get a bigger number?

Simple Ways to Lie – Missing Figures

- “Users report 23% fewer cavities with toothpaste X!”
  - From an independent laboratory
  - Certified by public accountant
- Are they lying? How was this number obtained?
- Through the fine print we find out:
  - 12 participants – Statistically inadequate sample size
- A small group switches to toothpaste X:
  1. Distinctly more cavities
  2. Distinctly fewer cavities
  3. About the same as before
- → Observer selection: Repeat the study, cherry-pick, discard the rest
- We don’t know how often an experiment has been repeated

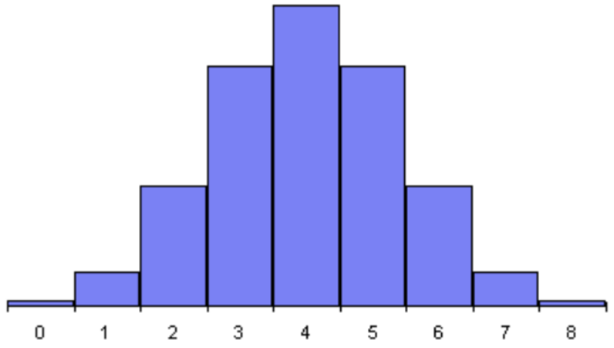
Simple Ways to Lie – Missing Figures

- We can “show” that a coin toss results in tails 75% of the time
- Tossing a coin 8 times:
    - Outcomes with 75% tails : $\binom{8}{6} = 28$
    - Total possible outcomes: $2^8 = 256$
    - Probability to get 75%: $\frac{28}{256} \approx \mathbf{0.109}$
- Tossing a coin 128 times:
    - Outcomes with 75% tails : $\binom{128}{96} \approx 1.48 \times 10^{30}$
    - Total possible outcomes: $2^{128} \approx 3.4 \times 10^{38}$
    - Probability to get 75%: $\frac{1.48\times10^{30}}{3.4\times10^{38}} \approx \mathbf{4.3 \times 10^{-9}}$
- Law of small numbers – unpredictable at the beginning
- Law of large numbers – predictable in the long run

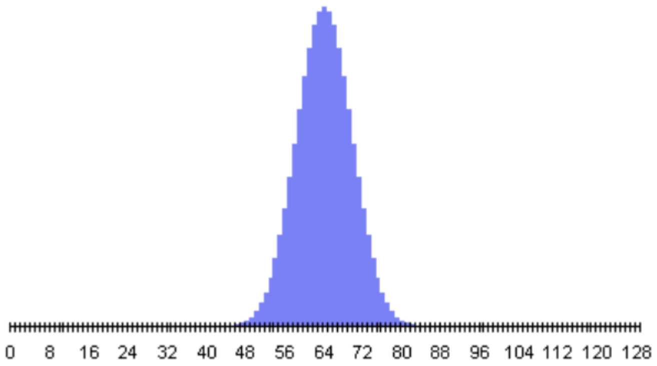
Binomial Coefficient:

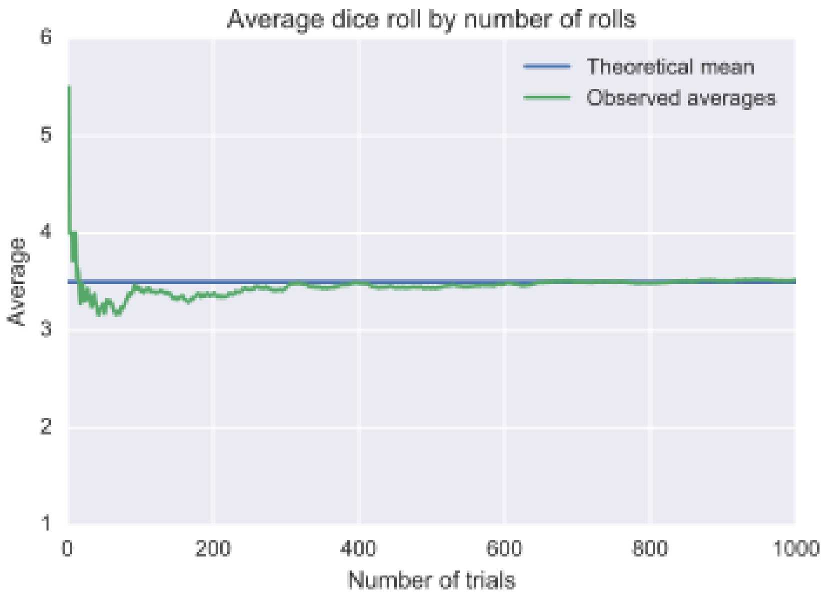
$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!}$$

8 Flips



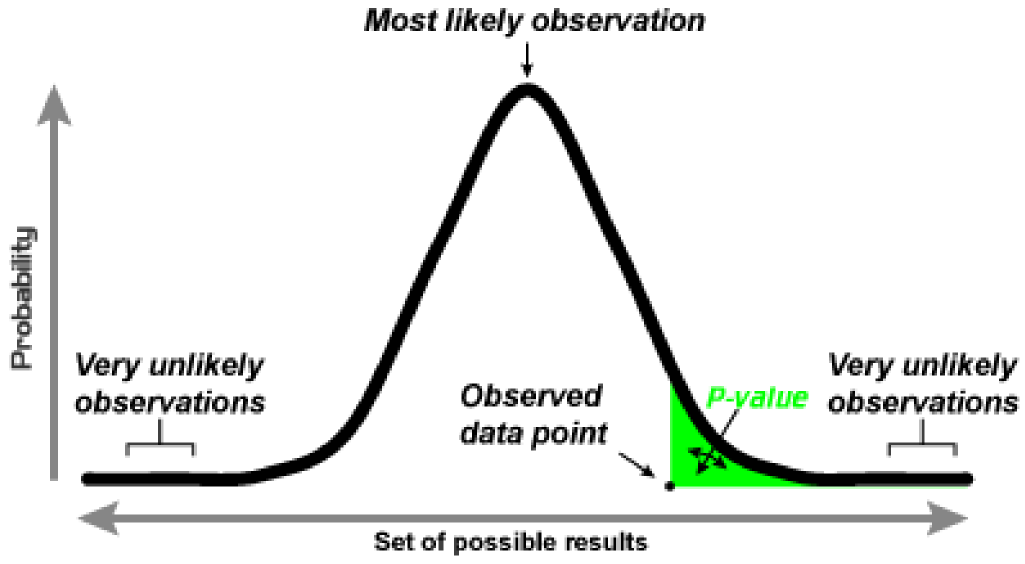
128 Flips





Simple Ways to Lie – Statistical Significance

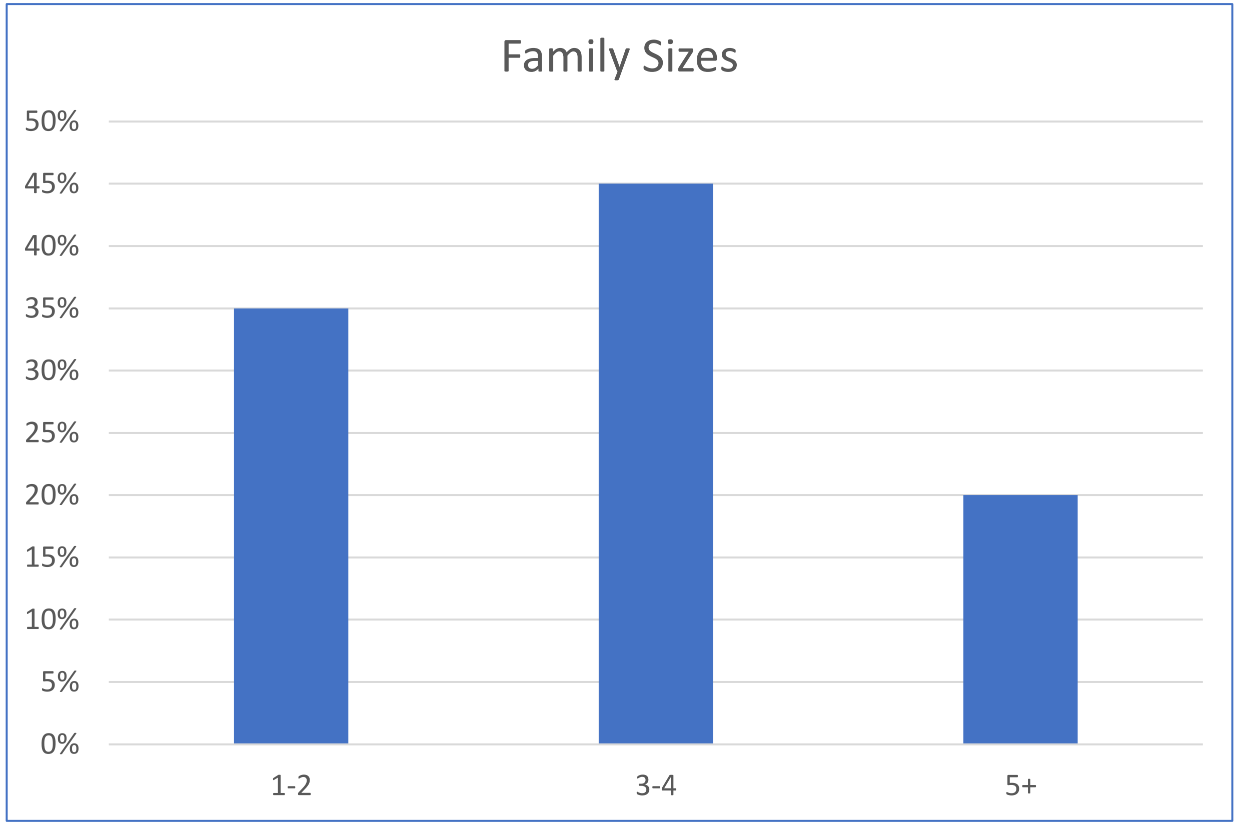
- Statistical significance can be expressed with a number
- Start with the null hypothesis
  - Every toothpaste is the same
  - 50% tails, 50% heads
- Given the null hypothesis, the p-value is the chance of getting the observed or a more extreme result
- p-values of the coin results:
  - 6 / 8 tails or more: 1.45×10^{-1}
  - 96 / 128 tails or more: 6.42×10^{-9}
- The toothpaste example would probably have a low statistical significance



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

Simple Ways to Lie – Missing Figures

- Knowing just an average can be worse than knowing nothing
- Example – American housing:
  - Mean of 3.6 people per family
  - → mainly build houses for 3-4 people
- Some more information:
  - 35% lie within 1-2
  - 45% lie within 3-4
  - 20% have 5 or more
- Many families are small, some are large
- Just the mean of 3.6 can distort the picture

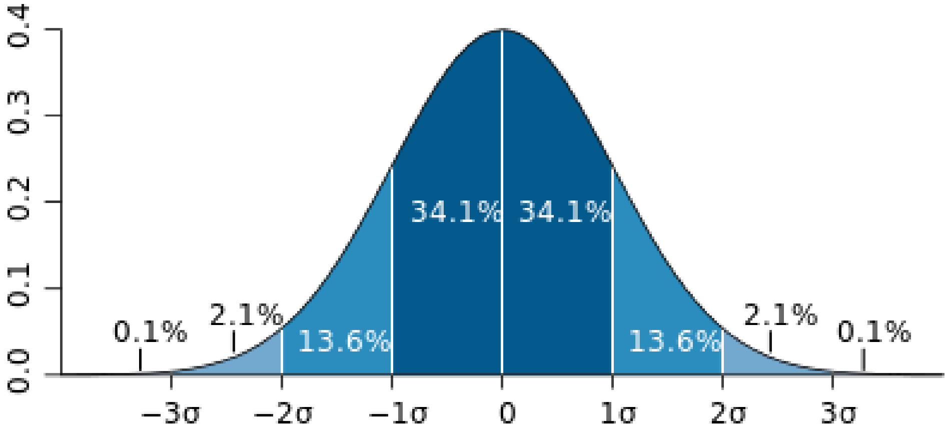


Simple Ways to Lie – Missing Figures

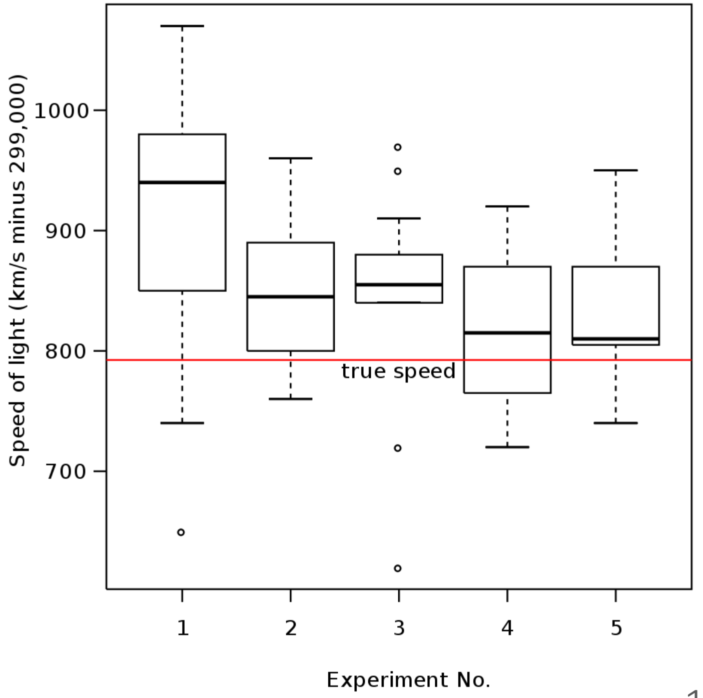
- A study about the harmful substances in different tobacco brands:
  - Virtually no difference between brands
  - One had to be at the bottom of the list
  - → A huge advertising campaign – "The healthiest cigarette of them all!"

- → Later in the this seminar: "The health effects of smoking"

Simple Ways to Lie – Indicators of Range



- There are multiple ways of providing a range
- Standard deviation
  - A measure for variation or dispersion of the set
  - $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$
  - Tobacco study: standard deviation when testing the same brand is very high
- Box plots – Box & Whiskers plots
  - Display of quartiles
  - 50% of the data is within the box
  - The median is displayed within the box
  - Whiskers can be limited in different ways
    - → Draw outliers as points



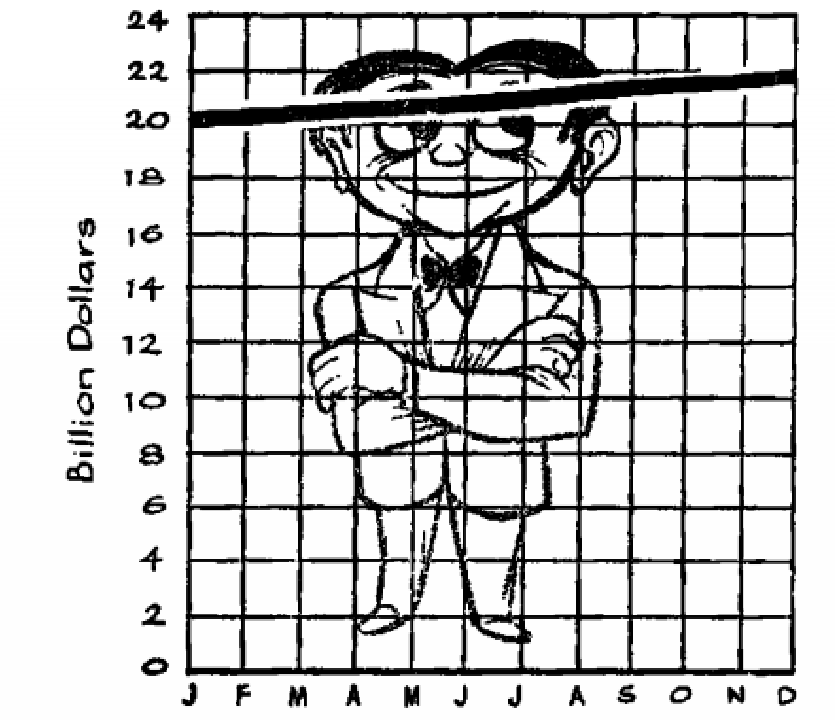
Heidelberg Collaboratory for Image Processing

Simple Ways to Lie – Missing Figures

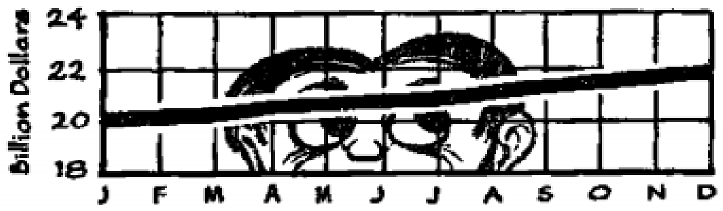
- “Electric power is available to more than $^3/_4$ of U.S. farms.”
- Could have been expressed as “Almost $^1/_4$ do not [..]”

  motivation ✓
- What classifies as “available”?
  - Do they have access to electricity in their homes?
  - Power lines in the vicinity?
    - Within meters?
    - Within kilometers?
- The statement isn’t false, but little information is conveyed
- The deceptive thing about the missing figures is, that their absence often goes unnoticed.

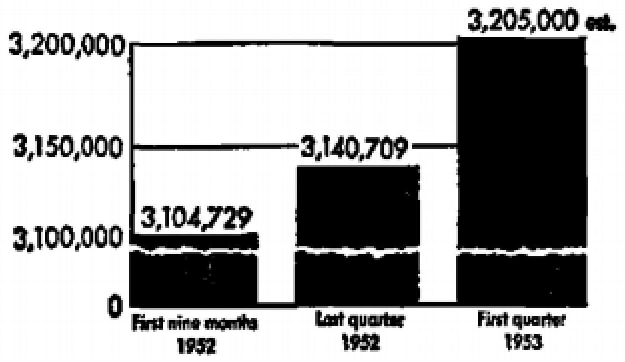
Simple Ways to Lie – Charts and pictographs



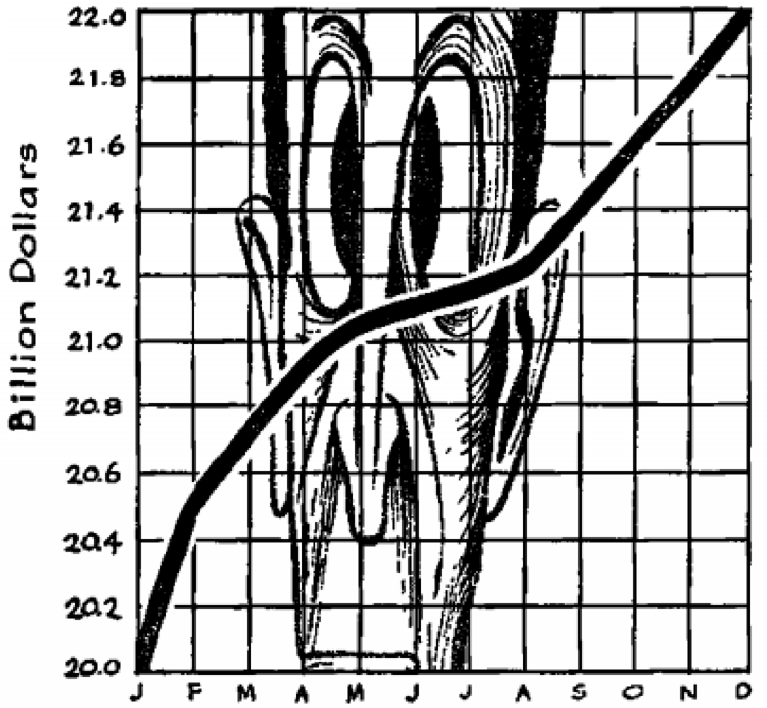
A small increase of 10% is perceived accurately



This version saves paper!
At a glance it seems to double

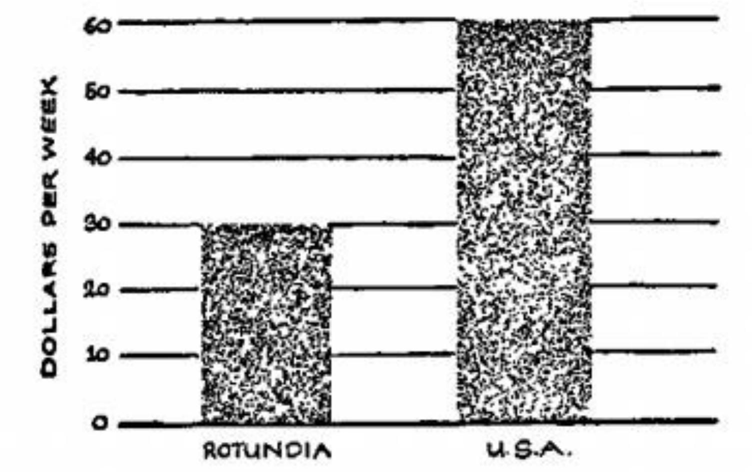


At least the cut is made obvious here

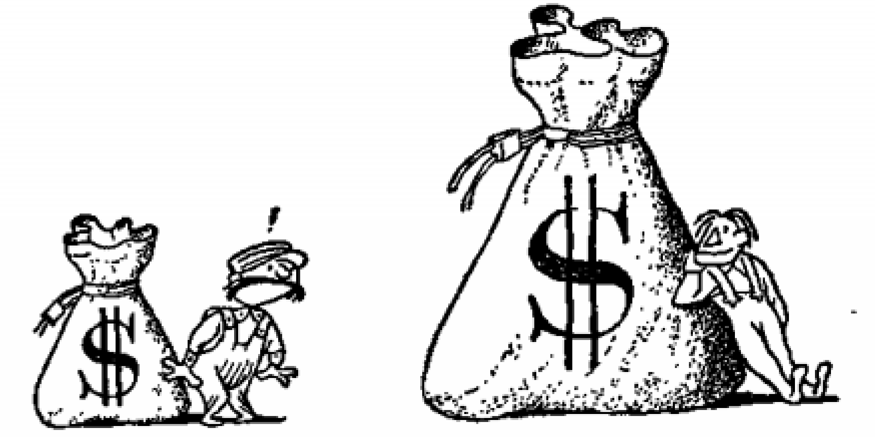


This seems newsworthy

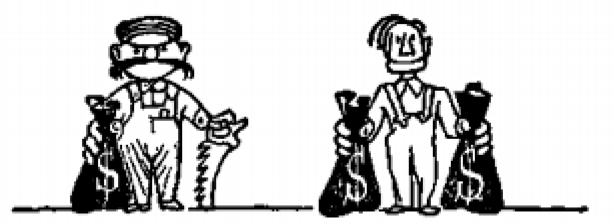
Simple Ways to Lie – Charts and pictographs

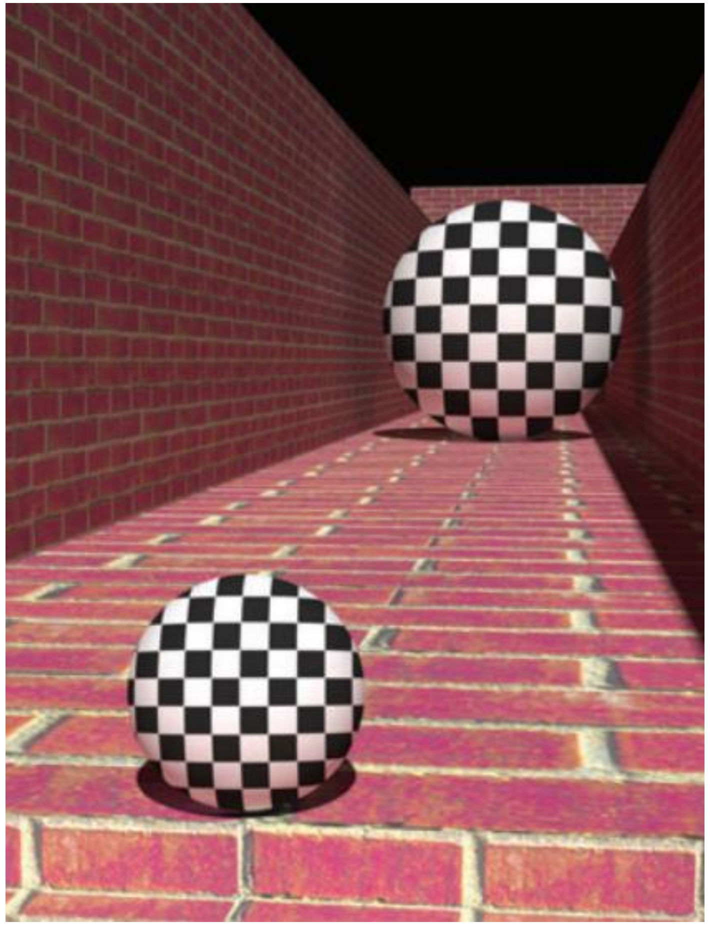


The increase of 100% can be presented as a pictograph



Instead 1 bag of twice the ~~size~~ height → 4 times the area, 8 times the volume





Sterzer, P. and Rees, G., 2006. Perceived size matters. *Nature Neuroscience*, *9*(3), p.302.

- Additionally, humans are easily fooled when perceiving size
- More about charts in the upcoming presentation next week: “How to Lie with Charts”

Simple Ways to Lie – The Semi Attached Figure

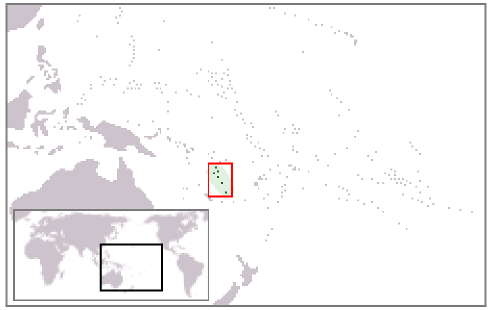
- An advertisement for a new electrical juicer:
  - “Extracts 26% more juice!”
  - 26% more juice than it’s competitors?
  - The figure comes from a comparison to a hand juicer
  - This information is almost irrelevant when buying an electric juicer
- More people died on airplanes this year than 100 years ago:
  - Does this mean airplanes are becoming more dangerous?
    - There are more people
    - There are more airplanes

Simple Ways to Lie – The Semi Attached Figure

- While navy personnel counted 9 deaths in 1000, for civilians in New York it was 16 in 1000
    - Is it safer to be in the navy?
    - Navy consists mostly of young and healthy people
- “If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.” – Darrell Huff, *How to Lie with Statistics*
- Things may sound the same at first, but they are not
- This is also known as the association fallacy
- More about fallacies in this seminar in Topic 4: Fallacies of Thinking

Simple Ways to Lie – Correlation and Causation

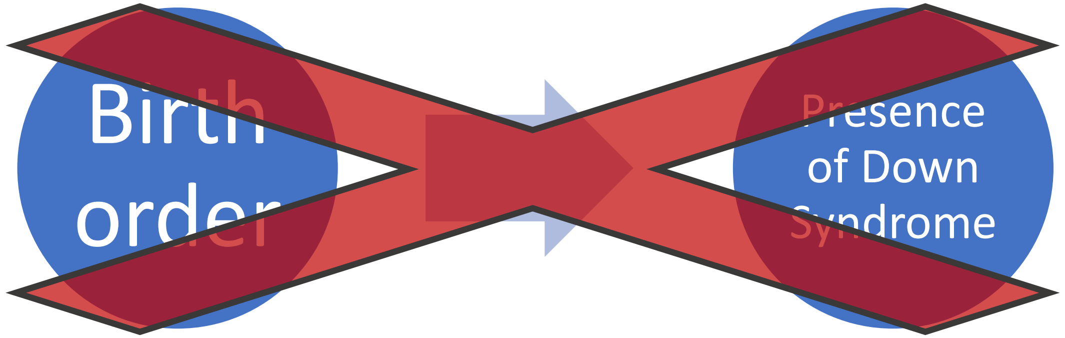
- Indigenous people from the island Vanuatu assumed having lice causes good health
- The evidence:
    - Everyone had lice
    - Most sick people had no lice



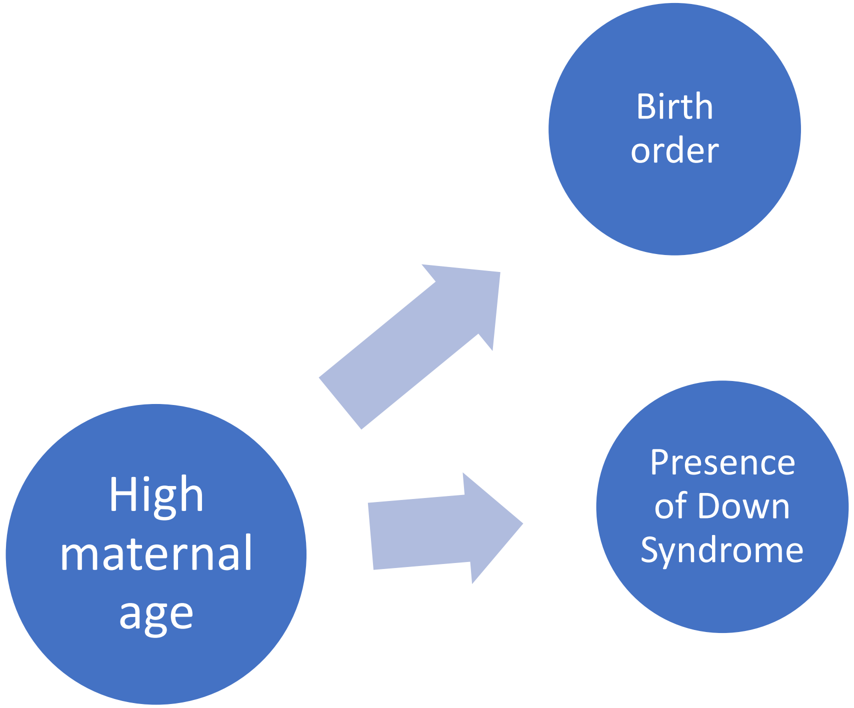
- A temperature change of 4-5 Degree is fatal for lice → People with fevers had no lice

Simple Ways to Lie – Correlation and Causation

- Is there a correlation between number of younger siblings and presence of Down Syndrome?
- Many younger siblings → presence of Down Syndrome?

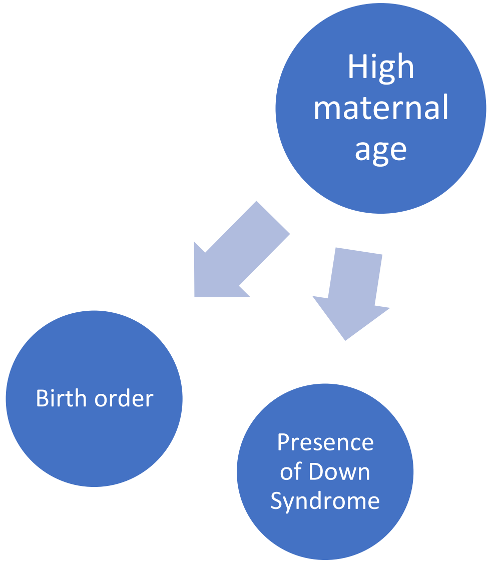


- More likely:
  - High maternal age → Birth order
  - High maternal age → Presence of Down Syndrome



Simple Ways to Lie – Types of Correlations

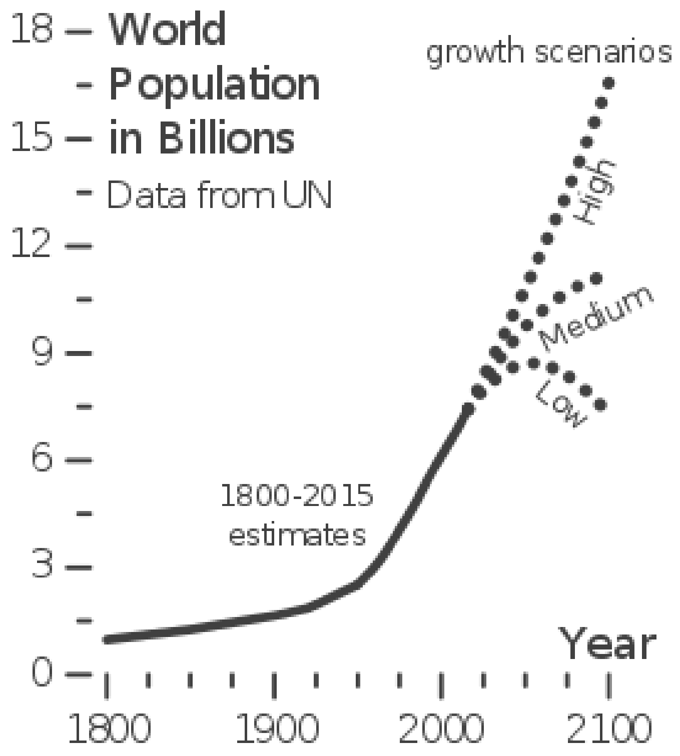
- Correlation by chance
  - Apparent correlation at first, but no correlation after multiple runs
  - Toothpaste example
- Real correlation, but what is the cause and what is the effect?
  - Income and ownership of stocks:
    - More income → more stock ownership
    - More stock ownership → more income
- Real correlation, but a third factor is the cause for both → Confounder
  - Birth order and Down Syndrome



Simple Ways to Lie – Unlimited Extrapolation

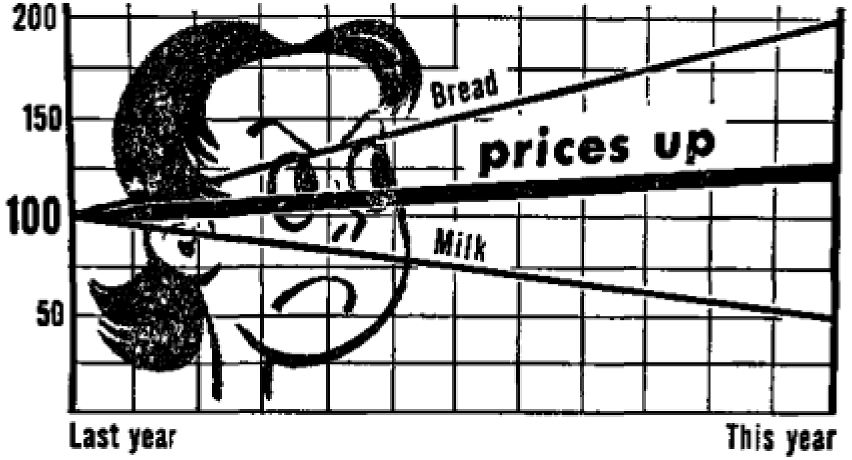
- Trends can be used to extrapolate on the data
- A lot of rain positively correlates with the quality of a harvest
  - But too much rainfall ruins the crop
  - Positive correlation may only hold to a point

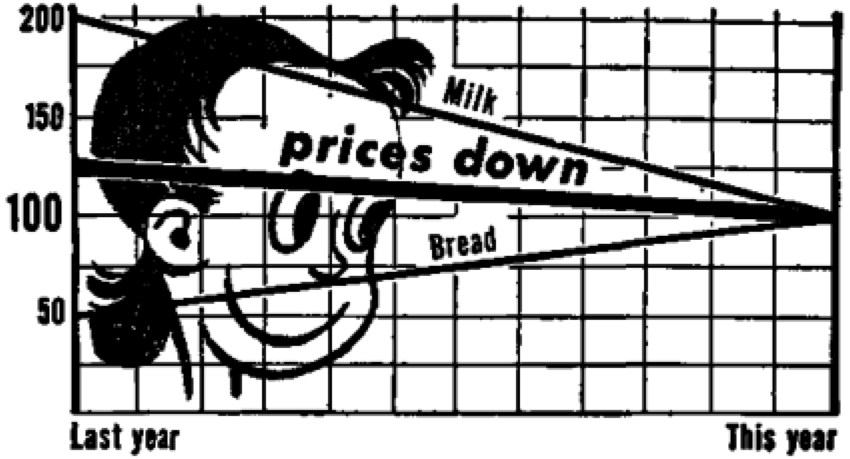
Estimates for the future world population



Causes for Lies

- Cost of living:
  - Milk price halves
  - Bread price doubles
- “Cost of living up!”
  - Past prices are the base
- “Cost of living down?”
  - New prices are the base
- → Confusion of base





In this case the geometric mean is the appropriate average:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \dots x_n}$$

$$\sqrt[2]{200\% \times 50\%} = 100\%$$

Causes for Lies – Confusion of base

- “Buy your Christmas presents now and save 100%”
  - based on the new price
  - → 50% cheaper
  - Today this counts as unfair business practice
- Pay cuts
  - “50% pay cut”
    - $\$100 \times 0.5 = \$50$
  - “50% pay cut restored”
    - ~~$\$50 + (\$100 \times 0.5) = \$75$~~
    - $\$50 \times 1.5 = \$75$
- Conveniently both cases sound better than they really are

Causes for Lies – Motive

- Are most lies a product of ill intent?
- Distortion and manipulation of statistics is not always the work of professional statisticians
- Legitimate work may be distorted, cherry-picked, and exaggerated for personal gain
- For most lies in statistics a motive can be found
  - Sensationalize
  - Inflate
  - Confuse
  - Oversimplify
- "Mistakes" are one-sided

Identifying Lies in Statistics

- Should only one measure be used to express averages?
  - Statistical methods shouldn't be rejected arbitrarily
  - Each measure has its place
- According to the author:
  - Statistics should be taken with a grain of salt
  - One should be able to recognize sound and usable data
    - Competence & integrity of the statistician
    - Competence & integrity of the writer
    - Competence of the reader
- Maybe statistics can be changed in a way that removes human error from the equation
  - → Topic 9: How to do Better?

Identifying Lies in Statistics

- Who says so?
  - A reputable name does not imply proper representation of the data
  - → Who is drawing the conclusions?
- How does he know?
  - Bad sampling?
- What’s missing?
  - The distribution might be very unnatural
  - → Averages may differ and don’t convey the underlying information well
- Did somebody change the subject?
  - Association fallacy / semi attached figure
- Does it make sense?
  - One statistic judged readability based on average word-length

Identifying Lies – Revisiting Nitrate Levels in Groundwater

- "From 2013 to 2017 the average nitrate concentration in the top 15 polluted regions has increased by 40mg/l"
- "Average":
  - In 2013 the "average" was the mean over the whole year
  - In 2017 the "average" was of maxima over multiple days
  - → **Unspecified and different averages**
- "Top 15 polluted regions":
  - Nitrate concentrations are measured to ensure safe levels
    - Regions with low concentration are not as interesting
    - Measurement devices are moved to regions with higher concentrations
  - The top 15 regions, the compared samples, change over time
  - When looking at the same regions, nitrate levels have decreased
  - → **Missing figure + selection bias**