How to Lie with Statistics. Getting acquainted with R

Корпусные методы исследований языковых процессов

Даша Попова

9.11.2022

Huff, 1954: The Sample with the Built-in Bias

- “The average Yaleman, Class of ’24, makes $25,111 a year”

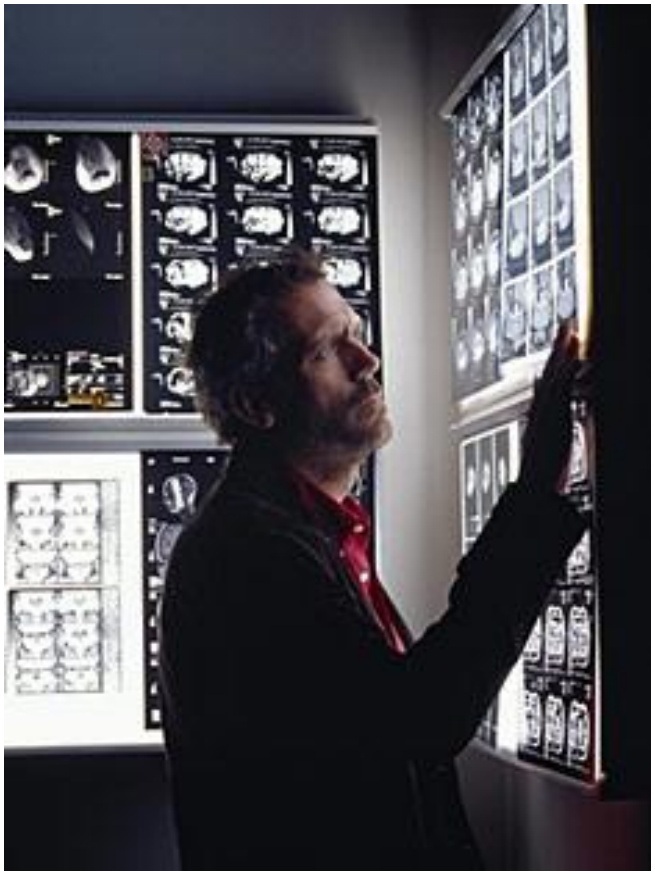
- Suspicious?
- Yes:

- the figure is too precise: people rarely know their income down to the dollar;
- btw, if you report your own findings, round them appropriately!

Huff, 1954: The Sample with the Built-in Bias

- "The average Yaleman, Class of '24, makes $25,111 a year"
- Suspicious?
- Yes:

- "Everybody lies" (Dr. Gregory House, Ep. 101)
- People lie by omission; by exaggeration; by minimization
- People lie, because they want to be polite; wish to provide complete answers; wish to meet the researcher's expectations; are ashamed of certain things
- Some questions invite lies



Schweigespirale, спираль молчания

- The **spiral of silence theory** is a political science and mass communication theory proposed by the German political scientist Elisabeth Noelle-Neumann.

- The following steps summarize how the process works:
- The model begin with individuals' inherent desire to blend with society. The fear of social isolation is necessary for the spiral to occur.
- Individuals who notice that their personal opinion is spreading will voice this opinion confidently in public. On the other hand, individuals who notice that their opinions are losing ground will be inclined to adopt a more reserved attitude when expressing their opinions in public.
- Representatives of the spreading opinion talk quite a lot while the representatives of the second opinion remain silent. An opinion that is being reinforced in this way appears stronger than it really is, while an opinion suppressed will seem to be weaker than it really is.
- The result is a spiral process which prompts other individuals to perceive the changes in opinion and follow suit until one opinion has become established as the prevailing attitude while the other opinion will be pushed back and rejected by most. The end of the spiral refers to the number of people who are not publicly expressing their opinions, due to the fear of isolation.

https://en.wikipedia.org/wiki/Spiral_of_silence

Huff, 1954: The Sample with the Built-in Bias

- “The average Yaleman, Class of ’24, makes $25,111 a year”
- Suspicious?
- Yes:

- the sample is biased, not representative: researchers tend to select respondents that are more easily accessible (physically, emotionally, intellectually) and/or that are more likely to meet their expectations (der Kluge Hans).

- “A psychiatrist reported once that practically everybody is neurotic”

- “60% учителей не сдали тест по математике” (Радио России, 14.11.18, 8-9)

Huff, 1954: The Well-Chosen Average

- What kind of average?
- Average of what?
- Average income in the neighborhood example:
  - $15,000 (when bragging) – the arithmetic average of the incomes of all the families in the neighborhood
  - $5,000 – the modal income
  - $3,500 (when petitioning for lower bus fare) – the median
- Average wages of employees example (a manufacturing business with three owners and 90 employees):
  - average wage of employees -- $2,200 (198,000/90)
  - average salary and profit of owners -- $26,000 (11,000 + 45,000/3)
  - bonuses for owners (part of the profit): $30,000
  - average wage or salary: $2,806.45 ((198,000 + 33,000 + 30,000)/93)
  - average profit of owners: $5,000 (15,000/3)

Huff, 1954: The little figures that are not there

- ‘The importance of using a small group is this: With a large group any difference produced by chance is likely to be a small one and unworthy of big type.’
- Example: tossing a penny ten times vs. a thousand times.
- Example: ‘users report 23% fewer cavities with Doakes' tooth paste’.

‘Let any small group of persons keep count of cavities for six months, then switch to Doakes'. One of three things is bound to happen: distinctly more cavities, distinctly fewer, or about the same number. If the first or last of these possibilities occurs, Doakes & Company files the figures (well out of sight somewhere) and tries again. Sooner or later, by the operation of chance, a test group is going to show a big improvement worthy of a headline and perhaps a whole advertising campaign.’

Huff, 1954: The little figures that are not there

- How many is enough?
- ‘It depends among other things on how large and how varied a population you are studying by sampling. And sometimes the number in the sample *is not what it appears* to be.’
- Example: polio vaccination (the control group, albeit large, was not affected due to the rarity of contraction)

Huff, 1954: The little figures that are not there

- the degree of significance
- the range of things or their deviation from the average that is given
- 'Knowing nothing about a subject is frequently healthier than knowing what is not so, and a little learning may be a dangerous thing.'
- Example: too much American housing that has been planned to fit the statistically average family of 3.6 persons.
- Example: baby stats – half the parents overjoyed, half the parents panicking.
- "Today, electric power is available to more than three quarters of U. S. farms... :" do they use it? Implicatures!!

Presuppositions and implicatures

(1) *Когда Ваня ездил в Париж?*
Пресуппозиция: Ваня ездил в Париж.

(2) *Король Франции лыс.*
Пресуппозиция: Существует король Франции.

(3) *Мой слон любит фисташковые торты.*
Пресуппозиция: У меня есть слон.

Пресуппозиция – условие осмысленности предложения, сохраняется под отрицанием и в вопросах.

(4) *У Вани две дочки.*
Импликатура: две, а не три, четыре …

(5) *Ваня промочил ноги и заболел.*
Импликатура: Ваня заболел после того, как промочил ноги.

(6) -- *Где кастрюли?*
-- *Посмотри в нижнем ящике.*
Импликатура: Возможно, кастрюли в нижнем ящике.

Presuppositions and implicatures

Заголовки газет часто содержат пресуппозиции. Какие тут содержатся пресуппозиции? Найдите триггеры пресуппозиций, если возможно, в следующих заголовках.

- Минфин Чечни попросил увеличить дотации на 13 миллиардов рублей
- Royal Bank of Scotland заплатит миллиард долларов за обман акционеров
- Чиновникам запретили автомобили с мощным двигателем
- В Минобороны подтвердили потерю Су-33 при посадке на «Адмирал Кузнецов»
- Хокинг предложил новое описание черных дыр
- Евгений Исаков о развитии серфинга в России и его главных опасностях
- Почему в России падает производство пушнины?
- *Гуаидо назвал дату начала направленной на свержение Мадуро операции* https://www.rbc.ru/politics/28/03/2019/5c9bf3549a794766d6e39428?from=from_main

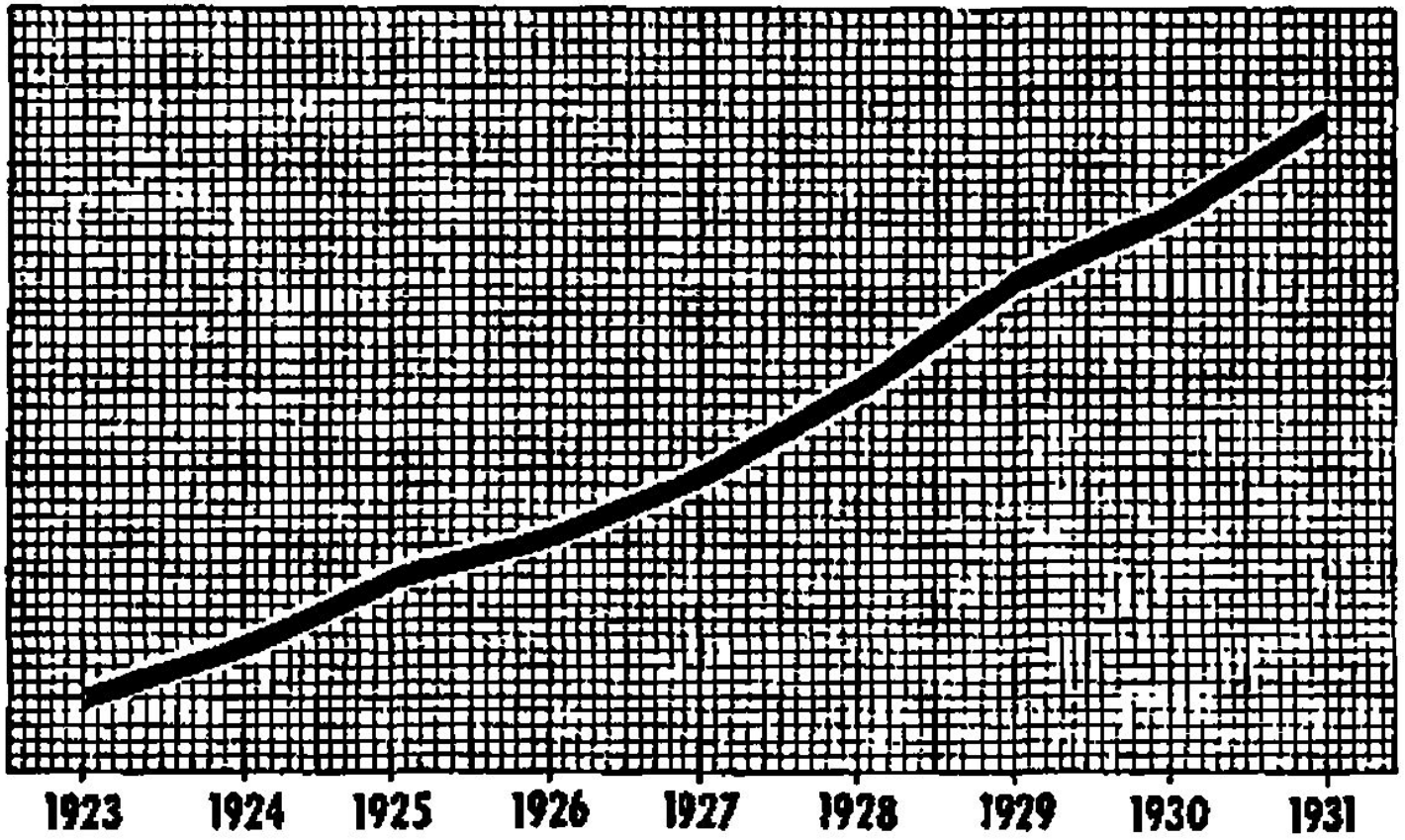
Presuppositions and implicatures

Заголовки газет часто содержат импликатуры. Какие тут содержатся импликатуры?

- К видеоблогеру в Кемерово пришли с обыском после ролика про банкира Тинькова https://www.novayagazeta.ru/news/2017/09/13/135229-k-videoblogeru-v-kemerovo-prishli-s-obyskom-posle-rolika-pro-bankira-tinkova

- Россиянка сходила в магазин и осталась без квартиры https://lenta.ru/news/2018/07/31/stayhome/

Huff, 1954: The little figures that are not there



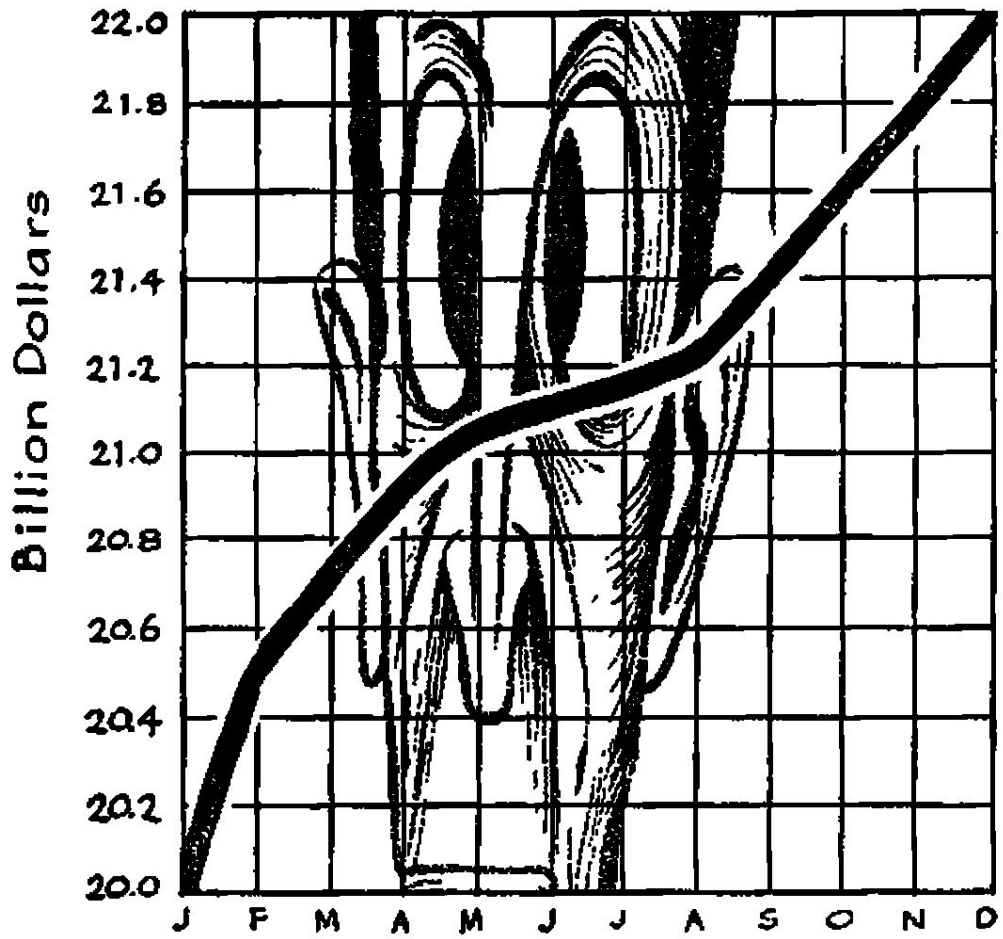
a graph used to advertise an advertising agency

Huff, 1954: Much ado about Practically Nothing

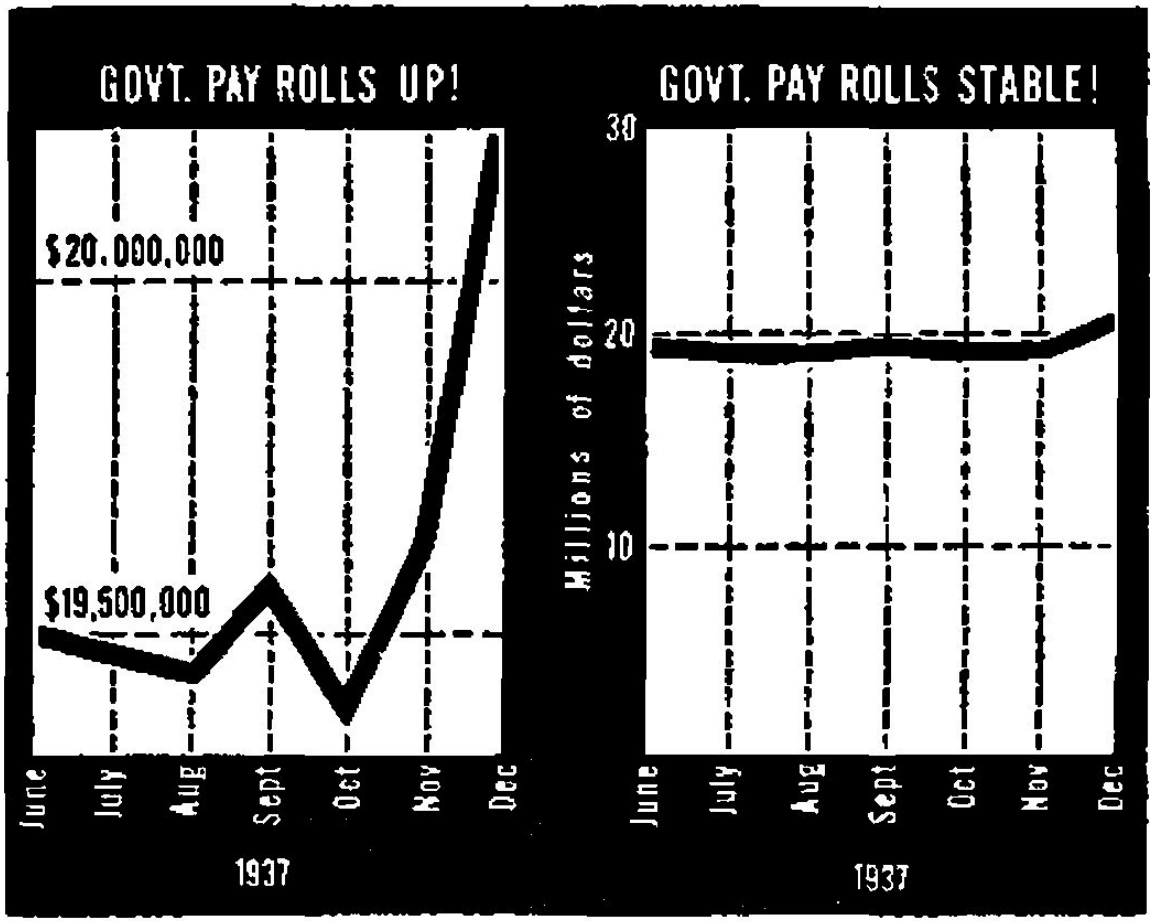
- ‘Sometimes the big ado is made about a difference that is mathematically real and demonstrable but so tiny as to have no importance. This is in defiance of the fine old saying that a difference is a difference only if it makes a difference.’

- Example: ‘The conclusion stated by the magazine and borne out in its detailed figures was that all the brands were virtually identical and that it didn't make any difference which one you smoked. But somebody spotted something. In the lists of almost identical amounts of poisons, one cigarette had to be at the bottom, and the one was Old Gold. Out went the telegrams, and big advertisements appeared in newspapers at once in the biggest type at hand. The headlines and the copy simply said that of all cigarettes tested by this great national magazine Old Cold had the least of these undesirable things in its smoke. Excluded were all figures and any hint that the difference was negligible.’

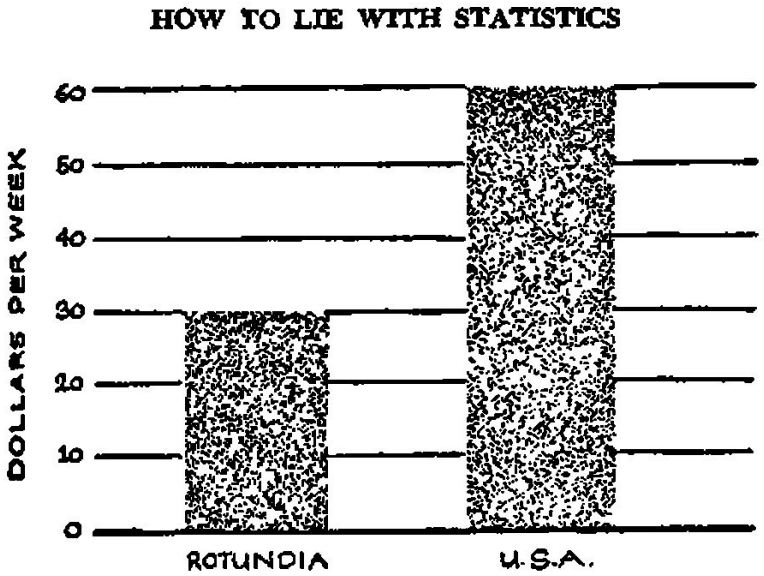
Huff, 1954: The Gee-Whiz Graph



Huff, 1954: The Gee-Whiz Graph



Huff, 1954: The One-dimensional Picture



If one moneybag holds $30, the other, having eight times the volume, must hold not $60, but $240.

Huff, 1954: The Semi-attached Figure

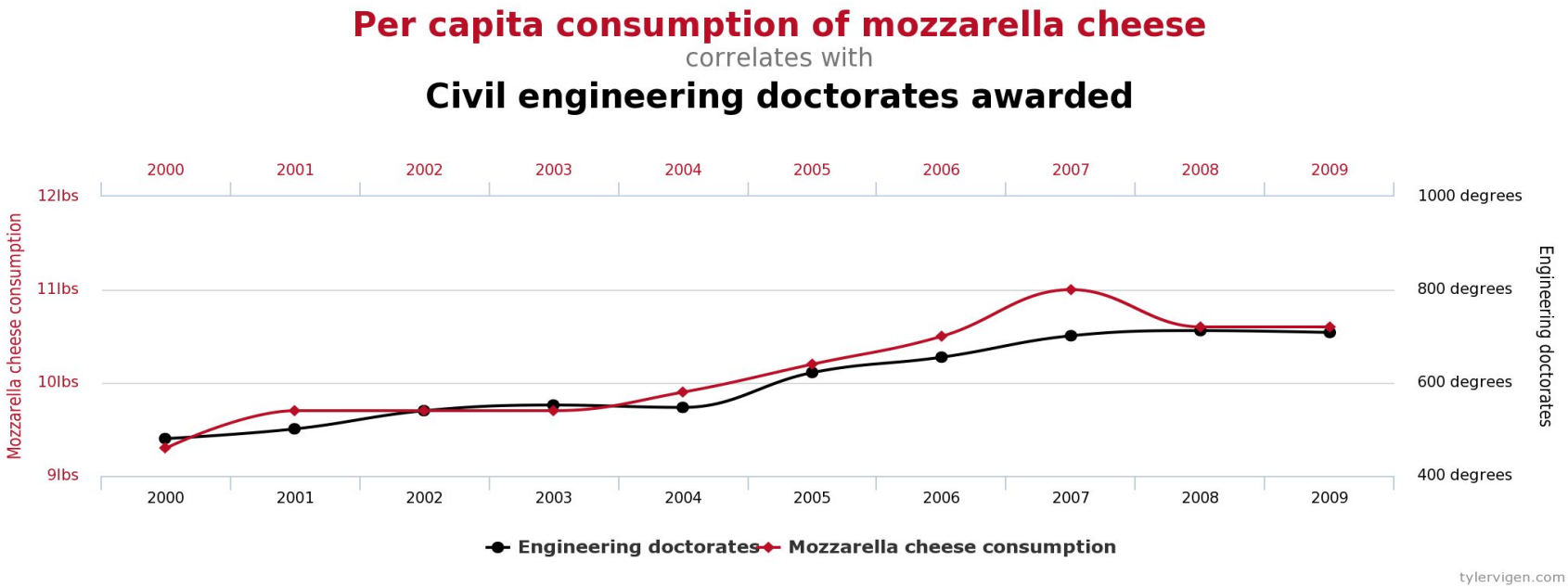
- ‘If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing.’
- Example: ‘You can't prove that your nostrum cures colds, but you can publish (in large type) a sworn laboratory report that half an ounce of the stuff killed 31,108 genus in a test tube in eleven seconds.’
- ‘There are often many ways of expressing any figure. You can, for instance, express exactly the same fact by calling it a one per cent return on sales, a fifteen per cent return on investment, a ten-million-dollar profit, an increase in profits of forty per cent (compared with 193539 average), or a decrease of sixty per cent from last year.’

Huff, 1954: The Semi-attached Figure

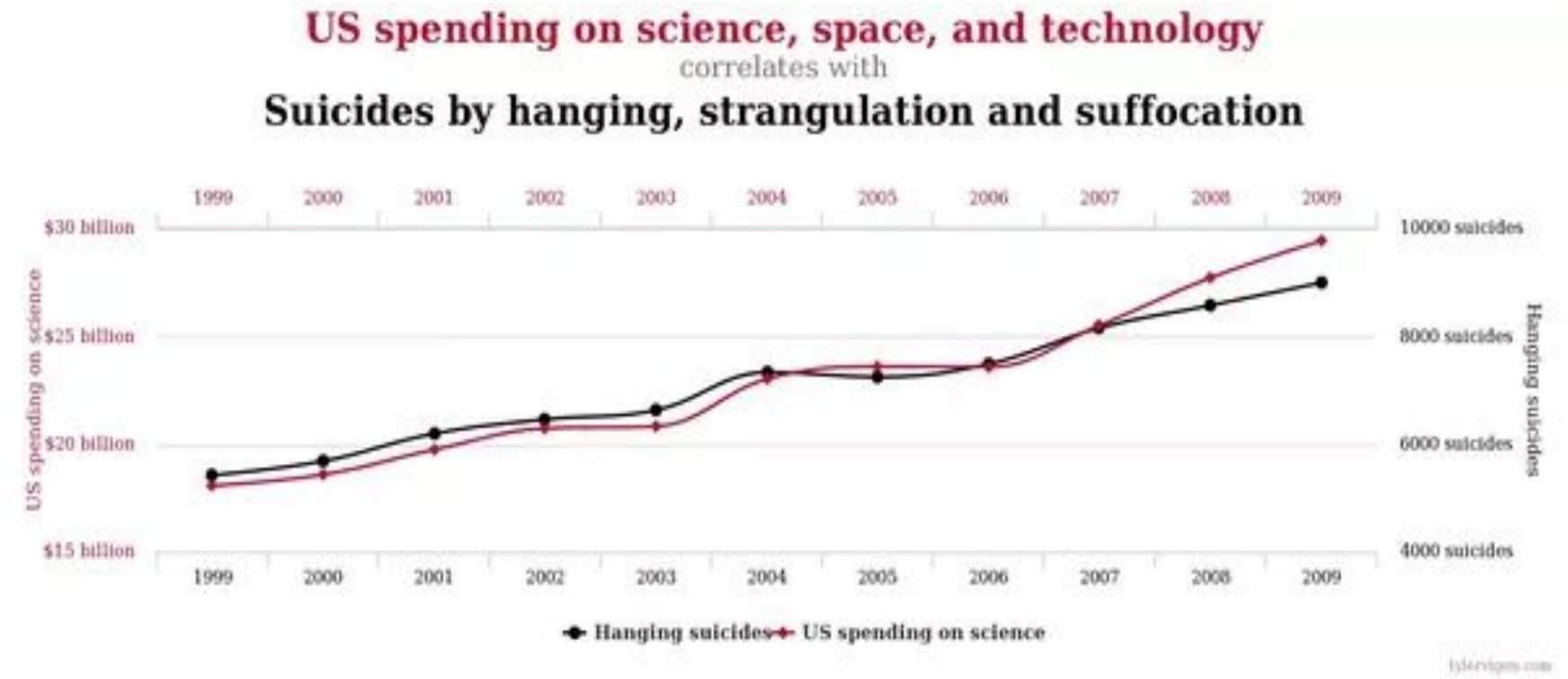
- ‘The death rate in the Navy during the Spanish-American War was nine per thousand. For civilian in New York City during the same period it was sixteen per thousand. Navy recruiters later used these figures to show that it was safer to be in the Navy than out of it.’

Huff, 1954: Post Hoc Rides Again

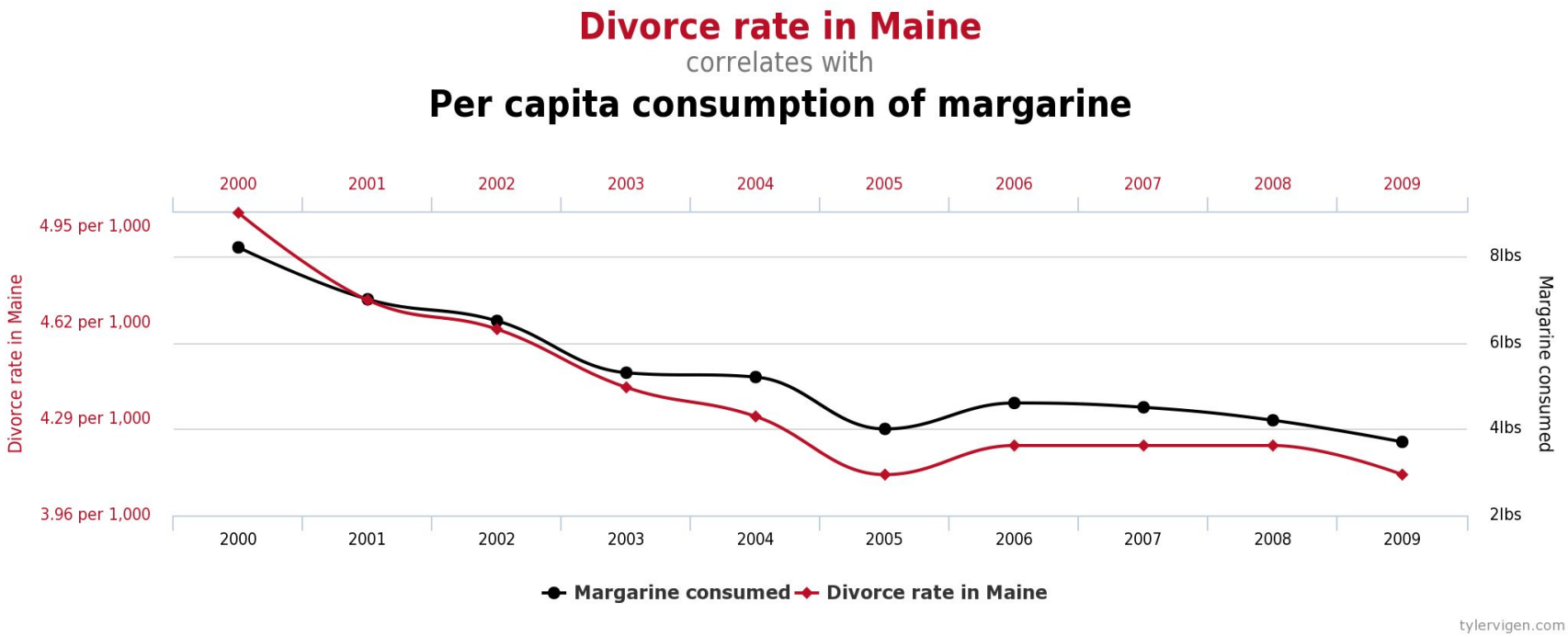
- The fallacy: if B follows A, then A has caused B.
- Example: smoking and low grades.
- A correlation by chance;
- A correlation where it is not possible to tell what caused what;
- A genuine correlation, without any one variable having any effect on the other variable.
- Another thing to watch out for is a conclusion in which a correlation has been inferred to continue beyond the data with which it has been demonstrated, e.g., rain and crops – a positive correlation holds up to a point and then quickly becomes a negative one.
- 'A correlation may be real and based on real cause and effect and still be almost worthless in determining action in any single case.'



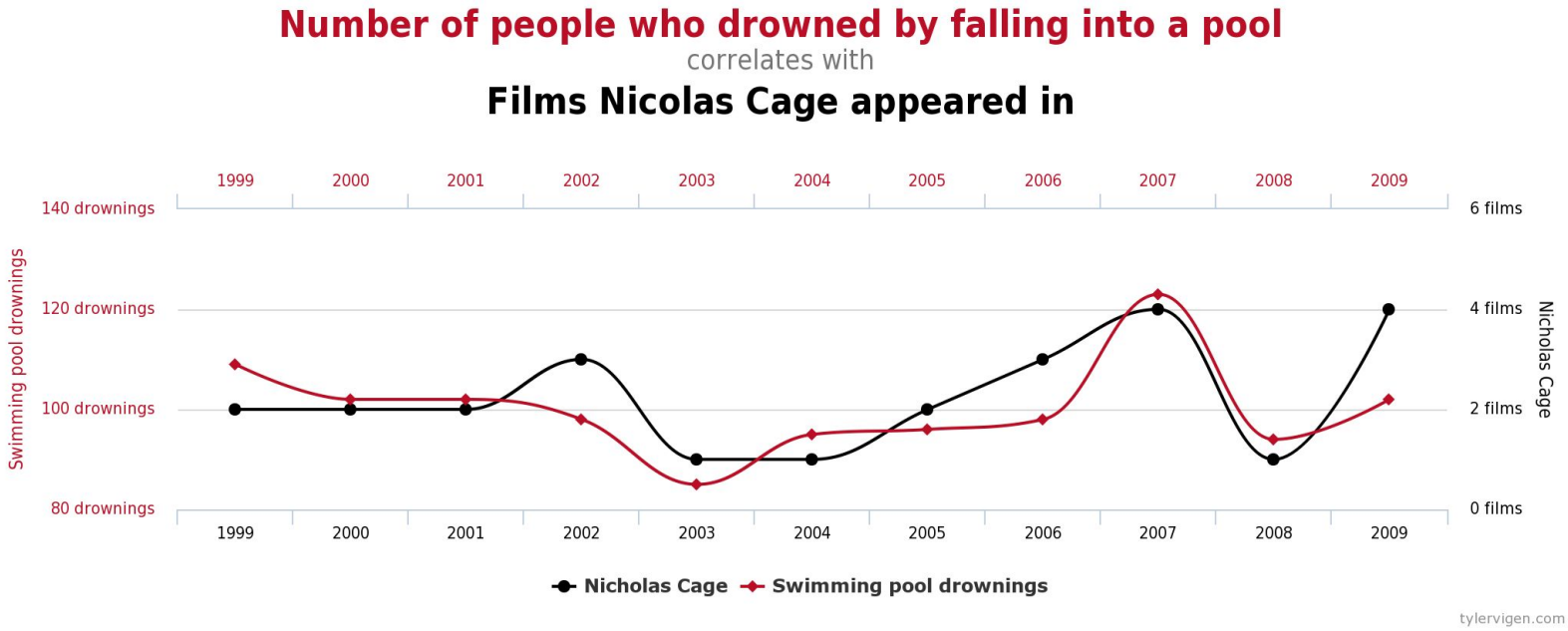
http://www.tylervigen.com/spurious-correlations



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

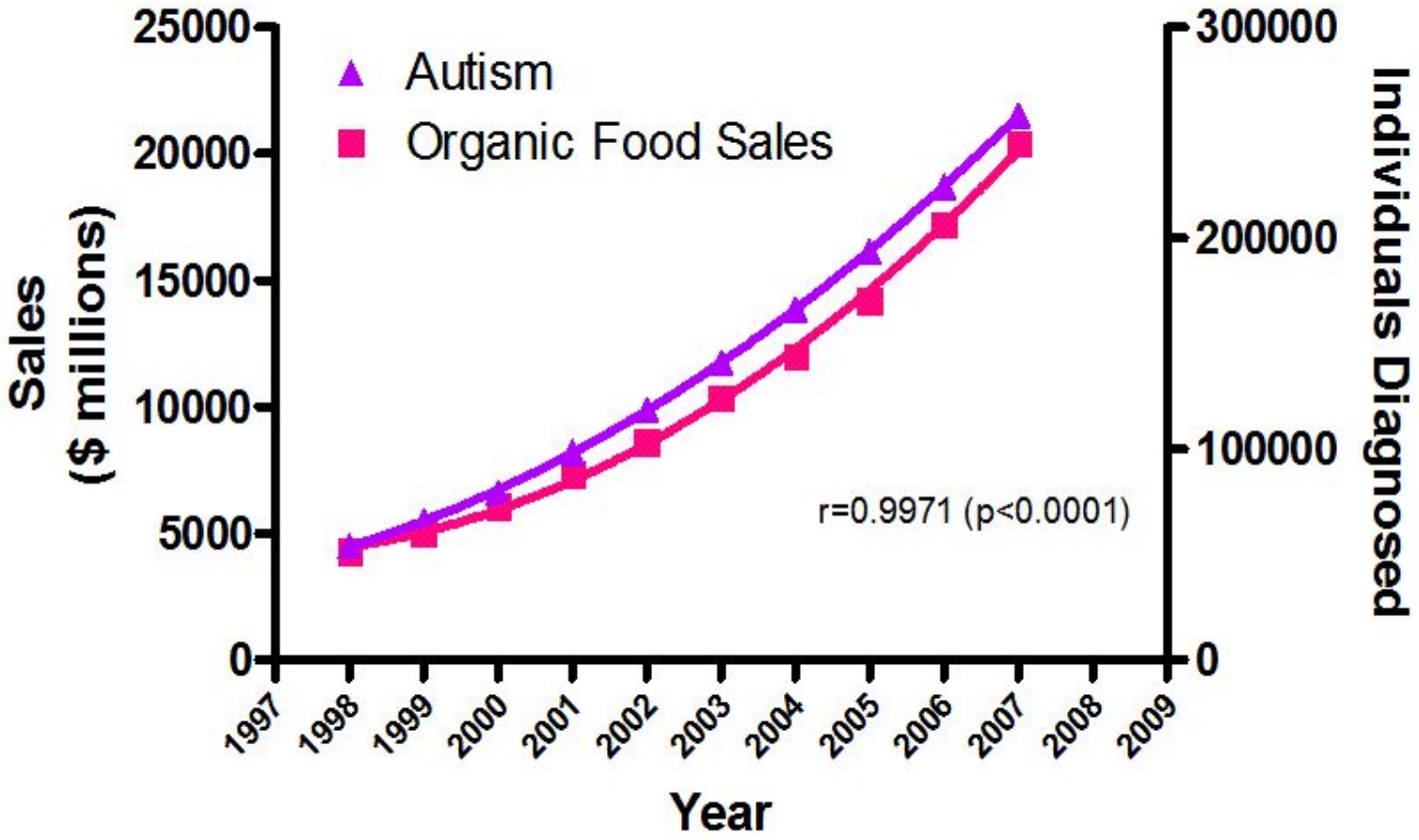


http://www.tylervigen.com/spurious-correlations



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

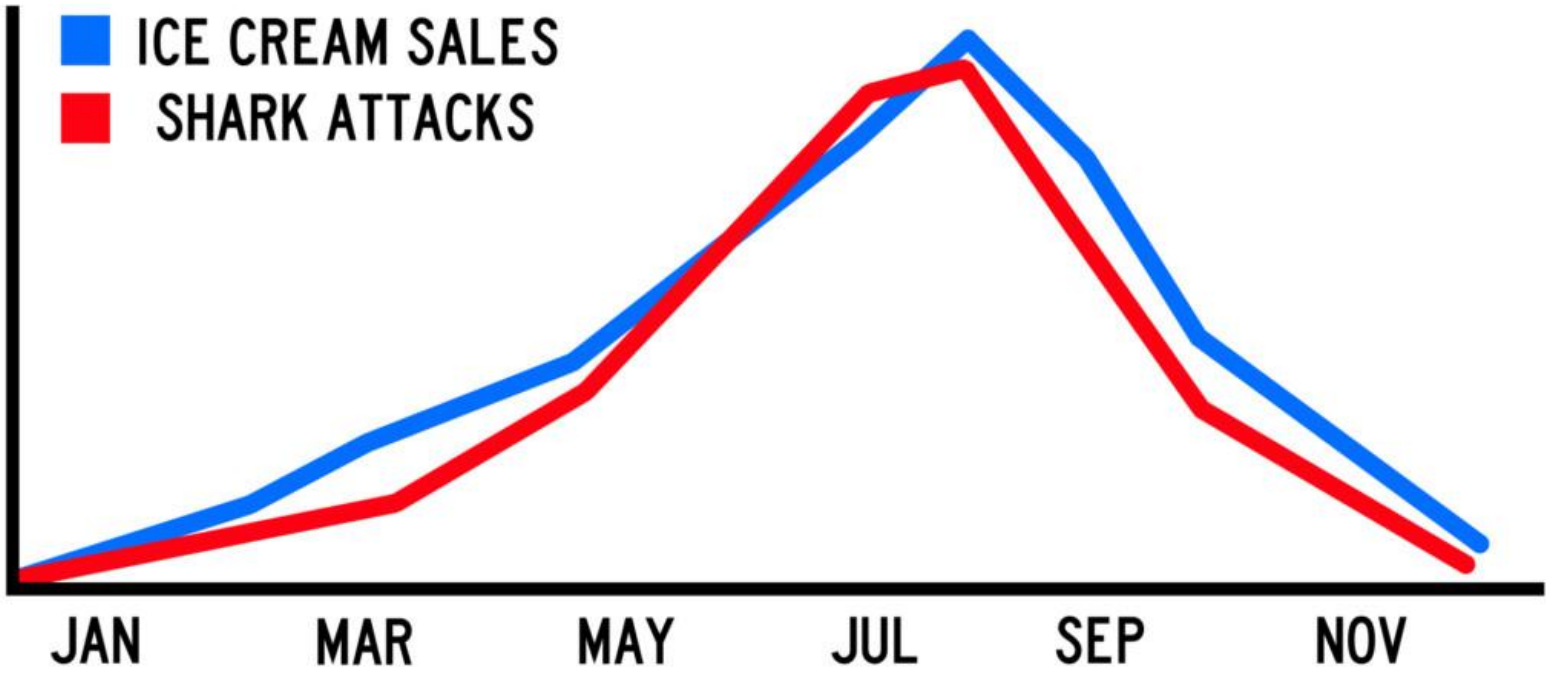
The real cause of increasing autism prevalence?



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

ThinkWell!

CORRELATION IS NOT CAUSATION!



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

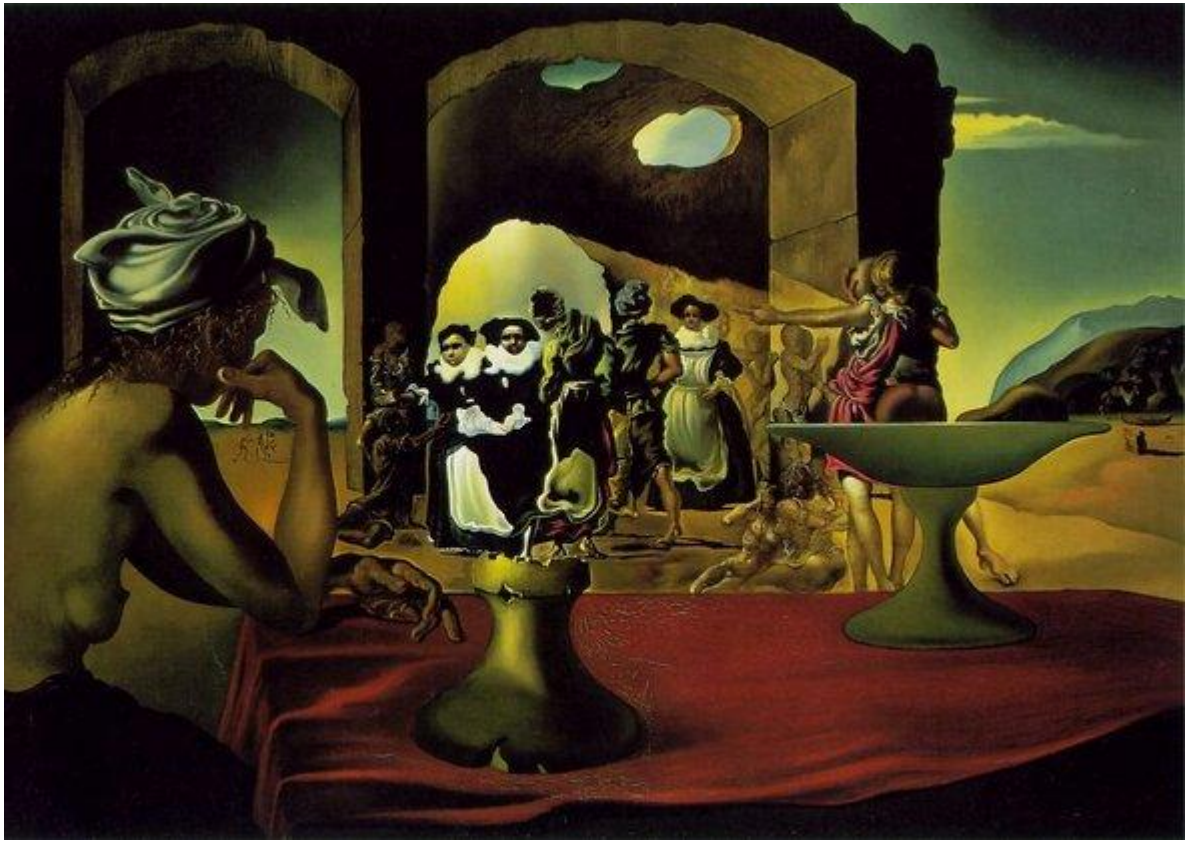
www.blog44.ca

Huff, 1954: How to Statisticulate

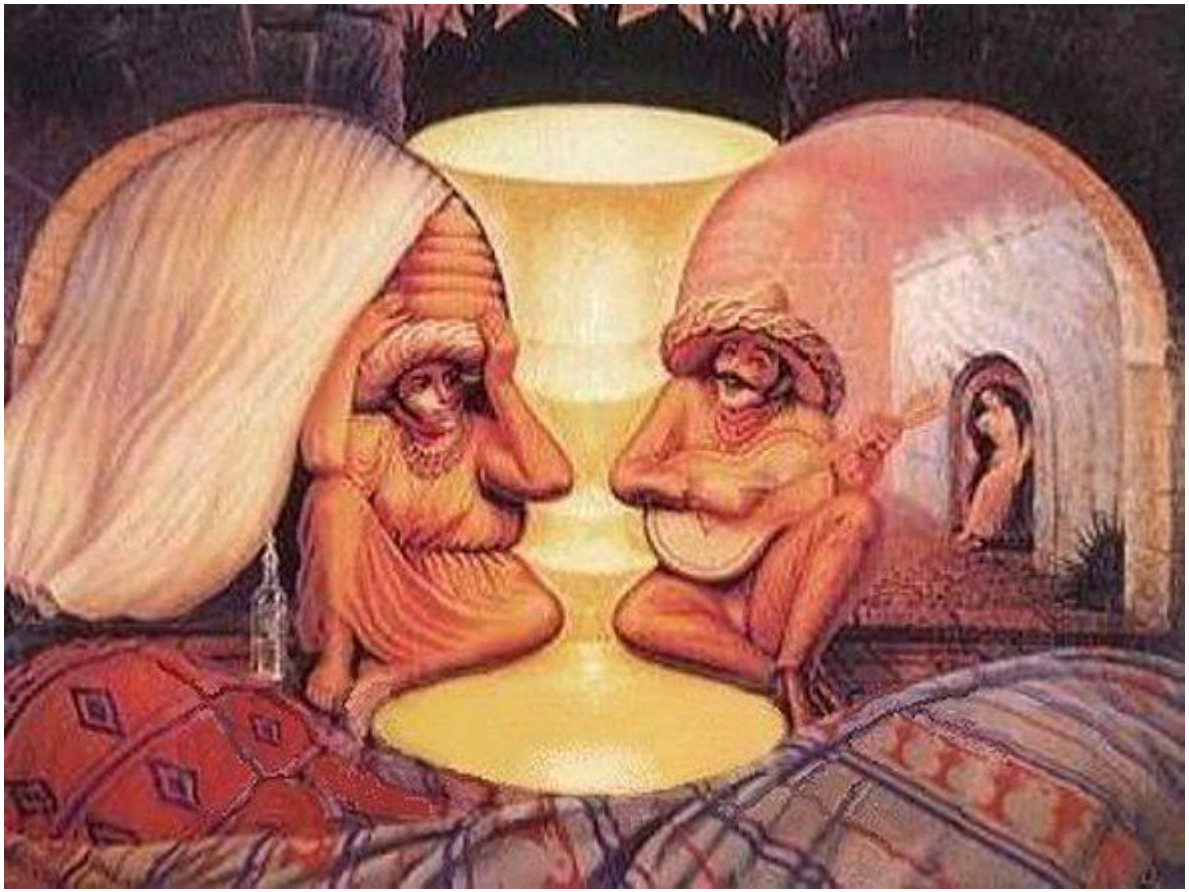
- ‘MISINFORMING people by the use of statistical material might be called statistical manipulation.’

Huff, 1954: How to Talk back to a Statistic

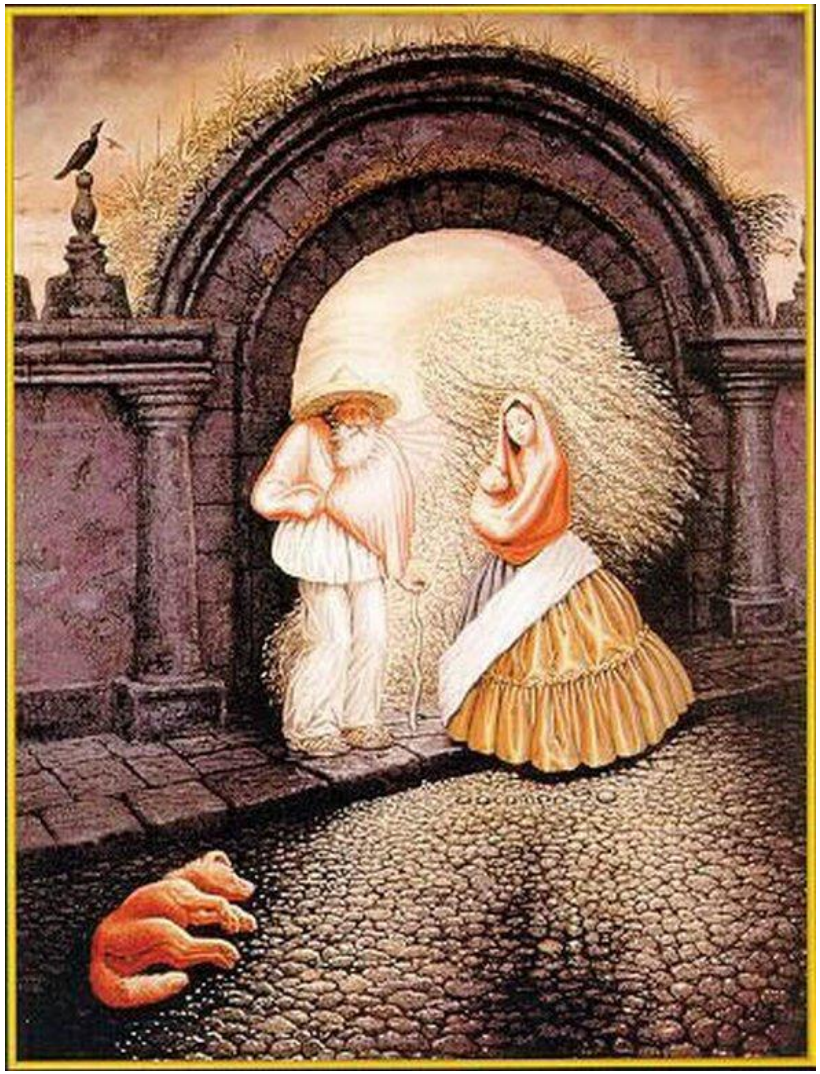
- Who says so?
- How does they know?
- What's missing?
- Did somebody change the subject?
- Does it make sense?



Salvator Dali. 1940. *Slave Market with the Disappearing Bust of Voltaire*.



Salvator Dali. 1930. *Old couple or musician*.



Salvator Dali. 1948. *Man/couple with sleeping dog*.

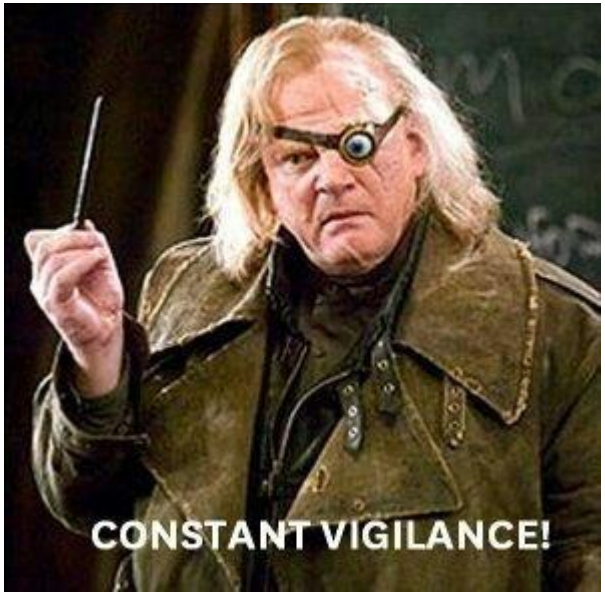
A note of caution:

"Averages and relationships and trends and graphs are not always what they seem. There may be more in them than meets the eye, and there may be a good deal less."

Huff, 1954: p.8

Важные навыки:

- если Вы смотрите на чужую статистику, подумайте, как она могла быть получена;
- если Вы проводите статистическое исследование сами,
  - будьте честны сами с собой и с читателем (помните, что все трюки всем практикующим хорошо известны);
  - продумывайте всю цепочку действий;
  - помните, что научное знание должно быть верифицируемо, воспроизводимо, фальсифицируемо;
- показатель хорошей статьи или отчёта – это возможность повторить исследование на основании прочитанного.



Литература

- Huff, Darrell. 1954. *How to Lie with Statistics*.

R

- язык статистического программирования
- плюсы:
  - ✔ широко распространен
  - ✔ хорошие возможности для визуализации
  - ✔ можно писать скрипты
- минусы:
  - ✔ овладеть в совершенстве сложно
  - ✔ медленный, если код длинный и неэффективный

R

- скачать: https://cran.r-project.org/bin/windows/base/
-
- консоль: поиграем с простыми вычислениями
- 1+2
- “Hi there, Console!”
- 2
- 4*5
- 6/3
- 2^3

R

- Console: поиграем с простыми вычислениями
- 9^0.5^3
- (9^0.5)^3
- 9^(0.5^3)
- Что Вы можете сказать про ()?

R: переменные

- x = 3
- y = 3 + 7
- x
- y
- x <- 3 + 6
- 3 + 6 -> x
- 4 ^ 3
- x = 4
- y = 3
- x^y
- Функции: sqrt(9)
- The keyboard shortcut for the assignment operator <- is **Alt + - (Windows) or Option + - (Mac)**

The mean

- **среднее арифметическое** (англ. *arithmetic mean, average*)
- сумма всех значений, деленная на их количество
- *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*
- *O*
- Какова средняя длина слов в немецком?
- В реальности, средняя длина слов в немецком 5-7 знаков
- Как Вы думаете, в русском больше или меньше?
- может быть дробным
- В чём опасность среднего арифметического?
- Найдите, как посчитать среднее в R.

- Ответ:
- >x = c(1,5)
- >mean(x)

The median

- Медиана – это значение признака, справа и слева от которого находится равное число наблюдений (по 50 %). Медиана (в отличие от среднего значения) устойчива к статистическим выбросам, то есть к резким индивидуальным отклонениям.

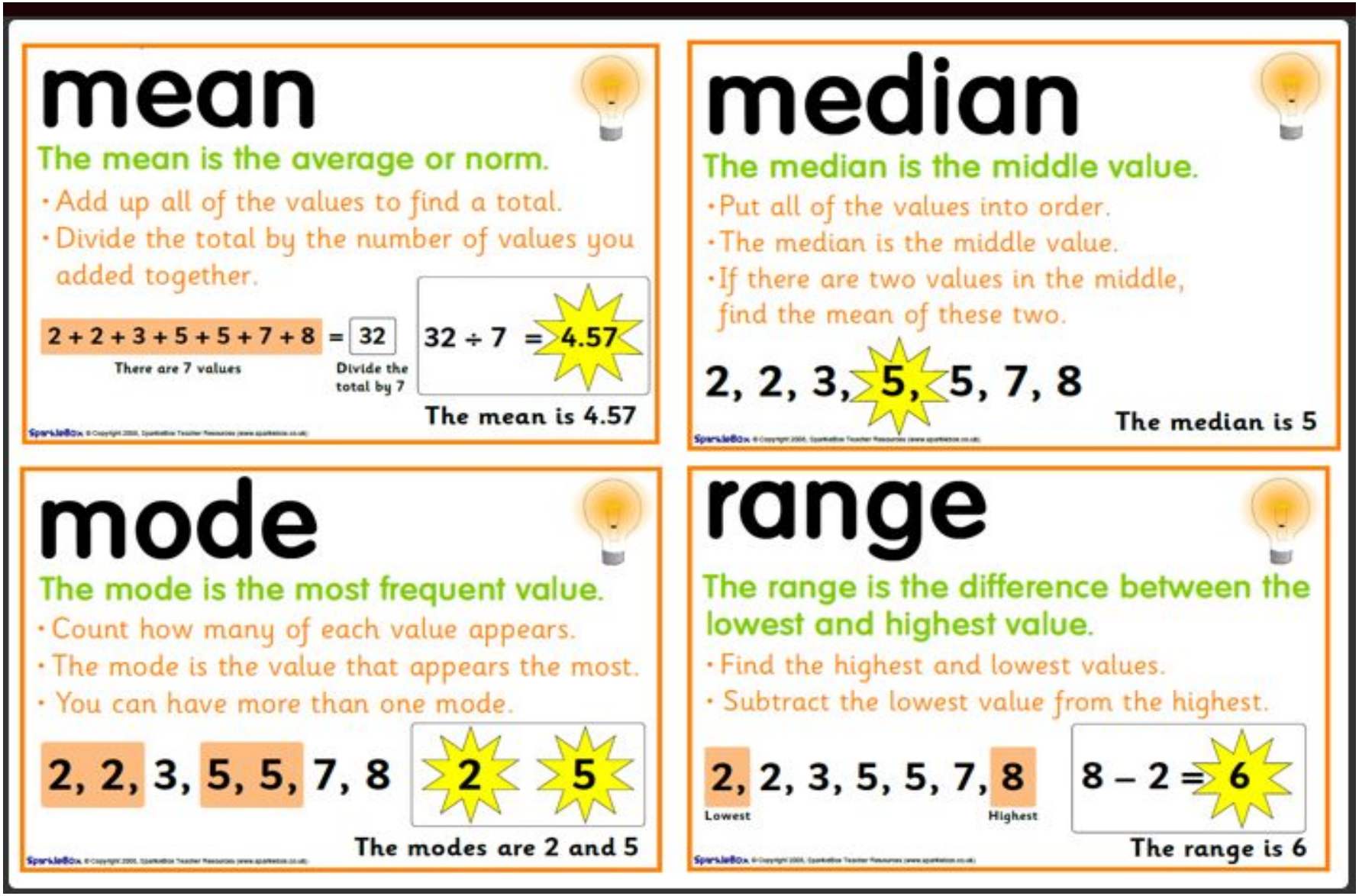
- Медиана и среднее арифметическое могут быть близки или даже совпадать, если в выборке нет выбросов.
- В R:
- >median(x)

Мода

- Величина, которая указывает не среднее, а самое часто встречающееся значение, называется **мода** (англ. *mode*).
- Использование моды особенно эффективно для анализа качественных данных, которые не могут быть сведены к среднему арифметическому.

Мода

- Распределение прилагательных мужского, женского и среднего рода в НКРЯ: 42% -- муж.род, 36% -- жен.род, 22% -- средний род.
- О чем это говорит?
- Среднее арифметическое для трех родов вычислить невозможно в силу того, что содержательно распределение по родам не может быть усреднено.
- В R нет встроенной функции для моды:
- https://www.tutorialspoint.com/r/r_mean_median_mode.htm



https://blogs.glowscotland.org.uk/re/bishopton/2017/02/17/p7-mean-median-mode-and-range/