Limits of statistical method

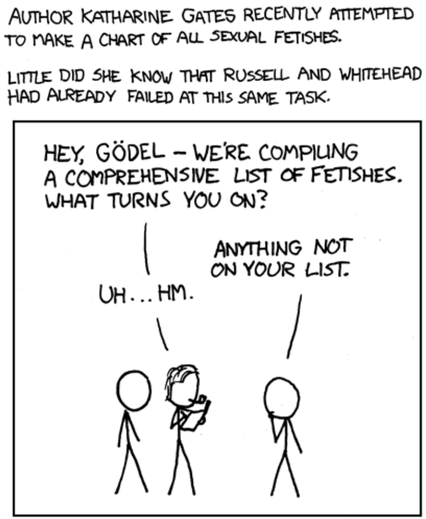
Petr Paščenko

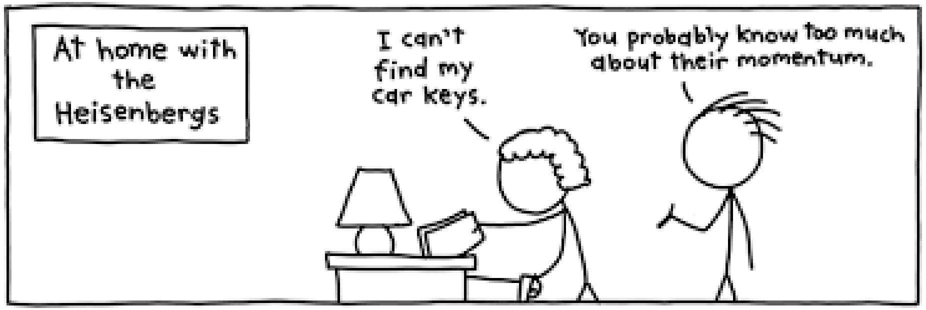
27. 11. 2023

Motivation: Limits of konwledge

It is good to know the limits…

- Second law of thermodynamics
  - *Heat does not spontaneously flow from a colder body to a hotter.*
- Gödel's incompleteness
  - *Any logical system is either consistent or complete but not both.*
- Heisenberg uncertainty principle
  - *the position and the velocity of an object cannot both be measured exactly, at the same time, even in theory*
- Many others
  - speed of light, P is not ? NP, …





Osnova

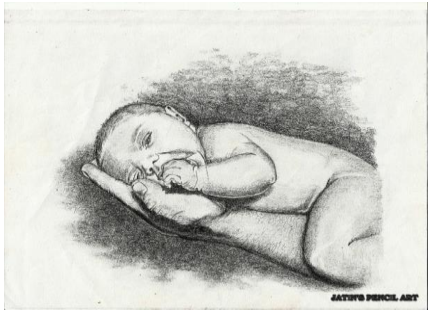
PROFINIT

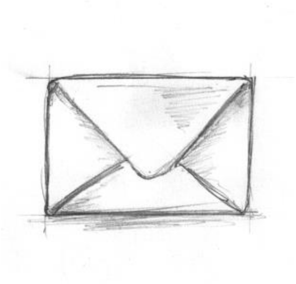
1. Mandelbrot seven states of randomness and why it matters
2. Central limit theorem assumtions
3. Correlation and causation
4. Statistical paradoxes

Mandelbrot seven states of randomness and why it matters (a lot)

10 most important days

Imagine, you delete 10 most important days from your life…



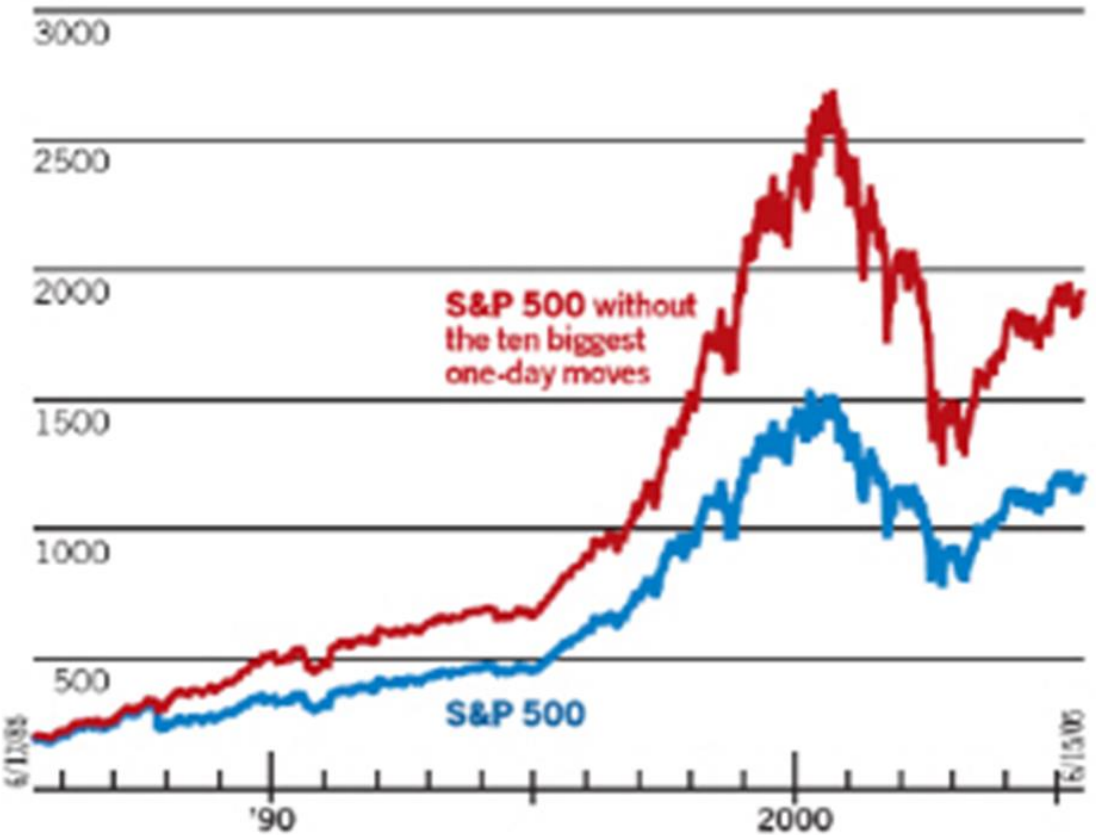


10 most important days

Imagine, you delete 10 most important days from the stock market.

- out of 20 years
  - i.e. 4000 busines days
  - i.e < 0.25%

- Good to know before you start building ML algo trading



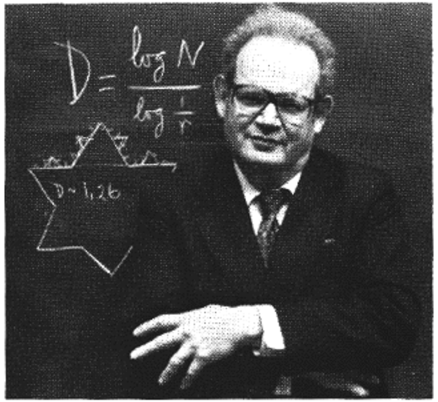
„Picking up nickels in front of steamroller“

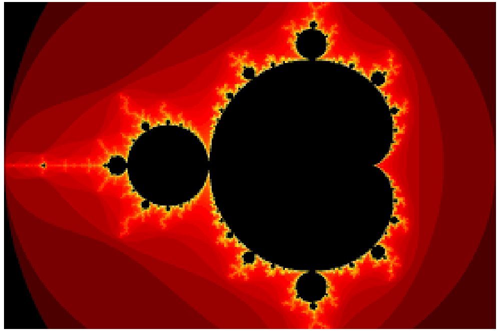
Unknown unknowns

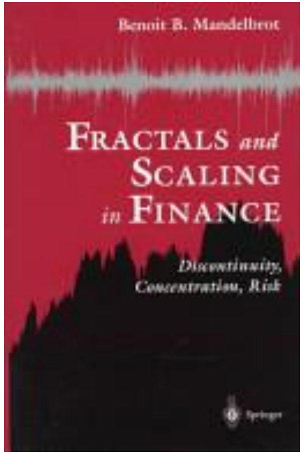
PROFINIT

Benoit Mandelbrot and Nassim Taleb

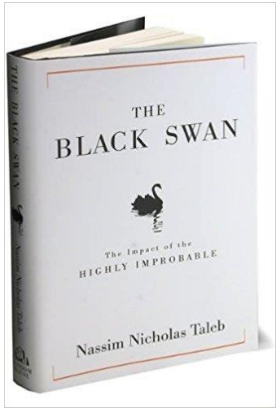
PROFINIT



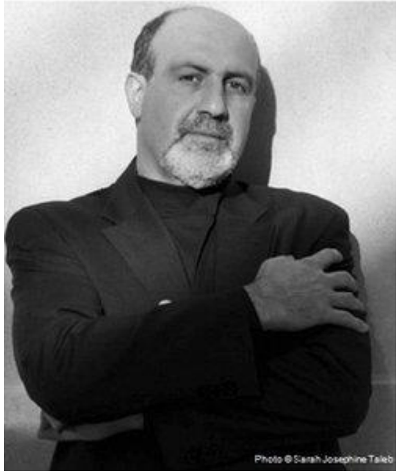




1997



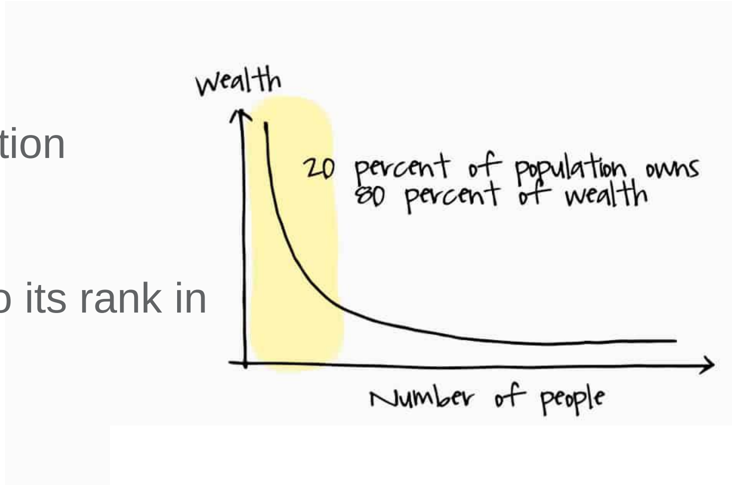
2010

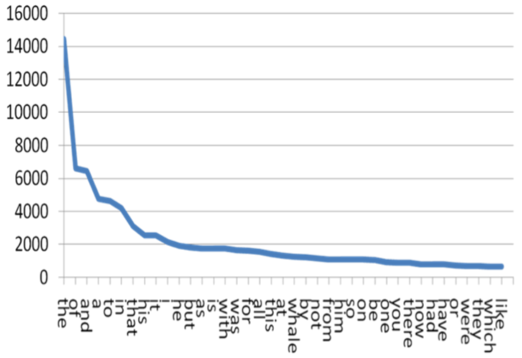


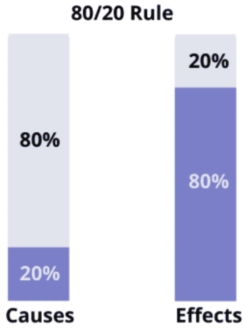
Pareto law and other empirical observations

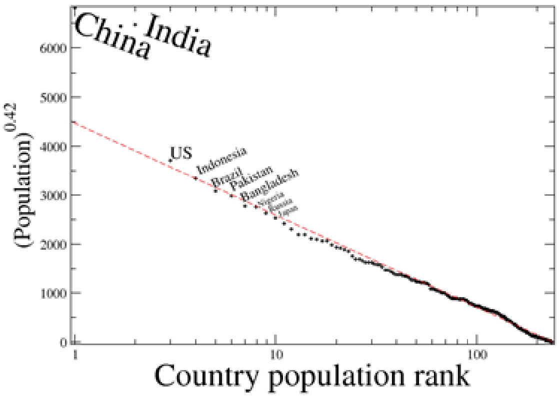
PROFINIT

- **Pareto principle** (1890 – economy)
  – 80% of Italy's land was owned by 20% of the population
- **Zipf's law** (1935 – mathematical linguistics)
  – the frequency of any word is inversely proportional to its rank in the frequency table
- **Jackson's law**: size of human settlements









Taleb: fable of two worlds

PROFINIT

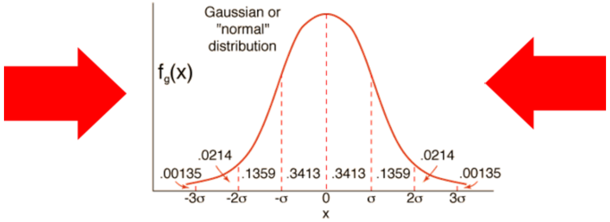
	Mediocristan	Extremistan
› Take 1000 random people on a stadium, sort them all by	**weight**	**wealth**
› Calculate average value in each group	**75 kg**	**$ 120 k**
› Add a single most **heavy** / **wealthy** person on the planet	**300 kg**	**$ 164 G**
› How does the average changed?	**75,2 kg**	**$ 164 M**
› What is the largest portion?	**4‰**	**99.93%**

The source proces on the background

PROFINIT

Mediocristan

- Evolutionary search for optimum
  - size, weight, height, etc.
  - natural panalization of extrems
  - negative feedback loop
- Aggregation
  - Central limit theorem
  - covergence to Normal distribution



Extremistan

- Winner takes all
  - join the winner
  - positive feedback loop
    - popularity, capital, gravity,…
- Matthew effect
  - *'For unto every one that have shall be given, and he shall have abundance; but him that have not shall be taken, even that which he have.'* Matthew 25:29

Key properties

PROFINIT

Mediocristan	Extremistan
Does not scale (dentist)	Does scale (google)
Physical limits	No limits
Physical measures (height)	A number (wealth)
Gaussian randomness	Power law (Pareto) randomness
Typical is close to average	No typical, no average
Winner takes a small piece	Winner takes (almost) all
Common in history	Common in current era
Black swan robust	Black swan vulnerable
Extremes can be neglected	Extremes is what matters
Easy to comprehend	Tricky to comprehend
Easy to predict	Impossible to predict
Slow gradual changes, continuity	Phase changes, discontinuities

Mandelbrot: Seven states of randomness
Key concepts

PROFINIT

- **even portioning** vs. **concentration portioning**
  - Having N random addends from a distribution, are their relative proportion of the same order of magnitude?
  - In other words, is maximum major portion of the sum?
- **scale factor of order q**
  - root of degree q of a q-th moment
  - **finite** or **infinite** moment
- **short run, long run, middle run**
  - limit theorems address long run mostly
  - short run is addressed by combinatorics
  - middle run is where most practical problems is

$$\mu'_q = \sum x_i^q \, p_i$$

$$\alpha_q = \sqrt[q]{\mu'_q}$$

30 random values from various distributions

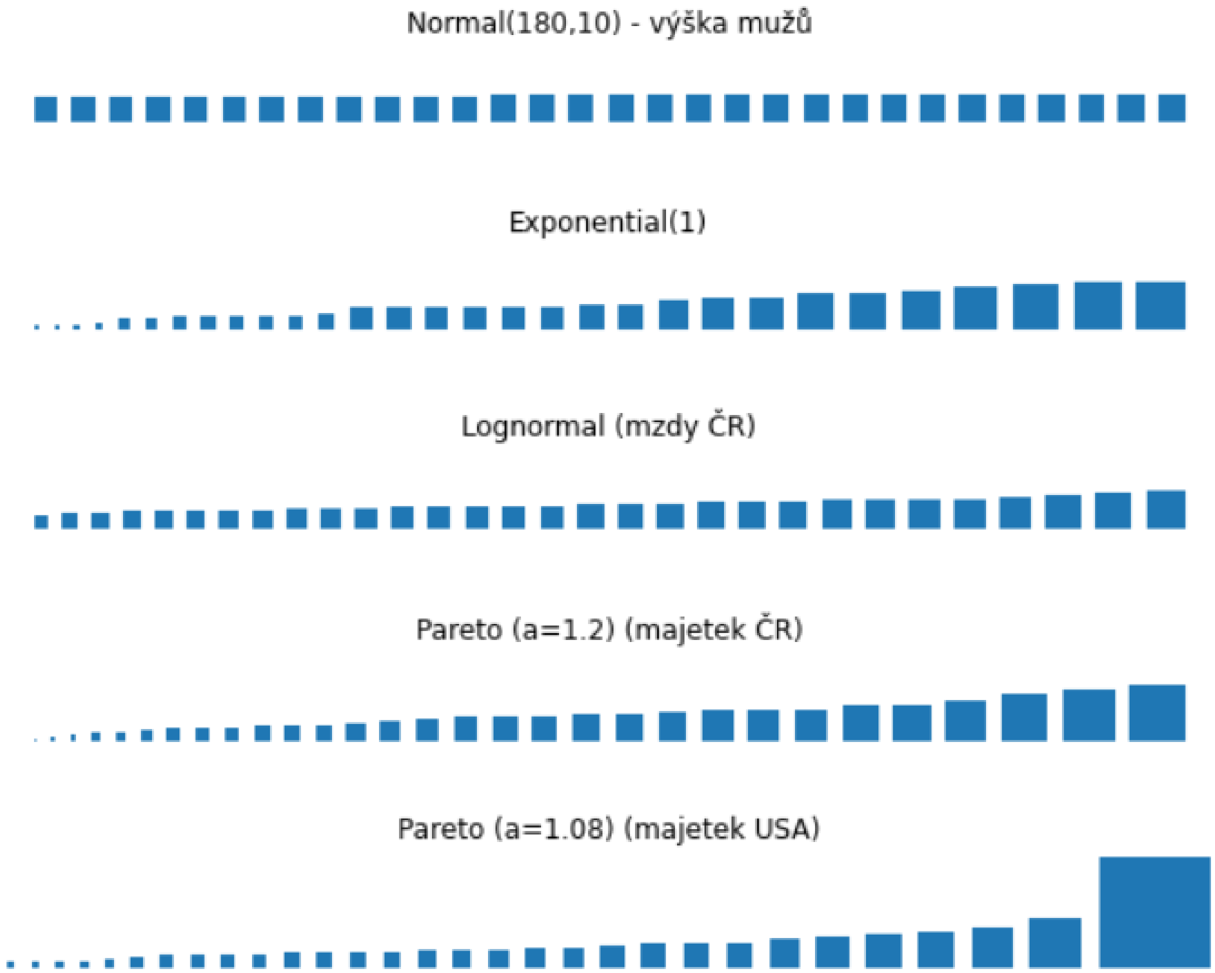
PROFINIT

1

2

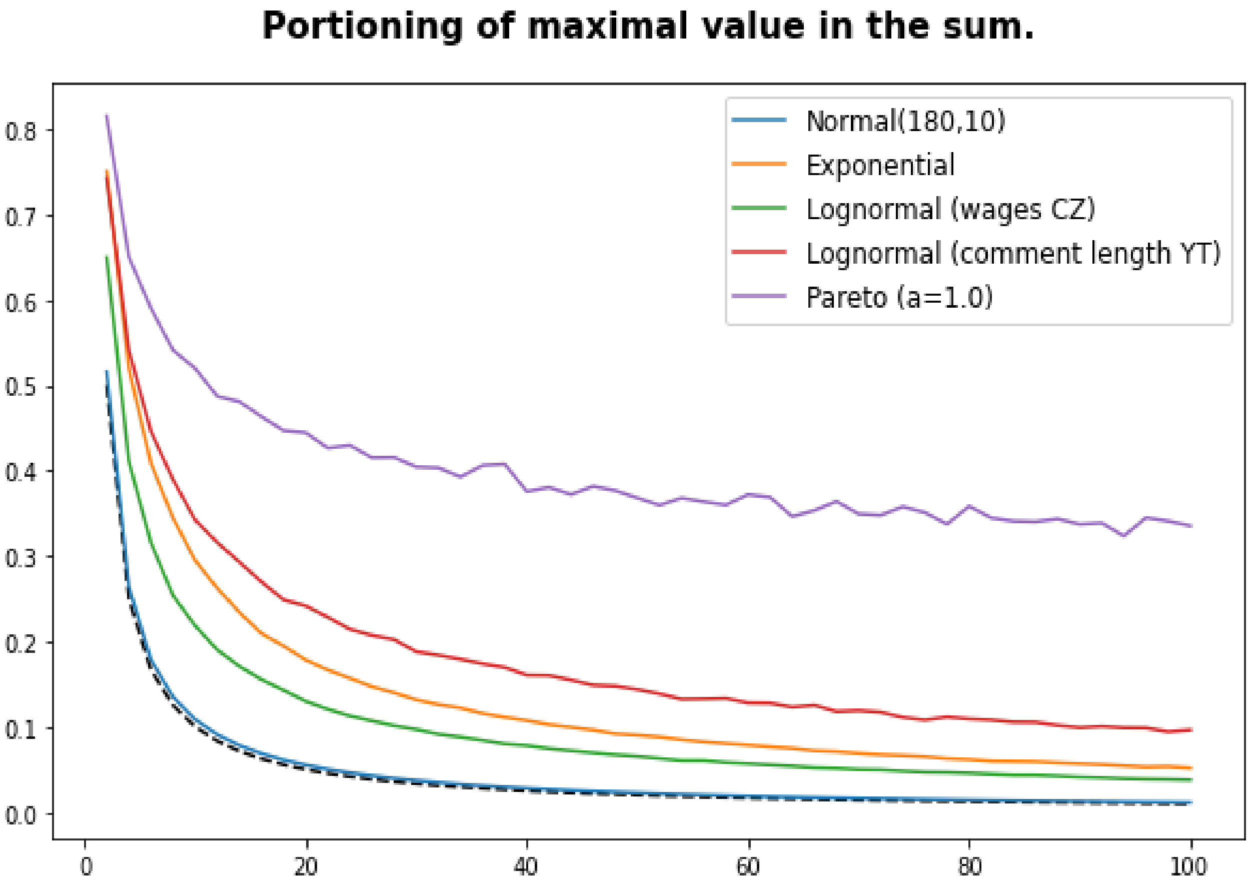
4

6



Portioning

PROFINIT

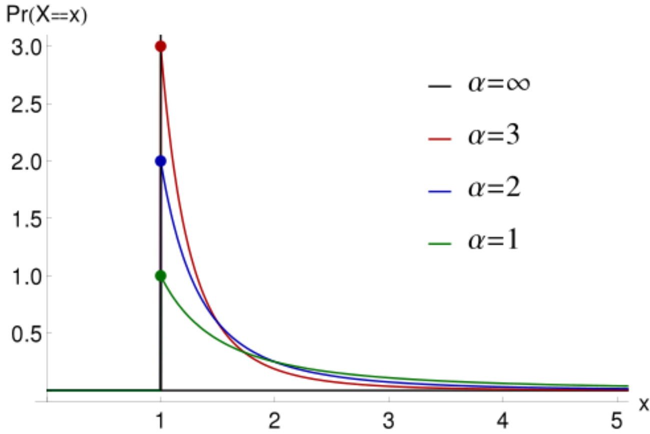


Mandelbrot: Seven states of randomness

PROFINIT

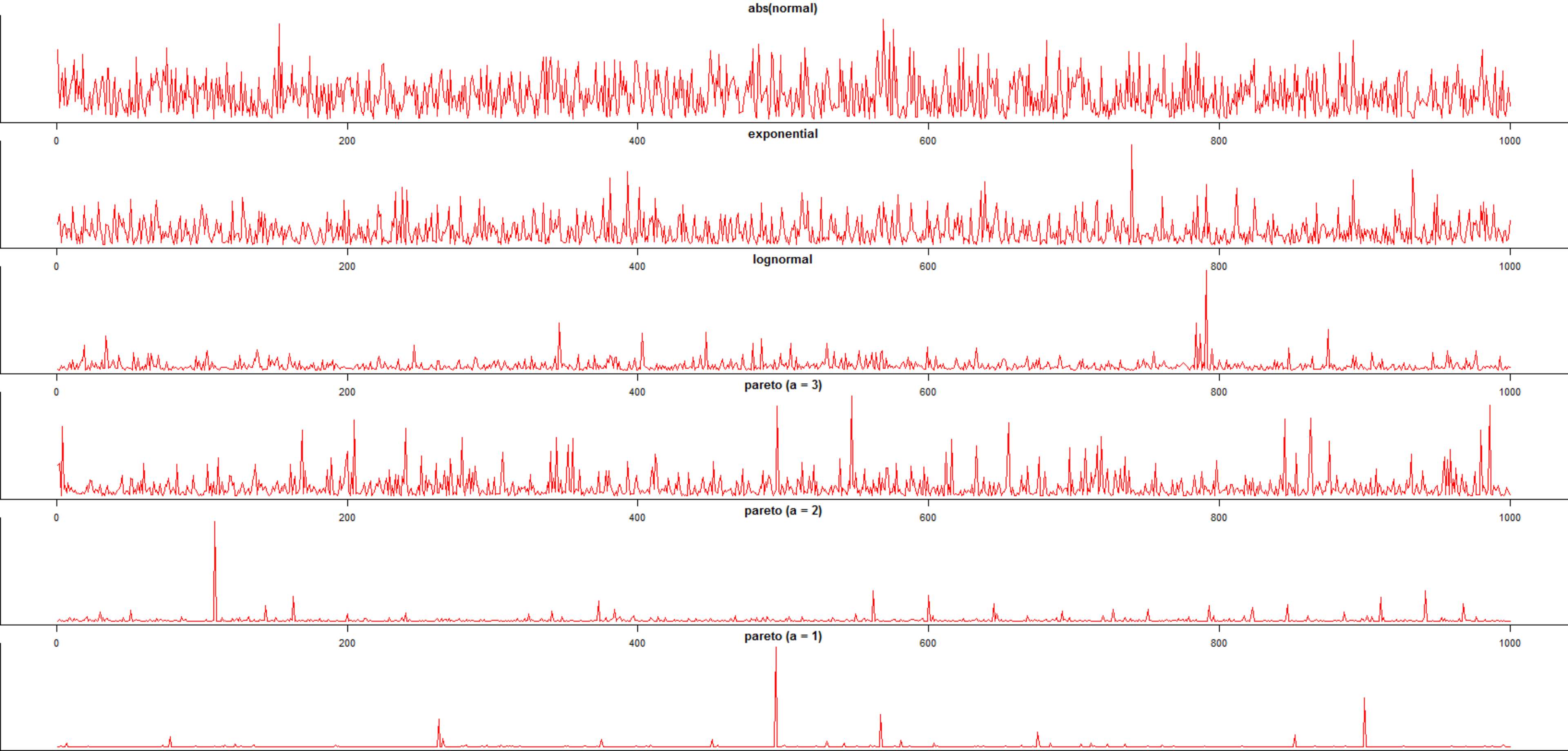
- **Mild randomness** (long term even portioning, all moments finite)
  1. **Proper mild randomness** (normal distribution)
     - Even portioning for N = 2.
  2. **Borderline mild randomness** (exponential)
     - Short term concentrated portioning, middle term even portioning
- **Slow randomness** (long term concentrated portioning, all moments finite)
  3. **Slow randomness with finite delocalized moments** $\alpha_q \in \omega(q)$
  4. **Slow randomness with finite and localized moments** $\alpha_q \in \omega(q^k)$ (lognormal)
- **Wild randomness**
  5. **Pre-wild randomness** (pareto $\alpha > 1$)
     - infinite moments for $q > 2$
  6. **Wild randomness** (pareto $0 < \alpha \leq 1$)
     - infinite variance, i.e. non convergent sample variance
  7. **Extreme randomness** (log-Cauchy $\mathrm{P}(x) \sim 1/\log(x)$)
     - infinite mean, i. e. non convergent sample mean

$$\frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}}$$



Mandelbrot: Seven states of randomness

PROFINIT



Pareto distribution, empirical parameter estimation

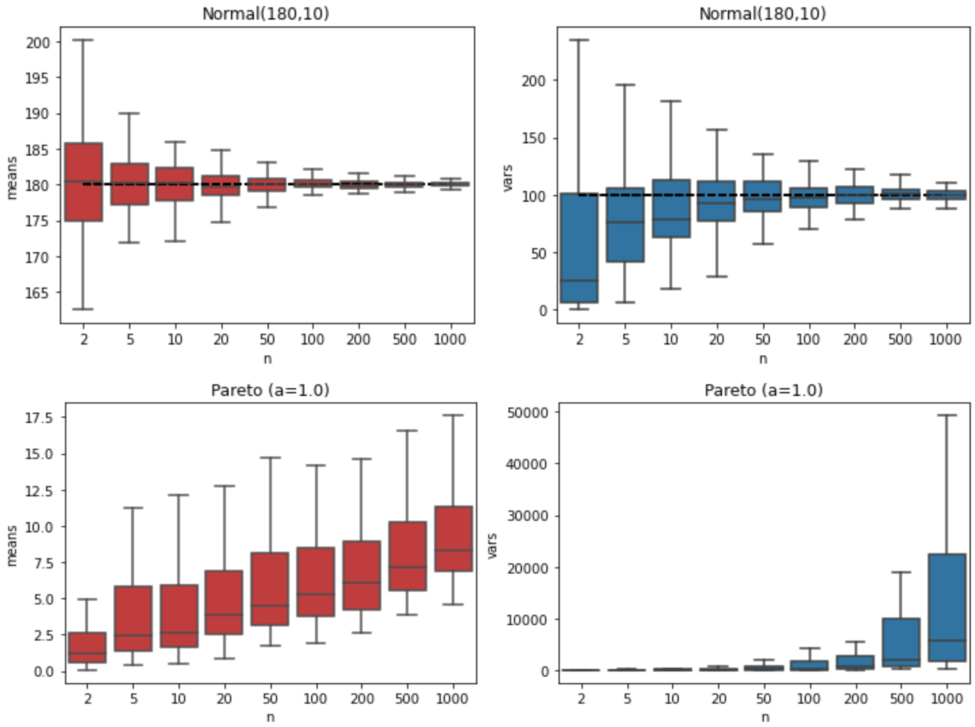
PROFINIT

Variable	Alpha
Word usage frequency	1,2
WWW visits per page (before FB)	1,4
Book title sell numbers	1,5
Earthquake magnitude	2,8
Moon crater size	2,14
Sun corona eruption sizes	0,8
War intensity	0,8
American citizen wealth	1,1
Surname frequency	1
Market movements	3 or less?
City sizes	1,3
Corporation sizes	1,5
Terrorist attack death counts	2

Consequences of wild randomness

PROFINIT

- Statistical inference does not work
  - we can not infer the parameters of distributions from data



Consequences of wild randomness

- Statistical inference does not work
  - we can not infer the parameters of distributions from data
- Central limit theorem does not work
  - we can not reduce the uncertainty by aggregation
- Prediction does not work
  - our forecast is systematically underestimated
  - our confidence interval is underestimated as well
- Black swan events
  - Unpredictable large scale events with usually negative consequences
  - Natural disasters, market crashes, political crises, epidemics, etc.
  - We can only prepare for foreseeable catastrophes

Central limit theorem assumptions

Central limit theorem and its assumptions

PROFINIT

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{N} X_i - n\mu}{\sqrt{\sigma^2 n}} \sim N(0,1)$$

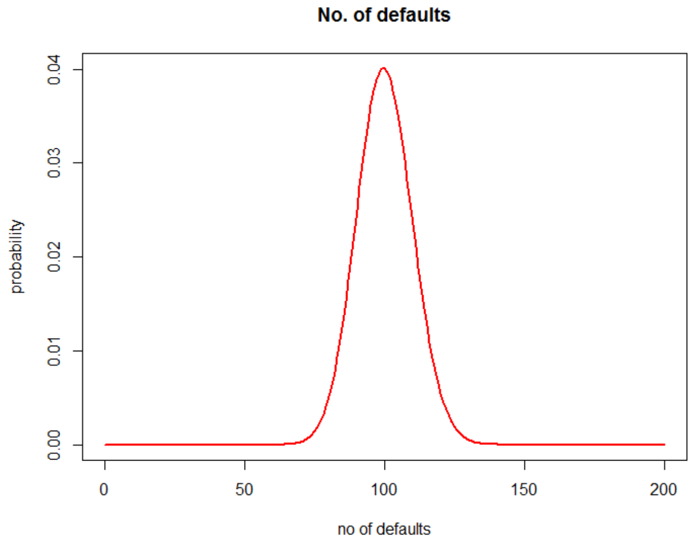
- Assumptions
  1. **X has finite mean and variance**
  2. **X is iid**
     - **independent**
       - random variables $X_1 \dots X_n$ are independent on each other
       - coins vs. sheeps
     - **identically distributed**
       - $X_1 \dots X_n$ are chosen from the same probabilistic distribution
       - there is no phase change or any other discontinuity in the process
       - almost never satisfied in practice
       - stability of model testing etc.

Example: local retail bank in a small town

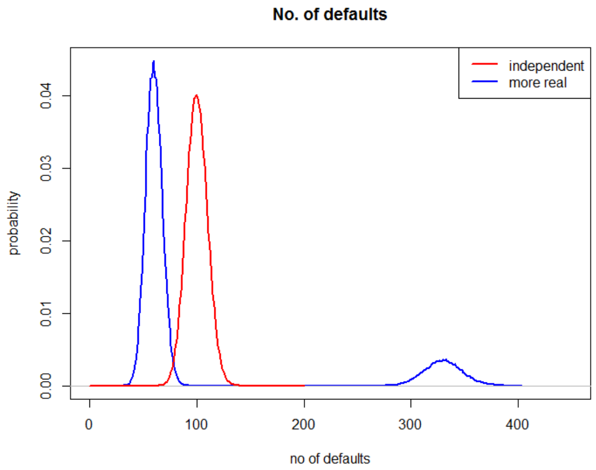
PROFINIT

- A retail bank
  - 50k people, 10k mortgages, $250k each
  - Priori probability of default is 1%
  - What is my expected worst case loss (85%, 98%, 99,9%)?
- Binomial distribution
  - $p_0 = 0.01$
  - $EX = p_0 \cdot N = 0.01 \cdot 10\,000 = 100$
  - $sd(X) = \sqrt{N \cdot p_0 \cdot (1 - p_0)} \cong 10$
  - Number of defaults in worst case:

Probability	85%	98%	99,9%
Worst Case	110	120	130



Example: local retail bank in a small town



- Small city
  - half of the people work for 1 factory
  - probability of bankruptcy 15%

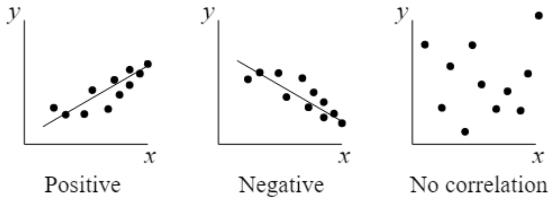
Situation		Number	Probability	Default
factory	bancrupt	5000	15%	1% → 5%
non fac.	bancrupt	5000	15%	1% → 1,6%
factory	non banc.	5000	85%	1% → 0,4%
non fac.	non banc.	5000	85%	1% → 0.8%

  - EX = $5000 \cdot (0.15 \cdot (0.05 + 0.016) + 0.85 \cdot (0.04 + 0.08) = 100$

Probability	85%	98%	99,9%
Worst Case independent	110	120	130
Worst Case real	66	357	375

Statistical paradoxes

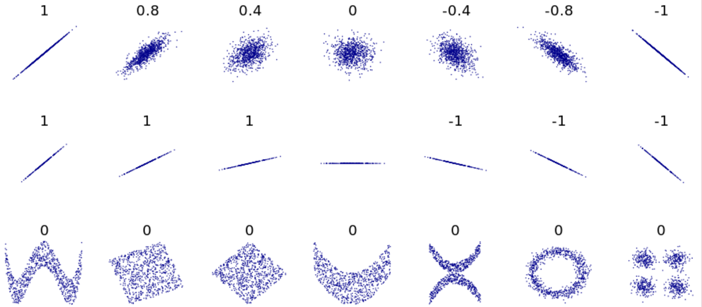
Correalation and Causation



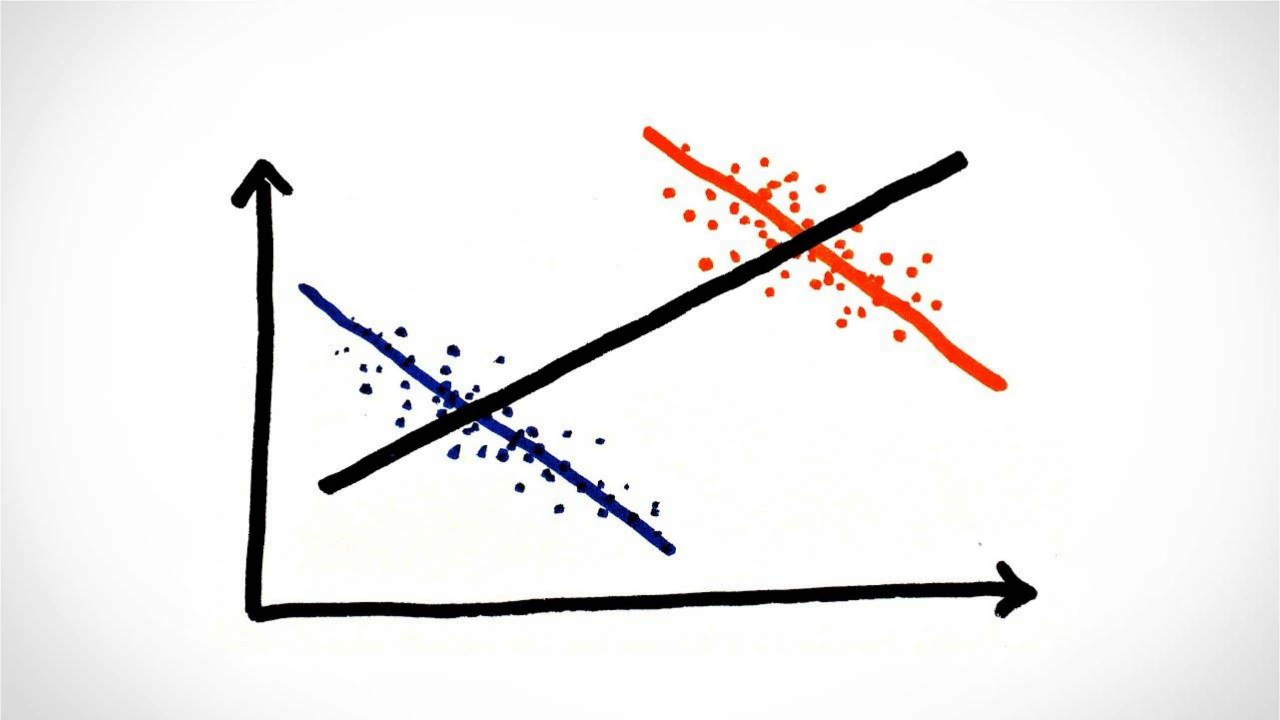
- Correlation
  - linear dependency of variables: A and B
- Causation = any form of dependence
  - visiting lectures implies passing the exam
- Correlation does not imply causation

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

- Correlation without causation
- Causation without correlation



Simpson paradox



- smoking versus life expectancy for male and female

Simpson paradox

PROFINIT

- What is the vaccine effectivness?

$$e = 1 - \frac{P(S|V)}{P(S|\neg V)}$$

$$e = 1 - \frac{\frac{5{,}3}{100\,000}}{\frac{16{,}4}{100\,000}} = 67{,}5\%$$

Severe cases		Efficacy
Not Vax per 100k	Fully Vax per 100k	vs. severe disease
214 16.4	301 5.3	**67.5%**

Simpson paradox

PROFINIT

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

- The classes are imbalaced
  - in both severe cases and vaccination

Diskuze